Universidade do Minho
Escola de Engenharia

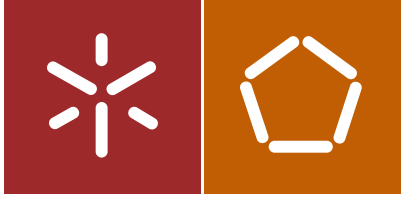
César Augusto da Silva Martins

Arquitetura de um Sistema de Análise de Dados Big Data no Modelo Cloud Computing

outubro de 2014

César Augusto da Silva Martins Arquitetura de um Sistema de Análise de Dados Big Data no Modelo Cloud Computing

UMinho | 2014



Universidade do Minho
Escola de Engenharia

César Augusto da Silva Martins

Arquitetura de um Sistema de Análise de Dados Big Data no Modelo Cloud Computing

Dissertação de Mestrado
Mestrado em Sistemas de Informação

Trabalho efetuado sob a orientação do
Professor Doutor Jorge Oliveira e Sá

outubro de 2014

Declaração

Nome: César Augusto da Silva Martins

Endereço eletrónico: cesar.martins.1988@gmail.com

Telefone: 915957760

Número de Cartão de Cidadão: 13360031 9 ZY3

Título dissertação/tese

Arquitetura de um Sistema de Análise de Dados Big Data no Modelo Cloud Computing.

Orientador: Professor Jorge Oliveira e Sá

Ano de conclusão: 2014

Designação do Mestrado: Mestrado em Sistemas de Informação

Universidade do Minho, 30 de outubro de 2014

Assinatura

Agradecimentos

A realização deste trabalho foi possível devido a disponibilidade da empresa Cloud365, que proporcionou a oportunidade de desenvolvimento deste projeto. É importante também salientar o apoio de todos os meus familiares e amigos que me incentivaram desde o início, bem como agradecer ao meu orientador, Professor Jorge Oliveira e Sá e ao Engenheiro Paulo Simões da Cloud365, por acreditarem no meu trabalho e nas minhas capacidades.

Resumo

Big Data está a provocar uma mudança na forma como as organizações e os recursos dentro das mesmas estão a trabalhar. As oportunidades que são criadas constituem novas alternativas para o negócio, influenciando o processo de tomada de decisão e estratégia das organizações no que diz respeito ao trabalho a realizar num futuro próximo. De facto a estratégia das organizações e o seu comportamento no mercado está muito dependente da enorme quantidade de informação que possuem armazenadas, mas também das informações que são obtidas através das redes sociais, e de dispositivos móveis, bem como das expetativas que os clientes têm em relação ao futuro das organizações.

O *Big Data* procura utilizar as grandes quantidades de informação que as organizações possuem e conseguem gerar, procurando organizá-las e utilizá-las como meio de suporte consistente, viável e capaz de se assumir como basilar no processo de tomada de decisão. Mas para que tudo isto seja possível é necessária uma adaptação a esta nova realidade. Este trabalho procura apresentar uma abordagem para o desenvolvimento de uma arquitetura para um sistema de análise de dados que seja capaz de dar resposta as necessidades atuais e futuras das organizações. A possibilidade de escolha do modelo *Cloud Computing* deve-se ao facto de as soluções de Cloud Computing permitirem uma elasticidade e uma disponibilidade que permite grande flexibilidade as organizações, devido ao benefício de se poderem utilizar em qualquer lugar e a qualquer hora, fazendo com que as organizações utilizem a informação quando dela necessitarem.

Tal como foi referido acima, o objetivo deste projeto é o estudo e desenvolvimento de uma arquitetura de um sistema de análise de dados *Big Data* no modelo *Cloud Computing*. Pretende-se também que este projeto possa auxiliar os profissionais que trabalham ou possam vir a trabalhar num futuro próximo com sistemas de análises de dados, pois podem encontrar nesta arquitetura um guia para a ajuda na procura de soluções, para colmatar os problemas que estão associados à seleção, adaptação e utilização das abordagens relacionadas com *Big Data* e que têm como base o modelo *Cloud Computing*.

Palavras-Chave: *Big Data,* Computação em Nuvem, Arquiteturas de Sistemas de Análises de Dados*,* Arquiteturas de Sistemas de Inteligência de Negócios, Arquitetura de Sistemas de Computação em Nuvem, Análises Big Data, Tecnologias de Análises de dados Big Data.

Abstract

Big Data is changing the way organizations and their resources are working. The opportunities created are new alternatives to do business and influence the decision making process and the strategy of organizations in what concerns the work that will be realized in future. In fact, strategy of organizations and their behavior in market depends on information's stored in their data bases. It also depends on social networks, mobile tools or costumer's expectations.

In fact, Big Data main purpose is take advantage of big volumes of information that organizations obtain through several means, organizing them and using them to create a consistent and a reliable basis to help decision make process.

However, to make all possible it is necessary an adaptation to a new reality. This work intends to present an approach to development an architecture concerning analysis system using Cloud Computing model. The choice of Cloud Computing model has several advantages, namely it can be used in any place at any hour by organization without need of any investment in infrastructures, once it is ready to be used.

The main purpose of this work is creating Big Data analytics by using Cloud Computing model. Finally, I hope this work helps who want to work with Big Data analytics in Cloud Computing model.

Keywords: *Big Data, Cloud Computing, Architecture of Database Analytical Systems, Architecture of Business Intelligence Systems, Cloud Computing Systems Architecture, Big Data Analytics, Technologies for Big Data.*

Índice

Índice de Figuras

Índice de Tabelas

Acrónimos e Abreviaturas

O presente documento apresenta e utiliza acrónimos localmente no âmbito de cada um dos capítulos que compõem o documento da dissertação.

- AWS – *Amazon Web Services*
- BI – *Business Intelligence*
- BGaaS – *Big Data as a Service*
- CEO – *Chief Executive Officer*
- COBIT – *Control Objectives for Information and Related Technology*
- CSP – *Cloud Service Provider*
- CSV – *Comma Separated Values*
- DBA – *Database Administrator*
- DBMS - *Database Management System*
- DER – *Diagrama de Entidades e Relacionamentos*
- DSR – *Design Science Research*
- ERP – *Enterprise Resource Planning*
- ETL – *Extract Transformation and Loading*
- GQM – *Goal Question Metric*
- HDFS – Hadoop *Distributed File System*
- HTTPS - *Hyper Text Transfer Protocol Secure*
- IaaS - *Infrastructure as a Service*
- ITIL – *Information Technology Infrastructure Library*
- JSDL – *Job Submission Description Language*
- KPI - *Key Performance Indicators*
- LDAP – *Lightweight Directory Access Protocol*
- NIST – *National Institute Standards and Technology*
- NoSQL – *Not Only Structured Query Language*
- ODBC – *Open Data Base Connectivity*
- OWL - *Web Ontology Language*
- PaaS – *Platform as a Service*
- PBS – *Product Breakdown Structure*
- RAM – *Random Access Memory*
- RDF - *Resource Description Framework*
- SaaS – *Software as a Service*
- SGBD - *Sistema de Gestão de Base de Dados*
- SOA – *Service Oriented Architecture*
- SQL – *Structured Query Language*
- SLA – *Service Level Agreement*
- SPI - *Service Provider Infrastructure*
- SSL – *Secure Socket Layer*
- SSO – *Single Sign On*
- TI – *Tecnologias de Informação*
- UDF – *User Defined Function*
- VPN – *Virtual Private Network*
- WBS - *Work Breakdown Structure*
- XML - *Extensible Markup Language*
- WSLA – *Web Service Agreement Language*

1 – Introdução

1.1 Descrição do problema

A Cloud365 é uma empresa relativamente recente, fundada em 2011 e situada no Polo Tecnológico do Lumiar em Lisboa. O foco da sua atividade reside na consultoria e implementação de soluções de *Cloud Computing*, possuindo também parcerias com importantes empresas da área da Tecnologias de Informação, como é o caso da Microsoft e da Colt. O tema de dissertação, "Desenvolvimento de Uma Solução de Análise de Dados Big Data ", procura perceber se é possível efetuar análises de grandes volumes de dados provenientes de diversas origens e diferentes estruturas de dados. Procurando viabilizar e validar uma arquitetura de suporte a uma solução tecnológica, que poderá trabalhar no ambiente de *Cloud Computing*, ponderando, num futuro relativamente próximo, o desenvolvimento de uma nova área de negócio. Dessa forma a Cloud365 propôs este projeto de dissertação.

1.2 Objetivos do Projeto

Após a definição do problema é necessário encontrar soluções para o mesmo. Assim impõe-se uma revisão da literatura, de modo a perceber o estado de arte do problema, procurando informação que permita descreve-lo e sustentar a resolução do mesmo.

Este levantamento do estado da arte também deverá ter em conta as tecnologias que existem no mercado, no sentido de perceber quais são as suas mais-valias e limitações. Depois de compreendidos e percebidos estes aspetos, é necessário proceder-se à escolha das tecnologias, podendo este processo integrar a realização de testes com o intuito de verificar como as diversas tecnologias de diferentes fornecedores podem ser integradas na solução desenvolver.

Assim, com a realização deste projeto pretende-se atingir os seguintes objetivos:

- Propor uma arquitetura para um sistema de análise de dados Big Data
- Identificar as competências necessárias para trabalhar com o *Big Data*.
- Identificar as tecnologias que podem ser utilizadas como o suporte a análise de dados Big Data.

1.3 Resultados Esperados

O resultado que se espera deste projeto de dissertação é o desenvolvimento de uma arquitetura, que permita fazer análise de dados *Big Data*. Dessa forma será desenvolvido um protótipo que permita validar a arquitetura da solução a desenvolver.

1.4 Planeamento do Projeto

A execução de um trabalho de investigação necessita de ser planeado, de forma a organizar as tarefas a executar, é necessário definir as tarefas a executar bem como o tempo que é necessário para a realização de cada uma delas, o planeamento do projeto pode ser consultado no anexo A – Planeamento do Projeto.

1.5 Abordagem Metodológica

A realização de um trabalho de investigação faz com que o seu autor deva ter em conta algum cuidado especial, na elaboração de todo o trabalho, procurando atingir os objetivos propostos, dentro do tempo disponível. A primeira tarefa a realizar, é a construção de um plano de trabalhos, que descreva o conjunto de etapas e atividades a executar durante o projeto de dissertação.

Neste sentido foi definida a seguinte questão de investigação: "De que forma uma solução de análise de dados Big Data pode ajudar a no processo de tomada de decisão?".

O planeamento de trabalho deve incluir uma metodologia de investigação que, inclui as etapas a executar, em conjunto com uma análise em relação às várias metodologias identificadas como passíveis de serem utilizadas/desenvolvidas no âmbito da dissertação. Após a análise, procede-se à seleção da metodologia que se considera mais adequada tendo em conta a natureza do problema definido.

É verdade que esta é uma dissertação que possui uma grande componente tecnológica, mas para perceber de que forma a componente tecnológica deve ser incorporada, é necessário realizar alguma investigação, nomeadamente para perceber o âmbito do problema e realizar pesquisas bibliográficas que se adequem ao tema em estudo. Após este trabalho, é necessário fazer o *design* da arquitetura e criar um protótipo que permita a sua implementação. Devido à complexidade deste trabalho e também devido aos recursos que a metodologia *Design Science Research* (DSR) nos oferece para lidar com estes problemas, optou-se por esta metodologia como suporte a todo o trabalho a realizar.

Com base neste contexto fez-se a opção pela metodologia *DSR*, que permite formalizar a estratégia a adotar, com vista à realização dos objetivos do projeto de dissertação, facilitando e promovendo, simultaneamente a compreensão do processo de investigação, desde a definição e contextualização do tema em estudo até à entrega da dissertação. Na tabela abaixo são apresentadas as atividades que serão realizadas:

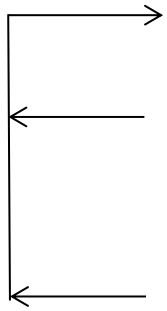
Fluxos de Conhecimento	Atividades do Processo	Metodologia e Técnicas	Resultados
	Definição do problema e idealizar uma solução	Pesquisa bibliográfica (ex.: matriz de conceitos) Análise de Literatura	Proposta e Tentativa de *Design*
	Desenvolvimento e implementação de uma solução	Pesquisa bibliográfica; Análise de literatura, Análise Critica; Modelos e *frameworks* conceptuais	Arquitetura de Sistema de Análise de Dados *Big Data* no Modelo *Cloud Computing*.
	Avaliação da solução e elaboração da conclusão	Principais Opiniões e comentários	Análise dos resultados obtidos, recomendações e conhecimentos adquiridos.

Tabela 1 Metodologia utilizada, adaptado de (Vaishnavi & Kuechler Jr., 2008)

Em seguida é apresentada de uma forma resumida a descrição das atividades associadas ao projeto de dissertação:

A primeira atividade define o problema procurando idealizar uma solução, É necessário existir uma clara definição do problema e um conjunto de possíveis soluções. O objetivo desta atividade é responder às seguintes questões: "Que tema abordar? Qual o problema a resolver? Quais as atividades necessárias para resolver esse problema? Quais os resultados esperados tendo em conta os objetivos a atingir?". De forma a concretizar esta atividade é necessário, realizar uma revisão da literatura, sobre o problema em questão (Webster & Watson, 2002).

A atividade anteriormente descrita é bastante importante. Em primeiro lugar, permite a realização de uma pesquisa e de um estudo sobre o tema, que ajudam na definição da estratégia a seguir no decorrer do projeto de investigação. Em segundo lugar possibilita uma definição sustentada, que será a base para a elaboração desta dissertação, por fim permite a criação de um ponto de referência que ajudará na elaboração da estratégia de gestão de riscos.

A segunda atividade foca-se no desenvolvimento de uma arquitetura conceptual, que depois será testada recorrendo a uma proposta de arquitetura tecnológica, com a implementação de um protótipo da solução A realização do primeiro objetivo necessita de uma revisão da literatura, esta deve estar focada nos conceitos de *Big Data, Cloud Computing, Architecture of Cloud Computing and Big Data Systems* (Webster & Watson, 2002). O segundo objetivo implica um estudo sobre tecnologias e arquiteturas conceptuais de sistemas de análise de dados existentes no mercado, de modo a perceber os requisitos (ex.: infraestruturas e necessidades de hardware), modo de funcionamento e implementação. Após a realização deste estudo será possível definir uma arquitetura conceptual da solução a desenvolver. Em

relação ao último objetivo, o mesmo depende da aprovação da arquitetura conceptual definida no segundo objetivo, só após a validação da mesma por parte da empresa Cloud365, se irá proceder à instanciação de uma arquitetura tecnológica que permita a implementação de um protótipo funcional recorrendo a tecnologias de análise de dados *Big Data*, e que funcionem no modelo *Cloud Computing*.

Por fim a terceira atividade, consiste na avaliação da solução e na elaboração da conclusão, que envolve a realização de uma análise crítica da solução alcançada e de todo o trabalho efetuado, procurando uma combinação dos resultados esperados com os resultados obtidos, daí resulta um conjunto de recomendações.

É esperado que o protótipo desenvolvido possa num futuro próximo constituir a base para uma solução comercial. O resultado final de todas as atividades a desenvolver será o documento do projeto de dissertação.

1.6 Estrutura do Documento

O presente documento define as linhas de orientação para o projeto de dissertação, com o objetivo de conseguir expor o assunto em estudo. A próxima etapa é a realizar é a descrição da organização do presente documento com uma sintetização dos capítulos que o constituem.

O Capítulo 1 iniciou-se com a contextualização do tema de dissertação e a importância do mesmo. Foram descritos os objetivos atingir e os resultados esperados. Foi descrita a abordagem metodológica a adotar como suporte para a realização de todo o trabalho. Finalmente descreveu-se a organização do documento, sendo então, apresentados os assuntos a abordar em cada um dos capítulos que constituem o projeto de dissertação.

O Capitulo 2 descreve uma perspetiva sobre os desafios atuais que são colocados ao Big Data e de que forma esses desafios podem ser superados. Este capítulo está organizado em três secções, a primeira secção apresenta uma visão sobre o estado de arte atual do *Big Data*, a segunda secção apresenta uma descrição sobre o estado de arte atual do *Cloud Computing*., na terceira secção é apresentada uma articulação entre o *Big Data* e o *Cloud Computing*, procurando perceber a importância e o impacto que estes temas terão no futuro das organizações.

O Capítulo 3 está organizado em três secções, a primeira secção apresenta a proposta da arquitetura conceptual com a descrição dos vários níveis associados a mesma, a segunda secção apresenta uma proposta de uma arquitetura física esta proposta tem como base os níveis e tarefas descritas na arquitetura conceptual, dando enfoque a uma possível solução da arquitetura conceptual com instanciação através de tecnologias que poderão ser o suporte tecnológico da arquitetura conceptual idealizada para a solução.

O Capítulo 4 foca-se na apresentação de alguns dos caso de teste que foram realizados, o seu objetivo é viabilizar a arquitetura desenvolvida. Desta forma através das análises realizadas consegue-se perceber a viabilidade da arquitetura e identificar quais as melhorias que necessitam de ser efetuadas. Os casos de teste que são apresentados abrangem diferentes áreas como a saúde, as telecomunicações e as compras online, procurando demonstrar a adequação da solução desenvolvida a diferentes contextos.

O Capítulo 5 apresenta as principais conclusões de todo o trabalho que foi realizado, com principal incidência nos resultados que foram obtidos nos caso de teste, procurando explicar de que forma o trabalho realizado pode ser aproveitado para futuros projetos. É também definido qual o trabalho futuro que deveria ser realizado e de que forma esse trabalho futuro poderia ajudar a melhorar o trabalho realizado com este projeto de dissertação.

2 – Estado da Arte de *Big Data* e *Cloud Computing*

Com o propósito de apresentar e clarificar alguns dos conceitos fundamentais no âmbito do trabalho de investigação a desenvolver, foi necessário realizar uma revisão da literatura, em que o resultado é descrito neste documento. O resultado obtido possibilitou a criação de uma base sólida para permitir o avanço do conhecimento e a realização dos objetivos de compreensão do tema em estudo, o que se constitui como um ponto de partida para a realização dos demais objetivos.

Para poder organizar a informação que foi recolhida na revisão de literatura foram seguidas as linhas de pensamento de Webster e Watson (2002). A matriz de conceitos resultante dessas sugestões pode ser consultada no anexo B – Matriz de Conceitos.

A secção 2.1 procura expor os conceitos de *Big Data* que se entendem por fundamentais para o trabalho ajudando a clarificar a área em estudo. Na secção 2.2 são analisados e explicados os conceitos de *Cloud Computing*. A secção 2.3 procura destacar a ligação entre os conceitos de *Big Data* e *Cloud Computing*, procurando explicar de que forma ambos os conceitos em conjunto podem oferecer como vantagens competitivas para os utilizadores quer de serviços quer de recursos de Tecnologias de Informação (TI).

2.1 *Big Data*

Todos os dias as organizações geram e têm à sua disposição uma elevada quantidade de dados, o que cria a dificuldade em maximizar os proveitos que podem ser obtidos com a análise dessa riqueza de dados. Por vezes devido à falta de conhecimento dentro das organizações no que respeita ao processamento da informação ou seja a recolha e disponibilização dos dados para processos analíticos, existe uma perda de capacidade no processo de suporte de decisões quer estas sejam de nível técnico ou tático. Todos os dias as organizações são confrontadas com dados provenientes de (Kimball & Ross, 2013):

- Aplicações de gestão tais como *Enterprise Resource Planning* (ERP), *Customer Relationship Management* (CRM), *Supply Chain Management* (SCM), entre outros;
- Folhas de cálculo e documentos, que podem ser internos ou externos à organização;
- Redes sociais, (o que se fala sobre a organização), imagens (normalmente de produtos acabados semiacabados e matérias primas, e vídeos (por exemplo explicativos da utilização de produtos...);
- Sensores acoplados a dispositivos eletrónicos, como por exemplo sensores acoplados a máquinas fabris (temperatura, vibração, interrupções em tempo e códigos de paragens, horas trabalhadas, etc.), alarmes (de intrusão, falhas de algum mecanismo, etc.), câmaras de vídeo, registos de eventos (*logs*) de *routers* e *firewall*s e
- Dispositivos móveis que são cada vez mais adotados nas organizações permitindo que a capacidade de processamento esteja cada vez mais distribuída.

Assim emerge a necessidade de uma solução que permita a recolha e integração de dados (com tipos distintos), provenientes de várias fontes mas para além disso que possibilitem a análise de dados em qualquer lugar (espaço) e momento (tempo) pois o rápido acesso aos dados é crucial para que os processos de tomada de decisão sejam mais rápidos, práticos e eficientes possíveis (Russom, 2013).

Devido à grande diversidade de dados surge o termo *Big Data* que engloba tanto os diferentes tipos de dados como as suas origens. O termo *Big Data* abarca múltiplos conceitos, gerando, por isso muita confusão e mal entendidos para todos os que trabalham com *Big Data* nomeadamente a comunidade académica.

É comumente aceite que o conceito *Big Data* foi pela primeira vez formulado por *Doug Laney*, em Fevereiro de 2001 onde analisou os desafios que as empresas enfrentavam na gestão dos dados, sendo que para tal definiu três dimensões as quais chamou de 3V's: **Volume** (grandes quantidade de dados) **Velocidade** (necessidade de captar, armazenar e analisar esses dados rapidamente, para suportar as atividades operacionais); e **Variedade** (capacidade de tratar, de forma integrada vários tipos de dados)

(Laney, 2001). Curiosamente, o termo *Big Data* não é referido nesse artigo, tendo vindo a ser adotado posteriormente.

De notar que já em finais do século XX, várias empresas reportavam o armazenamento e análise de grandes volumes de dados estruturados. Empresas como WallMart ou Bank of America reportavam, já nessa época a utilização de *Data Warehouse* (DW), com um volume de informação aproximado de várias dezenas ou centenas de *Terabytes* (TB). (Lohr, 2012).

No entanto se os dados a analisar fossem apenas os típicos dados estruturados (atributos numéricos ou alfanuméricos de relações no modelo entidade associação ou relacionamento (Modelo ER)) as tecnologias de bases de dados relacionais suportadas em processamento paralelo massivo *Massive Paralell Processing* (MPP), seriam suficientes para corresponder aos desafios (Kimball & Ross, 2013).

A estes três vetores iniciais (que passaram a ser referidos como "três V's") que originalmente definiam o conceito de *Big Data*, foram associadas outras características que contribuíram para proliferação do mesmo, nomeadamente o crescimento e massificação do uso das redes sociais e dispositivos móveis (Stonebraker, 2012)

Desta forma impõem-se a definição de uma arquitetura que permita dar resposta a todas estas necessidades.

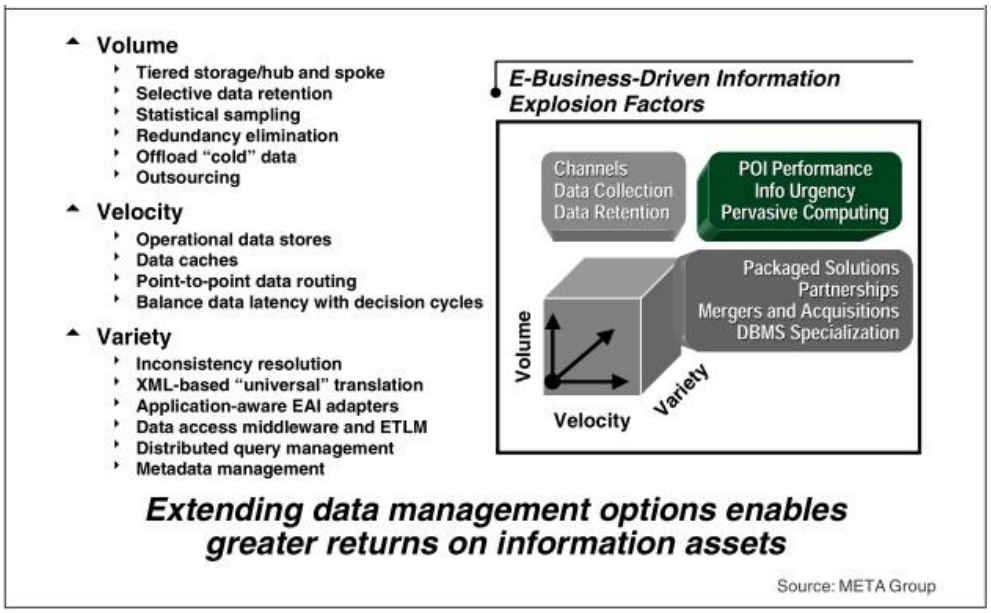


Figura 1 – 3D Data Management Controlling adaptado de (Laney, 2001)

A arquitetura de dados deve permitir uma integração de dados provenientes de diferentes fontes, o que exige um conhecimento exaustivo das fontes de dados, bem como de todo o processo necessário para extrair validar e transformar os dados de forma a auxiliar o processo de tomada de decisão (Russom, 2013).

É necessário que os dados sejam armazenados num repositório procedendo-se depois à sua análise. Devido aos dados serem do tipo *Big Data*, estes repositórios exigem tecnologias criadas especificamente para este tipo de soluções.

Finalmente, os dados devem ser disponibilizados para análises, onde os utilizadores (decisores, analistas, entre outros) contactam e exploram os dados existentes. Pretende-se que essa exploração seja o mais dinâmica possível, ou seja, o utilizador pode definir quais as variáveis e valores a explorar, podendo estas análises ser efetuadas através de *dashboards*, *reports*, *Online Analytical Processing* (OLAP), etc.

O termo *Big Data* tem sido muito utilizado, sobretudo porque as organizações estão a perceber o valor e as mais-valias que podem obter através das enormes quantidades de dados que possuem, podendo os mesmos ser catalisadores para o processo de inovação da organização através de "*insights* " (Stonebraker, 2012). Todas as organizações, umas mais rapidamente que outras estão a perceber a importância de *Big Data*, compreendendo as mudanças que necessitam de ser feitas, a fim de beneficiar, das vantagens competitivas da utilização de *Big Data* (Oswaldo, et al., 2010).

A ordem de grandeza associada a *Big Data* pode não ser verdadeira, no entanto as características mais interessantes são os dados estruturados, semiestruturados e não estruturados, bem como os dados brutos em diferentes formatos (Kimball & Ross, 2013).

O *Big Data* deve ser entendido como uma combinação de antigas e novas tecnologias que procuram ajudar as organizações a obter conhecimentos. Desta forma as organizações conseguem gerir a grande quantidade de dados e os problemas associados ao seu armazenamento. Os dados podem ter origem em diversas fontes internas e externas, *streaming* de dados, rede socais e médias sociais, dados com referenciação geográfica, entre outros, podendo estes dados ter diferentes tipos de estruturas (Halper & Krishnan, 2013), a saber:

- **Dados Estruturados:** São todos os dados que são organizados em blocos semânticos (entidades), sendo estas agrupadas através de relações e classes Todos estes dados são guardados em Sistemas de Gestão de Bases de Dados (SGBD). São chamados dados estruturados, porque possuem a estrutura rígida que foi previamente projetada através de um modelo, Diagrama de Entidades e Relacionamentos. Como exemplos de dados estruturados temos os dados provenientes dos ERP, de programas de gestão, históricos das últimas compras realizadas por um cliente, transações realizadas com cartões de crédito, etc. (Guoliang, et al., 2008).
- **Dados Semiestruturados:** São dados que não necessitam de estar armazenados num SGBD e que apresentam um elevado grau de heterogeneidade, são dados que não estão generalizados nem possuem um tipo de estrutura. Como exemplos de dados semiestruturados temos:

Extensible Makup Language (XML), *Resource Description Framework* (RDF), *Web Ontology Language* (OWL) e o conteúdo de um *email* etc. (Guoliang, et al., 2008).

- **Dados Não Estruturados:** São dados que não possuem necessariamente um formato ou sequência, não seguem regras e não são previsíveis. Estes dados são foco de muita atenção nos dias que correm devido principalmente à proliferação de dispositivos móveis responsáveis por uma grande variedade de dados, no entanto existem outras fontes de dados como: sensores de máquinas, dispositivos inteligentes, tecnologias de colaboração e redes sociais. Portanto, estes dados não são dados relacionados mas sim diversificados. Alguns exemplos deste tipo de dados são: textos, vídeos, imagens, etc. (Guoliang, et al., 2008).

As principais características que permitem a diferenciação entre os vários tipos de dados são apresentadas na tabela 2.

Big Data também proporcionou um crescimento da complexidade dos dados, devido à incapacidade dos atuais sistemas de bases de dados para armazenarem o crescente volume de informação como: acontece com vídeos, imagens, textos. O crescimento, proliferação e influência das redes sociais na sociedade foram os principais fatores de influência para o crescimento e importância do termo *Big Data* (Manyika, et al., 2011), a tabela 2 mostra explica as diferenças entre os vários tipos de dados.

Dados Estruturados	Dados Semiestruturados	Dados Não Estruturados
Estrutura predefinida	Nem sempre existe esquema	Não existe esquema
Estrutura regular	Estrutura irregular	Estrutura irregular
Estrutura independente dos dados	Estrutura embebida nos dados	A estrutura está dependente da fonte dos dados
Estrutura reduzida	Estrutura extensa (particular em cada dado visto que cada um pode ter uma organização própria)	Estrutura extensa depende muito do tipo de dado
Pouco evolutiva e bastante rígida	Muito Evolutiva, a estrutura pode mudar com muita frequência	Muito evolutiva, a estrutura muda com bastante frequência
Prescritiva com esquemas fechados e restrições de integridade	Estrutura descritiva	Estrutura descritiva
Distinção clara da estrutura de dados	Não é clara a distinção entre estrutura de dados	Não é possível distinguir entre as estruturas dos dados

Tabela 2 Dados estruturados, semiestruturados e não estruturados, adaptado de (Guoliang, et al., 2008)

Existem alguns estudos que citam um crescimento, e uma variedade de novos dados não estruturados na ordem dos 85% (Brown, et al., 2011) e associado a este efeito temos um problema que consiste em recolher este tipo de dados e armazená-los para posterior utilização (Russom , 2011).

Outro facto importante prende-se com a forma como devemos analisar estes dados para conseguirmos as melhores decisões e este é o grande desafio que se coloca aos profissionais que trabalham e poderão trabalhar com *Big Data* (Russom , 2011).

Big Data é um termo que não é fácil de definir, tal facto deve-se à grande diversidade de definições existentes, nomeadamente:

- Pode ser entendido como um conjunto de dados muito grande, o que não permite uma fácil utilização dos mesmos por parte dos sistemas de gestão de bases de dados relacionais (Sikka, et al., 2012).
- Consiste na mineração de grandes volumes de dados estruturados e não estruturados através da utilização de ferramentas não tradicionais como o Hadoop (Soares, 2013).
- Os dados Big Data devem estar de acordo com aos critérios da IBM que são descritos através dos 3V's: Volume, Variedade, Velocidade (Stonebraker, 2012).
- É uma inovação em computação na última década, pois muito recentemente perceberam o seu potencial na recolha, organização e exploração dos dados sobre diversas perspetivas. Exige um pouco de investimento mas todos concordam que o seu desenvolvimento e implementação proporcionarão enormes vantagens competitivas as organizações (Tran, et al., 2013).

O termo *Big Data* pode ter vários significados, mas é importante ter em conta que o mesmo está assente no princípio dos V's. Quando surgiram as primeiras abordagens teóricas relativas a *Big Data* eram 3 V's (**Volume**, **Variedade** e **Velocidade**), mas com o trabalho realizado por toda a comunidade científica e académica da área sugiram mais dois V's, V de **Veracidade** e o V de **Valor**. (Stonebraker, 2012):

- **Volume** - representa uma grande quantidade de dados a ser recolhida e analisada, sendo que a ideia é a utilização de *Structured Query* Language (SQL) *analytics (count, sum, max, min, average e group by)*, regressões, aprendizagem máquina e análises complexas em grandes volumes de dados.
- **Variedade -** corresponde a uma utilização de diferentes estruturas de dados como dados estruturados, não estruturados e semiestruturados.
- **Velocidade** - permite mostrar a rapidez com que os dados são processados.
- **Veracidade** – permite classificar as fontes de dados, tendo em conta aspetos como: qualidade, precisão e atualidades dos dados.
- **Valor** - corresponde ao valor que informação dos dados terá no processo de tomada de decisão.

No entanto é importante ter em conta que um em cada três processos de tomada de decisão é efetuado sem existir confiança na informação utilizada (Breternitz, da Silva, & Lopes, 2013). É importante realçar

que nem todos os sistemas de gestão de bases de dados relacionais devem ser considerados *Big Data*, (Manyika, et al., 2011).

Figura 2 - Estrutura do *Big Data* com os 5V's adaptado de (Stonebraker, 2012)

O uso de grandes quantidades de dados, necessita de uma estratégia de utilização de informação bem definida, sendo que existe um conjunto de etapas que devem ser seguidas para que o sucesso na utilização de grandes volumes de dados seja alcançado, e neste sentido, a recolha de dados o seu armazenamento e posteriores análises, deve incorporar uma avaliação da infraestrutura atual, aferindo, deste modo quais os sistemas de armazenamento que são utilizados, quais os dados provenientes de aplicações como CRM e ERP são importantes para o negócio e para o processo de tomada de decisão. Além disso é também necessário definir uma política que ajude na definição das linhas de orientação para a integração de todos os dados com o objetivo de garantir a sua qualidade.

Os grandes volumes de dados não podem apenas ser analisados através de SQL. Pelo contrário exigem uma mudança no paradigma, necessitam de análises mais complexas recorrendo a funções matemáticas e estatísticas. Atualmente pensa-se na importância dos dados, na recolha e análise da informação com o objetivo de encontrar ideias que permitam uma inovação e uma contínua criação de valor para o negócio que permita rentabiliza e sustentar as análises efetuadas. O movimento associado aos grandes volumes de dados ganhou impulso com o elevado número de casos que têm sido reconhecidos e que se enquadram na categoria de grandes análises de dados como: (Kimball & Ross, 2013):

- Busca e classificação.
- Monitorização.
- Procura da localização e da proximidade.
- Descoberta de factos casuais.
- Análise Sentimental.
- Teste de semelhança de documentos.
- Medidor de serviços públicos inteligentes.

- Comparação de imagens.

Big Data faz crescer a necessidade de um maior armazenamento, análise e integração de ferramentas de qualidade dos dados. A recolha de dados através das redes sociais, como Facebook ou Twitter, permite as organizações obter informações sobre os seus clientes, no que se refere às preferências dos mesmos sobre um determinado produto, mas também é possível perceber a satisfação dos clientes, através dos diversos comentários que os mesmos fazem sobre um determinado produto. As organizações podem desta forma aproveitar esta informação para acelerar o seu processo de inovação, pois conseguem obter maior segurança nas decisões que precisam de tomar e melhorar o seu poder de resposta em relação as necessidades dos clientes

Tendo em conta o âmbito anteriormente descrito existe uma ferramenta que ajuda na realização deste processo, o *MapReduce* que é uma *framework* de processamento distribuído que foi desenvolvida pela *Google* e que permite a realização de processamento de dados em máquinas fisicamente separadas.

Para demonstrar as diferenças entre o *MapReduce* e o SGBD estendido (Kimball & Ross, 2013), há que ter em atenção a informação da tabela 3:

SGBD Estendido	MapReduce/Hadoop
Principalmente proprietário	Fonte de código aberto
Caro	Menos Caro
Os dados devem ser estruturados	Os dados não necessitam de estruturação
Ótimo para pesquisas rápidas e indexadas	Ótimo para exames massivos de dados complexos
Apoio profundo para a semântica relacional	Apoio indireto para semântica relacional por exemplo HIVE
Apoio indireto para estrutura de dados complexos	Apoio profundo para estruturas de dados complexos
Apoio indireto para interação, ramificação complexa	Apoio profundo para interação, ramificação complexa
Apoio profundo para o processamento de transações	Pouco ou nenhum suporte para o processamento de transações

Tabela 3 - Comparação do SGBD estendido com o MapReduce/Hadoop, adotado de (Kimball & Ross, 2013)

Existe um conjunto de boas práticas que devem ser adotadas pelas organizações quando trabalharem com Big Data:(Kimball & Ross, 2013):

- Escolha das fontes de dados a importar para o sistema de armazenamento de dados que pode ser um DW, isto é com dados relevantes para o negócio e importantes para o processo de tomada de decisão.
- Foco na simplicidade da interface do utilizador e no desempenho.

- Pensamento numa ótica dimensional dividindo o mundo em dimensões e factos.
- Integração de fontes de dados separadas com dimensões conformadas.

- **Administração:** A estrutura de *Big Data* deve ser considerada tendo em conta análises e não consultas *ad hoc* ou relatórios padrão. Cada passo que é executado desde a origem dos dados ate à análise, deve estar suportado em rotinas analíticas complexas, implementadas através de funções UDF, com recurso a um ambiente de desenvolvimento orientado para a meta dado que pode ser adaptado de acordo com o tipo de análise. Este processo deve incluir carregamento, limpeza e tratamento de dados, integração com interfaces de utilizador e ferramentas de *Business Intelligence* (BI).
- **Arquitetura:** Deve-se optar pela construção de um sistema capaz de extrair factos. *Big Data* utiliza análises de dados, como uma forma de extrair factos que permitam análises de *tweets*, ou seja texto não estruturado, conduzindo a um conjunto de medidas de sentido numérico. A construção deste ecossistema abrangente, que possibilita a integração de dados estruturados dos sistemas SGBD, documentos, *e-mails business oriented* e *social networks* é o objetivo de *Big Data.* Trata-se pois, da integração de diferentes dados, provenientes de diferentes origens como dispositivos móveis, processos de automatização e alerta. Grande parte desta informação é significativa podendo ser analisada através de *querys*, que dividem a informação em dimensões.
- **Modelação de Dados:** O primeiro passo é criar as dimensões, dividir o mundo dos dados em dimensões e factos, para os utilizadores de negócio não importa o formato dos dados. As dimensões podem ser utilizadas para integrar dados provenientes de diferentes fontes. Antes de se efetuar a integração é imperativo identificar as dimensões associadas a cada fonte de dados. O processo de criar dimensões é muito importante para que se consiga obter sucesso na análise a grandes volumes de dados. Em situações em que o fluxo de dados é de baixa velocidade pode-se optar por uma criação de dimensões automáticas. Alguns dados de entrada podem ser unidimensionais e, nestes casos, a etapa de extração é efetuada logo após a recolha de dados, num processo que pode ser considerado quase em *real time*.
- **Governança:** No caso de *Big Data* não existe governança de dados: a mesma deve ser uma extensão da organização, pretendendo-se que abranja aspetos como: segurança, conformidade, qualidade dos dados, gestão de meta dados e dados mestres, glossário de negócios que permite expor as definições de contexto para a comunidade empresarial.

Com a adoção de *Big Data* por parte das organizações os processos de tomada de decisão irão sofrer algumas alterações, sobretudo porque os gestores não estão dispostos a esperar muito tempo (exemplo: semanas), por um relatório ou *dashboard*. Pelo contrário, é necessário que a informação esteja disponível o mais breve quanto possível, para que organizações disponham de um processo de tomada de decisão,

com informações fidedignas e atuais. Assim, é importante que *Big Data* possua as seguintes características (Stodder, 2013):

- **Adoção corporativa:** A organização deve adotar uma cultura que valorize o volume de dados, nos processos de tomada de decisão. É igualmente importante que exista alguma flexibilização e uma mente aberta o suficiente para acolher novos *insights* que estão associados a conjuntos de dados, sendo que em muitos casos isto implica uma mudança de mentalidade das pessoas para perceberem o potencial de novas oportunidades, que podem exigir uma grande mudança no método e na filosofia de trabalho da organização.
- **Incentivar a audácia:** Em lugar das organizações solicitarem análises a entidades externas as análises devem ser realizadas e estimuladas dentro da organização. Incentivando deste modo o esforço diário dos seus colaboradores. Impõe-se que o estímulo seja constante, de modo a potenciar a utilização e rentabilização das suas habilidades e ajudar no processo de tomada de decisão aproximando-o o mais possível do tempo real. Nesse sentido é necessário que as pessoas desenvolvam diversas capacidades como: colaboração com colegas e desempenho de funções em outros departamentos, procurando utilizar a informação no momento e no tempo adequados e tendo sempre em mente o objetivo de inovar.
- **Visualizar:** A apresentação da informação é muito importante, devendo esta ser concretizada de forma simples, clara, concisa, fácil de interpretar, procurando, assim, não causar enganos nem interpretações duplas.

Ao reunir novos dados de diferentes fontes a organização pode de uma forma rápida e fácil identificar padrões, conexões e perceções que anteriormente não seriam fáceis de identificar. Isto ajuda a organização a melhorar as suas decisões. Esta abordagem permite ir muito mais além da inteligência de negócios, possibilitando a criação de modelos preditivos com base nos novos modelos de negócio (Halper, 2013).

Os sistemas que manipulam *Big Data* são constituídos por centenas ou milhares de infraestruturas de rede de alta velocidade e grandes sistemas de armazenamento com discos rígidos de classe empresarial de elevada capacidade que são projetados para funcionarem em ambientes de computação de alta escalabilidade (Cheng, et al., 2012).

Contrariamente aos dados que se encontram armazenados nas bases de dados relacionais *Big Data* possui muitos dados que não se encontram organizados em estruturas, os sistemas relacionais possuem uma capacidade de armazenamento e análise limitados em determinadas gamas de dados como por exemplo números ou datas. O facto de *Big Data* compreender vários conjuntos de dados como texto,

vídeos dados provenientes de sensores, ficheiros de registo (*logs*), etc., as análises a estes tipos de dados permite que as organizações possam chegar a um conhecimento que anseiam (Borkar, et al., 2012).

Big Data será aplicado no sentido de recolha de cada vez mais dados de forma às organizações poderem aprofundar os seus conhecimentos, nomeadamente a nível do comportamento do consumidor, processos industriais, e fenómenos naturais. Tudo isto poderá conduzir a um desenvolvimento de aplicações para análise de dados mais detalhadas e precisas procurando explorar ao máximo todo o potencial dos dados recolhidos pelas organizações (Borkar, et al., 2012).

É verdade que grande parte da informação que temos não tem um valor significativo, o que faz com que se torne muito importante tratar a informação de forma a perceber qual é a informação útil, pois apenas desta forma se consegue conjugar o *Data Mining* e o *Big Data* através do *Big Data Analytics* para que a informação contida nos dados possa ajudar nos processos de inovação e criação de valor contínuo na organização.

O *Big Data Analytics* consiste na utilização apropriada de modelos estatísticos, com o objetivo de extrair valor a partir da análise de grandes volumes de dados, com tarefas como: a pesquisa dos dados em detalhe e a sua sumarização. Os dados necessitam de ser observados e as experiências que são realizadas com *Big Data* consistem na transmissão, armazenamento e interação, porém as atuais tecnologias que são utilizadas na maior das organizações dificultam este trabalho. As arquiteturas de *Big Data* necessitam de um crescimento na estrutura de armazenamento, devido à incapacidade dos sistemas relacionais em lidar com os novos tipos e formatos de dados.

Para a realização de análises *Big Data* é necessário um sistema longitudinal, escalável que permita um contínuo aumento da velocidade de processamento, memória e disco que permitam uma expansão e acessos a recursos de um modo simples. Pretende-se que as soluções de *Big Data* sejam soluções horizontais, que possibilitem um dimensionamento do sistema de acordo com as necessidades dos *clusters* que estão a ser utilizados pelo sistema. A velocidade com que o sistema, se consegue adequar às necessidades dos *clusters* será o segredo para o sucesso do *Big Data*, na medida em que se o sistema não conseguir acompanhar a velocidade com que os dados são fornecidos não irá conseguir fornecer respostas adequadas e atempadas para a organização.

Um DW trabalha com dados históricos armazenados de forma estruturada e centralizada e devido ao facto das empresas possuírem apenas 20% dos dados estruturados as decisões que são tomadas são baseadas em pequenas quantidades de dados do universo de dados que está disponível para apoiar o processo de decisão na organização. A grande maioria dos dados está armazenada em bases de dados relacionais e hierárquicas, sendo que os restantes dados estão presentes por exemplo nos planos de

apresentação de documentos. Atualmente as organizações já possuem dificuldades em tomar decisões com os dados disponibilizados pelos sistemas de informação (Demirken & Delen, 2012).

As estruturas atuais como o SGBD não são apropriadas para trabalhar com o *Big Data*, que exige uma infraestrutura independente de *hardware* e *software* para lidar com *Terabytes* de dados o que faz criar a necessidade de uma infraestrutura de armazenamento escalável, com alto desempenho e que utiliza tecnologias como o *Solid-State Drive* e *Direct-Attached Storage Drive*, ou então trabalhar com os dados em memória. Por exemplo tecnologias como o *Network-Attached Storage* são relativamente lentas para *Big Data*. Como a filosofia do *Big Data* se centra na divisão das pesquisas em várias frentes é necessário um grande número de processadores de forma a conseguir um alto desempenho, no processamento, bem como usar o processamento paralelo e distribuído.(Friedman, Beyer, & Thoo, 2010).

É verdade que os recursos de computação como o armazenamento e as redes em conjunto com os sistemas operativos de *cloud* e as estruturas de *framework* como o OpenStack. Permitiram o desenvolvimento de soluções integradas com a solução do *Cloud Operating System* da Microsoft e o VCloud da VMWare, que disponibilizam recursos e infraestruturas virtualizadas que são fundamentais para que possa existir um auto escalonamento no qual a *cloud* pode fornecer de forma automática recursos como as infraestruturas à medida que a procura de uma aplicação aumenta por parte dos utilizadores (Demirken & Delen, 2012).

Na atualidade o trabalho de TI é o desenvolvimento de soluções e aplicações cada vez mais fluídas, considerando que hoje é necessário existir uma colaboração muito mais próxima do negócio, com a realização de análises contínuas ao comportamento dos utilizadores e tudo isto obriga a revisões frequentes do ciclo de vida das aplicações. A ideia base aqui presente é a de um desenvolvimento central e rápido que possa ser adotado num vasto segmento de plataformas web e móveis (Qin & Li, 2013).

A integração de dados ainda é encarada por muitas organizações como um obstáculo, pois trata-se de uma tarefa que se pode tornar bastante cara e difícil.

A utilização de dados em *cloud* exige um ambiente de *cloud* com uma conectividade entre várias fontes de dados bem como a conectividade para fontes de dados baseadas na internet como é o caso das redes sociais. No entanto, existe um conjunto de técnicas que podem ser utilizadas para ajudar a minimizar ou mesmos e inibir a existência de problemas nos fluxos de dados entre as várias aplicações, sendo exemplo dessas técnicas as seguintes (Halper, 2014):

- Utilização de API's, que permitem a conectividade de dados entre os diferentes sistemas sendo que através destes adaptadores é possível contornar as diferenças nas origens e destinos dos dados evitando problemas de incompatibilidade de dados de um sistema para outro. Muitos dos

fornecedores destes serviços fornecem conectores que permitem transferir os dados de um local para outro sem problema.

- Plataformas de integração de dados que podem oferecer a integração de dados como um serviço neste caso os serviços de ETL, que permitem garantir a qualidade, validação e integridade dos dados, contudo a utilização da ferramenta tecnológica de ETL está dependente do fornecedor de serviços, tendo em conta que alguns permitem que os clientes possam utilizar as suas próprias ferramentas de ETL, outros preferem que os clientes utilizem a ferramenta de ETL presente na plataforma.

Antes de uma organização mover os seus dados é importante compreender a estrutura dos mesmos, o seu volume e a frequência de atualização, pois existem várias maneiras de mover os dados de um repositório local para a *cloud* sendo os mais comuns *File Transfer Protocol* e *Wide Network Area*. A portabilidade de dados permite que as organizações, consigam obter dados em execução no sistema de um fornecedor e implementar esses dados num outro local, no entanto estas são normas que ainda estão em evolução sendo uma questão importante que fornecedores de serviços de *cloud* esperam discutir, com o intuito de encontrar uma solução em breve (Halper, 2014).

A The DataWarehouse Institute (TDWI) revela que algumas organizações acreditam que BI deve estar maduro antes de optarem por uma mudança para serviços de BI na *cloud.* Na realidade BI pode ser utilizado na *cloud* sem ter um nível de maturidade elevado, por exemplo, se utilizarem ferramentas de CRM baseadas em *cloud* será possível efetuar análises na própria *cloud*. A opção por soluções deste tipo é benéfica para as organizações que querem criar um ponto de partida procurando amadurecer a sua solução de BI, podendo mais tarde optar por uma integração de dados provenientes de outros ambientes em *cloud* (Halper, 2014).

É verdade que muitas organizações nos últimos anos decidiram optar por criar soluções de BI móveis que lhes permitam analisar dados e fornecer resultados dessas análises em equipamentos como *Tablets* e *Smartphone*. No entanto, é importante determinar os custos associados com os sistemas de dados em *cloud*, contudo a maioria das organizações só tem em conta os custos de capital inicial versus as despesas operacionais (Dean & Ghemawat, 2010).

A maioria das organizações não tem uma imagem precisa dos seus próprios custos e dependendo da estratégia de BI, estes custos podem incluir os custos associados ao armazenamento de *backup*, arquivamento, continuidade de negócios e recuperação de desastres e software de manutenção. As organizações simplesmente compram os custos iniciais com as despesas operacionais, não tendo em conta dois pontos que são essenciais como a agilidade e flexibilidade (Dean & Ghemawat, 2010).

Existem organizações que pensam que o tempo de implementação e prevenção de riscos são fatores primordiais, mas não compreendem que os riscos de projetos em *cloud* são muito menores que os riscos de uma implementação local, logo, quando as organizações querem optar por este tipo de soluções devem ter em conta os custos anuais de serviços em *cloud* em comparação com os custos sem *cloud*, as políticas e o perfil de risco da sua própria organização. Para os serviços em *cloud* a governança deve ser tida em conta pois a mesma ajuda na definição de como os vários intervenientes se podem comportar porque várias pessoas em diferentes organizações farão parte de um plano de governança (Dean & Ghemawat, 2010).

Existe um conjunto de fatores que estão associados aos serviços e que permitem o cumprimento da Governança, é importante destacar dois desses fatores que são importantes para a execução da Governança (Tankard, 2012):

- **Cumprimento**: tem a ver com os requisitos de conformidade que estão regulamentados ou requisitos impostos pelo cliente ou parceiros e que a organização terá que cumprir, no entanto existem alguns requisitos que afetam quem fornece serviços de *cloud*, que é a garantia de que todas as empresas que processam, armazenam e transmitem dados de cartões de crédito possuem um ambiente com um elevado grau de segurança.
- **Visibilidade**: inclui as operações de ETL, arquivamento e afins, quer fornece os serviços têm que disponibilizar ferramentas tecnológicas que permitam apoiar e monitorar as operações, bem como possuir acordos de níveis de serviço.

A mesma análise da TDWI identifica a falta de segurança e de privacidade de dados, como os principais fatores de resistência a adoção de soluções em *cloud*, mas a opção por uma abordagem séria em relação às questões de segurança pode ajudar no sucesso da mitigação de muitos dos riscos de segurança. Por exemplo pode fazer algum sentido em primeiro lugar analisar o ambiente local no qual os dados estão armazenados com o objetivo de detetar os pontos fortes e fracos, é também importante reconhecer que muitos dos servidores de internet são vulneráveis. Deve-se falar com o fornecedor de serviços de *cloud* sobre a sua segurança, ter em conta a forma como são efetuadas as auditorias de segurança, e considerando, igualmente, questões como (Halper, 2014):

- Segurança Física: perímetro de segurança, acesso interno para servidores, detalhes sobre planos de tolerância a falhas.
- Operacionalidade: deve incluir prontidão em termos de vigilância do sistema de gestão de incidentes para o sistema ou outros incidentes como incêndios, ou outras crises e verificar a frequência com que são realizadas as avaliações das vulnerabilidades.

- Plataforma: a aplicação da segurança deve incluir intrusão deteção e suporte para conectividade segura via *Virtual Private Network* (VPN), *Secure Scoket Layer* (SSL), *Hyper Text Transfer Protocol Secure* (HTTPS) entre outros, bem como possuir um papel baseado em controlo de acessos.
- Segurança de dados: controlos que permitam manter a integridade dos dados e uma transmissão segura de dados usando por exemplo técnicas de criptografia.

As organizações que não conseguirem de uma forma rentável acompanhar esta inovação podem perder muitas oportunidades de se envolverem com os clientes, procurando otimizar processo complexos e tomar decisões tendo em conta os riscos os riscos associados a cada decisão (Dermiken & Delen, 2013).

O uso de diversos tipos de dados começa a ganhar força nas organizações tornando-se mesmo vital para o sucesso dos negócios das mesmas. Por exemplo as empresas que prestam serviços na área financeira estão a tentar ganhar introspeção com o cliente através de uma combinação dos conhecimento mercado e dados históricos, analisando os mesmos e procurando tirar conclusões sobre o que pode ou não acontecer a organização num futuro próximo (Demirken & Delen, 2012).

A importância que anteriormente era dada aos dados não estruturados e estruturados é agora maior, pois as organizações perceberam que estes podem ser uma parte importante das ações que a organização pode fazer tendo como base as decisões a tomar (Demirken & Delen, 2012).

No que se refere a BI, a grande maioria das organizações utiliza poucos tipos de dados não estruturados, com base nisso consegue-se perceber que cerca de 67%, dos responsáveis por tomar decisões utilizam bases de dados relacionais, o que faz todo o sentido, considerando que as tecnologias de análise de hoje foram desenvolvidas para dar resposta às necessidades das organizações, procurando melhorar a sua gestão financeira. Importa referir que existem importantes fatores de prioridade neste processo, nomeadamente a qualidade e a integridade dos dados, no entanto o uso de sistemas relacionais não permite a análise de novos tipos de dados com escalas extremas onde os principais requisitos para o sucesso são a velocidade e a agilidade. (Dermiken & Delen, 2013)

A Forrester conseguiu constatar que as pessoas responsáveis por tomar decisões nos Estados Unidos da América, reconhecem a importância da procura de mais dados e de novos formatos bem como a utilização de *analytics*, mas também alertam para uma restrição de dados de forma a poderem lidar com os obstáculos que enfrentam quando as tecnologias não conseguem dar resposta às necessidades que existem nos dados hoje em dia.

Os resultados obtidos com o inquérito que a Forrester realizou nos Estados Unidos da América aos gestores e responsáveis de departamentos de informática, revelou que quase 100% dos inquiridos

concordam que enfrentam cada vez mais um grande número de mudanças constantes, associadas a requisitos como uma maior variedade de dados novas visões que surgem sobre os dados. Mas também importa referir que existem inquiridos que equacionam limitar para 90% as quantidades de dados disponíveis, com o intuito de garantir análises com grande desempenho e qualidade. Estas duas visões distintas só entram em conflito se existir, por parte dos decisores de TI reconhecimento da crescente procura por mais dados, novos formatos e novas análises. A par disso existem algumas restrições de dados que procuram lidar com obstáculos difíceis, quando as tecnologias que utilizam não conseguem dar resposta aos problemas e aos novos desafios que os dados de hoje colocam. A razão que é apontada pelas organizações entrevistadas, tem a ver com o facto de as despesas associadas aos dados disponíveis para análise não permitirem a utilização de mais dados disponíveis encontrando-se a par destas limitações, outras questões de segurança e complexidade (Raj & Deka, 2012).

Outra grande questão está relacionada com o passado recente em que o armazenamento dos dados das organizações era feito num DW que por sinal é bastante dispendioso. Com as soluções de *Big Data*, as organizações podem guardar os dados e analisá-los, o que permite que um custo de oportunidade possa ser transformado numa oportunidade de negócio. Mas muitas organizações dizem "não" a esta oportunidade, pois simplesmente preferem ignorar as potencialidades que podem alcançar com a inovação que novos dados podem trazer ao negócio (Schönberger & Cukier, 2013).

Existem responsáveis de algumas organizações que manifestam algum interesse relativamente à opção por soluções *Big Data*, pois acreditam que estas tecnologias podem ajudar em larga escala na exploração e no acesso às capacidades dos requisitos de negócio. O Hadoop é vist0 como a tecnologia capaz de lidar com grandes volumes de dados e estruturas variáveis de dados, sendo que todo este trabalho que o Hadoop faz é realizado com a utilização de uma *framework* de código aberto e do processamento de operações paralelas de dados usando *hardware* de baixo custo. A solução Hadoop, permite oferecer uma alternativa aos dados estruturados armazenados em sistemas de DW e que a maioria das organizações acha mais rentável para a determinação de classes de análises de dados como os dados provenientes de ERP, CRM, SCM entre outros sistemas de apoio a gestão (Stuckenberg, et al., 2011).

Big Data é a oportunidade para as organizações poderem ganhar vantagens competitivas e a grande questão que se coloca hoje não se prende tanto com a opção de usar ou não o Hadoop, mas sim no que se refere a quando e como será melhor aproveitar o Hadoop como um complemento dos seus conjuntos de ferramentas tecnológicas para gestão de dados e análises existentes (Qin & Li, 2013).

O Apache Hadoop é um sistema que proporciona uma otimização das cargas de trabalho, melhorando as análises e permitindo um *streaming* de dados, análises de texto e de conteúdo, dando às organizações um maior poder de análise sobre cada um dos seus processos de negócio (Qin & Li, 2013).

O Apache Hadoop é a plataforma que oferece a governança, privacidade e segurança de que as organizações precisam além de ser uma plataforma aberta, modular e que pode ser integrada nas organizações a uma pequena escala e com o seu próprio ritmo. O objetivo destas soluções é além da utilização dos dados já conhecidos procurar e fomentar a utilização de análises que permitam levar a inovação, recorrendo às análises textuais a documentos e *emails*, utilizar as diversas informações da web como por exemplo o que é escrito nos media sociais sobre a organização os seus produtos, colaboradores e atividades que desenvolve. Utilizar as informações de navegação web como os *logs* procurando perceber que público procura informação na internet sobre a organização, o objetivo das pessoas quando procuram informação sobre uma determinada organização (Qin & Li, 2013).

Os sistemas de gestão de dados com características de escalabilidade, que permita uma atualização das cargas de trabalho e a utilização intensiva do sistema com grandes volumes e variedades de dados, num paradigma de *cloud* são considerados uma parte crítica. Isto deve-se às várias alterações que são necessárias nos dados padrão, principalmente no acesso a aplicações e à necessidade de dimensionar milhares de máquinas e de *commodities*, que aos poucos estão a ser amplamente adotados, por várias organizações. No domínio da análise de dados. O paradigma associado ao MapReduce e a sua implementação *open source* Hadoop, tem sido adotado amplamente por parte das organizações e universidades (Agrawal, et al., 2010).

A realização de análises preditivas, pode ajudar a perceber e a desenhar modelos e comportamentos dos diferentes utilizadores de modo a constatar as tendências que estão a ser implementadas na internet através das visitas e do comportamento dos utilizadores *online*.

A conjugação das tecnologias existentes de Armazenamento de dados DW com a utilização do *Apache* Hadoop e das suas capacidades de trabalho com diferentes tipos de dados faz com que o acesso aos dados se torne bastante rápido e eficiente pois a utilização de uma combinação de dados internos com a informação externa permite à organização obter uma maior sustentabilidade daquilo que o mercado espera que a organização seja capaz de fazer (Qin & Li, 2013).

O Hadoop pode entrar numa organização por causa de um pequeno caso de uso específico, mas a verdade é que dados atraem dados e uma vez dentro da organização o Hadoop pode tornar-se um centro de gravidade, estando este efeito associado ao *Big Data*, e que não pode estar apenas relacionado com o tamanho dos dados, mas também associado à agilidade que o sistema traz a organização.

Para viabilizar o Hadoop é necessário mais do que apenas um mecanismo de trituração de dados e um conjunto de pessoas dedicadas ao desenvolvimento. O Hadoop deve tornar-se uma plataforma empresarial que suporte o desenvolvimento de aplicações. Já no final de 2013 verificou-se a existência de um conjunto de fornecedores com estratégias e plataformas baseadas em Hadoop como é o caso da

Cloudera Enterprise Data Hub e do *Hortonworks Data Platform.* Existe mesmo uma frase protagonizada pelo *Chief Executive Office* (CEO) da Pitoval que descreve: "A concentração de dados em massa no *Hadoop Data Lake* " (Qin & Li, 2013).

A perspetiva do *Data Lake* (Lago de Dados) converge na descrição de uma arquitetura centrada nos dados na qual os silos são inimizados e o processamento acontece com pouco atrito através de um ambiente escalável e distribuído, em que os pedidos já não são ilhas que existem dentro de uma nuvem de dados que aproveita o acesso de banda larga para dados e recursos de computação escalável. Os dados por si não são mais restringidos por decisões de esquema, podendo ser explorados de forma mais livre (Wasniowski, 2014).

Tendo em conta a informação anterior, constata-se que a maturidade do Hadoop pode ser descrita em quatro níveis, partindo dos quais se consegue identificar em que nível uma organização se encontra, o que ajuda na identificação dos processos que devem ser melhorados para a organização aumentar a sua maturidade.

Os quatro níveis de maturidade do Hadoop são (Dumbill, 2014):

1. Inicialmente as aplicações eram autónomas e possuíam a sua própria base de dados, algumas dessas aplicações contribuíam com dados para o DW, desta forma os analistas conseguiam executar relatórios e obter resultados das análises aos dados que seriam utilizados pelos decisores, como demonstrado na figura 3 (Dumbill, 2014).

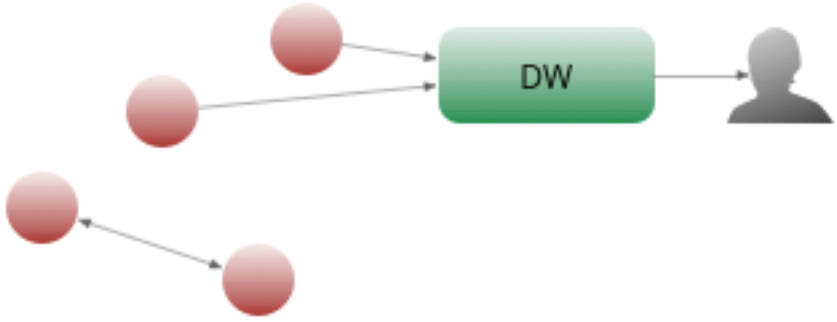


Figura 3 - Antes do Hadoop adaptado de (Dumbill, 2014)

2. O aparecimento do Hadoop, e a utilização do processamento paralelo através do Map Reduce, permite que os dados das organizações possam ser utilizados pelo Hadoop em conjunto com funções analíticas, tornando possível o processamento e analise de grandes quantidades de dados de uma forma rápida e eficiente, ver figura 4 (Dumbill, 2014).

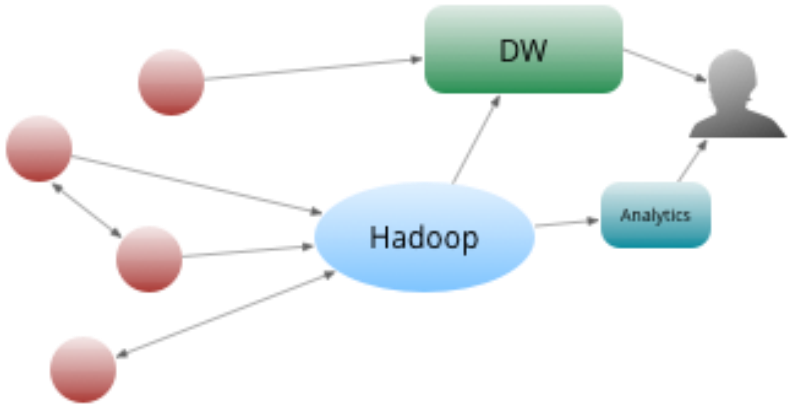


Figura 4 - Com o Hadoop dentro da Organização adaptado de (Dumbill, 2014)

3. Os sistemas mais recentes serão desenvolvidos tendo como base o Hadoop.. Por outro lado, os DW serão utilizados para exceções ou casos específicos de dados que as organizações necessitem, pois todos os dados externos a organizações serão integrados no Hadoop, como mostra a figura 5 (Dumbill, 2014):

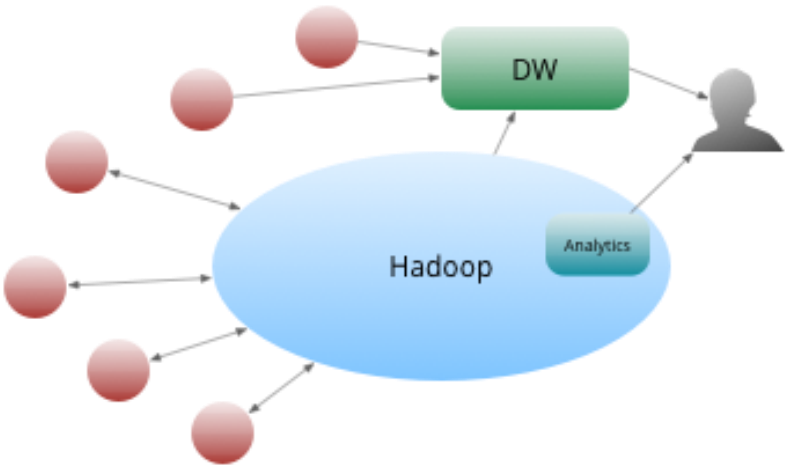


Figura 5 - Crescimento do Hadoop na Organização, adaptado de (Dumbill, 2014)

4. As novas aplicações serão construídas em Hadoop passando a possuir características como a elasticidade de computação de dados quer para dados operacionais, quer para funções analíticas. Permitem ainda adicionar diferentes níveis de segurança, manter os dados sempre disponíveis e uma rápida e fácil implementação de aplicações. No entanto algumas aplicações legadas poderão necessitar de uma execução de forma independente, ver figura 6 (Dumbill, 2014).

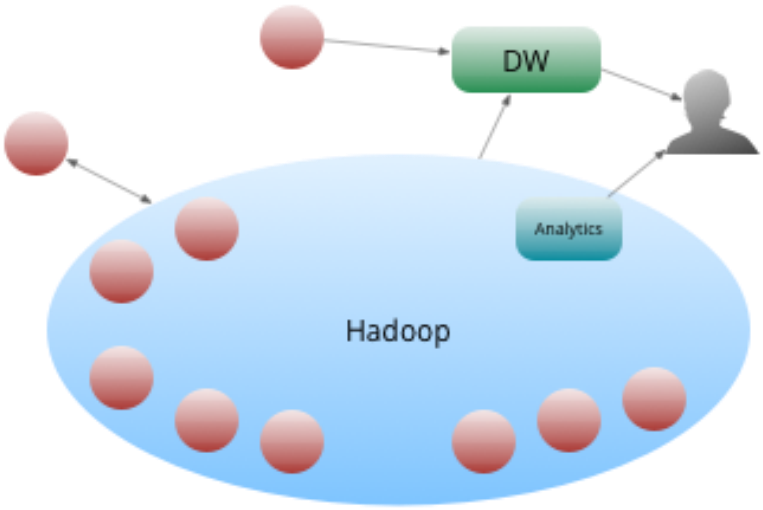


Figura 6 - *Data Lake* com aplicações em *Cloud*, adaptado de (Dumbill, 2014)

A realidade atual é que muitas organizações estão apenas a começar a utilizar Hadoop. O nível de maturidade corrente em maior parte das organizações é o nível 2 (com o Hadoop dentro da Organização), ou seja, o nível inicial (Dumbill, 2014).

As organizações que estão a optar pelo Hadoop têm desenvolvido um grande esforço em termos de investimento em soluções e oportunidades, pois os fornecedores ainda não oferecem soluções de gestão da infraestrutura de dados, descoberta de conhecimento nos dados e a reedificação de segurança.

A infraestrutura interna é desenvolvida em grande parte por duas empresas o Facebook e a Google, pois foram elas que desenvolveram vantagens da agilidade e elasticidade associadas ao Hadoop e a ideia do *Data Lake* (Dumbill, 2014).

Independentemente do estado no qual a organização se encontra é necessário dedicar tempo a perspetivar o futuro, a verdade é que o negócio das organizações é cada vez mais digital e o acesso aos dados é crítico e prioritário (Dumbill, 2014).

As pequenas empresas começam a ganhar competitividade, criam valor e transformam o seu negócio. Estão a captar todos os tipos de dados, tentando que este processo seja em tempo real, com o objetivo de estender os seus processos analíticos, incluindo as capacidades preditivas e cognitivas, procurando ser proactivas na sua privacidade, segurança e governança. Dessa forma, pretendem incluir os processos de análise em todo o lado com o objetivo de melhorar os seus resultados (Raj & Deka, 2012).

Neste sentido, as arquiteturas devem tornar o processo de análise de dados o mais rápido, eficaz e eficiente possível. À medida que se trabalha com uma crescente variedade de fontes de dados as cargas

de trabalho analíticas são maiores e implicam um conjunto de tecnologias que permitam lidar com a complexidade, mas ao mesmo tempo permitam um processamento rápido e eficiente dos dados, tornando os dados úteis de imediato para quem deles necessitam (Borkar, Carey, & Li, 2012).

Hadoop é introduzida como uma tecnologia de dados que permite o trabalho em *cluster* através de sistemas distribuídos de ficheiros, garantindo atomicidade, consistência, isolamento e durabilidade e que tornam estes sistemas uma alternativa viável e benéfica. A verdade é que os *clusters* são uma parte importante das arquiteturas de dados, pois possuem tecnologia que permite privacidade de dados e procura evitar problemas como a perda da integridade dos dados (Raj & Deka, 2012).

O Hadoop é uma tecnologia particularmente especial para dados não estruturados, que possibilita a comunicação com aplicações web, texto, rede e segurança de dados

No fundo a não existência de um modelo multidimensional com as tabelas de factos e as dimensões a ela associadas, obriga ao uso de HiveQL (linguagem de manipulação e acesso aos dados do Hive) para manipular dados, promovendo análises diretas aos dados. Por outro lado e em contraste com o modelo OLAP*,* que possui custos de armazenamento bastante altos, a utilização do MapReduce em combinação com o Hive possui custos de armazenamento bastante baixos.

De facto é notória a diferença entre os dados atualmente mais utilizados (dados estruturados) e o foco no volume, variedade e velocidade dos dados. A utilização de uma infraestrutura em *cluster* torna mais confiável e disponível a informação viabilizando uma margem de melhoria. A segurança de dados em *cluster* é mais fiável que as soluções de dados locais, devido principalmente à limitação que as soluções locais possuem em termos de análises de dados em grande escala em que o tempo de execução das análises é muito reduzido (Qin & Li, 2013).

As principais características associadas a *Big Data* como o processamento paralelo massivo e problemas de armazenamento redundantes e a resiliência de dados fora do aglomerado constituem as principais vulnerabilidades que necessitam de ser compreendidas e resolvidas num curto prazo (Tankard, 2012).

O *Big Data* é uma oportunidade e uma mudança na sociedade. A grande quantidade de dados é o mais conveniente para descobrir *insights* que permitam à organização melhorar o seu negócio. O problema das grandes quantidades de dados é que é irrealista confiar numa integração multidisciplinar.

2.2 *Cloud Computing*

O termo *Cloud Computing* nasce, por volta dos anos 60 como uma espécie de computação organizada, de utilidade pública com origem numa representação gráfica que tem em conta os recursos baseados na Internet e nos diagramas de sistemas (Francis, 2009).

Nos dias que correm o conceito é utilizado quando nos referimos a um novo paradigma ou tecnologia, flexível que oferece recursos e serviços de TI, com base na Internet (Böhm, et al., 2011).

Cloud Computing também pode ser considerado como um conjunto de conceitos associados a várias áreas de conhecimento como *Service-Oriented Architecture* (SOA), computação distribuída, computação em *Grid* (modelo computacional que divide as tarefas a executar por diversas máquinas) e virtualização (Youseff, et al., 2008).

Muito além da visão tecnológica que está associada ao conceito de *Cloud Computing*, o conceito pode também ser entendido como uma inovação principalmente na prestação de serviços de TI (Böhm, et al., 2011). Muitos acreditam que este é um potencial a ser explorado, principalmente no modo de desenvolvimento e implementação de recursos de computação e aplicações, procurando novos modelos de negócio principalmente para as empresas fornecedoras de *Software* (Youseff, et al., 2008) (Stuckenberg, et al., 2011).

O *National Institute Standards and Technology* (NIST) define o *Cloud Computing* como um conjunto partilhado de recursos configuráveis de computação como (redes, servidores, armazenamento, aplicações e serviços), que podem ser rapidamente oferecidos com um serviço (Olivier, et al., 2012).

Segundo o *NIST* existem muitos benefícios na adoção de *Cloud Computing* mas os mais importantes são os seguintes (Olivier, et al., 2012):

- Permite economias de escala no lado do fornecedor de serviços procurando promover uma maior produtividade do fornecedor de serviços de infraestrutura com uma flexibilidade contínua do lado do utilizador em relação às economias de escala diminuindo deste modo o investimento e custos de funcionamento, pois o retorno sobre o investimento cresce, levando ao aumento do ritmo e do nível de inovação geral.
- Permite às organizações encontrar as competências essenciais sustentáveis através de uma contínua modernização e reinvestimento nos serviços de TI, por parte do prestador de serviços.
- Sugere uma lógica de evolução complexa e contínua, com uma melhoria da informação procurando esconder a transparência da Lei de Moore que está no centro do desenvolvimento de TI. Permite a possibilidade de um sistema de controlo de políticas abstratas, sendo necessário ter em conta um controlo dos parâmetros de segurança, do *core business* (parte

mais importante ou cérebro do negócio) implementando uma política que evite a perda da informação relevante para o negócio.

Tendo em conta os benefícios anteriormente descritos em relação à adoção do *Cloud Computing,* importa ter em conta algumas características que lhe estão associadas, ver tabela 4.

Cloud Computing é um paradigma com muito sucesso no que diz respeito à orientação de serviços de computação. Este paradigma permitiu uma revolução na utilização da computação e das infraestruturas a ela associadas, mas quando falamos de *Cloud Computing* existem três conceitos de negócio que lhe estão associados: *Infrastructure as a Service* (IaaS), *Platform as a Service* (PaaS), e *Software as a Service* (SaaS) (Vaquero, et al., 2009).

O modelo *Cloud Computing* tem sido utilizado por grandes empresas para efetuar a gestão da sua infraestrutura. Dessa forma não existe a necessidade de efetuar atualizações constantes ao sistema, nem existe a necessidade de ajustar as cargas de trabalho e os recursos necessários ao sistema, pois todo este trabalho é feito por parte de quem está a gerir a infraestrutura, deixando esta de ser uma preocupação por parte de quem utiliza o sistema. (Qi, et al., 2010).

Cloud Computing mudou a forma como comunicamos e ampliou a oferta de serviços e de negócios (Foster, et al., 2008).

Características	Descrições
Serviços a Pedido	Usado como um serviço sempre disponível e sem necessidade de intervenção manual.
Amplo acesso a rede	O serviço é disponibilizado através de uma rede de forma independente do dispositivo e do utilizador final. A conexão de rede deve ser de alta performance e sempre disponível.
Partilha de Recursos	O fornecedor do serviço deve assegurar os recursos necessários, para que os consumidores de serviços possam utilizar a tecnologia de virtualização e *multi-tenancy*.
Elasticidade Rápida	Os recursos necessários devem ser disponibilizados rapidamente e libertados sem necessidade de intervenção manual quando deixarem de ser necessários.
Serviço medido	Um serviço consumido deve ser mensurável em termos de recursos usados. Desta forma a faturação é baseada no consumo tendo associados termo como "*pay per use"*.

Tabela 4 - Definição de *Cloud Computing* Segundo o Modelo NIST, adotado de (Mell & Grance, 2011)

Contudo *Cloud Computing* não tem só aspetos positivos. Existem alguns desafios a superar, de forma a poder alcançar o sucesso esperado (Olivier, et al., 2012):

1. Implementação de um modelo *cloud* de dados, que seja confiável e que permita às empresas, confiar os seus dados a terceiros de uma forma totalmente compatível com os respetivos enquadramentos regulamentares.
2. Implementação de uma *cloud* que permita aos utilizadores a combinação e obtenção de serviços *cloud* a partir de vários fornecedores com possibilidade de uma fácil mudança entre fornecedores de serviços de *cloud*.
3. Implementação de normas e diretrizes regulamentares, com vários critérios em relação aos serviços em *cloud*, tendo em conta aspetos como a segurança, proteção de dados, qualidade dos serviços e responsabilidade para os utilizadores específicos dos serviços.

A expetativa específica de negócio que está relacionada com o *Cloud Computing* numa organização depende das suas características mas também de algumas circunstâncias particulares como:

- **Custo de Contenção:** Com o *Cloud Computing* e a possibilidade de escalabilidade as organizações já não necessitam de investir dinheiro na construção e manutenção da sua própria infraestrutura, os serviços e os recursos necessários estão disponíveis no modo de *pay per use*. Assim as organizações não precisam de gastar dinheiro em recursos internos, que na maioria das organizações não seriam aproveitados. A economia que a organização obtém pode ser investida na ajuda a inovação do negócio. Antes de se optar por uma solução de TI em nuvem deve-se ter em conta o custo de TI que a empresa possui e os potenciais custos que uma solução em nuvem pode trazer para a organização.

- **Velocidade de Inovação:** Em comparação com a implementação de uma infraestrutura interna que pode ter um tempo de implementação de semanas ou meses para a organização, a opção por serviços em *cloud* pode ser efetuada em apenas algumas horas. Esta resposta rápida que a organização obtém permite uma rápida adequação às exigências do mercado com um custo acessível.
- **Disponibilidade:** A grande maioria das organizações que oferecem serviços em *cloud*, oferecem também maior capacidade de escala, interligação redundante e balanceamento de carga, ajudando a tempos de resposta a requisitos de negócio com alta disponibilidade. Esta disponibilidade de serviços deve ser assegurada por uma entidade independente em relação a organização que oferece os serviços de *cloud*, os níveis de serviço devem ser assegurados através de *Service Level Agreement* (SLA).
- **Escalabilidade:** A flexibilidade e escalabilidade de serviços em *cloud*, permitem a rápida adoção das TI as constantes mudanças de necessidades de negócio, procurando melhorar o

tempo de reação face à constante concorrência cada vez mais global. A resposta rápida as necessidades, com serviços disponíveis e escaláveis que permitem uma rápida adequação aos desafios que os mercados colocam são a chave para o sucesso da *cloud*.

- **Eficiência:** As organizações podem investir no seu *core business*, de uma forma inovadora, através de pesquisas e desenvolvimentos, mas a *cloud* permite uma maior sustentabilidade no crescimento e na competitividade de uma organização.
- **Elasticidade:** As organizações que oferecem serviços em *cloud* têm de possuir sistemas que podem ser utilizados por exemplo para a recuperação em caso de catástrofes, tudo isto é conseguido através do balanceamento de cargas e implementação da separação geográfica das salas de servidores, sendo esta uma forma de proteção da solução em nuvem contra desastres naturais.

A opção das organizações pela utilização de soluções em *cloud*, permite uma maior concentração das mesmas nos seus negócios e na inovação, ficando as preocupações com a infraestrutura delegadas para o fornecedor de serviços em *cloud*, o mesmo deve ser capaz de efetuar operações de melhoria rápida com custos eficientes, através de um processo de melhoria contínua.

A *cloud* também possui alguns modelos que permitem a sua implementação como soluções comerciais. Atualmente os modelos existentes no mercado são:

- **Private Cloud:** Neste tipo de solução o utilizador é uma organização específica ou uma unidade organizacional, esta solução pode ser interna à organização ou contratada a uma organização que forneça serviços em *cloud*. As vantagens da *cloud* não podem ser plenamente exploradas através deste modelo devido ao grau de personalização ser bastante limitado.
- **Community Cloud:** O serviço é utilizado por vários membros de um grupo, e estes serviços podem ser oferecidos por vários fornecedores que, por sua vez, podem ser internos ou externos à comunidade.
- **Public Cloud:** Serviços disponíveis para o público, em geral o mesmo é oferecido por um único fornecedor e neste modelo a estabilidade e os recursos como *pooling*, podem ser totalmente explorados.
- **Hibrid Cloud:** A *cloud* híbrida oferece uma conciliação que permite a organização várias possibilidade de combinação de vantagens e desvantagens, por exemplo os dados que necessitam de estar protegidos podem residir numa *Private Cloud*, enquanto os dados e aplicações públicas podem ser executados na *Public Cloud*.

Jim Reavis diretor executivo da *Cloud Security Alliance* (CSA), que é uma entidade não-governamental dedicada à segurança em ambientes de *cloud*, divulgou uma lista onde identifica sete itens de segurança em *cloud* que devem ser alvo de muita atenção (Yoon, 2011).

Os itens que devem ser tidos em atenção são (Nóbrega, 2010):

1. **Perda de dados**: De acordo com *Jim Reavis*, não está definido um nível mínimo de controlo de segurança em *cloud* e desta forma as aplicações podem perder dados, o que resulta principalmente do mau controlo das API's, problemas de armazenamento ou uma fraca gestão das chaves de acesso ao sistema. A juntar a tudo isto está a não existência de uma política de destruição de dados, sendo que na maioria dos casos o que existe é uma identificação falsa que dá a informação ao cliente de que os dados foram removidos quando na realidade os dados apenas são retirados do índice e não são devidamente apagados.
2. **Vulnerabilidades das tecnologias de partilha**: Em *cloud* uma configuração errada pode ser duplicada num ambiente que é partilhado por vários servidores e máquinas virtuais, e neste contexto devem existir acordos ao nível de serviços como os SLA que assegurem a gestão de atualizações e as melhores práticas possíveis no que diz respeito à manutenção da rede e configuração de servidores.
3. **Pessoas maliciosas dentro da equipa:** É necessário confiar na equipa existente, mas também é necessário confiar no fornecedor de serviços. O fornecedor de serviços possui os seus próprios níveis de segurança sobre acesso a *data centers*, o que faz com que existam diferentes níveis de controlo sobre a equipa.
4. **Desvio de tráfego, contas e serviços:** Existem muitos dados, aplicações e recursos presentes na *cloud* e a autenticação para este tipo de serviços é feita de forma insegura, pois o acesso pode ser feito a todos os itens e pode obter-se o acesso a uma máquina virtual de um cliente por exemplo:
    a. Acesso à conta do cliente, sendo que, neste caso o cliente tem todo o conteúdo da sua máquina virtual exposta ao invasor.
    b. Acessos ao administrador da *cloud*, em que o invasor tem poder sobre todas as máquinas virtuais de todo os clientes o que por si constitui uma ameaça bem maior.
5. **Interfaces de programação de Aplicativos Inseguras (API's):** As API's inseguras permitem que os utilizadores mal-intencionados possam utilizar estes serviços para invadir as contas. Podem existir ameaças de segurança, como acontece com os ataques de *bad loads* e *cross-scripting* às *apps* da *Adobe* da *Microsoft*. Outro exemplo foi a forma como o *Google* foi atingido pelo governo chinês, ninguém sabe se o ataque foi feito pelo governo chinês ou alguém

mal-intencionado, mas a mensagem a reter deste caso é que este tipo de ataques pode também acontecer com bancos, companhias de seguros e empresas fornecedoras de energia.

6. **Abuso do uso nefasto de *Cloud Computing*:** Quando os serviços de *cloud* sofrem acesso de pessoas não autorizadas e com fins mal-intencionados e que constituem uma ameaça para o negócio. A utilização nefasta da *cloud* deixou muitos utilizadores com medo de possíveis ataques *botnet* na Amazon *Web Services* (AWS). Isto não é apenas um impacto direto mas também afeta os utilizadores que dividem a nuvem com os transgressores.
7. **Perfil de risco desconhecido:** Se é verdade que a existência de transparência facilita muitas coisas para quem desenvolve a *cloud*, por outro lado a transparência faz com que os clientes apenas consigam ver uma interface, sem saber informações sobre as infraestruturas e níveis de segurança associados aos serviços que estão a contratar.

Quando decidimos optar pela utilização de *cloud* existe um conjunto de riscos que devem ser avaliados e tidos em conta. Por exemplo, em primeiro lugar, é importante avaliar a importância do ativo (dados e processo) imaginando vários cenários através de diversas perguntas que permitam perceber de que forma os ativos e os processos podem estar expostos (Kao, et al., 2012):

- Em que medida poderíamos ser prejudicados se o ativo fosse largamente divulgado e se fosse tornado público?
- Em que medida seríamos prejudicados se um funcionário do fornecedor do serviço de *cloud* acedesse a dados e processos indevidamente?
- Em que medida poderíamos ser prejudicados se alguém externo à organização executasse funções ou processos sem as devidas permissões?
- Qual seria o impacto no caso de acontecer uma quebra nos serviços que impossibilitasse o acesso a dados e processos?
- Qual o impacto de uma alteração inesperada nos dados?

Através da criação dos cenários referidos atrás conseguimos avaliar para cada ativo o impacto de confidencialidade, integridade e disponibilidade no caso de todo ou apenas uma parte ficar na nuvem. O passo seguinte será avaliar os fornecedores de serviço de *cloud* procurando reunir o máximo de informação como qual a arquitetura utilizada quais os mecanismos de segurança e quais as políticas quer de segurança quer de recuperação em caso de falhas ou desastres. Importa também perceber qual o modelo de *cloud* (exemplo: Pública, Privada, Híbrida ou Comunitária) que melhor corresponderá às necessidades do negócio e da organização. Também é importante saber qual o fluxo de dados que sai da organização para a nuvem e da nuvem para a organização, assim como é importante aferir o modo como os dados se movem para *cloud* (youseff, et al., 2008).

A necessidade emergente de assegurar uma transição gradual das aplicações de infraestrutura empresarial atual para a infraestrutura em *cloud* é um dos grandes desafios para esta próxima geração de informáticos. Emergiu uma necessidade significativa de adquirir um profundo conhecimento das soluções atuais com uma caracterização bem definida das mesmas, no sentido de garantir que os SGBD's em *cloud* tenham o sucesso e eficácia que as bases de dados relacionais tiveram e ainda continuam a ter nos ambientes corporativos atuais (Chang, et al., 2006).

É verdade que existem entraves que dificultam que as organizações possam optar por soluções em *cloud*. A maioria das organizações, referem as questões de segurança e de privacidade como os dois principais obstáculos para a adoção de soluções em *cloud*, mesmo com um grande esforço por parte dos fornecedores de serviços de *cloud*, sempre tendo em conta os aspetos como a governança procurando garantir que os seus serviços cumprem as exigências dos seus clientes.

A especificação de parâmetros que procuram garantir a qualidade de serviços constitui um mecanismo essencial em ambientes onde o *outsourcing* é bastante utilizado. Em seguida, será discutida a importância de acordos de níveis de serviço que têm como objetivo explorar as garantias do seu uso em serviços de tecnologia da informação e segurança especificando de que forma esses acordos podem ser representados.

A definição de *Service Level Agreement* (SLA), permite a garantia de níveis de serviço desejados, é um facto que por si só os SLA's não garantem a qualidade, mas o mesmo serviço suporta um conjunto de mecanismos que permitem a monitorização e a identificação das responsabilidades bem como punições e compensações caso o acordo não seja cumprido (Berberova, et al., 2009).

Os SLA's são muito utilizados nos serviços de telecomunicações com o intuito de especificar as características técnicas que permitem a garantia dos serviços ao utilizador como por exemplo largura de banda, disponibilidade e taxas de erros (Berberova, et al., 2009).

Nos serviços associados às TI os SLA's podem e devem ser adotados mas de uma forma diferente da que é utilizada nos serviços de telecomunicações. Assim existe um acordo que passa a representar quer as expetativas dos utilizadores quer as dos fornecedores de serviços e desta forma define as obrigações que podem ser especificadas para cada uma das partes envolvidas (Muller, 1999).

Importa referir que as informações que constam no acordo devem ser diferenciadas, sendo que um SLA de TI deve ter em conta as seguintes informações (Bianco, et al., 2008):

- descrição dos serviços;
- descrição das partes envolvidas, ou seja, o utilizador do serviço e o fornecedor do serviço;
- níveis de serviço desejados;

- métricas utilizadas para a monitorização do serviço;
- quem são os responsáveis pela monitorização;
- penalizações que serão aplicadas quando as obrigações especificadas não forem atendidas;
- mecanismo de evolução do SLA.

É muito importante referir que no contexto do SLA, o contínuo acompanhamento dos níveis de serviço bem como a sua especificação possui uma elevada importância daí que sejam utilizadas métricas que permitem avaliar o cumprimento das qualidades de serviço desejadas. A forma como estas métricas são avaliadas está dependente do tipo de serviço e dos tipos de características de qualidade que se deseja alcançar. Em resumo as qualidades de um SLA podem ser definidas como mensuráveis ou imensuráveis (Bianco, et al., 2008).

A tabela 5 permite identificar uma relação entre as qualidades mensuráveis e imensuráveis que estão associadas aos serviços de TI.

Consegue-se perceber a importância da definição de um SLA, mas também é importante compreender que um SLA não é um documento estático e a sua correta utilização depende do resultado de execução de várias atividades que são realizadas nos diferentes estágios da sua vida. Posto isto existe um conjunto de fases que estão associadas ao ciclo de vida de um SLA (TM Forum, 2005):

1. **Definição**: Nesta fase são identificadas as características do serviço e definidos os parâmetros de qualidade que devem ser disponibilizados aos utilizadores.
2. **Negociação**: A esta fase está associada a definição dos parâmetros do serviço, os custos para os utilizadores e as penalizações em cajo de incumprimento.
3. **Implementação**: Fase em que o serviço é preparado para o utilizador.
4. **Execução**: Esta fase está associada às operações de monitorização dos serviços, sendo o objetivo a avaliação dos parâmetros de qualidade que foram especificados e a verificação do cumprimento do SLA.
5. **Avaliação**: Nesta fase o fornecedor de serviços avalia a qualidade do serviço que fornece.
6. **Finalização**: É nesta fase que são tratadas as questões relacionadas com a finalização do serviço, que podem ser a expiração do contrato ou a violação do SLA definido.

Com a crescente importância da *cloud* também aumentarem as preocupações com questões como a segurança e a privacidade, no entanto a especificação de SLA's que envolvam características de segurança e que podem ser denominados de SecuritySLA, constitui alguns desafios que envolvem uma especificação dos níveis de segurança e a representação de acordo e a constante monitorização (Jaatun, et al., 2012)

Qualidades Mensuráveis	
Precisão	Limite de taxas de erros para os serviços durante um determinado período de tempo
Disponibilidade	Probabilidade de disponibilidade do serviço quando necessário
Capacidade	Número de solicitações concorrentes que o sistema é capaz de suportar
Custo	Custos dos Serviços
Latência	Tempo máximo entre a chegada das solicitações e a resposta a essas solicitações.
Tempo de provisionamento	Tempo necessário para que o serviço se torne operacional
Confiabilidade das mensagens	Garantia da entrega das mensagens
Escalabilidade	Capacidade do serviço aumentar o número de operações executadas com sucesso num determinado período de tempo.
Qualidades Não mensuráveis	
Interoperabilidade	Capacidade de comunicação com outros serviços.
Possibilidade de Mudança	Frequência com que as modificações ocorrem num determinado serviço
Segurança	Capacidade do serviço para resistir à utilização não autorizada, ao mesmo tempo que fornece serviços para clientes legítimos

Tabela 5 - Exemplo da qualidade nos serviços de TI

A definição de parâmetros de segurança pode ser efetuada de duas formas: 1 através de políticas de segurança, 2 a partir de métricas de segurança (Casola, et al., 2006).

A especificação de um SLA através de políticas de segurança, implica um conjunto de políticas que devem ser expressas em linguagens padrão como os, *web services*. Embora seja permitido especificar de forma clara os níveis de segurança desejados, existe uma falha ao nível da especificação dos mecanismos de monitorização, não tendo em conta muitas das informações que integram um SLA (Putri & Mganga, 2011).

Por outro lado as métricas de seguranças são bastante aceites pois permitem especificar os parâmetros de segurança mas também a maneira como a monitorização deve ser feita, pois as métricas possuem como base um conjunto de métricas teóricas de segurança que verificam se um determinado objetivo está ou não a ser cumprido (Krautsevich, et al., 2010).

A segurança que é uma qualidade não tem uma medida, tem como base uma premissa de que não existe um modelo não ambíguo e amplamente aceite que permite definir se um sistema é mais seguro que outro. No entanto as métricas de segurança dispõem de um conjunto de ferramentas que permite a obtenção de informações correta e com atualização contínua sobre o estado de segurança de um

determinado ambiente, permitindo avaliar operações e controlo de segurança num determinado ambiente. Desta forma as métricas de segurança podem ser classificadas da seguinte forma (Chew, et al., 2008):

1. **Implementação**: Mostram a evolução da implementação de programas, controlo, procedimento e políticas de segurança.
2. **Eficácia e Eficiência**: Permitem a monitorização correta de implementações de controlo e processos de segurança.
3. **Impacto**: Permite medir o efeito da segurança da informação na missão das organizações.

As métricas de segurança são também classificadas de acordo com a sua audiência, sendo as mesas divididas em três categorias (The Center for Internet Security., 2014):

- **Gestão**: Permite obter informações sobre o desempenho das funções de negócio e o seu impacto na organização.
- **Operacionais**: Ajudam na compreensão e otimização das atividades associadas às funções de negócio
- **Técnicas**: Permitem um detalhe técnico, fornecendo um suporte para outras categorias e métricas.

A tabela 6 apresenta um exemplo de métricas de segurança que estão relacionadas com a gestão de incidentes e classificadas de acordo com as categorias apresentadas:

Categorias	Métricas
Gestão	○ Custos dos incidentes ○ Custo médio dos incidentes ○ Percentagem de sistemas sem vulnerabilidades severas ○ Observância à política de atualização ○ Percentagem de observância de configuração ○ Percentagem de gastos com segurança de acordo com o orçamento de TI.
Operacional	○ Tempo médio gasto descobrir incidentes ○ Tempo médio entre ocorrências de incidentes ○ Tempo médio de recuperação de incidentes ○ Tempo médio para a realização de atualizações ○ Tempo médio para a eliminação de vulnerabilidades
Técnica	○ Número de incidentes registados ○ Número de vulnerabilidades conhecidas ○ Cobertura e níveis de verificação de vulnerabilidades ○ Observância da gestão de configurações ○ Observância da gestão de atualizações

Tabela 6 - Métricas de segurança e categorias

Mesmo com o criterioso processo de definição de métricas de segurança que procuram promover a transparência e apoiar os processos de decisão, a previsibilidade e planeamento pró-ativo, juntamente com os processos de definição destas métricas não é uma tarefa simples visto que as suas características de segurança não são facilmente quantificadas.

A especificação de métricas para serem utilizadas em *Security*-SLA tendo em conta as políticas de segurança são baseadas na metodologia de *Henning*. Esta metodologia utiliza um conjunto de políticas, normas e padrões de segurança como base para criação de métricas. O processo de decisão está associado a três etapas: 1 analisar as políticas especificando um conjunto de métricas básicas que devem ser trabalhadas, 2 analisar a arquitetura com o objetivo de validar os requisitos definidos na etapa anterior, de forma que as mesmas sejam implementáveis na arquitetura computacional existente, 3 realização de uma entrevista com os utilizadores do serviço de forma a perceber as suas preocupações com a segurança no ponto de vista de utilizador (Henning, 1999).

No que respeita à segurança a metodologia que é aplicada em diversos trabalhos é o modelo de *Goal Question Metric* (GQM) (Putri & Mganga, 2011) que é utilizado em conjunto com a *framework* de *Control Objectives for Information and Related Technology* (COBIT), (ISACA, 2012), que possibilita uma especificação de métricas para o *Security-SLA*, em *cloud*. A utilização do GQM permite a definição de uma hierarquia de métricas utilizadas que por sua vez, permite definir um índice geral de segurança em sistemas *cloud* (Silva, et al., 2012).

As formas de representação de um SLA são várias podem ser feitas através de um conjunto de formatos que vão desde a representação textual, em que os níveis de serviço são descritos através de linguagem textual, ou através de representações que utilizam linguagens para um determinado fim específico (Bianco, et al, 2008).

A utilização de linguagem natural pode ajudar em grande escala a representação deste tipo de acordos, pois a mesma garante uma maior flexibilidade e facilidade de entendimento dos mesmos, mas não é aconselhado utilizar esta linguagem quando o objetivo está relacionado com as questões ligadas à monitorização automática, portabilidade e manutenção destes documentos. A utilização de acordos que são desenvolvidos em linguagem padronizada permitem uma utilização direta em sistemas computacionais possui algumas vantagens (Bianco, etal, 2008):

1. Permite o suporte a ferramentas de negociação automáticas
2. Permite o controlo automático dos custos dos serviços
3. Permite a utilização de controlos automáticos que permitem uma verificação constante do cumprimento dos SLA's.

4. Permite detalhar os mecanismos automáticos que têm como função efetuar ações de notificação em resposta a eventos ou violações dos SLA.

A maioria das linguagens de representação de SLA é no formato XML, que permite uma portabilidade dos documentos produzidos. Assim sendo alguns dos possíveis SLA que pode ser obtidos através desta linguagem são:

- **SLAng**: (Lamanna, et al., 2003): é uma linguagem orientada para a definição de SLA's mais complexos e que envolvem diferentes tipos de serviços como rede, armazenamento e aplicações, tendo como base um modelo de aprovisionamento com três domínios (aplicações, camada intermédia e recursos subjacentes), e seis tipos de partes: aplicações, servidor web, componentes, *container*, armazenamento e rede. Neste SLA os acordos podem ser classificados de duas formas distintas, horizontais e verticais. No caso dos SLA's horizontais, os acordos são estabelecidos por duas partes ao nível da arquitetura e podem ter em conta serviços similares. Já os SLA's Verticais são descritos através de acordos entre as diferentes partes nos vários níveis da arquitetura.
- **Web Service Level Agreement Language *(WSLA)*** (Ludwig, et al., 2003): É uma especificação proposta pela IBM para criação de acordos para *web services* que permitam a descrição de serviços sendo de referir que as qualidades associadas a esses serviços que devem ser alvo de uma garantia. Um WSLA é constituído por três secções: 1 identificação das partes envolvidas no acordo, 2 definição dos serviços no qual são descritas as características associadas aos serviços, 3 definição dos objetivos que permitem descrever as obrigações tendo em conta os responsáveis pelas ações que são executadas em casos da existência de violação. Um WSLA possui uma estrutura que viabiliza a representação de métricas com alguma complexidade tendo como base expressões que relacionam outras métricas. A linguagem é constituída por operadores aritméticos e lógicas funções de manipulação de séries temporais, sendo utilizadas especificações de expressões lógicas para a monitorização de níveis de serviço.
- **WS-Agreement** (Andrieux, et al., 2007): Esta é uma linguagem que permite a publicação de capacidades por parte dos fornecedores de serviço, tais como: negociação, representação de acordos e monitorização em tempo real. É uma linguagem que pode ser utilizada para representar acordos em qualquer domínio, de serviços, utilizando desta forma outras linguagens para o domínio, onde pode ser utilizada, como por exemplo *WebService Definition Language (*WSDL*)*, *Job Submission Description Language (*JSDL*),* entre outras. Um acordo realizado nesta linguagem é composto por três componentes: 1 identificação de uma secção opcional que identifica o acordo: 2 contexto secção mandatária que especifica as partes envolvidas na

comunicação e a duração do acordo: 3 termos que descrevem o serviço que é oferecido e as suas garantias.

- **Sec-Agreement** (Hale & Gamble, 2012): Esta é uma extensão que está voltada para a representação de acordos de segurança que utilizam a linguagem *WS-Agreement*, possuindo um conjunto de novos objetos semânticos e operações que permitem a expressão de políticas de segurança nos termos de descrição do serviço e níveis de garantia associados ao serviço.
- **WS-Policy** (W3C, 2007): É uma especificação que possui mecanismos que abrangem aplicações baseadas em *web services* e que permitem especificar as capacidades, requisitos e características gerais das entidades através de declarações de XML. A *WSPolicy*, em particular, possui uma gramática flexível e extensível e que por isso proporciona uma especificação de diferentes tipos de políticas utilizando extensões. A representação de políticas de segurança é efetuada através da extensão *WS-Security Policy*, mesmo não sendo esta uma linguagem para a representação de SLA's, sendo que esta *framework*, utiliza a representação de *Security-SLA*.

As componentes de segurança de dados são aplicadas em vários pontos diferentes, incluindo a segurança na rede. Enquanto os dados estão em estado de passagem ou no ponto em que os dados estão num armazenamento com potencial de espionagem, os dados precisam de segurança a nível da base de dados. Os níveis de segurança vão desde detalhes pessoais, financeiros, até aos detalhes de segurança nacional (Casola, et al., 2006).

De forma a garantir que a segurança e a privacidade dos dados são garantidas é necessário que exista um conjunto de políticas e procedimentos que permitam aos utilizadores de sistemas de *cloud* confiar nas garantias de integridade, confiabilidade e segurança deste tipo de sistema, devendo ter em conta aspetos como:

- **Controlo de Acessos**: Geralmente quando nos referimos a controlo de acessos, estamos a referir-nos a um abrangente conjunto de funções de acesso e de requisitos quer para os utilizadores do sistema quer para os administradores de sistemas (privilégios dos utilizadores), nomeadamente quem pode aceder a rede, aos sistemas e aos recursos e aplicações. A gestão do controlo de acessos tem de ser capaz de responder às seguintes questões (Mather, et al., 2009):
    - Quem deve ter acesso ao sistema e a que recurso do sistema deve ter acesso? (Atribuição de direitos de acesso ao sistema por utilizador).
    - Porque é que o utilizador deve ter acesso a um determinado recurso? (Este acesso deve ser dado de acordo com as responsabilidade e funções do utilizador).

- De que forma o utilizador deve aceder aos recursos? (Qual o método de autenticação, e a segurança necessário para dar as permissões de acesso ao sistema por parte do utilizador).
- Quem tem acesso para o recurso? (Verificar se o utilizador tem ou não autorização para aceder a um determinado recurso).

Os aspetos referidos no domínio do controlo de acessos devem ser definidos para a organização através da implementação de políticas de acessos *standard* que estejam alinhadas de acordo com as regras e as responsabilidades atribuídas a cada um dos utilizadores, incluindo também os utilizadores com privilégios como é o caso do administrador de sistemas (Mather, et al., 2009).

- **Controlo de Acesso em Cloud:** Com a utilização de sistemas em *cloud* os diversos utilizadores podem aceder aos serviços através de um qualquer local desde que possuam uma ligação à internet, e por esta razão os acessos à internet baseados em controlo e focados na proteção de recursos, evitando que estes sejam acedidos por pessoas não autorizadas, e que associados a acessos baseados em atributos tornam-se inadequados, pois os mesmos não são exclusivos dos utilizadores e podem causar imprecisões. No caso da *cloud* o controlo de acessos possui um manifesto com uma *firewall* de políticas que tem foco num reforço do controlo de acesso, que, por sua vez é baseado no anfitrião de entrada e saída através de um conjunto de pontos de entrada para a lógica da *cloud* com um agrupamento de instâncias dentro da *cloud*. Usualmente são utilizadas políticas (regras), que utilizam o *Transaction Control Protocol/ Internet Protocol* (TCP/IP) standard, com parâmetros que incluem IP de origem, porta de origem, IP de destino e porta de destino (Mather, et al., 2009).

  Em contraste com o controlo de acesso baseado em rede, o controlo de acesso dos utilizadores deve ser mais fortemente enfatizado na *cloud*. Uma vez que podem ser adicionadas muitas ligações de entidades de utilizadores ao mesmo tempo a *cloud* deve ser capaz de possuir um controlo acesso com granularidade, tendo em conta o número de utilizadores a aceder ao serviço, procurando garantir deste modo o cumprimento dos requisitos de proteção de dados. A gestão do controlo de acesso dos utilizadores deve possuir (Mather, et al., 2009):

  - Forte sistema de autenticação
  - *Single Sign-On* (SSO), desta forma o utilizador ao efetuar o seu *Login*, uma vez possui logo o acesso a todos os sistemas sem a necessidade de se voltar autenticar novamente em cada um dos sistemas. Normalmente este trabalho é realizado através do uso de *Lightweight Directory Access Protocol* (LDAP) e posterior armazenamento na

base de dados LDAP no servidor. Um simples exemplo de utilização do SSO é a utilização de *cookies* mas estas apenas funcionam para os sites que se encontram no mesmo domínio.

- Privilégios de Gestão
- *Logging* e monitorização permanente dos recursos de *cloud*
- Desempenho significativo no que respeita à proteção da confidencialidade e integridade da informação que se encontra na *cloud*.

A utilização da norma ISO/IEC 27002 que define 6 objetivos do controlo de acessos para todos os utilizadores finais de soluções em *cloud*, contempla informações sobre privilégios dos utilizadores, rede, aplicações e informações de controlo de acesso que têm como objetivo ajudar a diminuir ao máximo os riscos para o negócio (Mather, et al., 2009):

O objetivo da utilização da ISO/IEC 27002 é assegurar a devida autorização de acesso e prevenir acessos não autorizados à informação. Os procedimentos formais devem estar definidos num conjunto de estratégias juntamente com a definição de um ciclo de vida atribuído aos acessos dos utilizadores, desde o registo inicial do utilizador, contemplando também os utilizadores que possuem acesso ao sistema mas que já não o fazem há algum tempo, evitando que estes mesmos utilizadores que não usam o sistema e serviços de forma regular sejam tidos em atenção. Neste sentido é necessário existir um controlo de atribuição de privilégios e verificar se os acessos estão a ser feitos de acordo com o que foi definido, procurando evitar que os utilizadores possam substituir os controlos do sistema (Mather, et al., 2009).

Em seguida são descritas as seis regras de controlo definidas pela ISO/IEC 27002:

- Controlo de acesso a informação
- Gestão correta dos acessos realizados por cada utilizador
- Encorajar a práticas de acesso corretos
- Controlo de acesso do serviço de rede
- Controlo de acesso do sistema operativo
- Controlo de acesso das aplicações e dos sistemas.

A *Information, Tecnology Infrastructure Library* (ITIL), que é dedicado às funções de gestão de acesso, que sofreu algumas alterações com o ITIL V3, inclui processos dedicados e orientados para a segurança de TI é uma das razões para que a perspetiva de segurança de TI permita uma garantia de acesso aos serviços de TI e aplicações apenas por quem possui autorização. Este deve ser um ponto de elevada importância que deve ser tido em conta por todos os utilizadores de serviços de TI (Mather, et al., 2009).

O objetivo da pessoa que desempenha estas funções é garantir ou conceder a autorização do acesso correto de um determinado utilizador a um serviço, sempre com o objetivo em mente de impedir os acesos não autorizados. O processo de gestão de acessos, é executado através de uma definição de políticas de gestão e segurança de TI (Mather, et al., 2009).

De forma a poder gerir a disponibilidade das aplicações são necessárias métricas, monitorização e níveis de gestão de serviços de acordo com a perspetiva de quem está a gerir. Infelizmente há falta de *standards* e capacidades como um CSP para ajudar os clientes. Devido à impossibilidade de poder contar com estes recursos e possibilidades, é necessário perceber o que o fornecedor de serviços oferece e que ajuda os utilizadores a efetuar da melhor forma a gestão dos níveis de serviço. Posto isto a seguinte tabela procura sumarizar as responsabilidades de monitorização de segurança de acordo com as perspetivas do cliente (Mather, et al., 2009), a tabela 7 apresenta as responsabilidades no que respeita a monitorização de segurança dos serviços.

Na perspetiva da gestão da segurança, os aspetos chave prendem-se com a falta de acesso dos responsáveis empresariais aos recursos de gestão desde os recursos de controlo de acessos que é apenas um dos muitos módulos dos serviços de prestação de serviços quer para fornecedores quer para clientes, que devem entender que os recursos de controlo de acessos estão disponíveis através de uma autenticação forte em que os utilizadores conseguem perceber quais são as suas responsabilidades no que diz respeito à gestão do ciclo de vida dos acessos aos serviços de *cloud*. Muitos fornecedores de serviços estão a fazer um grande esforço para manter os seus clientes informados sobre as novas tendências, procurando educá-los sobre as formas de proteger a informação que está armazenada na *cloud*. Por exemplo a Salesforce.com publicou ameaças e práticas de segurança de informação através do http://trust.salesforce.com. Contudo a maior parte está na forma como o cliente monitoriza e gere as ameaças e os riscos associados aos seus serviços de *cloud* (Mather, et al., 2009).

Atividades de Monitorização	IaaS	PaaS	SaaS
Monitorização de Rede	Monitorização das interfaces das máquinas virtuais	Responsabilidade do fornecedor de serviços (métricas não disponíveis para o cliente)	Responsabilidade do fornecedor (métricas não disponíveis para o cliente)
Monitorização do *Host*	Monitorização segura de eventos a partir de *host* IDS's tais como OSSEC Log eventos dedicados e log de persistência no servidor Monitorização de eventos de segurança da base de dados da máquina virtual e *logs* de sistemas.	Responsabilidade do fornecedor (métricas não disponíveis para o cliente).	Responsabilidade do fornecedor (métricas não disponíveis para o cliente).
Monitorização da Base de Dados	Instalar as ferramentas de monitorização segura da base de dados e dos *logs* com eventos dedicados e persistentes dos *logs* de servidor.	Responsabilidade do fornecedor (métricas não disponíveis para o cliente).	Responsabilidade do fornecedor (métricas não disponíveis para o cliente).
Monitorização de Aplicações	Monitorização das vulnerabilidades das aplicações (OWASP TOP 10), bem como as aplicações de eventos com os *logs* de intrusões	Monitorização das aplicações e dos *logs* para as vulnerabilidades (talvez disponíveis para a plataforma PaaS).	Responsabilidade do fornecedor

Tabela 7 - Responsabilidades do Cliente no que Respeita à Monitorização da Segurança dos Serviços

A utilização de sistemas virtualizados, com infraestruturas partilhadas em que os dados encontram-se misturados com os dados de outros clientes, permite que em todas as fases do ciclo de vida exista um processamento e armazenamento contínuo. Mas se não possuir as ferramentas corretas que possibilitam, através da identificação dos problemas e das ameaças, que a ação possa ser imediata de forma a evitar danos importantes para o sistema.

Em conclusão o escopo da gestão da segurança dos serviços de *cloud* variará de acordo com o modelo de serviços, as capacidades do fornecedor e a sua maturidade para lidar com os problemas que possam surgir. Os clientes apenas terão a possibilidade de optar pelos serviços de um determinado fornecedor tendo em conta aspetos como a flexibilidade e controlo de serviços que é oferecido pelo *Service Provider Infrastructured* (SPI). O serviço mais flexível (com menor abstração de serviços), com maior controlo que pode ser exercido sobre o serviço e que com tudo isto consiga uma gestão das responsabilidades de segurança e que consiga oferecer serviços sem falta de transparência na área dos SLA's, proporciona ao fornecedor de serviços capacidades de gestão e segurança com responsabilidades, gestão de funções com foco no processo de mudanças contínuas e utilização de ferramentas que permitam garantir características como confiabilidade, disponibilidade e segurança que possam ser estendidas aos serviços de *cloud*. A adoção de um *standard* de TI como o Isso 27002 ou o ITIL na organização irá permitir uma

revisão e um ajustamento contínuo das capacidades dos serviços baseados em *cloud* e, juntamente com a sensibilidade da informação e a utilização dos SLA's irá viabilizar uma grande variedade de funções de gestão e, por conseguinte, promover a implementação de um ciclo de vida e de melhoria contínua para os serviços de *cloud* (Mather, et al., 2009)

Em relação a privacidade dos dados a mesma pode e deve ser implementada através de um processo contínuo e de acordo com o que foi apresentado anteriormente em relação à segurança a proposta para controlo de objetivos de segurança de TI pode e deve ser definida de acordo com a Isso 27001 que possui uma descrição bastante ilustrativa sobre o controlo de objetivos em particular no que se refere à relevância do *Cloud Service Provider* (CSP). O controlo adicional dos objetivos e a sua aplicação estão em parte dependentes da natureza dos serviços que podem ser oferecidos pelo CSP's.

Desta forma a gestão de ativos e o controlo de acessos deve estar de acordo com a proteção de dados, segmentação e encriptação fornecendo:

- Lógica de segregação do CSP's para com os dados dos clientes
- Possibilidade de classificação dos clientes de acordo a sensibilidade dos seus dados
- Proteção dos dados de acordo com a classificação dos riscos definidos para a informação.

No que diz respeito à aquisição de sistemas de informação, desenvolvimento e manutenção dos mesmos, devem ser tidas em conta as encriptações *standard* com a possibilidade de utilização de um mecanismo que permita a encriptação de dados sensíveis (Mather, et al., 2009).

Deve também ser feito um esforço para que a comunicação das operações de gestão que sejam efetuadas seja eficaz, normalmente através da realização de um registo de todos os *logs*, o que permite que, quando necessário possa ser efetuada uma auditoria que tenha em conta os *logs* relevantes das ações que foram executas por um determinado utilizador, podendo, através desta informação, ser efetuada uma revisão aos procedimentos sobre o comportamento de um determinado utilizador perante o sistema. Neste contexto, a periodicidade das revisões, principalmente quando se trata de dados com alto riscos é muito importante pois a realização de auditorias permite detetar os risco e tomar medidas de prevenção que sejam necessárias para evitar que o risco possa prejudicar o sistema (Mather, et al., 2009).

2.3 Enquadramento de *Big Data* com *Cloud Computing*

A utilização de *Cloud Computing* e *Big Data*, podem num curto prazo tornar-se: soluções inovadoras, com baixos custos e acessíveis, a possibilidade de efetuar análises ao negócio em qualquer parte do mundo e através de uma qualquer dispositivo móvel ou portátil, torna-.se numa vantagem que permite a criação de soluções económicas e que não exigem nenhuma implementação complicada, devido à simplicidade de utilização.

As aplicações que estão voltadas para a internet normalmente originam uma grande quantidade de dados, os novos sistemas de gestão de dados devem ser capazes de lidar com um grande volume de atualizações.. É importante no entanto destacar o nível de mudança nas aplicações, nomeadamente no que se refere ao *design* dos sistemas, para que estes novos sistemas alcancem o sucesso (Curino, et al., 2010).

O primeiro passo para a realização deste novo tipo de análises é a definição das bases para uma gestão escalável de dados. É necessário a compreensão desta definição, tendo como foco a projeção de sistemas capazes de interagir com aplicações web, possibilitando a fácil integração dos mesmos com ambientes *cloud*. Existe dois aspetos que devem ser alvo de destaque tendo em conta o que foi anteriormente descrito (Das, et al., 2010):

- Em primeiro lugar o sistema deve possuir um DBMS escalável que permita grandes quantidades de dados.
- Em segundo terá que suportar um grande número de aplicações, podendo cada uma das aplicações possuir uma pequena quantidade de dados, sendo que o conjunto das aplicações dará origem a uma grande quantidade de dados, o que fará com que o DBMS além de crescer necessite de ser facilmente adaptado de acordo com as necessidades que aparecam.

A figura 7, procura demonstrar uma perspetiva sobre os diferentes modelos de *multi-tenancy*.

Os grandes desafios que são colocados ao *Big Data*, prendem-se com: a gestão de dados de grandes aplicações, que deve ser efetuada de acordo com o modelo de *Cloud Computing* (Das & Nishimura, et al., 2010).

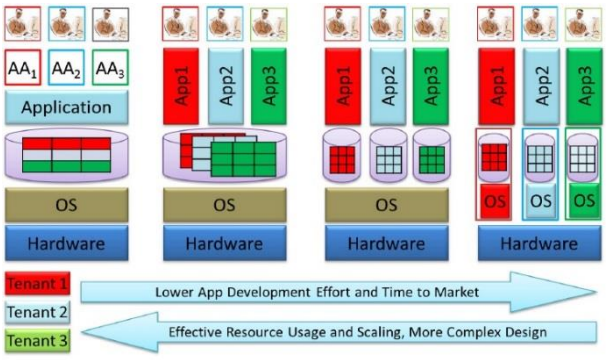


Figura 7 - Arquitetura de Um Sistema *Multi-Tenacy* adaptado de (Agrawal, et al, 2010).

Devido a o crescente aumento da quantidade de dados, tendo em conta estes factos, os sistemas de gestão de dados que trabalham de acordo com o modelo relacional, têm muitas dificuldades em lidar com novos tipos de dados, pois os mesmos aumentam a complexidade do sistema (Dean & Ghemawat, 2004). Desta forma nota-se um crescimento e popularidade de sistemas de código aberto como o Big Table, Hbase, Cassandra, MongoDB entre outros sistemas, que estão atualmente a ser implementados em larga escala principalmente nos modelos de *clouds* privadas, públicas e híbrida (Chang, et al., 2006).

Os sistemas anteriormente referidos têm como base o *Not Only Structured Query Language* (NoSQL), que é baseado em princípios de conceção como: escalabilidade e fácil comparação com os sistemas relacionais (Reinwald, 2010) .

As aplicações em *cloud* necessitam de apoiar um grande número de aplicações, que são acedidas por diferentes utilizadores. Neste âmbito, importa mencionar, os já falados sistemas *multi-tenancy*, dos quais é exemplo o SaaS (exemplo: *Salesforce.com)*, que é um sistema no qual muitos utilizadores partilham a mesma tabela de base de dados. Importa referir que existem diferentes modelos de *multi-tenancy* que revelam diferentes contextos e paradigmas de *cloud* (Jacobs & Aulbach, 2007).

A utilização de sistemas de análises de dados tem sido anunciado como a próxima evolução no que se refere à inteligência de negócios. Este facto tem como principais impulsionadores os serviços de *cloud*, que oferecem níveis de agilidade e elasticidade, através de um armazenamento escalável, com um poder de computação e de recursos flexíveis.

O constante crescimento do volume de dados antevê a necessidade de um aumento da utilização de *Data Mining*, para que se consiga encontrar valor na grande imensidão de informação que está ao nosso dispor.

A utilização de tecnologias como o Hadoop que permitem um escalonamento e distribuição do processamento por vários *clusters*, torna a solução de *Big Data* bastante escalável e com capacidade de responder a todas as necessidades que lhe sejam impostas. A adoção em larga escala de uma gestão de *clusters* através do *Big Data Analytics* requer uma variedade de mudanças e uma disponibilidade de sistemas e regras que ajudam no sucesso de utilização e implementação destas soluções nas organizações.

A manipulação de *Big Data* mudou um pouco o paradigma, no sentido em que as análises já não serão restringidas, existe uma inversão completa na gestão de dados em que os sistemas de processamento adicionam um novo modelo de processamento com o processamento paralelo e distribuído em que o

A tecnologia MapReduce presente no Hadoop e que apareceu em 2004, quando a *Google* percebeu as suas vantagens, o que permitiu o desenvolvimento do *clustering* como uma forma de resolver problemas associados as grandes quantidades de dados através de um modelo de processamento paralelo.

Existem ainda algumas incompatibilidades entre as ferramentas de BI e o MapReduce o principal problema está em adequar o modelo relacional ao MapReduce. No entanto podem existir ratefas de conversão através da utilização do Hive e da combinação de processamentos MapReduce com as ferramentas BI existentes (ex: Hortonworks e Cloudera).

3 – Arquitetura do Sistema de Análise de Dados *Big Data* no Modelo *Cloud Computing*

Este capítulo descreve o desenvolvimento da arquitetura de análise de dados *Big Data* no modelo de *Cloud Computing* que será explicada recorrendo a dois modelos:

1. **Arquitetura Conceptual:** descreve os níveis que constituem a arquitetura e a explicação das atividades que são realizadas em cada um dos níveis. Esta arquitetura é apresentada na secção 3.1.
2. **Arquitetura Física**: descreve uma solução tecnológica, através da instanciação de tecnologias para cada um dos níveis identificados na arquitetura conceptual. Realça-se que poderão existir outras soluções tecnológicas distintas da apresentada. Esta arquitetura é apresentada na secção 3.2.

As fontes de dados (***Data Sources***), representadas em ambas arquiteturas procuram identificar as diferentes origens e tipos de dados que podem ser utilizados, nomeadamente dados provenientes de ERPs, CRMs, SCMs, redes sociais, ficheiros de texto, vídeos, etc.. Numa fase inicial os dados poderão ter origem em Microsoft SQL Server, *Comma Separated Values* (CSV) e *Flat Files*, devido ao facto de, nas organizações de pequena e média dimensão, existirem uma enorme quantidade de dados com origem nas fontes anteriormente descritas, no entanto a arquitetura suporta dados provenientes de outras fontes.

3.1 Arquitetura Conceptual

A arquitetura proposta é composta por dois níveis, sendo que cada nível suporta um conjunto de atividades a ele associadas. Estas vão desde a recolha dos dados até à disponibilização ao utilizador dos resultados obtidos pelas análises realizadas aos dados. Os dois níveis que constituem a arquitetura são: ***Data Stanging Area*** e ***Data Analysis and Visualization***.

3.1.1 *Data Staging Area*

Este nível está responsável pelo processo denominado *Extract Transform and Load* (ETL) que compreende as atividades de extração dos dados de várias fontes, transformação e limpeza dos mesmos, de forma a assegurar que posteriormente os dados tratados são carregados para uma área de armazenamento. As diferentes fontes de dados podem ter várias origens e os dados podem ser estruturados, semiestruturados e não estruturados.

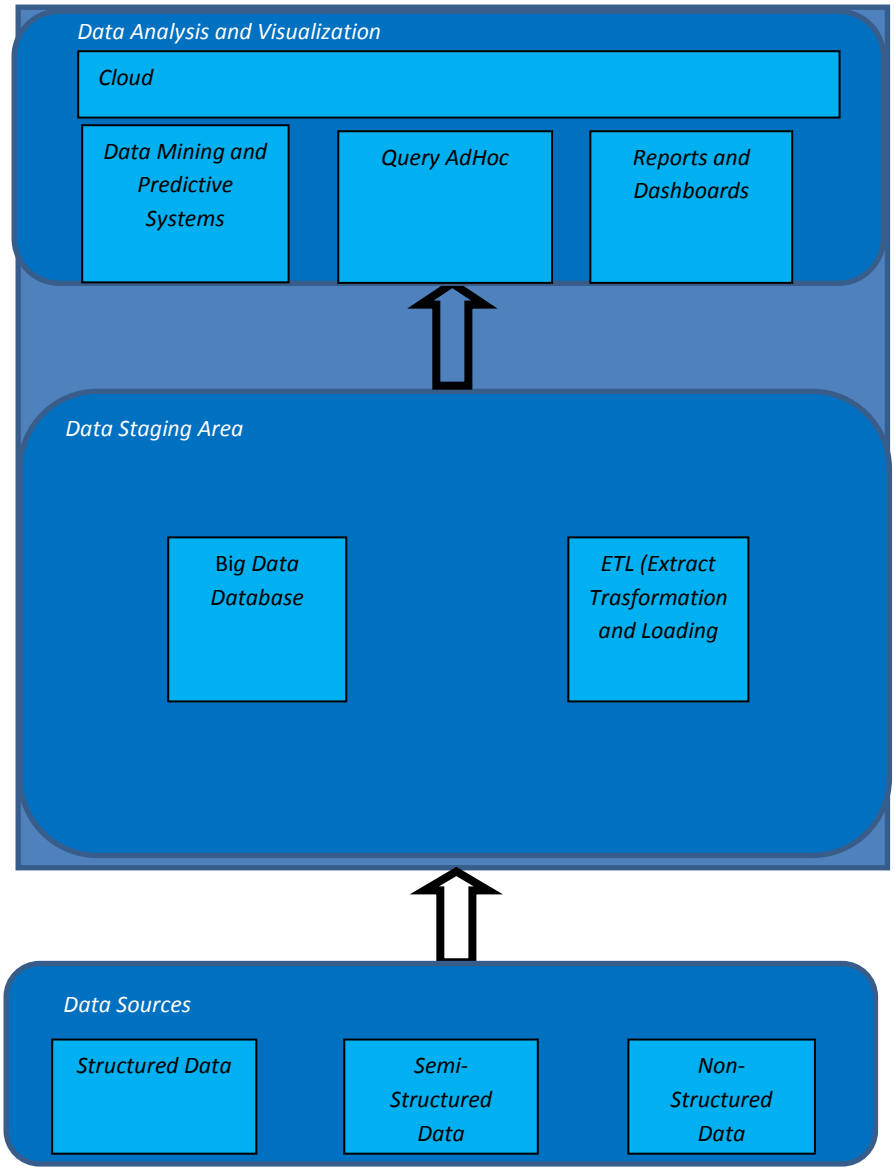


Figura 8 - Arquitetura Conceptual

As tarefas a executar são:

- Limpar os dados.
- Detetar e corrigir erros.
- Extrair dados para a realização de análises. Os dados são selecionados de acordo com as análises a realizar, mas também podem ser realizadas análises *ad-hoc*,.
- Armazenar os dados numa base de Dados *Big Data* (exemplo: Hive), que irá guardar os dados que serão utilizados pelo *Data Analysis and Visualization*.

O armazenamento dos dados que foram tratados com a utilização do processo ETL e que depois serão utilizados para o tratamento analítico no nível de *Data Analysis and Visualization*. Um aspeto importante associado à base de dados *Big Data* prende-se com o facto de a mesma se constituir como um repositório capaz de armazenar diversos tipos e origens de dados.

3.1.2 *Data Analysis and Visualization*

O nível de *Data Analysis and Vizualization*, é o responsável pela definição dos indicadores de negócio a analisar e desta forma é possível definir um conjunto de métricas que permitem suportar as decisões que a organização decida tomar. A possibilidade de disponibilização dos resultados das análises serem disponibilizados através da web é bastante pertinente, a mesma está relacionada com a necessidade de um acesso rápido aos dados a partir de um qualquer local a uma qualquer hora, permitindo que esses acessos possam ser feitos por um qualquer dispositivo com ligação à internet como por exemplo dispositivos móveis (*smartphones* e *tablets*)

Existem alguns requisitos que foram tidos em conta para a escolha das ferramentas tecnológicas, nomeadamente a opção por ferramentas *open source*, mas também serão exploradas ferramentas proprietárias, o que permite a opção por um maior grupo de ferramentas tecnológicas, procurando ter uma maior perceção de mercado sobre as opções tecnológicas que existem nesta área. No entanto existem alguns requisitos que devem ser tidos em conta na hora da escolha:

- Ferramentas tecnológicas *open source* com funcionamento em ambiente *cloud*.
- Ferramentas tecnológicas que analisem os dados e disponibilizem os resultados na web e que possuam mecanismos de segurança que permitam um acesso autorizado e seguro, viabilizando a consulta dos resultados através de diversos dispositivos (exemplo: computador pessoal, *tablet* e *smartphone*).

3.2 Arquitetura Física

Para cada um dos níveis que constituem a arquitetura são identificadas tecnologias que possibilitam a implementação da solução. A escolha dessas tecnologias baseou-se principalmente na facilidade de instalação, configuração e utilização das mesmas, uma vez que, se escolheram tecnologias de fabricantes conceituados e com valor no mercado. Será, ainda, realizada uma descrição das tarefas a efetuar por cada um dos níveis da arquitetura, a figura 9 mostra uma proposta da arquitetura tecnológica.

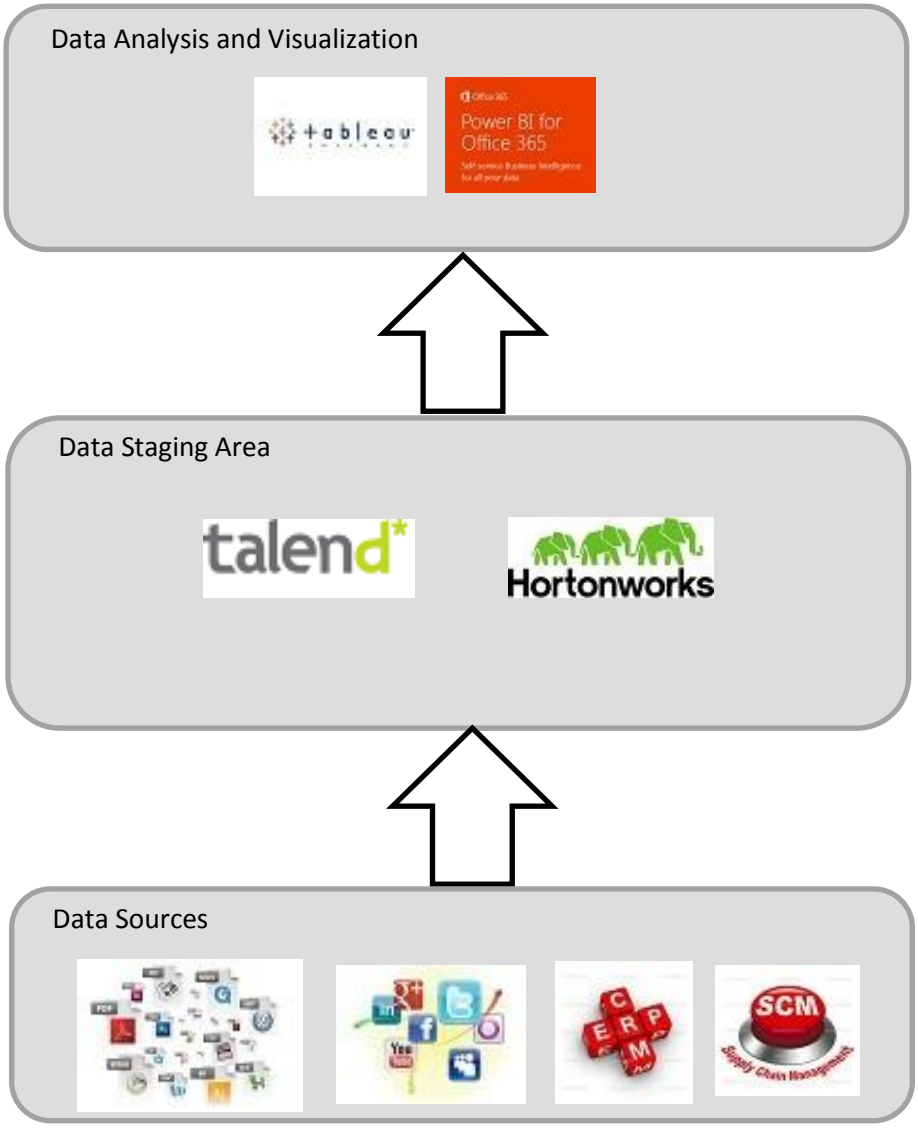


Figura 9 - Arquitetura Física

As tecnologias que suportam a solução desenvolvida, exigiram um elaborado trabalho de pesquisa e análise do mercado, sendo que a opção pelas ferramentas tecnológicas acima associadas teve em conta o quadrante mágico de tecnologias da *Gartner*, mas também a opinião dos responsáveis da empresa Cloud365, com o objetivo de escolher as ferramentas tecnológicas que melhor se adequassem e permitissem a melhor performance do sistema.

3.2.1 *Data Staging Area*

Para o nível de *Data Stanging Area*, a ferramenta escolhida foi do Talend Open Studio com os módulos de *Big Data*, *Data Quality* e *Data Integration*, devido à sua abrangente oferta de soluções, que possibilita a conexão a quase todos os tipos de tecnologias de base de dados. A boa usabilidade da ferramenta e os inúmeros componentes que possui permite a realização do processo ETL com grande rigor e exatidão, garantindo a consistência dos dados.

Neste nível também se optou pela utilização do Hortonworks a figura 10, apresenta toda a arquitetura da ferramenta Hortonworks. Os dados tratados através do processo ETL, realizado através do Talend Open Estúdio, foram armazenados através do sistema de ficheiros Hadoop *Distributed File System* (HDFS), sendo esta uma forma simples de armazenar dados para posteriores análises.

O HDFS é um sistema de ficheiros que permite um processamento simultâneo fornecendo uma gestão de recursos, com uma ampla variedade de métodos de acesso a dados, com um armazenamento eficiente em termos de custo, escalabilidade e tolerância a falhas.

O Hive que um componente de armazenamento de dados presente no Hortonworks foi escolhido como repositório dos dados tratados, devido à capacidade de integrar repositórios de dados de diferentes tipos e origens.

3.2.2 *Data Analysis and Visualization*

Este é o último nível da arquitetura responsável por realizar as análises aos dados e disponibilizar os resultados na *cloud*. A utilização da plataforma Azure da Microsoft permite que os dados sejam guardados numa ambiente de Cloud, podendo os mesmos ser utilizados para análises e disponibilização dos resultados em cloud. A utilização da plataforma Azure permite a instalação de software quer de tratamento quer de analise de dados num sistema operativo em que se encontra em Cloud e que pode ser acedido e utilizado em qualquer local.

O acesso aos dados é efetuado através de uma ligação *Open Data Base Connectivity* (ODBC) ao Hortonworks, importando os dados para o Microsoft Office Excel para depois serem analisados recorrendo a ferramenta Microsoft Power BI. Existe um conjunto de análises que podem ser realizadas como:

- ***Data Mining and Predictive Analysis*:** são análises que utilizam algoritmos de *Data Mining* para a previsão de determinados acontecimentos. Estas análises serão efetuadas através dos algoritmos presentes no *Addin* de *DataMining* (que são programas de computadores usado para adicionar funções a outros programas) que é utilizado no Excel 2013

- ***Query Adhoc*:** são análises realizadas através de *querys adhoc*, ou seja, não estruturadas na base de dados. Para realizar este tipo de *querys* vamos utilizar o Microsoft Power BI que permite através da importação dos dados para *Excel* ou também através de uma ligação ODBC, visualizar os resultados de forma interativa através das ferramentas de Power BI, compostas pelas Power Map, Power Pivot e Power Query. O Power BI oferece a possibilidade de visualização dos dados na web através do Office365 e utilizando esta solução existe uma componente de *Quality Assurance* (QA), que permite que o utilizador possa fazer perguntas em linguagem corrente (perguntas em inglês), sendo as respostas a estas perguntas obtidas através de gráficos que o Power BI fornece de acordo com a pergunta.

- ***Reports* e *dashboards*:** para a realização de cubos OLAP e apresentação dos resultados produzidos através dos mesmos, optou-se pela utilização da ferramenta Power View presente no Microsoft Power BI que possibilita uma apresentação agradável dos resultados através de gráficos interativos o que proporciona uma fácil navegação nos dados associados a uma determinada análise. Foi também explorada a solução de *dashboards* Tableau, uma ferramenta tecnológica que tem estado muito em voga, sendo alvo de muitas referências por parte de grandes consultoras como a *Gartner* e IDC como uma das tecnologias que terão o maior potencial de adoção nas organizações nos próximos anos no que se refere à visualização de dados.

<table>
<tr><td>Operações</td><td>Fornecimento de Recursos, Gestão e monitorização:
Ambari e Zookeeper</td><td colspan="2">Agendamento de Tarefas:
Oozie</td></tr>
<tr><td>Segurança</td><td colspan="3">Autenticação, Autorização, Utilização e Proteção dos Dados:
Armazenamento: HDFS,
Recursos: YARN,
Acesso: HIve
Pipeline: Falcon
Cluester: Knox</td></tr>
<tr><td rowspan="2">Acesso aos Dados</td><td>Código: Pig
SQL: Hive, Tez, HCatalog,
NoSQL: Hbase, Accumulo,
Pesquisa: Sorl,
Outros Spark</td><td rowspan="2">HDFS</td><td rowspan="2">Gestão de Dados</td></tr>
<tr><td>YARN: Sistema de operações sobre dados</td></tr>
<tr><td>Governança da Integração</td><td colspan="3">Fluxos de Dados e Governança
Falcon, webHDFS, NFS, Flume, Sqoop</td></tr>
</table>

Figura 10 - Arquitetura Hortonworks Sandbox versão 2.1

4 – Casos de Teste

4.1 Descrição do Funcionamento Operacional da Arquitetura

A arquitetura foi implementada recorrendo plataforma *online* Azure da Microsoft que permite a disponibilização de recursos como um sistema operativo, memória, disco e processador que podem ser acedidos a partir de qualquer local desde que possua uma ligação à internet. Recorrendo a esta solução foi instalado o Windows Server 2008 R2, dentro do qual foi depois instalada uma máquina virtual (Hortonworks Sandbox 2.1), e ainda as ferramentas de ETL e integração de dados Talend OpenStudio, *Big Data*, Data Quality e Data Integration, bem como as ferramentas de análise de dados Microsoft Power BI e Tableau. O objetivo da instalação de toda esta tecnologia foi a validação da capacidade da arquitetura, para posteriormente poder adotar a mesma num cenário real.

No entanto é importante destacar a utilização deste ambiente como uma mais-valia sobretudo porque permite que o trabalho possa ser executado em qualquer local independente da hora e sem limitação de tempo e espaço. Além disto não existe uma fronteira física com o limite de capacidade de processamento de uma máquina pois com este tipo de ambientes é possível adicionar recursos que deem resposta as necessidades. Um dos benefícios do uso do Azure tem a ver com o facto de não existir um grande investimento inicial na aquisição de máquinas e as posteriores necessidades de atualização de *hardware* para suprir as necessidades exigidas para a utilização de uma determinada ferramenta tecnológica.

A utilização de recursos em *cloud* exige também custos e embora estes sejam menores que os custos associados a aquisição de equipamentos de *hardware*, a disponibilização de recursos na plataforma Azure é instantânea e caso seja necessário uma atualização ou aumento de capacidades de armazenamento, memória *Random Acess Memory* (RAM) e processador pode ser feita e disponibilizada de forma rápida e bastante fácil.

Uma outra questão associada aos sistemas que se encontram em *cloud* é a segurança, verificando-se, de facto, que muitas empresas têm medo de optar por recursos em *cloud* precisamente por causa da alegada falta de segurança, a verdade é que todos os sistemas possuem falhas. Atualmente existe um conjunto de normas que são seguidas pelos fornecedores de serviços em *cloud* e que permitem aos utilizadores poderem confiar no sistema que estão a utilizar. No caso do Azure existe um mecanismo de autenticação dos utilizadores com nome de utilizador e palavra passe, para além de um conjunto de normas como o acesso ao servidor que disponibiliza o serviço em *cloud* e que é feito com recurso a certificados de segurança como o SSL e o HTTPS, que garantam a segurança e da privacidade dos utilizadores.

Este trabalho foi realizado na plataforma *cloud* Azure da Microsoft, a opção pela utilização desta plataforma permitiu que todo o trabalho fosse executado em qualquer local e com um acesso constante, sendo esta uma das grandes vantagens da utilização de recursos em *cloud*. A escolha do ambiente *cloud* teve como critério a capacidade de implementar níveis de segurança de acesso aos dados e desta forma foi definido um conjunto de utilizadores e permissões de acesso ao sistema, ou seja, cada utilizador tem uma credencial (nome utilizador e palavra chave) para aceder ao sistema, bem como um conjunto de permissões para efetuar tarefas como: aceder, visualizar ou analisar os dados.

Todo o trabalho realizado também poderia ser efetuado em grande parte, através de uma máquina local. Deste modo todo o trabalho ETL e armazenamento de dados poderia estar numa máquina local (servidor ou computador pessoal), sendo a disponibilização da informação feita através da *cloud* recorrendo as ferramentas do Azure apenas para a disponibilização dos resultados através da *Web*. Sendo que esta visualização de dados estaria dependente das permissões de acesso que seriam definidas para cada um dos utilizadores que tivessem acesso, apenas seriam disponibilizados os resultados de acordo com as permissões de cada utilizador, garantindo que não existe acesso indevido de um utilizador a dados aos quais não está autorizado aceder.

A base de dados utilizada (Hive) permite encriptação dos dados, mas não permite que os dados encriptados possam ser visualizados pelo *Database Administrator* (DBA), esta opção garante uma maior confiança na utilização do ambiente *Cloud Computing*.

4.2 Implementação do Caso de Teste Com Utilização de Dados OpenData

Para a realização do caso de teste foi selecionado um *dataset* da área da saúde obtido através da plataforma *OpenData* (https://health.data.ny.gov/). Esses dados correspondem a quatro tipos de cirurgias ocorridas durante os anos de 2010 a 2013 e com origem em hospitais de Nova Iorque, Estados Unidos da América. Este caso de teste deu origem a um artigo intitulado "Uma Arquitetura Moderna de Dados: Um Caso de Teste", apresentado na 14ª Conferência da Associação Portuguesa de Sistemas de Informação realizada nos dias 3 e 4 de Outubro em Santarém, o artigo pode ser consultado no anexo C – Paper CAPSI 2014.

O primeiro passo realizado foi a recolha dos dados a utilizar, que estavam no formato de ficheiros CSV. Após os dados serem importados para o ambiente do Windows Server, os mesmos foram tratados através do processo de ETL e para a realização deste processo ETL foi utilizado o Talend, no qual em primeiro lugar com o Talend Data Quality foi definido o modelo dos dados que estavam nos ficheiros CSV. Em seguida com o Talend Big Data, foi definido um processo que importava e tratada os dados para que estes pudessem ser introduzido no Hortonworks. O processo ETL realizado no Talend Big Data foi o passo seguinte, no qual foram escolhidos os ficheiros CSV a importar, sendo depois adicionada o modelo de dados de cada um dos ficheiros CSV. Seguidamente foram efetuadas a correções dos valores nulos e campos vazios que se encontravam nos ficheiros, que depois foram importados para o Hortonworks. Os atributos escolhidos tiveram em conta as possíveis análises que poderiam ser realizadas, sendo apenas importados os dados que interessavam, enquanto os restantes seriam descartados.

O segundo passo seria a colocação dos dados importados para o sistema de ficheiros HDFS dos Hortonworks, no Hive para que os mesmos pudessem depois ser refinados caso fosse necessário alguma refinação dos mesmos antes de serem acedidos pelas ferramentas de análise de dados. Uma vez importados os dados para o HDFS é necessário aceder aos dados através do HCatalog e colocá-los disponíveis para poderem ser utilizados pelas diversas ferramentas que compõem a solução Hadoop da Hortonworks. Desta forma com o HCatalog é criada uma tabela com os dados ficam disponíveis para ser acedida e manipulada pelo Pig e pelo Hive. O Pig é um editor de *scripting* que permite que caso seja necessário, possam ser programadas algumas ações a realizar aos dados. Em relação ao Hive trata-se de um DW que permite a integração de dados diversificados (dados estruturados, semiestruturados e não estruturados), num único local, fornecendo também a possibilidade de utilização da linguagem SQL para a manipulação dos dados. Para além disso o Hive permite a realização de operações SQL sobre os dados transformando as instruções SQL em instruções próprias que possibilitam a manipulação dos dados que estão armazenados no Hive.

Em seguida foi efetuada uma análise dos dados que se encontram armazenados no Hortonworks, processo este que exige uma ligação ODBC, para que se conseguir aceder aos dados armazenados e se poder trabalhar com os mesmos. No entanto importa mencionar que a utilização da ferramenta tecnológica Microsoft Power BI no Excel 2013 permite que o acesso aos dados possa ser efetuado através da ligação ODBC ou então possa utilizar o módulo Power Query e aceder aos dados em ficheiro HDFS. Para tal é necessário indicar o endereço IP (*Internet Protocol*) da máquina onde os ficheiros HDFS se encontram armazenados para ser efetuada uma conexão ao sistema HDFS e depois se proceder à importação dos dados para o Microsoft Excel, onde posteriormente serão analisados através das ferramentas Power View e Power Map, que permitem a criação de *dashboards* e *reports* bastante dinâmicos e com bastante usabilidade e compreensão por parte do utilizador.

De forma a obter o maior partido da informação que se encontrava nos comentários médicos foi necessário extrair essa informação de um ficheiro em formato XML, e após este processo os dados não estruturados foram tratados e importados para o Hortonworks, onde através de instruções SQL os dados dos comentários foram adicionados aos restantes dados.

Seguidamente, após adicionar a informação dos comentários aos restantes dados, foi definida uma estratégia de análise de dados. De forma a não serem efetuadas análises *ad hoc*, foram definidas 3 questões que seriam alvo de resposta, através das análises aos dados: "**1- Quais os hospitais com o maior número de intervenções cirúrgicas em Nova Iorque? 2 -Qual a evolução no número de mortes nas principais cirurgias realizadas nos hospitais de Nova Iorque? 3 - Qual a influência dos comentários para a melhoria dos serviços hospitalares?**".

Depois de carregados os dados anteriormente tratados seguimos para a fase de análise e visualização de resultados. Para executar a tarefa de análise dos dados, em primeiro lugar foi necessário estabelecer uma ligação ODBC e esta conexão permitiu o acesso aos dados através das ferramentas Microsoft Excel 2013. Com a solução Microsoft Power BI, uma vez estabelecida a conexão, os dados foram importados e analisados, sendo construídos *dashboards* e relatórios com o resultado das análises efetuadas, podendo os mesmos ser visualizados nas figuras 11, 12 e 13

De forma a provar a viabilidade da arquitetura anteriormente descrita, são apresentados alguns exemplos de análises que foram efetuadas aos dados de teste, com a intenção de comprovar as potencialidades da arquitetura anteriormente referida.

A primeira análise realizada, que se encontra representada na figura 11, mostra localização dos hospitais onde existiu o maior número de intervenções cirúrgicas, o que permite obter uma perceção das zonas de Nova Iorque, mais procuradas pelos pacientes.

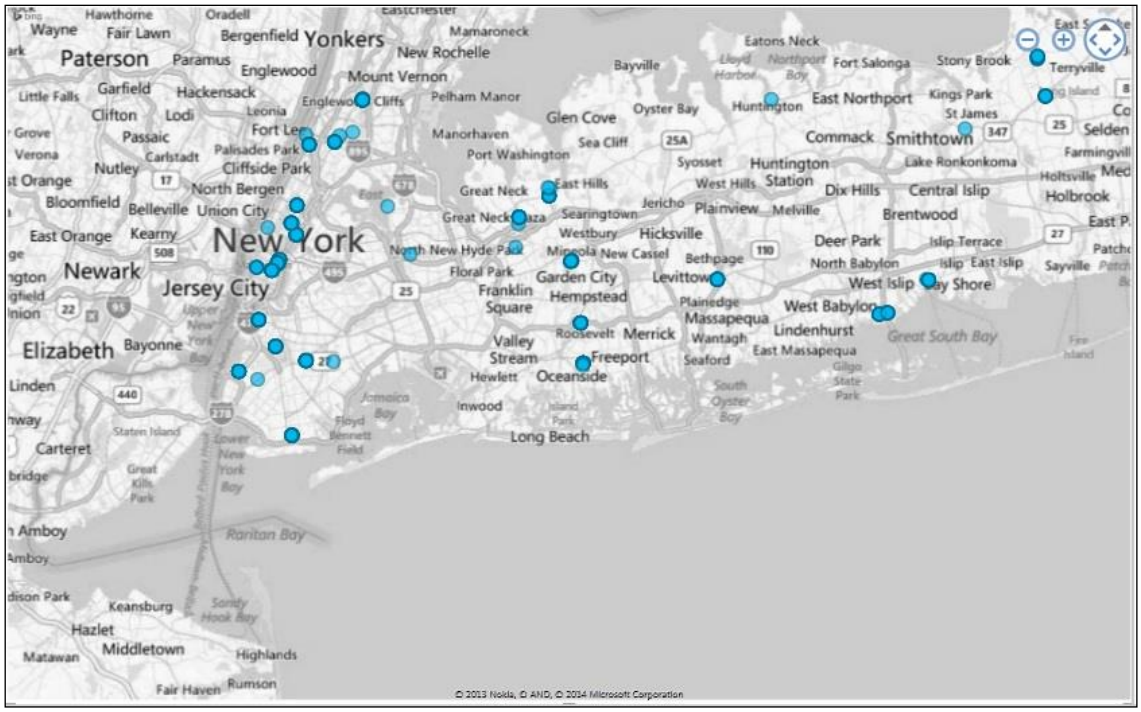


Figura 11 - Localização dos hospitais utilizados nas análises

Após aferir quais as zonas da cidade de Nova Iorque mais procuradas pelo pacientes. Decidiu-se analisar os vários tipos de cirurgias que eram realizadas aos pacientes. Esta análise permitiu-nos perceber que o número de mortes tem diminuído, sendo que esta informação pode ser visualizada na figura 12. Podemos observar que a diminuição do número de mortes pode dever-se à influência dos comentários realizados pelos pacientes que recorrem aos serviços dos hospitais da zona de Nova Iorque. Através desses comentários, os profissionais e o corpo clínico do hospital efetuaram um esforço com vista à melhoria dos serviços, foram tomadas medidas com o objetivo de melhorar os procedimentos clínicos, o que ajudou a incrementar a eficácia e eficiência na realização de quatro tipos de cirurgias: *ALLPCI* (cirurgia que consiste na retirada da tiroide), *CABG* (cirurgia que consiste no desvio da artéria coronária), *NON EMERGENCY PCI* (cirurgia que retira a tiroide não urgente) e *Valve or Valve*/CABG (cirurgia da válvula aórtica/cirurgia ponte da artéria coronária).

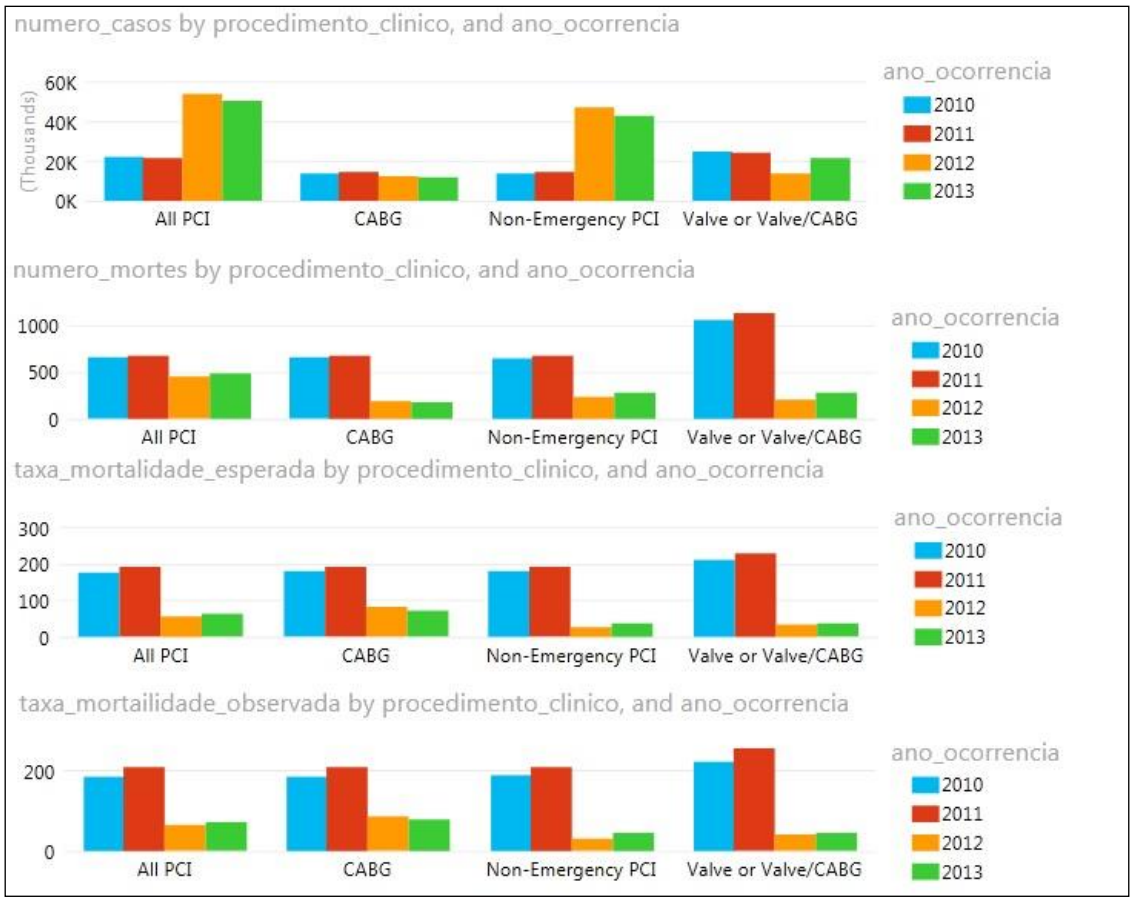


Figura 12 - Análise de Dados Por Ano e Por Procedimento Clínico

Também foi possível verificar a influência dos comentários dos pacientes na melhoria dos serviços, como podemos observar nos diferentes gráficos representados na figura 13. Os comentários permitiram que os hospitais diminuíssem o número de mortes e aumentassem a eficiência dos seus serviços melhorando a eficácia na resolução dos casos clínicos.

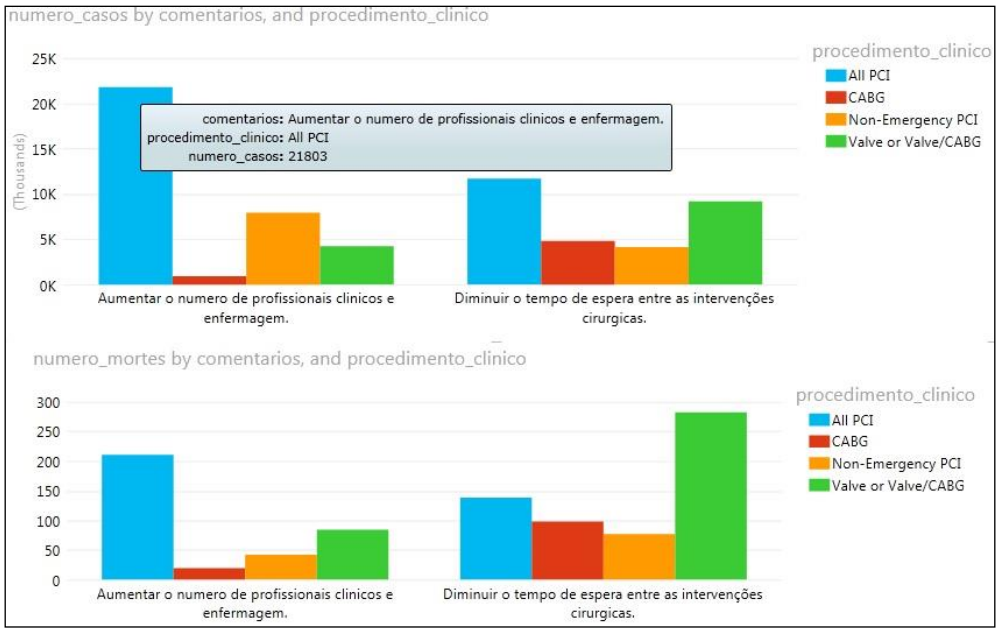


Figura 13 - Influência dos Comentários dos Pacientes na Melhoria dos Serviços Médicos

Os resultados obtidos através das análises realizadas convergiram numa conjugação de informação com origem em dados estruturados e dados não estruturados (comentários), de facto não é uma tarefa fácil relacionar estes tipos de dados. Este caso de teste serve também para mostrar que ainda existe muito trabalho a realizar nesta área, pois para uma correta utilização de dados não estruturados é muito importante perceber em primeiro lugar o que esperamos obter com análise dos dados e em segundo perceber o valor que esses dados não estruturados podem ter e em que medida podem ajudar no processo de decisão de uma organização. Um aspeto que se constatou foi que a informação que se encontrava nos dados não estruturados era um complemento a informação dos dados estruturados, o que se constitui como um fator de relevo no sentido em que dá confiança na hora de tomar decisões pois as bases que sustentam a mesmas são sólidas.

Por fim é importante referir que a adequação da arquitetura física aos resultados está dependente de muitos fatores, sendo de destacar dois deles:

1. os resultados dependem muito da riqueza dos dados, ou seja, se os dados não forem os mais adequados ou não tiverem grande riqueza os resultados não serão os melhores;

2. a importância do tratamento dos dados é fulcral, pois a sua adequada realização, ajuda a garantir a integridade dos dados, contribuindo para a obtenção dos melhores resultados quando efetuado o processo de análises de dados.

4.3 Caso de Teste Com Utilização de Dados de Logs Relativos às Compras Online

Este caso de teste foi realizado, partindo da seleção de um conjunto de dados do comércio eletrónico. Este conjunto de dados foi obtido a partir da empresa Americana *Omniture*, que foi adquirida pelo Adobe em 2009, e que fornece um conjunto de dados para analise em vários projetos.

O primeiro passo foi recolher e validar os dados, eliminando os valores nulos e duplicados. Este trabalho foi efetuado utilizando o Talend Open Studio com o módulo *Big Data*. Em seguida foram selecionados os dados que permitiram criar os indicadores de negócio, esses dados foram depois enviados para o repositório Hive presente no Hortonworks.. De forma a evitar a utilização de análises *ad hoc*, foram definidas duas questões a serem exploras: "**1 -Qual o comportamento dos utilizadores *online* em relação às compras efetuadas nos Estados Unidos da América? 2- Que faixa etária tem utilizado mais as compras *online* e quais os produtos comprados pelas pessoas de uma determinada faixa etária?**".

Com o objetivo de aproveitar ao máximo as informações dos clientes quando estes visitam um site de vendas, foi necessário efetuar uma verificação cruzada dos registos dos clientes e dos registos de compras, este trabalho de verificação cruzada foi realizado através do Talend Open Studio e da operação *join* sendo ambas funções que permitem agregação de todas as informações num único repositório de dados.

A análise e visualização dos resultados foram realizadas, através de uma conexão ODBC para possibilitar o acesso aos dados através de ferramentas Microsoft Power BI e Tableau, sendo as análises aos dados implementadas com recurso a *dashboards* e relatórios.

Para responder à primeira questão: "Qual o comportamento dos utilizadores *online* em relação às compras efetuadas nos Estados Unidos da América?". Foi criado um *dashboard*, o qual mostra quais os estados dos EUA em que os utilizadores optaram em grande número pela realização de compras *online*, sendo também possível verificar qual o artigo que é mais comprado online na maioria dos estados, e verificando, neste aspeto que a roupa é o artigo mais representativo no que diz respeito ao volume de compras. Em relação à segunda questão: "Que faixa etária tem utilizado mais as compras *online* e quais os produtos comprados pelas pessoas de uma determinada faixa etária?", Podemos verificar pelo gráfico que a faixa etária que mais compra efetuou *online* é a dos 46 aos 66 anos, sendo que os artigos mais comprados pelos utilizadores são: roupa, computadores, artigos de eletrónica, malas de mão, artigos de casa e jardim, filmes e calçado. Toda esta informação pode ser verificada na figura 14.

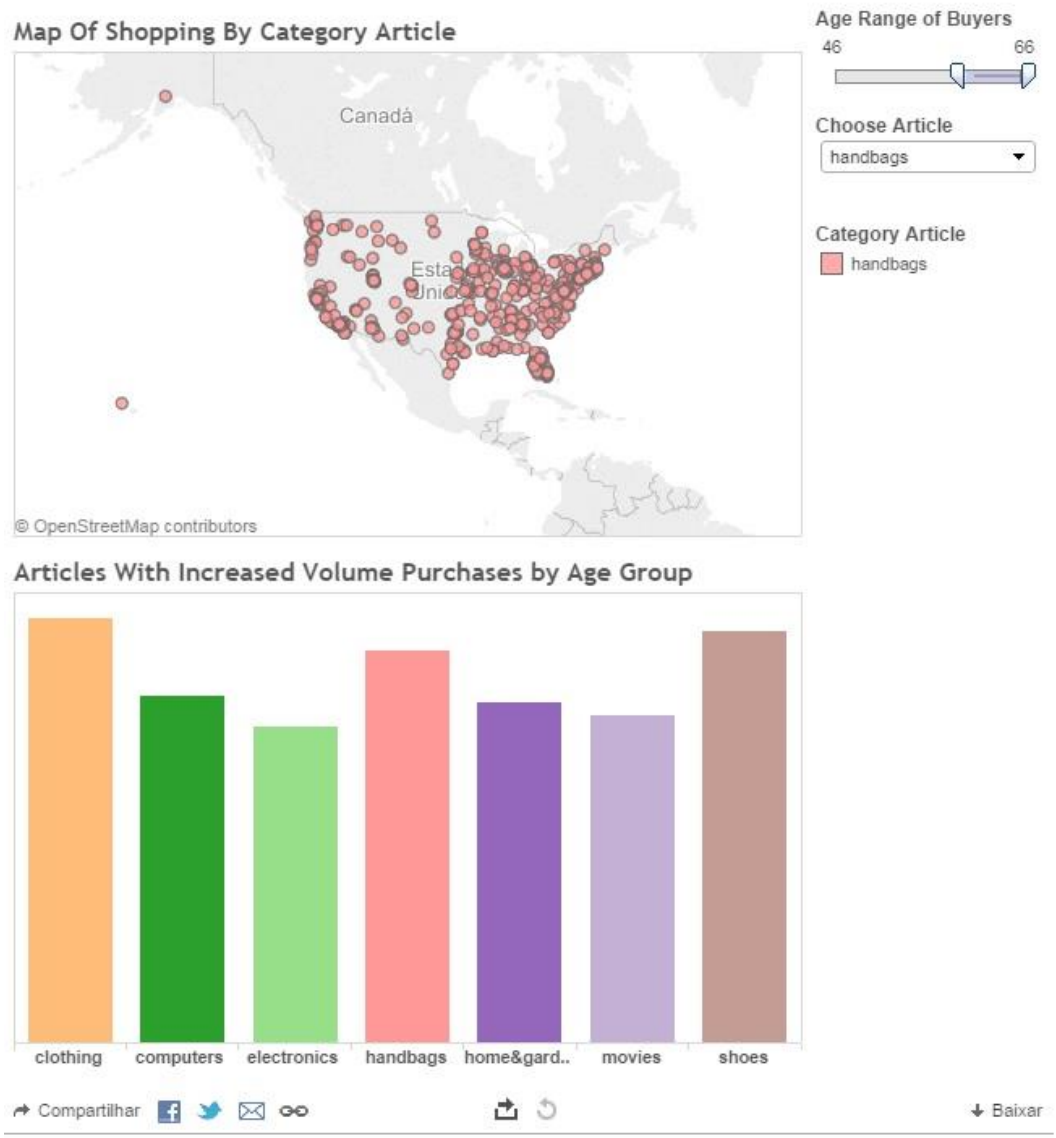


Figura 14 - Análise de Compras Online EUA, dados da Omniture

4.4 Caso de Teste Com Utilização de Dados das chamadas realizadas para Portugal

O seguinte caso de teste foi realizado com os registos que são produzidos pelos *Call Detail Record* (CDR) da Colt, que possuem informações sobre os detalhes das chamadas que foram efetuadas.

Após obter os dados da Colt, procedeu-se ao seu tratamento, pois os dados estavam num ficheiro de extensão CAT, que são ficheiros de sistema associados ao Todd Osorne Directory Catalog. De forma a conseguir abrir este ficheiro o mesmo foi importado para o Hortonworks, onde através da ferramenta File Browser foi convertido num ficheiro CSV. Em seguida com o Talend Open Studio módulo *Big Data* os dados do ficheiro CSV, forma validados, eliminado os valores nulos e duplicados. Depois foram definidos os indicadores de negócio, e por fim os dados foram importados para o Hortonworks e armazenados no repositório de dados Hive.

Como forma de obter informação útil dos CDR's, foi definida a estratégia de analisar a informação das chamadas tendo como base o custo e duração das mesmas. Assim foram definidas duas questões que seriam alvo de análise: "**1- Qual o tempo de duração de uma chamada para um determinado destino? 2 -Quantas vezes foram efetuadas chamadas para um determinado destino?**".

A análise e visualização dos resultados foram realizadas, através de uma conexão ODBC para permitir o acesso aos dados através de ferramentas Microsoft Power BI e Tableau, sendo as análises aos dados implementadas através de *dashboards* e *reports*.

O mapa da figura 15 mostra o tempo total de duração de uma chamada efetuada dos Estados Unidos da América com destino a Vila Franca de Xira em Portugal, podendo-se identificar os tipos de chamada através das diferentes cores atribuídas a cada deles.

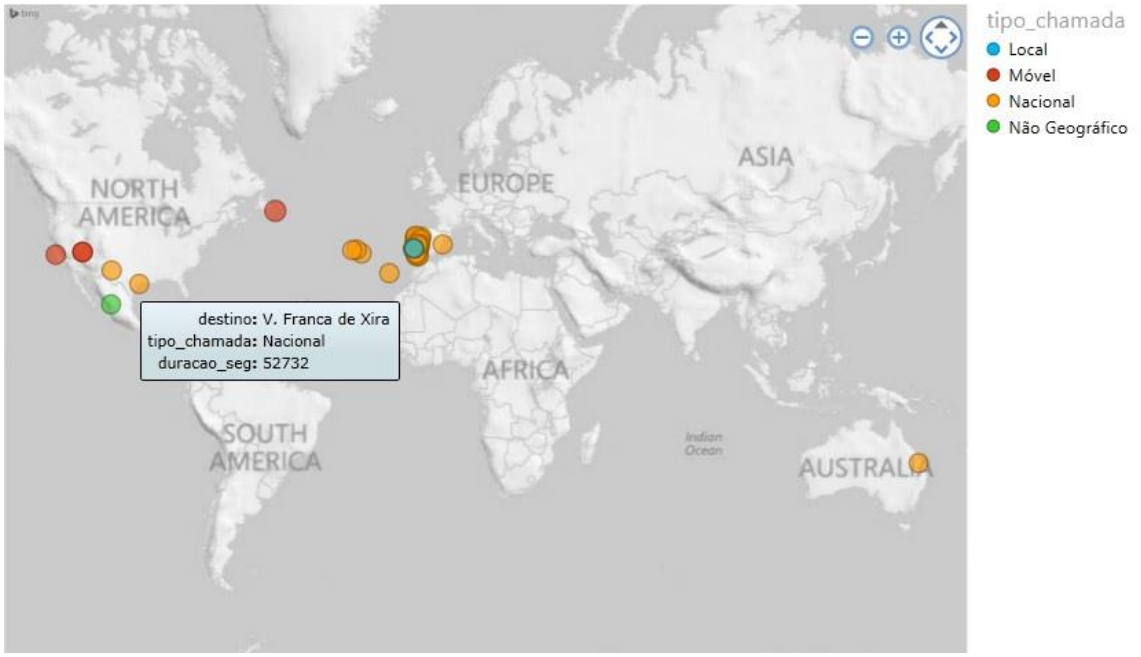


Figura 15 - Mapa Com a Identificação da Duração das Chamadas

Em relação à contagem da quantidade de chamadas que foram efetuadas para um determinado destino também podemos verificar que forma efetuadas 10 chamadas com origem na Austrália e que tiveram como Destino a cidade de Moura em Portugal, como pode ser observado pelo mapa da figura 16.

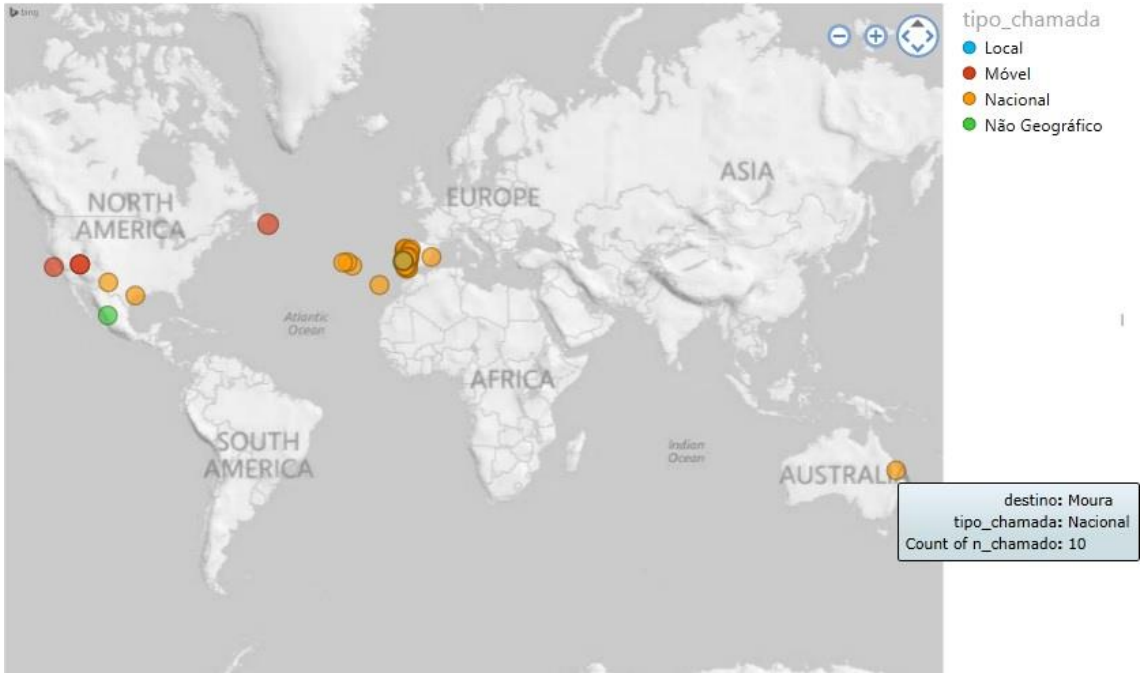


Figura 16 - Identificação do Número de Chamadas Efetuadas

O *dashboard* representado na figura 17 mostra o total de segundos que foram utilizados nas chamadas realizadas para um determinado destino, também é possível verificar o número total de chamadas que os

diferentes operadores móveis que operam em Portugal receberam num determinado dia e de acordo com os destinos selecionados.

As conclusões obtidas com as análises aos dados dos CDR's, podem ser relevantes, na medida em que podem ajudar os operadores de telecomunicações a perceber o comportamento dos utilizadores, sendo também possível identificar os destinos que recebam mais chamadas e qual o total de segundos que são utilizados nas chamadas para esses destinos.

A compreensão desta informação pode ajudar os operadores de telecomunicações a definir os tarifários tendo com base informação real sobre o comportamento dos utilizadores, ajudam na identificação dos locais onde devido ao grande fluxo de chamadas podem ocorrer maiores congestionamentos de tráfego, nas linhas de comunicação.

A informação que se encontra nos CDR's pode ser analisada de uma forma contínua e aproximada ao tempo real o que faz com que as decisões a tomar tenham como base informação sólida e atual.

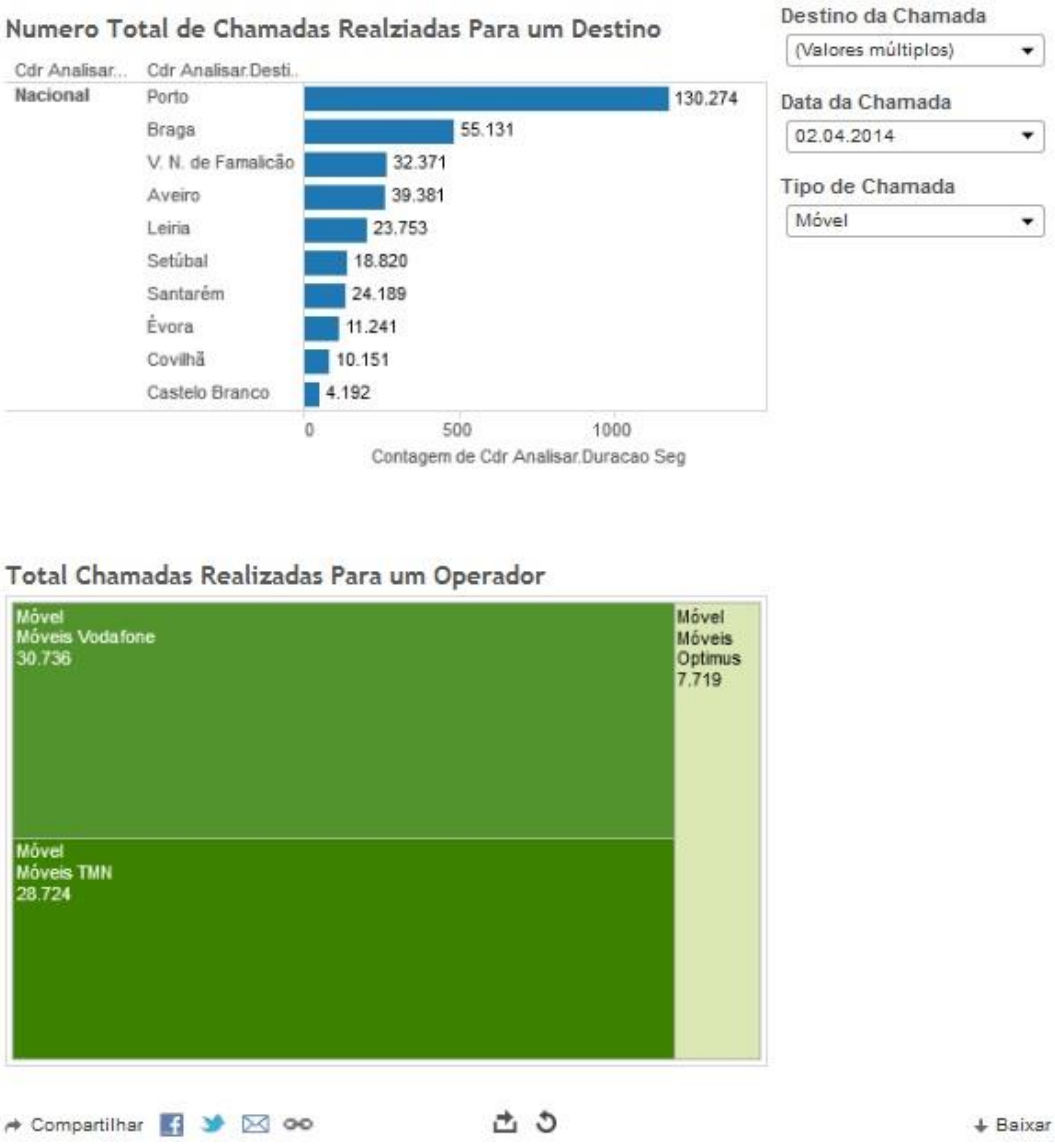


Figura 17 - Número Total de Chamadas Por Operadora Móvel

5 – Conclusões

Este capítulo procura resumir as conclusões mais relevantes do trabalho realizado, pois é importante justificar a importância do mesmo (secção 5.1). Além disso, é essencial refletir acerca das principais dificuldades que foram encontradas ao longo do desenvolvimento deste projeto (secção 5.2). Por fim são definidas as propostas a ter em conta na realização de um trabalho futuro (secção 5.3).

5.1 Síntese

O *Big Data* é uma das grandes inovações da atualidade que está a revolucionar a área da TI. Atualmente o número de organizações que fornecem serviços na área das TI (ex. empresas de desenvolvimento de software e empresas de serviços de TI), oferecem um vasto conjunto de soluções que cada vez mais têm o foco na *cloud*, nomeadamente empresas como *Google*, *Amazon*, *IBM* e *Microsoft*. Estas organizações também oferecem um conjunto de soluções de *Big Data* que procuram explorar os novos tipos de dados como uma forma de ajudar na inovação do negócio. No entanto existem alguns entraves à entrada destas tecnologias nas organizações, em primeiro lugar porque as organizações ainda não assimilaram de forma clara os conceitos e benefícios que as mesmas podem trazer para o seu negócio.

Foi na perspetiva de compreender de que forma a *cloud* pode suportar um sistema de análises de dados *Big Data* que se justificou a realização deste trabalho de investigação, podendo o mesmo servir como base de um contributo para as atividades de definição estratégica das organizações que pretendem desenvolver soluções de análises *Big Data* e alicerçar o processo no recurso à *cloud*, permitindo a disponibilização de um serviço de TI inovador, sólido e eficaz. Este trabalho de investigação visa também a identificação de um conjunto de abordagens que podem ser seguidas quer por quem necessitar de desenvolver soluções baseadas em *Big Data* e alinhadas com a necessidade de utilização da *cloud*, promovendo desta forma, a melhoria da estratégia de negócio das organizações, e, por conseguinte dos seus resultados e desempenho global.

Partindo do objetivo e da finalidade estabelecidos para este projeto, foi delineada uma estratégia de divisão do trabalho por etapas. Neste caso foram definidas 3 etapas de acordo com os objetivos definidos para este projeto de dissertação. A primeira etapa esteve associada à revisão e análise de vários documentos bibliográficos relacionados com *Big Data*, *Big Data Architecture Solutions and Tecnologies* e *Cloud Computing*, perspetivando a construção de uma base teórica de conhecimento que permitisse o suporte a todo o trabalho de investigação. Em seguida foi necessário compreender e identificar as diferentes abordagens relacionadas com o tema, de modo a conseguir obter informação que permitisse o suporte a construção de uma arquitetura conceptual e física que possibilitasse a criação de uma solução de análise de dados *Big Data* no modelo *Cloud Computing*. Ainda nesta fase procedeu-se à exploração das abordagens identificadas e à seleção das que pareceram mais adequadas para o âmbito deste

projeto em particular. Por fim, a etapa final da estruturação do projeto, consistiu na instanciação da arquitetura física com recurso a tecnologias existentes no mercado e que segundo alguns documentos bibliográficos são as mais indicadas para a construção de um sistema de análise de dados *Big Data* no modelo de *Cloud Computing*.

Todo o conhecimento obtido, na identificação de abordagens relacionadas com o *Big Data* e *Cloud Computing*, no contexto do trabalho realizado, ajudou na exemplificação dos casos práticos que permitiram a instanciação da arquitetura física com recurso a ferramentas tecnológicas, assumindo-se como uma importante contribuição para toda a investigação efetuada ao longo desta dissertação.

Este é um trabalho que apresenta muitas questões potenciais, quer a nível de negócio quer a nível de tecnologia nomeadamente no que respeita às tecnologias relacionadas com as ferramentas tecnológicas de *Big Data* em que existe uma convergência geral para a importância do Apache Hadoop. Por outro lado existe ainda muita indefinição relativa à escolha dos dados e às estratégias a utilizar para analisar e manipular os diferentes tipos e fontes de dados. Em relação à *cloud* verifica-se o problema e renitência das organizações em optarem por soluções de *cloud* devido às questões de segurança e privacidade de dados, o que faz com que seja necessário muito trabalho para melhorar este aspeto.

5.2 Discussão

Com o objetivo de alcançar a finalidade definida para este projeto de dissertação, procurou-se cumprir as três etapas anteriormente definidas: (E1) rever e conhecer os fundamentos da literatura, (E2) identificar, analisar, perceber e escolher a abordagem que melhor se adequa a construção de arquitetura de análise de dados *Big Data* no modelo de *cloud*, construção do modelo conceptual e modelo físico, (E3) instanciar a arquitetura definida com tecnologia e efetuar testes que permitam validar a arquitetura.

De forma a poder cumprir a primeira etapa (E1), procedeu-.se a uma revisão e análise de um conjunto de documentos bibliográficos relacionados com a área do *Big Data* e *Cloud Computing*, sendo de mencionar que, devido à abrangência de contribuições, foi necessário efetuar uma delimitação da área de estudo. Assim foram definidas quatro áreas de conhecimento com importância relevante para este trabalho (*Big Data*, arquiteturas de sistemas de análises de dados, arquiteturas de análises de dados no modelo C*loud Computing*, e tecnologias de análises de dados *Big Data*). Depois de realizada a revisão de literatura debruçamo-nos sobre a reflexão acerca dos principais aspetos do estado atual do tema em estudo no caso *Big Data* e *cloud*. Por fim procedeu-se à identificação das principais oportunidades e desafios para o processo de adoção de *Big Data* no modelo de *cloud*, nas organizações, quer a nível dos consumidores deste tipo de serviços quer a nível dos fornecedores de serviços e recursos de TI.

A realização da primeira etapa (E1) é considerada, e revelou-se com efeito, basilar para a constituição de uma base de conhecimento que permitiu suportar e sustentar a concretização das restantes etapas (E2 e E3). Mesmo com as dificuldades encontradas na realização da revisão da literatura, que foi devida principalmente a alguma confusão causada por diferentes pontos de vista por parte dos diferentes autores, pois são referidos e usados muitos conceitos e termos que são utilizados em *Big Data* e *cloud*, a primeira etapa proposta para este trabalho de dissertação foi alcançada.

A segunda etapa (E2), foi realizada de acordo com as sugestões de *Agrawal, Das e El Abbadi* (2011), que no que se refere a tendências do *Big Data* e *cloud*, aferidas com a revisão e análise aos documentos bibliográficos sobre os conceitos anteriormente definidos sugerem um conjunto de referências de como os sistemas devem ser pensados e construídos. A utilização das sugestões de *Agrawal, Das e El Abbadi* (2011), ajudou na definição sustentada de uma estratégia de desenho da arquitetura conceptual, com a definição dos vários níveis e conjuntos de atividades associadas, culminando depois da definição da arquitetura física na qual a cada um dos níveis foram associadas ferramentas tecnológicas que conseguissem responder às atividades definidas no modelo conceptual para cada nível. Seguidamente foi efetuada a instanciação da arquitetura física o que permitiu a criação de alguns cenários de teste que comprovam o funcionamento da arquitetura. O estudo e compreensão dos documentos bibliográficos sobre o estado atual do *Big Data* e adoção da *cloud* nas organizações, possibilitou o desenvolvimento de

uma visão abrangente acerca do estado atual do mercado, no que se refere a soluções que tenham em conta o *Big Data* e que podem ser entregues como um serviço através da *cloud*. Todo este estudo ajudou na compreensão das necessidades das organizações em relação ao *Big Data*, o que se constitui como um aspeto de relevo no sentido de comprovar a crescente necessidade de criar uma solução que fornecesse um serviço de análise de dados *Big Data* no modelo de *cloud*. Por fim a revisão e análise de documento bibliográficos sobre as arquiteturas de sistemas de análises de dados e conjunto com os documentos que identificam as ferramentas tecnológicas que estão associadas ao tema *Big Data*, ajudaram na identificação, análise e escolha da abordagem a seguir para a construção de uma arquitetura de análise de dados *Big Data* de acordo com o sistema de *cloud*.

Uma arquitetura moderna de dados deve possuir um sistema de armazenamento de dados diversificado, que permita armazenar dados estruturados, semiestruturados e não estruturados, devendo também possuir mecanismos que permitam uma distribuição dos dados por diferentes sistemas de armazenamento, nomeadamente bases de dados SQL e *Not only Structured Query Language* (NoSQL). Ao mesmo tempo deve possuir ferramentas que permitam efetuar consultas a dados de forma rápida e eficaz sem a necessidade de esperar muito tempo para obter as respostas às consultas efetuadas. A utilização de um ambiente como o Hortonworks que possuía configuração de todas as ferramentas associadas ao Hadoop foi o fator de relevo para a simplicidade de todo o processo de tratamento e acesso aos dados para posteriores análises.

A disponibilização de dados através da *cloud* obriga a definir um conjunto de restrições e de políticas de permissões de acesso aos dados por parte dos utilizadores, para que os dados sejam acedidos por quem deles necessita sem que os mesmos percam a sua integridade. Tudo isto pode ser implementado através de mecanismo de gestão de utilizadores que estão implementados nas plataformas de infraestruturas *online*, como é o caso do Microsoft Azure, a qual permite utilizar protocolos de segurança como *Secure Sockets Layer* (SSL) e *Hyper Text Transfer Protocol Secure* (HTTPS). A correta implementação de políticas de gestão de acesso leva a que os utilizadores confiem na utilização de soluções em *cloud*, sobretudo em soluções de análises de dados críticos. Uma das características importantes e que a solução adotada não permite totalmente prende-se com a possibilidade de as bases de dados terem níveis de segurança ao nível da encriptação dos dados, fazendo com que o acesso aos dados seja realizado apenas por quem tenha autorização (chave de desencriptação), e evitando assim que mesmo o *Database Administrator* (DBA) ou quem desenvolve a solução possa aceder aos dados.

A utilização de dados *Big Data* num ambiente *Cloud Computing* obriga à criação de uma arquitetura moderna de dados com as seguintes características: ter robustez, ser rápida, ter disponibilidade de dados para serem analisados e permitir a recolha e tratamento de dados de diversos tipos e origens como

dados estruturados, semiestruturados e não estruturados, procurando sempre salvaguardar os aspetos de segurança e integridade dos dados a analisar.

Considera-se que será necessário efetuar mais casos de teste perspetivando a validação da capacidade da arquitetura. Esses testes devem incluir grandes volumes de dados, nomeadamente um grande volume de dados não estruturados e semiestruturados. Outro aspeto a considerar é relativo à melhoria das análises dos dados não estruturados, por exemplo através de técnicas de *Text Mining.*

5.3 Limitações e Trabalho Futuro a Realizar

Este trabalho de dissertação, procurou que todo o trabalho fosse realizado de uma forma rápida e com recursos a um ambiente que não exige-se a necessidade de máquinas físicas por parte do utilizador, mas utilizar os recursos e os serviço *online* como o oferecido pela plataforma Azure da Microsoft como sendo o suporte para todo o trabalho a realizar.

Em relação aos objetivos que foram definidos para esta dissertação, todos foram cumpridos, no entanto a fase de validação da arquitetura carece de muito mais trabalho, visto que esta foi a fase onde os resultados esperados foram bons, infelizmente não se obteve o melhor partido das análise aos dados não estruturados, devido a dificuldade de encontrar dados não estruturados que permitissem a realização análises complexas. Algumas das análises realizadas neste trabalho podem ser consideradas simples, enquanto outras um pouco mais elaboradas e procuram demonstrar a importância que os dados não estruturados e semiestruturados podem ter para as organizações no seu dia-a-dia. O grande obstáculo encontrado foi a dificuldade em obter dados em empresas clientes da Cloud365, pois as empresas contactadas não disponibilizaram dados para análises, os dados que se pretendia analisar seriam os dados relativos a logs de website com informações como as páginas mais visitadas pelos utilizadores, os artigos mais visitados e mais comprados. No fundo os dados não estruturados permitem estudar o comportamento do utilizador *online*, podendo a empresa aproveitar esta informação para criar um sistema de recomendações de produtos ou serviços para os seus clientes.

Além de todas as questões que foram abordadas ao longo da realização desta dissertação, principalmente as relacionadas com as tecnologias de *Big Data*, nomeadamente referentes à escolha do sistema de *cloud* a adotar (Azure ou Amazon AWS), também foi debatida a questão da opção por uma solução totalmente em *cloud* ou por uma solução parcialmente em *cloud* na qual os repositórios de dados estariam em servidor local e a disponibilização das análises aos dados seria efetuada na *cloud*. A escolha de melhor opção tecnológica carece ainda de um estudo mais profundo de forma a perceber se as soluções baseadas na *cloud* são, efetivamente, a melhor opção. Tudo isto surge e assume extrema pertinência pois existem naturalmente muitas questões relacionadas com os modelos de negócio dos fornecedores de *software*, no que se refere à própria cultura organizacional, às políticas dos Sistema de Informação nas organizações, e ainda a aspetos relacionados com a legalização e regulamentação deste tipo de serviços, assim como com a segurança e privacidade dos dados que são neste momento, o grande entrave colocado pela grande maioria das organizações em optar por soluções em *cloud*. É neste contexto que emerge a necessidade de um estudo que procure avaliar de uma forma criteriosa os riscos da opção por um sistema de *cloud*, definindo simultaneamente os modelos e critérios de segurança e privacidade que devem ser seguidos, e procurando demonstrar que os critérios de segurança funcionam

e que a *cloud* é fiável, pois o fantasma da falta de segurança afasta muitas organizações da adoção de soluções em *cloud*.

Desta forma sugiro três propostas de trabalho futuro:

- Uma proposta é a utilização da arquitetura criada implementando-a num ambiente total de *cloud* e definindo os requisitos a seguir, principalmente a nível de segurança, no caso dos dados de diferentes tipos é necessário efetuar muitos mais testes com dados não estruturados e semiestruturados.
- Utilizar mais técnicas como *Data Mining* e *Text Mining* procurando provar as mais-valias e vantagens de inovação que as organizações podem ter em utilizar todos os tipos e fontes de dados.
- Por fim um trabalho interessante a realizar é perceber junto das organizações se as mesmas entendem o conceito de *Big Data* e percebem de que forma o *Big Data* pode ser uma mais-valia para o seu negócio, pois o que se verifica é que as empresas reconhecem o *Big Data* e acreditam que o mesmo pode ajudar no seu negócio, mas não percebem de que forma o *Big Data* pode ser utilizado para criar inovação no seu negócio por fim, sugiro que seja realizado um estudo sobre as tecnologias que podem ser combinadas com as tecnologias usadas atualmente, pois nota-se que todas as empresas responsáveis por ferramentas de tecnologia reconhecem o Apache Hadoop como o próximo passo para adoção de tecnologia *Big Data* nas organizações, na medida em que além de ser uma tecnologia *open source* e também por isso assume-se como um fator de estimulo ao desenvolvimento e possibilidade de otimização de soluções.

Referências

Abouzeid, A., Pawlikowski, K. B., Abadi, D. J., Rasin, A., & Silberschatz, A. (2009). HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads. *PVLDB, 2(1)*, pp. 922–933.

Agrawal, D., Das, S., & Abbadi, A. E. (2010). Big data and cloud computing: New wine or just new bottles. *PVLDB, 3(2)*, pp. 1647–1648.

Agrawal, D., Das, S., & El Abbadi, A. (March de 2011). Big Data and Cloud Computing: Current State and Future Opportunities. *ACM*, pp. 530-533.

Agrawal, D., El Abbadi, A., Antony , S., & Das, S. (2010). Data Management Challenges in Cloud Computing Infrastructures. *DNIS*, pp. 1-10.

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., . . . Stoica, I. (2010). A View of Cloud Computing. *Communications of the ACM 53*, pp. 50-58.

Andrieux, A., Czajkowski, K., Dan, A., Keahey, K., Ludwig, H., Nakata, T., . . . Xu, M. (Março de 2007). *Web services agreement specification (WS-Agreement),.* Obtido de http://scheme.ogf.org/documents/GFD.107.pdf.

Asif, Q. G. (2010). A Framework to Assist in the Assessment and Tailoring of Agile Software Development Methods. *PhD Thesis, UTS*.

Berberova, Diana; Bontchev, Boyan. (2009). Design of service level agreements for software. *CompSysTech '09*. New York, NY, USA: ACM, pp. 131-139.

Bianco, P., Lewis, G. A., & Merson, P. (2008). *Service level agreements in service-oriented architecture environments.* CMU/SEI-2008-TN- 021, Carnegie Mellon University - SEI.

Böhm, M., Leimeister, S., Riedl, C., & Krcmar, H. (2011). Cloud Computing–Outsourcing 2.0 or a new Business Model for IT Provisioning? Application Management.

Borkar, V. R., Carey, M. J., & Li, C. (2012). Big Data platforms: What's next. *XRDS: Crossroads, The ACM Magazine for Students* - Bi*g Data* , pp 44-49.

Breternitz, V. J., da Silva, L. A., & Lopes, F. S. (2013). O uso de Big Data em Computacional social Science: tema que a sociedade precisa discutir.

Brown, B., Chui, M., & Manyika, J. (October de 2011). Are you ready for the era of 'big data' M C K i n s e y G l o b a l I n s t i t u t e.

Casola, V., Mazzeo, A., Mazzocca, N., & Rak, M. (2006). A SLA evaluation methodology in service oriented architectures. In Quality of Protection, volume 23 of Advances in Information Security. Springer. pp. 119-130.

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., . . . Gruber, R. E. (2006). Bigtable: A Distributed Storage System for Structured Data. *OSDI*, pp. 205–218.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly: Business Intelligence and Analytics: From* Big *Data to* Big *Impact*.

Cheng, Y., Qin, c., & Rusu, F. (2012). Big Data Analytics made easy. *SIGMOD '12 Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* New York: pp. 697-700.

Chew, E., Swanson, M., Stine, K., Bartol, N., Brown, A., & Robinson, W. (2008). *Performance measurement guide for information security Technical Report SP-800-55-rev1.* National Institute of Standards and Technology.

Curino, C., Jones, E., Zhang, Y., Wu, E., & Madden, S. (2010). Relational Cloud: The Case for a Database Service. Technical Report 2010-14. *CSAIL, MIT http://hdl.handle.net/1721.1/52606.*

Das, S., Agarwal, S., Agrawal, D., & El Abbadi, A. (2010). ElasTraS: An Elastic, Scalable, and Self Managing Transactional Database for the Cloud. *Technical Report 2010-04 CS, UCSB*.

Das, S., Nishimura, S., Agrawal, D., & El Abbadi, A. (September de 2010). Live Database Migration for Elasticity in a Multitenant Database for Cloud Platforms. *Technical Report 2010-09, CS, UCSB*.

Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified data processing on large clusters. In Sixth Symposium on Operating System Design and Implementation (*OSDI):* 2004, San Francisco, pp. 137–150.

Dean, J., & Ghemawat, S. (2010). MapReduce: A Flexible Data Processing Tool. *Communications of the ACM - Amir Pnueli: Ahead of His Time*. New York pp. 72-77.

Demirken, H., & Delen, D. (2012). *Leveraging the capa*Bi*lities of service-oriented decision support systems: Putting analytics and* Big *data in cloud,* Department of Management Science and Information Systems, Spears School of Business, Oklahoma State University, United States, pp 412-421 .

Dumbill, E. (14 de Janeiro de 2014). *http://www.forbes.com/sites/edddum*Bi*ll/2014/01/14/the-data-lake-dream/*.

Foster, I., Zhao, Y., Raicu, I., & Shiyong, L. (November de 2008). Cloud Computing and Grid Computing 360-Degree Compared. *Grid Computing Environments Workshop, 2008. GCE '08*, pp. 1-10.

Francis, L. (2009). Cloud Computing: Implications for Enterprise Software Vendors (ESV), System Design and Management Program. *Massachusetts Institute of Technology*.

Friedman, T., Beyer, M. A., & Thoo, e. (2010). *Magic Quadrant for Data Integration Tools.* Gartner.

Ghemawat, S., Gobioff, H., & Leung, S. T. (2003). The Google file system. In 19th ACM Symposium on Operating Systems Principles.

Gualtieri, M. (2013). Bi*g Data Predictive Analytics Solutions, Q1 2013.* Massachusetts: Forrester.

Guolinag, L., Beng, C. O., Jianhua, F., Jianyoung, W., & Lizhu, Z. (2008). EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data. *SIGMOD '08 Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pp. 903-914.

Hagel, J., & Brown, J. S. (October de 2001). Yout Next IT Stratetgy.

Hale, M. L., & Gamble, R. (2012). SecAgreement: Advancing security risk calculations in cloud services. *SERVICES'12* (pp. 133-140). IEEE Eighth World Congress on Services.

Halper, F. (September de 2013). How To Gain Insight From Text. *TDWI Checklist Report*.

Halper, F. (August de 2014). Demystifying Cloud BI. *TDWI Checklist Report*.

Halper, F., & Krishnan, K. (2013). TDWI Big Data Maturity Model Guide Interpreting Your Assessment Score. *TDWI Benchmark Guide 2013–2014*.

Hayden, L. (2010). IT Security Metrics: A Practical Framework for Measuring Security & Protecting Data.

Henning, R. R. (1999). Security service level agreements: Quantifiable security for the enterprise? In Proceedings of the 1999 Workshop on New Security Paradigms. *NSPW'99* (pp. 54-60). ACM.

Hilbert, M. (15 de January de 2013). Big Data for Development: From Information to Knowledge Societies.

Howe, D., Costanzo, M., Fey, P., Gojobri, T., Hannick, L., Hide, W., . . . Rhee, S. Y. (2008). Big data: The future of biocuration. *Nature Internacional Weekly journal of science*, pp47-50.

ISACA. (Maio de 2012). *http://www.isaca.org/CO*BI*T/Pages/default.aspx,* obtido de http://www.isaca.org/COBIT/Pages/default.aspx.

Jaatun, M. G., Bernsmed, K., & Undheim, A. (2012). Security slas – an idea whose time has come? In Multidisciplinary Research and Practice for Information Systems, volume 7465 of Lecture Notes in Computer Science. (pp. 123-130). Berlin: Springer.

Jacobs, D., & Aulbach, S. (2007). Ruminations on multi-tenant databases. *BTW*, pp. 514–521.

Kao, T. C., Chang, C. Y., & Chang, K. C. (2012). Cloud SSDLC: Cloud Security Governance Deployment Framework in Secure System Development Life Cycle.. Liverpool: Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference, pp. 1143-1148.

Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit: The Defi nitive Guide to Dimensional Modeling, Third Edition.* Wiley 10475 Crosspoint Boulevard Indianapolis, IN 46256: John Wiley & Sons, Inc.

Krautsevich, L., Martinelli, F., & Yautsiukhin, A. (2010). Formal approach to security metrics: what does more secure mean for you? *In Proceedings of the Fourth European Conference on Software Architecture: Companion Volume* (pp. 162-169). ACM.

Lamanna, D. D., Skene, J., & Emmerich, W. (2003). SLAng: A language for defining service level agreements. In Proceedings of the 9th IEEE Workshop on Future Trends in Distributed Computing Systems, FTDCS'03. *Proceedings of the 9th IEEE Workshop on Future Trends in Distributed Computing Systems* pp. 100-106.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META Group Research Note, 6.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (21 de December de 2010). Big Data, Analytics and the Path From Insights to Value.

Lohr, S. (2012). *The Age of* Bi*g Data. .* The New York Times.

Ludwig, H., Keller, A., Dan, A., King, R. P., & Franck, R. (2003). *Web service level agreement (WSLA) language specification. Technical, IBM.* http://www.research.ibm.com/wsla/WSLASpecV1-20030128.pdf.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (May de 2011). Big data: The next frontier for innovation, competition, and productivity.

Mather, T., Kumaraswamy, S., & Latif, S. (2009). Cloud Security and Privacy: An Enterprise Perspective on Risks and Compliance. Em *Cloud Security and Privacy: An Enterprise Perspective on Risks and Compliance*. O'Reilly Media, Inc, pp. 109 - 164.

McAfee, A., & Bryjolfson, E. (Octeber de 2012). Big Data: The Management Revolution. *Harvard Business Review*.

Mell, P., & Grance, T. (2011). The NIST Definition of Cloud Computing. *NIST special publication800 (145)*.

Muller, N. J. (1999). Managing service level agreements. *International Journal of Network Management*, 155-166.

Olivier, B., Thomas, B., Heinz , D., Hanspeter , C., Babak , F., Markus , F., . . . Markus , Z. (6 de November de 2012). White Paper Cloud Computing. *Swiss Academy of Engineering Sciences*.

Oswaldo, T., Pjotr, P., Marc, S., & Ritsert, C. J. (March de 2010). Big data, but are we ready. pp. 647-657.

Pavlo, A., Paulson, E., Rasin, A., Abadi, D. J., DeWitt, D. J., Madden, S., & Stonebraker, M. (2009). A comparison of approaches to large-scale data analysis. *SIGMOD*, pp. 165–178.

Putri, N. R., & Mganga, M. C. (Janeiro de 2011). Enhancing information in cloud computing services using sla based metrics. Master's thesis, School Computing. *Blekinge Institute of Technology*.

Qi, Z., Lu, C., & Raouf, B. (20 de April de 2010). Cloud Computing state of the art and research challanges. *Cloud Computing state of the art and research challanges*, pp. 7 - 18.

Qin, H. F., & Li, Z. H. (2013). Research on the Method of Big Data Analysis. *2013 Asian Notwork for Scientific Information*, pp. 1-7.

Raj, P., & Deka, G. C. (2012). *Handbook of Research on Cloud Infrastructures for* Big *Data Analytics.* Information Science: IGI Global.

Reinwald, B. (2010). Database support for multi-tenant applications. *IEEE Workshop on Information and Software as Services*.

Russom , P. (May de 2011). Big Data Analytics. *TDWI Best Practices Report Fourth Quarter 2011:* Bi*g Data Analytics* .

Russom, P. (2013). Managing Big Data. *TDWI Best Practices Report Fourth Quarter 2013*.

Schönberger, V. M., & Cukier, K. (2013). Bi*g Data: A Revolution that Will Transform how We Live, Work, and Think.* Houghton Mifflin Harcourt - 242 pages.

Sikka, V., Färber, F., Lehner, W., Cha, S. K., Peh, T., & Bornhövd, C. (2012). Efficient transaction processing in SAP HANA database: the end of a column store myth. *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pp. 731-742.

Silva, C. A., Ferreira, A. S., & Geus, P. L. (2012). A methodology. *Proceedings of the IEEE Latin American Conference on Cloud Computing and Communications, LatinCloud'12* (p. for management of cloud computing using security criteria). Porto, Alegre, Brazil: IEEE Latin American Conference on Cloud Computing and Communications.

Siwach, G., & Esmailpour, A. (2014). Encrypted Search & Cluster Formation in Big Data. *ASEE 2014 Zone I Conference* (pp. 1-4). ASEE 2014 Zone I Conference.

Soares, E. (March de 2013). Big Data : Volume de dados no mundo crescerá 60% em 2012. *Computerworld 2012*.

Stodder, D. (2013). Data Visualization and Discovery for Better Business Decisions. *TDWI Best Practices Report*.

Stonebraker, M. (21 de September de 2012). What Does 'Big Data' Mean? *What Does 'Big Data' Mean? | blog@CACM | Communications of the ACM*.

Stubbs, E. (2014). Big *Data,* Big *Innovation: Enabling Competitive Differentiation through Business Analytics.* John Wiley & Sons - 256 pages.

Stuckenberg, S., Fielt, E., & Loser, T. (2011). The Impact Of Software-As-A-Service On Business Models Of Leading Software Vendors. *Experiences From Three Exploratory Case Studies. Proceedings of the 15th Pacific Asia Conference on Information Systems (PACIS 2011).*

Tankard, C. (2012). Big *data security*, Network Security, Volume 2012, Issue7, July 2012, Pages 5 -8, ISSN 1353-4858.

The Center for Internet Security. (23 de Setembro de 2014). Obtido de https://benchmarks.cisecurity.org/tools2/metrics/CIS_Security_Metrics_v1.1.0.pdf.

TM Forum. (2005). *Sla management handbook - volume 2. Technical Report GB9712, TeleManagement Forum.*

Tran, N., Sreenath, B., & Jaimin, D. (2013). Designing query optimizers for big data problems of the future. *Journal Proceedings of the VLDB Endowment Volume 6 Issue 11, August 2013*, 1168-1169.

Vaishnavi, V. K., & Kuechler Jr., W. (2008). *Design Science Research Methods and Patterns: Innovating Information and Communication Technology.* CRC Press.

Vaishnavi, V. K., & Kuechler, W. (2004). Design Science Research in Information Systems.

van Aalst, W. M., van Hee, K. M., van Werf, J. M., & Verdonk, M. (March de 2010). Auditing 2.0: Using Process Mining to Support Tomorrow's Auditor. *Computer (Volume:43 , Issue: 3)*, pp. 90-93.

van der Aalst, W., Adriansyah, A., & van Dongen, B. (30 de January de 2012). Replaying history on process models for conformance checking and performance analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 2, Issue 2*, pp. 182-192.

Vaquero, L. M., Merino, L. R., Caceres, J., & Lindner, M. (1 de January de 2009). A break in the clouds: towards a cloud definition. *ACM SIGCOMM Computer Communication Review*, pp. 50-55.

W3C. (Setembro de 2007). *Web Services Policy 1.5 - Framework. World Wide Web Consortium.* Obtido de http://www.w3.org/TR/ws-policy/.: http://www.w3.org/TR/ws-policy.

Wasniowski, R. A. (2014). A Cloud Oriented Framework for Scientific Data Processing.

Webster, J., & Watson, R. T. (2002). Analysing The Past to Prepare for the Future: Writing a Literature Review.

Yang, F., Shanmugasundaram, J., & Yerneni, R. (2009). A scalable data platform for a large number of small applications. *CIDR*.

Yoon, J. P. (2011). *Access Control And Trustiness for Resource Management in Cloud Databases.* Springer.

Youseff, L., Butrico, M., & Da Silva, D. (2008). Toward a Unified Ontology of Cloud Computing. Grid. *GCE'08*.

Anexos

Anexo A – Plano de Trabalho

1 Product Breakdown Structure (PBS)

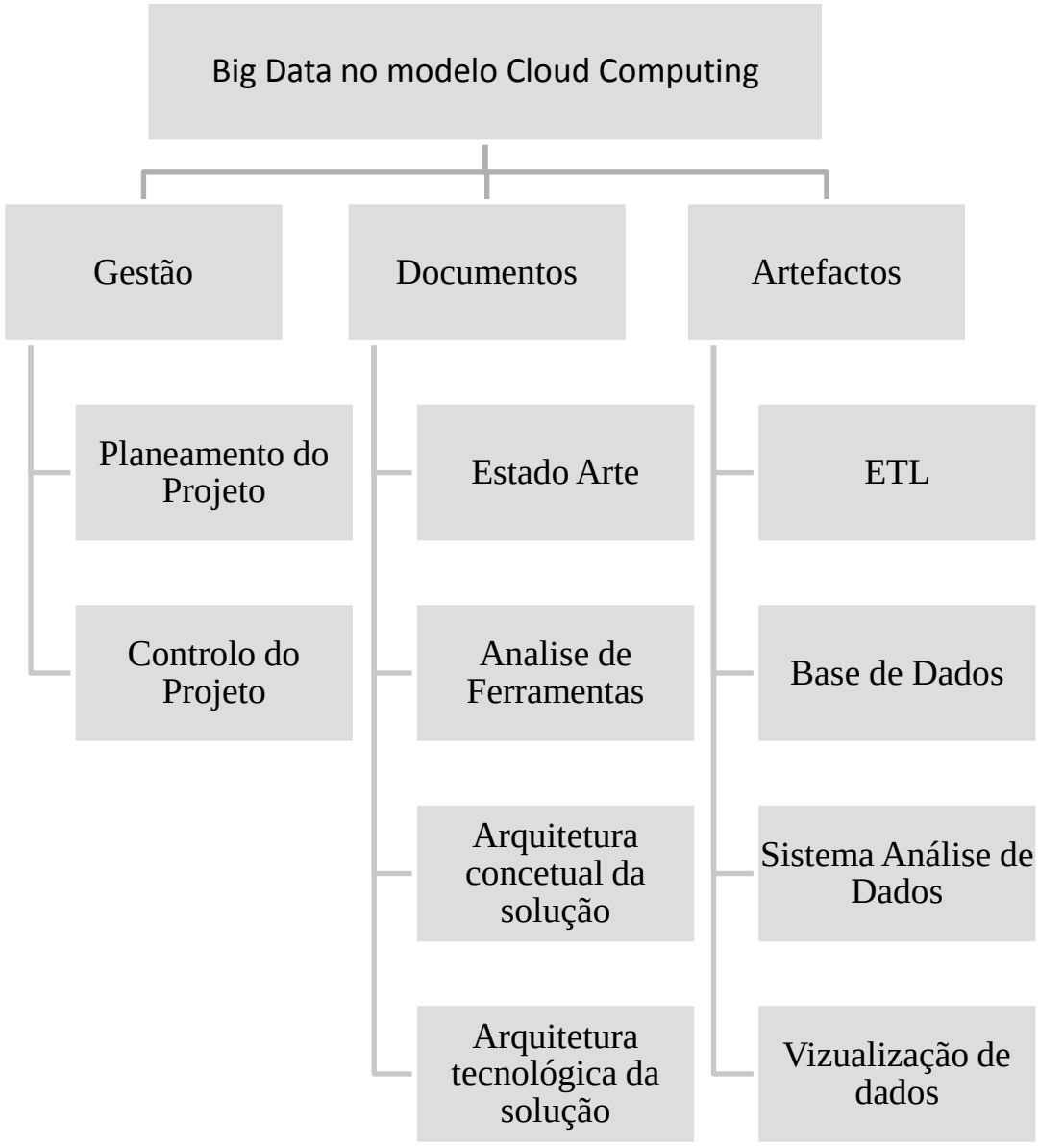


Figura 1 - *Product Breakdown Strusture*

Product Breakdown Structure (*PBS*) tem como objetivo descrever o produto que será desenvolvido no decorrer da realização do projeto de dissertação de mestrado. A *PBS* encontra-se dividida em três áreas: Gestão, Documentos, Artefactos:

- **Gestão:** Esta área procura explicar as tarefas associadas à gestão de todo o trabalho a efetuar no projeto, sendo necessário planear as atividades a desenvolver, estimar o tempo necessário para o desenvolvimento de cada uma, efetuar um controlo das atividades planeadas de forma a saber o ponto em que o projeto se encontra, se o planeamento do mesmo está ou não a ser cumprido, procurando, se tal se justificar, ajustar as atividades ou modificar o planeamento dependendo dos resultados da análise e do estado em que o projeto se encontra.
- **Documentos:** Esta área identifica todos os documentos a produzir, procurando mostrar o trabalho que é produzido, descrevendo como o mesmo foi realizado. Os documentos assumem um papel de extrema importância, na medida em que, ajudam a perceber o trabalho que foi realizado, em termos de resultados, fornecem simultaneamente as informações de relevo acerca do conteúdo e das etapas associadas a todo o processo.
- **Artefactos:** Esta área representa o resultado de todo o trabalho realizado. Os artefactos representam o produto final que é esperado obter com a realização deste projeto. O sucesso de um projeto depende em muito dos frutos do trabalho realizado, partindo do pressuposto de que se os artefactos estiverem de acordo com os objetivos do projeto o resultado final é bom. No caso dos artefactos não estarem de acordo com os objetivos do projeto o resultado é mau. Neste último caso será necessário executar novamente algumas atividades, procurando corrigir os erros detetados, obtendo um artefacto de acordo com os objetivos do projeto.

2 Work Breakdown Structure (WBS)

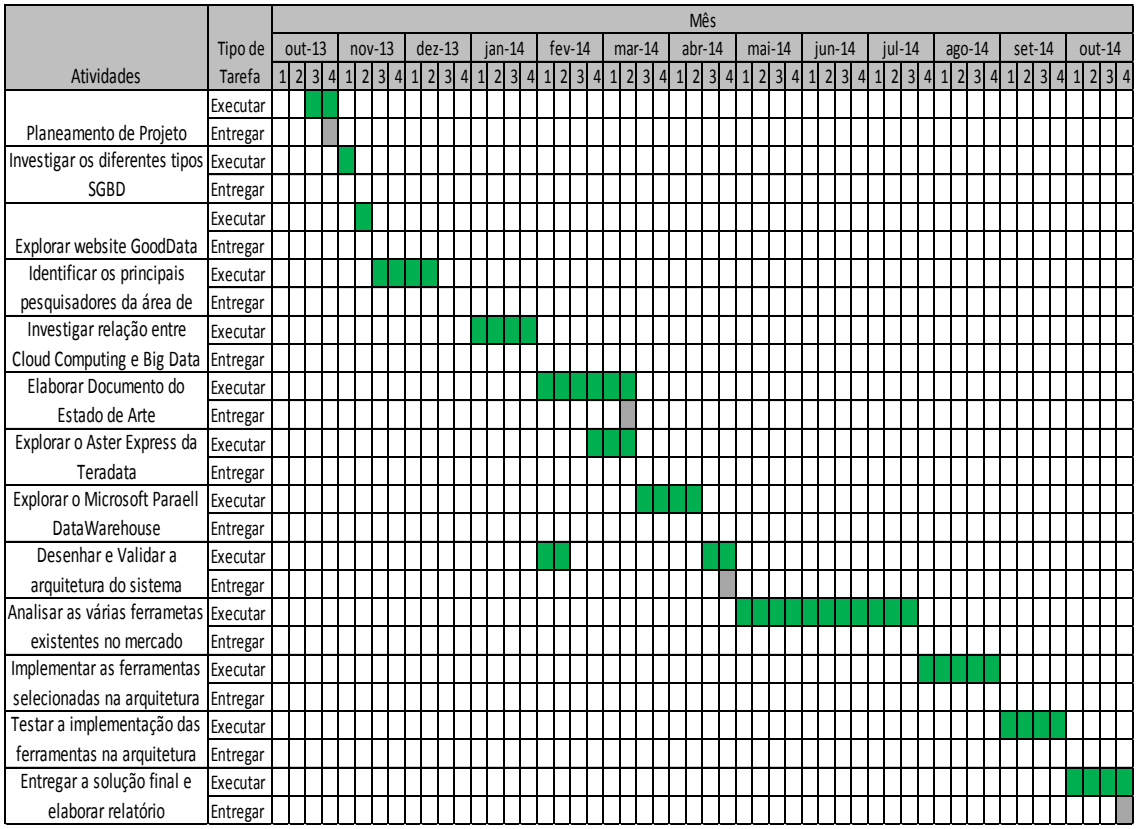


Figura 2 - Work Breakdown Structure

A *Work Breakdown Structure* (*WBS*), procura mostrar com maior detalhe as atividades a desenvolver ao longo do projeto com o objetivo de construir uma solução de acordo com os requisitos definidos na *PBS*. Através da *WBS* é possível saber a sequência de atividades a executar, permitindo a quem está a gerir saber a qualquer momento em que fase o mesmo se encontra, viabilizando assim a possibilidade de ajustar as atividades de acordo com o tempo disponível. A *WBS* apresentada anteriormente identifica as principais atividades a realizar ao longo do projeto. A cor verde representa o tempo estimado para a execução de uma determinada atividade, a cor cinzento identificada as datas das entregas dos documentos associados ao projeto. A WBS procura ajudar quem está a realizar o projeto consiga cumprir os prazos acordados.

3 Análise de Riscos

Risco	**Probabilidade (1 a 5)**	**Impacto (1 a 5)**	**Indicadores**	**Principais Consequências**	**Estratégia de Resolução do Problema**
Atrasos nas atividades de desenvolvimento	2	5	Dificuldade em cumprir os prazos estabelecidos	Atrasos na conclusão, aumento do tempo de realização do projeto	Aumentar o número de horas de trabalho
Gestão do âmbito do sistema	3	4	Atrasos nas tarefas a realizar	Atrasos na elaboração de documentação	Rever o plano de trabalho e executar as atividades em atraso
Alteração dos requisitos depois da análise inicial	1	4	Existência de dúvidas da parte do cliente em relação ao que pretende	Aumento do tempo de desenvolvimento do projeto	Realizar uma reunião para clarificar os objetivos do projeto
Falta de competências	2	5	Falta de conhecimento de quem executa o projeto	Atrasos no projeto	Obter o conhecimento necessários do Software
Falta de análise e relatório de erros a resolver	3	5	Atrasos na execução do projeto	Incumprimento dos prazos estabelecidos	Rever o plano de trabalho e executar as atividades em atraso
Rejeição do produto final por parte do cliente	1	5	Não satisfação do cliente em relação ao produto final	Aumento do tempo de desenvolvimento do projeto	Perceber quais os problemas e proceder a sua resolução
Má escolha dos softwares a utilizar	3	3	Incapacidade de resposta do *Software* ao pretendido	Insatisfação do cliente e necessidade de detetar as falhas	Acompanhamento contínuo do projeto por parte do *Stakeholder*
Perda de Dados	3	4	Dificuldade em executar as atividades	Atrasos na execução das atividades devido aos atrasos	Efetuar cópias de segurança de todo o trabalho realizado
Falta de conhecimento dos futuros utilizadores	3	3	Dificuldade em executar as tarefas	Utilização errada das várias ferramentas	Obter o conhecimento necessários das ferramentas a utilizar
Esconder a estado atual do projeto	1	4	Esconder a informação verdadeiros do estado do projeto	Atrasos e aumento do trabalho a executar	Assinalar as atividades executadas à medida que estas são concluídas
Pressão por parte do *Stakeholder*	1	2	Possibilidade de ocorrerem pressões que condicionem o desempenho	Possibilidade de uma diminuição do desempenho	Integrar continuamente o *Stakeholder* em todas as fases do projeto
Trabalho efetuado de forma defeituosa	1	4	Não concordância entre o estado atual do trabalho e a previsão inicial	Atrasos na entrega e necessidade de corrigir os erros detetados	Efetuar uma validação contínua do trabalho executado
Perda de credibilidade e confiança da empresa	1	5	Atrasos no projeto e indecisões do *Stakeholder*	Atrasos na tomada de decisões por parte do *Stakeholder*	Envolver o *Stakeholder* em todas as atividades do projeto
Avaria de Hardware	2	4	Problemas numa máquina que leva ao incumprimento de atividade	Atrasos na execução das atividades e necessidade de obter novo *Hardware*	Executar o trabalho em máquinas virtuais e fazer cópias de segurança
Indisponibilidade de acesso ao servidor em Cloud	1	3	Impossibilidade de acesso ao servidor *Online*	Limitação nas comunicações e dificuldade em realizar as atividades	Ter máquinas virtuais locais e poder atualizar depois a plataforma online
Má gestão do tempo e dos recursos do projeto	1	4	Execução de tarefas fora do planeado	Atraso na entrega do *Software*	Aumentar o número de horas de trabalho e cumprir o planeamento

Risco	Probabilidade (1 a 5)	Impacto (1 a 5)	Indicadores	Principais Consequências	Estratégia de Resolução do Problema
Atrasos na entrega da solução final	1	5	Atrasos nas tarefas e nas atividades planeadas	Insatisfação do cliente	Verificar o planeamento com frequência e reajustar o trabalho

Tabela 1 - Análise de Riscos

A Análise de Risco é muito importante num projeto, visto que se assume como estruturante na identificação dos aspetos que devem ser alvo de maior atenção no decurso de todo o processo. Isto, porque possibilita a aferição das atividades que podem ter risco de maior falha, permitindo, assim, perceber o risco que corremos (e as possíveis consequências e problemas que daí possam advir) se for cometido um erro ou existir um atraso numa determinada atividade. Neste sentido ao identificarmos as condicionantes, podemos atuar de forma mais preventiva, definindo as medidas a adotar para evitar que os riscos possam acontecer ou caso estes aconteçam evitar que os danos causados pelos mesmos sejam significativos.

Anexo B – Matriz de Conceito

			Conceitos				
			Evolução Tecnologica			Arquitetura de Sistemas Análises de dados Big Data em Cloud Computing	
Título	Autor	Ano	Organizations	Big Data and Innovation	Big Data Infrastuctured and Tecnologies	Cloud Computing	Arquiteturas de Sistemas Big Data em Cloud Computing
Clouds, big data, and smart	Jacques Bughin,	2011	X	X	X	X	
Data	Divyakant Agrawal,	2011	X	X			
State and Future Opportunities	Divyakant Agrawal	2012	X	X		X	X
MapReduce and DBMS Technologies for	Abouzeid, A.,	2009	X	X	X	X	X
Big data and cloud computing:	Agrawal, D., Das,	2010	X	X		X	
Data Management Challenges in	Agrawal, D., El	2010			X		X
new Business Model for IT Provisioning?	Böhm, M.,	2011	X			X	
social Science: tema que a sociedade	Breternitz, V. J.,	2013	X	X			
c k i n s e y g l o b a l i n s t i t u t e Are	Brown, B., Chui,	2011	X		X		
for Structured Data	Chang, F., Dean, J.,	2006			X	X	X
From Big Data to Big Impact	Chen, H., Chiang,	2012			X	X	X
Database Service. Technical Report	Curino, C., Jones,	2010			X		X
Managing Transactional Database for	Das, S., Agarwal,	2010			X	X	
a Multitenant Database for Cloud	Das, S., Nishimura,	2010			X	X	
on large clusters.	Dean, J., &	2004			X		X
Cloud Computing and Grid	Foster, I., Zhao, Y.,	2008				X	
Enterprise Software Vendors (ESV),	Francis, L.	2009			X	X	
method for unstructured, semi-structured	Guoliang, L., Beng,	2008		X			
How To Gain Insight From Text	Halper, F.	2013	X	X			
GUIDE Interpreting Your Assessment	Halper, F., &	2013	X	X			
Information to Knowledge Societies.	Hilbert, M.	2013	X	X	X		
Ruminations on multi-tenant databases.	Jacobs, D., &	2007		X			X
nitive Guide to Dimensional Modeling,	Kimball, R., &	2013		X		X	
innovation, competition, and productivity	Manyika, J., Chui,	2011	X	X			
Harvard Business Review.	McAfee, A., &	2012	X				
Computing. NIST special publication800	Mell, P., & Grance,	2011				X	
Academy of Engineering Sciences.	Olivier, B.,	2012				X	
Big data, but are we ready.	Oswaldo, T., Pjotr,	2010	X	X			
scale data analysis.	Pavlo, A., Paulson,	2009		X		X	
research challanges. Cloud Computing	Qi, Z., Lu, C., &	2010				X	
applications. IEEE Workshop on	Reinwald, B.	2010			X	X	
PRACTICES REPORT FOURTH QUARTER	Russom , P.	2011	X	X			
MANAGING BIG DATA	Russom, P.	2013	X				
HANA database: the end of a column	Sikka, V., Färber,	2012	X				X
crescerá 60% em 2012.	Soares, E.	2012	X	X			
FOR BETTER BUSINESS DECISIONS.	Stodder, D.	2013			X		X
'Big Data' Mean?	Stonebraker, M.	2012	X	X			
Business Models Of Leading Software	Stuckenberg, S.,	2011		X	X	X	
problems of the future.	Tran, N., Sreenath,	2013			X		X
Support Tomorrow's Auditor. Computer	van Aalst, W. M.,	2010	X	X		X	
conformance checking and performance	van der Aalst, W.,	2012	X	X		X	
definition.	Vaquero, L. M.,	2009				X	X
number of small applications.	Yang, F.,	2009			X		X
Computing. Grid.	Youseff, L.,	2008				X	X
http://www.forbes.com/sites/	Dumbill, E.	2014		X	X		
The Data Warehouse Toolkit:	Kimball, R., &	2013	X		X		X
3D data management:	Laney, D	2001	X	X			X

Anexo C - Paper CPASI 2014

Uma Arquitetura Moderna de Dados: Um Caso de Teste[1]

César Silva Martins[1], Paulo Simões[2], Jorge Oliveira e Sá[3]

[1] Universidade do Minho, Portugal
pg21441@alunos.uminho.pt

[2] ISEG – School of Economics and Management, Portugal
paulosimoes@iseg.utl.pt

[3] Universidade do Minho, Portugal
jos@dsi.uminho.pt

Resumo

Atualmente os dados são vistos como tendo tipos e origens distintas. Os tipos de dados podem ser estruturados, semiestruturados e não estruturados. As origens dos dados podem ser diversas como *Enterprise Resource Planning* (ERP), *Customer Relationship Management* (CRM), *Supply Chain Management* (SCM), folhas de cálculo, documentos de texto, redes sociais, imagens, vídeos, sensores entre outros.

Esta diversidade de dados exige uma arquitetura moderna que permita a recolha dos dados de várias origens e tipos, viabilizando igualmente a extração, transformação e limpeza dos mesmos através do processo de *Extract, Transform and Load* (ETL), bem como o armazenamento e integração dos dados para posteriores análises. Esta arquitetura deve ser suportada por um ambiente de *Cloud Computing*, garantindo assim a sua atualidade, ubiquidade e fácil acesso pelos utilizadores.

Este artigo propõe-se desenvolver uma arquitetura que permita ajudar a melhorar o processo de tomada de decisão.

Palavras chave: *Big Data, Cloud Computing, Cloud Computing Systems Architecture, Big Data Analytics, Technologies for Big Data*.

1. Introdução

Nos dias de hoje, as organizações geram e têm à sua disposição uma elevada quantidade de dados, o que cria a dificuldade em maximizar os proveitos que podem advir da análise dessa riqueza de dados. Por vezes não têm competência para os processar, ou seja, recolher, armazenar e disponibilizar os dados para processos analíticos, perdendo assim capacidade para suportar decisões de caráter técnico ou tático. As organizações são assim confrontadas com dados provenientes de [Kimball & Ross, 2013]:

Este trabalho foi financiado pela FCT - Fundação para a Ciência e Tecnologia no âmbito do projeto: PEst-OE/EEI/UI0319/2014

- aplicações de gestão, tais como: *Enterprise Resource Planning* (ERP), *Customer Relationship Management* (CRM), *Supply Chain Management* (SCM), entre outros;
- folhas de cálculo e documentos, que podem ser internos e externos à organização;
- redes sociais (o que se fala sobre a organização), imagens (nomeadamente de produtos acabados, semiacabados e matérias primas), vídeos (por exemplo explicativos da utilização de produtos, …);
- sensores acoplados a dispositivos eletrónicos, como por exemplo: máquinas fabris (temperatura, vibração, interrupções em tempo e códigos de paragens, horas trabalhadas, etc.), alarmes (de intrusão, …) e câmaras de vídeo, registos de eventos (*logs*) de *routers* e *firewalls*;
- dispositivos móveis que são cada vez mais adotados nas organizações, permitindo que a capacidade de processamento esteja cada vez mais distribuída.

Assim, verifica-se a necessidade de soluções que permitam a recolha e integração de dados (de tipos distintos) provenientes de várias fontes, mas, para além disso, que possibilitem a análise desses dados em qualquer lugar (espaço) e momento (tempo), pois o rápido acesso aos dados é crucial para que os processos de tomada de decisão sejam mais rápidos, práticos e eficientes possíveis [Russom, 2013].

Estes dados apresentam características distintas, na medida em que podem ser estruturados, semiestruturados ou não estruturados. Devido à grande diversidade de dados surge o termo *Big Data* que engloba tanto os diferentes tipos de dados como as suas origens. O termo *Big Data* abarca múltiplos conceitos, gerando, por isso muita confusão e mal entendidos.

É comumente aceite que o conceito *Big Data* foi pela primeira vez formulado por Doug Laney em Fevereiro de 2001, onde analisou os desafios que as empresas enfrentavam na gestão dos dados. Dessa forma foram definidas três importantes dimensões associadas aos dados, ver figura 1: Volume (caracteriza as grandes quantidade de dados); Velocidade (identifica a necessidade de captar, armazenar e analisar dados de forma rápida, ajudando a um maior suporte das atividades operacionais); e Variedade (capacidade de tratar, de forma integrada tipos de dados com características distintas) [Laney, 2001].

De notar que, já em finais do século XX, várias empresas reportavam o armazenamento e análise de grandes volumes de dados estruturados. Empresas como a *WallMart* ou o *Bank of America* reportavam, já nessa época a utilização de um *Data Warehouse* (DW) com várias dezenas ou centenas de *Terabytes* (TB) [Lohr, 2012].

Assim a inovação no conceito de *Big Data* não está tanto na necessidade de tratar muitos dados (Volume), nem em fazê-lo de forma muito rápida (Velocidade), mas mais na possibilidade de o fazer sobre dados com características diferentes (Variedade). Com efeito, se os dados a analisar fossem apenas os típicos dados estruturados (atributos numéricos ou alfanuméricos de relações no modelo entidade relacionamento(ER)) as tecnologias de bases de dados relacionais suportadas em processamento paralelo massivo *Massive Parallel Processing* (MPP), seriam suficientes para corresponder aos desafios. É a dimensão Variedade (necessidade de processar e analisar dados não estruturados), que introduz uma necessidade disruptiva com a informática atual [Kimball & Ross, 2013].

A estes três vetores iniciais (que passaram a ser referidos como “três V’s”) que originalmente definiam o conceito de *Big Data*, foram associadas outras características que contribuíram para a proliferação do mesmo, nomeadamente o crescimento e massificação do uso das redes sociais e dispositivos móveis [Stonebraker, 2012].

Desta forma surge a necessidade de uma arquitetura que permita dar resposta a todas estas necessidades.

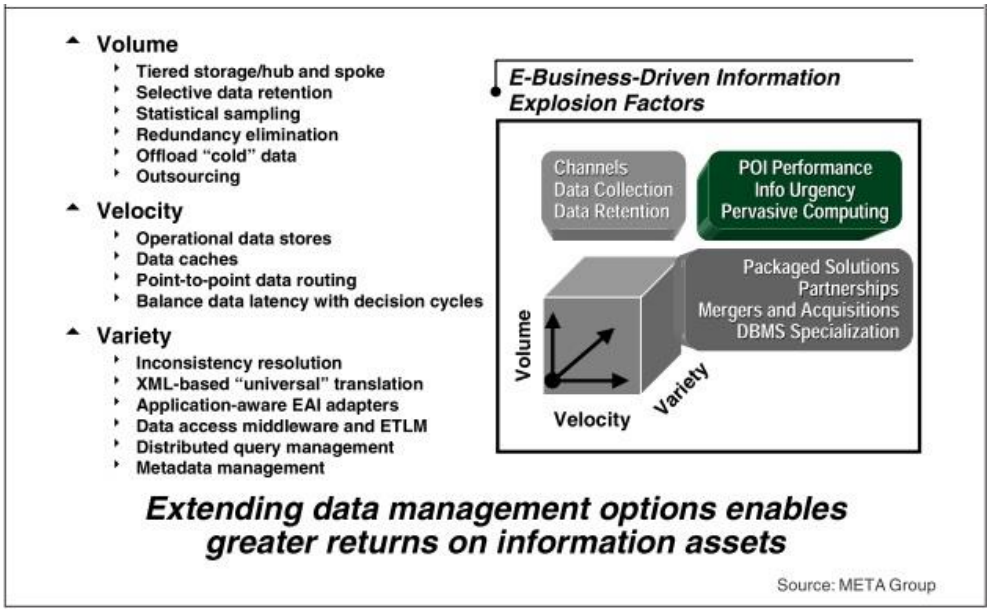


Figura 18 – 3D Data Management Controlling [Laney, 2001]

A arquitetura de dados deve permitir uma integração de dados provenientes de diferentes fontes, isto exige um conhecimento exaustivo das fontes de dados, bem como de todo o processo necessário para extrair, validar e transformar os dados de forma a auxiliar o processo de tomada de decisão [Russom, 2013].

É necessário que os dados sejam armazenados num repositório, procedendo-se depois a sua análise. Devido aos dados serem do tipo *Big Data*, estes repositórios exigem tecnologias criadas especificamente para este tipo de soluções.

Finalmente, os dados devem ser disponibilizados para análises, onde os utilizadores (decisores, analistas, entre outros) contactam e exploram os dados existentes. Pretende-se que essa exploração seja o mais dinâmica possível, ou seja, o utilizador pode definir quais as variáveis e valores a explorar, as análises podem ser efetuados através de *dashboards*, *reports*, *Online Analytical Processing* (OLAP), etc. Para se poder tomar decisões de forma rápida, sustentada e eficiente, é necessário que a arquitetura permita análises de dados em qualquer lugar e momento tudo isto pode ser obtido recorrendo-se ao ambiente de *Cloud Computing*. (Agrawal et al., 2010).

Assim o objetivo deste trabalho é a elaboração de uma arquitetura conceptual e física que permita cobrir as necessidades descritas atrás. Terá que ser uma arquitetura moderna para suportar *BigData* e *Cloud Computing*, nomeadamente ao nível da integração, armazenamento e disponibilização de dados.

Este artigo pretende responder à seguinte questão: “Quais as caraterísticas de uma arquitetura de dados para recolher, armazenar e disponibilizar dados para posteriores análises?“.

Assim, será inicialmente apresentada uma arquitetura conceptual, posteriormente essa arquitetura conceptual será instanciada e implementada com soluções tecnológicas existentes no mercado, permitindo suportar a solução pretendida. Finalmente será desenvolvido um caso de teste que servirá de prova de conceito e que tratará dados da área da saúde provenientes de um *dataset* aberto (*open data*).

2. Enquadramento teórico

As secções que constituem este capítulo, apresentam uma abordagem teórica sobre *Big Data* e *Cloud Computing*, o objetivo é fornecer uma perspetiva sobre o tema. Na secção 2.1 aborda-se o conceito de *Big Data* e apresentam-se algumas definições, aproveita-se para referir a importância dos dados que são utilizados diariamente nas organizações para suportar o processo de decisão. Na secção 2.2 aborda-se o conceito de *Cloud Computing*, este é um conceito que tem vindo a crescer de importância, pois nota-se a crescente necessidade das organizações em aceder aos seus dados de forma constante, ou seja, não depender de um determinado local, nem de um determinado horário, sendo essa uma das características que coloca o *Cloud Computing* no topo das preferências das organizações, no que concerne às Tecnologias de Informação (TI).

2.1 Big Data

O termo *Big Data* é cada vez mais utilizado, sobretudo porque as organizações estão a perceber o valor e as mais-valias que podem obter através das enormes quantidades de dados que possuem, podendo os mesmos ser catalisadores no processo de inovação da organização através de perceções e adequação da realidade atual face à estratégia da organização [Stonebraker, 2012].

Todas as organizações, umas mais rapidamente que outras, estão a perceber a importância do *Big Data*, compreendendo as mudanças necessárias a efetuar para beneficiar das vantagens competitivas da utilização do *Big Data* [Oswaldo et al., 2010]. O crescimento, proliferação e influência das redes sociais na sociedade foram os principais fatores de influência para o crescimento e importância do termo *Big Data* [Manyika et al., 2011].

A capacidade para tratar *Big Data* não pode ser conseguida com uma única tecnologia, mas sim como uma combinação de várias tecnologias, umas mais antigas e outras mais recentes e que procuram ajudar as organizações a obter uma maior eficácia na gestão de grandes quantidade de dados, bem como na resolução dos problemas associados ao seu armazenamento. Os dados podem ter origem em diversas fontes internas e externas, tais como *streaming* de dados, *logs* de navegação em *sites Internet,* redes sociais e *medias* sociais, dados geo-espaciais, textos, imagens, entre outros podendo estes dados ter diferentes tipos de estruturas [Halper & Krishnan, 2013], a saber:

- **Dados Estruturados:** São todos os dados que são organizados em blocos semânticos (entidades), as entidades são agrupadas através de associações e classes, uma

entidade de um determinado grupo pode possuir as mesmas descrições e atributos que podem ser efetuados para todas as entidades de um grupo, podendo desta forma conter o mesmo formato, tamanho e seguir a mesma ordem. Todos estes dados são guardados em Sistemas de Gestão de Bases de Dados (SGDB). São chamados dados estruturados, porque possuem a estrutura rígida que foi previamente projetada através de um modelo entidade associação. Como exemplos de dados estruturados temos os dados provenientes de ERPs, programas de gestão, históricos das últimas compras realizadas por um cliente, transações realizadas com cartões de crédito, etc. [Guoliang et al., 2008].

- **Dados Semiestruturados:** São dados que não necessitam de estar armazenados num SGBD e que apresentam um elevado grau de heterogenidade, o que faz deles dados não generalizados num qualquer tipo de estrutura. Como exemplos de dados semiestruturados temos: *Extensible Markup Language* (XML), *Resource Description Framework (RDF), Web Ontology Language* (OWL), o conteúdo de um *email* etc. [Guoliang et al., 2008].
- **Dados Não Estruturados:** São dados que não possuem necessariamente um formato ou sequência, não seguem regras e não são previsíveis. Os dados deste tipo são, atualmente, alvo de muita atenção devido principalmente à proliferação de dispositivos móveis responsáveis pela criação de uma grande variedade de dados. No entanto existem outras fontes de dados como: sensores de máquinas, dispositivos inteligentes, tecnologias de colaboração e redes sociais. Estes dados não são dados relacionados mas sim diversificados. Alguns exemplos deste tipo de dados são: textos, vídeos, imagens, etc. [Guoliang et al., 2008].

As principais características que permitem a diferenciação entre os vários tipos de dados são apresentadas na tabela 1.

Dados Estruturados	**Dados Semiestruturados**	**Dados Não Estruturados**
Estrutura predefinida	Nem sempre existe esquema	Não existe esquema
Estrutura regular	Estrutura irregular	Estrutura irregular
Estrutura independente dos dados	Estrutura embebida nos dados	A estrutura está dependente da fonte dos dados
Estrutura reduzida	Estrutura extensa (particular em cada dado visto que cada um pode ter uma organização própria)	Estrutura extensa depende muito do tipo de dados
Pouco evolutiva e bastante rígida	Muito Evolutiva, a estrutura pode mudar com muita frequência	Muito evolutiva, a estrutura muda com bastante frequência
Possui esquemas fechados e restrições de integridade	Não existe um esquema de dados associado	Não existe um esquema de dados associado
Distinção clara da estrutura de dados	Não é clara a distinção entre estrutura de dados	Não é possível distinguir entre as estruturas dos dados

Tabela 1 - Diferenças entre dados estruturados, semiestruturados e não estruturados, adaptado de [Guolinag et al., 2008]

Big Data também proporcionou um crescimento na complexidade dos dados, fazendo com que os atuais sistemas de gestão de bases de dados (baseados no modelo relacional) tenham dificuldade em armazená-los [Manyika et al., 2011].

O conceito de *Big Data* pode ter várias interpretações, mas é importante realçar que está assente no princípio das três dimensões iniciais **Volume**, **Variedade** e **Velocidade**, as quais se juntaram mais duas dimensões **Veracidade** e **Valor** fruto do trabalho realizado por toda a comunidade científica e académica da área [Stonebraker, 2012].

Passa-se a descrever os 5 V's, que também estão representados na figura 2:

- **Volume** - representa uma grande quantidade de dados a ser recolhida e analisada, tendo em vista a utilização de *Structured Query Language* (*SQL*) *analytics (count, sum, max, min, average e group by)*, regressões, aprendizagem máquina e análises complexas em grandes volumes de dados.
- **Variedade** - corresponde a uma utilização de diferentes estruturas de dados, podemos identificar a variedade de estruturas de dados como: estruturadas, não estruturadas e semiestruturadas.
- **Velocidade** – corresponde ao volume de informação que é gerado de forma rápida e crescente, o que diminui o espaço de tempo entre as tomadas de decisão, o grande desafio aqui é conseguir recolher e armazenar os grandes volumes de dados em tempo útil, procurando utilizar os dados históricos e *real-time* para suportar as decisões operacionais.
- **Veracidade** - permite a classificação das diversas fontes de dados (estruturados, não estruturados e semiestruturados) de acordo com a sua qualidade, de acordo com aspetos como: precisão e atualidade dos dados fornecidos.
- **Valor** - corresponde à importância que os dados utilizados terão nas decisões a tomar.

Considera-se que para um sistema ser considerado de *Big Data* tem que possuir pelo menos dois dos V's anteriormente referidos [Russom, 2013]. Constata-se que nem todos os sistemas de gestão de bases de dados relacionais devem ser considerados *Big Data*, [Manyika et al., 2011].

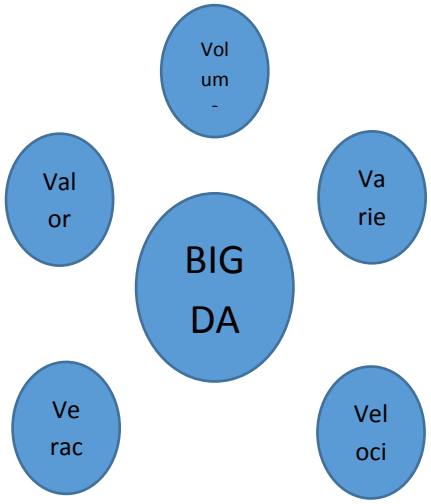


Figura 19 Estrutura do Big Data com os 5V's
adaptado de [Stonebraker, 2012]

2.2 Cloud Computing

O conceito relacionado com *Cloud Computing* surge nos anos 50 do século passado, sendo visto como uma espécie de computação partilhada, ou seja, devido ao elevado custo, o tempo de processador dos *mainframes* era partilhado por diversos utilizadores de forma a eliminar tempos mortos (denominado de *Time-Sharing*) [Francis, 2009].

Nos dias que correm o conceito é utilizado quando nos referimos a tecnologia, flexível que oferece recursos e serviços de TI, com base na *Internet* [Böhm et al., 2011]. *Cloud Computing* também pode ser considerado como um conjunto de conceitos associados a várias áreas de conhecimento como *Service-Oriented Architecture* (SOA), computação distribuída, computação em *Grid* (modelo computacional que divide as tarefas a executar por diversas máquinas) e virtualização [Youseff et al., 2008].

Muito além da visão tecnológica que está associada a *Cloud Computing*, o conceito pode também ser entendido como uma inovação, principalmente na prestação de serviços de TI [Böhm et al., 2011]. Muitos acreditam que este é um potencial a ser explorado, principalmente no modo de desenvolvimento e implementação de recursos de computação e aplicações, procurando novos modelos de negócio principalmente para as empresas fornecedoras de *Software* [Youseff et al., 2008], [Stuckenberg et al., 2011].

Segundo o *National Institute Standards and Technology* (NIST), existem muitos benefícios na adoção de *Cloud Computing*, sendo os mais relevantes: [Olivier et al., 2012]:

- Possibilita economia do lado cliente, na medida em que o fornecedor de serviços tem a responsabilidade de garantir serviços e o suporte das infraestruturas que o suportam, isto permite ao cliente melhorar a sua produtividade, além de diminuir o investimento em infraestruturas por parte do cliente, permitindo uma constante adequação do serviço as necessidades do cliente.
- Permite que as organizações cliente, se tornem mais sustentáveis através de um processo de modernização contínua, exigindo ao prestador de serviços de cloud uma melhoria contínua dos seus serviços através da adequação do hardware e software as necessidades exigidas pelos seus clientes. É também exigido ao fornecedor de serviços um conjunto de políticas de segurança que permitam a existência de parâmetros de segurança que salvaguardem a informação relevante para o negócio de cada um dos seus clientes.

A *Cloud* também possui alguns modelos que tornam possível a sua implementação como soluções comerciais. Atualmente os modelos existentes são:

- ***Cloud* privada:** O utilizador destas soluções é uma organização específica ou uma unidade organizacional, que pode ser interna à organização ou contratada a um fornecedor de serviços em *Cloud*. As vantagens da *Cloud* não podem ser plenamente exploradas através deste modelo devido à limitação do grau de personalização.
- ***Cloud* comunitária:** O serviço é utilizado por vários membros de um grupo, podendo ser oferecido por vários fornecedores, internos ou externos à comunidade.
- ***Cloud* pública:** Serviços disponíveis para o público em geral; o serviço é oferecido por um único fornecedor e neste modelo a estabilidade e os recursos como *pooling* podem ser totalmente explorados.

- ***Cloud* híbrida:** A *Cloud* híbrida oferece uma combinação variada à organização, um exemplo disso é o facto de os dados que necessitam de estar protegidos poderem residir numa *Cloud* privada, enquanto os dados e aplicações públicas podem residir numa *Cloud* pública.

3. Arquitetura de Dados

A arquitetura de dados será explicada recorrendo a duas etapas:

1. arquitetura conceptual – descreve os níveis que constituem a arquitetura e a explicação das atividades que serão realizadas em cada um dos níveis;
2. arquitetura física – descreve uma solução tecnológica, através da instanciação de tecnologias para cada um dos níveis identificados na arquitetura conceptual. Realça-se que poderão existir outras soluções tecnológicas distintas da apresentada.

As fontes de dados (*Data Sources*) representadas na arquitetura conceptual e física procuram representar as diferentes origens e tipos de dados que podem ser utilizados.

3.1 Arquitetura Conceptual

A arquitetura proposta é composta por três níveis, sendo que cada nível suporta um conjunto de atividades a ele associadas. Estas vão desde a recolha dos dados até à disponibilização ao utilizador, dos resultados obtidos pelas análises realizadas aos dados. Os três níveis que constituem a arquitetura são: *Cleaning and Modeling Data, Data Base Big Data e Data Analysis and Visualization*.

Na figura 3, apresenta-se a proposta da arquitetura conceptual da solução com uma descrição das atividades associadas a cada um dos níveis da arquitetura.

Cleaning and Modeling Data

Este nível está responsável pelo processo denominado *Extract Transform and Load* (ETL) que compreende as atividades de extração dos dados de várias fontes, transformação e limpeza dos mesmos de forma a assegurar que posteriormente os dados tratados são carregados para uma área de armazenamento. As diferentes fontes de dados podem ter várias origens e os dados podem ser estruturados, semiestruturados e não estruturados.

As atividades a executar neste nível são:

- Limpar os dados.
- Detetar e corrigir erros.
- Extrair dados para a realização de análises, os dados são selecionados de acordo com as análises a realizar, mas também podem ser realizadas análises *ad-hoc*, carregando desta forma um conjunto de dados aleatórios para serem analisados.
- Armazenamento dos dados numa base de dados *Big Data*, que guarda os dados tratados para mais tarde serem utilizados no *Data Analysis and Visualization*.
- As ferramentas tecnológicas a utilizar devem permitir:

- Importar dados de diversas origens e tipos.
- Efetuar o tratamento dos dados e as devidas correções se as mesmas se verificarem necessárias.
- Permitir a modelação, definindo os dados que são relevantes para análise.

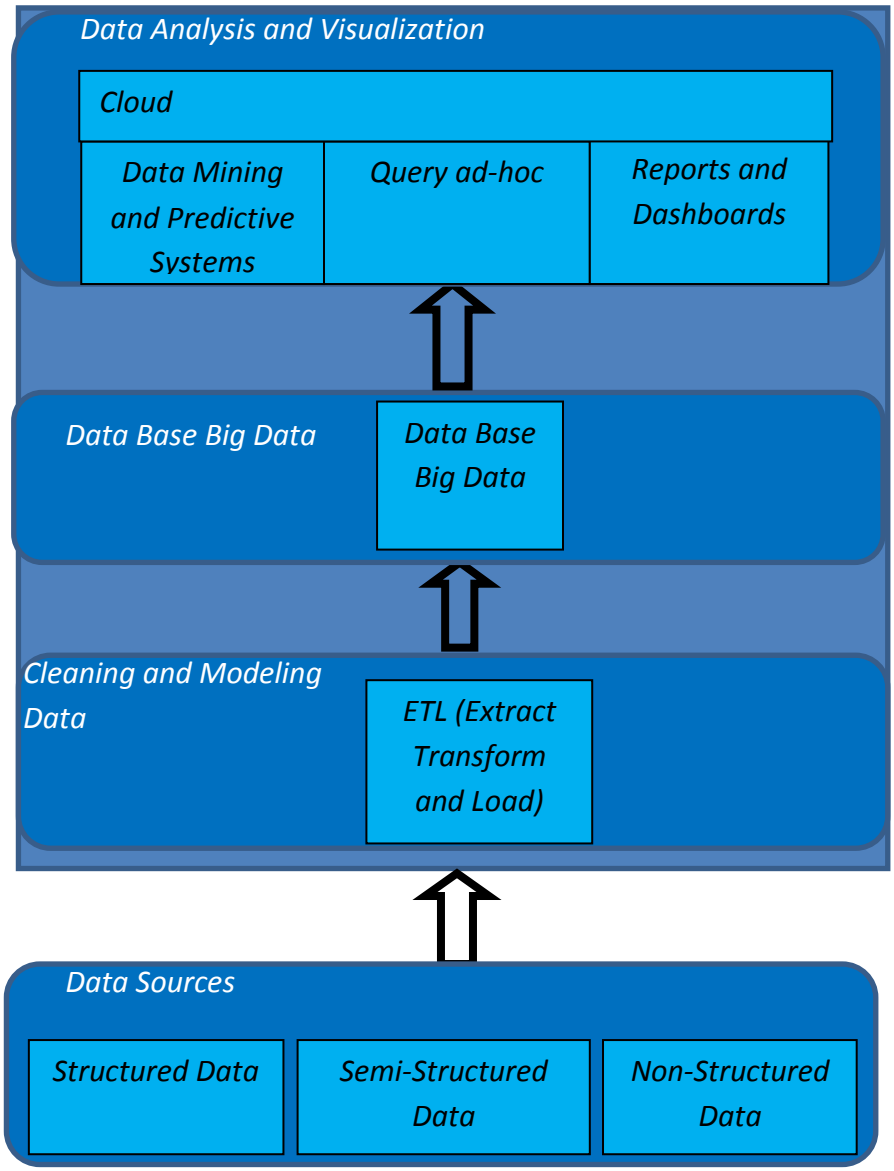


Figura 20 - Arquitetura Conceptual

Data Base Big Data

A *Data Base Big Data* armazena os dados que foram tratados e que serão utilizados para o tratamento analítico no nível de *Data Analysis and Visualization*. A base de dados *Big Data* é o resultado de todo o processo de ETL realizado no *Cleaning and Modeling Data*. Um aspeto importante associado à base de dados *Big Data* prende-se com o facto de a mesma se constituir como um repositório capaz de armazenar diversos tipos de dados.

Data Analysis and Visualization

O nível de *Data Analysis and Visualization*, é responsável pela definição dos indicadores de negócio a analisar, desta forma é possível definir um conjunto de métricas que permitem suportar as decisões que a organização decida tomar. Os resultados obtidos com as análises aos dados são apresentados através da *Cloud (ex:* com a disponibilização dos resultados através da *web*). A pertinência deste componente de visualização de dados está relacionada com a necessidade de um acesso rápido aos dados a partir de um qualquer local a uma qualquer hora, permitindo que esses acesos possam ser feitos por um quaisquer dispositivos com ligação a internet como por exemplo os dispositivos móveis (*smartphones* e *tablets*).

3.2 Arquitetura Física

Para cada um dos níveis da arquitetura são identificas soluções tecnológicas, que permitem a implementação da solução, como pode ser verificado no exemplo de implementação da figura 4. A escolha dessas tecnologias baseou-se principalmente na facilidade de instalação, configuração e utilização das mesmas, uma vez que foram escolhidas tecnologias de fabricantes conceituados e com valor no mercado. Será, ainda, realizada uma descrição das tarefas a efetuar por cada um dos níveis da arquitetura.

Cleaning and Modeling Data

Para o nível de *Cleaning and Modeling Data*, optou-se pela utilização de algumas das ferramentas do Hortonworks. O Hortonworks é uma máquina virtual que possui uma implementação completa de uma solução *Big Data* com todas as ferramentas Hadoop.

O Hortonworks possui um conjunto de ferramentas que são identificadas, com o nome *Data Access* na figura 5, estas ferramentas permitem o tratamento dos dados de diferentes origens e tipos, desta forma com alguma abstração e com ajuda da ferramenta PIG os dados podem ser tratados através de instruções SQL e *Not Only Structured Query Language* (NoSQL). Existe também um outro conjunto de ferramentas como o Apache Storm e o Sorl que permite pesquisas de *streaming* através de redes sociais da web.

Data base Big Data

Neste nível também se optou pela utilização do Hortonworks, na figura 6 está representada toda a arquitetura da ferramenta Hortonworks. Os dados tratados através do processo ETL, realizado no *Cleaning and Modeling Data*, foram armazenados através do sistema de ficheiros *Hadoop Distributed File System* (HDFS), sendo esta uma forma simples de armazenar dados para posterior análises.

O HDFS é um sistema de ficheiros que permite um processamento simultâneo, fornecendo uma gestão de recursos com uma ampla variedade de métodos de acesso a dados e um armazenamento eficiente em termos de custo, escalabilidade e tolerância a falhas.

A base de dados Hive, foi escolhido como o repositório dos dados tratados, devido à capacidade de integrar repositórios de dados de diferentes tipos e origens.

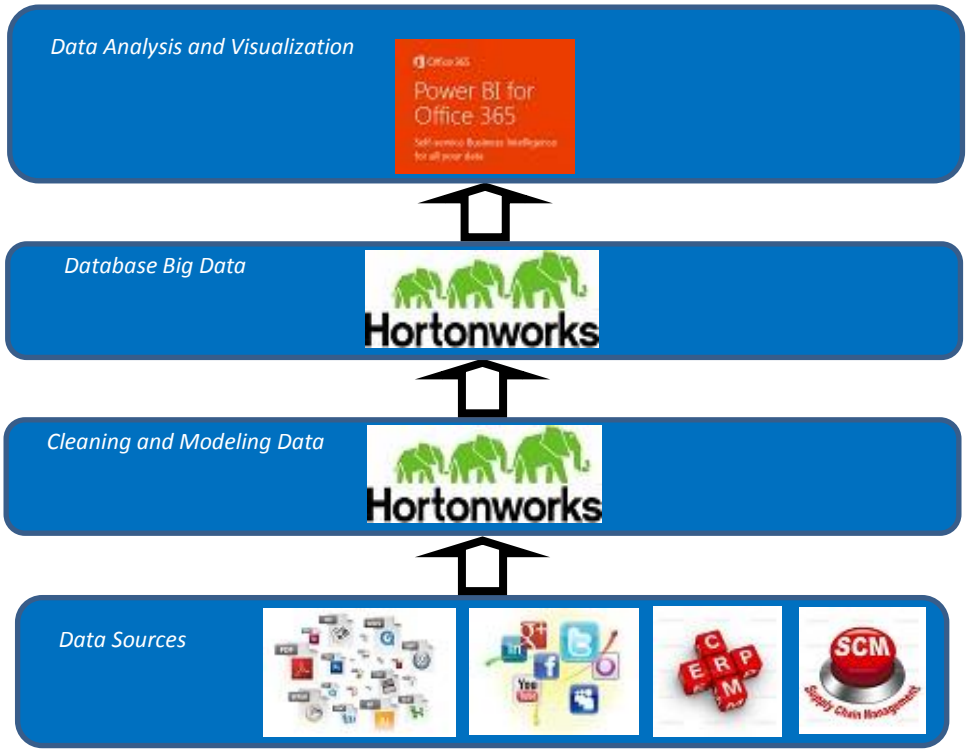


Figura 21 - Arquitetura Tecnológica

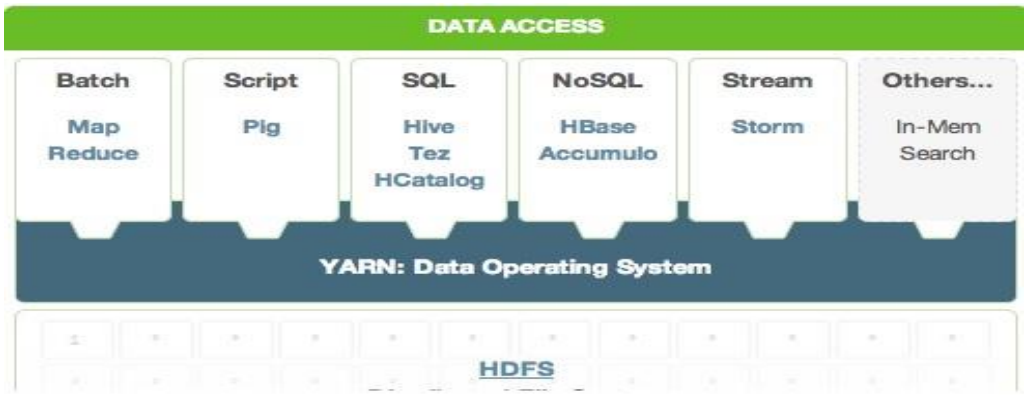


Figura 22 - Componente Data Access do Hortonworks

<table>
<tr><td>Operações</td><td>Fornecimento de Recursos, Gestão e Monitorização:

Ambari e Zookeeper</td><td colspan="2">Agendamento de Tarefas:

Oozie</td></tr>
<tr><td>Segurança</td><td colspan="3">Autenticação, Autorização, Utilizadores e Proteção dos Dados

Armazenamento: HDFS; Recursos: YARN; Acesso: Hive; Pipeline: Falcon, Cluster: Knox</td></tr>
<tr><td rowspan="2">Acesso aos Dados</td><td>Código: Pig

SQL: Hive, Tez, HCatalog

NoSQL: Hbase, Accumulo

Pesquisa: Solr</td><td rowspan="2">HDFS</td><td rowspan="2">Gestão de Dados</td></tr>
<tr><td>YARN: Sistema de operações sobre dados</td></tr>
<tr><td>Governança da Integração</td><td colspan="3">Fluxos de Dados e Governança

Falcon, WebHDFS, NFS, Flume, Sqoop</td></tr>
</table>

Figura 23 - Arquitetura Hortonworks Sandbox versão 2.1

Data Analysis and Visualization

Este é o último nível da arquitetura e é responsável por realizar as análises aos dados e disponibilizar os resultados na *Cloud*. Os dados a analisar são acedidos via ligação *Open Data Base Connectivity* (ODBC) ao Hortonworks, em seguida os dados são importando para o Microsoft Office Excel para depois serem analisados recorrendo a ferramenta Microsoft Power BI. Existe um conjunto de análises que podem ser realizadas como:

- *Data Mining and Predictive Analysis*, são análises que utilizam algoritmos de *Data Mining* para a previsão de determinados acontecimentos.
- *Query ad-hoc*, são análises realizadas através de *queries* não estruturadas na base de dados. Para realizar este tipo de *queries* vamos utilizar o Microsoft Power BI que permite através de uma ligação fornecendo o IP(Internet Protocol) do servidor onde os dados estão armazenados importar dados no formato HDFS para *Excel,* também podemos recorrer a uma ligação ODBC para importar os dados para o *Excel*. A visualização dos resultados pode ser efetuada de forma interativa através das ferramentas de Power BI, compostas pelas Power Map, Power Pivot e Power Query.
- *Reports* e *dashboards*, para a realização de cubos OLAP e apresentação dos resultados produzidos através dos mesmos, optou-se pela utilização na ferramentas Power View presente na Microsoft Power BI que possibilita uma apresentação agradável dos resultados através de gráficos interativos proporcionando uma fácil navegação nos dados associados a uma determinada análise.

4. Caso de teste

A arquitetura foi implementada recorrendo a plataforma Azure com recurso a uma máquina virtual (Hortonworks Sandbox versão 2.1) e à ferramenta de análise de dados Microsoft Power BI.

Para a realização do caso de teste foi selecionado um *dataset* da área da saúde obtido através da plataforma *Open Data* (https://health.data.ny.gov/). Esses dados correspondem a quatro tipos de cirurgias ocorridas durante os anos de 2010 a 2013 e com origem em hospitais de Nova Iorque, Estados Unidos da América.

4.1 Implementação

O primeiro passo realizado foi a recolha dos dados, de seguida procedeu-se à sua importação para o Hortonworks. No Hortonworks, através das ferramentas HCatalog e Pig, efetuou-se a validação dos dados, ou seja, eliminaram-se valores nulos e duplicados, realizou-se ainda a escolha de dados relevantes para análise através da definição de indicadores de negócio que serão alvo de análise, após este trabalho, procedeu-se ao carregamento dos dados já tratados para um repositório de dados HDFS, sendo escolhido o Hive como repositório de destino para armazenar os dados anteriormente tratados. O Hive permite a criação de um *dataset* agregado com informação de diversos tipos e formatos, trata-se de um *Data Warehouse* que permite agregar dados estruturados, semiestruturados e não estruturados, num só repositório.

De forma a obter o maior partido da informação que se encontrava nos comentários médicos foi necessário extrair essa informação de um ficheiro em formato XML, após este processo os dados não estruturados foram tratados e importados para o Hortonworks, onde através de instruções SQL os dados dos comentários foram adicionados aos restantes dados.

Em seguida, após adicionar a informação dos comentários aos restantes dados, foi definida uma estratégia de análise de dados de forma a não serem efetuadas análises *ad-hoc*, foram definidas 3 questões que seriam alvo de resposta, através das análises aos dados, as questões foram: "**Quais os hospitais com o maior número de intervenções cirúrgicas em Nova Iorque?**, **Qual a evolução no número de mortes nas principais cirurgias realizadas nos hospitais de Nova Iorque?**, **Qual a influência dos comentários para a melhoria dos serviços hospitalares?**".

Depois de carregados os dados anteriormente tratados seguimos para a fase de análise e visualização de resultados, para executar a tarefa de análise dos dados, em primeiro lugar foi necessário estabelecer uma ligação ODBC, esta conexão permitiu o acesso aos dados através das ferramentas Microsoft Excel 2013 com a solução Microsoft Power BI, uma vez estabelecida a conexão, os dados foram importados e analisados, sendo construídos *dashboards* e relatórios com o resultado das análises efetuadas, os mesmos podem ser visualizados nas figuras 6, 7 e 8

Este trabalho foi realizado na plataforma *Cloud* Azure da Microsoft, a opção pela utilização desta plataforma permitiu que todo o trabalho fosse executado em qualquer local e com um acesso constante, esta é uma das grandes vantagens da utilização de recursos em *Cloud*. A escolha do ambiente *Cloud* teve como critério a capacidade de implementar níveis de

segurança e acesso aos dados, desta forma foi definido um conjunto de utilizadores e permissões de acesso no qual cada utilizador possuía uma credencial (nome utilizador e palavra chave) para aceder ao sistema, bem como um conjunto de permissões para efetuar tarefas como: aceder, visualizar ou analisar os dados.

Todo o trabalho realizado também poderia ter sido feito em grande parte através de uma máquina local, desta forma todo o trabalho ETL e armazenamento de dados poderia estar numa máquina local (servidor ou computador pessoal), sendo a disponibilização da informação feita através da *Cloud* recorrendo as ferramentas do Azure apenas para a disponibilização dos resultados.

Por fim uma nota importante é o facto de a base de dados Hive permitir encriptação dos dados, mas não permitir que os dados encriptados possam ser visualizados pelo *Database Administrator*, esta opção garantiria uma maior confiança na utilização do ambiente *Cloud Computing.*

4.2 Exemplos de Análise Realizadas

Para comprovar a viabilidade da arquitetura anteriormente descrita, são apresentados alguns exemplos de análises que foram efetuadas aos dados de teste, estas análises pretendem comprovar as potencialidades arquitetura anteriormente referida.

A primeira análise realizada, que se encontra representada na figura 6, mostra localização dos hospitais onde existiu o maior número de intervenções cirúrgicas, isto permite obter uma perceção das zonas de Nova Iorque, mais procuradas pelos pacientes.

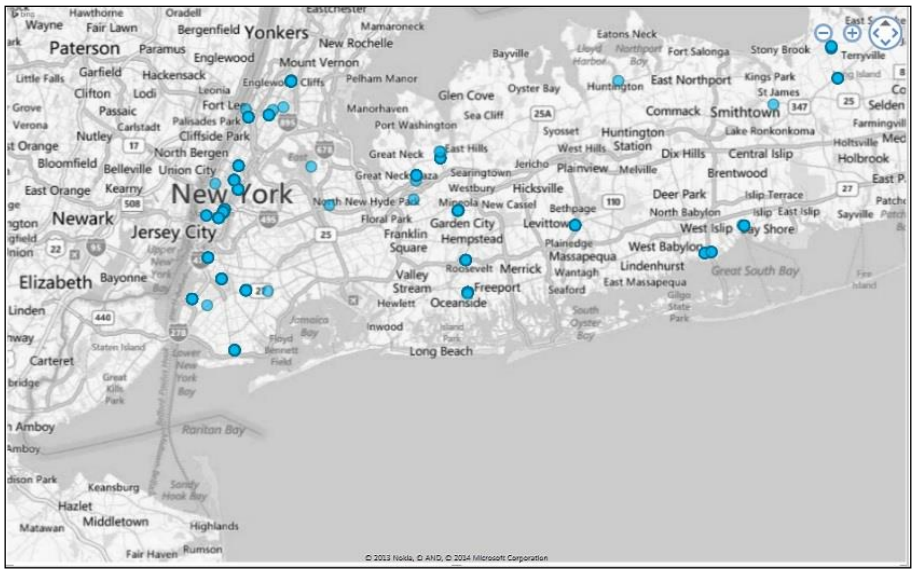


Figura 6- Localização dos hospitais utilizados na análise

Após percebemos as zonas da cidade de Nova Iorque mais procuradas pelo pacientes. Decidimos analisar tendo em conta os vários tipos de cirurgias que eram realizadas aos pacientes, esta análise permitiu-nos perceber que o número de mortes tem diminuído, esta informação pode ser visualizada na figura 7. Podemos observar que a diminuição do número de mortes pode dever-se à influência dos comentários realizados pelos pacientes que

recorrem aos serviços dos hospitais da zona de Nova Iorque. Através desses comentários, os profissionais e o corpo clínico do hospital efetuaram um esforço com vista a melhoria dos serviços, foram tomadas medidas com o objetivo de melhorar os procedimentos clínicos, o que ajudou a incrementar a eficácia e eficiência na realização de quatro tipos de cirurgias: *ALLPCI* (cirurgia que consiste na retirada da tiroide), *CABG* (cirurgia que consiste no desvio da artéria coronária), *NON EMERGENCY PCI* (cirurgia que retira a tiroide não urgente) e *Valve or Valve*/CABG (cirurgia da válvula aórtica/cirurgia ponte da artéria coronária).

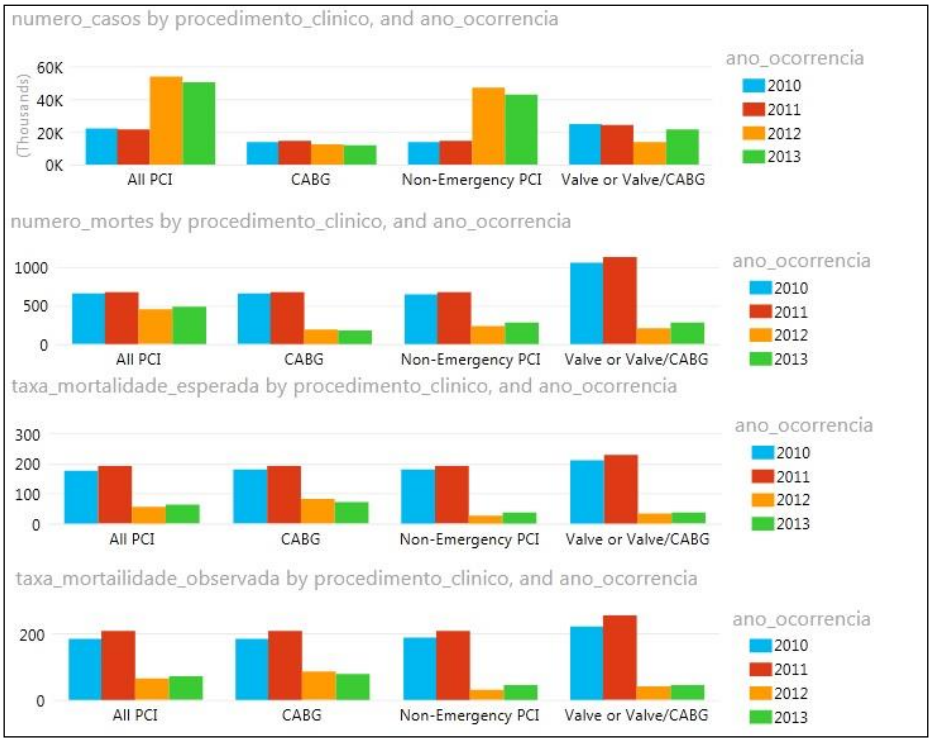


Figura 7 - Análise de Dados Por Ano e Por Procedimento Hospitalar

Também foi possível verificar a influência dos comentários dos pacientes ajudaram na melhoria dos serviços, como podemos observar nos diferentes gráficos representados na figura 8. Os comentários permitem que os hospitais diminuíssem o número de mortes e aumentassem a eficiência dos seus serviços melhorando a eficácia na resolução dos casos clínicos.

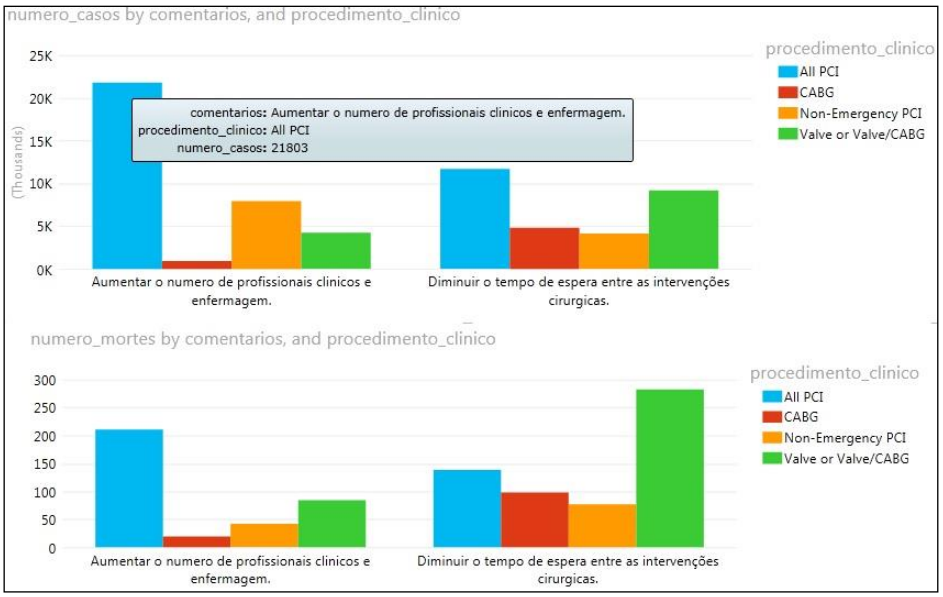


Figura 8 - Influência dos Comentários dos Pacientes na Melhoria dos Serviços Médicos

4.3 Avaliação de Resultados

Os resultados obtidos através das análises realizadas permitem uma conjugação de informação com origem em dados estruturados e dados não estruturados (comentários), de facto não é uma tarefa fácil relacionar estes tipos de dados. Este caso de teste serve também para mostrar que ainda existe muito trabalho a realizar nesta área, pois para uma correta utilização de dados não estruturados é muito importante perceber em primeiro lugar o que esperamos obter com análise dos dados e em segundo perceber o valor que esses dados não estruturados podem ter e em que medida podem ajudar o processo de decisão de uma organização. Um aspeto que pude constatar neste caso é que a informação que se encontrava nos dados não estruturados era um complemento a informação dos dados estruturados, isto é importante pois ajuda a ter mais confiança na hora de tomar decisões pois as bases que sustentam a mesmas são sólidas.

Por fim é importante referir que a adequação da arquitetura física aos resultados está dependente de muitos fatores, mas é importante destacar dois destes fatores:

- os resultados dependem muito da riqueza dos dados, ou seja, se os dados não forem os mais adequados ou não tiverem grande riqueza os resultados não serão os melhores;
- a importância do tratamento dos dados, pois um adequado tratamento dos dados efetuados no nível de *Cleaning and Modeling Data* da arquitetura física bem como a garantia da integridade dos dados, contribui para a obtenção de bons resultados após as análises aos dados.

5. Conclusões

Uma arquitetura moderna de dados deve permitir um armazenamento de dados diversificado, permitindo armazenar dados estruturados, semiestruturados e não estruturados, da mesma forma devem existir mecanismos que permitam a interação com esses mesmos dados, nomeadamente a utilização de SQL e NoSQL. A organização e distribuição dos dados deve ser feita recorrendo a sistemas de armazenamento de dados que permitam um armazenamento de dados diversificados como é o caso do HDFS e do Hive, algumas características muito importantes que estes sistema de armazenamento devem possuir são o rápido acesso aos dados, garantia da consistência e integridade dos dados através de mecanismos de segurança que permitam um acesso aos dados apenas as pessoas que possuam autorização para manipular esses mesmos dados.

A utilização de um ambiente como o Hortonworks em conjugação com a utilização de ferramentas Hadoop, permite simplificar todo o processo de tratamento e acesso aos dados para posteriores análises.

O uso da *Cloud* para a disponibilização de dados, obriga a definir um conjunto de restrições e políticas de permissão de acesso aos dados por parte dos utilizadores. Para que o acesso aos dados seja efetuado por quem está autorizado e deles necessita, procurando sempre garantir a sua integridade, tudo isto pode ser implementado através de mecanismo de gestão de utilizadores, que se encontram implementados nas plataformas de infraestruturas *online*, como é o caso do Microsoft Azure, Amazon AWS, OpenStack entre outros. Estas plataformas recorrem a utilização de protocolos de segurança como *Secure Sockets Layer* (SSL) e *Hyper Text Transfer Protocol Secure* (HTTPS).

Uma correta implementação de políticas de gestão de acesso permite aos utilizadores confiar na utilização de soluções em *Cloud*, sobretudo em soluções de análises de dados devido a importância que os dados possuem na vida e nas decisões tomadas organizações. É importante dar ênfase a uma característica que os sistemas deveriam adotar, que é a insistência de um mecanismo que possibilite à base de dados ter níveis de segurança ao nível da encriptação dos dados. Isto faria com que o acesso aos dados apenas seria permitido a quem tivesse autorização, através da utilização de uma chave de desencriptação, evitando assim que o *Database Administrator* (DBA) ou quem desenvolve a solução possa aceder aos dados.

A utilização de dados *Big Data* num ambiente *Cloud Computing* obriga à uma reformulação e um pensamento mais elaborado das arquiteturas de dados, colocando em evidência as seguintes características: robustez, rápida disponibilidade de dados para serem analisados, permitir a recolha e tratamento de dados de diversos tipos e origens como dados estruturados, semiestruturados e não estruturados, salvaguardar sempre os aspetos de segurança e integridade dos dados.

No entanto considera-se que será necessário, efetuar mais casos de teste, para validar a capacidade da arquitetura. Esses testes devem incluir grandes volumes de dados, nomeadamente dados do tipo não estruturados e semiestruturados. As análises também podem ser melhoradas utilizando por exemplo técnicas de *text mining.*

Referências

Böhm, M., Leimeister, S., Riedl, C., & Krcmar, H. (2011). Cloud Computing–Outsourcing 2.0 or a new Business Model for IT Provisioning? Application Management.

Francis, L. (2009). Cloud Computing: Implications for Enterprise Software Vendors (ESV), System Design and Management Program. *Massachusetts Institute of Technology*.

Guolinag, L., Beng, C. O., Jianhua, F., Jianyoung, W., & Lizhu, Z. (2008). EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data. *SIGMOD '08 Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 903-914.

Halper, F., & Krishnan, K. (2013). TDWI Big Data Maturity Model Guide Interpreting Your Assessment Score. *TDWI Benchmark Guide 2013–2014*.

Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Defi nitive Guide to Dimensional Modeling, Third Edition. *Wiley: The Data Warehouse Toolkit: The Defi nitive Guide to Dimensional Modeling, Third Edition*.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META Group Research Note, 6.

Lohr, S. (2012). The Age of Big Data. *The New York Times*.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (May de 2011). Big data: The next frontier for innovation, competition, and productivity.

Olivier, B., Thomas, B., Heinz , D., Hanspeter , C., Babak , F., Markus , F., . . . Markus , Z. (6 of November, 2012). White Paper Cloud Computing. *Swiss Academy of Engineering Sciences*.

Oswaldo, T., Pjotr, P., Marc, S., & Ritsert, C. J. (March, 2010). Big data, but are we ready. pp. 647-657.

Russom, P. (2013). Maning Big Data. *TDWI Best Practices Report Fourth Qaurter 2013*.

Stonebraker, M. (21 of September, 2012). What Does 'Big Data' Mean? *What Does 'Big Data' Mean? | blog@CACM | Communications of the ACM*.

Stuckenberg, S., Fielt, E., & Loser, T. (2011). The Impact Of Software-As-A-Service On Business Models Of Leading Software Vendors. *Experiences From Three Exploratory Case Studies. Proceedings of the 15th Pacific Asia Conference on Information Systems (PACIS 2011).*

Youseff, L., Butrico, M., & Da Silva, D. (2008). Toward a Unified Ontology of Cloud Computing. Grid. *GCE'08*.