Uma Arquitetura Moderna de Dados: Um Caso de Teste*

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 e implementar uma solução que será validade através de um caso de teste com dados da área da saúde.

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]:

- 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;

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

- 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 *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, por exemplo, 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.

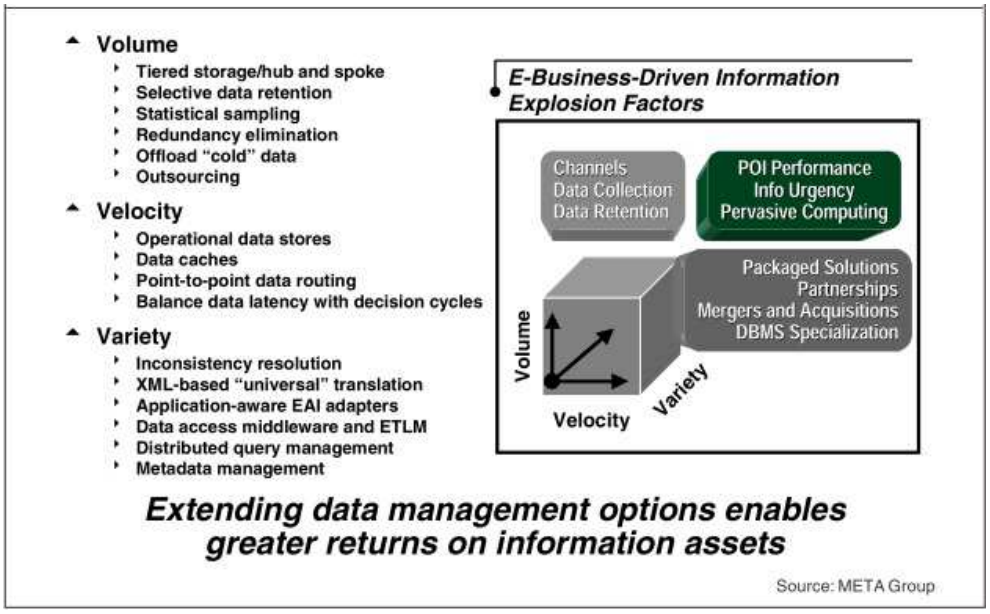

Figura 1 – 3D Data Management Controlling [Laney, 2001]

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 efetuadas 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 atual, moderna que suporte *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 ER. 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].

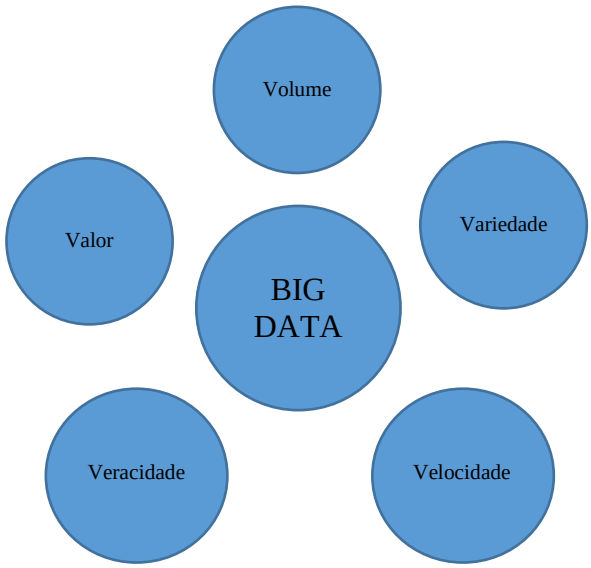

Figura 2 - Estrutura do Big Data com os 5V's adaptado de [Stonebraker, 2012]

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 economias do lado cliente, na medida em que o fornecedor de serviços tem a responsabilidade de garantir serviços e o suporte das infraestruturas que os suportam, isto permite que o cliente se foque naquilo que é objetivo do seu negócio e assim melhore a sua produtividade, por outro lado, garante que o cliente tem uma constante adequação do serviço às suas necessidades, muitas das vezes através de uma redução do custo dos serviços e infraestruturas.
- O fornecedor de serviços de *Cloud* procura garantir uma melhoria contínua dos seus serviços através da melhor adequação do *hardware* e *software* às necessidades exigidas pelos seus clientes. É também exigido que o fornecedor de serviços disponibilize um conjunto de políticas de segurança que salvaguardem a informação relevante 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.

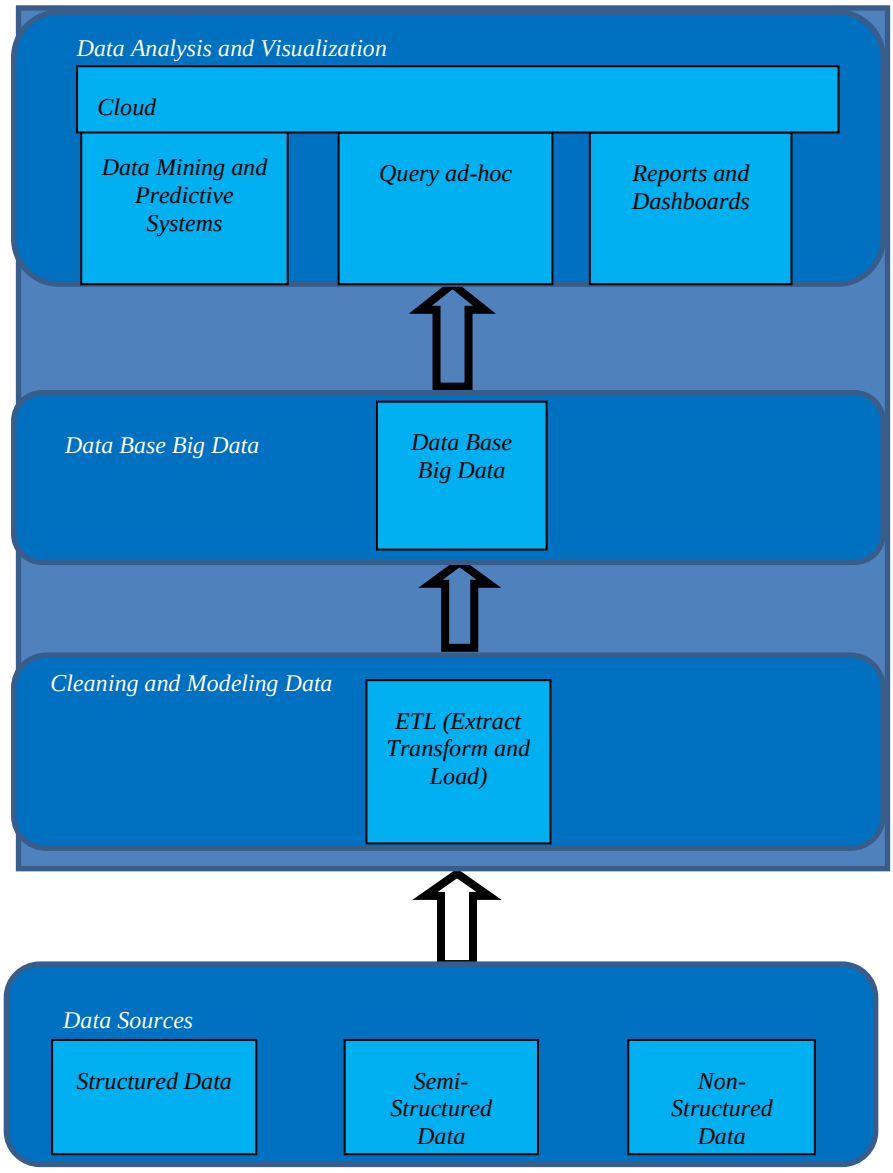

Figura 3 - Arquitetura Conceptual

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.

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.

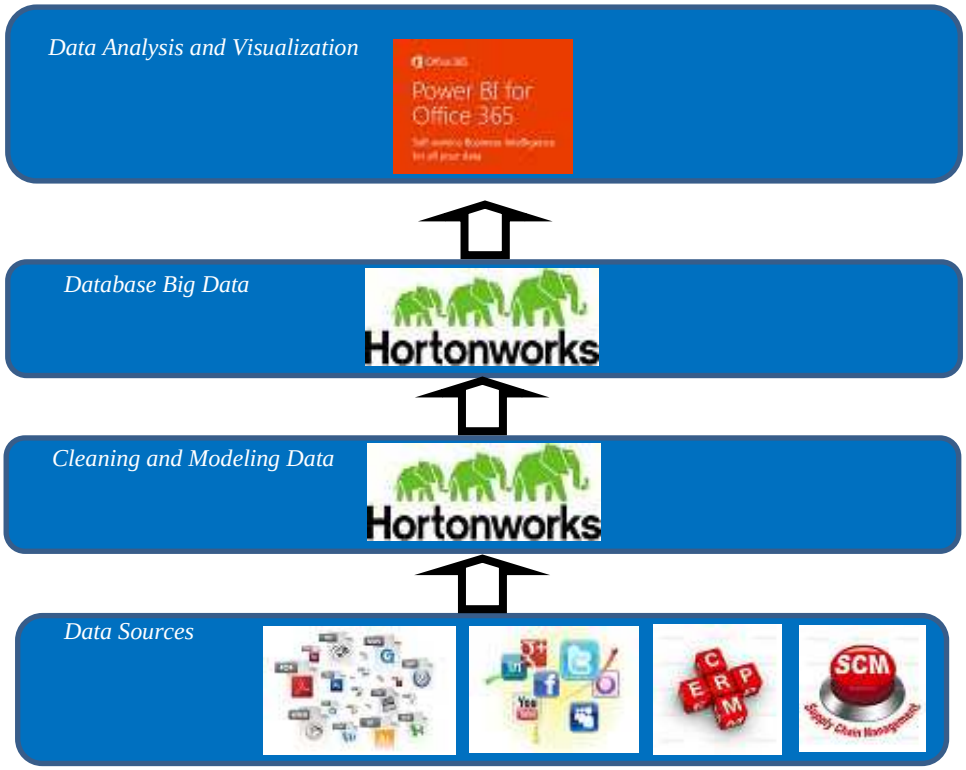

Figura 4 - Arquitetura Tecnológica

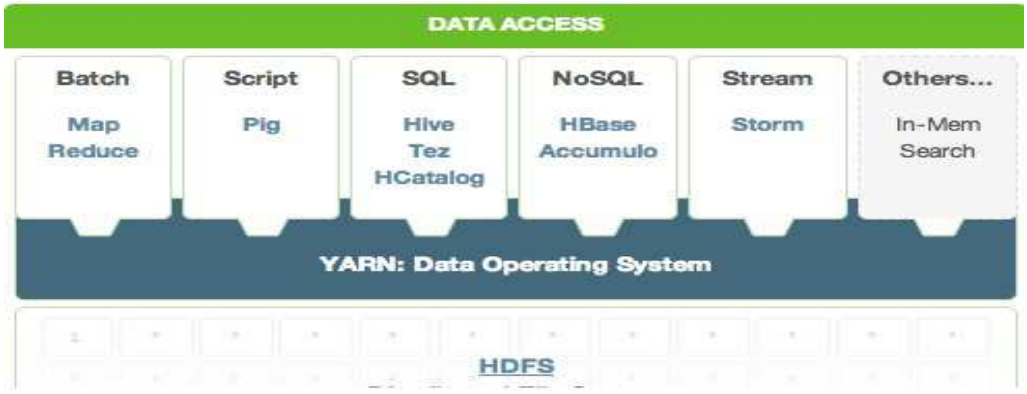

Figura 5 - Componente Data Access do Hortonworks

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 escolhida como o repositório dos dados tratados, devido à capacidade de integrar repositórios de dados de diferentes tipos e origens.

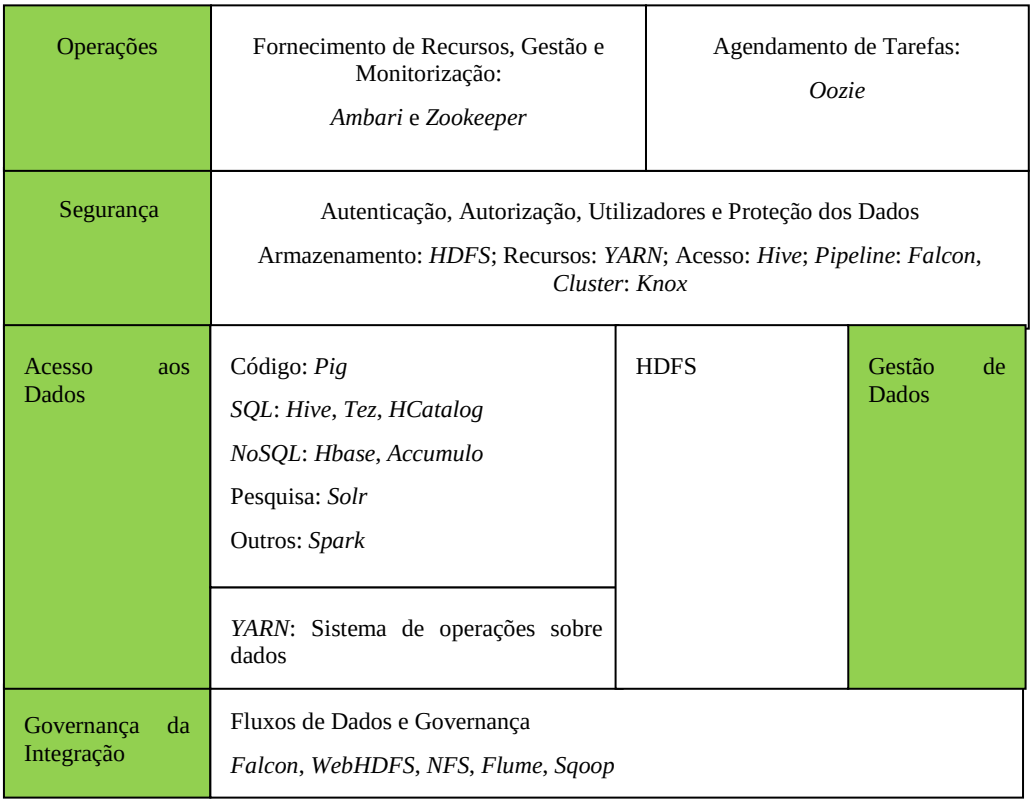

Figura 6 - Arquitetura Hortonworks Sandbox versão 2.1

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 das 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.

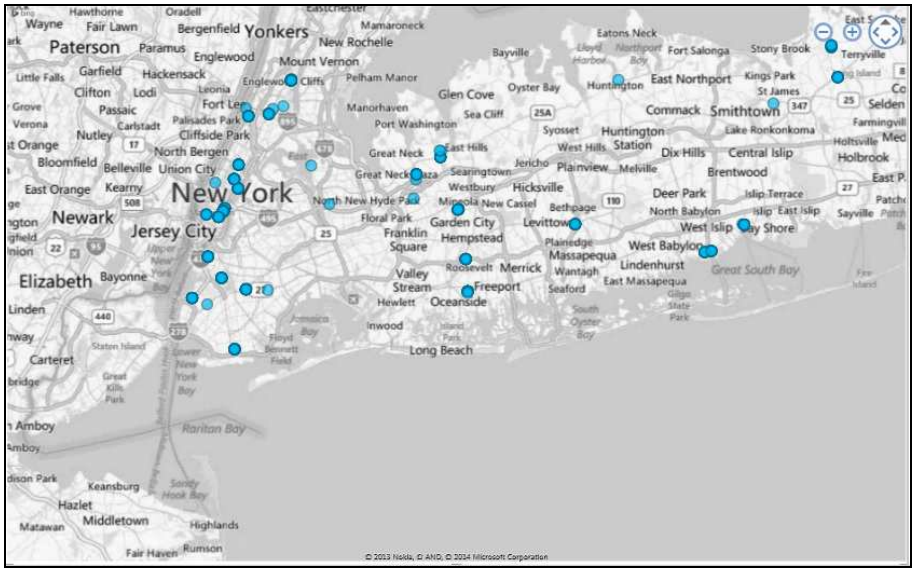

Figura 6- Localização dos hospitais utilizados na análise

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).

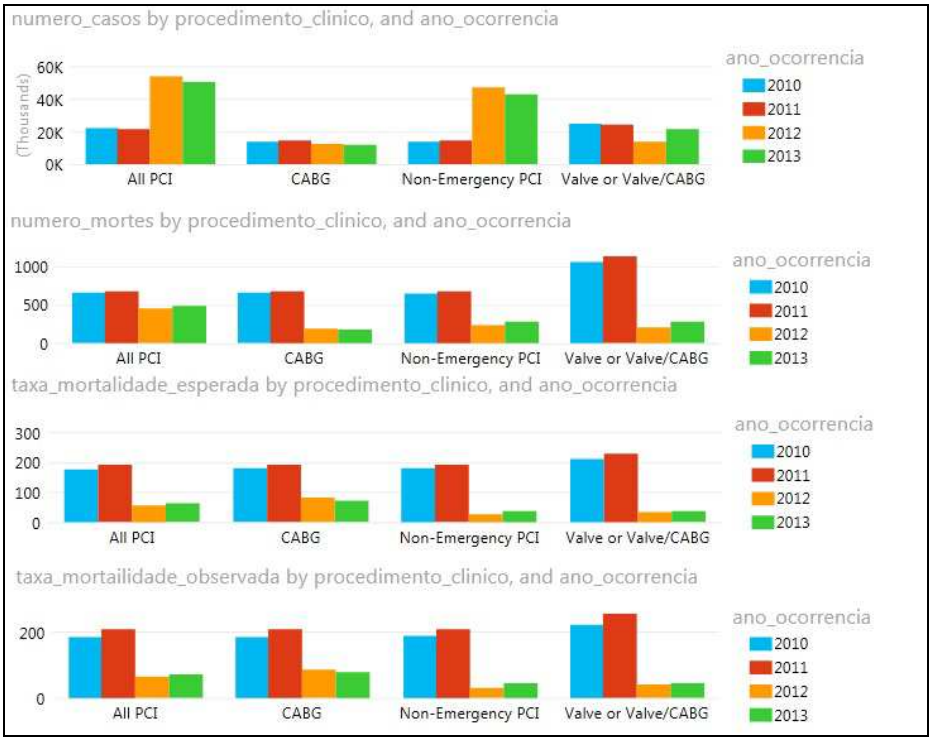

Figura 7 - Análise de Dados por Ano e por Procedimento Hospitalar

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.

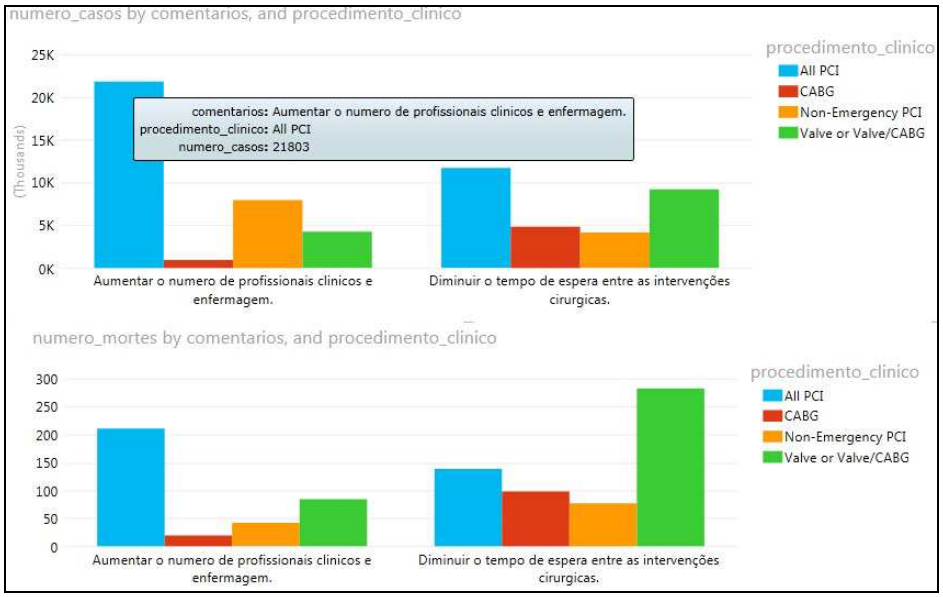

Figura 8 - Influência dos Comentários dos Pacientes na Melhoria dos Serviços Médicos

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. 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 a 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. Pode-se constatar que, neste caso, a informação que se encontrava nos dados não estruturados era um complemento da 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:

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, 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 pelos utilizadores. É importante dar ênfase a uma característica que os sistemas deveriam adotar, que consiste num mecanismo que possibilite à base de dados ter níveis de segurança ao nível da encriptação dos dados, fazendo com que o acesso aos dados apenas seja permitido a quem tiver autorização

através da utilização de uma chave de desencriptação, evitando assim que até o *Database Administrator* (DBA) ou quem desenvolve a solução possa ter acesso a determinados 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, 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*, 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*.