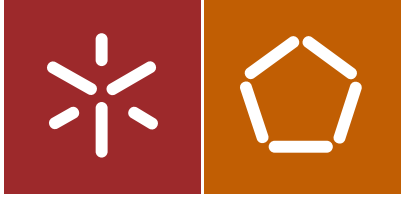
Universidade do Minho
Escola de Engenharia

Cátia Sofia Vaz Crasto de Menezes

Big City Data:
Uma nova dimensão sobre as cidades

outubro de 2014

Big City Data:
Uma nova dimensão sobre as cidades
Cátia Sofia Vaz Crasto de Menezes
UMinho | 2014



Universidade do Minho
Escola de Engenharia

Cátia Sofia Vaz Crasto de Menezes

Big City Data:
Uma nova dimensão sobre as cidades

Dissertação de Mestrado
Ciclo de Estudos Integrados Conducentes ao Grau de
Mestre em Engenharia e Gestão de Sistemas de Informação

Trabalho efetuado sob a orientação da
Professora Doutora Maribel Yasmina Santos

outubro de 2014

DECLARAÇÃO

Nome Cátia Sofia Vaz Crasto de Menezes

Endereço electrónico: catiamenezes.1991@gmail.com Telefone: 915569748

Número do Bilhete de Identidade: 13903375

Título dissertação :

Big City Data – Uma nova dimensão sobre as cidades

Orientadora:

Professora Doutora Maribel Yasmina Santos

Ano de conclusão: 2014

Ciclo de Estudos Integrados Conducentes ao Grau de Mestre em Engenharia e Gestão de Sistemas de Informação

Nos exemplares das teses de doutoramento ou de mestrado ou de outros trabalhos entregues para prestação de provas públicas nas universidades ou outros estabelecimentos de ensino, e dos quais é obrigatoriamente enviado um exemplar para depósito legal na Biblioteca Nacional e, pelo menos outro para a biblioteca da universidade respectiva, deve constar uma das seguintes declarações:

É AUTORIZADA A REPRODUÇÃO INTEGRAL DESTA TESE/TRABALHO APENAS PARA EFEITOS DE INVESTIGAÇÃO, MEDIANTE DECLARAÇÃO ESCRITA DO INTERESSADO, QUE A TAL SE COMPROMETE;

Universidade do Minho, ___/___/_____

Assinatura: _____

"Porque eu sou do tamanho do que vejo e não do tamanho da minha altura!" (Alberto Caeiro)

Agradecimentos

Agradeço a todas as dificuldades que enfrentei, pois foram adversárias dignas e tornaram as minhas vitórias muito mais agradáveis.

Estas vitórias nunca tinham sido conquistadas sem todo o apoio de pessoas importantes que sempre estiveram cá para me lembrar que era capaz. Só me resta agradecer a estas pessoas por todos os momentos proporcionados.

À minha orientadora, Professora Doutora Maribel Yasmina Santos, pela sua disponibilidade, pela sua paciência, pela sua dedicação e por acreditar que eu era capaz de cumprir com os objetivos desta dissertação de mestrado.

Ao CEIIA pela oportunidade de realizar esta dissertação de mestrado no seu ambiente, nomeadamente, ao Luís Cunha pela disponibilidade que demonstrou em todo o decorrer do projeto, pela sua paciência e por fazer sempre com que pensasse mais à frente, com novas perspetivas e desafios. Agradeço também a toda a equipa do S3C pelo apoio que prestaram quando era possível, pela integração na equipa e pelo ótimo ambiente de trabalho.

Aos meus pais que sempre lutaram na vida para tornar possível esta realidade e por permitirem que tivesse uma vida cheia de oportunidades e realizações.

Às minhas irmãs, aos meus cunhados e às minhas sobrinhas por sempre me lembrarem de como era importante seguir este caminho.

Ao Lau, por estar sempre do meu lado, por ser sempre o meu apoio incondicional, por me corrigir quando estava errada, por me fazer sempre acreditar que era possível chegar onde cheguei e principalmente pela paciência e motivação quando os momentos menos bons apareciam.

À Lu, a amiga de todas as horas, o meu ombro amigo, o meu alicerce e, principalmente, a pessoa que tornou todos estes anos importantes, marcando cada um de uma forma muito especial.

À Trice, pela sua paciência, pela sua prontidão em qualquer momento, pela disponibilidade em me ouvir e pela sua presença em todos os marcos importantes deste percurso.

Ao Zé Miguel, à Céline e ao Toni pela amizade, pela boa disposição, pela vida académica fantástica que sempre proporcionaram e principalmente pela motivação que sempre deram.

Obrigado!

Resumo

Atualmente as organizações deparam-se com um enorme crescimento da quantidade de dados que recolhem e armazenam, os quais são extremamente importantes para o sucesso do negócio no que diz respeito à análise de indicadores. Ao mesmo tempo, estes dados surgem por vezes de forma semi-estruturada ou não estruturada, com diversos tipos e formatos. Este facto leva as organizações a sentirem dificuldade ao nível do armazenamento e processamento desses dados.

Surge aqui uma nova era denominada de *Big Data*. Este novo conceito caracteriza-se pela sua capacidade de armazenamento de grandes quantidades de dados, bem como o seu processamento.

Paralelamente a este tema aparece o conceito de *Smart Cities.* Estas cidades utilizam as tecnologias para disponibilização de informação útil para os seus cidadãos e ao mesmo tempo impulsionar a participação nas suas atividades diárias com a partilha de informação. Desta forma é gerada uma densa quantidade de dados e, por conseguinte, existe a necessidade da utilização das características que o *Big Data* oferece.

O objetivo desta dissertação de mestrado é desenvolver um protótipo de um sistema Big Data que apoie as atividades diárias do CEIIA - Centro de excelência e inovação para Indústria Automóvel, no que diz respeito à recolha, processamento, armazenamento, análise e visualização de dados recolhidos das redes sociais, entre outros dispositivos que possam captar informação urbana existente. Por este motivo, o CEIIA sentiu a necessidade de adotar um sistema que suportasse todos estes dados de forma a poder ter uma melhor perspetiva das cidades e dos cidadãos, melhorando o seu serviço ao cliente. A construção e desenvolvimento deste sistema permite novas formas de análise e visualização dos dados de uma forma mais abrangente e com novas perspetivas de negócio.

Abstract

Nowadays, organizations face itself with an increasing amount of data that collect and store, which are extremely important for the business success regarding to indicators analysis. At the same time this data emerge in semi-structured or non-unstructured forms with several types and formats. This fact leads the organizations to feel difficulties with the processing and storage of this data.

Here rises a new era named *Big Data*. This new concept is characterized by its huge capacity of storage and process a big amount of data.

Along with this theme appears the concept of *Smart Cities.* These cities use technologies to provide useful information for its citizens and, at the same time, boost its participation in daily activities, with the share of information. So it is generated a dense amount of data and therefore exist the necessity of use Big Data characteristics.

The purpose of this master thesis is to develop a prototype of a Big Data system that could support daily activities on CEIIA – Centre for Excellency and Innovation in Mobility Industry, with regard to social networks, between other devices that could gather the existing urban information. For that reason, CEIIA felt the need to adopt a system that could support all the data in order to have a better perspective about the cities and its citizens, improving its services to the client. The design and development of this system allow new ways of data analysis and visualization opening new business perspectives.

Índice

Índice de figuras

Índice de tabelas

Lista de acrônimos e abreviaturas

API – *Application Programmer Interface*

BI – *Business Intelligence*

CEIIA – Centro de excelência e inovação para Indústria Automóvel

DSR – *Design Science Research*

ETL - *Extract, Transform, Load*

HDFS – *Hadoop File System*

JSON - *JavaScript Object Notation*

NoSQL – Not Only SQL

SDC – *Smart Devices Cloud*

SQL - *Structured Query Language*

SSC - *Smart Service Cloud*

TI – Tecnologias de Informação

Capítulo 1 - Introdução

Neste capítulo é introduzido o tema a desenvolver descrevendo os objetivos, a finalidade e a motivação para a realização do mesmo. Numa segunda parte é descrita a metodologia seguida e a estrutura que o documento apresenta.

1.1. Enquadramento

Um dos grandes desafios existentes para as organizações hoje em dia é o armazenamento e tratamento dos dados que são gerados constantemente nas suas atividades e que têm sofrido um forte crescimento.

O crescimento massivo dos dados deve-se a diversos fatores dos quais Krishnan (2013) salienta a inovação, a transformação do modelo de negócio, a globalização e a personalização dos serviços.

A inovação transformou a forma como a organização se envolve nos negócios, na prestação de serviços e na medição do valor e rentabilidade associados.

A globalização levou a uma mudança drástica do comércio mundial desde a fabricação até ao atendimento ao cliente, bem como a variedade e o formato dos dados.

Devido à globalização houve necessidade de reestruturar o modelo de negócio seguido pelas organizações passando-se de uma orientação para o produto para a orientação ao serviço, onde o valor para a organização é medido pela eficácia do serviço prestado ao cliente e não pela utilidade do produto. Desta forma, existe uma maior necessidade de produzir mais dados relativamente aos produtos e serviços para atender às necessidades de cada grupo de clientes.

De forma a satisfazer as necessidades dos clientes e, passando a existir uma orientação ao serviço, é indispensável a personalização da oferta e o valor percebido por esses clientes de forma a medir a maturidade da transformação realizada.

Para além destes fatores, Krishnan (2013) refere que existem novas fontes de dados geradas pelas redes sociais, dispositivos móveis e sensores. Além disto, as organizações usufruíam de informação dispersa que analisada individualmente não acrescentava valor, mas sob uma perspetiva contextual, estes dados agora conjugados oferecem perspetivas importantes para o processo de tomada de decisão.

O aparecimento de novos modelos de negócio e o agressivo crescimento da capacidade das tecnologias abriu caminho para a integração de todos os dados da organização numa plataforma global, que originou a criação de uma plataforma de suporte a processos de tomada de decisão significativas e contextualizadas.

O facto é que todas estas tendências têm adicionado complexidade aos processos organizacionais e, ao mesmo tempo, criou a necessidade de aquisição de dados necessários para obtenção de uma visão mais abrangente sobre áreas que até então não eram passíveis de análise.

Podemos assim dizer que as organizações têm acesso a informação com grande riqueza mas não sabem tirar valor dessa mesma informação devido a sua quantidade e a falta de estruturação (Zikopoulos, Eaton, & Zikopoulos, 2011).

Toda esta problemática leva a uma nova era denominada *Big Data*. Este conceito nasce pela sua capacidade de acesso a grandes volumes de dados que podem ser úteis para visualização e análise de padrões únicos ou repetidos relativamente a comportamentos. É uma infraestrutura prática combinada com novas *frameworks* e plataformas de processamento (Hadoop e NoSQL), de baixo custo e alta escalabilidade (Krishnan, 2013).

Associado a este tema surge um novo conceito com grande importância, as *Smart Cities*. Estas cidades estão destinadas a tornar-se uma das ferramentas públicas mais poderosas nos próximos anos. A integração das Tecnologias de Informação e Comunicação na evolução de uma cidade, além das melhorias na prestação de serviços, constitui também um desenvolvimento económico e social sustentável (Telefónica, 2011).

As *Smart Cities* geram um conjunto de dados de larga escala de forma estruturada e não-estruturada, com necessidades de um processamento e armazenamento em tempo real e de forma eficiente.

O CEIIA – Centro de Excelência e Inovação para a Indústria Automóvel, é um centro de pesquisa e inovação, com desenvolvimento de produtos e processos para a indústria da mobilidade através de recursos completos para desenvolvimento de produtos incluindo estilo, *design* e engenharia, prototipagem e testes.

Dentro da área automóvel e da mobilidade, o CEIIA tem como objetivo a conceção e desenvolvimento de projetos focados na industrialização e marketing de novos produtos e

serviços. Este conjunto de projetos tem como objetivo o desenvolvimento de infraestruturas e equipamentos de carregamento para veículos de duas e quatro rodas.

Dentro desta área surge o mobi.me (plataforma explicada com mais pormenor na secção 3.1) como uma plataforma de gestão inteligente da mobilidade elétrica, com a função de gerir a integração de diferentes serviços dentro de uma cidade (integração com parques de estacionamento, gestão cruzada de transportes públicos, utilização e partilha de carros e bicicletas) e ao mesmo tempo gerir a energia através de uma rede inteligente de carregamentos. É assim responsável por toda a gestão da complexidade do negócio, pois possui toda a informação acerca dos utilizadores, operadores e postos de carregamento, num único ecossistema.

Com a evolução que se tem verificado desta plataforma e do ambiente que a constitui, surge a necessidade de adotar um novo sistema que traga novas fontes de dados e que suporte o aumento da informação gerada pelos dipositivos.

1.2. Finalidade e objetivos

A finalidade deste projeto é assim, conceber um protótipo de um sistema *Big Data*, capaz de suportar todos os dados recebidos, armazená-los e processá-los, bem como obter dados de novas fontes (redes sociais), criando indicadores úteis para a visualização e a análise dos resultados obtidos. Desta forma, será possível uma melhor perceção da realidade diária das cidades conseguindo proporcionar uma melhoria do serviço prestado e o bem-estar dos cidadãos. Com a implementação deste sistema será possível adquirir novas perspetivas de negócio a nível mundial bem como novos pontos de interesse para atuar.

Para que esta finalidade seja alcançada com sucesso, é necessário cumprir um conjunto de objetivos:

- ✓ Identificar atuais implementações na área de *Big Data* e analisar as suas vantagens e desvantagens;
- ✓ Identificar quais as fontes de dados disponíveis;
- ✓ Identificar a estrutura em que os dados são fornecidos bem como o seu tipo e formato;
- ✓ Identificar técnicas de processamento de texto e análise;
- ✓ Construção do protótipo do sistema para todo o processo, iniciando-se na extração dos dados, passando ao seu processamento e armazenamento, concluindo com a visualização dos resultados;
- ✓ Validação do protótipo e avaliação da qualidade do sistema.

Como resultados dos objetivos enumerados espera-se:

- ✓ Extração de dados de novas fontes de dados como redes sociais;
- ✓ Análise do sentimento dos utilizadores perante determinadas palavras-chave enquadradas no ambiente do negócio;
- ✓ Novas formas de visualização dos dados, como por exemplo infografias.

1.3. Metodologias de investigação

Na área dos Sistemas de Informação, destaca-se a metodologia *Design Science Research* (DSR) que disponibiliza um conjunto de técnicas e perspetivas de pesquisa, bem como princípios, práticas e procedimentos para que haja sucesso no final do projeto (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007).

Esta metodologia está dividida em seis passos, seguidamente representados (Figura 1) e explicados.

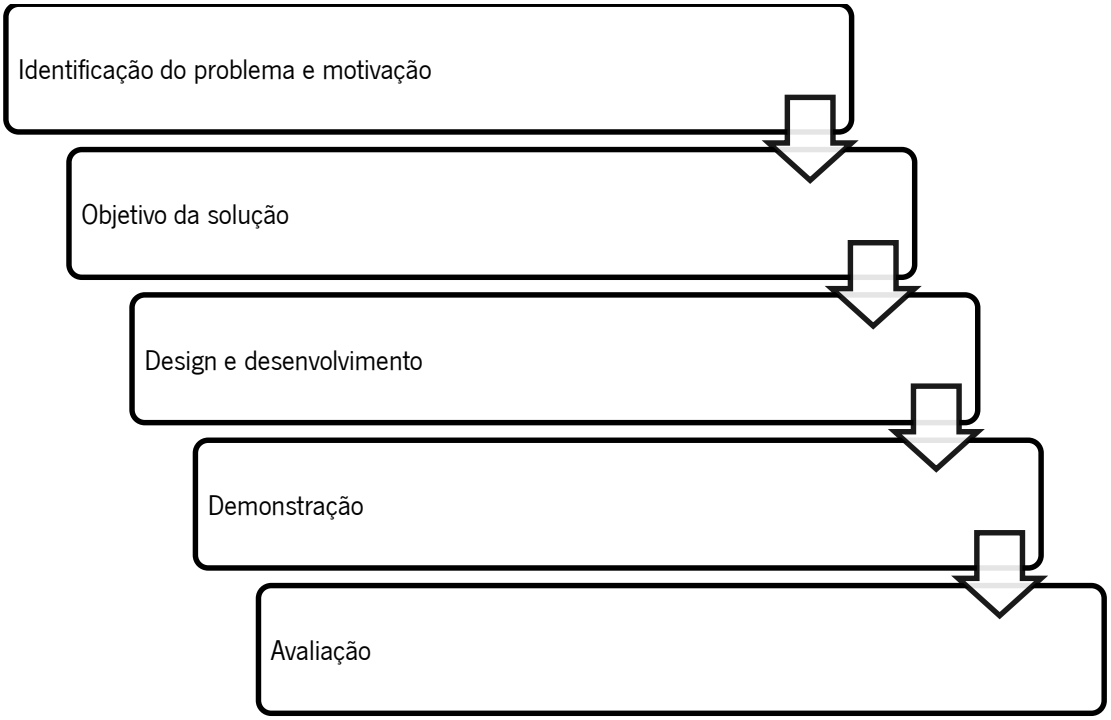


Figura 1 - Metodologia *Design Science Research*

Identificação do problema e motivação – Neste primeiro passo é necessário definir o problema e justificar qual o valor da futura solução. Além disto, é necessário conhecer todo o estado da arte da temática em causa de forma a conseguir focar o principal problema e aquilo que já foi feito para o tentar resolver. Para tal, enquadrando com o presente projeto, este passo diz respeito à revisão de literatura acerca do conceito de *Big Data*, das suas características e arquiteturas já existentes, bem como do conceito de *Smart Cities* e a necessidade da integração destes dois conceitos, demonstrando casos de sucesso da sua implementação. Um outro conceito também importante para o desenvolvimento do projeto consiste no *sentiment analysis* o qual também foi revisto.

Objetivo da solução – Depois de adquirir todo o conhecimento do estado da arte sobre o tema, é possível definir os objetivos do projeto. Assim, depois da aquisição do conhecimento dos conceitos acima descritos é possível definir os objetivos a cumprir neste projeto.

***Design* e desenvolvimento** – Neste passo, pretende-se definir as funcionalidades do sistema a desenvolver bem como a arquitetura que o irá suportar. Após isso será possível desenvolver o sistema *Big Data* pretendido pelo CEIIA com as respetivas funcionalidades de suporte às suas atividades.

Demonstração – Após definir os objetivos, a arquitetura, as funcionalidades do sistema e quais as tecnologias a utilizar, foi feita uma análise de casos de estudo de sucesso, onde foram integrados estes dois conceitos, de forma a garantir que o problema inicialmente encontrado seja resolvido.

Avaliação – Neste passo, é necessário reunir todo o conhecimento adquirido nos passos anteriores bem como os objetivos definidos para o sistema a desenvolver:

- Funcionalidades;
- Arquitetura
- Requisitos,

No final do projeto, será feita uma comparação dos objetivos definidos inicialmente com o produto final. Caso os resultados não sejam os esperados, o sistema terá de voltar ao passo do *design* e desenvolvimento.

Comunicação – O último passo diz respeito à publicação do projeto e dos resultados. Este projeto foi descrito numa dissertação de mestrado.

1.4. Estrutura do documento

Para que seja possível uma melhor organização e compreensão, este documento foi dividido em capítulos que por sua vez estão divididos em secções e subsecções.

O primeiro capítulo é composto pelo enquadramento do contexto onde se insere esta dissertação de mestrado, seguido da sua finalidade e objetivos que se pretende cumprir, bem como a metodologia usada para que seja possível a sua concretização.

No segundo capítulo é feito todo o enquadramento concetual acerca dos temas *Big Data*, *Smart Cities*, a integração destes dois temas, *Sentiment Analysis* e arquiteturas *Big Data*. Além disto, é feito também o enquadramento tecnológico, fazendo um levantamento do mercado e as respetivas características.

O terceiro capítulo contém a descrição do contexto em que este sistema foi desenvolvido, seguido da proposta de arquitetura. Além disto, contém a descrição de todas as etapas seguidas para a sua elaboração.

O quarto capítulo é composto pelos resultados finais resultantes de todas as etapas de construção do sistema. Aqui são apresentados indicadores definidos bem como a análise e visualização dos mesmos. É feita também uma pequena discussão em forma de conclusão acerca dos objetivos definidos e resultados cumpridos.

O último capítulo diz respeito à conclusão deste projeto. Aqui apresenta-se o trabalho realizado, as dificuldades encontradas e trabalho que futuramente possa ser realizado.

Capitulo 2 - Enquadramento Concetual e Tecnológico

Neste capítulo, será apresentada toda a revisão de literatura acerca do tema da dissertação aqui apresentada bem como toda a análise tecnológica efetuada para a realização do protótipo do sistema desenvolvido.

2.1. *Big Data*

2.1.1. O conceito

A quantidade de dados gerada em todo mundo sofreu um crescimento explosivo tornando-se uma base essencial para o mercado competitivo, apoiando o crescimento da produtividade, inovação (Lohr, 2012). A este fenómeno chamamos *Big Data* definido já por alguns autores apesar da recente expansão do tema.

Uma das definições é proposta pela Oracle (2012) que se refere a estes sistemas como um conjunto de dados de grande dimensão, que têm como desafio o armazenamento, a pesquisa, a partilha a visualização e a análise da informação gerada.

Uma outra definição é dada por Vieira et al. (2012) que refere *Big Data* como uma coleção de bases de dados complexas e volumosas onde é impossível executar operações simples de forma eficiente usando sistemas de gestão de base de dados tradicionais.

Krishnan (2013) define *Big Data* como sendo volumes de dados disponíveis em diferentes graus de complexidade e ambiguidade gerada em diferentes velocidades. Refere também a impossibilidade de serem processados usando tecnologias, métodos de processamento e algoritmos tradicionais.

Lohr (2012) considera que o *Big Data* tem potencial para ser uma "*dashboard* da humanidade" e que é uma ferramenta inteligente que ajudará a combater a pobreza, o crime e a poluição. Refere também que *Big Data* é um termo de marketing mas também um avanço na tecnologia que permite entender o mundo.

Analisando cada definição podemos verificar que é destacada a capacidade de armazenar e processar grandes quantidades de dados, em diversos formatos e tipos, bem como a vantagem de visualização e análise dos dados em tempo real, com uma velocidade superior. É assim um conceito definido de forma idêntica por diferentes autores, onde existe consenso no que diz respeito às capacidades deste tipo de sistema.

Demchenko et al (2013) acrescenta que apesar de todas as definições dadas para esta tecnologia, não existe uma análise detalhada da mesma. Diz ainda que, no momento, a maioria das discussões assentam nas características mais significativas do *Big Data* bem como nos incentivos para a sua existência.

2.1.2. Características

Inicialmente o *Big Data* era caracterizado e definido por três características essenciais, sendo elas o volume, a variedade e a velocidade dos dados (Laney, 2001). Com o passar dos anos e com todos os estudos já efetuados até à atualidade podemos neste momento identificar cinco características importantes nesta nova tecnologia. Às três já existentes, Demchenko et al (2013) acrescenta a veracidade e o valor (Figura 2).

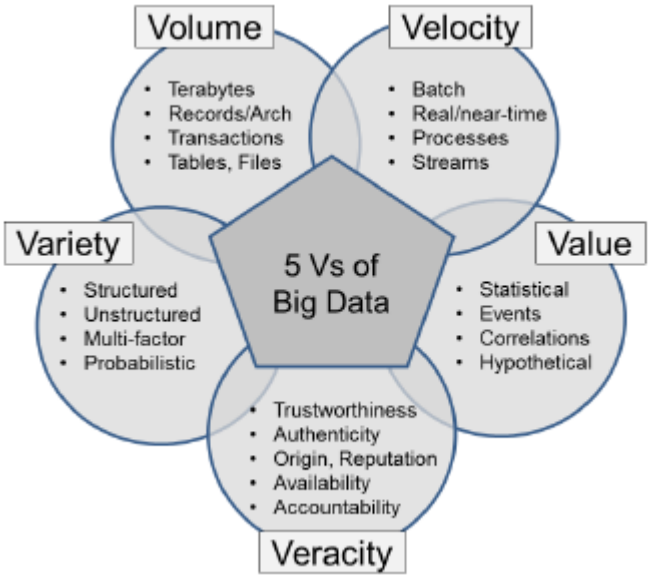


Figura 2 – 5 V's - Características *Big Data* (Demchenko et al., 2013)

Como já foi referido, a quantidade de dados armazenada atualmente é de larga escala, o que torna impossível a sua análise na totalidade com os sistemas tradicionais de gestão de Base de Dados. Esta definição diz respeito a uma das características do *Big Data*, o volume.

Estamos constantemente a guardar registos sobre tudo o que fazemos. O mesmo acontece com as organizações que diariamente nas suas atividades possuem informação rica, da qual não sabem tirar valor devido a sua quantidade e não estruturação. A verdade é que todos esses dados são úteis para perceção do negócio, dos clientes e do mercado (Zikopoulos et al., 2011). Assim, o volume consiste na quantidade de dados gerada continuamente sendo estes de diferentes tipos

com diferentes tamanhos (Krishnan, 2013). Demchenko et al. (2013) distingue esta característica como a mais importante, que impõe requisitos adicionais e específicos para todas as tecnologias e ferramentas tradicionais utilizadas.

Uma outra característica deste tema é a variedade. Com a explosão da utilização de sensores e dispositivos inteligentes, os dados gerados pelas empresas tornam-se complexos devido à falta de estruturação e ao facto de não seguirem o modelo relacional tradicional (Zikopoulos et al., 2011). Segundo Krishnan (2013), esta complexidade associada à variedade de formatos tem a ver com a disponibilidade de informação adequada acerca do que está contido nos dados reais recolhidos, sendo fundamental para processamento de imagens, vídeo e grandes quantidades de texto. Para que seja possível processar estes novos formatos, o sistema *Big Data* requer alta escalabilidade, capacidade de processamento distribuído, capacidade de processamento de imagens, gráficos, vídeo e som. Resumidamente, a variedade representa todos os tipos de dados como parte do processo de decisão.

O *Big Data* tem também como característica a velocidade que consiste na rapidez em que os dados chegam e são armazenados. Esta característica torna-se essencial no aparecimento desta temática.

Os fluxos de dados de forma contínua e o conjunto de resultados são úteis quando a aquisição e processamento têm um tempo de atraso curto. É nesta situação que se torna crítica a recolha e processamento com velocidades escaláveis, com diferentes tamanhos, numa quantidade de tempo mínima (Zikopoulos et al., 2011).

Para que seja garantida a qualidade dos dados recebidos, o *Big Data* caracteriza-se pela sua veracidade (Sridhar & Dharmaji, 2013). Esta qualidade é garantida pela consistência dos dados, nomeadamente, a sua origem, recolha e métodos de processamento, utilizando infraestruturas confiáveis (Demchenko et al., 2013).

Por último, esta tecnologia caracteriza-se pelo valor que origina. Esta é uma característica de grande importância, definindo-se pelo valor agregado que os dados recolhidos podem trazer para o processo, atividade ou análise a que se destina. Relaciona-se diretamente com a variedade e o volume (Demchenko et al., 2013).

Segundo Krishnan (2013) a par destas características surgem os conceitos de ambiguidade, viscosidade e viralidade como características adicionais (Figura3).

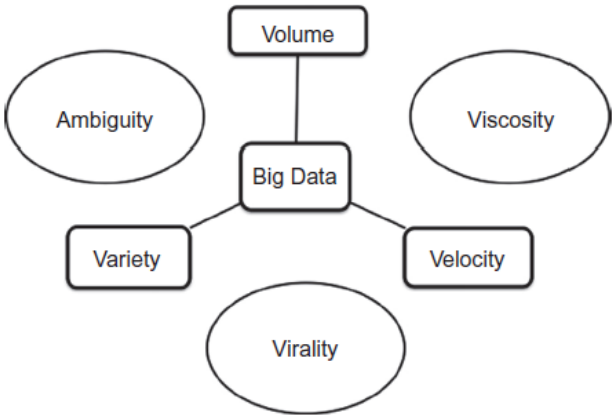


Figura3 - Características adicionais do *Big Data* (Krishnan, 2013)

A ambiguidade aparece quando há falta de informação acerca dos dados, manifestando-se na maioria das vezes nas características variedade e volume, ou seja, quando existe grande quantidade de dados e ao mesmo tempo estes têm grande variedade. A viscosidade mede a resistência de navegação no volume de dados, manifestada em fluxos de dados e regras de negócio, podendo se tornar uma limitação da tecnologia. Por último, a viralidade mede e descreve a rapidez com que os dados são compartilhados.

2.1.3. Análise de dados e *Big Data*

A análise de dados torna-se uma parte essencial do processo de tomada de decisão num mercado cada vez mais competitivo.

O referido crescimento da quantidade de dados e da forma como é gerido, leva as organizações a despenderem muitos recursos nas infraestruturas de *Business Intelligence* (BI), de Tecnologias de Informação (TI), aplicações de transação e ferramentas de análise de negócio. No entanto, toda esta massiva quantidade de dados será cada vez mais usada na análise de negócio para melhoria das operações e serviços (Sridhar & Dharmaji, 2013).

Sridhar & Dharmaji (Sridhar & Dharmaji, 2013) consideram a análise de negócio, uma função do BI que, quando combinados, proporcionam uma mais rápida e eficiente tomada de decisão, aumentando a produtividade global através do acesso aos dados. Através desta análise é possível compreender o funcionamento da empresa, as suas ligações com o mercado e os indicadores

chave de desempenho para impulsionar as decisões de negócio e as melhorias no desempenho dos processos de negócio. Para tal, o *Big Data* proporciona uma análise preditiva em tempo real composta pela análise avançada, constituída por um conjunto de técnicas analíticas utilizadas para prever resultados futuros. Prevê e disponibiliza técnicas que permitem identificar tendências atuais e futuras, de forma a guiar as decisões para que os resultados futuros sejam melhorados.

2.1.4. Etapas de construção de um sistema *Big Data*

Num sistema tradicional de gestão de base de dados, inicia-se todo o processo pela análise dos dados disponíveis e definição dos requisitos, estabelecendo o modelo de dados. Estes sistemas são bastante eficientes sob uma perspetiva de desempenho na escrita, pois são dados com estrutura e forma.

O processamento de dados num sistema *Big Data* inclui etapas muito semelhantes ao processamento de dados de um sistema transacional (Krishnan, 2013).

Neste tipo de sistema o processamento de dados é iniciado pela recolha e carregamento dos dados para um local de armazenamento, onde os mesmos são estruturados e contextualizados e posteriormente transformados e analisados. O resultado final deste processo fornece uma nova visão sobre os dados, associado ao seu contexto segundo as regras de negócio definidas.

Segundo Krishnan (2013), existem quatro etapas distintas de processamento. Na Figura4, estão representas as etapas respeitantes a este processo.

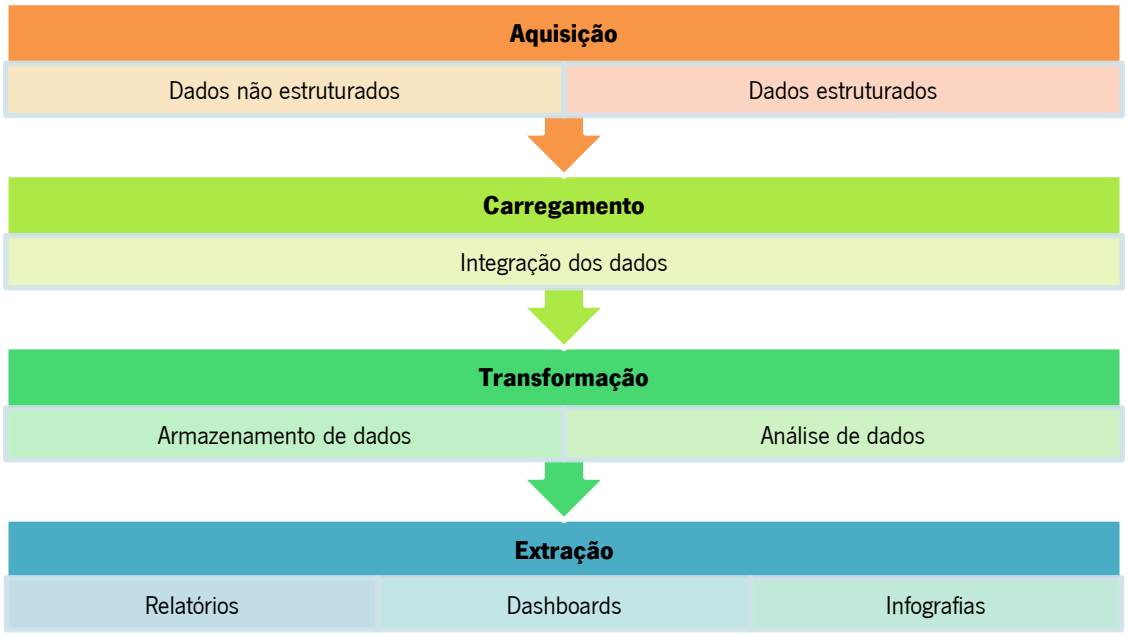


Figura4 - Ciclo de vida de um sistema *Big Data* (Adaptado de: (Krishnan, 2013))

A primeira etapa consiste na aquisição de dados onde é feita a sua recolha nas diferentes fontes (ficheiros, sensores, imagens, som, vídeos e sistemas em tempo-real) e depois carregados para um sistema de ficheiros (*Hadoop* ou *NoSQL*), normalmente classificados em subdiretorias, segundo o seu tipo.

Segundo Dijcks (2012), estes dados são essenciais para o negócio e, para que a organização possa tirar partido deles para o processo de tomada de decisão, é imprescindível implementar um sistema que seja distribuído devido à necessidade de carregar os dados tão rápido quanto possível.

Depois de adquiridos, os dados são carregados e estruturados, ficando preparados para serem transformados. Durante o carregamento, os dados são repartidos por vários ficheiros, que depois passam a ser catalogados num conjunto de ficheiros, associando o seu contexto. Esta partição dos dados pode ser feita de forma horizontal ou vertical, dependendo dos requisitos definidos e dos seus utilizadores. Entenda-se como horizontal como uma organização em linhas e vertical como uma organização em colunas.

Depois de organizados os dados, e como já referido, é feita a sua transformação aplicando regras de negócio e processando o contexto em que se inserem. Esta etapa é constituída por vários passos de execução, tornando a sua gestão complexa. Aqui é também feito o processamento dos dados onde se realiza uma análise dos dados, de forma a prepará-los para a integração com o

Data Warehouse e com as ferramentas de análise através da sua marcação, classificação e categorização. O resultado desta etapa traduz-se num conjunto de chaves associadas a métricas (par chave-valor).

Numa última etapa, são extraídos os resultados das etapas anteriores para análise, elaboração de relatórios e visualização dos dados.

2.2. *Big Data* em *Smart Cities*

2.2.1. Necessidade de informação

O aumento populacional das cidades está em constante evolução e, juntamente com este fenómeno tem crescido o congestionamento do trânsito, os trajetos têm uma maior distância e a poluição é cada vez maior (Latamore, 2013).

Este aumento da população urbana tem-se tornado num papel central da evolução da humanidade. Consequentemente, os esforços para facilitar melhores condições de vida são imensos e, segundo Vilajosana et al (2013), as tecnologias podem ser a solução a este problema tendo em conta os seus avanços.

As cidades são uma fonte de dados gerados em tempo real e de grande dimensão. Estes dados têm origem nos tradicionais sensores, GPS inseridos nos autocarros, dados meteorológicos e camaras de trânsito. Depois de recolhidos são analisados para que seja possível fazer previsões de tráfego em diferentes partes da rede rodoviária e posteriormente ser possível intervir resolvendo problemas detetados (Latamore, 2013).

O conceito de *Smart Cities* tem sido adotado por vários grupos de pesquisa, empresas e entidades públicas para condensar um conjunto de serviços avançados, com o objetivo de tornar as cidades mais eficientes, sustentáveis e favoráveis ao cidadão.

Este paradigma pode ser visto como um sistema distribuído, com diferentes fontes de informação de onde derivam dados úteis para um conjunto de aplicações. Estes dados são usados para elaboração de respostas a nível estratégico e tático (Villanueva, Santofimia, Villa, Barba, & Lopez, 2013).

Pike Research (Vilajosana et al., 2013) define *Smart City* como a integração da tecnologia com uma estratégia de sustentabilidade, bem-estar dos cidadãos e desenvolvimento económico. Um modelo de *Smart City* que corresponde a esta definição deve ser multidimensional, englobar

diferentes aspetos da inteligência e salientar a importância da integração e interação em vários domínios.

Mitchell & Casalegno (2009) definem *Smart City* como a utilização eficiente e responsável dos recursos escassos que são necessários para o funcionamento da cidade.

A empresa espanhola Telefónica (2011) no seu livro acerca das *Smart Cities*, descreve-as como sendo uma cidade que usa as TI para que a sua infraestrutura crítica, os seus componentes e serviços públicos oferecidos sejam mais interativos e eficientes tendo em conta que os seus cidadãos tenham consciência da sua utilização.

Para reduzir dados isolados, os processos de negócio e encorajar a partilha bem como o fluxo de dados, as cidades bem administradas estabelecem capacidades de negócio comuns, nomeadamente infraestruturas ricas em dados e arquiteturas orientadas ao serviço. O objetivo para as TI é provar que esta centralização ou reestruturação dos serviços é prático e útil (Hedlund, 2011).

Segundo Hedlund (2011), quando é permitida a integração de dados e processos das cidades em produtos comerciais de terceiros, facilita aos cidadãos encontrar e consumir informação quando e como querem acerca do tráfego local, transportes públicos ou serviços públicos.

Os processos das cidades eram otimizados para os funcionários e agências públicas, e não para os cidadãos, refletindo a forma como estas prestavam os serviços. As novas tecnologias oferecem melhores formas de medir os indicadores de desempenho que realmente importam para os cidadãos, permitindo que seja seguido um modelo de *Smart City* que coloca o cidadão no centro dos serviços (Hedlund, 2011).

A mudança de uma cidade para uma *Smart City* melhora a sua eficiência, cria um impacto ambiental positivo, aumenta a segurança e melhora a saúde. Estes resultados tornam a cidade mais habitável, atraente, economicamente viável e, por conseguinte, atrai novos cidadãos e novos negócios. Hedlund (2011) salienta a sustentabilidade ambiental, social e económica (Tabela 1).

As *Smart Cities* convertem-se numa plataforma digital que permite maximizar a economia, a sociedade, o ambiente, o bem-estar das cidades e facilita a mudança para um comportamento mais sustentável entre todos os utilizadores. Além disto, visa o aproveitamento máximo dos orçamentos públicos em consequência da melhoria dos processos das cidades e dos seus

habitantes. Por outro lado, permite capacitar novos modelos de negócio constituindo uma excelente plataforma de inovação (Telefónica, 2011).

Tabela 1 - Cidades Tradicionais vs *Smart Cities* (Hedlund, 2011)

Cidades tradicionais	***Smart Cities***
Compram ou desenvolvem aplicações personalizadas	Investimento e reutilização de capacidades comuns das TI e do negócio; Configuram aplicações personalizadas em vez se comprarem ou desenvolverem.
Mudança cara e lenta devido a problemas legais ou de aquisição; Serviços estáticos, raramente utilizados.	Serviços ágeis e rapidamente atualizados; Redução dos custos com TI; Redução dos custos das empresas.
Fácil utilização para o utilizador que conhece a cidade; Dificuldade na utilização por parte dos cidadãos ou por alguém fora do contexto da cidade.	Centralização do cidadão; Facilidade de utilização.
Integração e reutilização dos dados dispendiosa	Fácil integração e reutilização de dados; Alta qualidade dos dados.
Baixos custos com infraestruturas comuns; Custos altos em TI.	Altos custos com infraestruturas comuns; Baixos custos em TI.
Indisponível para serviços da *Cloud*.	Disponível para serviços da *Cloud*.
Cidadãos excluídos dos processos das cidades.	Transparência, otimização e responsabilização dos cidadãos.
Pouca empregabilidade de trabalhadores da cidade; Baixa automatização.	Emprega normalmente cidadãos; Centralização dos serviços de atendimento ao cidadão reduzindo os custos.

Podemos assim considerar vantagens nestas cidades como: a redução de gastos públicos, nomeadamente no fornecimento e gestão dos serviços, aumento da eficiência e qualidade destes serviços através da gestão eficiente dos recursos e da qualidade do serviço prestado, suporte à tomada de decisão facilitando o reconhecimento das necessidades da cidade e o planeamento de novos serviços, promoção da inovação gerando novos negócios e desenvolvimento social, disponibilização de informação em tempo real (Telefónica, 2011).

Em suma, as *Smart Cities* apoiam o desenvolvimento das cidades e cidadãos tanto na resolução dos problemas atuais como na identificação e gestão de problemas futuros.

2.2.2. Serviços das *Smart Cities*

No processo de implementação de uma *Smart City* deve-se ter sempre uma visão das necessidades existentes. Neste tipo de projeto surge como serviços a mobilidade urbana, a eficiência energética, a gestão sustentável dos recursos, a gestão das infraestruturas da cidade, a participação do governo, segurança pública, a saúde, a educação e a cultura.

A mobilidade urbana é definida pela sustentabilidade, segurança e eficiência das infraestruturas e sistemas de transporte, assim como pela acessibilidade local, nacional e internacional. Este serviço é considerado pela Telefónica (Telefónica, 2011) um dos maiores problemas nas cidades. Aqui destaca-se o congestionamento do tráfego que tem um impacto negativo considerável na qualidade de vida da cidade e do ambiente, provocando a diminuição da produtividade.

Para esta problemática a Telefónica (2011) sugere algumas soluções.

A primeira solução sugere a gestão do tráfego em tempo real através da disponibilização da informação aos utilizadores do estado do tráfego, de ocorrências nas estradas, de obras rodoviárias, sinalização, entre outras. Ao mesmo tempo será necessária a atualização de mapas, sugestões de rotas alternativas com tempo ou distância idêntica e sugestões de condução ecológica.

A segunda solução passa pela gestão dos meios de transporte que, com o aumento de população nas áreas urbanas, tem sofrido um impulso para o crescimento em termos de capacidade, de forma a satisfazer a procura. Aqui, o utilizador dispõe de informação de locais de paragem, circuitos, distâncias.

A gestão de estacionamentos é também uma forma de resolver o congestionamento através de sensores que informem ao utilizador os locais onde existem lugares de estacionamento livres. Quanto mais rápido os condutores estacionarem o carro maior será a circulação do trânsito.

Um segundo serviço das cidades diz respeito à eficiência energética e meio ambiente onde as *Smart Cities* têm muito para contribuir. O aumento do preço da energia e o aumento da poluição são temas prioritários nas sociedades modernas pois existe necessidade de tornar eficiente a gestão energética bem como a gestão ambiental.

A gestão de infraestruturas e edifícios públicos passa pela eficiência na utilização dos recursos que aqui se encontram disponibilizados. Associado a esta gestão temos o governo e a cidadania

onde o cidadão tem acesso a serviços de participação ativa de forma a contribuir no processo de tomada de decisão, através de plataformas públicas (ex.: e-Administração, e-Participação).

O referido crescimento das cidades e da sua população traz consigo a necessidade do aumento da segurança pública, onde é preciso coordenar grande quantidade de agentes e recursos. As TI entram aqui como um importante benefício.

Em áreas como serviços de emergência e proteção civil, aplicações que permitam otimizar o tempo de resposta são de grande utilidade. Neste contexto, podem ser utilizados serviços de videovigilância centrados em determinadas áreas, sensores que localizam pessoas de forma a prever situações de aglomeração.

A partir de toda a análise efetuada ao contexto das *Smart Cities* podemos verificar que em todas as atividades e serviços são gerados, armazenados, explorados e analisados grandes quantidades de dados.

Bert Latamore (2013) numa iniciativa da IBM acerca deste tema refere três partes essenciais para que este sistema seja eficaz. Numa primeira fase é necessário identificar o problema, de seguida informar os condutores para que saibam o que vão encontrar e por último implementar um sistema que mantenha o tráfego em movimento. Como capturar e responder aos dados em tempo real não é viável, o objetivo passa por prever o que vai acontecer nos minutos seguintes.

Nasce aqui a necessidade de ligação entre o conceito de *Smart Cities* e *Big Data*. Pois como verificamos acima nos seus conceitos, além do suporte a grande quantidade de dados, temos também dados de forma não estruturada, gerados em tempo real.

France (2013), dentro deste âmbito, refere que as *Smart Cities* estão a ser impulsionadas pelo *Big Data* e por sistemas humanos que quando combinados formam uma solução eficaz.

Os dispositivos de mobilidade estão cada vez mais conectados e geram muita informação que pode ser utilizada para melhorar as condições do tráfego e a experiência da condução. No caso deste estudo, os sensores colocados nos alertas dos carros emitiam avisos através de *Dashboards* de forma a alertar para a baixa pressão dos pneus, luzes partidas ou falhas internas, ABS partido, forte aceleração e estradas escorregadias. Os dados recolhidos nestes dispositivos eram enviados para a IBM *Smart Traffic Center* juntamente com a localização GPS. Assim, será possível usar

esses dados para avisar os respetivos condutores e as autoridades sobre as condições perigosas das estradas, acidentes ou densidade do trânsito em tempo real.

Nesta vertente de carros conectados há necessidade deste novo conceito de *Big Data* para a segurança, eficiência e fluxo do trânsito.

2.2.3. Casos de estudo

Apesar da atualidade destas temáticas existem já casos de implementação onde estes dois conceitos são integrados com sucesso.

Caso de estudo 1

O primeiro caso surge de uma cooperação entre a NXP *Semiconductors*, IBM, Beijer *Automative*, Nokia, TASS, *Technische Universiteit Eindhoven*, TNO, ANWB, CibaTax, KPN and the *Dutch Directorate-General for Public Works and Water Management*. O objetivo era melhorar as condições do tráfego e a segurança rodoviária através de uma tecnologia que monitoriza e reporta as condições da estrada em tempo real, alertando os motoristas nas proximidades de forma a evitar acidentes.

A NXP *Semiconductors*, a IBM e os seus parceiros equiparam os veículos com um dispositivo, que contém um chip que reúne dados relevantes do sistema de comunicação central do carro. Estes dispositivos recolhem e convertem os dados relevantes do sensor anonimamente e envia para a um centro de tráfego da IBM, juntamente com a localização GPS. Através destes dados recebidos, serão feitas análises para que os motoristas e as autoridades rodoviárias sejam informados sobre condições adversas das estradas, acidentes ou densidade do tráfego em tempo real.

Toda esta circulação de informação traz benefícios em vários aspetos, nomeadamente para os condutores e para as autoridades. Sendo disponibilizada a informação em tempo real, em caso de condições adversas, as autoridades poderão agir mais rapidamente não deixando aumentar o fluxo rodoviário. Ao mesmo tempo, os condutores são informados das condições da via não correndo riscos de acidente ou trânsito congestionado.

Podemos verificar que estas atividades em cidades necessitam de ser conjugadas com sistemas que suportem todos os dados gerados pelos dispositivos referidos neste caso, e ao mesmo tempo a sua forma não estruturada.

Caso de estudo 2

O segundo caso de integração destes conceitos foi feito em Barcelona, Espanha onde havia a necessidade de armazenar e analisar *Big Data* de forma a perceber as necessidades dos cidadãos e às empresas (Enbysk, 2013).

Os dados foram recolhidos do sistema autárquico, de novas fontes públicas, das redes sociais e de sinais GPS com o objetivo de aumentar as oportunidades de negócio e os serviços, aumentar a segurança e encorajar a colaboração da cidade, dos cidadãos e das empresas. Foram assim criados três novos serviços como *Open Data*, *bigov Better City Indicators* e *La Mercé*.

O serviço *Open Data* tinha o objetivo de recolher informação pública acerca da cidade como mapas da cidade e de instalações públicas, população, perfis dos contratantes, eventos da cidade, economia, empresas e eleições.

O segundo serviço, *bigov Better City Indicators*, consistia numa *Dashboard* com dados acerca de procedimentos administrativos, serviços da cidade incluindo a utilização de bicicletas públicas e o número de pessoas que utilizam cada linha de autocarros, economia e demografia da população. Este serviço foi disponibilizado aos funcionários públicos e aos cidadãos e fornece dados em tempo-real.

O último serviço, *La Mercé*, tinha como função o planeamento do maior festival nacional com o mesmo nome do serviço. Os dados recolhidos eram acerca das atividades de entretenimento do festival, locais para alimentação, informação do interesse e para a satisfação do cidadão, apoio à mobilidade das pessoas e deteção de problemas. Estes dados encontravam-se tanto na forma estruturada como não estruturada, o que justifica a utilização de um sistema *Big Data*. Ao mesmo tempo, era incluída informação das redes sociais onde os participantes publicavam fotografias ou opiniões acerca dos eventos.

Além disto, os participantes teriam informação dos parques de estacionamento, estado do trânsito, do tempo e informação GPS.

Com estes serviços, foi possível aumentar o nível de participação dos cidadãos bem como uma visão transparente e pessoal do governo.

2.3. Sentiment Analysis

Nos últimos anos, as redes sociais tornaram-se importantes no que diz respeito à partilha de conteúdos e à informação que a partir daí é gerada. Desde então, estas fontes de dados podem ser apelidadas como uma forma de sabedoria coletiva onde é possível a previsão de resultados através de informação real (Asur & Huberman, 2010). A verdade é que o valor do conteúdo destes dados é de bastante importância, podendo ser analisados de diversas formas.

Uma das formas de análise possíveis neste contexto é a exploração e análise de opiniões ou sentimentos dos utilizadores, através do conteúdo que publicam.

Esta técnica de análise chama-se *Sentiment Analysis* e consiste principalmente numa área de estudo que lida com linguagem natural orientada ao processamento de opiniões e sentimentos (Gebremeskel, 2011). Este conceito assume uma grande importância pois, a partir daqui, é possível para as diferentes áreas de negócio obter oportunidades estratégicas a partir do estudo e descoberta de conhecimento dentro do conteúdo expresso pelos utilizadores da Web (Bermingham & Smeaton, 2010). Este tema não é recente mas, devido à referida explosão das redes sociais, tem sofrido um crescimento substancial.

A exploração das opiniões e sentimentos permitirá às organizações analisar como os seus clientes se sentem ou qual a sua opinião em relação aos seus produtos e serviços. Da mesma forma, torna possível analisar novos mercados, novos clientes e novos produtos (Liu, 2011b). Estas opiniões são uma importante influência nos seus comportamentos.

Segundo Liu (2011a), as crenças e perceções da realidade, bem como as decisões tomadas são condicionadas pela forma como o mundo é visto e avaliado. Por esta razão, quando existe necessidade de tomar decisões, é importante analisar e perceber qual a opinião que existe sobre determinado tema. Assim, as organizações cada vez mais usam a informação daí gerada como apoio à tomada de decisão.

Além destes, as organizações possuem também os seus próprios dados resultantes da sua atividade e de onde podem também extrair este tipo de análise.

Os dados extraídos vêm normalmente em formato de texto, identificado e classificado através do conteúdo emocional criado pelos utilizadores, expressando opiniões positivas, negativas ou neutras. Daqui pode então ser retirada a polaridade da opinião ou a orientação do sentimento

(Chew, 2010). Ao mesmo tempo este texto pode ser implícito ou explícito conforme a expressão usada pelo individuo.

Segundo Wilson et al (2005), a tarefa de *sentiment analysis* é feita ao nível do documento. Contudo, dados como respostas a questões, extração de opiniões comentários de produtos requerem uma análise a nível da frase.

A análise de sentimentos expressos em texto é uma tarefa difícil no que diz respeito à forma como cada indivíduo se expressa, podendo o sentido dado à palavra ou frase não ser totalmente claro. Um exemplo desta situação poderá ser o uso de ironias ou sarcasmos.

Tipicamente, esta análise inicia-se com um dicionário de palavras ou frases, previamente identificadas como positivas e negativas que classificam o texto com a sua polaridade. No entanto, existem limitações na atualização destes dicionários pois não estão preparados para identificar a semântica dada ao texto.

É importante aqui perceber a semântica das palavras fora do contexto em que estão inseridas para que seja feita uma análise correta do seu sentido. Por exemplo, nos países orientais a cor preta significa morte e tristeza o que transmite um sentimento negativo, mas se esta palavra for analisada nos países ocidentais não terá o mesmo significado pois usam a cor branca associada a esse sentimento.

Ao nível de análise de frases, deve ser considerado se a negação é feita ao nível local e se tem dependências como a negação da proposição ou negação do sujeito. Além disso, certas frases contêm palavras de negação para intensificar a mudança da polaridade. O contexto da polaridade pode também ser influenciado pela modalidade, sentido das palavras e a função sintática da palavra na frase.

2.4. Arquiteturas *Big Data*

Após o enquadramento do conceito de *Big Data* e das suas características, passamos agora a análise das arquiteturas associadas a este novo tema.

Como já vimos, o processamento de um sistema *Big Data* é bastante complexo necessitando de ser rápido e eficaz. Surgem aqui duas abordagens distintas que consistem no processamento centralizado e no processamento distribuído, sendo a segunda a mais relevante para este tipo de sistema.

O processamento centralizado consiste na recolha de dados para uma única área de armazenamento onde os mesmos são processados por um único computador, frequentemente com uma ampla arquitetura de memória, processador e armazenamento (Krishnan, 2013).

No processamento distribuído, os dados e o seu processamento são distribuídos geograficamente ou por *data centers,* localizados juntamente com os resultados num armazenamento centralizado. Este tipo de arquitetura veio superar as limitações do processamento centralizado e é esta que contém as características necessárias para suportar um sistema *Big Data* (Krishnan, 2013).

Dentro do processamento distribuído, Krishnan (2013) refere arquiteturas do tipo Cliente-Servidor, onde o cliente faz toda a recolha e apresentação dos dados, enquanto o servidor faz o processamento e gestão dos dados, do tipo *Cluster*, onde as máquinas estão ligadas numa arquitetura em rede para processar os dados em paralelo e cada *cluster* está associado a uma tarefa, e por último, do tipo *Peer-to-Peer*, onde não existem clientes nem servidores dedicados e todo o processamento é distribuído por máquinas que podem desempenhar estes papéis.

Segundo este autor, a vantagem do processamento distribuído passa pela sua escalabilidade e pela capacidade de processamento paralelo dos dados com a redução do tempo de latência. Pelo contrário existe redundância dos dados e dos processos, sobrecarga dos recursos e volume de dados.

Foram desenhadas combinações de infraestruturas práticas com o processamento distribuído para resolver problemas da utilização desta técnica com grandes quantidades de dados. Para tal, a arquitetura que suporta o processamento distribuído com grandes quantidades de dados deve ter características como: processamento paralelo dentro de um ou vários sistemas, utilização mínima da base de dados com a remoção das suas dependências, armazenamento distribuído

baseado em ficheiros pois é menos dispendioso, escalabilidade linear, alta velocidade de replicação, disponibilidade/viabilidade, o processamento localizado dos dados e armazenamento dos resultados e a tolerância a falhas devido à replicação extrema. Com estes recursos e capacidades, as limitações do processamento distribuído em bases de dados relacionais desaparece. Este novo modelo de arquitetura criou um ambiente de processamento distribuído escalável e extensível (Krishnan, 2013).

O processamento de um sistema *Big Data*, para que seja rápido e eficaz, necessita de cumprir requisitos como o volume, a velocidade, a variedade, a ambiguidade e a complexidade, sendo que os três primeiros foram já referidos como características destes sistemas.

O primeiro requisito, o volume, deve-se à necessidade de serem processados em paralelo grandes quantidades de dados, através de vários sistemas e vários módulos. A velocidade prende-se à necessidade de uma rápida recolha dos dados, bem como o processamento dos dados a partir de várias fontes e de diversos formatos, tipos, estruturas e localizações, surgindo aqui o requisito da variedade. Estes tipos de dados são também considerados ambíguos pois surgem sem contexto e sem informação da sua origem. O último requisito consiste na complexidade que diz respeito à necessidade de processar os dados rapidamente e eficientemente através de algoritmos.

O processamento de dados de larga escala requer um ambiente computacional de grande capacidade para que a sua gestão seja simples com uma escalabilidade linear.

Além do referido acima e para que o processamento destes sistemas seja conseguido com sucesso, é necessário cumprir requisitos específicos que o *Big Data* exige. Estes requisitos dizem respeito ao processamento dos dados e às infraestruturas de suporte. No processamento de dados é necessário um modelo de arquitetura menos complexo, a aquisição de dados quase em tempo real, a transformação mínima dos dados, uma eficiente leitura dos dados, a capacidade para múltiplas partições, o armazenamento em ficheiros ou sistemas de gestão de base de dados não relacionais e, por último, a partilha dos dados a partir de vários pontos de processamento.

Quanto às infraestruturas, estas devem ser linearmente escaláveis, com capacidade para grandes quantidades de dados, com tolerância a falhas, com capacidade de recuperação automática, alta capacidade de paralelismo, suportar processamento de dados distribuído e utilizar um interface de linguagem de programação.

2.5. Tecnologias *Big Data*

Nesta secção encontra-se a análise das ferramentas disponíveis, descrevendo as suas funcionalidades e qual a etapa de construção do sistema *Big Data* que melhor satisfaz. Encontra-se dividida em subsecções, onde a primeira representa as ferramentas e suas funcionalidades e a última uma comparação das funcionalidades das ferramentas, para que seja possível satisfazer os requisitos de um sistema deste tipo.

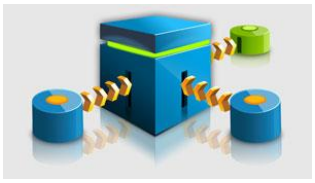
2.5.1. Ferramentas *open source* e comerciais

Para a elaboração do sistema *Big Data* em causa, há necessidade de cumprir vários requisitos para que o mesmo funcione da melhor forma.

Como já referido anteriormente, estes sistemas necessitam principalmente de processamento distribuído em tempo real, de grande escalabilidade e de armazenamento de grandes quantidades de dados, estruturados ou não estruturados. Desta forma, é necessário selecionar as ferramentas que melhor conseguem satisfazer estes requisitos tornando o sistema o mais eficiente possível.

Na Tabela 2 podemos verificar as ferramentas *Open Source e* comerciais disponíveis, bem como as suas funcionalidades:

Tabela 2 - Análise e descrição das ferramentas

	Ferramenta	Funcionalidades
Aquisição 	*Flume (Cloudera Virtual Machine CDH)*	✓ Distribuído; ✓ Recolha de dados eficiente; ✓ Arquitetura simples e flexível baseada em *streaming* de dados; ✓ Robusto e tolerante a falhas; ✓ Modelo de dados expansível.
	Knime Palladian	✓ Algoritmos de extração de dados baseados em Java.
Processamento 	*Knime*	✓ Integração de dados a partir de diversas fontes (ficheiros, bases de dados); ✓ Transformação de dados; ✓ *Data Mining*; ✓ Análise de Redes Sociais; ✓ Processamento de imagens; ✓ Processamento de texto; ✓ *Text Mining*.
	Hive (Cloudera Virtual Machine CDH)	✓ Suporte a *queries ad-hoc;* ✓ Usa *MapReduce* para processar dados; ✓ *Text Mining*; ✓ Indexação de documentos; ✓ Utiliza algoritmos *MapReduce*; ✓ Modelos preditivos; ✓ Usa o *MySQL* como base de dados dos metadados.
	Pentaho Community Edition	✓ Integração dos dados; ✓ Conectado com *Hadoop*, *NoSQL* e bases de dados analíticas; ✓ Utiliza *MapReduce* para reduzir ciclos de desenvolvimento; ✓ Preparação, modelação e exploração de dados não estruturados; ✓ Integração de dados com processamento paralelo com rápida execução.
	Oracle Big Data Appliance (comercial)	✓ Processamento em tempo real; ✓ Utiliza *Cloudera* do *Apache Hadoop*, *MapReduce*, *Oracle NoSQL Database*, *Cloudera Impala* e *Cloudera Search* para satisfazer os requisitos destes sistemas;

		✓ Escalabilidade linear; ✓ Processamento paralelo.
Armazenamento 	*Hive* (*Cloudera Virtual Machine CDH*)	✓ Armazenamento dos dados no *Hadoop* HDFS; ✓ Tabelas com partições; ✓ Utiliza algoritmos *MapReduce*; ✓ Armazenamento de dados em ficheiros; ✓ Usa o *MySQL* como base de dados dos metadados.
	MySQL	✓ Armazenamento estruturado dos dados; ✓ Integrada com o Hadoop;
	Hadoop HDFS (Cloudera Virtual Machine CDH)	✓ Escalabilidade horizontal rápida; ✓ Processamento paralelo; ✓ Armazenamento de grandes quantidades com baixo custo; ✓ Suporte de dados não-estruturados e semiestruturados; ✓ Utiliza o *MapReduce* para organização dos dados.
	NoSQL	✓ Não relacional; ✓ Distribuído; ✓ Escalabilidade Horizontal; ✓ Modelo de dados mais flexível;
Visualização 	*Tableau* (comercial)	✓ Suporte a grandes quantidades de dados; ✓ Rápido e fácil; ✓ Interativo; ✓ Novas formas de visualização
	Qlik View Sense Desktop	✓ Rápido e fácil carregamento dos dados; ✓ Interativo; ✓ Intuitivo.
	Pentaho Community Edition	✓ Análise interativa; ✓ Detalhe; ✓ Relatórios interativos para grandes volumes de dados; ✓ *Dashboards* a partir de grandes quantidades de dados; ✓ Escalabilidade da memória para uma rápida análise.
	Oracle Business Intelligence (comercial)	✓ Decisões em tempo real; ✓ Escalabilidade.

2.5.2. Requisitos *vs* ferramentas

Após a pesquisa e análise das características e funcionalidades das ferramentas é possível fazer a sua comparação. Note-se que esta análise foi feita a partir da informação disponível e não da sua experimentação.

Podemos verificar que, na sua maioria, correspondem a um dos grandes requisitos de sistemas *Big Data* no que diz respeito ao processamento distribuído de grandes quantidades de dados e o seu armazenamento. Na seguinte matriz (Tabela 3) podemos ver a comparação entre as ferramentas, segundo as etapas em que se aplicam, e os requisitos principais para o correto funcionamento de um sistema deste tipo.

Tabela 3 – Requisitos *vs* ferramentas

Ferramentas/Requisitos	**Suporte a dados estruturados e não-estruturados**	**Processamento distribuído, paralelo**	**Escalabilidade linear e horizontal**
Aquisição			
Flume	**X**	**X**	**X**
Knime	**X**		
Processamento			
Knime	**X**	**X**	**X**
Hive	**X**	**X**	
Pentaho	**X**	**X**	
Oracle Big Data Appliance	**X**		
Armazenamento			
Hive	**X**	**X**	**X**
MySQL			**X**
Hadoop HDFS	**X**	**X**	
NoSQL		**X**	
Visualização			
Tableau	**X**		**X**
Qlik View Sense Desktop	**X**		**X**
Pentaho	**X**		**X**
Oracle Business Intelligence			**X**

Verifica-se assim uma grande capacidade de diversas tecnologias que suportam todas as necessidades de um sistema *Big Data* para que este possa ser rápido e eficiente.

Através da matriz pode-se concluir que, na sua maioria, as ferramentas permitem uma escalabilidade linear ou horizontal que tornará o sistema e a análise dos dados mais rápida. Sendo estes dados em grandes quantidades e, por vezes sem forma ou estrutura, as ferramentas permitem também processamento distribuído ou paralelo para que os dados sejam processados ao mesmo tempo em diversos dispositivos.

No ponto 3.2 encontra-se a escolha das ferramentas selecionadas bem como a sua função neste projeto.

Capítulo 3 - *Big City Data* – Uma nova dimensão sobre as cidades

O objetivo deste capítulo é fazer uma descrição de todo o processo de desenvolvimento bem como as decisões tomadas para a realização do mesmo, nomeadamente ferramentas, fontes de dados e tipos de análise a realizar. Note-se que apenas será representado o processo para uma fonte de dados devido à extensão do processo. É aqui também resumida a plataforma mobi.me, referida no enquadramento.

3.1. O sistema *Big City Data* e o mobi.me

Como já referido no enquadramento, o desenvolvimento desta dissertação de mestrado está inserida no contexto do CEIIA, mais propriamente na plataforma mobi.me.

A plataforma mobi.me é constituída por diferentes sistemas responsáveis por gerir cada área do negócio. Na Figura 5 apresenta-se a arquitetura atual desta plataforma:

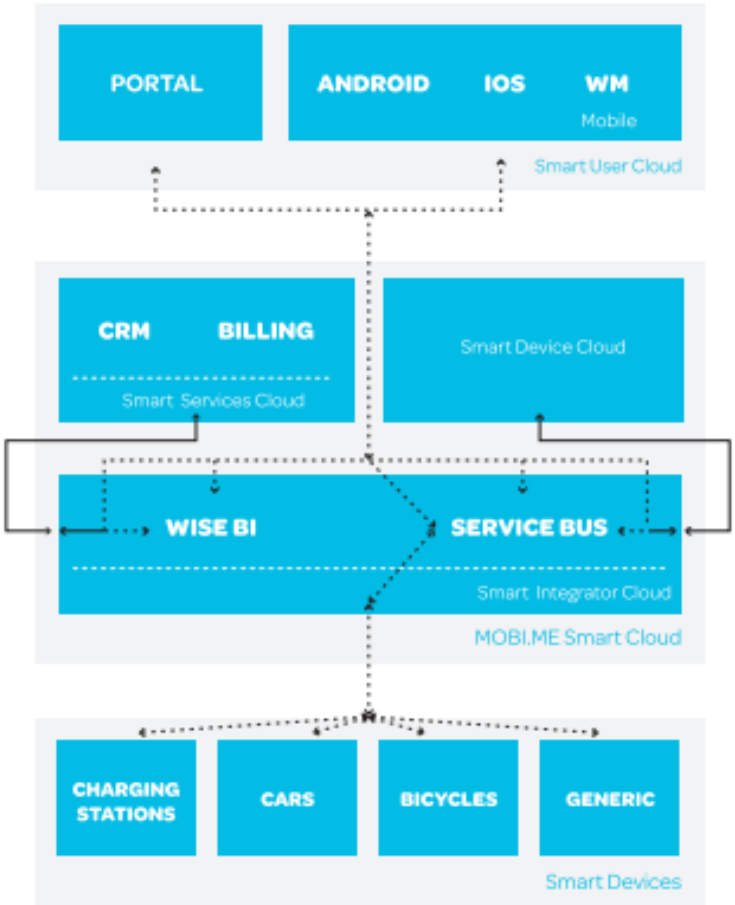


Figura 5 - Arquitetura da plataforma mobi.me (CEIIA, 2013)

Esta arquitetura está dividida em três camadas principais sendo que a primeira diz respeito aos *Smart Devices*, ou seja, os dispositivos que comunicam com os sistemas embebidos na plataforma mobi.me. Numa segunda camada denominada *Smart Coud* estão contidos os sistemas onde é feita toda a gestão da informação. A última camada, *Smart User Cloud*, representa os sistemas destinados ao utilizador final.

O primeiro sistema denomina-se *Smart Devices Cloud* (SDC) que é responsável por gerir todos os *smart devices* (postos de carregamento, veículos), as suas configurações, comunicações, alarmes e manutenções, agregando toda a informação operacional. O principal objetivo deste sistema prende-se com a necessidade de abranger o conceito geral de um *smart device*, permitindo uma escalabilidade fácil e intuitiva dos mesmos, e ao mesmo tempo diminuir o impacto no modelo de dados existente. Os *smart devices* podem ser de qualquer um dos tipos embebidos, desde que estejam ligados à rede e consigam enviar e receber pedidos ou eventos ao SDC. Na Figura 6 e Figura 7 está representado esse mesmo sistema:

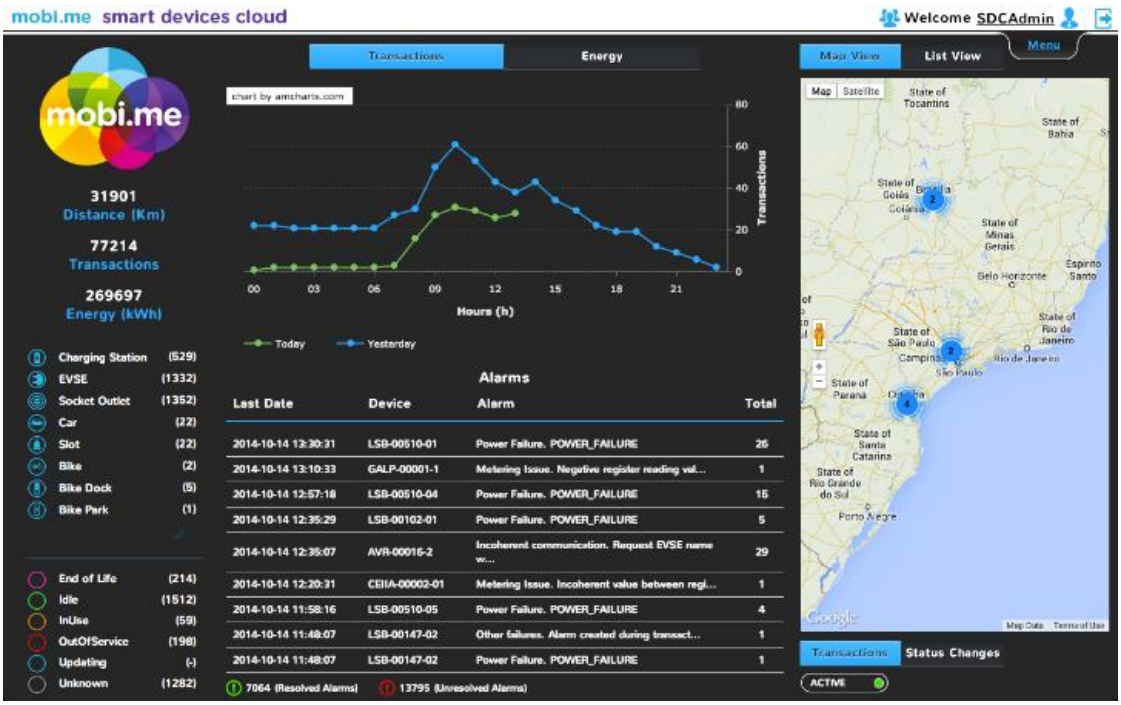


Figura 6 - mobi.me *Smart Devices Cloud* (SDC)

Figura 7 – mobi.me *Smart Devices Cloud* (SDC)

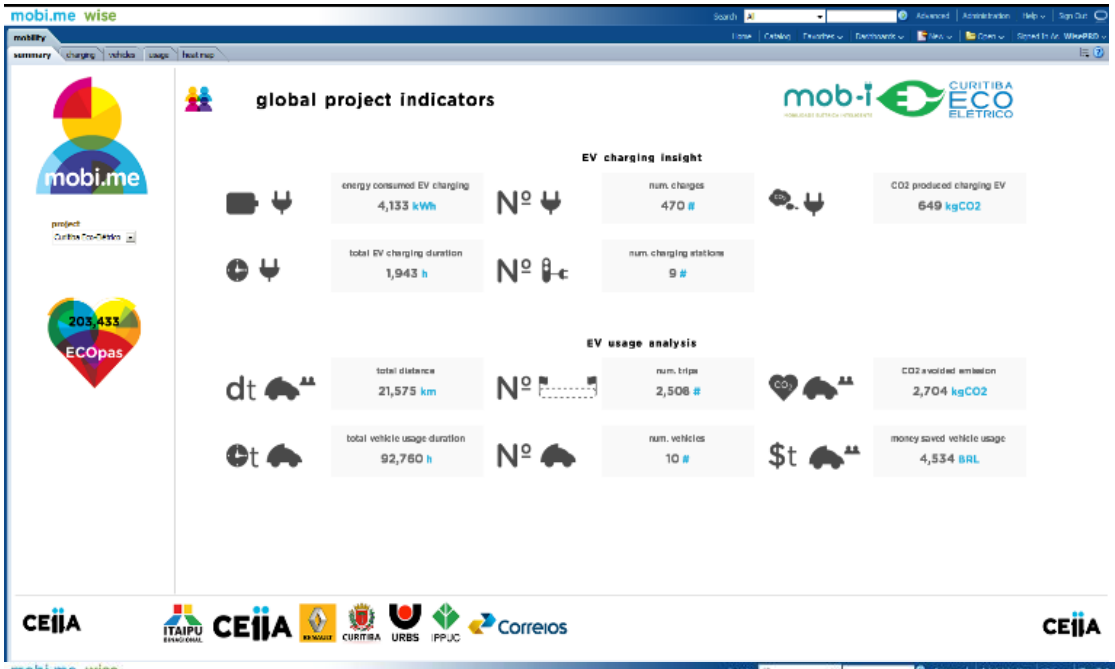
O SDC permite aos operadores aceder a toda a informação desta rede bem como definir diferentes perfis de utilizador e diferentes permissões de acesso para cada área de atuação.

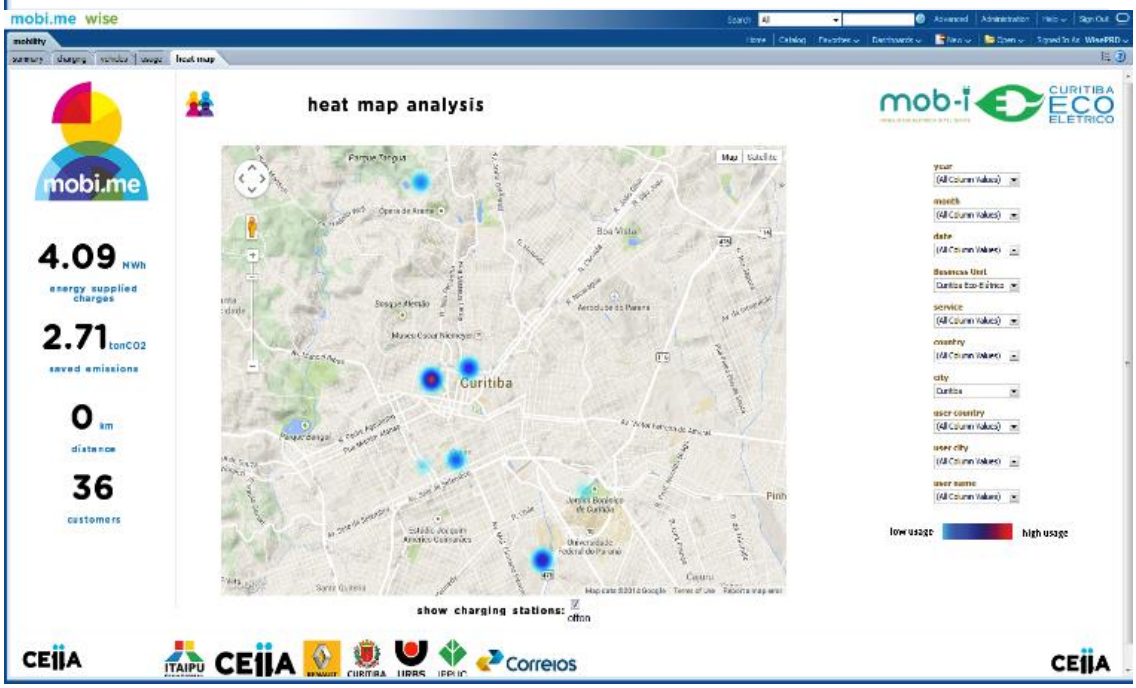
Um outro sistema essencial denomina-se *Smart Service Cloud* (SSC) que é responsável por agregar toda a informação relacionada com a complexidade do negócio e do serviço. Inclui funções de gestão de clientes, faturação e pagamentos. Do ponto de vista funcional, inclui componentes de *Customer Relationship Management* (CRM) e *Billing* (faturação) para toda a plataforma.

Além disto inclui processos como gestão de contratos, transações, notificações, reservas, produtos e serviços, tarifas, encomendas, faturação, métodos de pagamento e campanhas de marketing.

Assim, o principal objetivo do SSC é antecipar as necessidades dos clientes através da informação disponibilizada e da interação com a plataforma mobi.me.

De forma a integrar toda a informação que provém dos dois sistemas acima referidos surge o sistema mobi.me WISE. O objetivo deste sistema passa por projetar indicadores técnicos e de negócio relevantes para os operadores de negócio, para os municípios e outros *stakeholders*. Na Figura 8 apresenta-se um exemplo deste sistema:


mobi.me wise
global project indicators
mob-i CURITIBA ECO ELETRICO
EV charging insight
energy consumed EV charging
4,133 kWh
num. charges
470 #
CO2 produced charging EV
649 kgCO2
total EV charging duration
1,943 h
num. charging stations
9 #
EV usage analysis
total distance
21,575 km
num. trips
2,508 #
CO2 avoided emission
2,704 kgCO2
total vehicle usage duration
92,760 h
num. vehicles
10 #
money saved vehicle usage
4,534 BRL
203,433
ECOpas
CEIIA
ITAIPU
CURITIBA
URBS
IPPUC
Correios


mobi.me wise
heat map analysis
mob-i CURITIBA ECO ELETRICO
4.09 MWh
energy supplied charges
2.71 tonCO2
saved emissions
0 km
distance
36
customers
Curitiba
show charging stations:
low usage
high usage
CEIIA
ITAIPU
CURITIBA
URBS
IPPUC
Correios

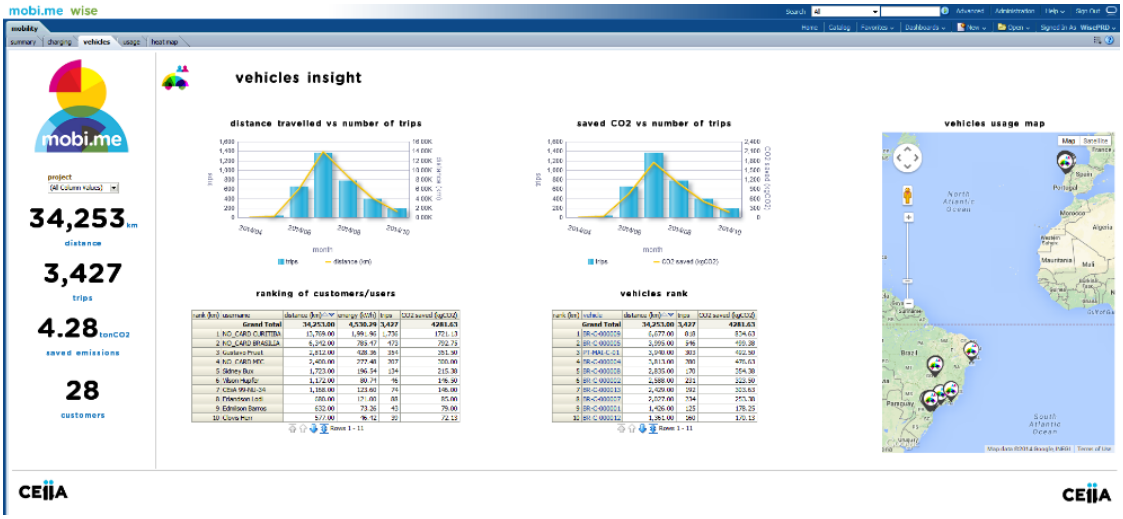


Figura 8 – MOBI.ME WISE *Business Intelligence Solution*

A utilização deste sistema possibilita analisar graficamente indicadores como consumos, quilómetros percorridos, quantidade de dióxido de carbono poupado, a distribuição geográfica de todos os *smart devices*, entre outros indicadores de negócio importantes para a gestão desta rede.

Como camada intermédia de ligação entre os *smart devices* e os sistemas referidos surge o sistema *Service Bus*. Este sistema é responsável por fazer comunicar corretamente toda a informação gerada pelos dispositivos com cada sistema, para que cada um trate da informação adequada.

Por último, foi criado o sistema *Portal and Mobile Apps* como um interface para o utilizador final através de um portal WEB e de aplicações móveis (IOS, Android). Na Figura 9 é apresentado um exemplo desse sistema:

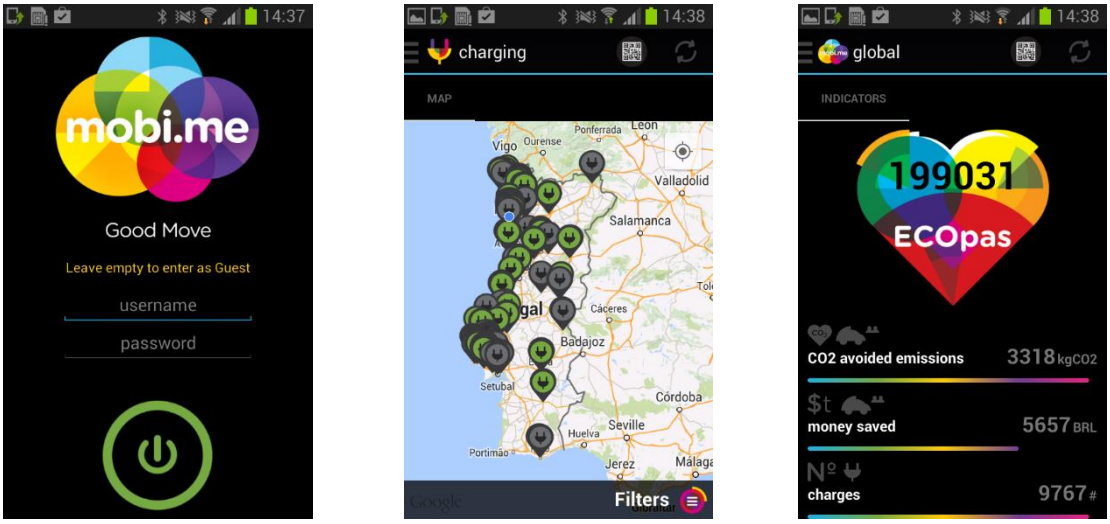


Figura 9 - MOBI.ME Portal and Mobile Apps

A partir do portal WEB e das aplicações móveis o utilizador poderá, através do seu *login* e consoante as suas permissões, consultar os *smart devices* disponíveis (postos de carregamento), gerir as suas reservas (criar ou editar), aceder aos dados da sua conta, aceder ao histórico de informação, comunicar anomalias e receber mensagens.

A plataforma mobi.me tem vindo a ser continuamente desenvolvida para que possa continuar a comunicar em tempo real com uma série de dispositivos físicos (veículos, medidores inteligentes, postos de carregamento, parquímetros, portagens e sistemas de controlo de tráfego), fornecendo uma monitorização avançada.

Além da agregação técnica de toda a informação originada pelos *smart devices*, pretende-se que seja possível identificar todos os utilizadores do sistema para que possa existir uma visão integrada de todas as interações com a rede.

A verdade é que esta plataforma e todos os seus sistemas têm crescido constantemente devido à evolução que tem havido nesta área.

Além da gestão desta rede em Portugal, existe neste momento um projeto no Brasil, em parceria com a Itaipu Binacional. Este projeto visa a implementação de uma rede inteligente de mobilidade elétrica, monitorizada pela plataforma mobi.me.

Existe assim a necessidade de analisar novas fontes de dados bem como novos pontos de atuação. Posto isto, surgiu a oportunidade de implementar um sistema *Big Data* com os seguintes objetivos:

- Obter e armazenar dados de fontes como *Facebook*, *Twitter*, *Instagram* e *Google News*;
- Armazenar e processar esses dados de forma a tratá-los, torná-los estruturados e passíveis de análise;
- Usar técnicas de análise como o *Sentiment Analysis*, *Text Mining* e *Text Processing* de forma a perceber o que os utilizadores falam acerca desta área;
- Visualizar os resultados em Relatórios e *Dashboards* através de infografias e outros métodos de visualização.

Desta forma, será possível visualizar onde e o que é falado nestas redes acerca de um conjunto de palavras-chave dentro da temática da mobilidade, de forma a enriquecer a análise do negócio bem como analisar novos pontos de atuação e novos produtos.

Da mesma forma, surge também a oportunidade de utilizar novas formas de visualização geográfica da informação gerada pelos *smart devices*. Desta forma, decorreu em paralelo a este projeto, um outro no mesmo contexto, que além de projetar esta informação irá suportar a visualização geográfica dos dados provenientes da extração realizada neste projeto.

Na Figura 10 apresenta-se um enquadramento deste sistema na plataforma mobi.me:

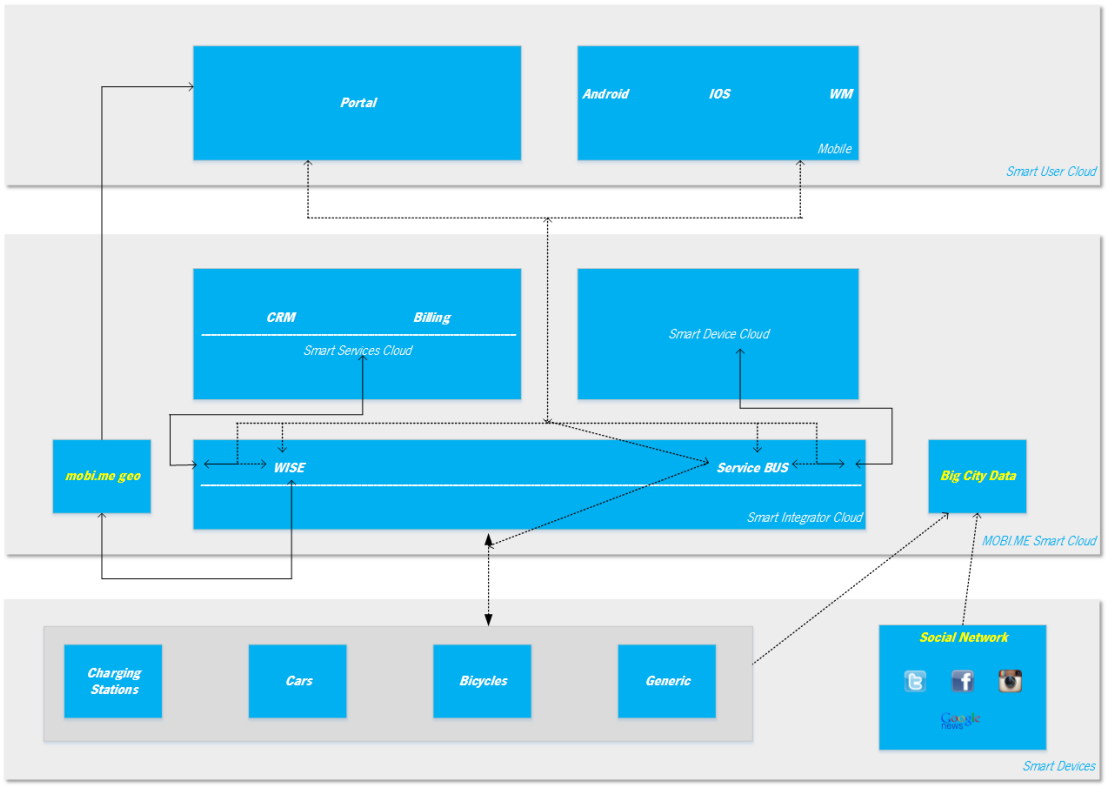


Figura 10 - Enquadramento do sistema Big City Data na arquitetura mobi.me

Como se pode verificar, será acrescentada uma nova fonte de dados (*Social Network*) aos *smart devices*, que irá ser processado e armazenado no sistema *Big City Data*. No entanto, este sistema irá também ter como fonte os atuais *smart devices* já referidos acima.

3.2. Arquitetura do sistema

Após toda a análise feita às arquiteturas *Big Data* existentes e analisando a arquitetura apresentada da plataforma mobi.me, segue-se a proposta de arquitetura para o sistema *Big City Data* (Figura 11).

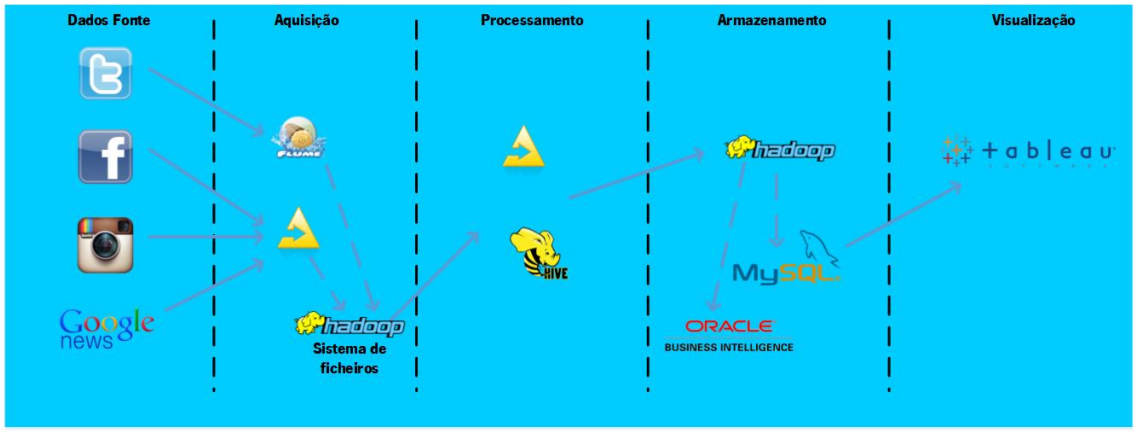


Figura 11 - Arquitetura do sistema Big City Data

Como já mencionado, um dos objetivos deste projeto passa por extrair dados de novas fontes. Assim, foram retirados dados das redes sociais (*Facebook, Twitter, Instagram e Google News*) de forma a perceber o que os utilizadores destas redes partilham e qual a sua opinião acerca de temas relacionados com as *smart cities*, com o CEIIA, com a plataforma mobi.me, com mobilidade inteligente, entre outros dentro desta área ou de outras que se venham a tornar pertinentes.

Antes de utilizar qualquer ferramenta de extração foi necessário o registo em cada uma das redes sociais com uma conta de *developer* de forma a obter os códigos de acesso aos dados dos utilizadores. Note-se que todos os acessos são feitos através das suas API's (*Application Programmer Interface*).

Para extrair estes dados foram selecionadas duas ferramentas distintas: o Flume e o Knime.

O Flume é um serviço capaz de movimentar grandes quantidades de dados através de *streaming* usando um cliente, uma fonte, um ou vários canais e uma sincronização. O fluxo de dados desta ferramenta inicia-se na definição de um cliente que irá transmitir um evento para um agente. Este agente é constituído pela fonte dos dados, que receberá o pedido do evento e de seguida o distribui por um ou mais canais. Estes canais são então sincronizados e inicia-se o processo de *streaming* dos dados.

O Knime, para este processo de extração, usa uma ferramenta chamada Palladian. Esta ferramenta é baseada em Java e fornece uma funcionalidade de pesquisa na Web e a respetiva extração dessa informação. Para este processo é necessário colocar os códigos de acesso às fontes nas preferências da ferramenta (Figura 12).

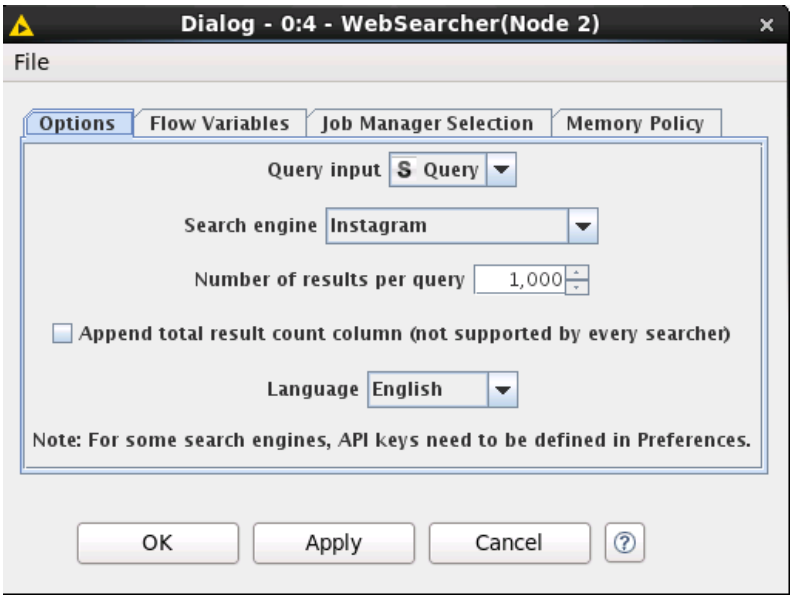


Figura 12 - Configuração do *WebSearcher*

Para esta extração são preenchidos os campos com a fonte de dados pretendida, o número de linhas a extrair e qual a sua língua.

A vantagem de utilizar o *Flume* centra-se na quantidade de informação, acerca de determinada publicação, que este trás. Os dados chegam em formato JSON (*JavaScript Oriented Notation*) com toda a informação disponibilizada pelo utilizador (Figura 13). Já o *Knime* devolve a informação apenas com os campos definidos pela ferramenta.


```
{"filter_level":"medium","retweeted":false,"in_reply_to_screen_name":null,"possibly_sensitive":false,"truncated":false,"lang":"tl","in_reply_to
{"w":340,"resize":"fit","h":226},"medium":{"w":600,"resize":"fit","h":399},"large":{"w":1024,"resize":"fit","h":682}},"id":517928422077919233,"
"expanded_url":"http://twitter.com/Amlin_Aguri/status/517928422455377920/photo/1","source_status_id_str":"517928422455377920","indices":[139,14
"url":"http://t.co/tR7d2rbema"}]},"in_reply_to_user_id_str":null,"timestamp_ms":"1412319669003","in_reply_to_status_id":null,"created_at":"Fri
"text":"RT @Amlin_Aguri: Takuma Sato at the Beijing Formula E race. #Formulae #ePrix @TakumaSatoRacer #GoTakuGo #ElectricCars #DriveTheFuture h
```


Figura 13 - Exemplo do ficheiro JSON extraído pelo *Flume*

O passo seguinte passa por processar os dados recebidos usando o *Hive* e o *Knime*.

O *Hive* é uma ferramenta que facilita a gestão e a consulta de bases de dados de grande dimensão permitindo assim utilizar técnicas de ETL (*Extract, Transform, Load*), estruturar os dados em vários

formatos, aceder a dados em ficheiros armazenados no *Hadoop* e executar consultas usando *MapReduce*.

Esta ferramenta usa uma linguagem idêntica ao SQL (*Structured Query Language*) mas denomina-se neste caso de *Hive Query Language* (HQL). Esta linguagem permite fazer seleção, projeção, junção, agregação e união dos dados bem como a criação de tabelas serializadas ou particionadas. Uma outra característica desta linguagem é a sua extensibilidade no que diz respeito à transformação e agregação de colunas. Além disso, todas as tarefas executadas no *Hive* utilizam o algoritmo *MapReduce* para a leitura das linhas de entrada transformando-as em *strings* na saída (Thusoo et al., 2013). Este algoritmo consiste num modelo de processamento de grandes quantidades de dados através da definição de uma função de mapeamento onde é gerado um conjunto de pares chave-valor como intermediários. Além disto, é também definida uma função de redução que combina todas as chaves dos intermediários que existem em comum (Dean & Ghemawat, 2004).

Além disto, esta linguagem traz vantagens no que diz respeito à leitura de dados em forma de *arrays*, *maps* e *struct*, sendo uma característica essencial para a leitura dos dados acima referidos. Foi também usado o *Knime* para tratamento destes dados de forma a facilitar a leitura dos mesmos.

Para armazenar os dados, foi usado o sistema de ficheiros do *Hadoop* HDFS e o *MySQL*. Como referido anteriormente, o *Hive* usa o *Hadoop* para armazenar os dados em formato de ficheiro, devidamente particionados e organizados. Depois de processar os dados e ter sido feito o seu tratamento, foram armazenados numa base de dados *MySQL* para que fiquem organizados em forma de tabelas facilitando a sua leitura. À parte disto, foram também armazenados dados numa base de dados Oracle para que seja possível fazer uma visualização georreferenciada.

Finalmente, foi possível fazer a análise e visualização dos dados graficamente. Para isso, foi usado o *Tableau* sendo esta uma ferramenta que permite visualizações interessantes e interativas como, por exemplo, mapas, *tag cloud*, mapas de calor e infografias.

Para a realização deste projeto foi usada a máquina virtual da *Cloudera*. Esta máquina utiliza como sistema operativo uma distribuição *Linux Red Hat 6*, trazendo incorporado todos os componentes do *Hadoop*. Além disto, traz consigo o *Cloudera Manager* com a função de gerir e monitorizar os serviços destes componentes, mantendo o seu correto funcionamento.

A ferramenta *Knime*, usada para tratamento dos dados, foi também instalada nesta máquina. Já o *Tableau*, a ferramenta utilizada para a análise e visualização dos dados, foi instalado na máquina física com o sistema operativo *Windows* pois não existe ainda uma versão para distribuições *Linux*.

Na secção 3.3 serão descritos ao pormenor todos os passos seguidos em cada etapa desta arquitetura.

3.3. Desenvolvimento do protótipo do Sistema

A presente secção é constituída pela descrição de todo o desenvolvimento do sistema *Big City Data*. Serão então explicados todos os passos, desde a extração dos dados até à visualização dos resultados.

3.3.1. Fase 1 - Configuração das ferramentas

Numa primeira fase foi necessário realizar um conjunto de configurações para a utilização da máquina virtual da *Cloudera* para que esta funcionasse corretamente, nomeadamente, espaço utilizado pelos serviços e utilizadores do sistema.

Além disto, houve necessidade de configurar alguns componentes para que fosse possível a extração de dados e o seu armazenamento.

As configurações realizadas estão detalhadas nas próximas subsecções, contendo cada tarefa realizada em cada fase.

3.3.2. Fase 2 – Extração dos dados

Como referido na secção 3.2, antes de ser possível extrair os dados houve necessidade de criar aplicações nas redes sociais utilizando contas de *developer* de forma a obter os códigos de acesso aos dados. Esses códigos são fornecidos com permissões de acesso aos dados públicos mediante a autorização dos utilizadores.

Após obter os códigos, foi necessário configurar as ferramentas de extração.

A ferramenta *Flume* faz a extração dos dados utilizando um ficheiro de configuração onde são especificadas as fontes de dados, os canais de comunicação e a sua sincronização. A Figura 14 apresenta o conteúdo desse ficheiro:

Figura 14 - Configuração do Flume para extração de dados do *Twitter*

Dentro do ficheiro é definido o agente (*SocialAgent*), a fonte de dados (no caso o *Twitter*), o canal e a sincronização com o sistema de ficheiros do *Hadoop*. Depois é feita a conexão à fonte através do canal criado e dos códigos fornecidos pelo *Twitter* seguida da definição da pesquisa e extração utilizando um conjunto de palavras-chave mediante a área pretendida. No caso foram definidas palavras na área da mobilidade elétrica e do ambiente do CEIIA:

- ✓ Curitiba, Itaipu e ecomóvel – A escolha destas palavras deve-se ao facto de existir o projeto em conjunto com a cidade de Curitiba e com a Itaipu Binacional na área da mobilidade elétrica, como referido no ponto 3.1;
- ✓ Ceiia e mobi.me- Perceber o que é falado da organização e da plataforma bem como a opinião dos utilizadores acerca das mesmas;
- ✓ Carros elétricos ou *electric cars* – Perceber o que é falado no mundo acerca de carros elétricos bem como novidades acerca desta área;
- ✓ *Smart Meters* – Perceber a evolução deste conceito de medidores inteligentes e áreas de aplicação;
- ✓ *Intelligente mobility* – Perceber o que é falado acerca de mobilidade inteligente e qual a evolução que está a ter.

O Flume utiliza um ficheiro com extensão .jar que contém todo o código Java para fazer esta extração. Este ficheiro é parte integrante da livraria da ferramenta.

Os dados extraídos chegam em formato JSON e ficam armazenados numa diretoria com o caminho indicado na figura (Figura 14) dentro do sistema de ficheiros do *Hadoop*, organizados por ano, mês, dia e hora. Esta extração é feita em tempo real, começa no momento em que o agente do flume é ativado e acaba no momento em que o agente é desligado.

O processo de extração de dados na ferramenta *Knime* é feito de forma mais simples. Primeiramente, configuramos a ferramenta com os códigos de acesso às redes sociais para que a extração seja autorizada. Depois são utilizados dois *nodes* (nome dado aos objetos da ferramenta) onde são definidas as palavras-chave no primeiro e a rede social que se pretende no segundo (Figura 15).

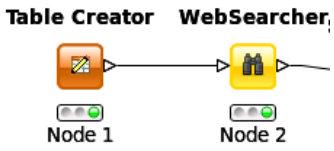


Figura 15 - Extração de dados no *Knime*

Dentro do *WebSearcher* define-se qual a rede social em que se pretende fazer a pesquisa e extração dos dados, o número de linhas e em que linguagem.

3.3.3. Fase 3 – Processamento dos dados

Para que seja possível tirar partido dos dados recebidos foi necessário fazer o seu processamento de forma a torna-los legíveis e passíveis de análise.

Como referido anteriormente os dados chegam em formato JSON com duas estruturas, ou seja, numa coleção de pares chave-valor definida como um objeto, registo ou estrutura, ou numa lista ordenada de valores em forma de *array*, vetor ou lista. Houve então necessidade de organizar estes dados numa estrutura de tabela de forma a facilitar a sua leitura. Para tal, foi usado o *Hive* executando um conjunto de *queries*. Como referido no ponto 3.2, esta ferramenta permite uma fácil leitura de dados em forma de *array*, *struct* ou listas, daí ser vantajoso no processamento dos ficheiros JSON extraídos.

Numa primeira etapa, foi criada uma tabela com todos os *tweets*, organizando-os por colunas. O exemplo da *query* executada para este efeito tem a seguinte forma (Figura 16):

```
create table twitteranalytics (
   created_at string,
   entities struct <
      hashtags:ARRAY<STRUCT<text:STRING>>>,
   geo struct <
      coordinates:ARRAY<string>,
      type:string>,
   place struct <
      name:string,
      country:string,
      full_name:string>,
   id bigint,
   source string,
   text string,
   user struct <
      name: string,
      profile_image_url_https: string,
      protected: boolean,
      screen_name: string,
      verified: boolean>
)
PARTITIONED BY (datehour INT)
    ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe'
    LOCATION '/user/hive/warehouse/bigcitydata.db/twitteranalytics';
```

Figura 16 - *Query* utilizada para organizar os dados numa tabela no Hive

Através da *query* representada na figura (Figura 16), os dados foram organizados por colunas fazendo a extração do conteúdo dos ficheiros. O resultado deste foi organizado por partições em forma de pastas nomeadas com o formato data/hora, utilizando um ficheiro com extensão .jar integrante da livraria do *Hive*, que permite este tipo de formatações. O local de armazenamento para estes resultados é predefinido na configuração da ferramenta, dentro do sistema de ficheiros do *Hadoop*. Na Figura 17 apresenta-se um exemplo dos resultados após a execução da *query.*

entities	geo
{"hashtags":[{"text":"SmartCities"}]}	{"coordinates":
{"hashtags":[{"text":"cwb"},{"text":"curitibacool"},{"text":"curitiba"},{"text":"brazil"},{"text":"landscape"},{"text":"city"},{"text":"bus"}]}	{"coordinates":
{"hashtags":[{"text":"SmartCities"},{"text":"utilities"}]}	{"coordinates":
{"hashtags":[{"text":"SmartCities"}]}	{"coordinates":

Figura 17 - Exemplo do resultado da *query*

Depois de organizados os dados houve necessidade de extrair o conteúdo dos *arrays* para facilitar a leitura dos dados. Esse processo foi feito através da criação de uma nova tabela, selecionando os dados guardados na tabela criada inicialmente (Figura 18).

```
CREATE TABLE tweetsanalise
    STORED AS TEXTFILE
    AS
    SELECT twitteranalytics.id,
            twitteranalytics.created_at,
            twitteranalytics.entities.hashtags.text AS hashtags,
            twitteranalytics.geo.coordinates[0] as latitude,
            twitteranalytics.geo.coordinates[1] as longitude,
            twitteranalytics.place.name as cidade,
            twitteranalytics.place.country as pais,
             twitteranalytics.place.full_name as localizacao,
            twitteranalytics.text,
            twitteranalytics.user.name,
            twitteranalytics.user.screen_name
            FROM twitteranalytics

            ;
```

Figura 18 - Organização dos dados

Usando esta *query*, foi possível extrair cada posição dos *arrays* transformando-os em colunas e ao mesmo tempo fazer uma limpeza do texto de forma a torná-lo legível. O resultado deste processo está representado na Figura 19.

hashtags	latitude	longitude	cidade	pais
["Arenão","Curitiba","VemComigo"]	-25.395836	-49.153272	Pinhais	Brasil
["curitiba","curitilover","primavera","sunset","sun","sky","spring"]	-25.442448	-49.24017	Curitiba	Brasil
["curitiba","Paraná","Brasil","instaboy","beard","instabeards","barba","beard","imloveit"]	-25.4167	-49.25	Curitiba	Brasil
["evento","jib","patricinador","Curitiba","fotografia"]	-25.435143	-49.260933	Curitiba	Brasil

Figura 19 - Resultado da organização dos dados

A coluna que contém as *hashtags* (palavras-chave encontradas no texto recebido relacionadas com as palavras definidas na pesquisa) encontrava-se ainda numa estrutura que não permitia fazer a sua análise separadamente. Para tratar esta situação foi utilizado então o Knime para transformar esta linha em várias linhas, mantendo-se associadas a informação restante à qual correspondia. Neste processo foram separadas as palavras usando como delimitador a vírgula. O fluxo deste processo é seguidamente apresentado (Figura 20) bem como o respetivo resultado (Figura 21):

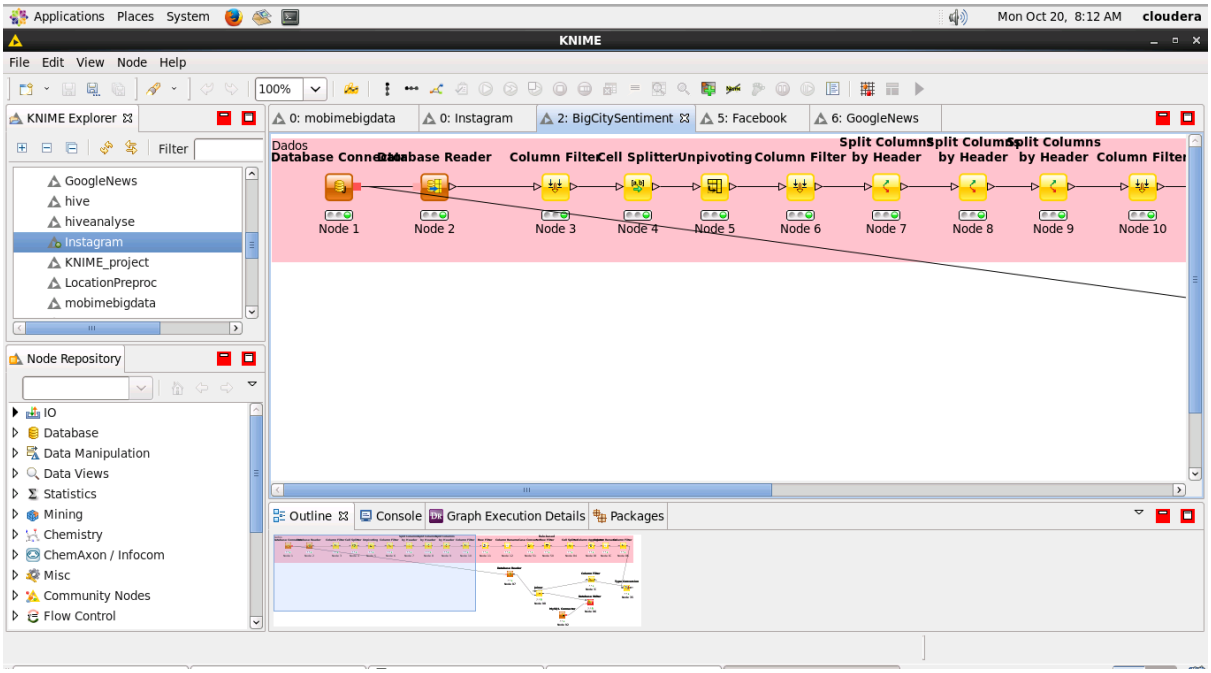


Figura 20 - Fluxo do processo de tratamento dos dados no *Knime*

Neste processo, começou-se por fazer a separação das palavras usando a vírgula como delimitador, colocando cada palavra numa coluna. Depois disto, estas colunas foram transformadas em linhas como era o objetivo. No entanto, algumas palavras continham símbolos associados sendo necessária a sua remoção.

Hashtags	text
SmartCities	Very proud to join @abbyandrietsch and @Whughes338 in @TVanderArk
cwb	□□□□#cwb #curitibacool #curitiba #brazil #landscape #city #bus http:/
curitibacool	□□□□#cwb #curitibacool #curitiba #brazil #landscape #city #bus http:/
curitiba	□□□□#cwb #curitibacool #curitiba #brazil #landscape #city #bus http:/
brazil	□□□□#cwb #curitibacool #curitiba #brazil #landscape #city #bus http:/
landscape	□□□□#cwb #curitibacool #curitiba #brazil #landscape #city #bus http:/
city	□□□□#cwb #curitibacool #curitiba #brazil #landscape #city #bus http:/
bus	□□□□#cwb #curitibacool #curitiba #brazil #landscape #city #bus http:/

Figura 21 - Resultado do *Knime* para tratamento das *hashtags*

No mesmo processo foram também tratadas as datas dos *tweets* de forma a conseguir visualizar os resultados por dia. Utilizando o mesmo raciocínio obteve-se o resultado apresentado na Figura 22.

Hashtags	Data
smartcities	Sep 23
curitiba	Sep 23
curitiba	Sep 23
curitiba	Sep 24
curitiba	Sep 24
smartcities	Sep 24
smartcities	Sep 24

Figura 22 - Resultado do *Knime* para tratamento das datas

O processo de tratamento das datas assemelha-se ao descrito acima em relação à separação das *hashtags*.

Após todo este processo de extração e tratamento dos dados, foi possível carregá-los e armazená-los. Esse passo está detalhado na próxima secção.

3.3.4. Fase 4 – Armazenamento dos dados

Nesta fase foi feito o armazenamento dos dados para posteriormente serem analisados. Como já foi dito o *Hive* guarda os resultados em forma de ficheiro, usando um delimitador para separação dos dados, no sistema de ficheiros do *Hadoop*. Assim sendo, os resultados de todas as *queries* referidas anteriormente ficaram armazenadas neste local de armazenamento.

No entanto, após ter realizado o tratamento dos dados no *Knime*, foi criada uma base de dados *MySQL* para o seu armazenamento. Dentro desta base de dados foram guardadas as tabelas com os dados do *Twitter*, *Facebook*, *Instagram* e *Google News* (Figura 23).

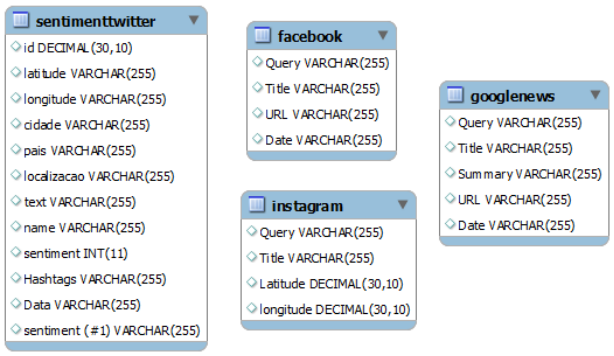


Figura 23 - Base de dados *MySQL bigcitydata*

Na Figura 23 estão representadas as tabelas criadas na base de dados *MySQL* denominada *bigcitydata*. Os campos destas tabelas correspondem à informação recebida de cada uma das fontes.

Todas estas tabelas têm em comum o campo que contem as palavras-chave definidas para a pesquisa e análise.

Em paralelo a este armazenamento, foi criada uma tabela com as coordenadas existentes nos dados, bem como as respetivas palavras-chave. Esta tabela foi guardada numa base de dados Oracle, para que fosse possível uma visualização georreferenciada no sistema que está a ser também desenvolvido no contexto de uma tese de mestrado na mesma área de estudos.

A fase seguinte diz respeito à análise feita aos dados e a visualização dos mesmos. Essa fase está descrita na próxima secção.

3.3.5. Fase 5 – Análise e Visualização dos dados

Como já foi supracitado, cada vez mais se assiste ao aumento da utilização das redes sociais para partilha de opiniões e de informação. Posto isto, as organizações usam cada vez mais estas fontes de dados para análise das opiniões e sentimentos dos seus clientes ou possíveis clientes acerca dos produtos e serviços fornecidos.

Desta forma, foi realizada uma análise dos sentimentos dos utilizadores destas redes sociais em relação às palavras-chave acima mencionadas.

Este processo foi realizado no *Hive* que através de funções próprias permite este tipo de análise, utilizando um dicionário da *MPQA Opinion Corpus* que contém palavras positivas e negativas previamente classificadas. A escolha deste dicionário prende-se ao facto de ser o mais usado neste tipo de análise.

A primeira etapa passou por criar uma tabela onde foi guardado o dicionário mencionado. De seguida foram criadas três *views*:

- Na primeira, todos os caracteres maiúsculos foram transformados em minúsculos;
- Na segunda, as palavras foram separadas por linhas;
- Na terceira, cada palavra do texto foi classificada segundo a sua polaridade comparando com as palavras do dicionário. A classificação passa por atribuir às palavras o valor -1 se forem negativas, 1 se forem positivas e 0 neutras.

Para as duas primeiras foi usada a função *lateral view explode()*. Esta função permite transformar uma coluna com palavras ou números em várias linhas.

Depois de classificadas as palavras, foi criada uma tabela onde se guardou o texto classificado através da soma da polaridade atribuída na terceira *view*, guardando o id do *tweet*. O resultado desta soma classificou o texto da seguinte forma:

- Quando a soma da polaridade fosse maior do que 0, esta seria positiva;
- Quando a soma da polaridade fosse menor do que 0, esta seria negativa;
- Quando a soma da polaridade fosse igual a 0, esta seria neutra.

Estando o texto classificado, criou-se então uma última tabela onde estão presentes todas as colunas iniciais adicionando o sentimento atribuído. Nesta tabela o sentimento encontra-se classificado com um intervalo de inteiros de 0 a 2, inclusive, sendo que 0 é negativo, 1 é neutro, e 2 é positivo. Na Figura 24 apresenta-se um exemplo do resultado dessa classificação.

	name	screen_name	sentimen
ɔd @Whughes338 in @TVanderArk's new #SmartCities book. Check it out here: http://t.co/e8GeBbDWwG	Isral DeBruin	israldebruin	2
ɔa #brazil #landscape #city #bus http://t.co/3Ns0zz48H6	Ivo Stankiewicz	ivostank	1
ın Diego, CA http://t.co/FwIEX5EvRG http://t.co/62O3GMealr #SmartCities #utilities	Veracity	veracityamg	2
ıst education system" - @TVanderArk in new book #SmartCities that Work for Everyone...	Steven Hodas	stevenhodas	2
onamentos do #TCE sobre o Edital do #Metrô de #Curitiba	Giovanni Marquesi	GCMARQ	1
ietsch & @israldebruin in @TVanderArk's new book, #SmartCities that Work for Everyone: http://t.co/H7EFTw65GF	STC Milwaukee	STCMilwaukee	1

Figura 24 - Classificação dos sentimentos dos *tweets*

Para visualizar esta análise, bem como outro tipo de análises que podem ser efetuadas nestes dados, foi utilizada a ferramenta de análise e visualização *Tableau*. Esta ferramenta preza-se pela sua qualidade, quantidade e variedade de tipos de visualizações disponíveis.

No próximo capítulo, são demonstradas todas as visualizações e análises efetuadas neste projeto.

Capitulo 4 - O sistema Big City Data

Neste capítulo o objetivo é apresentar os resultados obtidos através da análise e visualização dos dados recolhidos.

Na segunda parte do capítulo, são apresentados esses resultados graficamente através de *Dashboards*, relatórios e infografias.

No final do capítulo estão são apresentadas algumas conclusões acerca deste trabalho bem como algumas dificuldades encontradas.

4.1. Indicadores

Para terminar este processo de análise e visualização dos dados recolhidos e transformados, foram definidos indicadores de análise de forma a perceber o que é falado nas redes sociais acerca da área da mobilidade.

Para análise e visualização dos dados será seguido apenas o exemplo da rede social *Twitter*.

Sendo o foco da pesquisa entender a presença dos conceitos acima descritos na rede social, os indicadores escolhidos assumem a forma de questão. Foram assim definidos os seguintes indicadores:

- ✓ Quantas vezes por dia surgem tópicos acerca de cada palavra-chave?
- ✓ Qual a frequência em que estas palavras-chave são faladas?
- ✓ Qual a opinião ou sentimento dos utilizadores em relação a estes temas?
- ✓ Em que locais do mundo estas palavras são faladas?
- ✓ Qual o país que mais fala de determinada palavra-chave?

Além de responder a estas questões, foi construída uma infografia que permita ao utilizador combinar os diferentes atributos e aplicar filtros aos mesmos, conseguindo responder a qualquer questão.

A definição de indicadores de negócio torna-se um fator de grande importância para as organizações. O facto é que através da análise e visualização de dados utilizando estes recursos permite que seja conhecida informação de negócio extremamente importante para o processo de tomada de decisão.

Neste caso específico, torna-se uma mais-valia na medida em que permite ao CEIIA ter conhecimento da visão dos utilizadores das redes sociais acerca do serviço que presta e qual a aceitação que tem no mercado. Da mesma forma, permite conhecer a opinião e o sentimento que estes mesmos utilizadores transmitem nas suas publicações, em qualquer parte do mundo.

As opiniões e sentimentos expressos nas redes sociais têm um impacto bastante significativo nas organizações pois estes tornam-se rapidamente virais.

Neste caso específico, torna-se uma mais-valia na medida em que permite ao CEIIA ter conhecimento da visão dos utilizadores das redes sociais acerca do serviço que presta e qual a aceitação que tem no mercado. Da mesma forma, permite conhecer a opinião e o sentimento que estes mesmos utilizadores transmitem nas suas publicações, em qualquer local do mundo.

4.2. Resultados Obtidos

Para responder às questões de análise definidas na secção 4.1 foram elaboradas diversas formas de visualização, das quais *Dashboards*, infografias e relatórios. Note-se que as representações a seguir apresentadas referem-se à rede social *Twitter*, sendo que os dados extraídos são apenas uma amostra para análise, englobando apenas publicações acerca dos tópicos definidos e não todas as publicações existentes nesta rede social.

De forma a organizar esta secção, são apresentadas as questões colocadas com a respetiva imagem da visualização gráfica e a sua análise.

1. Quantas vezes por dia surgem tópicos acerca de cada palavra-chave?

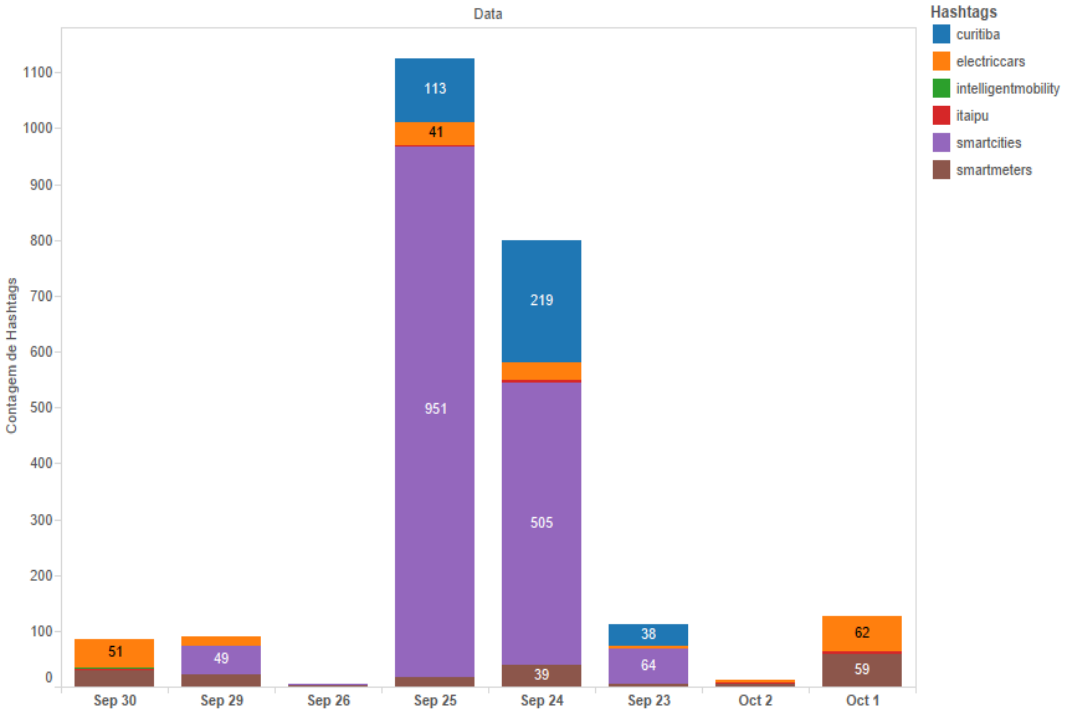


Figura 25 - Quantas vezes por dia surgem tópicos acerca de cada palavra-chave

Através do gráfico (Figura 25) verifica-se alguma variação nos assuntos falados na rede social do *Twitter* em relação às palavras-chave definidas.

Pode-se também verificar através da projeção destes dados no mapa a ocorrência de acontecimentos relacionados com estes tópicos bem como a publicação realizada (Figura 26).

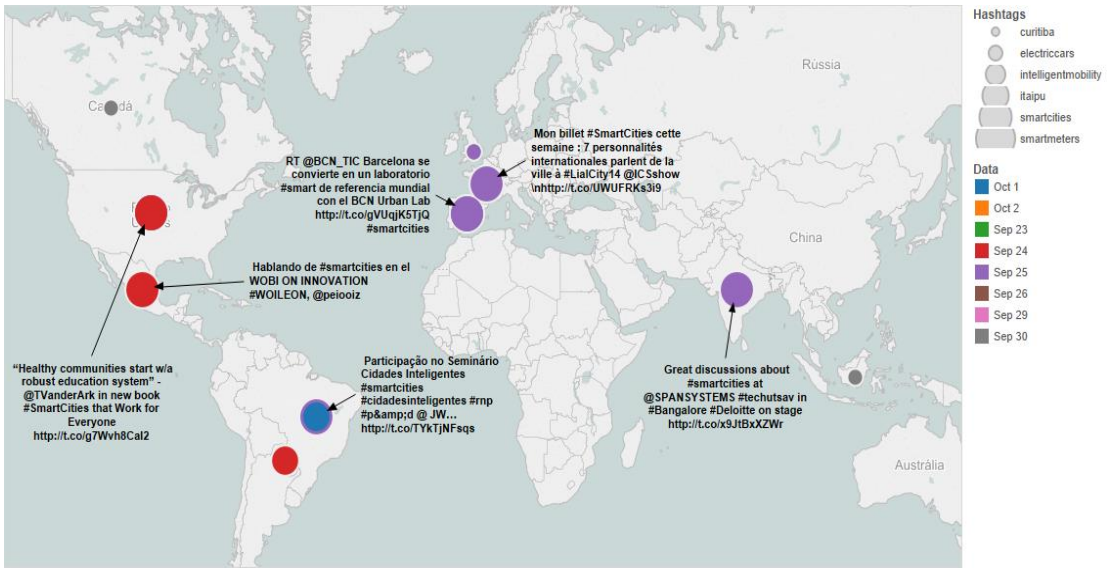


Figura 26 - Acontecimentos que levaram ao crescente número publicações relacionadas com *smart cities*

Conclui-se desta forma que nos dias 24 e 25 de setembro verifica-se um grande volume de tópicos relacionados com *smart cities* devendo-se ao acontecimento de alguns eventos centrados nesta temática. Por exemplo, no dia 25 de Setembro decorreu no Brasil um seminário onde a temática principal foi *smart cities.*

2. Qual a frequência em que estas palavras-chave são faladas?

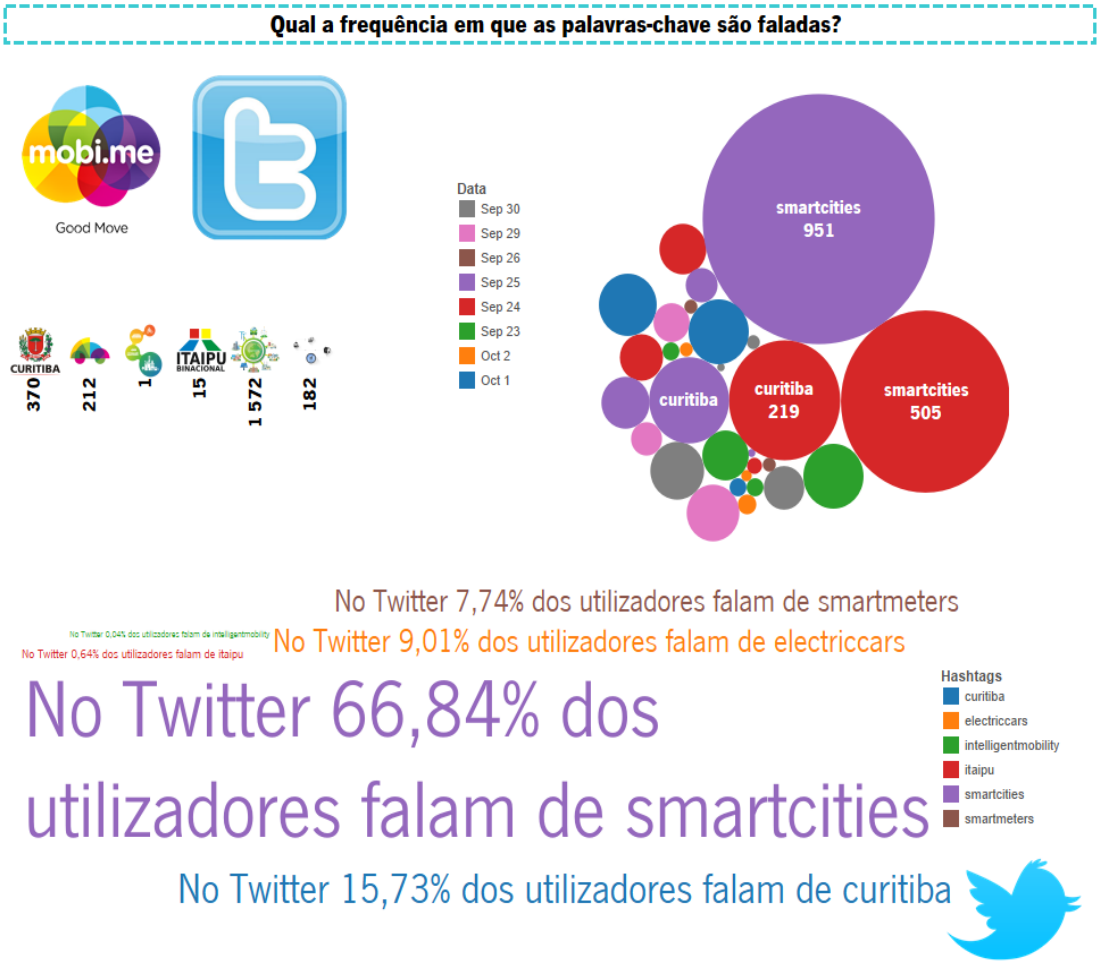


Figura 27 - Infografia: Qual a frequência em que as palavras-chave são faladas

Para responder a esta questão foi elaborada uma infografia de forma a representar vários tipos de análises. Note-se que os dados utilizados dizem respeito a publicações referentes apenas às *hashtags* representadas na legenda, por exemplo pode perceber-se que a *hashtag smart cities* tem uma maior percentagem de publicações pelo facto de ser um tema atual, bastante discutido e específico. Contrariamente, quando se procuram publicações acerca de *intelligent mobility* a percentagem é bastante menor por ser um tópico mais abrangente

O cálculo destas percentagens foi feito utilizando a função disponibilizada no *Tableau*, que permite calcular a percentagem de cada ocorrência ao longo da tabela.

Analisando os gráficos pode-se então verificar que a palavra-chave mais popular é *smart cities* representando 66,84% das publicações, ou seja, esta surge em 1572 das publicações extraídas.

3. Qual a opinião ou sentimento dos utilizadores em relação a estes temas?

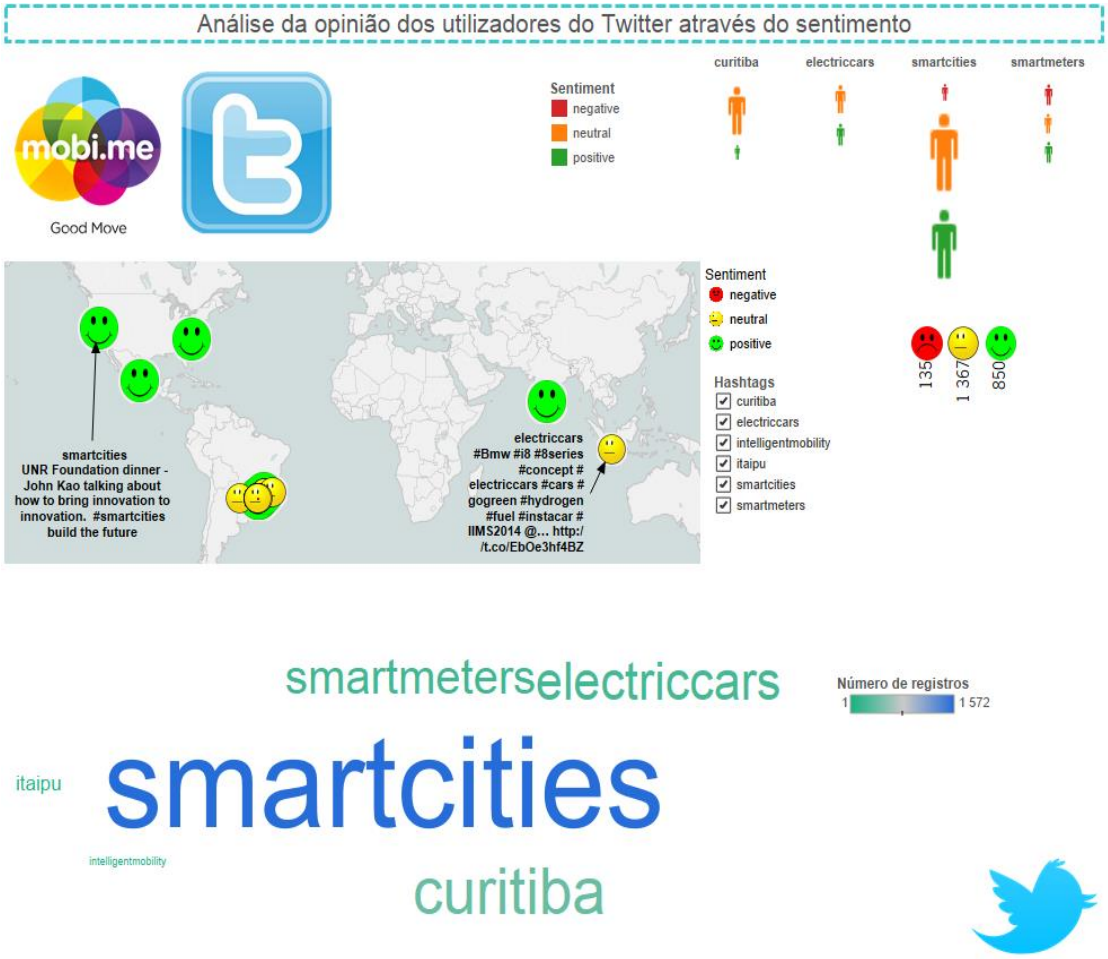


Figura 28 - Análise do sentimento dos utilizadores do *Twitter* relativamente às palavras-chave

Após a classificação do sentimento dos utilizadores do *Twitter* referido na subsecção 3.3.5, foi possível concluir que as opiniões dos utilizadores em relação a estes tópicos são na sua maioria vagas. No entanto, verifica-se algum positivismo em relação ao aparecimento de temas como *smart cities* e *electric cars*.

Na Figura 29 aparecem os totais representados para cada sentimento relativamente a cada palavra-chave.

	Hashtags					
Sentiment (#1)	curitiba	electriccars	intelligentmobility	itaipu	smartcities	smartmeters
negative	9	23		1	39	63
neutral	328	118	1	12	852	56
positive	33	71		2	681	63

Figura 29 - Total de cada sentimento por *hashtag*

Note-se que nem todas as opiniões estão representadas no mapa presente na infografia (Figura 28) devido ao facto de que nem sempre o utilizador disponibiliza a sua localização no momento em que publica o respetivo texto.

4. Em que locais do mundo estas palavras são faladas?

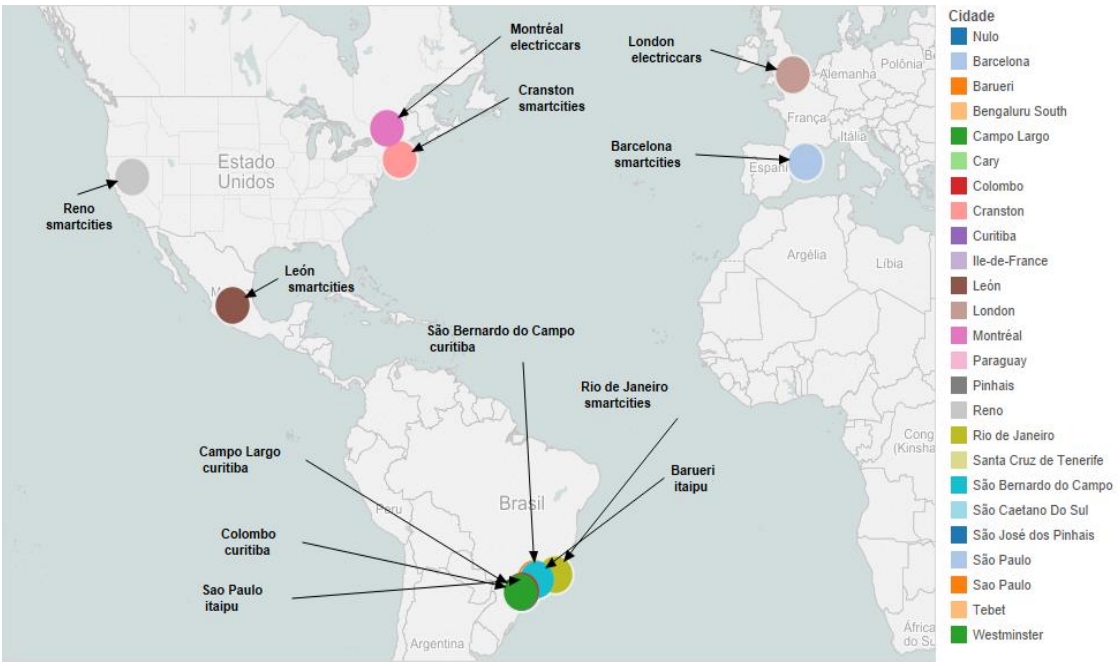


Figura 30 - Distribuição das palavras-chave pelo mundo

Na Figura 30 é apresentado um mapa com a localização das palavras-chave pelo mundo. O facto é que tanto as *smart cities* como os *electric cars* são um assunto recente, no entanto são já temas de debate em várias partes do mundo. Verifica-se, por exemplo, em Barcelona a presença deste tema devido ao projeto de implementação de tecnologias que tornaram já esta cidade numa *smart city*.

5. Qual o país que mais fala de determinada palavra-chave?

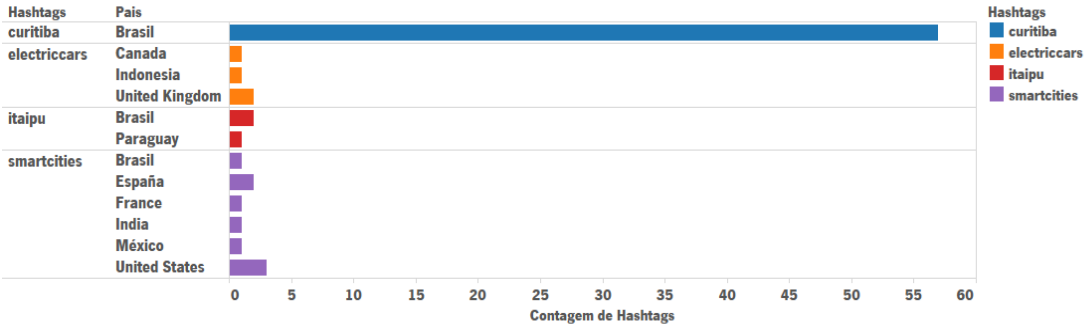


Figura 31 - Presença das palavras-chave por país

A frequência em que é falado cada tópico apresenta-se na Figura 31. As publicações relativas à cidade de Curitiba aparecem na sua maioria no Brasil. Já as publicações relativas a *smart cities* aparecem como tópico principal no Brasil, na Espanha, na França e nos Estados Unidos. Este último deve-se ao acontecimento de eventos, como já referido na descrição da Figura 25 e Figura 26, e à implementação deste conceito em diversas cidades destes países.

Relativamente ao tópico de *electric cars*, verifica-se uma maior frequência no Canadá, na Inglaterra e na Indonésia. Este facto deve-se ao recente envolvimento destes países em projetos de implementação deste tipo de veículos em algumas das suas cidades.

Além de responder a cada uma das questões, o utilizador pode combinar todos os atributos disponíveis de forma a responder às questões que pretende analisar. Na Figura 32 apresenta-se uma infografia com as diversas formas de análise e visualização, com os respetivos filtros que podem ser aplicados.

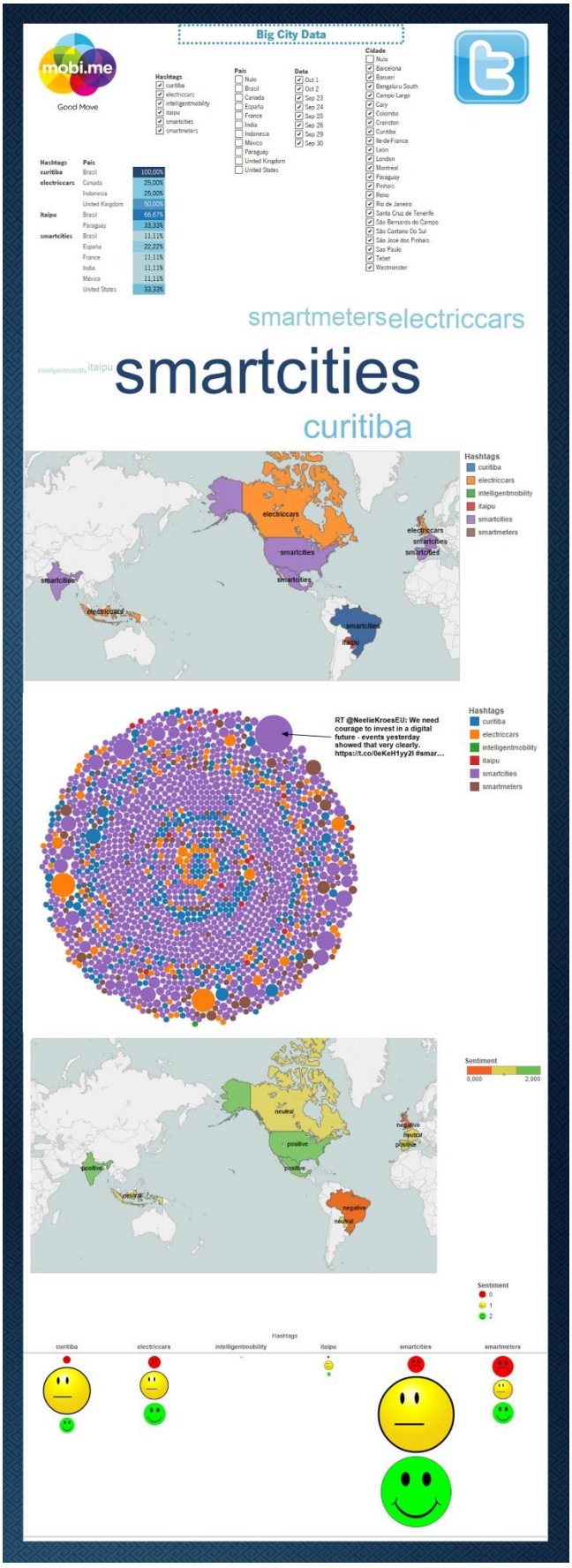


Figura 32 - Combinação de todos os atributos disponíveis para análise e visualização

Com esta infografia é possível para o utilizador perceber, por exemplo, que *hashtag* foi mais debatida num determinado dia e num determinado país, selecionando o dia que pretende nos filtros correspondentes.

4.3. Avaliação dos resultados

Depois de apresentar os resultados obtidos, é importante analisar alguns aspetos importantes de todo este trabalho.

Todo o estudo e desenvolvimento desta tese de mestrado, bem como do protótipo do sistema *Big City Data,* trouxeram resultados bastante interessantes.

Uma grande vantagem deste projeto foi o facto de conseguir trazer novos dados, de novas fontes, com uma perspetiva diferente dos dados existentes neste momento na plataforma mobi.me. Estes dados permitiram perceber o que é falado neste momento no mundo, acerca de temas que são importantes para a evolução desta plataforma, e como é que as pessoas aceitam a evolução das cidades em que vivem.

Um outro aspeto importante é conseguir perceber a evolução do conceito de carros elétricos no nosso país e no mundo. O facto é que este tipo de veículos tem crescido substancialmente, pelo facto de haver cada vez mais preocupação por parte das cidades com o desenvolvimento sustentável, com a qualidade de vida das sociedades e com a diminuição da poluição.

A nível tecnológico foi possível retirar que a era das *Smart Cities* está em constante evolução, e cada vez mais serão implementadas tecnologias nas cidades de forma a conseguir tornar a vida dos cidadãos melhor e com mais recursos de apoio às suas atividades diárias.

Consequentemente, os sistemas *Big Data* tornam-se aliados a esta evolução, permitindo suportar os dados que daí advêm, bem como poder analisar esses dados de diversas formas. Seguindo o exemplo usado no sistema apresentado, a utilização de sistemas deste tipo permitem às cidades perceber a opinião dos cidadãos, qual a sua satisfação em relação aos serviços disponibilizados tendo sempre a preocupação e a oportunidade de conseguir melhorar o seu ambiente.

Capitulo 5 – Conclusões

Este capítulo tem como objetivo apresentar as conclusões relativas ao global do projeto, nomeadamente, o trabalho realizado desde a recolha bibliográfica até ao desenvolvimento do sistema, as principais dificuldades sentidas, as limitações do projeto e, por último, as propostas de melhoria e implementações futuras.

5.1. Trabalho realizado

Neste projeto foi desenhado e implementado um protótipo de um sistema *Big Data* com o objetivo de adquirir dados de novas fontes para a plataforma mobi.me, bem como novas formas de análise e visualização dos mesmos.

Para tal, e numa primeira fase, foram definidos todos os objetivos para o desenvolvimento do projeto bem como os resultados esperados, tendo em conta a finalidade do projeto.

Depois, foi necessário todo um trabalho de pesquisa e revisão de literatura acerca do *Big Data*, incluindo conceitos propostos por diversos autores, visto ser ainda um tema recente e muito discutido. Neste contexto foram também descritas as características necessárias para que seja implementado com sucesso e para que corresponda aos requisitos das diferentes arquiteturas. Aqui destaca-se a necessidade da velocidade no processamento de dados em que estes têm uma grande variedade de formatos.

Além disso, foram revistos também os conceitos de *Smart Cities*, as suas características, serviços que disponibilizam e a necessidade de utilizar sistemas *Big Data* no seu contexto. Para tal foram também analisados casos de implementação, integrando estes conceitos.

Para o desenvolvimento dos objetivos houve também necessidade de fazer um levantamento do estado da arte do conceito de *Sentiment Analysis*, de forma a perceber como este é aplicado e quais as técnicas usadas para que seja feita uma análise correta do texto. Neste seguimento, foram também descritos os problemas subjacentes a este tipo de análise.

Terminando o enquadramento concetual, foram estudadas e descritas as arquiteturas disponíveis para sistemas *Big Data*, de forma a perceber que os seus requisitos.

Por último foi feito um levantamento do mercado das tecnologias, *open source* e comerciais, que cumprem com os requisitos para o desenvolvimento destes sistemas. Para tal, foi feito um resumo

das suas características e depois elaborada uma matriz de comparação com os requisitos principais necessários para a extração, processamento, armazenamento e visualização dos dados.

Depois disto foi possível iniciar o desenvolvimento de todas as etapas deste sistema. Numa primeira fase, foram feitos testes às ferramentas escolhidas de forma a perceber se correspondiam às necessidades. De seguida foram extraídos os dados, analisando as suas características (forma, tamanho, tipo) para que fosse feito um correto tratamento dos mesmos.

A fase seguinte passou pelo processamento dos dados recebidos de forma a torná-los mais legíveis, visto que estes chegavam em ficheiros em formato JSON. Depois de processados foi possível fazer análises como o sentimento e opinião dos utilizadores em relação a um conjunto de tópicos.

Por último, foram elaboradas diferentes formas de visualização como infografias, *Dashboards* e relatórios, de forma a responder a um conjunto de indicadores de análise

A implementação deste tipo de sistemas trazem inúmeras vantagens para as organizações, permitindo a aquisição de conhecimento acerca dos seus clientes, que através da informação tradicional não é possível. Torna-se assim possível uma análise do negócio e do mercado muito mais abrangente e com novas perspetivas.

De uma forma geral, e tendo em conta os objetivos inicialmente definidos e os resultados obtidos, conclui-se que o projeto terminou com sucesso.

5.2. Dificuldades e limitações

O desenvolvimento deste projeto teve algumas limitações, na sua maioria pela complexidade de configuração de ferramentas e pela dificuldade de acesso aos dados.

No que diz respeito às ferramentas, a maior dificuldade foi na configuração dos componentes da máquina virtual utilizada. Isto porque, na sua maioria, as configurações eram feitas em ficheiros dentro das suas pastas de instalação onde eram requeridas permissões e onde existiam dependências.

No que respeita ao acesso aos dados, a dificuldade passou principalmente pela sua extração e pela obtenção de permissões para a sua aquisição.

Por último, houve dificuldade na aplicação de um dicionário na análise de sentimentos. Isto porque a utilização de um dicionário neste tipo de análise nem sempre é capaz de identificar o sentido verdadeiramente dado ao texto.

5.3. Trabalho futuro

Apesar de os objetivos terem sido cumpridos e sendo este um protótipo de sistema, existem ainda várias melhorias a ser feitas para que seja cada vez mais completo e valioso.

Como trabalho futuro seria interessante fazer um dicionário para análise de sentimentos aplicado a esta área, para que os resultados sejam mais específicos e melhor classificados. A construção de um dicionário aplicado ao contexto traz vantagens ainda maiores ao negócio.

Uma outra vantagem seria habilitar este sistema com uma vertente de análise dos telefonemas recebidos quando existem problemas com os postos de carregamento ou com outros *smart devices* de forma a perceber o problema do cliente, o seu grau de satisfação/insatisfação recolhendo *feedback* importante de forma a contextualizar dados gerados pelas plataformas operacionais. Esta análise tem como objetivo melhorar o atendimento aos clientes bem como ajudar os operadores da rede a suportar, cada vez melhor, problemas que possam surgir.

Uma última proposta tem a ver com a integração das aplicações móveis com as redes sociais. Um cliente que use a aplicação, e que a conecte com a rede social preferida, poderá fornecer a este sistema novos dados, aplicados mais especificamente a este contexto. Os dados que poderão ser adquiridos daqui possibilitam novas perspetivas de negócio.

Bibliografia

Asur, S., & Huberman, B. a. (2010). Predicting the Future with Social Media. Computers and Society; Physics and Society. doi:10.1016/j.apenergy.2013.03.027

Bermingham, A., & Smeaton, A. F. (2010). *Crowdsourced Real-world Sensing : Sentiment Analysis and the Real-Time Web*. Dublin City University. Retrieved from http://doras.dcu.ie/15585/

CEIIA. (2013). MOBI.ME. Porto: CEIIA – Centro de excelência e inovação para Indústria Automóvel.

Chew, C. M. (2010). *Pandemics in the Age of Twitter : A Content Analysis of the 2009 H1N1 Outbreak by A Content Analysis of the 2009 H1N1 Outbreak*. University of Toronto. Retrieved from https://tspace.library.utoronto.ca/bitstream/1807/25454/1/Chew_Cynthia_M_201011_MSc_thesis.pdf

Dean, J., & Ghemawat, S. (2004). MapReduce : Simplified Data Processing on Large Clusters, 1–13. Retrieved from http://static.googleusercontent.com/media/research.google.com/pt-PT//archive/mapreduce-osdi04.pdf

Demchenko, Y., Grosso, P., de Laat, C., & Membrey, P. (2013). Addressing Big Data Issues in Scientific Data Infrastructure. Retrieved from http://uazone.org/demch/papers/bddac2013-bigdata-infrastructure-v06.pdf

Dijcks, J.-P. (2012, January). How to Implement a Big Data System: Understanding a big data infrastructure by looking at a typical use case. Retrieved from http://www.oracle.com/technetwork/articles/bigdata/implementing-bigdata-1502704.html

Enbysk, L. City Deploys Big Data BI Solution to Improve Lives and Create a Smart-City Template (2013). Retrieved from http://smartcitiescouncil.com/system/tdf/public_resources/Barcelona deploys big data solution.pdf?file=1&type=node&id=712

France, J. (2013). VERGE SF 2013: How Big Data, human systems make smart cities. Retrieved from http://www.greenbiz.com/blog/2013/10/17/verge-sf-2013-how-big-data-human-systems-make-smart-cities

Gebremeskel, G. (2011). *Sentiment Analysis of Twitter posts about news*. University of Malta. Retrieved from http://www.lct-master.org/getfile.php?id=114&n=1&dt=TH&ft=pdf&type=TH

Hedlund, J. (2011). *The Smart City: Using IT to Make Cities More Livable*. Retrieved from http://www.microsoft.com/global/sv-se/offentlig-sektor/PublishingImages/The Smart City_Using IT to Make Cities More Livable.pdf

Krishnan, K. (2013). *Data Warehousing in the Age of Big Data*. Newnes.

Laney, D. (2001). *Controlling Data Volume, Velocity, and Variety*. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Latamore, B. (2013, December). IBM Smarter Cities initiative: Using big data for better traffic flow | SiliconANGLE. Retrieved from http://siliconangle.com/blog/2013/12/26/ibm-smarter-cities-initiative-using-big-data-for-better-traffic-flow/

Liu, B. (2011a). Opinion Mining and Sentiment Analysis. In *Web Data Mining* (p. 459).

Liu, B. (2011b). *Web Data Mining* (2ª edição.).

Lohr, S. (2012). Big Data's Impact in the World. *The New York Times*.

LOHR, S. (2012). The Age of Big Data. *The New York Times*. Retrieved from http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all&_r=0

Mitchell, W. J., & Casalegno, F. (2009). *connected sustainable cities* (2ª edição.). MIT Mobile Experience Lab Publishing. Retrieved from http://www.connectedurbandevelopment.org/pdf/connected_sustainable_cities.pdf

Oracle. (2012). *Oracle Information Architecture: An Architect's Guide to Big Data*. Retrieved from http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77.

Sridhar, P., & Dharmaji, N. A Comparative Study on How Big Data is Scaling Business Intelligence and Analytics (2013). Retrieved from http://www.erpublications.com/uploaded_files/download/download_03_09_2013_17_59_21.pdf

Telefónica, F. (2011). *Smart Cities: un primer paso hacia la Internet de las Cosas*. Fundación Telefónica.

Thusoo, A., Sarma, J. Sen, Jain, N., Shao, Z., Chakka, P., Anthony, S., ... Murthy, R. (2013). Hive - A Warehousing Solution Over a Map-Reduce Framework. Retrieved from http://www.vldb.org/pvldb/2/vldb09-938.pdf

Vieira, M., Figueiredo, J., Liberatti, G., & Viebrantz, A. (2012). *Bancos de Dados NoSQL: Conceitos, Ferramentas, Linguagens e Estudos de Casos no Contexto de Big Data*. Retrieved from http://data.ime.usp.br/sbbd2012/artigos/pdfs/sbbd_min_01.pdf

Vilajosana, I., Llosa, J., Martinez, B., Domingo-Prieto, M., Angles, A., & Vilajosana, X. (2013). Bootstrapping smart cities through a self-sustainable model based on big data flows. *IEEE Communications Magazine*, *51*(6), 128–134. doi:10.1109/MCOM.2013.6525605

Villanueva, F. J., Santofimia, M. J., Villa, D., Barba, J., & Lopez, J. C. (2013). Civitas: The Smart City Middleware, from Sensors to Big Data. In *2013 Seventh International Conference on Innovative*

Mobile and Internet Services in Ubiquitous Computing (IMIS) (pp. 445–450). doi:10.1109/IMIS.2013.80

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*, 347–354. doi:10.3115/1220575.1220619

Zikopoulos, I. P., Eaton, C., & Zikopoulos, P. (2011). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw Hill Professional.