unesp UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
CAMPUS DE GUARATINGUETÁ

ENZO VIVIANI GARCIA

Estrutura e características de um sistema de *Business Intelligence* baseado em *data lake* em uma empresa do setor de energia

Guaratinguetá/SP
2020

Enzo Viviani Garcia

Estrutura e características de um sistema de *Business Intelligence* baseado em *data lake* em uma empresa do setor de energia

Trabalho de Graduação apresentado ao Conselho de Curso de Graduação em Engenharia de Produção Mecânica da Faculdade de Engenharia de Guaratinguetá, Universidade Estadual Paulista, como parte dos requisitos para obtenção do diploma de Graduação em Engenharia de Produção Mecânica.

Orientador: Prof. Dr. José Roberto Dale Luche

Guaratinguetá/SP
2020

Garcia, Enzo Viviani
G216e Estrutura e características de um sistema de *Business Intelligence* baseado em *data lake* em uma empresa do setor de energia / Enzo Viviani Garcia – Guaratinguetá, 2021.
41 f : il.
Bibliografia: f. 38-41

Trabalho de Graduação em Engenharia de Produção Mecânica – Universidade Estadual Paulista, Faculdade de Engenharia de Guaratinguetá, 2021.
Orientador: Prof. Dr. José Roberto Dale Luche

1. Armazenamento de dados. 2. Gestão do conhecimento.
3. Tecnologia da informação. I. Título.

CDU 681.3.07

Luciana Máximo
Bibliotecária CRB-8/3595

unesp UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
CAMPUS DE GUARATINGUETÁ

ENZO VIVIANI GARCIA

ESTE TRABALHO DE GRADUAÇÃO FOI JULGADO ADEQUADO COMO PARTE DO REQUISITO PARA A OBTENÇÃO DO DIPLOMA DE "**GRADUADO EM ENGENHARIA DE PRODUÇÃO MECÂNICA**"

APROVADO EM SUA FORMA FINAL PELO CONSELHO DE CURSO DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO MECÂNICA

Profª. Drª. Andreia Maria Pedro Salgado
Coordenadora

BANCA EXAMINADORA:

Prof. Dr. JOSÉ ROBERTO DALE LUCHE
Orientador/UNESP-FEG

Prof. Dr. CLEGINALDO PEREIRA DE CARVALHO
UNESP-FEG

Prof. Dr. WELLINGTON DA ROCHA GOUVEIA
UNICEP

Fevereiro de 2021

RESUMO

Desde que os computadores pessoais, smartphones e principalmente a internet se popularizaram, os assuntos relacionados aos dados e o poder de informação que carregam consigo têm sido cada vez mais abordados. Atualmente, no mercado de trabalho é comum se exigir um conhecimento básico em ferramentas de análise de dados pois está função está cada vez mais deixando de ser do time da tecnologia da informação e passando a ser do próprio usuário que deseja a informação. Com esse cenário em mente, esse trabalho buscou realizar um estudo de caso em uma empresa do setor de energia que realizou recentemente a implementação de uma ferramenta de *data lake*, buscando atender o objetivo de caracterizar uma solução de *business intelligence* seguindo um paradigma de *data lake*. Comparando o projeto de implementação realizado com um framework proposto na literatura foi encontrada uma convergência nos passos realizados pela empresa e nos passos sugeridos pela literatura.

PALAVRAS-CHAVE: Big data. Data lake. Business intelligence.

ABSTRACT

Since the personal computers, smartphone and, mainly, the internet got popular, the idea of data and how much information power they carry within themselves is each year a trend topic. Currently in the job market, it is common to demand a basic knowledge in data analysis tools since this function is ceasing to be responsibility of the IT team and becoming the job of the user that want that information. Considering that scenario, this study tried to accomplish a case study in an energy sector company that performed recently an implementation of a data lake tool, looking for accomplish the objective of characterize a business intelligence solution fallowing a data lake paradigm. Comparing the accomplished implementation project with a theoretical framework suggested in literature, the study find the both converged to the same path.

KEYWORDS: Big data. Data lake. Business intelligence.

LISTA DE ABREVIATURAS E SIGLAS

Benchmarks – Avaliação de desempenho, com base em comparação com outra empresa

Big Data – Grande volume de dados produzidos atualmente pela humanidade

Business Intelligence – Sistema de apoio a decisão baseado em dados

Churn – Rotatividade do cliente

Cluster – Agrupamento de dados com base em uma característica

Costumer Relationship Management (CRM) – Sistema para registro de dados dos clientes

Data Driven – Empresa que toma suas decisões baseadas em dados

Data Lake – Repositório para uma quantidade massiva de dados

Data Mart – Partição do *Data Warehouse*

Data Warehouse – Repositório de dados

Flatfile – Formato de arquivo simples, para armazenamento de dados

Foreign Key (FK) – Campo para conexão entre duas tabelas

Hot data – Dado quente, ou seja, dado mais atualizado que a empresa tem acesso. Geralmente gerado no próprio dia

Holding – Empresa dona de outra

Lookup – Tabela auxiliar de informações em um modelo *Snowflake*

Machine Learning–Algoritmo desenvolvido que se aprimora automaticamente por meio de um sistema de *feedback*

Open Source–Programa que possui o código aberto para que qualquer pessoa possa fazer adições de funcionalidades

Outliers – Dados muitos distantes da curva de tendência dos outros dados da amostra

Primary Key (PK) – Campo para conexão entre duas tabelas

Poc (proof of concept) – Pequeno piloto de um projeto para teste iniciais

Self Service Analytics – Estrutura empresarial em que os funcionários do negócio têm a possibilidade de desenvolverem análises de dados

Snow flake Schema – Modelo dimensional de armazenamento de dados

Star Schema - Modelo dimensional de armazenamento de dados

Structured Query Language (SQL) – Linguagem mais popular para consulta em bancos de dados

Tags – Marcações feitas para categorizar componentes

SUMÁRIO

1 INTRODUÇÃO

Por onde se olha, há pessoas mexendo em seus smartphones. Seja conversando com amigos e familiares, fazendo compras ou lendo notícias. E, com esse cenário, uma observação se mostra muito clara: nunca foram produzidos tantos dados tão rapidamente como é produzido hoje (KEIM et al., 2006), produzindo a já popular, *big data*. Os dados, inclusive, deixaram de ser exclusividades das grandes empresas, que eram as únicas que tinham capacidade de armazenamento e processamento de dados (DAVENPORT; DYCHÉ, 2013a). Desde alguns anos atrás, as novas *startups* já "nascem" com os dados como seu *core business*, empresas de médio porte possuem cada vez mais ferramentas, a preços cada vez mais acessíveis, para se tornarem *data driven* (BEIER, 2016) e as empresas tradicionais passaram a dar importância muito maior aos dados, visualizando-os até mesmo como seus mais valiosos ativos (FARID et al., 2016; HINDLE; VIDGEN, 2018).

Já se foi o tempo em que apenas empresas que possuíam uma parte do negócio online tinham informações para extrair da internet. Hoje, até empresas que não possuem nenhuma parte "digital" em seu negócio podem extrair *insights* diretamente da rede mundial de computadores. Por exemplo, uma ação que é muito praticada hoje em dia é a análise de sentimento. Essa técnica de *machine learning* analisa um texto, que pode ter sido escrito por um consumidor em uma rede social ou pode ser a transcrição de uma conversa que o consumidor teve com o serviço de atendimento ao consumidor (SAC) da empresa, por exemplo, e com base em um treinamento prévio do modelo que está analisando o texto, fornece uma percepção geral de qual emoção o autor do texto transmitiu por meio do mesmo. Assim, a empresa consegue ter uma ideia de como está a avaliação de sua imagem (O'LEARY, 2014).

Além da utilização da *big data* na área de marketing, existem vários outros modos de agregar valor ao negócio utilizando-a, até mesmo em atividades chave dentro da empresa, como atividades jurídicas (BEAN; KIRON, 2013), na linha de produção (LEE et al., 2013) e no sistema de cadeia de suprimentos (KAUR; SINGH, 2018). Nesse último cenário, Bose, Pal e Ye(2008) encontraram em seu estudo que uma integração dos sistemas que compõem a cadeia de suprimentos com tecnologias da *big data* pode trazer muitos benefícios a empresa, como o aumento das vendas, uma redução do custo de estoques e uma melhor organização interna.

Com todos esses dados disponíveis, as empresas passaram a carecer de um repositório que aceitasse dados em um volume absurdo, de várias fontes e em diferentes formatos

(FANG, 2015; MALYSIAK-MROZEK; STABLA; MROZEK, 2018; O'LEARY, 2014). Foi criado então, o conceito de *data lake*, um repositório de dados com arquitetura e modelo diferente dos populares *Data Warehouse* (DW), que eram e continuam sendo muito utilizados nas empresas que trabalham com dados de uma forma mais "tradicional" (FANG, 2015; O'LEARY, 2014).

Junto com o conceito de big data e data lake, surge um novo jeito de analisar os dados: a análise de *hot data*. Nas empresas mais tradicionais, que trabalhavam com DW e análise de dados, era muito comum a geração de relatórios e *dashboards* baseados em *cold data*, isto é, dados do dia anterior (d-1). Isso faz com que a empresa apenas "veja" o passado e certas partes do presente, porém torna muito difícil a previsão do futuro (KAMBATLA et al., 2014). Isso se deve ao fato de os dados terem que ser tratados antes de entrar no DW, por um processo de *Extract, Transform and Load* (ETL), que será explicado mais à frente. Como para armazenamento no data lake não é necessário esse processo, os dados são armazenados direto de suas fontes e já ficam disponíveis para o analista utiliza-los, podendo assim "ver" o hoje claramente e fazer previsões e suposições sobre o futuro (KAMBATLA et al., 2014; WAMBA et al., 2017).

1.1 OBJETIVOS

O objetivo geral neste trabalho consiste em caracterizar uma solução de *business intelligence* seguindo um paradigma de *data lake* em uma empresa do setor de energia.

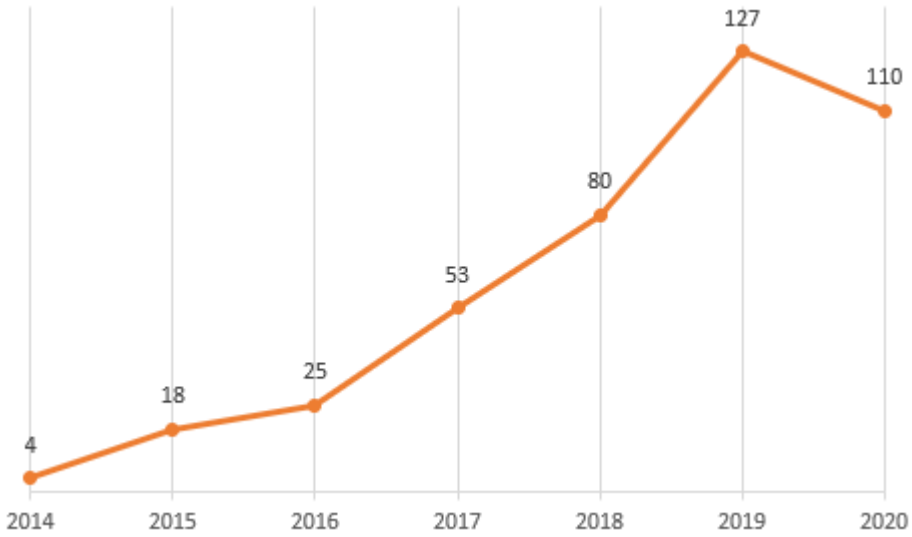
Como objetivos específicos tem-se o mapeamento do uso de data lake na indústria e a definição de ferramentas para o framework proposto para suporte do objetivo principal.

1.2 JUSTIFICATIVA

Segundo estudo publicado em 2015 pela consultoria em tecnologia Management Solutions, o volume dos dados produzidos pelo ser humano duplica a cada 18 meses (MANAGEMENT SOLUTIONS, 2015; GUAMÁN; VACA; YUQUILEMA, 2018), sendo que grande parte desses dados é proveniente de mídias sociais, e pode estar tanto em formatos estruturados quanto em formato não estruturado (BAARS; KEMPER, 2008; STIEGLITZ et al., 2018). Para que se possa trabalhar com esse volume massivo de dados, é necessário que empresas de tecnologia desenvolvam novas ferramentas, mas também que a academia estude novos modos teóricos de armazenamento, tratamento e análise desses dados. Dessa forma,

insumos teóricos são fornecidos as partes responsáveis pelo desenvolvimento de novas tecnologias, para tentarem transformar a teoria em prática (CARAÇA; LUNDVALL; MENDONÇA, 2009). Seguindo essa linha de pensamento, foi feito um estudo da quantidade de artigos publicados nas bases de dados Scopus e Web of Science, a partir de 2014. Essa data foi escolhida, pois o conceito de *data lake* foi cunhado por James Dixon em 2010 (O'LEARY, 2014) e notou-se que, pesquisando a expressão "data lake" no título, palavras chave ou resumo, em qualquer data e apenas trabalhos escritos em inglês ou português, até 2014 não houveram registros de trabalhos. O resultado da pesquisa, considerando a primeira aparição de trabalhos sobre o tema como data inicial, pode ser observado na Figura 1.

Figura 1 – Evolução das publicações contendo o termo "Data Lake" em seu título, palavras chave ou resumo



Fonte: Scopus (2020) e Web of Science (2020).

Como se pode observar na Figura 1, a quantidade de artigos publicados relacionados ao termo "*data lake*" vem crescendo de forma muito rápida. Esse crescimento pode ser associado a uma maior acessibilidade que as pessoas em geral possuem das ferramentas de computação em nuvem (BANSAL, 2014) e ao valor que as empresas estão dando aos dados (FARID et al., 2016), tornando-os, um assunto cada vez mais relevante.

Analisando mais profundamente os documentos presentes na base de dados estudada, foram identificados poucos trabalhos que estudaram o caso real de uma empresa que está em processo de migração de um sistema de armazenamento e análise de dados tradicional para um sistema mais moderno. Alguns estudos, como o de Davenport e Dyché (2013) analisam a

implementação de ferramentas de *Big Data* em empresas, enquanto outros, como o de O'leary (2014) comparam *data warehouses* com *data lakes*, em relação a estrutura, capacidade, confiabilidade dos dados, entre outros componentes das estruturas.Um dos poucos estudos de caso encontrados, é dos autores indianos Dutta e Bose (2015) que realizam um estudo para implementação de tecnologias relacionadas a *big data* em uma grande empresa de cimento indiana. Como resultados, desenvolveram um framework, que nas próprias palavras dos autores, "é mais prático do que teórico" (DUTTA; BOSE, 2015).

Cabe ressaltar que foi encontrado apenas um estudo brasileiro relevante (com pelo menos uma citação). Em seu trabalho, Couto et al. (2019) fazem um mapeamento da literatura e respondem algumas perguntas relacionados à definição mais aceita de *data lake* e as arquiteturas possíveis do mesmo. Porém, não é conduzida nenhuma aplicação real dos conceitos

Neste trabalho é desenvolvido um referencial teórico embasado nos diversos estudos encontrados e, em seguida, são aplicados os conceitos num estudo de caso que estuda os motivos que levaram uma empresa já consolidada no mercado a mudar seu sistema de *business inteligence* (BI) tradicional para um baseado em *data lake* e *self-service analytics*. O acompanhamento é realizado por seis meses, junto aos responsáveis pelo processo e passos tomados, desde a escolha da empresa fornecedora do ambiente, passando pelas dificuldades, até a implementação final da tecnologia. O estudo será delimitado à realidade e desafios da empresa líder do setor de óleo e gás brasileiro, que se encontra em migração para o setor de energia como um todo, com matriz em São Paulo, capital, no ano de 2020.

Em função do exposto, a questão de pesquisa que norteia esse estudo é: "Como modernizar o sistema de *business intelligence* de uma empresa do setor de energia?".

1.3 ESTRUTURA DO TRABALHO

Este estudo está dividido em seis capítulos. O primeiro capítulo apresenta uma introdução sobre o tema, juntamente com os objetivos e justificativa. O segundo capítulo apresenta um referencial teórico sobre conceitos importantes dentro do tema BI, como *data warehouse*, ETL, *data lake*, entre outros. O terceiro capítulo discorre sobre o método e as ferramentas de pesquisa utilizadas. O quarto capítulo apresenta o framework teórico, encontrado na literatura, no qual esse estudo se baseia e o quinto capítulo mostra o estudo de caso feito na empresa do setor de energia. Por fim, o sexto capítulo apresenta as considerações finais junto com sugestões para estudos futuros.

2 REFERENCIAL TEÓRICO

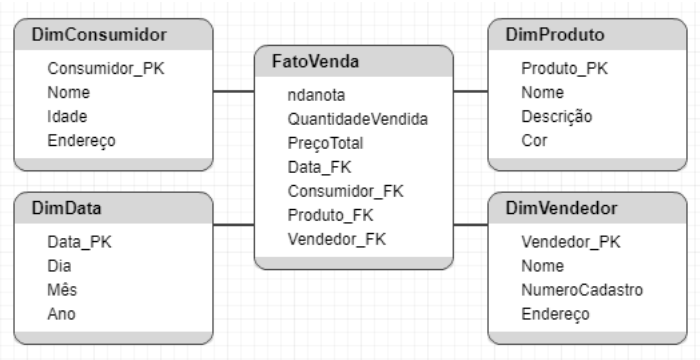
Neste capítulo serão apresentados os seguintes conceitos: *Data Warehouse* (DW), *Extract, Transform and Load* (ETL), *Business Intelligence* (BI), *Big Data*, *Data Lake* e *Serf-Service Analytics*. Primeiramente a análise dadostradicional e em seguida, a análise de dados mais atual.

2.1 DATA WAREHOUSE (DW)

Segundo Chaudhuri e Dayal (1998) e Inmon (2002), DW é "uma coleção de dados orientada ao assunto, integrada, variada com o tempo e não volátil, usada principalmente para auxiliar na tomada de decisões pela organização". Ou seja, é um banco de dados com muita capacidade de armazenamento, bem organizado, que possui dados já tratados e consolidados. Dentro dos DW é possível subdividir ainda em *Data Marts* (DM), que nada mais são que pequenas repartições do DW em bancos de dados menores e focados em assuntos específicos (CHAUDHURI; DAYAL, 1997).

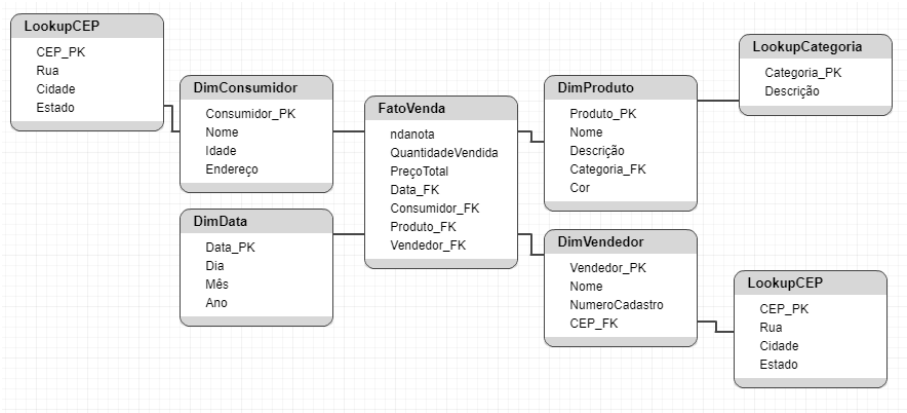
Dentro dos DWs, o mais comum é que os dados estejam organizados de acordo com um modelo multidimensional. O esquema mais comum de representação do modelo multidimensional, é o esquema-estrela (*star schema*). Nele, existe a tabela fato, na qual estão os eventos que acontecem, e as tabelas dimensões, nas quais estão dados "imutáveis". Os dados dessas tabelas são relacionados por uma chave-estrangeira (*foreignkey*) (CHAUDHURI; DAYAL, 1997). Um exemplo da aplicação desse esquema: quando ocorre uma venda na empresa, o número, o valor, a quantidade da nota e o código do fornecedor são lançados na tabela fato; por meio de chaves no sistema, o código do fornecedor é buscado na tabela dimensão referente ao fornecedor e seus dados, como endereço, nome, razão social, etc são associados àquela linha da tabela fato. Outro esquema que representa o modelo multidimensional, e é uma derivação do *star-schema,* é o esquema floco de neve (*snowflake schema*). Nele, a estrutura de tabelas fato e dimensões é mesma, porém existem também as tabelas *lookup,* as quais estão relacionadas as tabelas dimensão por meio de chaves estrangeiras também. Na Figura 2 e 3 é possível observar uma representação gráfica de ambos os esquemas.

Figura 2 - Star Schema



Fonte: Adaptado de Chaudhuri e Dayal (1997).

Figura 3 – Snowflake Schema



Fonte: Adaptado de Chaudhuri e Dayal (1997).

É possível observar na Figura 2 a relação da tabela fato com as tabelas dimensão (DIM), por meio da *foreign key* (FK), na qual um campo contendo essa chave na tabela fato busca informação em outra tabela que tem essa mesma chave como chave primária (*primary key*, PK), isto é, um identificador único de cada linha. Já na Figura 3, a mesma relação é observada, conjuntamente com a relação entre as tabelas dimensão e as tabelas *lookup*, que se relacionam do mesmo modo.

Devido a essa estrutura bem específica de *schemas*, os dados armazenados em DW's geralmente são muito assertivos e confiáveis, pois os dados passam por processos de transformação que podem ser constantemente revisados e aprimorados (FANG, 2015; WALKER; ALREHAMY, 2015). Os analistas de dados que acessam os DW's também se beneficiam de sua rapidez, uma vez que os dados já foram tratados para serem armazenados e quando chamados em consultas, os bancos precisam apenas exibi-los ao usuário (FANG, 2015). Fora essas vantagens, Fang (2015) ainda menciona a facilidade de acesso aos dados

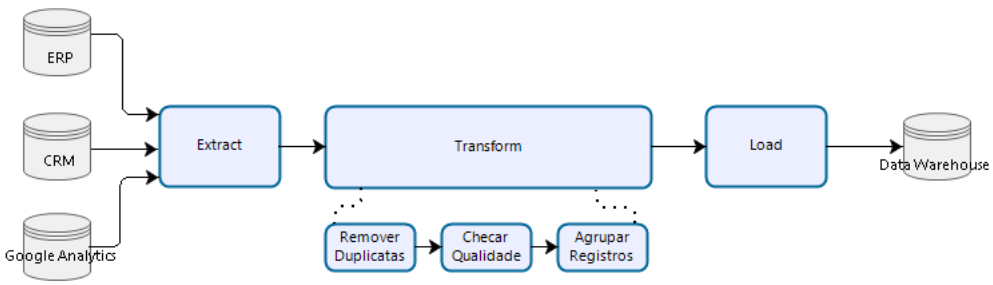
armazenados nos DW's, que podem ser acessados por meio de simples consultas utilizando a Linguagem Estruturada de Querys (*Structured Query Language*, SQL), e a integração de dados de diferentes fontes, que ocorre nas transformações do dado, antes do mesmo ser armazenado.

Porém, uma estrutura bem específica não gera só vantagens, já que os DW's só conseguem armazenar um tipo de dado: os dados estruturados. Isso torna o processo de tratamento para armazenamentos desses dados longo, uma vez que são necessário vários processos de transformação específica para o dado chegar ao modelo exato para ser armazenado (WALKER; ALREHAMY, 2015). Além disso, a estrutura dos DW's é pensada e desenvolvida para responder apenas uma série de perguntas, de modo que, uma vez que o banco foi construído, é muito difícil pensar em análises que divergem muito das análises "pré-pensadas" ou até mesmo pensar em novos modos de análise (O'LEARY, 2014; WALKER; ALREHAMY, 2015).

2.2 EXTRACT, TRANSFORM AND LOAD (ETL)

A extração, transformação e carregamento é um processo realizado antes de armazenar os dados nos DW's, responsável por ingerir dados de fontes externas, transforma-los no formato desejado – seja integrar dados de fontes (pode ser a mesma fonte ou fontes diferentes), converter medidas, agrupar dados, deixá-los no formato do *schema* utilizado, etc – e carregá-los no DW para armazenamento (BANSAL; KAGEMANN, 2015; PERWEJ, 2017). Esse processo é feito por meio de ferramentas integradas com as fontes de dados (ERP, CRM, etc) e com os DW's e pode ser observado na Figura 4.

Figura 4 – Processo de ETL



Fonte: Adaptado de Bansal (2014).

Conforme apresentado na Figura 4, na primeira etapa – extração – os dados são extraídos das fontes, geralmente em formatos *flat* (CSV, XLS, TXT), isto é, registros em formatos tabulares sem índices ou relacionamentos explícitos entre si, com suas colunas separadas por vírgulas, por exemplo. Na segunda etapa, transformação, os dados são limpos, integrados entre si e convertidos para o esquema do banco. Tipos comuns de transformação são: normalizar os dados, remover duplicatas, checar a qualidade dos dados (se nos campos da tabela destino no DW estão em formato *date*, por exemplo, é necessário checar se todos os registros que irão para esse campo estão nesse formato), aplicar filtros (por data, região, sexo, etc) e agrupar campos parecidos e somar suas métricas. Por fim, a última etapa, carregamento, se refere ao processo de carregar os dados nas tabelas do DW (BANSAL, 2014; BANSAL; KAGEMANN, 2015).

2.3 BUSINESS INTELLIGENCE

O termo *Business Intelligence*, ou BI, é utilizado para caracterizar todo o conjunto de processos e ferramentas que transformam os dados coletados e armazenados de uma empresa em informações prontas para suportar decisões (MZAGHLOUL; ALI-ELDIN; SALEM, 2013). Seu modelo tradicional é composto por uma ferramenta que realiza o processo de ETL e outra ferramenta responsável pela disponibilização dos dados gerados, geralmente um software de relatórios (DAVENPORT; DYCHÉ, 2013b). Além disso, também é possível haver uma ferramenta de visualização de dados.

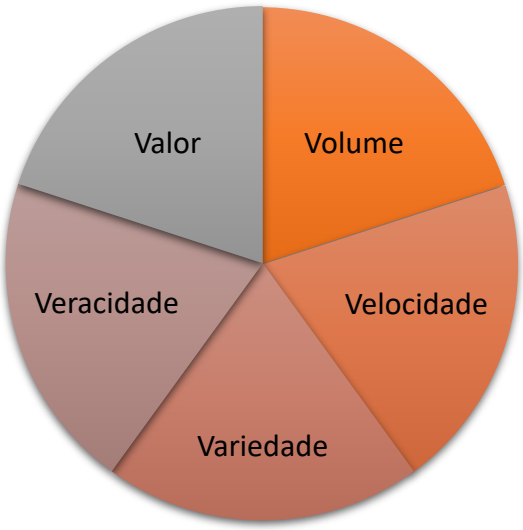
Apesar da vasta disponibilidade dos relatórios gerados pelo sistema de BI, há uma complexa estrutura por traz desse sistema (LÖNNQVIST; PUHAKKA, 2009). Primeiramente porque todo o sistema, na maioria dos casos, trabalha com um modelo dimensional específico: ou seja, é necessário que a arquitetura das ferramentas seja feita de acordo com esse modelo. Outro problema é relacionado à limitação de arquitetura dos DW's: o modo como ele será construído deve ser previamente pensando conforme as perguntas que se quer responder com os dados. Portanto, o sistema de BI tradicional é limitado a responder apenas as perguntas para as quais ele foi previamente construído para responder (O'LEARY, 2014; WALKER; ALREHAMY, 2015).

2.4 BIG DATA

Big data pode ser considerada como um volume tão grande de dados e com estrutura tão difusa que as ferramentas tradicionais de programação, como repositórios de dados, softwares de processamento e análise, etc, não teriam condições de lidar com todo esse dado(BEIER, 2016; PROVOST; FAWCETT, 2013).

A *Big Data* também é definida geralmente com base em seus "V". Inicialmente, eram três, volume, velocidade e variedade, que adiante foram acrescentados de mais quatro: veracidade, variabilidade, valor e visibilidade (MILOSLAVSKAYA; TOLSTOY, 2016). Outros autores, como Perwej (2017) e Malysiak-Mrozek, Stabla e Mrozek (2018) consideram apenas cinco "V": volume, velocidade, variedade, veracidade e valor, que podem ser observados graficamente na Figura 5.

Figura 5 – Os cinco "V" da *Biga Data*



Fonte: Adaptado de Perwej (2017) e Malysiak-Mrozek, Stabla e Mrozek (2018).

Na Figura 5 é possível observar os cinco "V" mais presentes nas definições de *Big Data*: volume se refere à quantidade de dados; velocidade se refere à rapidez na qual os dados são gerados; variedade se refere aos vários tipos possíveis de dados; veracidade se refere à acurácia dos dados; e valor se refere ao que os dados podem gerar de valor para empresa.

Outra característica marcante da *Big Data* é sua estrutura, composta de dados em suas três possíveis estruturas (MILOSLAVSKAYA; TOLSTOY, 2016; PERWEJ, 2017):

- Dados Estruturados - Qualquer dado que possa ser armazenado, retirado e processado em um formato único e definitivo. Esse é o tipo de dado mais organizado, o que beneficia ferramentas de busca e armazenamento. Ex: dados provenientes de sistemas *Costumer Relationship Management* (CRM), *Enterprise Resource Planing* (ERP) e dados contidos em *Data Warehouse* (DW);
- Dados Semi-estruturados - São dados que não estão organizados de um jeito específico, mas possuem uma organização associada a metadados ou cabeçalhos. Geralmente precisam de um tratamento especial antes de serem armazenados em bancos de dados tradicionais. Ex: JSON, XML, email; e
- Dados Não-Estruturados - Quaisquer dados que não possuem uma estrutura pré-definida ou que não são organizados de uma maneira pré-definida. Podem ser textuais ou não textuais. Ex: Imagens, arquivos de aúdio, SMS.

É importante ressaltar que devido ao grande volume de dados, é necessário realizar um saneamento nos mesmos, de forma a "filtrar" apenas dados relevantes. Mesmo em dados provenientes internamente da própria empresa podem haver sujeiras ou *outliers*, que necessitam de tratamento antes de serem armazenados e analisados (BIZER et al., 2011; ZHANG et al., 2018).

2.5 DATA LAKE

O conceito de *data lake* surge em oposição ao conceitos de DW (O'LEARY, 2014), apresentando-se como um enorme repositório de dados brutos (ALSERAFI et al., 2016; GUAMÁN; VACA; YUQUILEMA, 2018). Esse repositório possui uma enorme escalabilidade de armazenamento, geralmente utilizando a plataforma Hadoop como base, e pode ser consultado por aplicações analíticas dinâmicas, e não apenas aplicações pré-definidas como ocorre nos DW (MILOSLAVSKAYA; TOLSTOY, 2016).

O Apache Hadoop, criado em meados de 2006, é uma plataforma *open source* desenvolvida para armazenamento e processamento de um grande volume de dados PERWEJ, 2017), e que hoje serve como um dos componentes da arquitetura de vários *data lakes.* Seu framework é baseado em vários *clusters* (conjuntos de hardware) e em 2 principais tecnologias: o sistema de arquivo distribuído Hadoop (Hadoop's Distributed File System,

HDFS) e o MapReduce. O HDFS é uma tecnologia de armazenamento de dados, em que os mesmos são divididos em blocos e armazenados separadamente em diversos nós (espaços dentro dos *clusters* reservados para armazenamento); já o MapReduce, é responsável pelo processamento dos dados, fazendo isso dentro dos nós já povoados pelo HDFS e em vários nós simultaneamente (MUNSHI; MOHAMED, 2018).

O nome "data lake" foi cunhado por James Dixon, em 2010, em seu blog pessoal. Segundo Dixon (2010):

> Se você pensar em um *data warehouse* como uma água engarrafada – limpa, empacotada e estruturada, fácil para consumo – o *data lake* é um imenso corpo de água num estado mais natural. O conteúdo do *data lake* flue para dentro do mesmo através das fontes que alimentam o *lake*, e vários usuários do *lake* podem vir examinar, mergulhar ou pegar amostras.

A estrutura do *data lake*, diferentemente dos DW's, é uma estrutura chamada flat, na qual os dados são armazenados em seu formato bruto e possuem um identificador único e algumas *tags* de metadados (GUAMÁN; VACA; YUQUILEMA, 2018). Esses metadados, ou dados que referenciam outros dados, são utilizados para entendimento dos relacionamentos e informações que cada dado faz e representa.

Outra diferença muito marcante com o DW, é relacionada à rapidez em disponibilizar os dados, após eles terem sido gerados nos sistemas que alimentam o *lake.* Enquanto no DW os dados passam por um processo de ETL que geralmente é complexo e custoso antes de serem armazenados, no *lake* esse processo de modelagem dos dados só ocorre quando os dados são utilizados (FANG, 2015; GUAMÁN; VACA; YUQUILEMA, 2018). Esse modelo de processamento apenas na utilização é chamado de "*No Schema"*, ou sem esquema, pois os dados são armazenados sem padrão de esquema definido. Segundo FANG (2015), esse modelo apresenta vantagens e desvantagens. As vantagens são relacionadas a flexibilidade que o analista de dados possui, já que o dado está ali do jeito que foi gerado e pode ser analisado de várias formas e ao suporte de armazenamento aos mais variados tipos de dados. Já as desvantagens estão ligadas a uma regra: em algum momento os dados vão ter que ser modelados, pois precisam estar enquadrados em algum esquema para serem analisados.

2.6 SELF-SERVICE ANALYTICS

Junto com as novas formas de armazenar dados, surgem novas ferramentas para analisá-los. Apesar da ferramenta mais utilizada em quase todas as companhias ainda ser o Microsoft

Excel, existem vários softwares de fácil uso disponíveis no mercado (DINSMORE, 2016). Esse softwares, como o Microsoft Power BI ou o Tableau, por exemplo, buscam atingir todos os tipos de usuários: desde o usuário *expert* (Ex: desenvolvedores, estatísticos, etc) até os usuários que são apenas consumidores de informação, conhecidos como usuários da ponta (CLARKE; TYRRELL; NAGLE, 2016; DINSMORE, 2016).

O *self-service analytics* surge em oposição ao modelo tradicional de BI, no qual após os dados serem tratados, geralmente um expert em análise de dados, por meio de um software específico constrói as análises que a área focadano negócio solicitou (MZAGHLOUL; ALI-ELDIN; SALEM, 2013). Porém o que ocorre caso o que foi construído não satisfaça o negócio? Bom, aí a tarefa de construir uma nova visão do dado entra na fila de tarefas do analista de dados novamente e será entregue assim que finalizado. Esse processo, apesar de ser feito por um especialista, além de ser muito lento pode demorar a gerar o resultado que a área de negócio deseja (MZAGHLOUL; ALI-ELDIN; SALEM, 2013). O *self-service analytics* buscar acelerar esse processo, fornecendo ferramentas de fácil manuseio para o usuário da ponta construir visões de dados e ter seus próprios *insights* (CLARKE; TYRRELL; NAGLE, 2016; DINSMORE, 2016; MZAGHLOUL; ALI-ELDIN; SALEM, 2013).

O usuário final só consegue essa liberdade de criação, devido ao modo como as ferramentas são construídas hoje em dia (DINSMORE, 2016). Não é mais necessária uma linha de código para se construir análises simples-moderadas nos softwares mais atuais, tornando-os tão práticos como o, já comentado e amplamente utilizado, Microsoft Excel. Nesse cenário segundo Dinsmore (2016), o especialista em dados deixa de construir indicadores para as pontas (que são construídos pelos usuários finais agora) e constrói indicadores mais gerais, para o negócio como um todo, além de ser o principal responsável pela integração dos dados, isto é, o controle e organização das fontes de dados que estão alimentando as análises da ponta.

3 MÉTODO

Neste capítulo, será tratado o método utilizado no estudo.

3.1 CARATERIZAÇÃO DA PESQUISA

O Quadro 1 apresenta a classificação desta pesquisa.

Quadro 1 – Classificação da pesquisa

Natureza	Aplicada
Objetivos	Exploratória
Abordagem	Qualitativa
Procedimentos Técnicos	Pesquisa bibliográfica, pesquisa documental e estudo de caso

Fonte: Adaptado de Kothari (2004).

Conforme apresentado no Quadro 1, a natureza da pesquisa pode ser considerada como aplicada, pois visa estudar um fenômeno, sem interferência no mesmo, formar uma teoriae melhorar uma situação real (KOTHARI, 2004; PROVDANOV; FREITAS, 2013). Quanto ao seu objetivo, pode ser classificada como uma pesquisa exploratória, uma vez que possibilitará que o pesquisador explore o universo definido e observe seus elementos para compor uma nova explicação para o fenômeno estudado (MIGUEL, 2018; PROVDANOV; FREITAS, 2013).

A abordagem pode ser considerada qualitativa, pois além do ambiente natural ser a fonte das coletas de dados, há todo um contexto a ser considerado na análise dos dados como motivações, cenário macroeconômico, etc (KOTHARI, 2004; PROVDANOV; FREITAS, 2013)

Por fim, os procedimentos técnicos utilizados serão três: uma pesquisa bibliográfica inicial, a fim de contextualizar o leitor dos assuntos trabalhados no estudo e servir como base de comparação para as decisões tomadas a frente; uma pesquisa documental realizada nos documentos da empresa estudada, onde se buscará as motivações para mudança da visão adotada sobre os dados, além de critérios técnicos para tal mudança e escolha da plataforma a ser utilizada; e um estudo de caso, no qual se analisará a realidade de uma empresa que está buscando realizar a modernização de seu sistema de BI(PROVDANOV; FREITAS, 2013).

3.2 EMPRESA ESTUDADA

Para realização deste estudo, foi escolhida uma empresa do setor de óleo e gás que está em processo de migração para o setor de energia como um todo. Esta empresa, já está no mercado a mais de 80 anos, possui ações na bolsa de valores brasileira, através de sua *holding*, e é referência mundial no ramo, atraindo diversas outras empresas para sua matriz em São Paulo, capital, para realização de *benchmarks.*

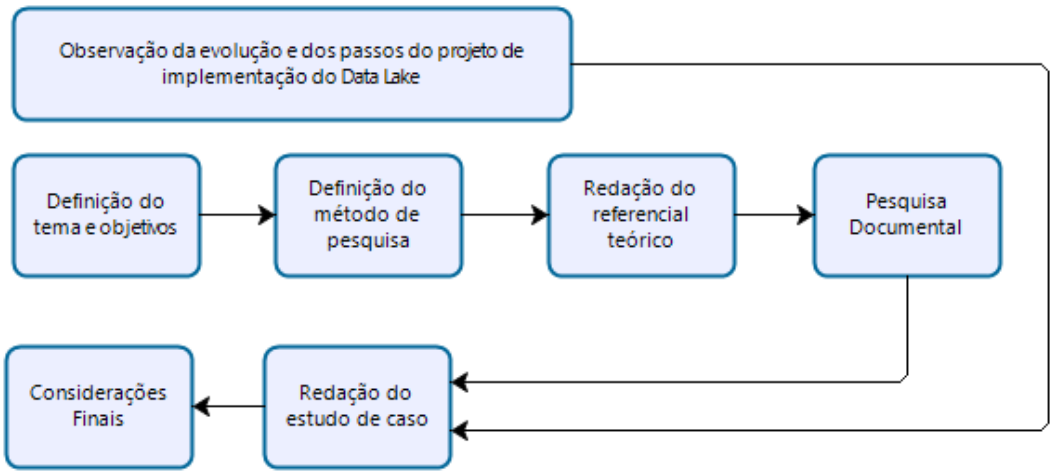
Após passar por uma mudança em sua diretoria em 2018, a empresa passou a ter um foco voltado na importância que os dados têm dentro da companhia. Somado a isso, por oferecer um produto muito tradicional, a empresa busca uma renovação no modo de venda de seu produto, investindo cada vez mais em mídias sociais, aplicativos mensageiros e aplicativos próprios.

Com todo esse cenário e uma mudança na gerência da área de Tecnologia da Informação no ano de 2020, foi iniciado um projeto voltado à construção de uma maior capacidade de armazenamento, processamento e análise de dados, agora provento em diversos formatos e de diversas fontes como redes sociais, aplicativos de compras, etc. Esse projeto está sendo executado primeiramente através da contratação de uma ferramenta de data lake para armazenamento e realização de algumas análises de dados.

3.3 FLUXO METODOLÓGICO

O fluxo metodológico o qual esse trabalho irá seguir encontra-se na Figura 6.

Figura 6 – Fluxo metodológico



Fonte: O próprio autor.

O estudo iniciou com a definição do tema e dos objetivos pelo autor, tendo em vista as oportunidades de estudo encontradas na literatura. O segundo passo foi a definição do método de pesquisa. Em seguida, foi redigido o referencial teórico embasado em trabalhos presentes nas bases de dados Scopus e Web of Science.

O passo seguinte é o estudo de caso, que está sendo realizado desde o começo desta pesquisa. A observação da evolução e das decisões tomadas no projeto está sendo feita desde a definição dos objetivos desse trabalho.

Além da observação, o estudo de caso foi adensado com uma pesquisa documental, feita nos documentos autorizados pela empresa e que buscou identificar os motivos para o início da mudança no sistema de BI da empresa, conjuntamente com os *benchmarks* realizados e com os critérios para escolha de ferramentas.

Por fim, foi redigido o texto final com as considerações finais do autor, conjuntamente com oportunidades de pesquisas futuras.

4 FRAMEWORK TEÓRICO

Com base na leitura e análise da literatura, foi escolhido o framework mais completo e relacionado ao tema, implementação de um projeto relacionado a *big data*, encontrado. Esse framework, sugerido por Dutta e Bose (2015), é exibido na Figura 7 e em seguida, tem todas suas fases e etapas explicados.

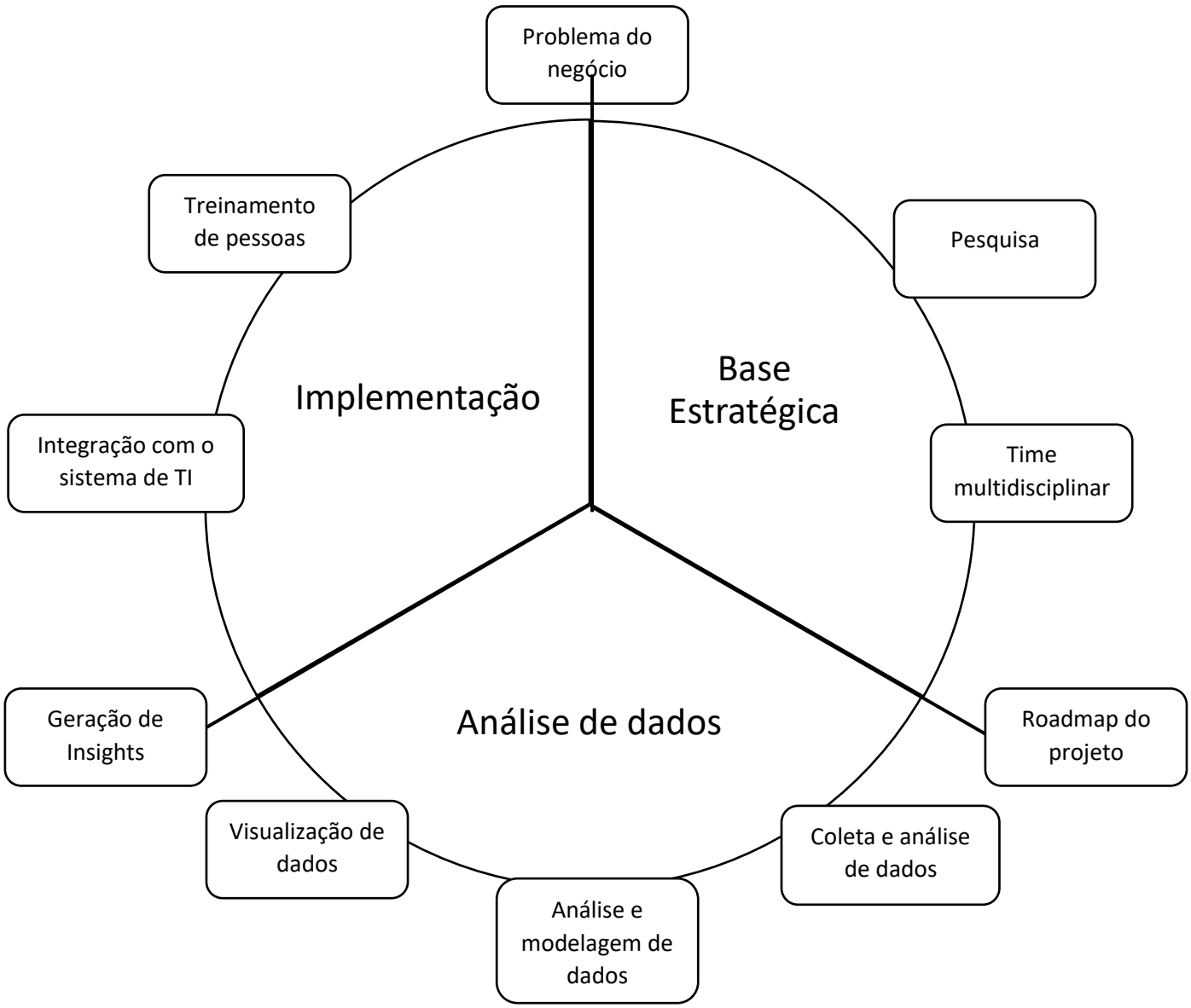
4.1 PROBLEMA DO NEGÓCIO

A primeira fase consiste em definir qual é o problema do negócio que visa ser melhorado ou resolvido. Este entendimento é de extrema importância, pois define o escopo de todo projeto, alinha as expectativas dos *stakeholders* e desmistifica qualquer receio em relação a *big data*. É importante considerar nessa fase a opinião de vários pontos de vista, para que o entendimento seja o mais completo possível (CHANG et al., 2020; DUTTA; BOSE, 2015; KOHAVI; ROTHLEDER; SIMOUDIS, 2002).

4.2 PESQUISA

Em seguida, os profissionais responsáveis pelo projeto devem realizar uma pesquisa tanto interna como externa. A pesquisa interna deve avaliar os serviços de TI atuais e como eles podem ser melhorados. Já a pesquisa externa, deve realizar*benchmarks* com outras empresas e buscar quais são as tecnologias presentes no mercado (DUTTA; BOSE, 2015; FLYVBJERG, 2013; HETEMI; JERBRANT; MERE, 2020). A contratação de empresas especializadas na realização de avaliações é prática comum nessa fase.

Figura 7 – Framework teórico proposto



Fonte: Adaptado de Dutta e Bose (2015).

4.3 TIME MULTIDISCIPLINAR

Segundo Dutta e Bose (2015) essa é uma das fases mais importantes do projeto, o momento de montar o time para realização do projeto.É necessário considerar um time diverso com: funcionários técnicos da área de TI, como engenheiros e arquitetos de dados; funcionários funcionais do TI, que possuem mais interface com o negócio, como cientistas e analistas de dados; funcionários do negócio, que entendem quais são as regras a serem utilizadas e quais dados são relevantes; e outros especialistas em área específicas, como profissionais da segurança da informação, profissionais da área jurídica (principalmente levado em conta as novas regras relacionadas a LGPD), entre outros. Essa fase é de extrema importância, pois pode limitar ou expandir a visão que o time terá do projeto.

4.4 ROADMAP DO PROJETO

Com as etapas base do projeto realizadas, é necessário planejar todos os passos a serem tomados. Um *roadmap* de projeto é importante para acompanhar a evolução de cada frente de atuação. Nessa fase, após a construção dos passos, as tarefas são divididas entre os membros e é comum que sejam formados pares ou até times menores para focar em frente específicas. Outro ponto importante desta fase é a definição de metas e datas, para que haja planejamento por parte dos integrantes do projeto (ASHRAFI et al., 2019; DUTTA; BOSE, 2015; VIAENE; VAN DEN BUNDER, 2011).

4.5 COLETA E ANÁLISE DE DADOS

Após o planejamento, o projeto se inicia com a tarefade levantar quais são as fontes de dados a serem tratadas. Todas as fontes de dados relevantes ao projeto devem ser analisadas do ponto de vista do negócio (Esta fonte de dados me traz informações importantes?) e do ponto de vista da área de TI (Quais são os tipos de dados? Qual a volumetria dos dados? Em qual tipo de banco de dados está?). Novas fontes de dados podem ser consideradas, como fontes de dados não estruturados, por exemplo e a relação entre esses novos dados e os dados estruturados presentes na companhia deve ser planejada (DUTTA; BOSE, 2015).

4.6 ANÁLISE E MODELAGEM DE DADOS

A etapa seguinte, geralmente realizada pelo cientista de dados, tem por objetivo entender quais informações os dados podem trazer para empresa (DUTTA; BOSE, 2015). Diferentemente do modelo antigo de BI, em que as informações que os dados poderiam trazer para a empresa eram definidas previamente para que todo o modelo de armazenamento dos dados fosse construído com base nessa definição, nos projetos de *big data* a análise do poder de informação dos dados é feita posteriormente, gerando maior capacidade de geração de novos insights (O'LEARY, 2013; WALKER; ALREHAMY, 2015).

4.7 VISUALIZAÇÃO DE DADOS

Após os dados serem analisados e modelados conforme suas características é importante definir como eles serão visualizados pelos usuários. Dinsmore (2016) comenta que apesar do modelo de relatórios através de tabelas ser bem difundido, hoje existem diversas ferramentas capazes de transformarem um grande volume de dados em gráficos e outras formas de visualização, de forma a deixar a identificação de padrões, por exemplo, mas rápida. Nesse momento, *softwares* específicos de visualização de dados gerar muitos benefícios para o projeto e para empresa (CLARKE; TYRRELL; NAGLE, 2016).

4.8 GERAÇÃO DE INSIGHTS

Em complemento a etapa anterior, nessa etapa as regras de negócio são introduzidas nos dados, de forma a criar novas bases "pré-prontas" que fornecem informações específicas sobre um único indicador (DUTTA; BOSE, 2015). É interessante ressaltar que esses insights produzidos são relacionados a indicadores já conhecidos do negócio, ao passo que, na fase de análise e modelamento de dados o cientista de dados pode descobrir novas visões e indicadores que não haviam sido considerados ou imaginados anteriormente (FANG, 2015).

4.9 INTEGRAÇÃO COM O SISTEMA DE TI

Outro ponto crucial de ponderamento é relacionado a estrutura do time de TI, porque afinal, após a realização do projeto de implementação, o time de TI terá um novo ambiente e possivelmente novas ferramentas para sustentar (DUTTA; BOSE, 2015). Então, é importante realizar o projeto de infraestrutura minuciosamente para que o máximo de *bugs* e possíveis falhas sejam resolvidas o mais breve possível, pois segundo Bose, Pal e Ye(2008)o sucesso de integração geralmente leva a uma melhora na performance do negócio como um todo também.

4.10 TREINAMENTO DE PESSOAS

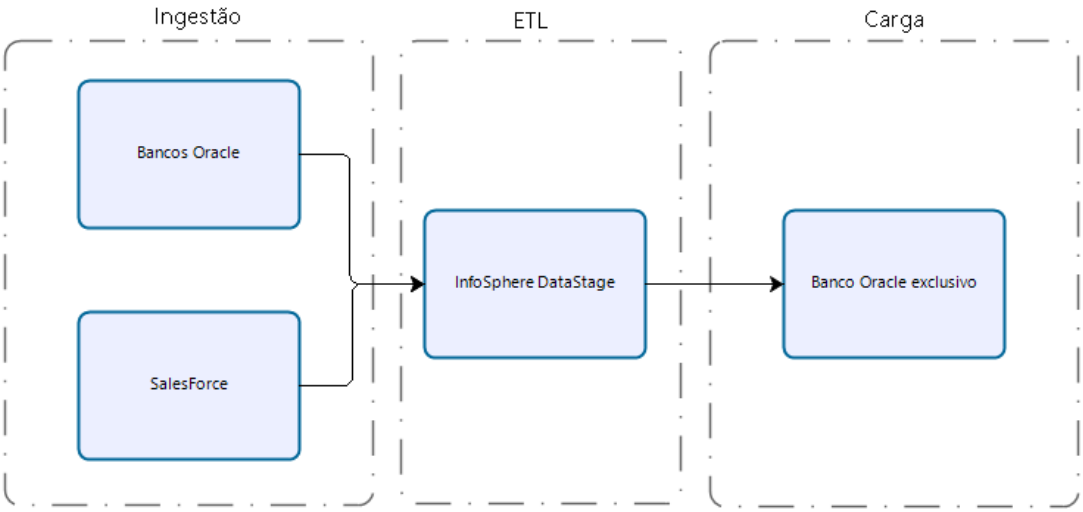
Por fim, é necessário treinar os funcionários da empresa para conseguirem utilizar todo o potencial disponível da ferramenta. Nessa fase, caso uma empresa terceira especializada na tecnologia tenha sido contratada para auxiliar na implementação, é importante que haja passagem de conhecimento para que o pessoal de TI interno da empresa possa administrar e sustentar tudo que foi implementado. Também é importante que os funcionários mais voltados ao negócio tenham acesso as ferramentas e tenham treinamentos (WATSON; WIXOM, 2007) para produzirem seus próprios insights e a empresa possa caminhar para um cenário de *self-service analytics* (DUTTA; BOSE, 2015).

5 ESTUDO DE CASO

5.1 ESTRUTURA DO SISTEMA DE BI

A empresa estudada já está no mercado a mais de 80 anos e possui um sistema de BI implementado a aproximadamente 10 anos. Esse sistema utiliza como fonte de origem dados de bancos Oracle de produção, com seus diversos módulos de emissão de nota, de planejamento financeiro, etc e do sistema de CRM SalesForce; em seguida, a ferramenta do IBM InfoSphereDataStage é responsável por realizar o processo de ETL; por fim, os dados são armazenados em outro banco Oracle, específico para área de BI da empresa. Na Figura 8 é possível observar o processo realizado de forma gráfica.

Figura 8 – Processo de carga, tratamento e carregamento dos dados na empresa estudada



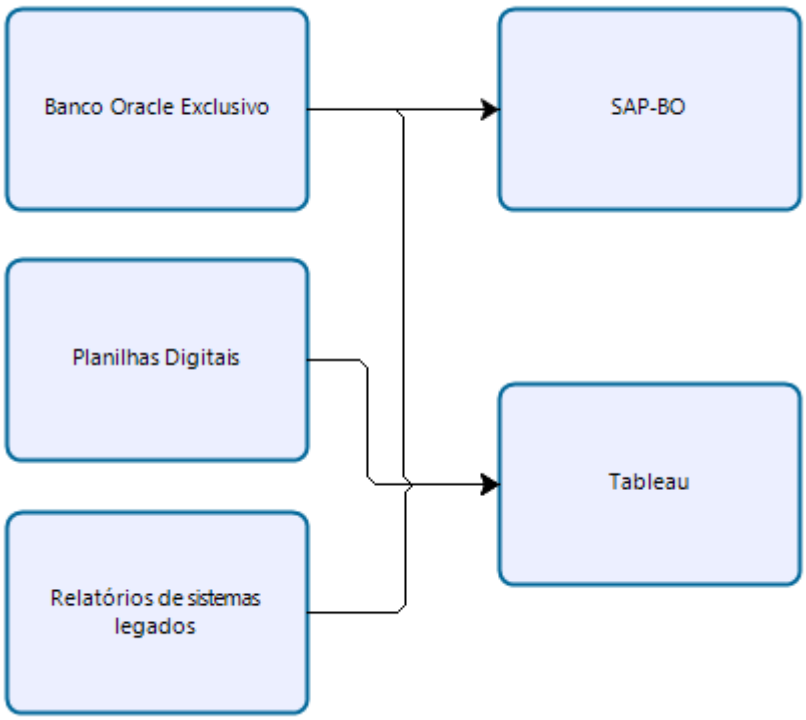
Fonte: O próprio autor.

Além desse processo estruturado automatizado de obtenção dos dados, existem muitos sistemas legados (aplicações utilizadas por áreas específicas e sem integração) e entradas de informações manuais, por meio de planilhas digitais.

Na parte de visualização de dados, a empresa trabalha oficialmente com duas ferramentas: SAP-BO e Tableau. Os relatórios produzidos e disponibilizados no SAP-BO possuem como fonte de dados apenas os dados presentes no banco específico para a área de BI. Já os dados que alimentam o Tableau são provenientes de várias fontes, desde bancos Oracle, passando por informações dos sistemas legados, até as próprias planilhas provenientes

de entradas de dados manuais. O fluxo de dados para as ferramentas de visualização pode ser observado na Figura 9.

Figura 9 – Fluxo de dados para as ferramentas de visualização



Fonte: O próprio autor.

A iniciativa para início de todo o projeto de implementação do *data lake* e modernização das ferramentas de BI teve início no ano de 2019, depois da percepção que era necessário a empresa conhecer mais seus clientes.

Ainda no final de 2019 novos projetos baseados em dados nasceram, como projetos de governança, indicadores e *churn*. É necessário dar destaque ao projeto de *churn*, que visa prever, com base em dados, quando algum cliente tem tendência em encerrar a parceria com a empresa, pois devido à alta demanda de *hot data* para esse projeto, foi contratado um provedor de serviços em nuvem para armazenar dados especificamente para essa iniciativa. Também é interessante destacar que a *holding* da empresa estudada possui um *data lake* e que essa *expertise* interna foi levada em conta nas análises internas da necessidade de modernização.

Entrando no ano de 2020, a gerência da área de tecnologia da informação (TI) foi renovada e os projetos iniciados estavam mais evoluídos, destacando ainda mais a necessidade de utilização de novas ferramentas. O novo gerente de TI trouxe consigo muitas

ideias e objetivos de modernização, como a transformação da empresa em um negócio *data driven* e a criação de um sistema de *self-service* BI

5.2 PROVA DE CONCEITO

O projeto de implementação do *data lake* foi iniciado com uma prova de conceito (POC), na qual foram escolhidos dois concorrentes: a Microsoft com sua plataforma de computação em nuvem Azure e a Amazon com sua plataforma AWS (Amazon Web Services). Outros concorrentes, como Google por exemplo, foram descartados devido a fatores como preço muito elevado, disponibilidade limitada e pouca presença no mercado brasileiro.

A POC foi composta de três fases: uma apresentação comercial, que deveria abordar teoricamente pontos previamente estabelecidos; uma apresentação técnica, onde deveriam ser apresentados, na prática, serviços previamente estabelecidos; e uma sessão de *benchmarks*, conduzida com três clientes de cada concorrentes, cujos nomes foram sugeridos pelos próprios concorrentes. Para avaliação da POC, foram chamados quatro integrantes da equipe de BI e 3 integrantes da equipe de negócio.

A primeira fase, da apresentação comercial, foi feita através de uma apresentação de cada concorrente, de até 4h. Nessa apresentação, deveriam ser apresentados os serviços disponíveis na plataforma, além de ser explicado teoricamente como é o funcionamento de algumas funcionalidades como ferramentas de integração, de ciência de dados, como são calculados os custos, entre outros. Ademais, nessa fase, todos os presentes tinham direito a voto com o mesmo peso, e o sistema de votação foi feito com base em pontos e características da apresentação, como tempo de apresentação, domínio sobre o conteúdo, clareza nas explicações, etc.

Na segunda fase, referente a apresentação técnica, foram passados alguns desafios para os concorrentes. Alguns deles eram: conexão, integração e extração de dados de um sistema de CRM (*CostumerRelationship Management*), aplicação das ferramentas de *machine learning* para análise de imagens e áudio, execução de código desenvolvido internamente para previsão de *churn*, entre outros. Após um período de duas semanas, os concorrente deveriam apresentar os resultados obtidos em uma apresentação de duas horas, na qual apenas a equipe de BI tinha direito a voto, pois foram avaliados pontos técnicos, como: facilidade de conexão, velocidade de execução, facilidade de uso das ferramentas disponíveis, entre outros.

Por fim, na última fase, referente aos *benchmarks*, foram feitas conversas com representantes técnicos de seis empresas, três sugeridas pela Microsoft e três pela Amazon. Nessa conversa, foram levantados pontos como satisfação da empresa contata com os serviços oferecidos por cada concorrente e grau de utilidade das ferramentas. Cada representante das empresas contatadas fez uma avaliação da empresa que os indicou.

Para avaliação final, também foram levados em conta alguns outros pontos significativos, como tecnologia já utilizada no grupo e na empresa. Por fim, foi preparada uma proposta comercial por parte de cada concorrente para a empresa estudada, que foi entregue à gerência e diretoria juntamente com uma consolidação da avaliação da POC.

5.3 PROBLEMA DO NEGÓCIO

A gerência da empresa estudada, apesar de possuir vários anos no mercado, começou a perceber que estava perdendo a vanguarda por possuir processos muito demorados e custosos, tanto internamente quanto externamente. A coleta de dados era feita do mesmo jeito a muitos anos, e foi percebido que a oportunidade de coletar novos dados que poderiam gerar mais insights estava sendo perdida.

Outro problema é relacionado ao modo como os dados eram tratados. A empresa trabalha com um sistema de CRM muito customizado e o tratamento de dados é feito por uma sequência de sistemas com estrutura construída a mais de uma década no passado. Por esse motivo, os indicadores analisados diariamente são os mesmo que foram pensados quando o sistema foi projetado.

Por fim, o outro problema a ser considerado é relacionado a qualidade e confiabilidade dos dados. Apesar de a empresa possuir um banco de dados específico para o sistema de BI, que retira informações de várias origens, ainda existiam muitas fontes de dados distintas na empresa e era comum que esses dados não estivessem completamente iguais.

Portanto, os problemas do negócio que o projeto de implementação do *data lake* visa resolver, são três:

- Obtenção e armazenamento de novos dados a respeito do cliente;
- Possibilidade de criação de novos indicadores e exploração dos dados; e
- Consolidação das fontes de dados da empresa, criando um repositório de dados central.

5.4 PESQUISA

Com os problemas a serem resolvidos definidos, os responsáveis pelo projeto iniciaram a parte de pesquisa. Nesse momento é importante ressaltar que devido ao tamanho da empresa estudada e a quantidade de dados, foi decidido no escopo do projeto que a migração de dados para nuvem seria realizada de forma gradual. Assim, nesta fase de pesquisa, foi estudado quais áreas estavam mais propensas a serem as primeiras a terem dados na nuvem, considerando uma série de fatores, como: quantidade de dados, facilidade de integração com as fontes de dados, tempo de retorno para o negócio, horas economizadas, entre outros.

Foram escolhidas 4 áreas para iniciarem o povoamento do *data lake:* marketing, experiência do cliente, governança empresarial e segurança, saúde e meio ambiente (SSMA). Dentro de cada uma dessas áreas foi feito um estudo profundo dos processos, de forma a estabelecer quais dados seriam os primeiros a serem levados para nuvem. Este estudo, consistiu em levantar quais processos poderiam ser beneficiados pelas informações estando na nuvem, levantar todas as fontes de dados, tabelas, tipo de integração, volumetria, etc, além do levantamento de todas as regras de negócio aplicada aos dados.

Por fim, foi desenhada uma arquitetura a ser seguida na ingestão dos dados.

5.5 TIME MULTIDISCIPLINAR

A etapa de formação do time foi realizada desde a idealização do projeto. Era de conhecimento de todos que um time multidisciplinar era essencial, portanto, desde o início havia integrantes de várias áreas da TI. Conforme as áreas atendidas pelo projetos foram sendo decididas na etapa de pesquisa, integrantes dessas áreas foram adicionados ao time e ao projeto. Somado aos colaboradores internos, foi contratada uma consultoria especialista em implementação de projetos de *big data* para auxiliar na etapa inicial de estruturação do ambiente e início de ingestões de dados.

5.6 *ROADMAP DO PROJETO*

Juntamente com a etapa anterior, a etapa de definição do *roadmap* do projeto também já estava sendo pensada desde o início. Foi decidido que o projeto seria realizado através de uma metodologia de gerenciamento de projetos ágil, isto é, o projeto é composto de várias *sprints* (fases que tem duração de 1 ou 2 semanas) em que no final delas deve ser entregue um

objetivo previamente definido. Assim, as sprints são dividias e criadas de forma a dividir objetivos grandes, como realizar a ingestão de dados da fonte “x”, em objetivos pequenos, como mapear o tipo de banco da fonte “x”, levantar qual o método de integração com essa fonte “x”, levantar qual a volumetria de dados da fonte “x”, etc. Na Figura 10 é possível observar uma ferramenta chamada Kanban, utilizada para gestão da metodologia ágil.

Figura 10 – Exemplo de Kanban utilizado para gestão de projetos ágeis

Backlog	*To do*	*Doing*	*Done*	*Block*

Fonte: o próprio autor.

Na Figura 10 é possível observar a estrutura básica de um Kanban: no quadro *backlog* são colocados todos os passos importantes do projeto, destrinchados no menor nível possível de tarefas; no quadro *To do* são inseridos quais itens do *backlog*serão tratados na*sprint* seguinte; no quadro *Doing*os responsáveis por uma tarefa do quadro *To do* anotam quais tarefas estão realizando; no quadro *Done*são colocados as tarefas já concluídas; e por fim, no quadro *Block* são colocadas as tarefas que não puderam ser realizadas devido a algum bloqueio fora da atuação dos integrantes do time.

É importante destacar o Kanban pois ele foi utilizado para gestão do projeto por duas das integrantes do time, intituladas de Product Owner (PO) e *Scrum Master* (SM). O PO é responsável por acompanhar o projeto como um todo e garantir que ele atenda as expectativa do(s) cliente(s) do projeto, ao passo que a função do SM é realizar a gestão e manutenção do Kanban e resolver quaisquer anotações que estejam no quadro *Block.*

O *roadmap* do projeto foi divido em basicamente 3 partes, e em seguida cada parte teve suas tarefas destrinchadas em *sprints*: *setup* da infraestrutura do ambiente, ingestão de dados e consumo de dados. As partes de ingestão e consumo de dados foram divididas por área entre os integrantes do time, de forma que todas pudessem ser desenvolvidas simultaneamente.

5.7 COLETA E ANÁLISE DE DADOS

Uma vez com o *roadmap* definido e as *sprints* começando, as ingestões de dados tiveram início após o *setup* da infraestrutura. Quais fontes teriam seus dados ingeridos foram decididas com base em muitas informações levantadas na etapa de pesquisa. Porém, nessa

fase, diferentemente da fase de pesquisa, as fontes de dados foram estudadas mais profundamente.

Com os parâmetros e particularidades de cada fonte de dados mapeadas, foi iniciada a ingestão de dados na área de homologação do *data lake*. Foram criados processos para extração automática dos dados e teste de conexão e capacidade de transferência. Então, foram testados alguns tratamentos nos dados, utilizando osingeridos na primeira camada do *data lake*, além de aplicadas as regras de negócio sugeridas por cada área. Os dados foram então validados, tanto pelos funcionários de TI quanto pelos funcionários do negócio responsáveis pelos dados.

5.8 ANÁLISE DE MODELAGEM DE DADOS

Partindo dos dados examinados na etapa anterior, foi definido que eles possuíram dois destinos: uma área de *sandbox* para cientista e analista de dados e a pré construção de indicadores já utilizados.

A camada *sandbox,* é um ambiente não produtivo para desenvolvimento. Nesse ambiente, os cientistas e analistas de dados podem ver vários tipos de dados sem ou com muito pouco tratamento, permitindo que novas visões e *insights* sejam gerados. Porém, esse ambiente será melhor explicado na fase de geração de *insights.*

Nessa fase, os dados foram tratados com regras de negócio e foram construídos indicadores e métricas muito utilizados pela empresa. Esses indicadores foram disponibilizados em um banco a parte, com modelo dimensional, que se aproxima muito dos bancos *on-premise* utilizados pela empresa. Desta forma, relatórios gerenciais tabulares, contendo informações imprescindíveis para a companhia ainda podem ser gerados e distribuídos para quem prefira essa visualização.

5.9 VISUALIZAÇÃO DE DADOS

Seguindo com o projeto, foi importante definir quais seriam as ferramentas de visualização de dados. Essa definição é importante pois dita como os usuários, sejam da área de TI, analistas de dados da área de negócio ou da ponta do negócio vão visualizar tanto os *insights* gerados na *sandbox* quanto os indicadores já existentes.

Essa fase apresentou um ponto de decisão importante para o time do projeto. Isso porque, a empresa possui uma ferramenta oficial de visualização de dados que é muito

utilizada internamente. Porém a empresa proprietária das tecnologias utilizadas no *data lake* possui uma tecnologia de visualização de dados própria, que apresenta vantagens em relação a integração com o armazenamento na nuvem. Foi decidido então que a empresa continuaria utilizando a ferramenta de visualização de dados oficial para indicadores estratégicos ou disponibilizados para a companhia inteira; ao passo que caso as áreas desejassem utilizar a outra ferramenta de visualização de dados para indicadores internos, seria permitido, porém, o custo para adquirir as licenças seria subtraído do orçamento daquela área em específico.

5.10 GERAÇÃO DE INSIGHTS

Um dos últimos passos na implementação de um projeto de *big data* é a geração de *insights*. Isso ocorre, porque é necessário que muita coisa esteja funcionando bem e que haja mão de obra qualificada para isso.

Como a companhia já tinha alguns cientistas e analistas de dados, o ambiente de *sandbox* para geração de novos *insights* foi disponibilizado assim que havia certeza que os dados disponibilizados lá eram confiáveis e concordantes com os outros sistemas da companhia. Essa fase se mostrou um marco grande para o projeto, pois um dos objetivos principais do projeto, de possibilitar a criação de novos indicadores e exploração de dados, foi alcançado.

Nesse ponto, também é importante destacar que com a disponibilização desse ambiente, a empresa se aproxima mais de um sistema de *self-service analytics*. Isso porque, apesar de não estar no escopo do projeto, existem planos para treinamento de pessoal interno do negócio, de forma que um usuário da ponta possa se conectar na camada *sandbox* e ele mesmo desenvolver algum indicador ou visão que julgue interessante, tirando quase que totalmente a dependência com a área de TI.

5.11 INTEGRAÇÃO COM O SISTEMA DE TI

Essa foi mais uma fase que foi desenvolvida desde o início do projeto e não apenas após as outras. A integração com a estrutura já existente de TI sempre foi um ponto de atenção durante o projeto e uma pauta constantemente levada para reuniões.

O time de infraestrutura da *holding* sempre esteve presente nas reuniões de infraestrutura e arquitetura, uma vez que alguns recursos são compartilhados. O time interno de infraestrutura também sempre participou ativamente das reuniões e definições. Porém,

como a responsabilidade de sustentação do ambiente será quase que total do time de BI, todas as decisões quanto as frentes sempre foram tomadas levando em consideração a opinião de todos do time de BI e do time responsável pelo projeto.

5.12 TREINAMENTO DE PESSOAS

A última etapa do projeto, pode-se dizer que sempre esteve presente durante o projeto também. Como foi contratada uma consultoria especializada para fornecer o *know-how* de uma implementação da *big data*, o time responsável pelo projeto foi capacitado pela consultoria para, no futuro, realizar processos e a sustentação do ambiente.

Vários usuários da ponta também foram capacitados em algumas novas ferramentas em que eles serão os responsáveis pelo manuseio, através de cursos e workshops. Usuários mais antigos foram treinados para adaptação as novas ferramentas e as possibilidades que elas levaram consigo.

5.13 CONCLUSÕES

Apesar deste ser o primeiro projeto de implementação de big data da empresa estudada e de os responsáveis pelo projeto não terem feito pesquisas sobre o tema na literatura, o projeto seguiu de forma quase que completa o framework proposto por Dutta e Bose (2015). Alguns pilares não foram desenvolvidos conforme recomendações dos autores e estes serão discorridos a seguir.

Na primeira etapa, de definição do problema do negócio, foi decidido por atacar três problemas de uma vez. Isso pode fazer com que o projeto se torne mais complexo (e consequentemente cada etapa se torne mais complexa também) que o necessário, se tratando de uma fase inicial que envolve a implantação de toda uma infraestrutura também. Dutta e Bose (2015) sugerem que seja definido e trabalhado apenas um problema de negócio e em seguida, quando a solução já estiver implantada, que sejam feitos outros projetos para resolução de outros problemas.

A formação do time multidisciplinar também não foi feita com muitas pessoas de áreas diferentes. Além dos funcionários do time de *business intelligence*, juntaram – se ao time uma gestora de projetos e 4 colaboradores do negócio (um de cada uma das áreas escolhidas para ingestão). Outros especialistas, sejam da área de TI ou do negócio, juntavam com o time

apenas pontualmente, ou seja, não podendo dar seu ponto de vista frequentemente em todos os temas, ao contrário do que sugerem os autores que propuseram o framework.

6 CONSIDERAÇÕES FINAIS

O objetivo principal proposto para este trabalho foi atendido, com a caracterização da solução de *business intelligence* implantada na empresa feita. Os objetivos específicos foram atendidos parcialmente: a definição das ferramentas utilizadas no framework foi realizada, porém o mapeamento da utilização de data lake na indústria se limitou a relatos presentes na literatura, não sendo realizado nenhum estudo de campo para se determinar a utilização dessas ferramentas na prática.

Com isso, este estudo pretende ser um dos pontos de partida para pesquisadores que anseiem por estudar projetos de aplicação de *big data* em empresas brasileiras, servindo como um caso real de aplicação. Para estudos estrangeiros, também pode servir como base de conhecimento da aplicação do projeto em uma empresa tradicional e de grande porte.

Por fim, este estudo limitou-se a comparar um projeto de implementação de *big data* com um framework proposto para implementação de projetos dessa natureza na literatura. Houveram outras limitações, como a utilização de um framework proposto em apenas um artigo e a utilização de apenas duas bases de dados para pesquisa de artigos. Para estudos futuros, sugere-se que pesquisadores tentem guiar um projeto de implementação de *big data* em uma empresa brasileira com base nesse mesmo framework, para que seja estudada a eficácia do mesmo. Também é interessante que seja feito um estudo na literatura comparando-se os frameworks propostos para implementação desse tipo de projeto.

REFERÊNCIAS

ALSERAFI, A.; ABELLO, A.; ROMERO, O.; CALDERS, T. Towards information profiling: data lake content metadata management. *In:* IEEE INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOPS (ICDMW), 16., 2016, Barcelona. **Proceedings** [...]. Barcelona, 2016. p. 178–185. Disponível em: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7836664. Acesso em: 12 abr. 2020.

ASHRAFI, A.; ZARE RAVASAN, A.; TRKMAN, P.; AFSHARI, S. The role of business analytics capabilities in bolstering firms' agility and performance. **International Journal of Information Management**, Oxford, v. 47, n. Jan., p. 1–15, 2019. Disponível em: https://doi.org/10.1016/j.ijinfomgt.2018.12.005. Acesso em: 15 mar. 2020.

BAARS, H.; KEMPER, H. Management support with structured and unstructured data: an integrated business intelligence framework. **Information Systems Management,** Nova Iorque, v. 25, n. 2, 2008, p. 132-148. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/10580530801941058. Acesso em: 12 abr. 2020.

BANSAL, S. K. Towards a semantic extract-transform-load (ETL) framework for big data integration. *In*: IEEE INTERNATIONAL CONGRESS ON BIG DATA**,** 14., 2014, Anchorage. **Proceedings** [...]. Anchorage, 2014. p. 522–529. Disponível em: https://ieeexplore.ieee.org/document/6906824. Acesso em: 14 abr. 2020.

BANSAL, S. K.; KAGEMANN, S. Integrating big data: a semantic extract-transform-load framework. **Computer**, Piscataway, v. 48, n. 3, p. 42–50, 2015. Disponível em: http://ieeexplore.ieee.org/document/7063172/. Acesso em: 15 abr. 2020.

BEAN, R.; KIRON, D. Organizational alignment is key to big data success. **MIT Sloan Management Review**, Massachusetts, n. 1, 2013. Disponível em: https://shop.sloanreview.mit.edu/store/organizational-alignment-is-key-to-big-data-success. Acesso em: 21 abr. 2020.

BEIER, M. Startups' experimental development of digital marketing activities: a case of online-videos. *In:* INTERDISCIPLINARY EUROPEAN CONFERENCE ON ENTERPRENEURSHIP RESEARCH (IECER), 10., 2016, Chur. **Proceedings** [...]. Chur, 2016. p. 1–8. Disponível em: https://ssrn.com/abstract=2868449. Acesso em: 01 mar. 2020

BIZER, C.; BONCZ, P.; BRODIE, M. L.; ERLING, O. The meaningful use of big data: four perspectives: four challenges. **SIGMOD Record**, Nova Iorque, v. 40, n. 4, p. 56–60, 2011. Disponível em: https://dl.acm.org/doi/10.1145/2094114.2094129#sec-ref. Acesso em: 04 mar. 2020.

BOSE, I.; PAL, R.; YE, A. ERP and SCM systems integration: the case of a valve manufacturer in China. **Information & Management**, Amsterdam, v. 45, p. 233–241, 2008. Disponível em: https://www.sciencedirect.com/science/article/pii/S0378720608000335. Acesso em: 20 mar. 2020

CARAÇA, J.; LUNDVALL, B. Å.; MENDONÇA, S. The changing role of science in the

innovation process: from queen to cinderella? **Technological Forecasting and Social Change**, Nova Iorque, v. 76, n. 6, p. 861–867, 2009. Disponível em: http://dx.doi.org/10.1016/j.techfore.2008.08.003. Acesso em: 26 mar. 2020.

CHANG, V.; VALVERDE, R.; RAMACHANDRAN, M.; LI, C. S. Toward business integrity modeling and analysis framework for risk measurement and analysis. **Applied Sciences**, Basel, v. 10, n. 9, p. 1-24, 2020. Disponível em: https://www.mdpi.com/2076-3417/10/9/3145. Acesso em: 13 maio 2020.

CHAUDHURI, S.; DAYAL, U. An overview of data warehousing and olap technology. **SIGMOD Record**, Nova Iorque, v. 26, n. 1, p. 65–74, 1997. Disponível em: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/sigrecord.pdf. Acesso em: 02 mar. 2020.

CLARKE, P.; TYRRELL, G.; NAGLE, T. Governing self service analytics. **Journal of Decision Systems**, Abingdon, v. 25, p. 145–159, 2016. Disponível em: http://dx.doi.org/10.1080/12460125.2016.1187385. Acesso em: 23 mar. 2020.

COUTO, J.; BORGES, O.; RUIZ, D.; MARCZAK, S.; PRIKLADNICKI, R. A mapping study about data lakes: an improved definition and possible architectures. *In:* INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING AND KNOWLEDGE ENGINEERING (SEKE), 31., 2019, Lisboa. **Proceedings** [...]. Nova Iorque, July 2019. p. 453–458. Disponível em: http://ksiresearchorg.ipage.com/seke/seke19paper/seke19paper_129.pdf. Acesso em: 01 maio 2020.

DAVENPORT, T. H.; DYCHÉ, J. Big data in big companies. **International Institute For Analytics**, Portland, v. 1, n. 1, p. 1–31, 2013. Disponíevel em: https://www.iqpc.com/media/7863/11710.pdf. Acesso em: 03 mar. 2020.

DINSMORE, T. W. **Disruptive analytics**. Berkeley: Apress, 2016.

DIXON, J. **Pantaho, Hadoop, and Data Lakes**. 2010. Disponível em: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/. Acesso em: 9 mar. 2020.

DUTTA, D.; BOSE, I. Managing a big data project: the case of ramco cements limited. **International Journal of Production Economics**, Amsterdam, v. 165, p. 293–306, 2015. Disponível em: http://dx.doi.org/10.1016/j.ijpe.2014.12.032. Acesso em: 14 mar. 2020.

FANG, H. Managing data lakes in big data era: what's a data lake and why has it became popular in data management ecosystem. *In:* IEEE INTERNATIONAL CONFERENCE ON CYBER TECHNOLOGY IN AUTOMATION, 15., 2015, Shenyang. **Proceedings**[...]. Nova Iorque, 2015. p. 820–824 Disponível em: https://ieeexplore.ieee.org/document/7288049. Acesso em: 03 mar. 2020

FARID, M.; ROATIŞ, A.; ILYAS, I. F.; HOFFMANN, H. F.; CHU, X. CLAMS: Bringing quality to data lakes. *In:* INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 8., 2016, São Francisco. **Proceedings** [...]. São Francisco, 2016. p. 2089–2092. Disponível em: https://dl.acm.org/doi/10.1145/2882903.2899391. Acesso em: 21 abr. 2020.

FLYVBJERG, B. Quality control and due diligence in project management: getting decisions

right by taking the outside view. **International Journal of Project Management**, Oxford, v. 31, n. 5, p. 760–774, 2013. Disponível em: http://dx.doi.org/10.1016/j.ijproman.2012.10.007. Acesso em: 21 abr. 2020.

GUAMÁN, M. A. A.; VACA, M. J. N.; YUQUILEMA, J. Fr. B. Mapeo sistematico de literatura de un data lake. **mktDESCUBRE**, Riobamba, v. 11, n. 1, p. 50–66, 2018. Disponível em: http://revistas.espoch.edu.ec/index.php/mktdescubre/article/view/153. Acesso em: 03 mar. 2020.

HETEMI, E.; JERBRANT, A.; MERE, J. O. Exploring the emergence of lock-in in large-scale projects: a process view. **International Journal of Project Management**, Oxford, v. 38, n. 1, p. 47–63, 2020. Disponível em: https://doi.org/10.1016/j.ijproman.2019.10.001. Acesso em: 15 mar. 2020.

HINDLE, G. A.; VIDGEN, R. Developing a business analytics methodology: a case study in the foodbank sector. **European Journal of Operational Research**, Amsterdam, v. 268, p. 836–851, 2018. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0377221717305702. Acesso em: 16 mar. 2020.

INMON, W. H. **Building the data warehouse**. 3. ed. Nova Iorque: John Wiley & Sons, 2002.

KAMBATLA, K.; KOLLIAS, G.; KUMAR, V.; GRAMA, A. Trends in big data analytics. **Journal of Parallel and Distributed Computing**, Maryland Heights, v. 74, n. 7, p. 2561–2573, 2014. Disponível em: http://dx.doi.org/10.1016/j.jpdc.2014.01.003. Acesso em: 06 maio 2020.

KAUR, H.; SINGH, S. P. Heuristic modeling for sustainable procurement and logistics in a supply chain using big data. **Computers and Operations Research**, Oxford, v. 98, p. 301–321, 2018. Disponível em: https://doi.org/10.1016/j.cor.2017.05.008. Acesso em: 04 mar. 2020.

KEIM, D. A.; MANSMANN, F.; SCHNEIDEWIND, J.; ZIEGLER, H. Challenges in visual data analysis. *In:* INTERNATIONAL CONFERENCE ON INFORMATION VISUALISATION, 10., 2006, Londres. **Proceedings** [...]. Londres, 2006. p. 9–14. Disponível em: http://kops.uni-konstanz.de/bitstream/handle/123456789/5515/IV2006.pdf?sequence=1. Acesso em: 04 abr. 2020.

KOHAVI, R.; ROTHLEDER, N. J.; SIMOUDIS, E. Emerging trends in business analytics. **Communications of the ACM**, Nova Iorque, v. 45, n. 8, p. 45–48, 2002. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.456.5008&rep=rep1&type=pdf. Acesso em: 15 mar. 2020.

KOTHARI, C. R. **Research methodology:** methods and techniques. 2. ed. Nova Delhi: New Age International, 2004.

LEE, J.; LAPIRA, E.; BAGHERI, B.; KAO, H. Recent advances and trends in predictive manufacturing systems in big data environment. **Manufacturing Letters**, Amsterdam, v. 1, n. 1, p. 38–41, 2013. Disponível em: http://dx.doi.org/10.1016/j.mfglet.2013.09.005, Acesso em: 01 abr. 2020.

LÖNNQVIST, A.; PUHAKKA, V. The measurement of business intelligence. **Information Systems Management**, Nova Iorque, v. 40, n. 3, p. 1–14, 2009. Disponível em: https://www.researchgate.net/publication/220630156_The_Measurement_of_Business_Intelligence. Acesso em: 03 mar. 2020.

MALYSIAK-MROZEK, B.; STABLA, M.; MROZEK, D. Soft and declarative fishing of information in big data lake. **IEEE Transactions on Fuzzy Systems**, Piscataway, v. 26, n. 5, p. 2732–2747, 2018. Disponível em: https://ieeexplore.ieee.org/document/8314734. Acesso em: 15 mar. 2020.

MANAGEMENT SOLUTION. **Data science y la transformación del sector financiero**. Barcelona: Management Solution, 2015. *E-book.*

MIGUEL, P. A. C. Processos: uma abordagem da engenharia para a gestão de operações. *In:* MIGUEL, P. A. C. (coord). **Metodologia de pesquisa em engenharia de produção.** 3. ed. Rio de Janeiro: Elsevier, 2018. v. 1, cap. 9, p. 197-213.

MILOSLAVSKAYA, N.; TOLSTOY, A. Big data, fast data and data lake concepts. **Procedia Computer Science**, Amsterdam, v. 88, p. 300–305, 2016. Disponível em: http://dx.doi.org/10.1016/j.procs.2016.07.439. Acesso em: 06 mar. 2020.

MUNSHI, A. A.; MOHAMED, Y. A. R. I. Data lake lambda architecture for smart grids big data analytics. **IEEE Access**, Piscataway, v. 6, p. 40463–40471, 2018. Disponível em: https://ieeexplore.ieee.org/document/8417407. Acesso em: 06 mar. 2020.

MZAGHLOUL, M.; ALI-ELDIN, A.; SALEM, M. Towards a self-service data analytics framework. **International Journal of Computer Applications**, Nova Iorque, v. 80, n. 9, p. 41–48, 2013. Disponível em: https://research.ijcaonline.org/volume80/number9/pxc3891840.pdf. Acesso em: 01 abr. 2020.

O'LEARY, D. E. Artificial intelligence and big data. **IEEE Intelligent Systems**, Piscataway, v. 28, n. 2, p. 96–99, 2013. Disponível em: https://ieeexplore.ieee.org/abstract/document/6547979. Acesso em: 18 mar. 2020.

O'LEARY, D. E. Embedding AI and crowdsourcing in the big data lake. **IEEE Intelligent Systems**, Piscataway, v. 29, n. 5, p. 70–73, 2014. Disponível em: https://ieeexplore.ieee.org/document/6949519. Acesso em: 15 mar. 2020

PERWEJ, Y. An experiential study of the big data. **International Transaction of Electrical and Computer Engineers System**, Newark, v. 4, n. 1, p. 14–25, 2017. Disponível em: http://pubs.sciepub.com/iteces/4/1/3. Acesso em: 03 mar. 2020.

PROVDANOV, C. C.; FREITAS, E. C. **Metodologia do trabalho científico:** métodos e técnicas da pesquisa e do trabalho acadêmico. 2. ed. Novo Hamburgo: Universidade Fevale, 2013.

PROVOST, F.; FAWCETT, T. Data science and its relationship to big data and data-driven decision making. **Big Data**, Nova Rochelle, n. 1, p. 51–59, 2013. Disponível em: https://www.liebertpub.com/doi/full/10.1089/big.2013.1508. Acesso em: 06 mar. 2020.

SCOPUS. **Base de dados bibliográficos.** Disponível em: https://www.scopus.com. Acesso em: 03 mar. 2020.

STIEGLITZ, S.; MIRBABAIE, M.; ROSS, B.; NEUBERGER, C. Social media analytics: challenges in topic discovery , data collection , and data preparation. **International Journal of Information Management**, Oxford, v. 39, p. 156–168, 2018. Disponível em: https://doi.org/10.1016/j.ijinfomgt.2017.12.002. Acesso em: 15 mar. 2020.

VIAENE, S.; VAN DEN BUNDER, A. The secrets to managing business analytics projects. **MIT Sloan Management Review**, Cambridge, v. 53, n. 1, p. 65–69, 2011. Disponível em: https://www.researchgate.net/publication/230800948_The_Secrets_to_Managing_Business_Analytics_Projects. Acesso em: 06 mar. 2020.

WALKER, C.; ALREHAMY, H. Personal data lake with data gravity pull. *In:* INTERNATIONAL CONFERENCE ON BIG DATA AND CLOUD COMPUTING, 5., 2015, Shanghai. **Proceedings** [...]. Piscataway, 2015. p. 160–167. Disponível em: https://www.researchgate.net/publication/283053696_Personal_Data_Lake_With_Data_Gravity_Pull. Acesso em: 16 mar. 2020.

WAMBA, S. F.; GUNASEKARAN, A.; AKTER, S.; REN, S. J.; DUBEY, R.; CHILDE, S. J. Big data analytics and firm performance: effects of dynamic capabilities. **Journal of Business Research**, Nova Iorque, v. 70, p. 356–365, 2017. Disponível em: http://dx.doi.org/10.1016/j.jbusres.2016.08.009. Acesso em: 06 mar. 2020.

WATSON, H. J.; WIXOM, B. H. The current state of business intelligence. **Computer**, Piscataway, v. 40, p. 96-99, 2007. Disponível em: https://ieeexplore.ieee.org/document/4302625. Acesso em: 03 mar. 2020.

WEB OF SCIENCE. **Base de dados bibliográficos.** Disponível em: http://login.webofknowledge.com. Acesso em: 03 mar. 2020.

ZHANG, Q.; YANG, L. T.; CHEN, Z.; LI, P. A survey on deep learning for big data. **Information Fusion**, Amsterdam, v. 42, p. 146–157, 2018. Disponível em: https://doi.org/10.1016/j.inffus.2017.10.006. Acesso em: 01 abr. 2020.