UNIVERSIDADE DO SUL DE SANTA CATARINA

GIOVANI REINERT JUNIOR

CAMINHANDO PARA O *BIG DATA*: ESTRUTURAÇÃO DE DADOS EM UMA EMPRESA DO SEGMENTO INDUSTRIAL

Palhoça

2021

GIOVANI REINERT JUNIOR

CAMINHANDO PARA O *BIG DATA*: ESTRUTURAÇÃO DE DADOS EM UMA EMPRESA DO SEGMENTO INDUSTRIAL

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Sistemas da Computação da Universidade do Sul de Santa Catarina, como requisito parcial à obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Daniella Pinto Vieira, MEng

Palhoça

2021

GIOVANI REINERT JUNIOR

CAMINHANDO PARA O *BIG DATA*: ESTRUTURAÇÃO DE DADOS EM UMA EMPRESA DO SEGMENTO INDUSTRIAL

Este Trabalho de Conclusão de Curso foi julgado adequado à obtenção do título de Bacharel em Sistemas da Computação e aprovado em sua forma final pelo Curso de Graduação em Sistemas de Informação da Universidade do Sul de Santa Catarina.

Palhoça, 14 de junho de 2021.

_____

Professor e orientador Daniella Pinto Vieira, MEng .
Universidade do Sul de Santa Catarina

_____

Prof. Flavio Ceci, Dr.
Universidade do Sul de Santa Catarina

_____

Prof. Aran Bey Tcholakian Morales, Dr.
Universidade do Sul de Santa Catarina

Dedico este trabalho à minha família, que sempre esteve do meu lado, deram-me amor, compreensão em todos os momentos e ajudaram a construir este projeto que muda a minha vida. A professora Daniella pela ajuda e atenção ao orientar este trabalho.

AGRADECIMENTOS

A esta universidade, seu corpo docente, direção e administração pela oportunidade que me foi dada em ampliar o meu conhecimento na minha área de atuação profissional.

A minha orientadora Prof. Daniella Pinto Vieira, pelo suporte no tempo que lhe coube, pelas suas correções e incentivos.

À nossa coordenadora do curso de Sistemas de Informação, Prof. Vera Schuhmacher, por sempre pensar nos alunos e buscar o melhor para nós.

A minha namorada, Jaqueline da Silveira, por toda atenção e paciência que teve para me amparar.

Aos meus pais, pelo amor, incentivo e apoio incondicional. E a todos que direta ou indiretamente fizeram parte da minha formação, o meu muito obrigado.

A minha equipe de trabalho, que me acolheu de uma forma a qual jamais vou esquecer, e toda ajuda e apoio que me foi dado.

Agradeço a todos os meus professores do curso de Sistemas de Informação da Universidade do Sul de Santa Catarina, por todo o conhecimento compartilhado ao longo dessa jornada.

A todos os meus amigos, por entenderem minhas ausências, meus momentos de estresse e cansaço.

RESUMO

Após as grandes progressões provenientes das revoluções industriais, o segmento fabril está passando por uma exponencial transformação digital, na qual, utiliza de tecnologias e ferramentas que unificam cada vez meus seus segmentos em busca de resultados mais assertivos perante a concorrência desenfreada. Com todo investimento dedicado ao crescimento tecnológico da indústria, fica evidente a necessidade de ferramentas capazes de ler, processar e demonstrar todos os dados gerados de inúmeras formas possíveis. Para desenvolver uma arquitetura *big data* que se adeque a necessidade e as limitações da empresa industrial, foi necessário entender conceitos, atributos de *big data*, assim como ferramentas e linguagens que contribui para a implementação de um ecossistema. Neste contexto, o presente trabalho de conclusão de curso tem como objetivo explorar o campo de conhecimento da arquitetura de *big data,* visando demonstrar como aplicá-la no segmento industrial de modo a demonstrar os benefícios da aplicação desta tecnologia. Os resultados apresentados na realização da prova de conceito demonstraram que a implementação de um *big data* é expressiva e soluciona problemas reais de empresas do segmento industrial, deixando as análises simples e trazendo a garantia da integridade das informações para tomadas de decisões, trazendo ganho na apresentação de dados confiáveis para tomada de decisão.

Palavras-chave: *Big Data*, ETL, Indústria.

ABSTRACT

After the great progressions arising from the industrial revolutions, the manufacturing segment is undergoing an exponential digital transformation, in which it uses technologies and tools that unify my segments each time in search of more assertive results in the face of unbridled competition. Any investment dedicated to the technological growth of the industry, the need for tools capable of reading, processing and demonstrating all the data generated in possible ways is evident. To develop a *big data* architecture that suits the needs and limitations of the industrial company, it was necessary to understand the concepts, attributes of *big data*, as well as tools and languages that contribute to the implementation of an ecosystem. In this context, the present work of completion of course aims to explore the field of knowledge of *big data* architecture, describes how to apply it in the industrial segment in order to demonstrate the benefits of applying this technology. The results achieved in carrying out the proof of concept demonstrated that the implementation of a *big data* is expressive and solves real problems of companies in the industrial segment, leaving it as simple analysis and bringing the guarantee of integrity of the information to obtain decisions, bringing gain in the presentation of data taken for decision making.

Keyword: *Big Data*, ETL, Industry.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIAÇÕES E SIGLAS

API..... APPLICATION PROGRAMMING INTERFACE
BI.....BUSINESS INTELLIGENCE
CNI..... CONFEDERAÇÃO NACIONAL DA INDUSTRIA
CRM..... CUSTOMER RELATIONSHIP MANAGEMENT
IA……..... INTELIGENCIA ARTIFICIAL
JSON……... JAVASCRIPT OBJECT NOTATION
SQL……..... STRUCTURED QUERY LANGUAGE
ERP.....ENTERPRISE RESOURCE PLANNING
FTP.....FILE TRANSFER PROTOCOL

SUMÁRIO

1 INTRODUÇÃO

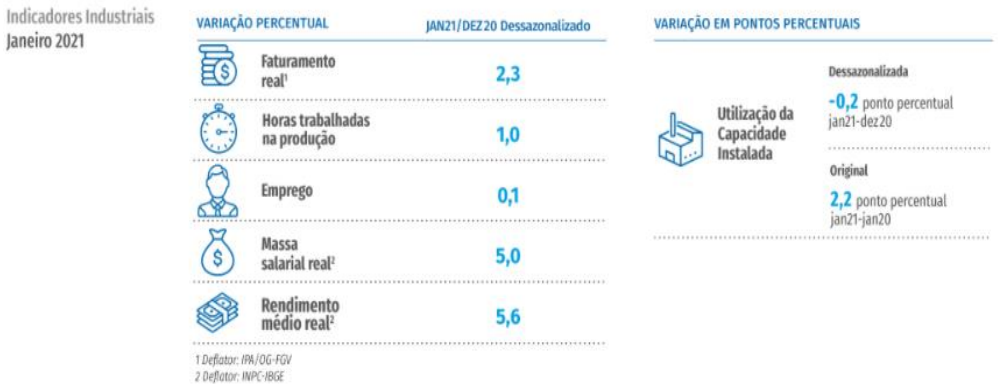
No contexto da primeira revolução industrial até os dias atuais, segundo Gonçalves (1994), pode-se afirmar que a forma de se trabalhar sofreu muitas mudanças, levando em consideração os adventos da tecnologia da informação em relação aos aspectos sociais das empresas (TAPSCOTT, 2010). Com a crescente globalização e evolução das tecnologias percebe-se a expansão exponencial do volume de dados gerados (SOMASUNDARAM; SHRIVASTAVA, 2011). Dependendo da quantidade de dados e suas características é possível denominar esse grande volume de informações como *big data*.

A definição de Gartner (2012), sobre *big data* é a união de "ativos de alto volume, velocidade e variedade de informação que demandam custo-benefício, formas inovadoras de processamento de informação para maior visibilidade e suporte à tomada de decisão".

Conforme a imensa quantidade de dados que as empresas geram é perceptível o quanto a utilização de ferramentas de processamento em massa vem crescendo. Davenport e Kim (2014, p.2) destacam que o crescimento do *big data* está ocorrendo em praticamente todas as áreas econômicas mundiais.

Colocando em ênfase o segmento industrial, a Confederação Nacional da Indústria (CNI) afirma que a trajetória das atividades industriais do início de 2021 são superiores ao mesmo período de 2020, demonstrando na **Erro! Fonte de referência não encontrada.** indicadores que comprovam esse crescimento.

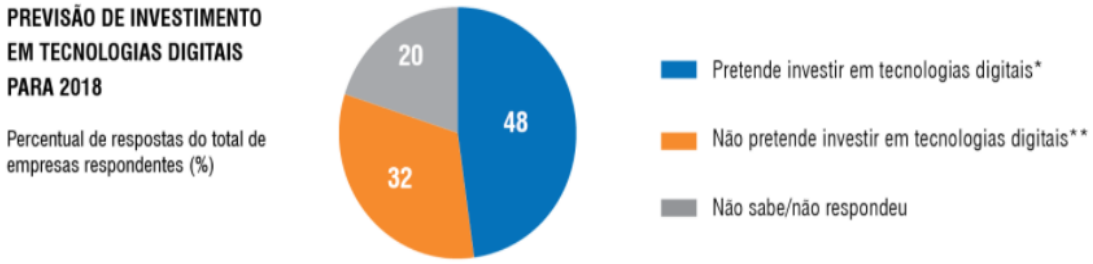
Figura 1 - Indicadores de crescimento das Indústrias



Fonte: Confederação Nacional da Indústria (2021)

De acordo com dados da CNI (2018), entre o início dos anos de 2016 e 2018, a quantidade de indústrias brasileiras que utilizam tecnologias digitais cresceu 10%, além do apontamento onde 48% das industrias pretendem investir em tecnologias digitais conforme a figura 2, mostrando uma forte tendência de crescimento.

Figura 2 - Previsão de investimentos em tecnologias digitais

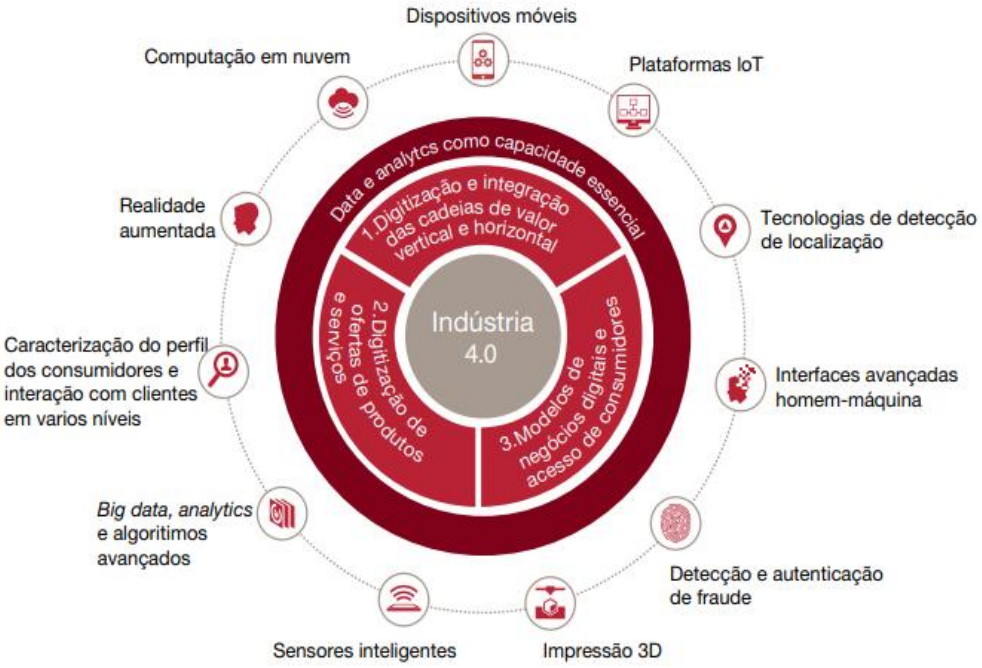


Fonte: Confederação Nacional da Indústria (2018)

O site intitulado Ind 4.0 (2020), informa que cada dia mais indústrias estão utilizando análises eficientes provenientes de informações contidas em *big data*, as quais trazem apoio à inovação, bem como desenvolvimento, e, gerenciamento orientado a dados.

Segundo a PwC Brasil (2020), está acontecendo uma transformação digital nas principais empresas industriais e de manufatura do mundo, o qual é resultado de uma necessidade exponencial por informações que resultem crescimento. Na **Erro! Fonte de referência não encontrada.**, pode-se observar que "*big data*, *analytics* e algoritmos avançados", fazem parte do contexto de tecnologias digitais da evolução industrial 4.0.

Figura 3 - Tecnologias digitais colaborativas da Indústria 4.0



Fonte: PwC Brasil (2020).

Em decorrência ao grande volume de dados armazenados de diversas formas possíveis, no ambiente corporativo, fica evidente a necessidade de encontrar tecnologias que permitam tirar o melhor proveito das informações guardadas, principalmente as que contribuam para o processo de tomada de decisão mais assertivas.

Com isso, pretende-se, com essa pesquisa, fazer uma análise para compreender as possibilidades de se aplicar uma arquitetura de *big data* em empresas do segmento industrial. Esta arquitetura deve ter ênfase em uma estrutura projetada para atender de forma correta o processo de extração, tratamento, e análise de dados, para, por fim, auxiliar de forma efetiva o processo de tomada de decisão.

Neste sentido, o presente trabalho trata da proposta da implementação de uma arquitetura *big data* no segmento industrial para melhor utilização das informações.

1.1 PROBLEMA DE PESQUISA

Perante as evoluções de pensamentos comerciais, Gantz e Reinsel (2012) consideram que a maioria das organizações requerem uma linguagem e informações de forma nítida sobre como utilizar *big data* em seus negócios e como essa arquitetura implementada pode influenciar de forma positiva nas tomadas e decisões. Ao encontro com este último argumento, a pesquisa da Unit (2014) aponta que 75% das lideranças organizacionais têm em mente que as empresas necessitam de uma reestruturação de pensamentos, voltada a orientar suas decisões por meio de dados, com o intuito de terem o controle das informações relevantes para a empresa.

Todavia, é importante ressaltar que com esses inúmeros dados armazenados, pode-se fazer com que em uma empresa do segmento industrial, estruture uma organização de produção mais eficiente, tornando a empresa como um todo mais eficaz nos demonstrativos e na lucratividade.

Os dados podem possuir diversas formas, como arquivos e imagens, podendo não haver padrões, sendo estruturados ou não estruturados, dificultando ainda mais, uma consulta de informações de forma simples.

A tecnologia de *big data* pode lidar com uma grande quantidade de dados, com isso, proporcionando meios mais fáceis para a análise de dados, os quais são de suma importância para empresas tomarem decisões que proporcionar grande evolução em sua cadeia de valor.

Com isso, o segmento industrial brasileiro poderia se beneficiar de análises de dados mais precisas de modo a aumentar a produtividade em toda a sua cadeia de valor. Neste sentido, é possível estabelecer perguntas de pesquisa que norteiam este trabalho, a saber: (a) Como a arquitetura de *big data* pode ser desenhada para oferecer uma visão sistêmica relacionada a gestão da indústria? (b) Como implantar uma arquitetura de *big data* que atenda as características das empresas do segmento industrial?

1.2 OBJETIVOS

Para que as perguntas de pesquisa propostas possam ser respondidas, é necessário ter a definição de objetivos tangíveis para haver um direcionamento eficaz. Assim, são definidos como objetivo geral e específicos:

1.2.1 Objetivo geral

O objetivo geral da pesquisa é explorar o campo de conhecimento da arquitetura de *big data* visando demonstrar como aplicá-la no segmento industrial.

1.2.2 Objetivos específicos

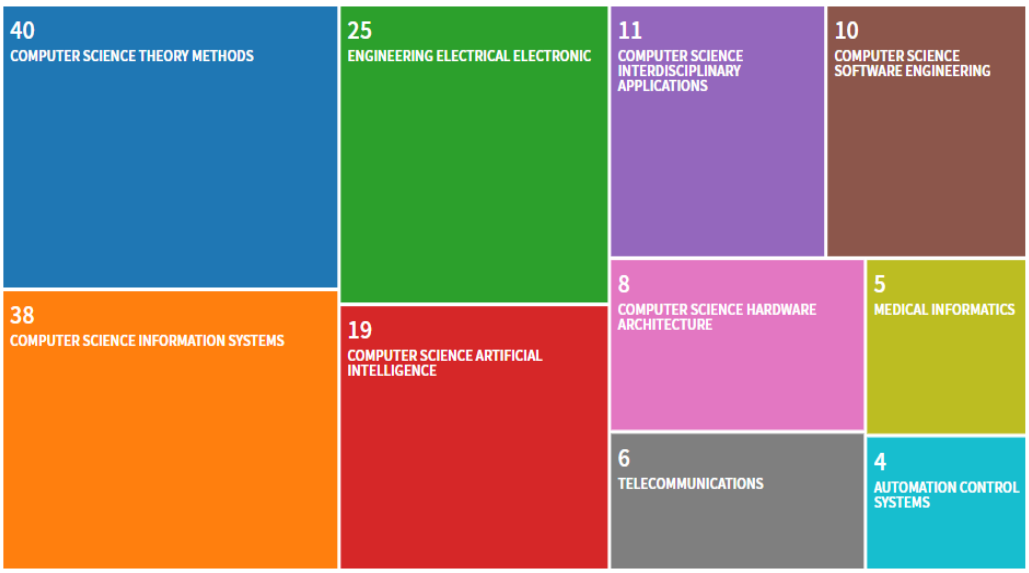
Para a conclusão do objetivo geral foram definidos um conjunto de objetivos específicos, sendo eles:

a) realizar uma pesquisa bibliográfica sobre os principais conceitos de *big data*;
b) demonstrar os benefícios da implementação da arquitetura *big data*, bem como seus desafios no cenário de uma empresa do segmento industrial.
c) demonstrar, por meio de uma prova de conceito, a aplicação do modelo da arquitetura *big data* proposto para uma empresa do segmento industrial.

1.3 JUSTIFICATIVA

Segundo a plataforma *Web of Science*, na pesquisa por palavras-chaves "*extract*" + "*transform*" + "*load*" + "*big data*", foram identificados 102 trabalhos publicados. Tais trabalhos em sua grande maioria categorizam-se nas áreas de conhecimento de sistemas da informação e ciências da computação, conforme ilustra a **Erro! Fonte de referência não encontrada.**. O que demonstra a convergência do tema proposto com a área de sistemas de informação.

Figura 4 - Artigos publicados por área segundo a Web of Science



Fonte: *Web Of Science* (mar, 2021).

Das publicações citadas,102 no total da pesquisa pode-se observar que o *h-index*[1] é igual a 11, e a média de citações por item é igual a 4,94, conforme a **Erro! Fonte de referência não encontrada.**. O que demonstra que o tema está na vanguarda das publicações técnico-cientificas.

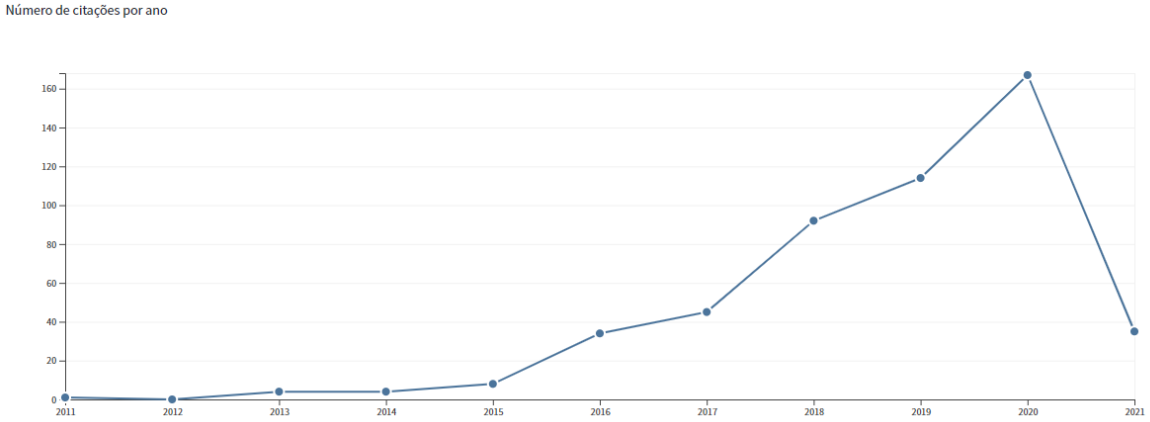
Figura 5 - Índices de medição de busca

Fonte: *Web Of Science* (mar, 2021).

Além disso, conforme demonstrado no gráfico da **Erro! Fonte de referência não encontrada.**, a partir de 2014 é possível observar uma forte tendência na publicação de trabalhos que conjugam as palavras-chave citadas na busca.

[1] O h-index, segundo a Web Of Science, é baseado em uma lista de publicações classificadas em ordem decrescente pela contagem de números citados. O valor de h é igual ao número de artigos. Esta métrica é útil porque desconta o peso desproporcional de artigos altamente citados ou artigos que ainda não foram citados.

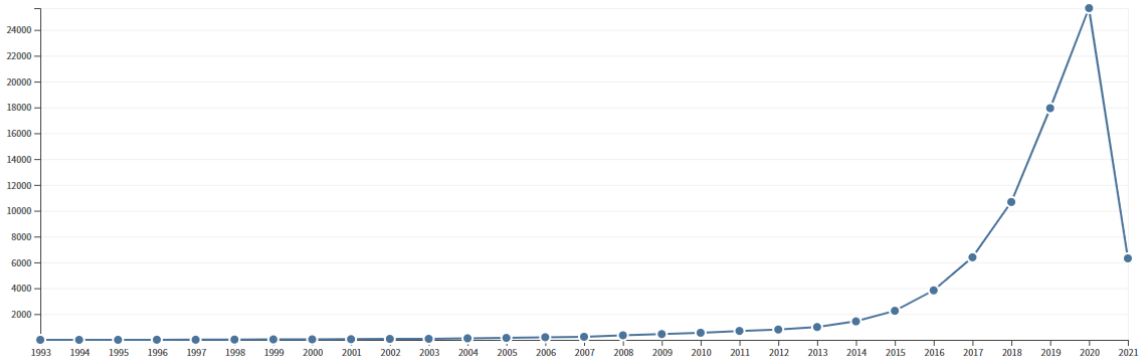
Figura 6 - Gráfico de buscas crescente sobre *big data* e ETL



Fonte: *Web Of Science* (mar, 2021).

A consultoria PwC (2020) aponta que mais de 70% das fábricas de todo o mundo vão utilizar algo de tecnologia *big data* na indústria 4.0. Também de acordo com a *Web of Science* (2021) é possível observar a crescente expoente das palavras "*big data*" + "industry" na **Erro! Fonte de referência não encontrada.**, onde a tendência demonstra fortes indícios de estudos e implementações na convergência destas áreas.

Figura 7 - Gráfico de buscas crescente sobre *big data* e indústria



Fonte: Web Of Science (mar, 2021).

Dado o exposto, justifica-se o tema deste trabalho considerando que (a) o mesmo está contido no domínio de conhecimento do curso de sistemas de informação, e, (b) instiga a produção de trabalhos de pesquisa técnicos-científicos devido à forte tendência de crescimento de publicações nesta temática.

1.4 ESTRUTURA DA MONOGRAFIA

O presente trabalho é composto por seis capítulos que retratam o desenvolvimento da pesquisa. O Capítulo 1 descreve a proposta do trabalho por meio da descrição dos seus objetivos, tanto o geral, quanto os específicos, junto com a problemática e a justificativa. Logo após, no Capítulo 2, apresenta o resultado da pesquisa bibliográfica, onde são abordados os princípios do segmento industrial, os atributos e tecnologias do *big data*, visando atender ao primeiro objetivo específico. No Capítulo 3, é apresentado o método de desenvolvimento do trabalho, com a arquitetura proposta. Já no Capítulo 4, é apresentada uma proposta de solução após análise do modelo atual utilizado pela empresa e seu ambiente organizacional. No Capítulo 5, é apresentado o desenvolvimento da solução, explicando a utilização de cada ferramenta e a metodologia escolhida, afim de colher dados e apresentar resultados. Por fim, no Capítulo 6 são apresentadas as conclusões e trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo apresenta o resultado da pesquisa bibliográfica realizada considerando os temas vinculados ao presente trabalho. A seguir são abordos os princípios do segmento industrial, os conceitos e as tecnologias relacionadas ao *big data*, bem como as ferramentas para sua implementação.

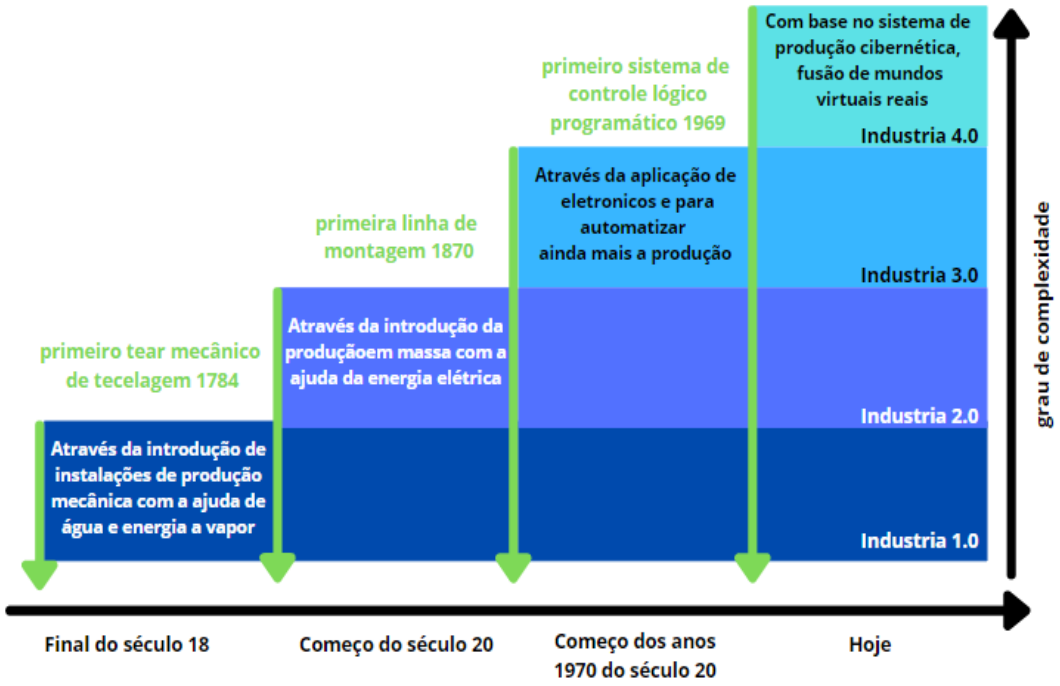
2.1 SEGMENTO INDUSTRIAL

Segundo Drath e Horch (2014), a indústria que conhecemos hoje é o resultado de três grandes revoluções. A Primeira Revolução Industrial ocorreu na Grã-Bretanha, durante o século XIII, a qual, como resultado, obteve um grande avanço na economia agraria, que começou a utilizar métodos de produção mecânicos. A Segunda grande revolução, se passou por volta do final do século XIX, com o desenvolvimento de motores a combustão, utilização do petróleo como combustível e o maior marco, a produção industrial em larga escada graças a energia elétrica, e pelas linhas de produção baseadas na separação de trabalho, o que permitiu a fabricação em serie para atingir o novo mercado sedento pelo consumismo. No final da década de 1960, a introdução da tecnologia da informação e microeletrônica no processo do desenvolvimento industrial, abriu novas portas para a produção automatizada e com o menor esforço e desperdício, e isso foi chamado de a Terceira Revolução Industrial (DRATH; HORCH, 2014).

Atualmente estamos vivenciando a Quarta Revolução Industrial, onde, o principal elemento de mudança, é a introdução de tecnologias na indústria. A utilização da internet como plataforma de transmissão de informações permite a comunicação de inúmeros dispositivos, unindo de forma simples a produção com outros setores da "Internet das Coisas" (CNI, 2016). Na indústria, agora as máquinas, sistemas e redes baseados em tecnologias da informação e comunicação, são capazes de trocarem informações, assim, melhorando cada vez mais para gerenciar os processos de produção industrial (GTAI, 2016).

A Quarta Revolução Industrial leva a automação dos processos de fabricação a um novo nível, introduzindo tecnologias de produção em massa (MARTIN, 2017). A **Erro! Fonte de referência não encontrada.** apresenta de forma ordenada as evoluções industriais, pela data e a principal modificação de paradigma.

Figura 8 - Grandes Revoluções Industriais



Fonte: Adaptado de Deloitte, 2015.

Essa nova revolução industrial, também é chamada de Indústria 4.0 e promete um aumento substancial da eficácia operacional, assim como a implantação de novos modelos de negócios, serviços, produtos, os quais, cada vez mais vão gerar informações em formato de dados, que podem ser aproveitados posteriormente (HERMANN; PENTEK; OTTO, 2016).

2.2 *BIG DATA*

Não há um consenso único sobre o conceito de *big data* aplicado no mercado. Entretanto, vários autores apresentam definições sobre perspectivas distintas aos quais, alguns apresentam similaridades em alguns pontos, permitindo um certo entendimento sobre a definição. Ter de fato essa definição bem desenvolvida é de extrema valia para o entendimento concreto deste trabalho.

Para Manyika (2011), o conceito de *big data* pode ser definido como “conjuntos de dados cujo tamanho é além da capacidade de ferramentas de software de banco de dados típicos para capturar, armazenar, gerenciar e analisar”. Segundo Nist (2015) o termo *big data* é um enorme conjunto de dados, em que a grande massa de dados é não estruturada e necessita de análise em tempo real.

Percebe-se, com as definições dos autores citados, pontos de convergências entre suas definições, tais quais, a necessidade em ter um grande volume de dados, a alta velocidade para o processamento e as inúmeras variedades de fontes de dados.

Já Taurion (2013), faz a união dessas informações semelhantes e apresenta uma fórmula conceituada sobre a definição de *big data*:

BigData = Volume + Variedade + Velocidade + Veracidade + Valor

O autor complementa a fórmula descrevendo:

> *Big Data* não é apenas um produto de software ou hardware, mas um conjunto de tecnologias, processos e práticas que permitem às empresas analisarem dados a que antes não tinham acesso e tomar decisões ou mesmo gerenciar atividades de forma muito mais eficiente. (TAURION, 2013, p.32).

Com base nas definições apresentadas é possível definir o conceito *de big data* como a implementação de um conjunto de práticas e tecnologias que tem como foco processar, analisar e armazenar uma grande variedade e volume de dados com uma grande velocidade, permitindo que as empresas utilizem de forma eficiente os resultados para tomada de decisões.

2.2.2. Atributos do *big data*

Segundo Taurion (2013), existem cinco atributos principais que formam o conceito de *big data*, são eles: volume, variedade, velocidade, veracidade e valor.+

2.2.2.1 Velocidade

O atributo velocidade está diretamente relacionado a um tempo de resposta rápido perante as informações que são consumidas e analisadas (RIFFAT, 2014).

> Velocidade é um critério que vai se tornar cada vez mais importante, devido à crescente rapidez com que as empresas precisam reagir às mudanças no cenário de negócios, bem como é necessária para tratar os dados em tempo real, interferindo na execução do próprio processo de negócios (TAURION, 2013, p.70).

Tendo como ponto de vista, que algumas empresas, possuem como prioridade tomadas de decisões com base em tendências e mudanças em algum cenário, o atributo velocidade acaba se tornando o principal, se comparado aos outros (MCAFEE; BRYNJOLFSSON, 2012).

2.2.2.2 Volume

Para Riffat (2014) o atributo volume é a quantidade de dados a qual é consumida e processada. Perante o *big data*, o volume será todas as informações necessárias que deverão ser manipuladas, sendo elas, modificações, cortes ou até mesmo novos dados gravados.

2.2.2.3 Valor

Esse atributo é um que deve ser muito bem analisado pela empresa que deseja começar a usar o *big data*, afinal, "*Big Data* só faz sentido se o valor da análise dos dados compensar o custo de sua coleta, armazenamento e processamento" (TAURION, 2013). O retorno em forma de dados, deve ser maior que o custo que tem que ser investido para realizar a análise, porque, caso contrário, será um valor totalmente desperdiçado.

2.2.2.4 Veracidade

Uma informação que não é autêntica não possui valor algum, e nada poderá ser feito com ela (TAURION, 2013). Esse atributo pode ser destacado dos outros, pois, caso não haja veracidade dos dados, nenhuma análise terá credibilidade.

2.2.2.5 Variedade

Com o crescimento do universo digital, praticamente tudo que fazermos conectado à internet é gravado em inúmeros bancos de dados, com dados nos formatos mais diferentes possíveis. Fazendo com que seja criado tecnologias capazes de armazenar e processar diversas fontes diferentes de dados (MCAFEE; BRYNJOLFSSON, 2012).

Taurion (2013), explica a importância de conectar dados de locais diferentes e o quanto isso pode ter valor para a análise das informações:

> A variedade é um parâmetro importante, pois, com diversas fontes de dados aparentemente sem relações, podemos derivar informações extremamente importantes e fazer análises preditivas mais eficientes. Por exemplo, conectando dados meteorológicos com padrões de compra dos clientes podemos planejar que tipo de produtos deverão estar em destaque nas lojas quando for detectado que haverá um período de alguns dias de temperatura elevada, daqui a três dias. Ou conectar dados geográficos com detecção de fraudes (TAURION, 2013, p.32).

2.2.3 Tecnologias do *big data*

A quantidade de dados gerados diariamente por pessoas e empresas criou a necessidade da existência do *big data*. Os dados, com informações, acabaram se tornando humanamente impossíveis de serem analisados, tratados e reestruturados de forma manual. (MICHAEL.; MILLER, 2013). Existem três camadas que sustentam o Big Data: (1) as tecnologias de infraestrutura que coletam e gravam os dados; (2) tecnologias de processamento, como *Hadoop* e *MapReduce*; e, (3) as tecnologias de análises, que pertencem ao princípio do *Big Data Analytics* (TAURION,2013). Nesta seção, serão abordadas algumas das tecnologias mais utilizadas para soluções big data, focando nas camadas de processamento e analises, além de ferramentas facilitadores para a elaboração deste trabalho.

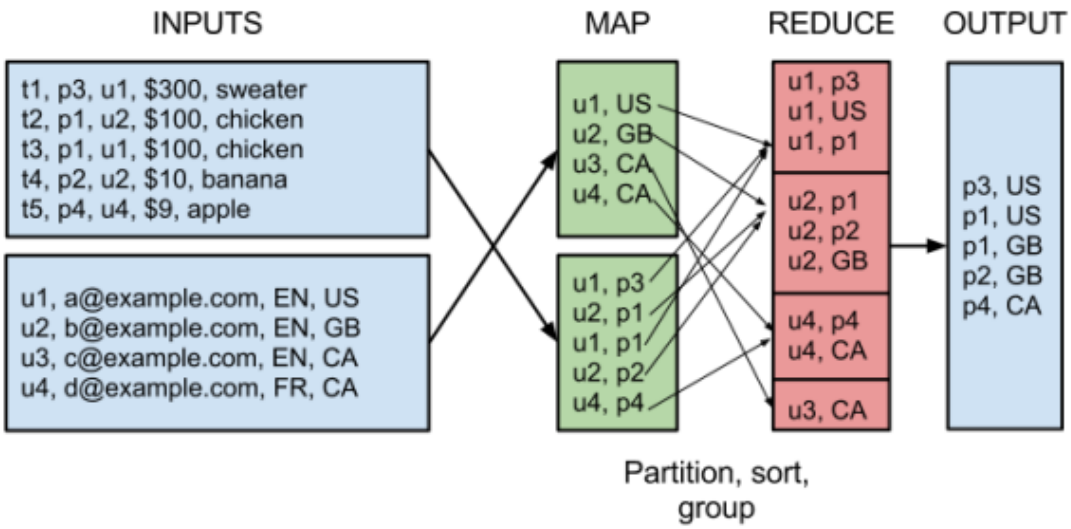
2.2.3.1 MapReduce

Devido ao grande volume de dados em ambientes distribuídos, surgiu o *MapReduce*, uma solução para aproveitar o aumento do poder de processamento de sistemas que possuem ligação (CONDIE et al., 2010).

A utilização de *MapReduce*, permite a implementação em sistemas distribuídos, processar grandes quantidades de dados, automatizando o escalonamento, particionamento de disco, além das tolerâncias a falhas, tornando o foco do trabalho apenas a tratativa de dados (DEAN; GHEMAWAT, 2010).

Segundo Condie et al (2010), para implementação do *MapReduce*, é necessário a implementação de funções pré-definidas. A primeira função, é o *map*, que é executada a cada entrada de dupla de chave, diminuindo uma lista de pares intermediários. A segunda função é a *reduce*, onde, é chamada uma vez para cada chave distinta da lista criada na etapa anterior, e retorna o resultado do processamento. A **Erro! Fonte de referência não encontrada.** ilustra o fluxo descrito nesse processamento de um *MapReduce*.

Figura 9 - Exemplo MapReduce



Fonte: Rathbone (2013)

A solução mais difundida no mercado atual, referente a *MapReduce*, é o Apache Hadoop, o qual, possui um sistema chamado *Hadoop Distributed File System (HDFS)*. Essa função permite receber um grande volume de dados e possibilita a execução dos processos do *MapReduce* para sistemas distribuídos. O mesmo armazena apenas os dados de entrada da função *map* e a saída da função *reduce*, deixando os dados intermediários em cada nodo do sistema (CONDIE et al., 2010).

2.2.3.2 NoSQL

O termo *NoSQL*, segundo Sadalage e Fowler (2013), foi usado pela primeira vez por Carlo Strozzi (1998) durante um debate, onde ele se referiu a um banco de dados relacional, chamada Strozzi *NoSQL*, que não utilizava a linguagem SQL, e com o passar do tempo, a palavra virou referência a essa nova categoria de banco de dados.

De acordo com Mpinda, Bungama e Maschietto (2015), a demanda pelo uso de sistemas *NoSQL*, acontece principalmente quando alguma aplicação necessita manter serviços com um alto desempenho, perante a um grande volume de dados, explorando o campo de conhecimento do *big data*.

Para Machado (2018), em bancos de dados *NoSQL*, as tabelas são denominadas de tabelas *hash* distribuídas, no qual, armazenam objetos indexados

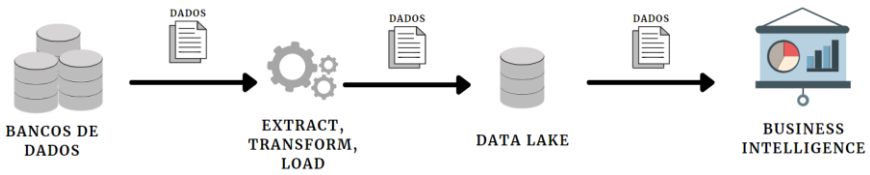
por chaves, onde possibilita o encontro dos objetos utilizando apenas esta chave. Os bancos NoSQL são projetados para terem um aumento de escala em sentido horizontal, por meio de clusters distribuídos em hardwares de baixo custo, proporcionando uma escalabilidade de forma simples.

2.3 ESTRUTURAÇÃO DE DADOS

As tomadas de decisões baseadas em dados, não exclui as habilidades analíticas humanas, mas complementa em seus pontos fracos, tornando a assertividade maior (CRUZ,2007). Porém é necessário que os dados sejam verídicos, sem falhas e tratados, para que não hajam análises baseadas em dados falsos.

Nesta seção, serão abordadas algumas ferramentas e estruturas que contribuem para a implementação de estruturação e controle de dados de forma organizada.

Figura 10 - Exemplo de Estruturação de Dados



Fonte: Autor (2021)

Na figura 10, é possível verificar a ilustração das etapas de uma estruturação de dados, onde é utilizado um processo de *extract, transform, load*, e armazena os dados tratados em um *data lake* para e então alimentar um *business intelligence*, que será a ferramenta final para o usuário.

2.3.1 *ETL*

ETL, do inglês, *Extract, Transform, Load*, é um conjunto de processos para realizar a etapa de trazer dados de um sistema para uma base de dados, não restritos de sistemas, mais também de websites, bases de e-mails, redes sociais, arquivos de texto e até bases de dados pessoais (TANAKA, 2015).

O grau de dificuldade depende diretamente de como será o cenário a enfrentar nos sistemas de origem, que podem estar armazenados em esquemas

comuns (homogêneos) ou em estruturas diferentes (heterogêneas); em um banco de dados comum ou com os dados espalhados por diferentes bancos (NETO, 2012). Sendo pilar para a estruturação dos dados de forma correta, o qual é necessário para ter uma implementação de uma arquitetura *big data* confiável.

Taurion (2013) destaca “manipulação de dados distribuídos em clusters de servidores usados de forma massivamente paralela”, como uma das vantagens na utilização de transformações eficientes.

2.3.2 Data Lake

O termo *Data Lake* foi criado pelo CTO do Pentaho, James Dixon, e pode ser considerado uma área que guarda e processa dados com formatos diferentes, de forma performática, e se trata de um conceito, e não uma tecnologia. Matos (2018), afirma que o verdadeiro valor do *data lake*, vem da capacidade e das habilidades de ciência de dados de quem está o utilizando, dando ênfase as competências dos envolvidos.

Ribeiro (2018), explica que o dado armazenado em um *data lake* está pronto para o uso final, entretanto, essa informação pode estar equivocada, porém, se o dado está em seu formato bruto, não faz sentido usá-lo para fazer uma análise antes de fazer as devidas tratativas.

2.3.3 Business Intelligence

Para Côrtes (2002 apud Fontana, 2009), *Business Intelligence (BI)* é um conjunto de conceitos e metodologias que contribuem com a tomada de decisões na medida em que transforma o dado em informação. Utilizando as ferramentas que o *BI* disponibiliza, é possível obter informações através de vínculos de dados, que podem vir de um *Data Lake* ou *Data Warehouse*, além da demonstração através de indicadores (BATISTA, 2004 apud REGINATO, 2007).

Barbieri (2001, p. 34) cita que o *BI* é a utilização de várias fontes de informação de forma a auxiliar na definição de estratégias de negócio no mercado competitivo.

Segundo o mesmo autor, Barbieri (2001, p. 34) “os Sistemas legados e os emergentes Enterprise Resource Planning (ERP), sistemas integrados corporativos, não trazem as informações gerencias na sua forma mais palatável”,

ou seja, as empresas podem ter os dados, mais não necessariamente sabem o que esses dados querem dizer.

As citações e afirmações constituem o presente trabalho, assim, a partir da finalização dos conceitos das ferramentas e tecnologias que o baseiam, na próxima seção é apresentado a metodologia de pesquisa para a realização do mesmo.

3 MÉTODO DE PESQUISA

A definição da palavra método, segundo Garcia (1998, p. 44 apud HEERDT, 2007):

> Representa um procedimento racional e ordenado (forma de pensar), constituído por instrumentos básicos, que implica utilizar a reflexão e a experimentação, para proceder ao longo do caminho (significado etimológico de método) e alcançar os objetivos preestabelecidos no planejamento da pesquisa (Garcia, 1998, p. 44 apud HEERDT, 2007).

Desta forma, por ser uma pesquisa de natureza técnica-científica descreve-se neste capítulo os métodos utilizados para a elaboração do projeto.

3.1 CARACTERIZAÇÃO DO TIPO DE PESQUISA

Para caracterização do presente trabalho é necessário definir o conceito de pesquisa. Para Gil (2008), pesquisa é um "processo formal e sistemático de desenvolvimento do método científico". Já para Silva e Menezes (2005), pesquisa "é um conjunto de ações, propostas para encontrar a solução para um problema, que tem por base procedimentos racionais e sistemáticos".

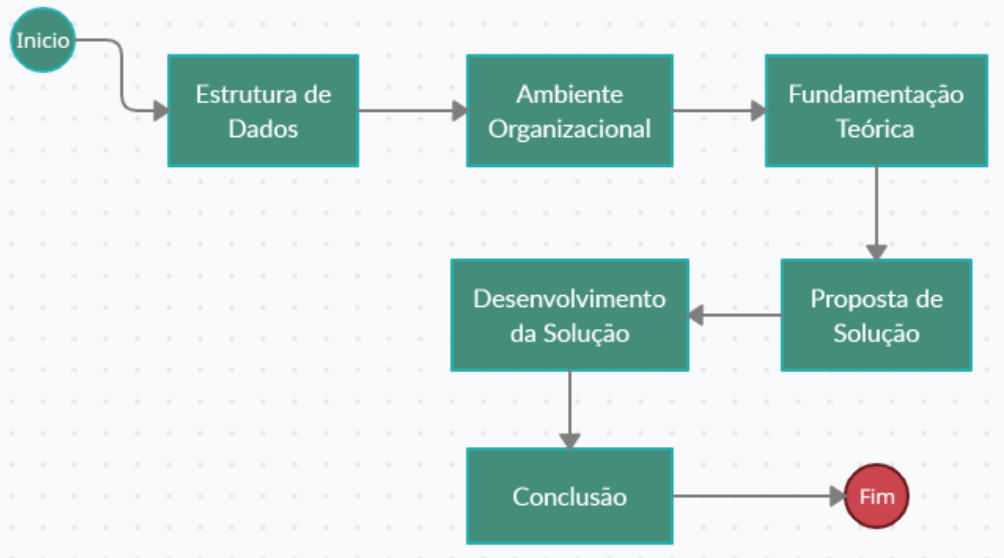
O trabalho possui a abordagem de pesquisa básica e aplicada com natureza prática, porque, "objetiva gerar conhecimentos para aplicação prática e dirigidos à solução de problemas específicos" (SILVA; MENEZES, 2005). A metodologia de pesquisa adotada neste trabalho é pesquisa aplicada, pois existe a realização de uma pesquisa bibliográfica sobre as características arquiteturais de *big data*, demonstrando os benefícios da implementação, seus desafios e a implementação em uma empresa de segmento industrial. Além dessa definição, como a pesquisa é qualitativa, não há uma apresentação de resultados em demonstrados por números, e sim, por meio de modelos de dados e protótipos (MEZZAROBA; MONTEIRO, 2004).

A pesquisa bibliográfica é a proposta de pesquisa mais adequada para ser implementado neste trabalho. Segundo Gil, a pesquisa bibliográfica pode ser definida pelo desenvolvimento de uma linha de raciocínio baseada em materiais já criados, principalmente de livros e artigos científicos (GIL, 2008).

3.2 ATIVIDADES METODOLÓGICAS

As etapas metodológicas do presente trabalho são as listadas na figura 11:

Figura 11 - Fluxograma de processos



Fonte: Autor (2021)

1) Identificação do problema: Etapa de levantamento do funcionamento dos processos de uma empresa com segmento industrial.
2) Ambiente Organizacional: Etapa na qual será possível entender melhor as funcionalidades atuais da empresa e como os usuários utilizam as informações.
3) Fundamentação Teórica: Etapa onde é apresentado os fundamentos do problema, o qual, servirá de embasamento para consultas e reforçar o conhecimento teórico de *big data* e processamento de dados.
4) Proposta de Solução: Etapa onde será apresentado as soluções com a utilização de arquiteturas *big data* para responder à pergunta do projeto.
5) Desenvolvimento da Solução: Etapa de conferência das propostas de solução, e implementação definitiva do projeto que atende as necessidades atuais da empresa referente a dados.
6) Conclusão: Etapa final, onde será apresentado os resultados gerados por esta monografia, com base na implementação de arquiteturas de *big data* em uma empresa com segmento industrial, dando oportunidade para trabalhos futuros utilizarem como base nos objetivos estudados do trabalho.

As etapas citadas, levam a conclusão da exploração do campo de conhecimento da arquitetura de *big data* e demonstram como essas informações podem ser

aplicadas no segmento industrial demonstrando os benefícios da aplicação desta tecnologia.

3.3 DELIMITAÇÕES

A solução apresentada por esta pesquisa está restrita somente ao desenvolvimento de uma solução com arquitetura *big data* que atendam às necessidades e limitações da empresa do segmento industrial em questão.

4 PROPOSTA DE SOLUÇÃO

Neste capítulo são apresentados detalhes da arquitetura proposta. Desta forma, contextualiza-se o cenário da empresa e demonstra-se como os dados eram utilizados no ambiente.

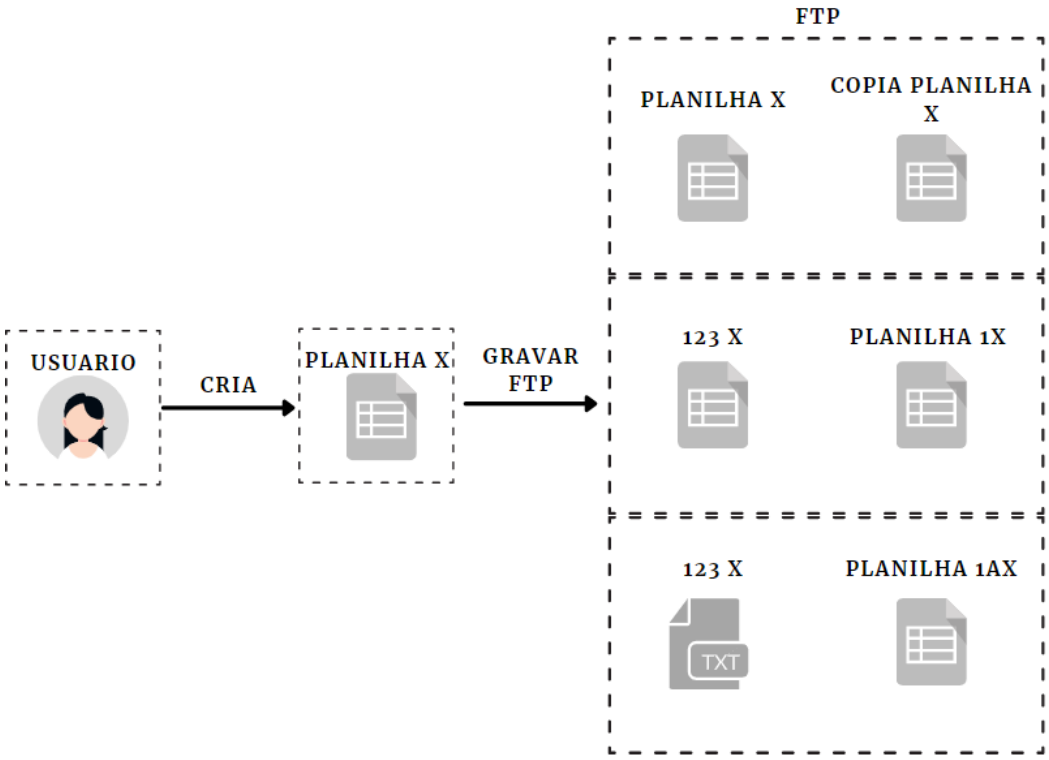
4.1 AMBIENTE ORGANIZACIONAL

A empresa, fundada em 1976 e objeto deste estudo, atua no setor industrial, a qual se destaca no segmento de tecnologia e segurança. O seu portfólio de produtos é variado e contempla desde câmeras de monitoramento e segurança à equipamentos para conexões em rede. Atualmente, a empresa possui mais de três mil funcionários separados entre fábrica e administrativo.

Para controle do trabalho a empresa utilizava dados de planilhas, os quais, eram totalmente descentralizados e cada usuário possuía um controle próprio das informações. Por muitas vezes, mantinham dados errados ou desatualizados.

Essa estrutura de dados em planilhas ficava disponível em um diretório compartilhada via o protocolo FTP (*File Transfer Protocol*). Na maioria das vezes não havia controle de acesso dos usuários ao conteúdo de modo que algumas informações trafegavam livremente. Além disso, neste diretório haviam arquivos duplicados, desatualizados, ou até mesmo totalmente fora do contexto da pasta. Processo ilustrado na figura 12:

Figura 12 - Exemplo falta de organização com as Planilhas

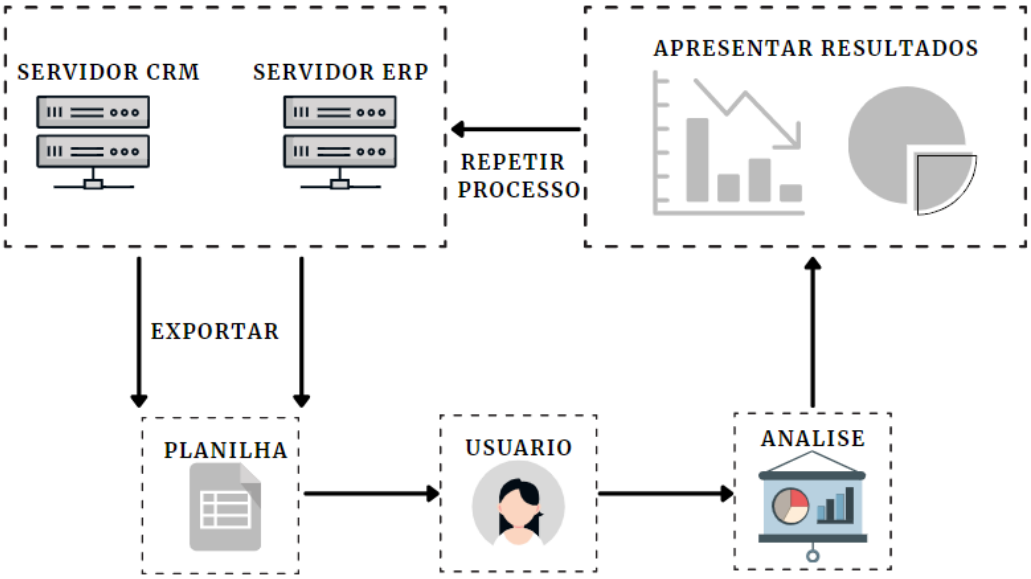


Fonte: Autor (2021)

Outro problema recorrente era a perda de informações por exclusão de alguma planilha. A empresa não possuía qualquer sistema de backup ou plano de ação de recuperação de dados.

Além das planilhas criadas e alimentadas pelos usuários, era possível exportar informações do *ERP* ou *CRM (Customer relationship management*), para ter acesso às informações alimentadas nessas bases. Essas informações exportadas eram utilizadas para fazer análises manuais, tornando o processo uma rotina repetitiva (figura 13): de exportação o dado, implementação da análise, conferência do processo e por fim utilização do resultado.

Figura 13 - Processo repetitivo de análise manual



Fonte: Autor (2021)

Com esses processos manuais e repetitivos sempre há perda de produtividade e desempenho, visto que, a repetição de tarefas leva o usuário a rotina e a comodidade, deixando de investir tempo em algo mais produtivo para crescimento profissional e maior desempenho para a empresa.

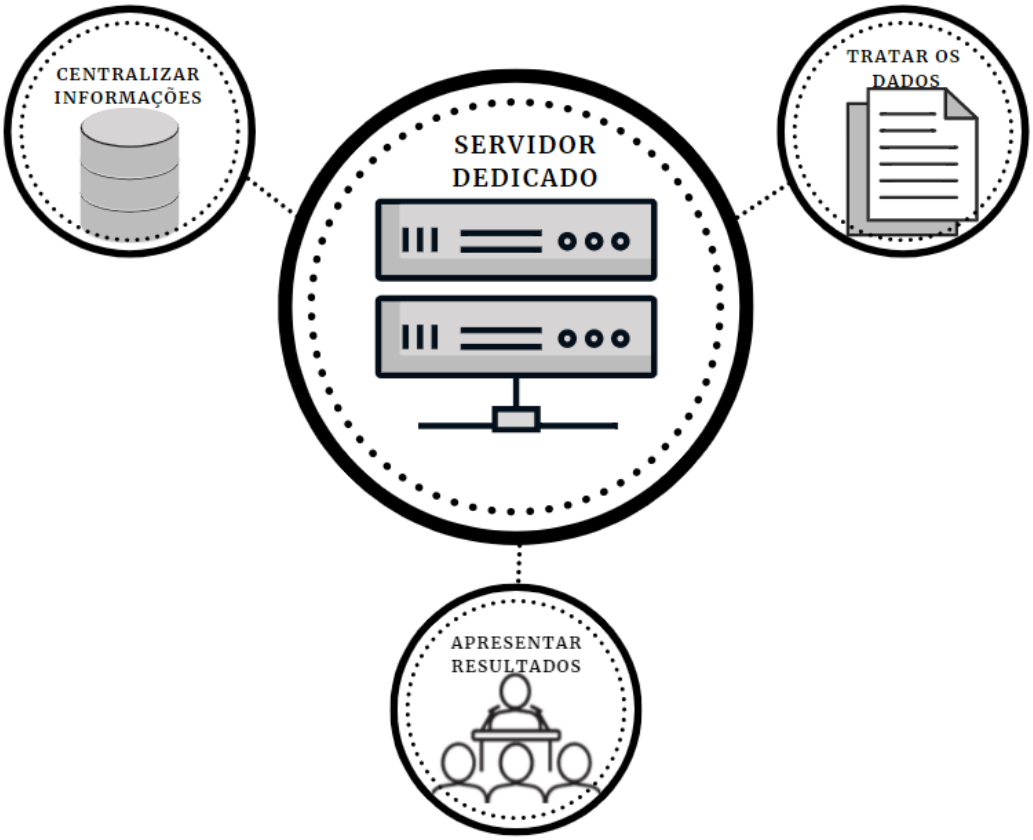
4.2 MODELO DE ESTRUTURA DE DADOS

Após analisar as condições que os dados eram submetidos, foi realizada uma série de pesquisas para identificar a melhor forma de conduzir um projeto de estruturação dos dados da empresa. Como a parte fabril da empresa está implementando sistemas de controle de produção e acompanhamento por item produzido, há a expectativa de que futuramente haverá uma grande quantidade de informações. Pensando nisso, apresentou-se a proposta de implementar uma arquitetura de dados que, futuramente poderá ser escalável.

Tendo essas informações como base o primeiro ponto necessário seria um servidor dedicado com poder de processamento elevado e com memória escalável para não haver problemas futuros.

Nesse servidor, existem três processos principais: (1) centralizar as informações; (2) tratar os dados; e, (3) apresentar resultados. Cada um dos processos é de extrema importância, conforme ilustrado na figura 14. Somente com a união entre eles, é possível melhorar a estrutura de dados.

Figura 14 - Principais processos iniciais



Fonte: Autor (2021)

A etapa de centralização das informações consiste na estruturação de um banco (*data lake*), que receberá os dados de diversas fontes de dados diferentes, centralizando os dados em um único lugar, algumas informações brutas e outras já estruturadas prontas para utilização da etapa final.

A etapa de tratativa de dados é formada pela criação de projetos em R, Python e Pentaho Data Integration para organizar, estruturar e ajustar todos os dados necessários.

A etapa final, apresentação de dados, é a etapa onde será utilizado os dados já estruturados, e armazenados no *data lake* para alimentar e estruturar análises com a ferramenta de *BI*, para então os usuários finais tomarem decisões com base nas informações.

4.2.1 Servidor Dedicado

O pilar para o início da estruturação é possuir um ambiente adequado. No qual deverá comportar todas as ferramentas e processos para contemplar a centralização dos dados e tratativas necessárias. Visto que, é nítido que dados são extremamente voláteis, variando muito em formas, tipos e quantidade. Por isso, é necessário a utilização de um servidor escalável para ser maleável conforme as demandas e necessidades futuras não previstas.

O container em um servidor *Docker*[2] atende as necessidades, porque além de ser escalável (variando conforme a distribuição), haverá uma estrutura eficiente para realizar backups e simples para recuperar o container, em caso de algum problema, perca de informação ou histórico.

4.2.2 Centralizar os Dados

Com o objetivo de possuir a governança de dados de forma centralizada e ordenada é necessária alguma forma de armazenar essas informações. De forma resumida, temos duas opções: um *data lake* ou um data Warehouse.

Como já há uma noção base de que tipos de dados e estruturas de dados que a empresa possui, a melhor escolha possível seria a utilização de um banco de dados não relacional, onde pode-se gravar dados brutos, dados tratados e históricos de dados criados que não possuem registros armazenados.

Conforme a descrição, um *Data Lake* MongoDB[3] atenderia as necessidades, entrando em concordância com a implementação futura de uma arquitetura *big data*, que poderia utilizar os dados gravados em tempo real nessa estrutura, para mapear e monitorar as informações.

4.2.3 Tratar os Dados

Com um local para centralizar os dados pode-se ter autonomia para iniciar o processo de tratamento de informações. Essa será uma etapa constante, dependendo da necessidade de novas implementações que necessitem de novas informações.

---

[2] Segundo a página oficial do *Docker* (2021), é uma plataforma aberta para desenvolvimento, envio e execução de aplicativos, permitindo a separação dos mesmos na infraestrutura para que seja possível entregar o software de forma rápida, gerenciando a infraestrutura da mesma forma gerencial dos aplicativos.

[3] Segundo a página oficial do MonngoDB(2021), é considerado um banco de dados distribuído com proposito amplo, baseado em documentos, podendo ser na nuvem.

Inicialmente, o ideal seria a utilização de linguagens de programação, as quais possuem amplas variedades de bibliotecas para realização de processos de ETL. Além disso, deverá ser levado em conta, as comunidades dessas linguagens, quais são ativas e poderão ajudar em possíveis dúvidas posteriores.

Levando em consideração os aspectos comentados acima, foi identificado duas linguagens que serão validas, R e Python. Segundo o site Oficial da linguagem (R 2021), R é um ambiente de software livre para computação estatística e demonstrativos através de gráficos. Também uma ampla variedade de técnicas estatísticas (modelagem linear e não linear, testes estatísticos clássicos, análise de série temporal, classificação, agrupamento).

A linguagem de programação R é voltada a cálculos estatísticos e complexos, com bibliotecas robustas de tratativa de dados, que crescem diariamente, além da visualização de dados direto de *datasets* através de um esquema relacional apresentado pelo Rstudio. Um dos pontos fortes de R é a facilidade com que gráficos de qualidade de publicação bem projetados podem ser produzidos, incluindo símbolos matemáticos e fórmulas quando necessário (R 2021).

Em sua documentação, fornecida no site oficial (R 2021), é possível apontar as seguintes características:

- Manuseio eficaz de dados e instalação de armazenamento;
- Conjunto de operadores para cálculos em matrizes, em matrizes particulares;
- Recursos gráficos para análise e exibição de dados na tela;
- Linguagem de programação bem desenvolvida, simples e eficaz que inclui condicionais, loops, funções recursivas definidas pelo usuário e recursos de entrada e saída.

Segundo o site oficial da Linguagem Python é uma linguagem de programação que permite trabalhar mais rapidamente e integrar seus sistemas de forma mais eficaz (PYTHON, 2021). Na documentação (PYTHON, 2021) é possível observar as seguintes características:

- linguagem fortemente *tipada;*
- alto nível, funcional, interpretada, orientada a objetos, imperativa;
- simples na escrita e compreensão;

- versatilidade no uso de scripts e ou programas sem interface gráfica;
- documentação completa e comunidade ativa;
- pode alcançar alto grau de desempenho e estabilidade.

A linguagem Python é muito utilizada na área de ciência de dados, tendo um leque de bibliotecas focadas nas necessidades do projeto.

4.2.4 Apresentar Resultados

Após possuir os dados tratados pelas linguagens de programação e ferramentas mencionados, e armazenados, será necessário apresentar os resultados. Alguns dados já estarão calculados e polidos pelas ferramentas de ETL e não terão a necessidade de sofrerem algum tipo de ajustes na demonstração, porém, outros deverão ser vinculados antes apresentar ao usuário final.

Para atender a qualidade esperada na demonstração de resultados, a utilização de um sistema de *BI* seria ideal. A empresa já possui licenças do *BI* Qlik Sense, o qual é utilizado por algumas áreas que demonstram dados simples, sem cálculos e tratativas, apenas apresentação de informação de forma mais clara para o usuário final, se comparado a um SQL ao banco de dados.

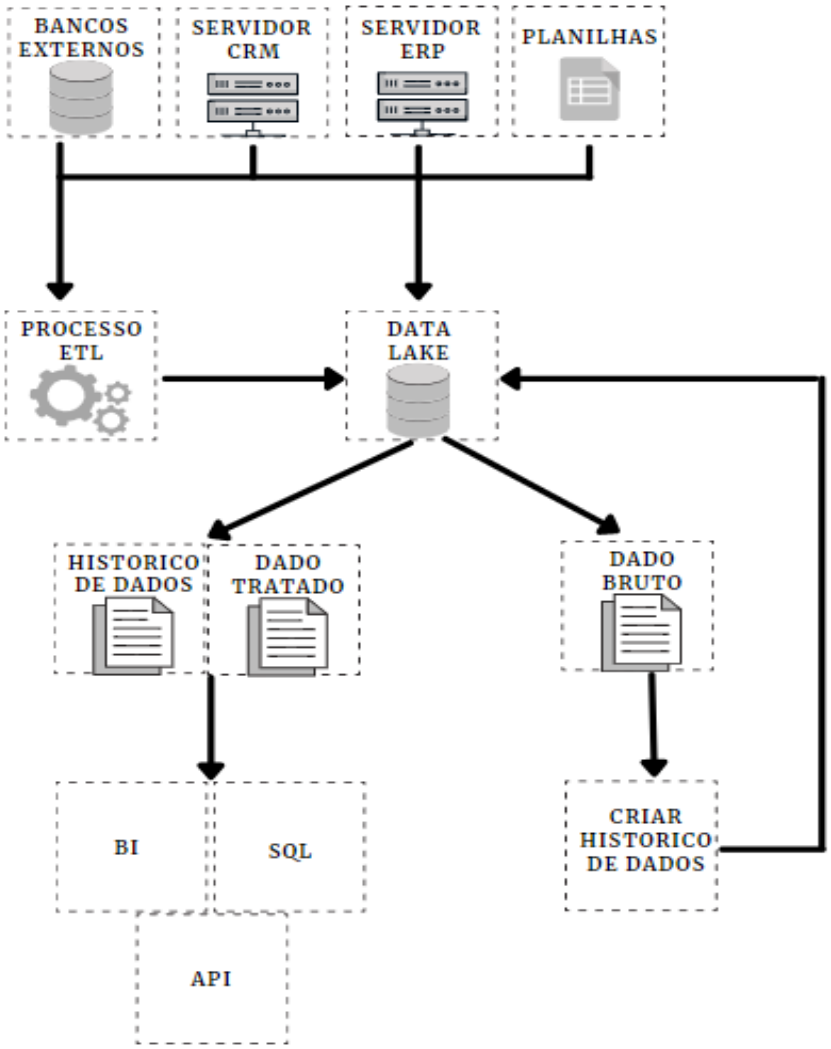
Um painel de *BI* além de apresentar os dados de forma clara e objetiva, em formato de gráficos, vínculos e análises personalizados, possui um sistema facilitador, que não demandam de muito desenvolvimento, poupando tempo na etapa final.

Dado o exporto acima, para atender ao objetivo de demonstrar os benefícios da implementação da arquitetura big data, bem como seus desafios no cenário de uma empresa do segmento e o objetivo de demonstrar, por meio de uma prova de conceito, a aplicação do modelo da arquitetura big data, será implementado no próximo capitulo a solução para os desafios levantados.

5 DESENVOLVIMENTO DA SOLUÇÃO

Com o levantamento das necessidades apresentadas no capítulo anterior, com base nos atuais problemas de má distribuição e gestão de informações, foi desenvolvido uma estrutura para atender todo o fluxo dentro de um *Docker* (figura 15).

Figura 15 - Projeto da estruturação de dados



Fonte: Autor (2021)

Nessa estrutura, todas as informações que veem de inúmeras formas diferentes são centralizadas em um *data lake*, o qual recebe dados tratados, não tratados e ainda serve como armazenamento de histórico de processos, alimentando o *BI*, podendo ser consultado via consultas no banco e através de API.

5.1 *DATA LAKE*

O objetivo inicial era utilizar como *data lake* um banco de dados não relacional MongoDB, entretanto, esse banco estava em processo de homologação pela empresa. Para acelerar o processo de implementação, optou-se por um banco de dados já era homologado, PostgreSQL 11. No momento não será um problema usar uma base relacional, porque as integrações serão com arquivos e outros bancos também relacionais ou API. Assim que o banco MongoDB for homologado, será possível migrar as informações já criadas no banco relacional.

Para manter o controle e a organização das tabelas no *data lake*, foi criado um padrão para gravação de nome das tabelas de cada projeto. Esse padrão consiste em uma abreviação de três caracteres do nome do projeto, seguido de um ponto final e o nome da tabela, possibilitando associar de qual projeto é cada tabela:

Projeto Proof of Concept = "POC" + "." + "Nome_Tabela"

Mantendo esse padrão, foram gravados no *Data Lake* 12 tabelas de projetos, os quais estão apresentados no quadro 1:

Quadro 1 - Lista de projetos

Abreviação	Projeto
ADP	Acompanhamento de Projetos
AML	Análise do Mercado Livre
AUX	Auxiliar
CRM	Customer relationship management
DDB	Diário de Bordo
PLC	Política Comercial
POC	Proof of Concept
SCH	Schedule
SOT	Sellout
SPO	Solution per Objetive
TOT	Enterprise Resource Planning
WEP	Web of Price

Fonte: Autor (2021)

Os dados gravados no *Data Lake*, são constituídos de dados brutos tratados, ordenados e históricos. Todas as informações gravadas são oriundas das soluções do PDI, Python ou R.

Quando o PDI é utilizado, os dados com erros de digitação são tratados, formatados no padrão utf8, une-se as informações e é realizado os cálculos simples. Como essa plataforma apresenta um alto desempenho, atualmente ela é a mais utilizada para o processo de ETL.

Quando o projeto envolve múltiplos cálculos, ou é complexo, é utilizado Python ou R para desenvolver as soluções, isso porque, utilizando as linguagens de programação, é possível ter maior autonomia para explorar inúmeras formas de desenvolver algo que atenda todas as necessidades. Visto que, em alguns momentos, é necessário realizar extrações de bancos diferentes, armazená-los em variáveis e desenvolver uniões de informações e cálculos que se interligam, para formar um resultado que atenda o esperado, e por fim gravar as informações prontas para uso final no *Data Lake*. A figura 16 ilustra uma amostra das tabelas gravadas no banco:

Figura 16 - Amostra das tabelas no *Data Lake*

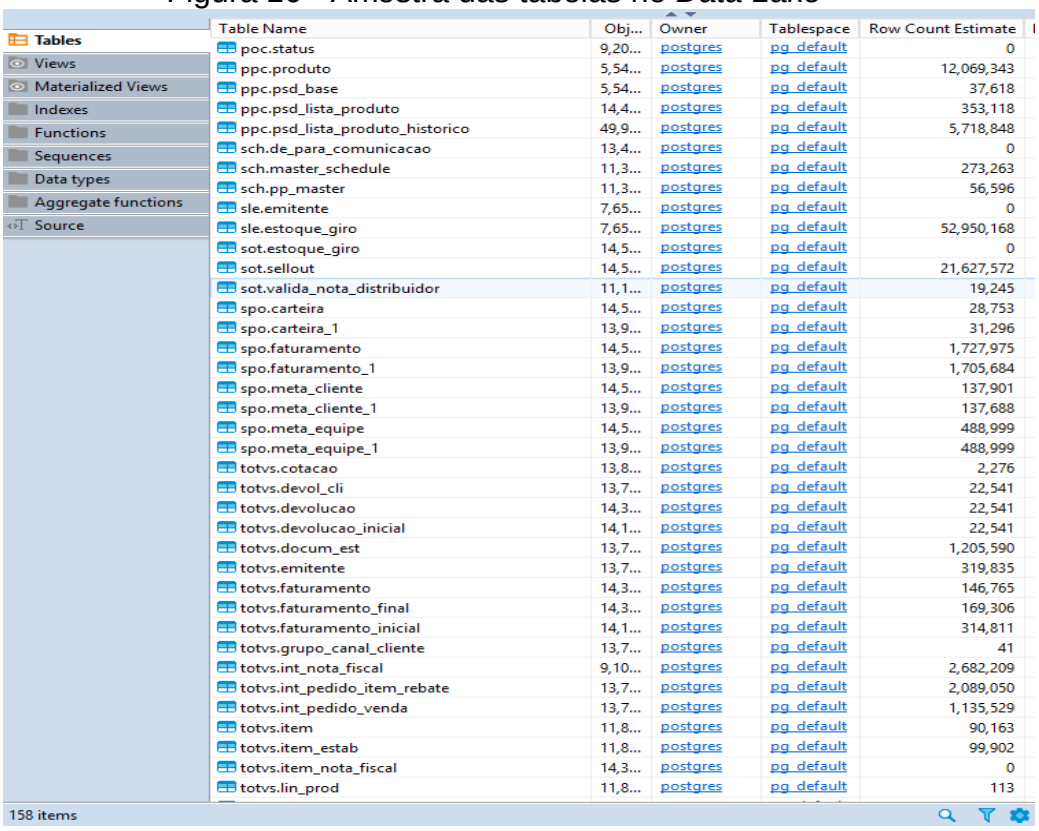
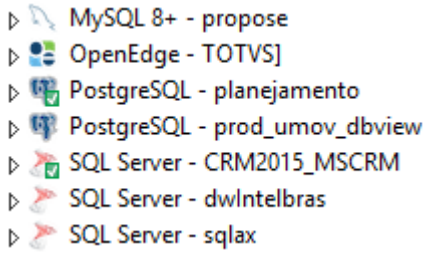


Table Name	Obj...	Owner	Tablespace	Row Count Estimate
poc.status	9,20...	postgres	pg_default	0
ppc.produto	5,54...	postgres	pg_default	12,069,343
ppc.psd_base	5,54...	postgres	pg_default	37,618
ppc.psd_lista_produto	14,4...	postgres	pg_default	353,118
ppc.psd_lista_produto_historico	49,9...	postgres	pg_default	5,718,848
sch.de_para_comunicacao	13,4...	postgres	pg_default	0
sch.master_schedule	11,3...	postgres	pg_default	273,263
sch.pp_master	11,3...	postgres	pg_default	56,596
sle.emitente	7,65...	postgres	pg_default	0
sle.estoque_giro	7,65...	postgres	pg_default	52,950,168
sot.estoque_giro	14,5...	postgres	pg_default	0
sot.sellout	14,5...	postgres	pg_default	21,627,572
sot.valida_nota_distribuidor	11,1...	postgres	pg_default	19,245
spo.carteira	14,5...	postgres	pg_default	28,753
spo.carteira_1	13,9...	postgres	pg_default	31,296
spo.faturamento	14,5...	postgres	pg_default	1,727,975
spo.faturamento_1	13,9...	postgres	pg_default	1,705,684
spo.meta_cliente	14,5...	postgres	pg_default	137,901
spo.meta_cliente_1	13,9...	postgres	pg_default	137,688
spo.meta_equipe	14,5...	postgres	pg_default	488,999
spo.meta_equipe_1	13,9...	postgres	pg_default	488,999
totvs.cotacao	13,8...	postgres	pg_default	2,276
totvs.devol_cli	13,7...	postgres	pg_default	22,541
totvs.devolucao	14,3...	postgres	pg_default	22,541
totvs.devolucao_inicial	14,1...	postgres	pg_default	22,541
totvs.docum_est	13,7...	postgres	pg_default	1,205,590
totvs.emitente	13,7...	postgres	pg_default	319,835
totvs.faturamento	14,3...	postgres	pg_default	146,765
totvs.faturamento_final	14,3...	postgres	pg_default	169,306
totvs.faturamento_inicial	14,1...	postgres	pg_default	314,811
totvs.grupo_canal_cliente	13,7...	postgres	pg_default	41
totvs.int_nota_fiscal	9,10...	postgres	pg_default	2,682,209
totvs.int_pedido_item_rebate	13,7...	postgres	pg_default	2,089,050
totvs.int_pedido_venda	13,7...	postgres	pg_default	1,135,529
totvs.item	11,8...	postgres	pg_default	90,163
totvs.item_estab	11,8...	postgres	pg_default	99,902
totvs.item_nota_fiscal	14,3...	postgres	pg_default	0
totvs.lin_prod	11,8...	postgres	pg_default	113

Fonte: Autor (2021)

Os dados brutos, são extraídos de sete fontes diferentes de banco de dados, duas API, e algumas planilhas. Dentre os bancos, estão listados OpenEdge, MySQL, PostgreSQL e SQL Server, conforme a figura 17:

Figura 17 - Listagem dos Bancos



Fonte: Autor (2021)

Todas essas informações centralizadas e tratadas, servem de repositório, que, serão utilizados para o desenvolvimento de painéis informativos e analíticos, que serão de extrema valia para tomada de decisões de gerentes e diretores.

5.2 EXTRACT, TRANSFORM, LOAD - ETL

Para garantir a integridade das informações, é necessário realizar uma etapa antes da gravação de informações em um banco de dados. Essa etapa, chamamos de processo ETL, o qual terá como foco, extrair as informações dos locais necessários, transforma-las, calculara-las e modifica-las para ter a garantia do que o dado, poderá ser usado como ferramenta informativa, e por fim, gravar em um repositório, o qual denominamos de *data lake*, o qual poderá ser consultado pelos analistas para ser utilizados em analises de resultados em um *business intelligence*.

Com o levantamento prévio da utilização de linguagens de programação para realização do processo ETL, foram definidas as linguagens R e Python. Entretanto, além dessas, a ferrramenta o *Pentaho Data Integration* (PDI) também é extremamente eficiente para realizar o processo de ETL em tabelas com bilhares de linhas, além de ter como foco o desempenho, é uma ferramenta que possui uma versão gratuita que atende todos as necessidades atuais.

5.2.1 Pentaho Data Integration

Pentaho Data Integration versão *comunity*, é uma ferramenta de ETL, de código aberto, distribuída pela *Kettle*. Contribuindo para que as empresas adotem uma gestão da informação voltada a dados e criação de estratégias competitivas (VARGAS, 2008, p. 7).

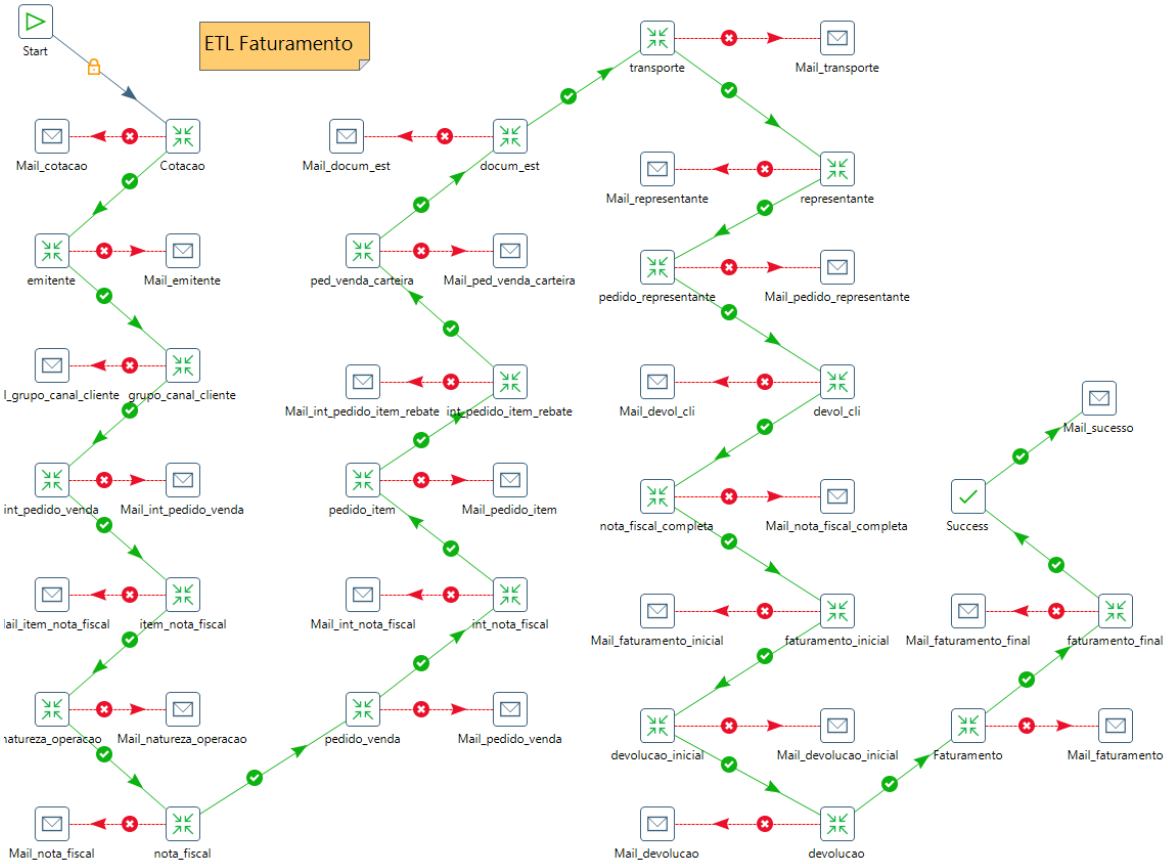
O PDI funciona com fluxos de etapas em blocos, os quais executam um ou várias etapas, que podem ser*:*

- *Job*: tem um início e fim pré-definido, com um fluxo de etapas de *transformations*, podendo ter dois tipos de saídas de cada etapa, verdadeiro, para caso o processo de certo, continuando normalmente, ou false, para caso o processo de algum problema, podendo parar e enviar uma notificação de falha.
- *Transformations:* podem possuir várias entradas de dados de fontes diferentes carregando de forma paralela, além de ter vários *streps* que contribuem para facilitar os processos de ETL. Caso precise rodar várias transformações, é necessário chamá-las por um *job*.

Como o PDI é uma ferramenta com alto desempenho para processamento de dados, ela é utilizada para a maioria dos projetos para fazer o processo ETL, os quais é necessário carregar muitos dados, tratar erros, fazer uniões de campos ou tabelas diferentes, cálculos e consultas, economizando tempo, pela simplicidade da ferramenta e pelo processamento de dados de forma otimizada.

Um exemplo de projeto, é o projeto referente ao faturamento, foi inteiramente tratado utilizando o PDI, essa estruturação consiste em vinte e três transformações de 26 tabelas diferentes. Todas essas transformações são realizadas em etapas de entrada de dados, via SQL com tratativas simples, depois podem passar por uma etapa de ordenação, junção de tabelas, filtros, renomeações até reestruturações. Por fim, a etapa final grava a tabela gerada tratada no *Data Lake*. Cada uma dessas transformações é executada em formato ordenado por um *job*, o qual possui um aviso de envio de e-mail automático caso de alguma falha no processo ou ao término da gravação de todas as etapas realizadas com sucesso. Todo esse processo pode ser observado na figura 18:

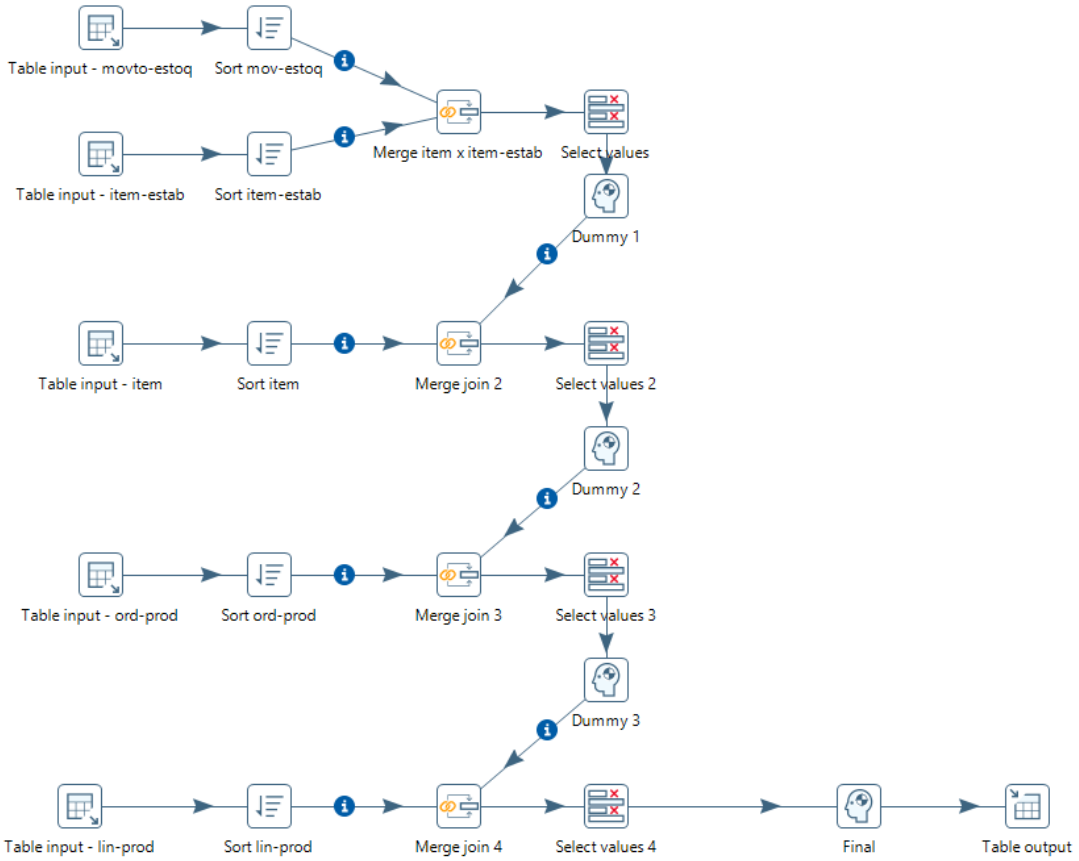
Figura 18 - Job com todas as transformações do faturamento



Fonte: Autor (2021)

Em uma visão míope do processo, é possível observar as etapas de transformação de forma separada de um *job*. Nesses processos, é possível receber várias entradas de informação ao mesmo tempo, o qual, por meio de etapas de ETL, tratam os dados, conectam as informações e gravam no *Data Lake*, conforme demonstrado na figura 19:

Figura 19 - Exemplo de um transformador no projeto do estoque



Fonte: Autor (2021)

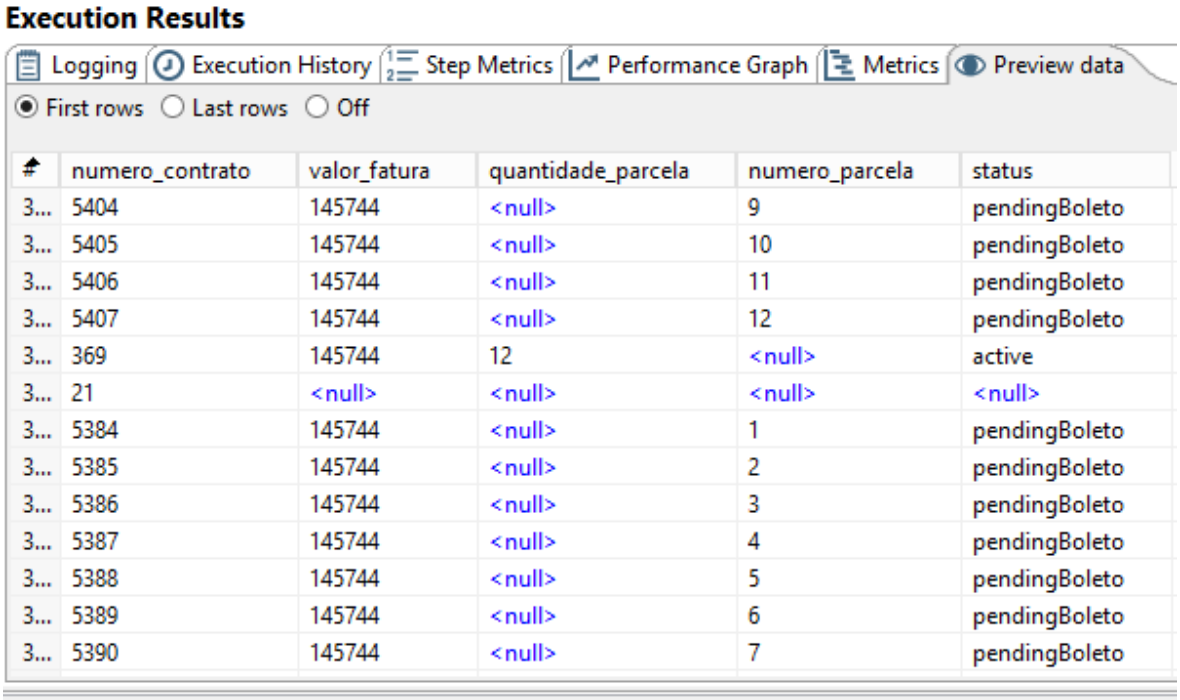
Além da utilização do PDI para realizar os processos de ETL, também utilizamos para realizar integração com API. De uma forma prática e simples, é criado uma etapa de *Rest Client* para buscar as informações na API necessária, para em seguida, ler as informações em um formato JSON, que é possível selecionar os campos desejados e definir seu tipo, nome e alterações necessárias, conforme exemplo na figura 20:

Figura 20 - Exemplo de um transformador fazendo integração com uma API

Fonte: Autor (2021)

Outro facilitador do PDI, é a apresentação das informações diretamente pelas etapas, no qual é possível verificar se as tratativas estão adequadas ao necessário e ter uma breve visão de uma amostra de dados, evitando a execução de processos repetidas vezes para validação das informações. Demonstrado de forma objetiva na figura 21.

Figura 21 - Resultado em dados de uma API

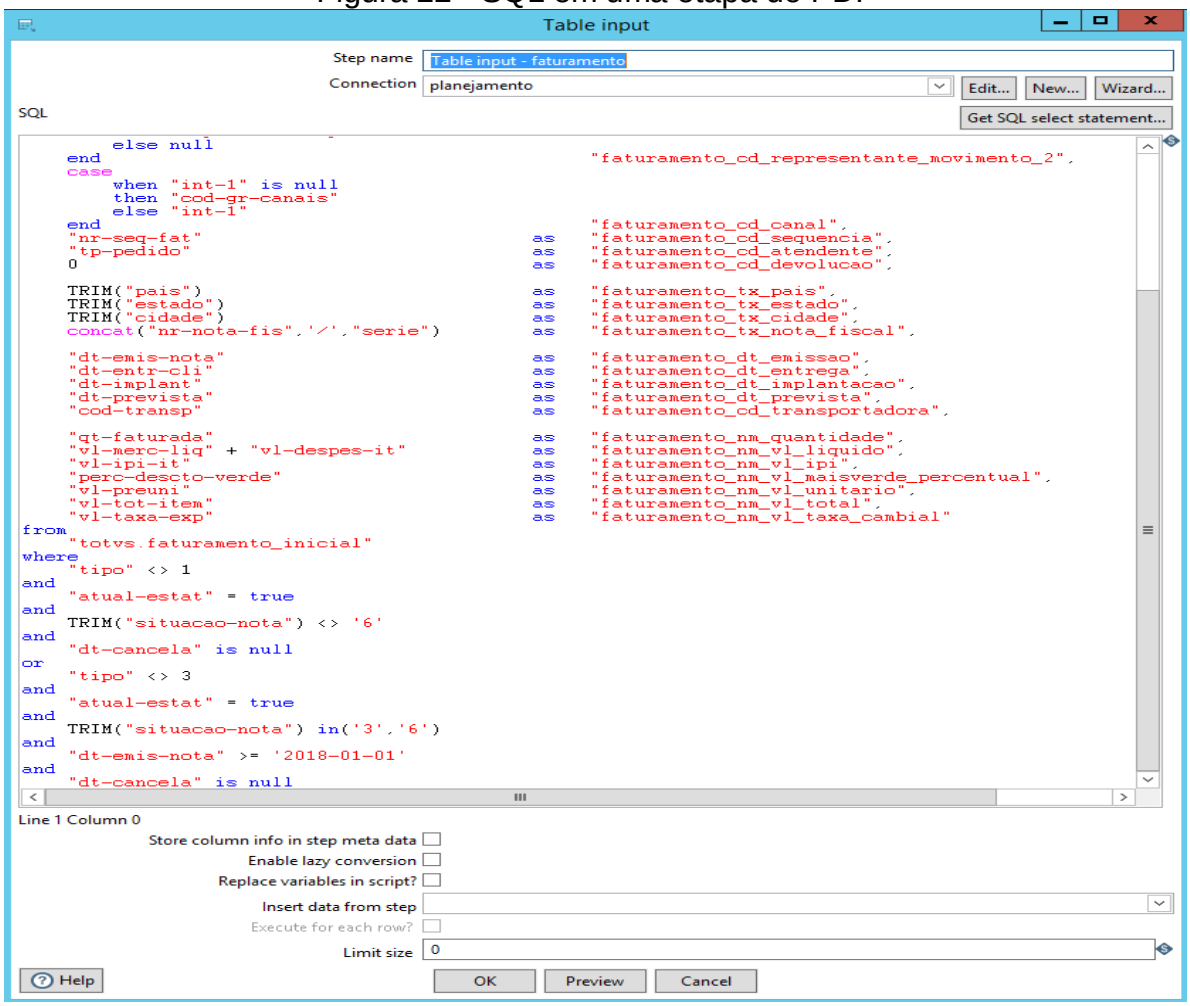


#	numero_contrato	valor_fatura	quantidade_parcela	numero_parcela	status
3...	5404	145744	<null>	9	pendingBoleto
3...	5405	145744	<null>	10	pendingBoleto
3...	5406	145744	<null>	11	pendingBoleto
3...	5407	145744	<null>	12	pendingBoleto
3...	369	145744	12	<null>	active
3...	21	<null>	<null>	<null>	<null>
3...	5384	145744	<null>	1	pendingBoleto
3...	5385	145744	<null>	2	pendingBoleto
3...	5386	145744	<null>	3	pendingBoleto
3...	5387	145744	<null>	4	pendingBoleto
3...	5388	145744	<null>	5	pendingBoleto
3...	5389	145744	<null>	6	pendingBoleto
3...	5390	145744	<null>	7	pendingBoleto

Fonte: Autor (2021)

Dentre os inúmeros passos em blocos do PDI, o mais utilizado em todos os projetos é o *Table input*, pois além de fazer a conexão com os bancos necessários, é possível desenvolver uma série de filtros e tratativas iniciais direto no carregamento dos dados brutos. Tornando o processo menos oneroso e trazendo um desempenho maior ainda referente as próximas etapas, as quais, irão receber apenas os dados necessários para as demais modificações. Segue abaixo exemplo desta função na figura 22:

Figura 22 - SQL em uma etapa do PDI



Fonte: Autor (2021)

5.2.2 Python

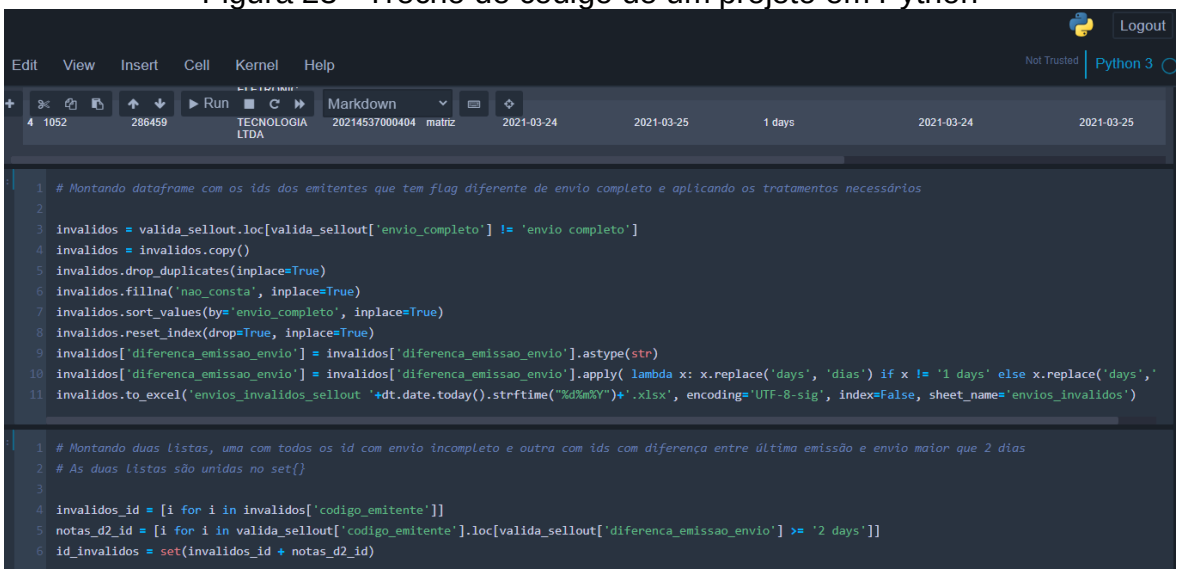
O desenvolvimento utilizando a linguagem Python, se torna uma opção extremamente viável, quando a necessidade do projeto é mais complexa. Além de possuir uma das maiores comunidades ativas da área de dados, existem bibliotecas completamente focadas em análise e manipulação de dados, como a biblioteca pandas.

Isso aliado à sintaxe direta da linguagem fazem com a implementação de projetos na área de dados se deem com mais celeridade, sendo possível ter um retorno de produtividade em pouco tempo quando comparado à outras tecnologias.

Os projetos selecionados para a implementação em Python, são projetos os quais possuem muitas necessidades específicas, assim evitamos usar o PDI, porque o foco dele são situações rotineiras de tratamentos e grande volume de dados.

Com a flexibilidade da linguagem, junto com toda a estrutura as suas bibliotecas disponibilizam, fica evidente o motivo pelo qual é utilizado o Python nesses projetos, de forma clara apresentado um trecho do código na figura 23.

Figura 23 - Trecho de código de um projeto em Python

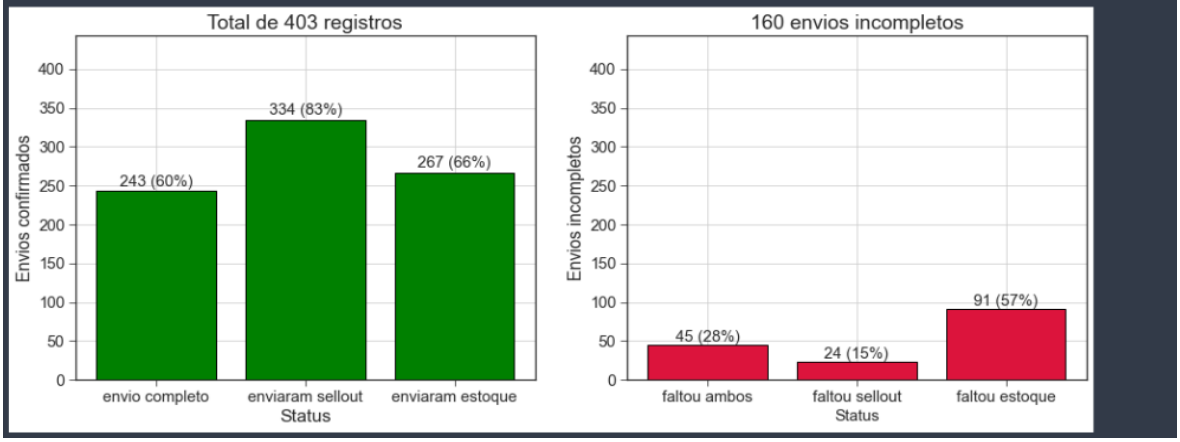


Fonte: Autor (2021)

Os projetos selecionados para serem desenvolvidos em Python, necessitam de uma série de extrações e formatações de dados, as quais devem ser armazenadas em variáveis locais para realização de cálculos e normalizações. Entre esses dados armazenados, várias vertentes são formadas, entre cálculos de percentuais às análises preditivas simples. No final, da maioria desses projetos, é enviado uma análise em formato de gráfico, demonstrado na figura 24, para os responsáveis via e-mail, o qual são utilizados para tomadas de decisões, e todo o histórico é armazenado no *data lake* para utilização de dados com maior assertividade graças a grande quantidade de insumos históricos.

Figura 24 - Gráfico gerado em Python

```
    yval = bar.get_height()
    porcentagem = round((yval / invalidos.shape[0]) * 100)
    string_yval = str(yval)+' ('+str(porcentagem)+'%)'
    plt.text(bar.get_x() + bar.get_width()/2.0, 1.01* yval, string_yval, ha='center', va='bottom', rotation=0, fontsize=14)

plt.title(f'{invalidos.shape[0]} envios incompletos', fontsize=18)
plt.ylabel('Envios incompletos', fontsize=16)
plt.xlabel('Status',  fontsize=14)
plt.ylim(0,valida_sellout.shape[0]+40)
plt.grid(True)

# Salvando a figura com nome e data atual
plt.savefig("validacao de envio de sellout "+dt.date.today().strftime("%d%m%Y")+".png", bbox_inches='tight')

#fig, axes = plt.subplots()
```



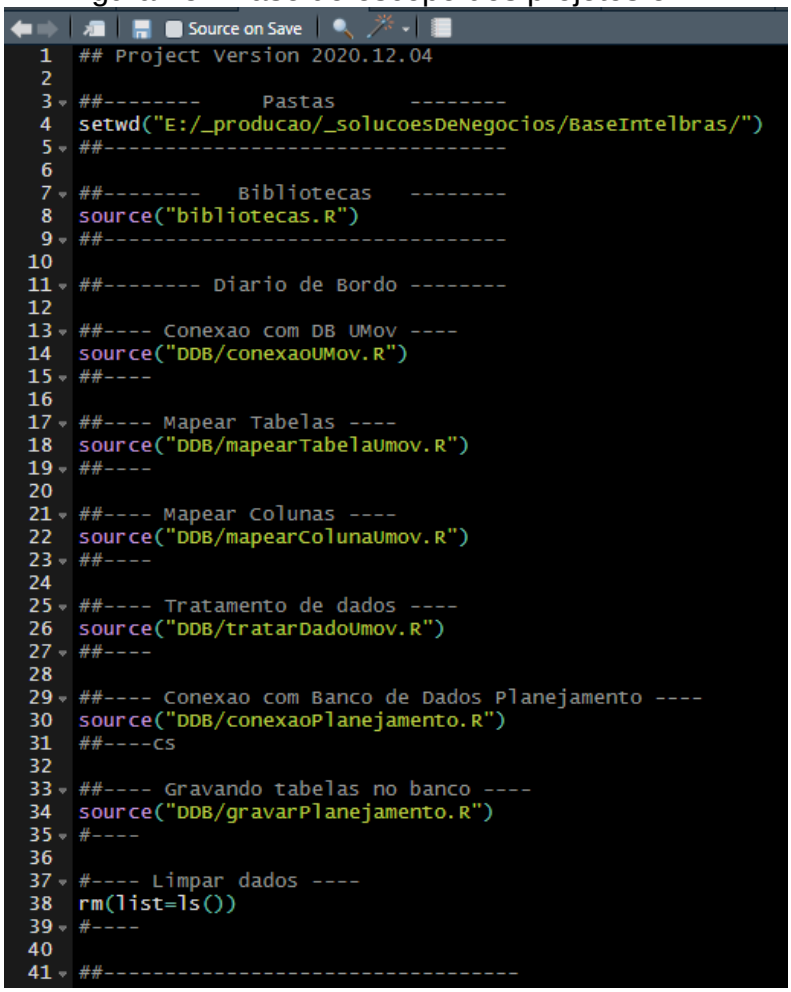
Fonte: Autor (2021)

5.2.3 R

Outra linguagem muito utilizada para análise e tratamento de dados é a linguagem R. Além de ser uma linguagem que vem crescendo bastante, ela possui como foco a manipulação de dados, cálculos e exibição gráfica (R 2021).

Dentre os projetos, foi utilizado a linguagem R nos que haviam a necessidade de cálculos complexos entre grandes volumes de informações, conforme a figura 25. Como não há a necessidade de criação de gráficos diretamente do R, devido a utilização de um *BI* para demonstração dos resultados, esse processo muito utilizado em R, não será necessário.

Figura 25 - Base do escopo dos projetos em R

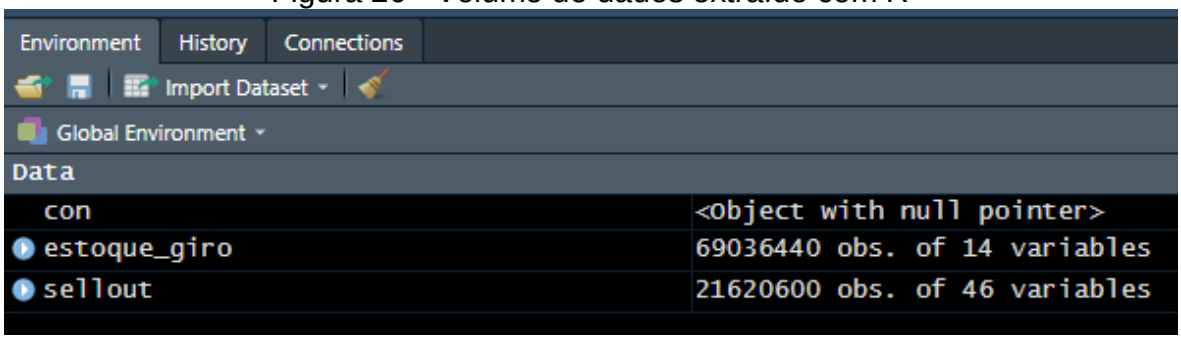


```
## Project Version 2020.12.04

##--------     Pastas      --------
setwd("E:/_producao/_solucoesDeNegocios/BaseIntelbras/")
##---------------------------------

##--------    Bibliotecas   --------
source("bibliotecas.R")
##---------------------------------

##-------- Diario de Bordo --------

##---- Conexao com DB UMov ----
source("DDB/conexaoUMov.R")
##----

##---- Mapear Tabelas ----
source("DDB/mapearTabelaUmov.R")
##----

##---- Mapear Colunas ----
source("DDB/mapearColunaUmov.R")
##----

##---- Tratamento de dados ----
source("DDB/tratarDadoUmov.R")
##----

##---- Conexao com Banco de Dados Planejamento ----
source("DDB/conexaoPlanejamento.R")
##----cs

##---- Gravando tabelas no banco ----
source("DDB/gravarPlanejamento.R")
#----

#---- Limpar dados ----
rm(list=ls())
#----

##---------------------------------
```

Fonte: Autor (2020)

Após a implementação dos projetos em R, é possível observar a facilidade em que a IDE RStudio junto com a linguagem trazem perante aos cálculos de regras de negócios com grande quantidade de volume. Um exemplo prático, é o projeto demonstrado na da figura 26 que possui um *dataset* com 69 milhões de linhas, o qual, sempre é carregado para servir como base de histórico, para a implementação de cálculos diversos para gravação dos dados de forma estruturada no *Data Lake*, e todo esse processo, é rodado automaticamente uma vez ao dia, e leva menos de quatro minutos.

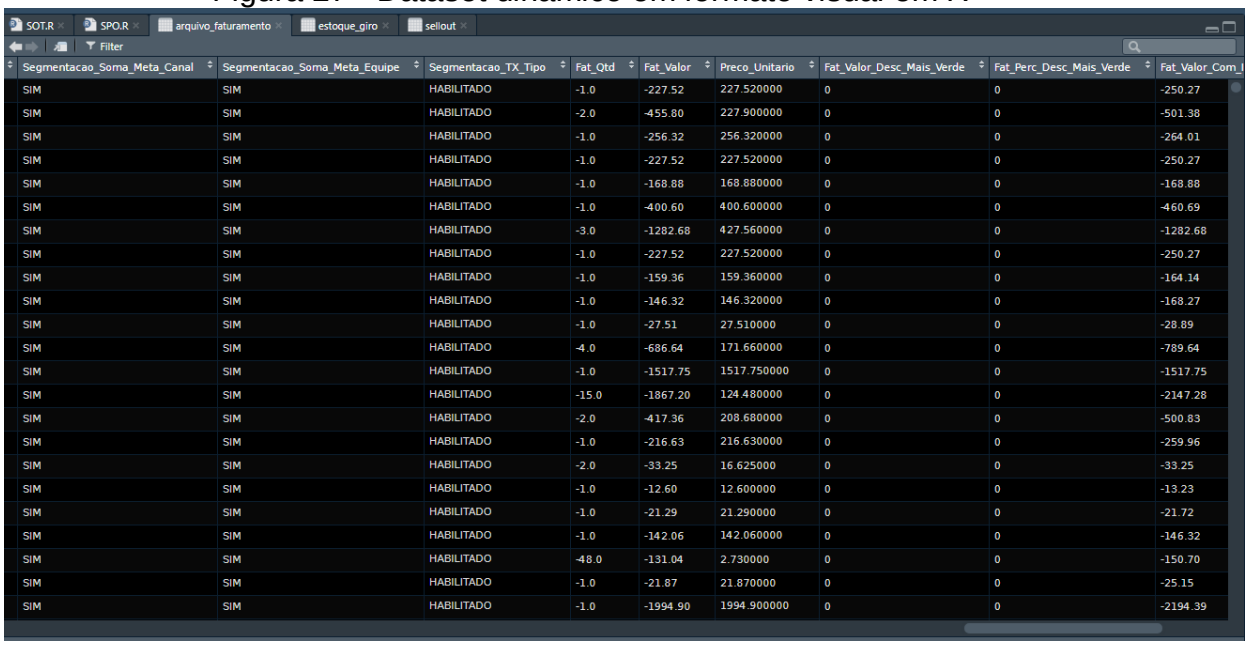
Figura 26 - Volume de dados extraído com R



Fonte: Autor (2021)

Outro facilitador que contribui muito na hora da análise das informações para entender melhor as informações como um todo, é a apresentação do *dataset* em formato visual. Muito útil enquanto desenvolve uma aplicação e confere o resultado das tratativas e cálculos em um formato visual final antes da gravação no *data lake,* conforme a figura 27.

Figura 27 - Dataset dinâmico em formato visual em R



SOT.R | SPO.R | arquivo_faturamento | estoque_giro | sellout

Filter

Segmentacao_Soma_Meta_Canal	Segmentacao_Soma_Meta_Equipe	Segmentacao_TX_Tipo	Fat_Qtd	Fat_Valor	Preco_Unitario	Fat_Valor_Desc_Mais_Verde	Fat_Perc_Desc_Mais_Verde	Fat_Valor_Com_I
SIM	SIM	HABILITADO	-1.0	-227.52	227.520000	0	0	-250.27
SIM	SIM	HABILITADO	-2.0	-455.80	227.900000	0	0	-501.38
SIM	SIM	HABILITADO	-1.0	-256.32	256.320000	0	0	-264.01
SIM	SIM	HABILITADO	-1.0	-227.52	227.520000	0	0	-250.27
SIM	SIM	HABILITADO	-1.0	-168.88	168.880000	0	0	-168.88
SIM	SIM	HABILITADO	-1.0	-400.60	400.600000	0	0	-460.69
SIM	SIM	HABILITADO	-3.0	-1282.68	427.560000	0	0	-1282.68
SIM	SIM	HABILITADO	-1.0	-227.52	227.520000	0	0	-250.27
SIM	SIM	HABILITADO	-1.0	-159.36	159.360000	0	0	-164.14
SIM	SIM	HABILITADO	-1.0	-146.32	146.320000	0	0	-168.27
SIM	SIM	HABILITADO	-1.0	-27.51	27.510000	0	0	-28.89
SIM	SIM	HABILITADO	-4.0	-686.64	171.660000	0	0	-789.64
SIM	SIM	HABILITADO	-1.0	-1517.75	1517.750000	0	0	-1517.75
SIM	SIM	HABILITADO	-15.0	-1867.20	124.480000	0	0	-2147.28
SIM	SIM	HABILITADO	-2.0	-417.36	208.680000	0	0	-500.83
SIM	SIM	HABILITADO	-1.0	-216.63	216.630000	0	0	-259.96
SIM	SIM	HABILITADO	-2.0	-33.25	16.625000	0	0	-33.25
SIM	SIM	HABILITADO	-1.0	-12.60	12.600000	0	0	-13.23
SIM	SIM	HABILITADO	-1.0	-21.29	21.290000	0	0	-21.72
SIM	SIM	HABILITADO	-1.0	-142.06	142.060000	0	0	-146.32
SIM	SIM	HABILITADO	-48.0	-131.04	2.730000	0	0	-150.70
SIM	SIM	HABILITADO	-1.0	-21.87	21.870000	0	0	-25.15
SIM	SIM	HABILITADO	-1.0	-1994.90	1994.900000	0	0	-2194.39

Fonte: Autor (2021)

Com a conclusão os processos de ETL, utilizando as ferramentas e linguagens mencionadas, com as informações estruturadas e gravadas no *data lake*, resta a etapa de apresentação para analises.

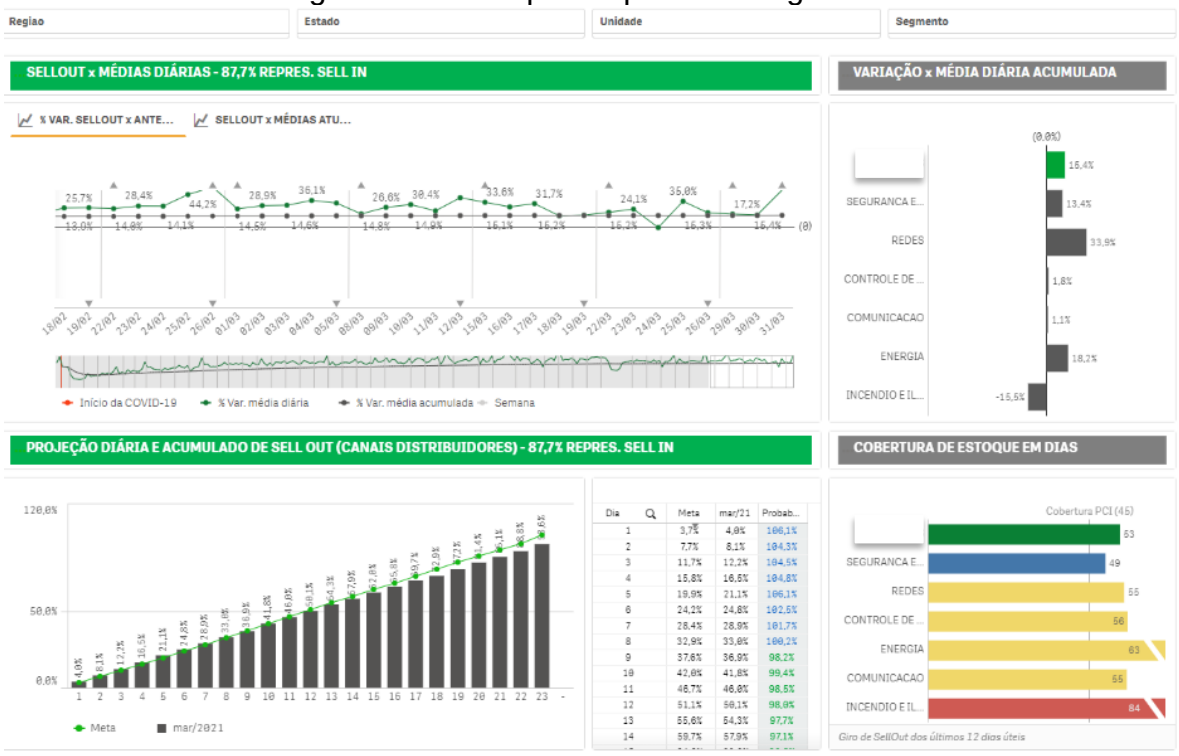
5.3 APRESENTAÇÃO DOS RESULTADOS

Após realizar as tratativas necessárias com a ferramenta predeterminada, dependendo das especificações já citadas nos tópicos 5.2, e realizar a gravação no *data lake*, se inicia a etapa da criação de painéis informativos na plataforma de *BI* Qlik Sense.

Com a utilização de uma plataforma de *BI*, com os dados já tratados e calculados no *data lake*, a maior complexidade que resta é a estruturação de um painel legível, informativo e simples.

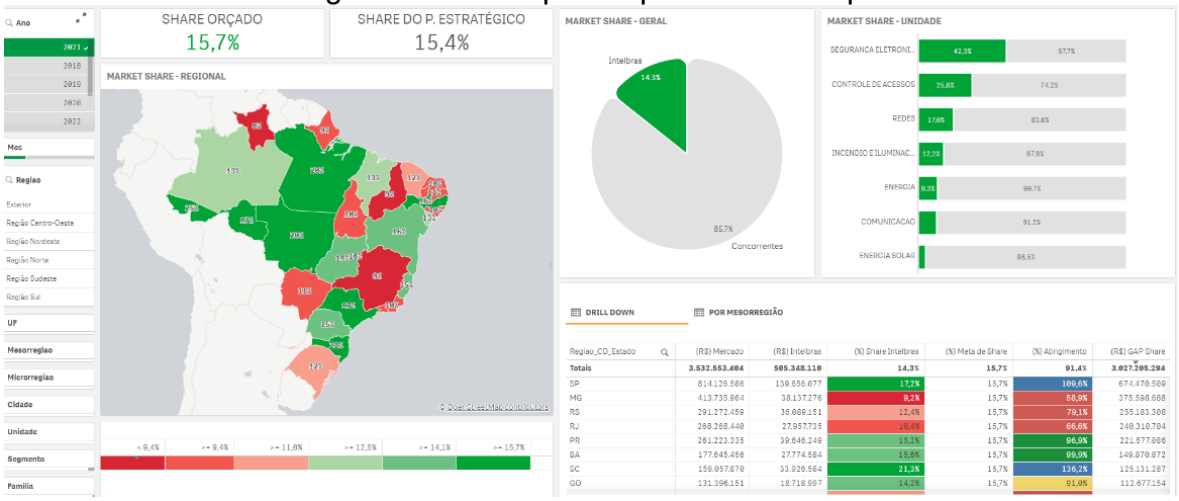
Esses dados, demonstrados em formato de análises pela plataforma de *BI*, como demonstrado na figura 28 e 29, são responsáveis pela orientação de tomadas de decisões, ou seja, a garantia que esses dados estejam calculados e manipulados são de extrema importância, por isso, todos os processos anteriores são de extrema importância.

Figura 28 - Exemplo de painel com gráficos



Fonte: Autor (2021)

Figura 29 - Exemplo de painel com mapa



Fonte: Autor (2021)

O usuário final das plataformas irá atuar com um perfil consultivo, o qual, terá como prioridade visualizar os dados, não tendo mais a necessidade de realização de análises e cálculos repetitivos, trazendo um desempenho em formato de economia de tempo e esforços desnecessários os quais eram realizados de forma manual e repetitiva.

Desta forma, é possível afirmar que a implementação da solução em formato de arquitetura *big data* armazenado em um *data lake*, com a implementação de projetos em python, R e *pentaho data integration*, apresenta dados confiáveis para apresentação de forma simples no *BI*.

6 CONCLUSÃO E TRABALHOS FUTUROS

Neste capítulo é apresentado a conclusão adquirida deste trabalho, após o levantamento de informações e desenvolvimento da arquitetura, e as próximas etapas para trabalhos futuros.

6.1 CONCLUSÃO

Este trabalho apresentou como o desenvolvimento de uma arquitetura de *big data* pode ser desenhada para oferecer uma visão sistêmica relacionada a gestão da indústria de forma prática.

Inicialmente, foi realizado uma busca por informações esclarecedoras referente ao conceito *big data* na indústria, para comprovar com fatos, a real necessidade de uma implementação. A partir das pesquisas, foram apresentadas tecnologias e ferramentas as quais poderiam vir a contribuir com o desenvolvimento da solução. De encontro a parte, foram apresentados estudos e delimitações para centrar o real foco do trabalho.

Posterior as pesquisas, foi realizado uma apresentação do estado atual dos processos manuais o qual a empresa do segmento industrial analisada se encontrava, juntamente com exemplos gráficos de fluxogramas. Com base na atual necessidade foi projetado uma ideia de como resolver esta problemática com *big data*.

Durante o desenvolvimento, algumas limitações ao plano inicial fizeram com que mudanças fossem necessárias, entretanto, sem limitar a obtenção do resultado final esperado. Com a criação da arquitetura para armazenamento do *big data*, com ferramentais previamente explicadas, foi possível utilizar um sistema de *BI* para a apresentação de resultados, trazendo assim, uma maior produtividade para os usuários e juntamente com isso, as respostas para as problemáticas deste trabalho.

Portanto, é possível afirmar que a implementação da arquitetura atende as características de empresas com o segmento industrial, utilizando ferramentas e estruturas que proporcionam o armazenamento de dados em um servidor dedicado.

Com a finalização, é respondido as duas perguntas da pesquisa:

a) A arquitetura *big data* desenhada e implementada com a utilização de ferramentas ETL, linguagens de programação, centralização de dados,

data lake e *BI* oferecem uma visão dos dados através das análises estruturadas, trazendo uma assertividade mais eficaz nas tomadas de decisão.

b) As implementações das tecnologias foram restritas perante as limitações da empresa do segmento industrial, porém, mesmo assim, foi possível estruturar de forma simples e eficiente as ferramentas apresentadas descritas no desenvolvimento da solução.

6.2 TRABALHOS FUTUROS

Com o desenvolvimento deste trabalho, foi identificado alguns pontos os quais podem serem implementados em trabalhos futuros:

- Implementação de um banco de dados não relacional Mongo DB para maior abrangência de captação de dados;
- Implementar um ecossistema *hadoop*;
- Desenvolver uma análise econômica e qualitativa da utilização da atual arquitetura de um *docker* em servidor local para a migração em nuvem;
- Utilização do *big data* para criação de novas tecnologias munidas por IA e *machine learning;*

REFERÊNCIAS

BARBIERI, Carlos. ***BI* - Business Intelligence – Modelagem e Tecnologia.** 1. ed. Rio de Janeiro: Axcel Books, 2001.

CRUZ, E.P.; COVA, C. J. G. Teoria das Decições: Um Estudo do Método Lexicográfico. **RPCA**, Rio de Janeiro, v. 1, n.1, p.26-35, set./dez. 2007.

CONDIE, T. et al. ***MapReduce* Online**. 2010. Disponível em: <http://db.cs.berkeley. edu/papers/nsdi10-hop.pdf>. Acesso em: 27 Out. 2020.

CONFEDERAÇÃO NACIONAL DA INDÚSTRIA - CNI. **Desafios para a indústria 4.0 no Brasil**. Brasília: CNI, 2016.

DAVENPORT, T. H. ***Big data* no trabalho: derrubando mitos e descobrindo oportunidades**. 1º .ed. Elsevier Ltd, 2014

DEAN, J.; GHEMAWAT, S. ***MapReduce***: Simplified data processing on large clusters. 2004. Disponível em: <http://research.google.com/archive/mapreduce.html>. Acesso em: 20 Nov. 2020.

DEAN, J.; GHEMAWAT, S. ***MapReduce***: A flexible data processing tool. 2010. Disponível em: <http://doi.acm.org/10.1145/1629175.1629198>. Acesso em: 20 Nov. 2020.

DELOITTE. ***Industry 4.0****: Challanges and solutions for the digital transformation and use of exponential technologies. Zurich*: Deloitte AG, 2015. Disponível em: Acesso em: 17 Nov 2020.

DOCKER. **Site Oficial Docker**. Disponível em: < https://www.docker.com/>. Acessado: 22 Mai 2021

DRATH R.; HORCH A. ***Industrie 4.0****: Hit or Hype?* IEEE Industrial Eletronics Magazine, 2014.

FONTANA, Glaucio. Data Warehouse. **Material didático da disciplina Business Intelligence.** Palhoça: UnisulVirtual, 2009.

GARTNER. ***Big Data***. 2012. Disponível em: <http://www.gartner.com/it-glossary/ big-data/>. Acesso em: 18 Out. 2020.

GANTZ, J.; REINSEL, D. The digital universe in 2020: *Big data*, bigger digital shadows, and biggest growth in the far east. Framingham**-**USA: IDC iView: IDC Analyze the Future, 2012.

GERMANY TRADE & INVEST - GTAI. **Industrie 4.0**: Smart manufacturing for the future.

Berlin: GTAI, 2016. Disponível em: <https://www.manufacturingpolicy.eng.cam.ac.uk/policies-documents-folder/germany-industrie-4-0-smart-manufacturingfor-the-future-gtai/view>. Acesso em: 18 Nov. 2020.

GIL, A. C. **Métodos e técnicas de pesquisa social**. São Paulo: Atlas, 2008. v. 6.

HABITZREITER, Piter. ***Big data*: uma análise conceitual, abordando suas aplicações, técnicas e ferramentas.** 2014. 72 f. Monografia (Graduação) - Curso de Engenharia da Produção, Universidade do Vale do Rio dos Sinos, São Leopoldo, 2014.

HEERDT, Mauri Luiz. LEONEL, Vilson. **Metodologia Científica e da Pesquisa**, UNISUL, 2007. Disponível em: <http://www.fatecead.com.br/mpc/aula01_ebook_unisulvirtual.pdf>. Acesso em: 29 nov. 2020.

REGINATO, Luciane; NASCIMENTO, Auster Moreira. Um estudo de caso envolvendo Business Intelligence como instrumento de apoio à controladoria. **Rev. contab. financ**., São Paulo, v. 18, jun 2007. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext &pid=S1519-70772007000300007>. Acesso em: 21 Jan 2021.

RIBEIRO, R. Linkedin. **ETL vs ELT? Quando duas letras fazem muita diferença**.,2018. Disponivel em: <https://www.linkedin.com/pulse/etl-vs-elt-quando-duas-letrasfazem-muita-diferença-rodrigo-r-g-/>. Acesso em: 24 mai. 2021.

RIFFAT, M. *Big data*: Not a panacea. **ISACA Journal**, v. 3, 2014.

MACHADO, F.N.R. **Big Data: O Futuro dos Dados e Aplicações**. 1. ed. São Paulo: Saraiva, 2018

MANYIKA, J. et al. ***Big data*: The next frontier for innovation, competition, and productivity**. 2011. Disponível em: <www.mckinsey.com/mgi>. Acesso em: 13 set.
2020.

MARTIN. **Industry 4.0**: Definition, Design Principles, Challenges, and the Future of Employment. 2017. Disponível em: <https://www.cleverism.com/industry-4-0/> Acesso em: 17 Nov 2020.

MATOS, D. *Data Lake*, a Fonte do *Big Data*. **Ciência e Dados**, 2018. Disponível em:<http://www.cienciaedados.com/data-lake-a-fonte-do-big-data> . Acesso em: 21 mai. 2021.

MATOS, D. NoSQL Database. **Ciência e Dados**, 2019. Disponível em: <http://www.cienciaedados.com/nosql-database/>. Acesso em: 02 mai. 2021.

MCAFEE, A.; BRYNJOLFSSON, E. ***Big Data*: The management revolution**. 2012.
Disponível em: <https://hbr.org/2012/10/big-data-the-management-revolution/>. Acesso em: 27 Out. 2020.

MEZZAROBA, O.; MONTEIRO, C. S. **Manual de Metodologia da Pesquisa no Direito**. São Paulo: Saraiva, 2004. ISBN 8502048694.

MICHAEL., K.; MILLER, K. W. ***Big Data***: New opportunities and new challenges. p.22-24, 07 junho 2013. **IEEE**. Disponível em: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6527259>. Acesso em: 18 Out. 2020.

MONGODB. **Site Oficial MongoDB**. Disponível em: < https://www.mongodb.com/>. Acessado: 22 Mai 2021

MPINDA, A.; BUNGAMA, P.; MASCHIETTO, L. ***From relational database to columnoriented NoSQL database: migration process***. Disponível em: . Acesso em: 19 Nov de 2020.

NETO, Trajano C. M. **Avaliação das Ferramentas ETL open-source Talend e Kettle para Projetos de Data Warehouse em Empresas de Pequeno Porte**. Lauro de Freitas, BA: 2012. Disponível em: <http://www.ambientelivre.com.br/downloads/doc_download/87-tcc-ferramentas-deetl-open-source-talend-e-kettle.html>. Acesso em: 30 Set 2020.

NIST. **NIST *Big Data* Interoperability Framework**: Volume 1, Definitions. v. 1, p. 32, 2015.

PYTHON. **Site Oficial Pybrain**. Disponível em: < https://www.python.org/>. Acessado: 8 Jan 2021

RATHBONE, M. **Real World Hadoop: Implementing a left outer join in map reduce**. 2013. Disponível em: <http://blog.matthewrathbone.com/2013/02/09/ real-world-hadoop-implementing-a-left-outer-join-in-hadoop-map-reduce.html>. Acesso em: 26 Out. 2020.

R. **Site Oficial R**. Disponível em: < https://www.r-project.org/>.Acessado: 8 Jan 2021

SADALAGE, Pramod J.; FOWLER, Martin. ***NoSQL Essencial***: um guia conciso para o mundo emergente da persistência poliglota. São Paulo: Novatec, 2013. 216 p.

SILVA, Ivan Menerval da; CAMPOS, Fernando Celso de. **Novas perspectivas utilizando o *big data*: um estudo bibliométrico** 2000-2012. Proceedings of the 11th contecsi international conference on information systems and technology

management, São Paulo, p.4150-4172, 30 maio 2014. **TECSI**. http://dx.doi.org/10.5748/9788599693100-11contecsi/ps-1035.

SILVA, E. L. da; MENEZES, E. M. **Metodologia da Pesquisa e Elaboração de Dissertação**. Florianópolis: UFSC, 2005. Disponível em: <https://projetos.inf.ufsc.br/ arquivos/Metodologia_de_pesquisa_e_elaboracao_de_teses_e_dissertacoes_4ed . pdf>. Acesso em: 29 Nov. 2020.

SOMASUNDARAM, G; SHRIVASTAVA, Alok; EMC EDUCATION SERVICES. **Armazenamento e gerenciamento de informações: Como armazenar, gerenciar e proteger informações digitais.** Porto Alegre: Artmed Editora S.A., 2011.

TANAKA, Asterio. **Tópicos Avançados de Banco de Dados (Business Intelligence):** Integração de Dados e ETL. Disponível em: <http://www.uniriotec.br/~tanaka/SAIN/03-ETL2015.1.pdf>. Acesso em: 31 out. 2020.

TAURION, C. ***Big Data***. Rio de Janeiro: Brasport Livros e Multimídia Ltda, 2013.

UNIT, EIU–Economist Intelligence. *Big data* - Harnessing a game-changing asset. Londres:**The Economist**, 2011.

VARGAS, Marcelo F. **Construção De Data Warehouse para Pequenas e Médias Empresas Usando Software Livre. Lages**, SC: 2009. Disponível em: <https://revista.uniplac.net/ojs/index.php/tc_si/article/download/826/536>. Acesso em: 04 abr. 2021.

ZIKOPOULOS, P.; EATON, C. ***Understanding big data***: Analytics for enterprise class hadoop and streaming data. New York: Mc Graw Hill, 2012.

APÊNDICE A – CRONOGRAMA DO PROJETO

Cronograma TCC – Aluno: Giovani Reinert Junior – Orientador: Daniella Pinto Vieira

TCC1 (2020)																				
ATIVIDADE	**Agosto**				**Setembro**				**Outubro**				**Novembro**				**Dezembro**			
Semana	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Revisão da literatura			x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Tema da Pesquisa			x	x																
Capitulo 1					X	X														
Capitulo 2							x	x	x	x	x	x	x							
Capitulo 3														x	x	x				
Apêndice A																x				
Correções do Avaliador Externo																	X			
Entrega Versão final TCC1																		X		
TCC2 (2021)																				
ATIVIDADE	**Março**				**Abril**				**Maio**				**Junho**				**Julho**			
Semana	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Revisão da literatura	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Modelagem/Proposta	x																			
Entrega da Proposta	x	x																		
Desenvolvimento		x	x	x	x	x	x	x												
Entrega Capitulo 4								x												
Conclusão									x	x	x	x								
Defesa																		x		
Correções e Entrega																			x	x