Fundamentos de Big Data

Leandro Lessa

2023

SUMÁRIO

Capítulo 1. Introdução ao Big Data

Bem-vindos ao módulo de Fundamentos de Big Data! Nosso objetivo neste módulo são dois. Primeiro, queremos te contextualizar sobre as oportunidades e desafios que o mundo do *big data* oferece – e porque isso tem revolucionado as organizações e empresas desde o começo dos anos 2000.

Em segundo lugar, queremos que você desenvolva desde o primeiro módulo a habilidade de construir códigos e modelos que extraiam valor dos dados; por isso, vamos abordar também algoritmos clássicos empregados para tal. Em particular, você vai conhecer e aplicar o k-means, um algoritmo que segmenta pontos em grupos. Imagine o quanto isso é útil: se você trabalha com marketing, pode descobrir quais os perfis mais comuns de cliente você tem. Em seguida, você vai conhecer a regressão linear, um modelo clássico da estatística que permite fazer previsões a partir dos dados – como estimar quanto vale a sua casa, por exemplo.

Ao fim do módulo, você terá uma visão teórica e prática das possibilidades do *big data*, o que vai te ajudar a começar a perceber oportunidades de aplicá-lo no dia-a-dia na sua carreira e atividade profissional.

Este capítulo tem como objetivo apresentar os principais conceitos básicos relacionados à terminologia do Big Data, bem como características que se relacionam entre inteligência artificial e ciências de dados.

1.1. Introdução ao cenário do Big Data

Com o surgimento da internet, o volume de dados produzidos ao redor do mundo cresceu de forma expressiva com o passar dos anos. A utilização em larga escala das mídias sociais e de dispositivos móveis aumentaram ainda mais a quantidade de dados gerados diariamente.

Segundo a *International Data Corporation* (IDC) - principal fornecedora global de inteligência de mercado, serviços de consultoria e eventos - os dados digitais produzidos no mundo dobram a cada 2 anos. O mundo se move através dos dados e do que se pode extrair deles. Todo esse avanço da tecnologia nos levou ao local onde produzimos milhões de dados por minuto.

Conseguimos ter uma ideia sobre a quantidade de dados produzidos quando temos acesso a alguns comparativos, como a DOMO (empresa especializada em computação na nuvem) por exemplo, que anualmente realiza um estudo que estima a quantidade de informação que é produzida a cada minuto todos os dias. Além de estimar esses dados, o estudo indica hábitos e consumos da população nas principais plataformas sociais. A Figura 1 mostra a quantidade de dados produzidos por minuto no ano de 2020.

Figura 1 - Dados produzidos por minuto em 2023.

Fonte: eDiscovery (2023).

Conforme ilustrado na Figura 1, observamos as principais plataformas que mais produziram dados por minuto na Internet em 2023. Veja, por exemplo, a quantidade de mensagens postadas no Twitter a cada minuto, ou quantas buscas são realizadas no Google.

Com tantos dados produzidos a cada minuto, os métodos tradicionais de coleta, armazenamento e processamento de dados começaram a não ser suficientes, causando problemas e gastos cada vez maiores para suprir as necessidades do negócio. Diante disso, surge o conceito do Big Data. Definimos Big Data como um termo genérico para

qualquer coleção de conjuntos de dados tão grandes ou complexos que se torna difícil processá-los usando técnicas tradicionais de gerenciamento de dados, como, por exemplo, o SGBDs (sistema de gerenciador de banco de dados relacional) (CIELEN; MEYSMAN; ALI, 2016).

Em outras palavras, podemos dizer que o Big Data é uma área do conhecimento que tem como finalidade estudar maneiras de tratar, analisar, processar, armazenar e gerar conhecimento através de grandes volumes de dados de várias fontes, com o objetivo de acelerar a tomada de decisão e trazer vantagem competitiva.

Antes da explosão dos dados, os dados eram armazenados e gerenciados de forma estruturada no formato de tabelas com linhas e colunas nos SGBDs. Para isso, utilizava máquinas com grande capacidade de processamento e armazenamento. Quando surgia a necessidade de expandir a capacidade operacional dessas máquinas, era necessário realizar a instalação de novos componentes de *hardware.*

Os problemas começam a aparecer quando se alcança um grande volume e variedade de dados. Os sistemas gerenciadores de banco de dados começam a apresentar dificuldades de escalabilidade, disponibilidade e flexibilidade de dados. Por exemplo, é muito custoso o aprimoramento dessas máquinas de maneira vertical, no qual aprimoramos uma máquina adicionando mais recursos como memória e processamento, não garante uma efetividade quando se trata de Big Data. Além disso, toda vez que é necessário realizar um *upgrade*, a máquina fica indisponível devido ao processo de manutenção.

Para contornar esse problema do alto custo de manter uma máquina com poder computacional mais elevado, grandes empresas pesquisaram um novo sistema que fosse possível, escalável e que tivesse um custo reduzido. Surge então o *Hadoop*, uma forma de armazenamento e processamento distribuído. A ideia principal é utilizar um *Cluster* (conjunto/agrupamento)

de máquinas que realizam armazenamento e processamento distribuído. Ou seja, no lugar de centralizar o custo operacional em apenas uma “super máquina”, passou a utilizar máquinas com capacidade menor de processamento trabalhando em conjunto. Dessa forma, é possível processar com um grande volume de dados de forma distribuída. Assim, uma única máquina nesse cluster não tem poder de processamento elevado, mas, em conjunto com outras, conseguem fornecer um poder de processamento e armazenamento elevados para os dados do Big Data. Além disso, para aumentar a capacidade de processamento de um cluster, é possível adicionar novas máquinas sem que haja problemas de indisponibilidade. Essa é uma solução chamada de escalabilidade horizontal, encontrada para resolver os problemas de Big Data.

1.2. A origem dos dados do Big Data

As origens dos dados que compõem o Big Data são as mais variadas: Telefones celulares, mídias sociais, inteligências artificiais, tecnologias de imagem e muitos outros sistemas produzem dados que precisam ser armazenados em algum lugar e servir a alguma finalidade. Quando falamos em Big Data, nos referimos não apenas aos dados específicos gerados por uma amostra tímida de usuários, mas principalmente sobre a enorme quantidade de dados produzidos e armazenados diariamente e que não estão necessariamente padronizados e esse é um dos grandes desafios encarados atualmente. Quando pensamos na geração dos dados em si, temos os principais fatores em:

1.2.1. Dados gerados por pessoas

São dados criados pela atividade humana na tecnologia, como:

- Postagens nas redes sociais;
- Mensagens enviadas em aplicativos;
- Textos escritos em *blogs*, revistas ou páginas da *web*;
- Áudios ou vídeos compartilhados;
- E-mails e afins.

As redes sociais merecem destaque no quesito produção de dados. Cada vez que um indivíduo posta qualquer informação, compartilha *links*, realiza compras ou classifica um conteúdo, ele está gerando incontáveis dados que serão utilizados dentro do Big Data.

1.2.2. Dados gerados por máquinas

São aqueles criados a partir das máquinas e sem a necessidade da intervenção humana, pois já são programados para extrair tais dados, como por exemplo:

- Sensores em veículos, eletrodomésticos e máquinas industriais;
- Câmeras e sistemas de segurança;
- Satélites;
- Dispositivos médicos;
- Ferramentas pessoais, como aplicativos de *smartphone* e afins.

1.2.3. Dados gerados por empresas

São dados gerados por empresas: aqueles que as organizações obtêm à medida que administram seus negócios, como por exemplo:

Registros gerados toda vez que você faz uma compra em uma loja on-line ou física - registros como números exclusivos de clientes, os itens que você comprou, a data e hora em que você comprou os itens e quantos de cada item você comprou.

No mundo do Big Data você verá esse dado ser chamado de "dado transacional". O que precisamos ter em mente é que nem todo dado gerado é necessariamente considerado Big Data. O fato de enviar mensagens de texto ao longo do dia, por exemplo, só poderia ser considerado Big Data caso o número esteja perto dos milhões.

1.3. Estrutura dos dados

Como estamos falando de várias fontes geradoras de dados, precisamos levar em conta que os dados podem vir estruturados ou não estruturados, como arquivos de texto ou numéricos e até mesmo em arquivos de multimídia.

Um diferencial entre o Big Data e a análise tradicional dos dados é que em sua maioria ele analisa dados de natureza não estruturada ou

semiestruturada, o que requer diferentes técnicas e ferramentas de processamento e análise.

Para isso, os ambientes de computação distribuída e arquiteturas de processamento paralelo maciço (MPP), que permitem a ingestão e análise de dados paralelizados, são a abordagem preferida para processar esses dados complexos (EMC EDUCATION SERVICES, 2015).

Nos estudos realizados sobre o Big Data, podemos encontrar até três tipos de dados (não estruturados, semiestruturados e estruturados), mas eles são comumente generalizados em dois grupos: estruturados e não estruturados.

1.3.1. Dados estruturados

Refere-se a todos os dados que estejam em conformidade com um determinado formato. Tem como característica ser bem definidos, inflexíveis, pensados antes da própria criação dos dados. Dessa forma, não é possível que tipos de dados diferentes das estruturas preestabelecidas sejam carregados.

Por exemplo, se uma coluna de uma tabela for criada com o tipo de dado numérico, essa coluna não aceitará dados textuais. Um exemplo básico são as planilhas, onde geralmente há linhas e colunas que seguem um determinado padrão.

Os dados estruturados sempre são claramente organizados e mais fáceis de analisar. Em uma planilha, por exemplo, você conseguiria facilmente indicar valores e quantidades listadas, por essa razão muitos dos dados com os quais as organizações trabalham podem ser categorizados como estruturados.

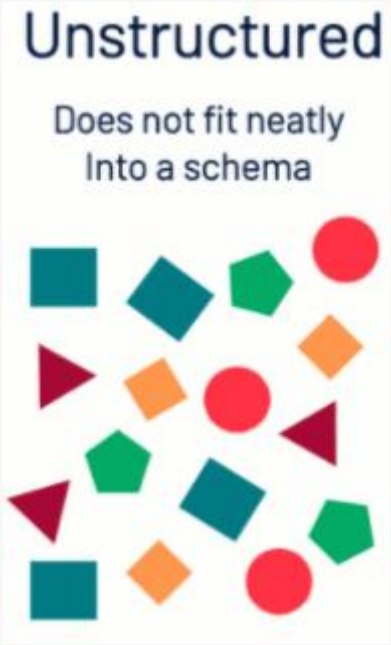
Figura 2 - Representação dos Dados Estruturados.

1.3.2. Dados não estruturados

Os dados não estruturados são o oposto dos dados estruturados. Eles não possuem uma estrutura pré-definida, alinhada ou padronizada. Os dados não estruturados se caracterizam por possuir uma estrutura flexível e dinâmica ou, até mesmo, nenhuma estrutura. Esses dados podem ser compostos por vários elementos diferentes, como: imagens, áudios, vídeos, gráficos e textos. Eles são difíceis de processar devido a sua complexibilidade e formatação. Os dados não estruturados podem ser encontrados em mídias sociais, e-mails, fotos, vídeos, chats, arquivos de logs, sensor de IoT, entre outros.

Hoje os dados não estruturados são os mais difundidos, alcançando cerca de 90% do total dos dados produzidos. Por isso, muitas organizações têm lutado para tentar entender esses dados, principalmente para usá-los em estratégias e ideias em seus negócios. Nesse ponto a Inteligência Artificial tem um grande papel na análise dos dados, já que as análises conterão vídeos, postagens em mídia social, fotografias, e-mails, arquivos de áudio e imagens.

Figura 3 - Representação dos Dados Não Estruturados.



1.3.3. Dados semiestruturados

Os dados semiestruturados se encaixam entre as duas definições anteriores. Eles não residem em uma tabela formatada, porém possuem um certo nível de organização. Esses dados possuem uma estrutura heterogênea, não sendo uma estrutura completamente rígida e nem exclusivamente flexível. Um exemplo desse nível de organização é o código HTML, onde você consegue extrair muitas informações dentro de uma forma específica de expressar os dados. Em muitos casos, os dados dispõem de uma definição regular (por exemplo, um catálogo de produtos), em outros, um padrão estrutural que pode ser identificado ou não existem informações descritivas relacionadas (por exemplo, um arquivo de imagem) (ABITEBOUL, S., 1997).

Figura 4 - Representação dos Dados Não Estruturados.

Fonte: Imagem ilustrativa.

1.4. Os V's do Big Data

Como você pôde perceber pelo explicado até aqui, o Big Data trabalha com uma quantidade surpreendente de dados que estão em variadas formas. A proposta de uma solução de Big Data precisa oferecer informações precisas, mesmo levando em consideração a complexidade dos dados que possui. Por isso o Big Data leva o conceito dos 5 V's: Volume, Velocidade, Variedade, Veracidade e Valor. Existem outras definições que adicionam outros V's, entretanto, vamos focar nas principais características que são essenciais para o processo de análise do Big Data.

Volume: Dentro do Big Data, o volume é mais bem evidenciado por fatos cotidianos, como a quantidade de transações bancárias realizadas por dia, os registros de chamadas, troca de e-mails ou interações em redes sociais. Esses exemplos ajudam a ter uma ideia do volume de dados presente no mundo atualmente.

A IDC prevê que a quantidade de dados que existe no mundo está crescendo de 33 zetabytes, em 2018, para 177 zetabytes, em 2025. Todos os dias é criado um número impensável de dados de forma que não valeria a

pena informar a quantidade de produzidos aqui, uma vez que ao acabar de ler esse texto os dados já terão se alterado.

Velocidade: Refere-se literalmente à rapidez com que os dados são gerados. É o grande diferencial competitivo dentro das empresas, afinal, quanto mais rápido você processa o dado, mais rápido ele se torna uma informação estratégica para sua empresa. É o mesmo conceito de utilizar um mapa desatualizado para conhecer uma nova cidade, provavelmente as informações, ruas e comércios serão diferentes e por isso sua experiência poderá ser desagradável.

A velocidade do Big Data garante uma melhor assertividade nas informações, já que elas são analisadas quase em tempo real.

Estima-se que haverá um momento em que a tecnologia permitirá que os dados sejam analisados em tempo real, e assim as informações serão atualizadas instantaneamente para um fim específico.

Variedade: O volume é o primeiro dos desafios, seguido pela variedade dos dados produzidos e captados atualmente. Como informamos anteriormente, hoje no Big Data a maior parte dos dados não estão estruturados, ou seja, não se encontram agrupados em ordem e separados por assunto. Temos uma infinidade de dados dispersos na rede e isso produz variados pontos de vista sobre uma mesma situação.

As empresas que conseguem usar essa variedade em seu favor têm um valor específico agregado em seu negócio

Veracidade: Um item que está intimamente ligado à velocidade, já que um dado desatualizado não pode ser considerado um dado verídico, uma vez que não condiz com o que está acontecendo naquele momento. É preciso ter em conta que para colher bons resultados dentro do Big Data é imprescindível obter dados verdadeiros e relevantes, ou seja, verificados e ponderados para o propósito da análise em questão.

A veracidade está associada à confiabilidade do dado. Para isso, esses dados precisam possuir as seguintes características:

- Qualidade e consistência:
  - Os dados coletados são realistas? Eles são confiáveis? Ou eles são inventados?
- Origem conhecida:
  - A fonte dos dados é confiável?
- Dados devem ser verdadeiros e não dados fakes:
  - Dados de opinião não devem ser considerados.
- Dados internos e externos à instituição:
  - Dados internos são mais fáceis de verificar sua veracidade do que dados externos.

Valor: A troca do resultado pelo custo do investimento. Para saber realizar todo o processo de Big Data dentro do negócio, precisa levar em conta o custo do mesmo, ter uma visão realista sobre onde aplicar os resultados e, principalmente, saber exatamente qual informação se procura. É necessário esse foco para obter o valor real do processo e assim pesar o custo e benefício. Em outras palavras, os dados agregam valor para a empresa? A aplicação do Big Data aumentou a receita da empresa, reduziu custos? Encontrou alguma nova oportunidade de negócios? Melhorou a qualidade do produto ou serviço? Aumentou a satisfação do cliente? Garantiu melhores resultados? Todas essas perguntas são fundamentais para saber se os dados possuem valor para a empresa.

1.5. O potencial do Big Data

O potencial do Big Data está relacionado à capacidade de interpretar e aplicar dados externos e não estruturados. O grande desafio é processar dados não estruturados de forma eficaz e inteligente e, ao mesmo tempo, ter a capacidade de transformar e direcionar a organização baseado nos dados coletados. Dessa forma, as empresas precisam desenvolver uma nova forma de aplicar o *Business Intelligence* (BI) trazendo informações inteligentes e em larga escala.

A importância do Big Data nas empresas está diretamente relacionada à vantagem competitiva. No mundo corporativo, existem exemplos notáveis de criação de vantagem competitiva a partir de estratégias baseadas em técnicas de Big Data. Por exemplo, é possível definir perfil de consumidores e oferecer produtos a partir de uma análise em tempo real. Sem dúvida, as empresas que conseguem tirar proveito do Big Data conseguem sair na frente da sua concorrência.

Existem diversas vantagens ao se utilizar toda a capacidade dos dados coletados e processados, podemos citar, por exemplo, a otimização dos custos ou um aumento significativo nos lucros. Abaixo contém alguns outros exemplos de vantagens do uso do Big Data.

- Tomada de decisão mais rápida e assertiva;
- Acesso às informações privilegiadas;
- Interpretação de tendências de mercado a partir da análise dos eventos;
- É possível prever situações futuras e criar plano de ação;
- Acompanhar a aceitação de novos produtos ou serviços no mercado e tomar decisões a fim de potencializar ou corrigir problemas que possam acontecer de forma rápida e eficaz;

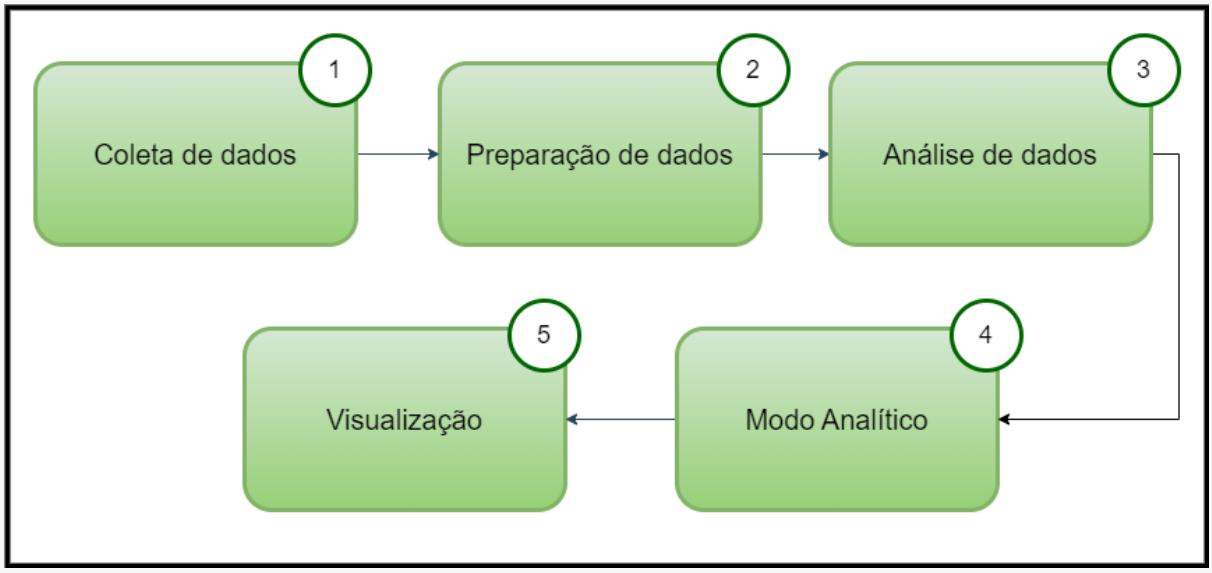
- Analisar o perfil de comportamento dos clientes e identificar oportunidades de vendas contínuas;
- Descobrir novas oportunidades de demanda, produtos e serviços;
- Acompanhar as rotas das frotas e identificar problemas em tempo real, como excesso de velocidade e caminhos obstruídos;
- Acompanhar a experiência do cliente em relação à fidelização, nível de satisfação;
- Atuar rapidamente em casos de detecção de fraudes, clonagem de cartão, crimes de forma geral;
- Monitorar ações de *marketing* em tempo real.

1.6. Etapas do processamento do Big Data

O Big Data necessita de um processo bem elaborado para transformar dados variados em informações úteis, pois só assim ele estará cumprindo sua função. Nesse processo, um fluxo de atividades é estabelecido para que o processamento do Big Data leve a construção de informações e geração de *insights*.

São necessárias 5 etapas para a realização desse processamento, conforme você poderá ver na imagem abaixo:

Figura 5 - Etapas do processamento do Big Data.



Fonte: Adaptado pelo autor (Bahga e Madisetti, 2016).

1.6.1. Coleta de dados

Aqui são definidas as fontes a serem utilizadas para gerar armazenamento, podendo ser bancos de dados SQL, NoSQL, *Frameworks* e muitos outros, precisando que se leve em conta o objetivo final da coleta/projeto (OLSEN, 2015).

No caso de um projeto voltado para o marketing, podem ser utilizadas fontes que indiquem o comportamento do consumidor através de cliques em anúncios, logins de site, sistemas operacionais dos dispositivos utilizados, localização e históricos de compra e pesquisa. Essa é a fase que gera suporte para todas as outras.

1.6.2. Preparação dos dados

É a etapa minuciosa onde os dados são "limpos" e se tornam aptos para a análise. Essa etapa tem o objetivo de reduzir discrepâncias, evitar valores nulos, indicar as anomalias e preencher lacunas. Essas informações, se não tratadas, acabam prejudicando o processo como um todo, já que a análise poderá ser comprometida caso os dados não sejam verdadeiros.

Além de evitar os ruídos, ou seja, os resultados indesejados, essa limpeza proporciona uma pesquisa mais específica, aumentando a

qualidade dos resultados apresentados e permitindo novos *insights* no resultado.

1.6.3. Análise de dados

É a etapa onde selecionamos os algoritmos de *Machine Learning* que serão inseridos para alcançar o resultado, levando em consideração o conjunto de dados obtidos na etapa anterior. Por exemplo, podemos utilizar a clusterização caso nosso objetivo seja agrupar o perfil de certos clientes.

1.6.4. Modo analítico

O modo analítico define como os dados devem ser analisados. Levando sempre em conta o objetivo final do processo de Big Data, aqui é definido se os dados devem ser analisados em tempo real, pelos dados históricos ou pela interação com o usuário. Quando se decide realizar uma análise em tempo real, a aplicação deve coletar a cada momento os dados e analisá-los. Por outro lado, quando se define realizar uma análise com dados de histórico, utilizamos dados por um período maior de coleta. Por exemplo, um grande volume de dados coletados durante um mês. E por fim, a análise interativa ocorre devido a maior flexibilidade de interações com os usuários. À medida que o usuário vai interagindo com o sistema, é gerado as análises que ele deseja. Por exemplo, analisar um conjunto de vendas por estado, conjunto de vendas por item, valor vendido no mês etc.

1.6.5. Visualização

Essa última etapa trata sobre a exibição dos dados ao usuário da aplicação, podendo ser, por exemplo, uma visualização estática, interativa ou até mesmo dinâmica.

O poder transformador do processo de Big Data está na boa escolha e execução desses cinco passos. A escolha correta dos algoritmos, da base de dados e dos filtros a serem aplicados para gerar o resultado são de suma importância para que o processo ocorra bem e alcance o objetivo esperado.

1.7. Aplicações do Big Data

As aplicações do Big Data podem ser diversas. Vamos imaginar o seguinte cenário: uma empresa de varejo possui um processo de Big Data bem definido no qual consegue entender o perfil dos seus clientes e prever o que farão a seguir. Imagine o potencial competitivo que essa empresa tem e o poder decisório, tendo em vista a análise preditiva dos clientes. Isso tudo é possível com a análise de Big Data. Dessa forma, é possível conseguir *insights* importantes, tais como:

- Entender melhor os clientes:
  - Quem são nossos clientes?
  - Do que eles gostam?
  - Como eles usam nossos produtos?
- Melhorar produtos:
  - É necessário fazer alguma alteração ou promoção nos produtos?
  - Que tipos de mudanças devemos fazer?
  - O que as pessoas mais gostam em nosso produto?
- Planejamento do negócio:
  - Estamos investindo nas coisas certas?
  - Os riscos que assumimos valerão a pena?
- Concorrência:
  - Quais são nossos maiores concorrentes?
  - O que fazer para destacar no nosso setor?

Com o processamento do Big Data, as empresas possuem a capacidade de encontrar respostas às perguntas que desejam conhecimento e, ao mesmo tempo, conseguem entender padrões nos dados que antes das análises não eram possíveis identificar. A seguir, vamos apresentar alguns exemplos de como o Big Data está sendo aplicado em projetos reais.

- Automotivo: Aplicam análises avançadas de dados de motorista, veículo, suprimentos e IoT para melhorar a eficiência de fabricação de peças automotivas, a segurança e a Inteligência artificial para carros com motoristas autônomos;

- Financeiros: Utilizam dados de clientes e transações para reduzir o risco de fraudes, aumentar retorno e melhorar a satisfação de clientes;

- Mídia e entretenimento: Análise de dados de público e conteúdo para aprofundar o envolvimento do público nas programações, reduzir a rotatividade e otimizar as receitas de publicidade;

- Saúde: Utilizam dados para monitoramento de pacientes em tempo real; Análise de padrões de doenças, extração de informação em imagens médicas, descoberta e desenvolvimento de novos medicamentos, análise de dados genéticos.

- Telecomunicações: Utilizam os dados de clientes e da rede para melhorar os serviços e o desempenho da rede, analisar registros de chamadas, alocação de banda em tempo real, planejamento da rede, redução de rotatividade de clientes.

- Varejo: Utilizam dados de clientes e produtos para realizarem análise de sentimento, segmentação de mercado e cliente, marketing personalizado, previsão de demandas.

> Capítulo 2

Capítulo 2. Definição e Conceitos do Big Data

Neste capítulo, vamos abordar as principais definições e conceitos relacionados ao Big Data.

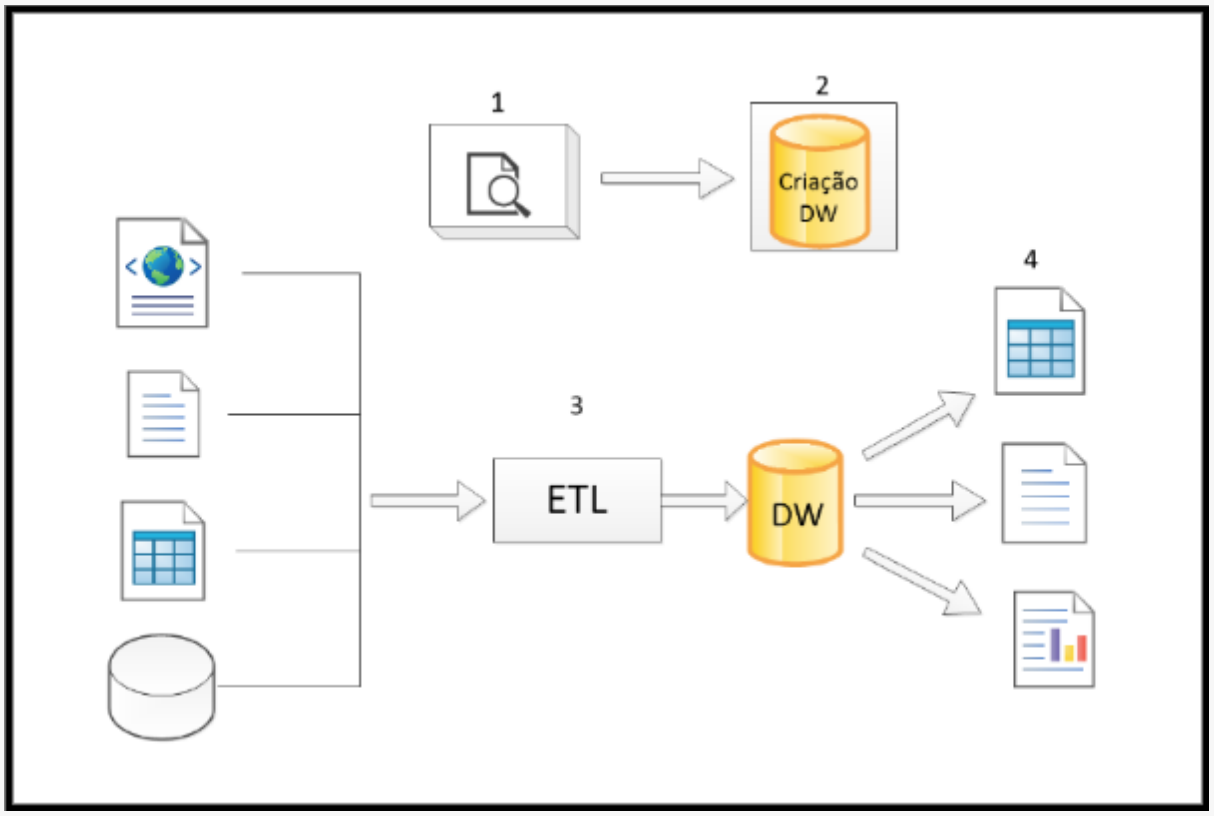
2.1. Business Intelligence (BI)

Atualmente, o grande desafio das empresas líderes de mercado é tomar em tempo hábil decisões baseadas nas coletas de informações essenciais e relevantes disponibilizadas ao mercado. Porém, com a grande quantidade de dados e informações produzidas pelas organizações, analisá-los torna-se cada vez mais complexo, pois muitas vezes essas fontes precisam ser estruturadas, tratadas e entregues com tempo hábil. Por esses fatores surgiu a necessidade de criar plataformas de inteligência de negócios (BI – *Business Intelligence*) (DUANAND XU, 2012).

A terminologia BI refere-se ao processo de coleta, organização, análise, compartilhamento e monitoração de informações que oferecem suporte à gestão de negócios. Segundo Batista (2004), BI é "um conjunto de ferramentas e aplicativos que oferecem aos tomadores de decisão a possibilidade de organizar, analisar, distribuir e agir, ajudando a organização a tomar decisões melhores e mais dinâmicas".

Carlos Barbieri (2001) mostra um conceito mais amplo onde BI pode ser entendido como a utilização de várias fontes de informação para se traçar estratégias de competitividade para negócios da empresa. Um dos seus objetivos está relacionado ao apoio e auxílio aos processos de tomadas de decisões baseados em dados trabalhados especificamente na busca de vantagens competitivas. Para Audy e colaboradores (2005), o BI é um sistema de informação que dá suporte à análise de dados do processo decisório, empregando tecnologias como Data Warehouse, ETL, OLAP e *Data Mining* nos níveis tático e estratégico (AUDY, 2005). A Figura 6 ilustra as etapas da criação de um projeto de BI.

Figura 6 - Etapas da criação de um projeto de BI.



1. Antes de iniciar a criação de um DW, é muito importante conhecer o banco de dados que será utilizado e as tabelas relacionadas. Para isso, é necessário fazer um mapeamento completo e detalhado de todo do modelo, a fim de identificar a fonte dos dados que o alimenta e conhecer cada detalhe. Nessa etapa de mapeamento é feito o levantamento de todos os dados que serão utilizados e exibidos no modelo de DW a ser criado.

2. Após definido os mapeamentos, os bancos de dados, tabelas e diversas fontes de dados a serem utilizadas, o próximo passo é criar o esquema de DW. O esquema mais utilizado é o *Star Schema* (Esquema Estrela) proposto por Dr. Ralph Kimball e que tem como característica básica a presença de dados altamente redundantes para se obter um melhor desempenho (COLAÇO JR., 2004).

3. O terceiro passo da etapa é criar processos de ETL. Esse processo é o mais crítico e demorado na construção de um Data Warehouse, pois consiste na extração dos dados de bases heterogêneas, na transformação e limpeza destes dados, e na carga dos dados na base do DW.

4. O último passo da etapa é a etapa de criação dos relatórios. Essa etapa é o produto final da análise. Aqui podemos criar diversos tipos de relatórios que irão exibir o conhecimento extraído do DW. Os relatórios são criados de acordo com a demanda da instituição.

2.1.1. BI e Big Data

Essas duas estratégias são bastante diferentes e podem se relacionar para gerar um excelente resultado na geração de *insights* e informações estratégicas, aumentando a competitividade da empresa. (SUN; ZOU; STRANG, 2015)

Ao passo que o Big Data trabalha correlacionando um número incontável de dados e gerando informações que talvez ainda não tenham sido pensadas, o BI está coletando, monitorando, filtrando e organizando as informações já conhecidas de forma que o gestor receba esses dados tratados para a tomada de decisões.

Enquanto as soluções de Big Data estão minerando os dados de forma precisa, as ferramentas do BI estão condensando e analisando essas informações para a tomada de decisão.

2.2. Data Warehouse

O Data Warehouse (DW) é uma tecnologia que foi desenvolvida nos anos 80 com o objetivo de simplificar as pesquisas através de um banco de dados organizado e com alta capacidade de armazenamento. Pode ser visto como um grande banco de dados que contém dados históricos relativos às atividades de uma instituição ou organização de forma consolidada. Para

Barbieri (2001), um Data Warehouse (DW) é um banco de dados histórico, separado em estruturas lógicas dimensionais, concebido para armazenar dados extraídos dos sistemas legados e ERP da empresa. Segundo Colaço (2004), antes de serem armazenados no DW, os dados são selecionados, organizados e integrados para que possam ser acessados de forma mais eficiente, auxiliando assim o processo de tomada de decisão.

A tradução literal do termo Data Warehouse seria "armazém de dados", já que ele foi pensado para agir exatamente dessa forma, fornecendo ao usuário um agrupamento de dados estruturados e produzindo relatórios que facilitam a tomada de decisões (INMON, 1995).

Os benefícios do Data Warehouse são:

- Vantagem competitiva;
- Simplicidade nas informações;
- Facilidade de uso;
- Infraestrutura computacional;
- Custo de operação;
- Acesso rápido;
- Qualidade dos dados.

Os desafios com o Data Warehouse hoje são:

- Perda de material valioso pela não utilização dos dados não estruturados;
- A complexidade do desenvolvimento;
- O alto custo para o desenvolvimento;

- Necessidade de administração e treinamento.

2.2.1. Data Mart

Data Mart é um banco de dados que representa um segmento (assunto) de um Data Warehouse. Pode ser representado como um subconjunto de dados dentro do conjunto do Data Warehouse, normalmente se identifica com um setor (departamento) específico do negócio. Os Data Warehouses são criados para servir como o armazenamento central de dados para toda a empresa, enquanto um Data Mart atende à solicitação de uma divisão ou função comercial específica. O Data Mart por ser um conjunto menor de dados, oferece acesso mais fácil aos dados do setor (departamento) em questão. A Tabela 1 ilustra algumas diferenças entre Data Warehouse e Data Mart.

Tabela 1 - Diferenças entre Data Warehouse e Data Mart.

Data Warehouse	Data Mart
Armazena dados de várias áreas de assunto.	Um data mart carrega dados relacionados a um departamento, como RH, marketing e finanças, etc.
Ele atua como um repositório central de dados para uma empresa.	É uma subseção lógica de um DW no qual os dados são depositados em servidores baratos para aplicativos departamentais específicos.
É complicado projetar e usar um data warehouse porque geralmente inclui uma grande quantidade de dados, mais de 100 GB.	Projetar e usar um data mart é comparativamente mais fácil devido ao seu tamanho pequeno (menos de 100 GB).
Projetado para suportar o processo de tomada de decisão de toda a empresa	Desenvolvido para grupos de usuários ou departamentos corporativos específicos. Assim, oferece interpretação departamental e armazenamento descentralizado de dados.
Tem grandes dimensões e integra dados de um grande número de fontes, o que pode causar risco de falha.	Tem dimensões menores e os dados são integrados a partir de um número menor de fontes, portanto, há menos risco de falha.
É orientado ao assunto e variante no tempo, no qual os dados existem por um período mais longo.	Destina-se a áreas específicas relacionadas a uma empresa e retém dados por um período mais curto.

2.3. ETL

O ETL (*Extract, Transformation and Load*) são ferramentas responsáveis pela extração, transformação e carregamento dos dados no DW. Em um projeto de construção de um DW, os processos ETL consomem

mais de 70% do tempo de desenvolvimento. Todo esse processo é devido à diversidade existente em termos e estrutura de dados nas bases de dados de origem (COLAÇO JR., 2004).

Por meio do ETL, é possível definir a qualidade dos dados e a forma como eles são manipulados, a fim de transformá-los em uma informação inteligível e confiável.

Na primeira etapa, é realizada a extração dos dados de todas as fontes relevantes. As fontes podem incluir banco de dados local, sistemas de CRM, Data Warehouse, planilhas eletrônicas, entre outras fontes de dados que possam fornecer insights. Como os dados provenientes de várias fontes têm um esquema diferente, cada conjunto deve ser transformado de maneira diferente antes de serem utilizadas análises de BI.

A segunda etapa é realizada a transformação dos dados. Nela é realizado todo o tratamento nos dados coletados. Tais como limpeza, correção, compilação, conversão de tipos. Após isso, os dados estão prontos para serem armazenados no repositório de dados.

Na última etapa é realizado o carregamento dos dados. Nessa etapa, os dados que foram extraídos e transformados na etapa anterior são armazenados nos repositórios de dados destino.

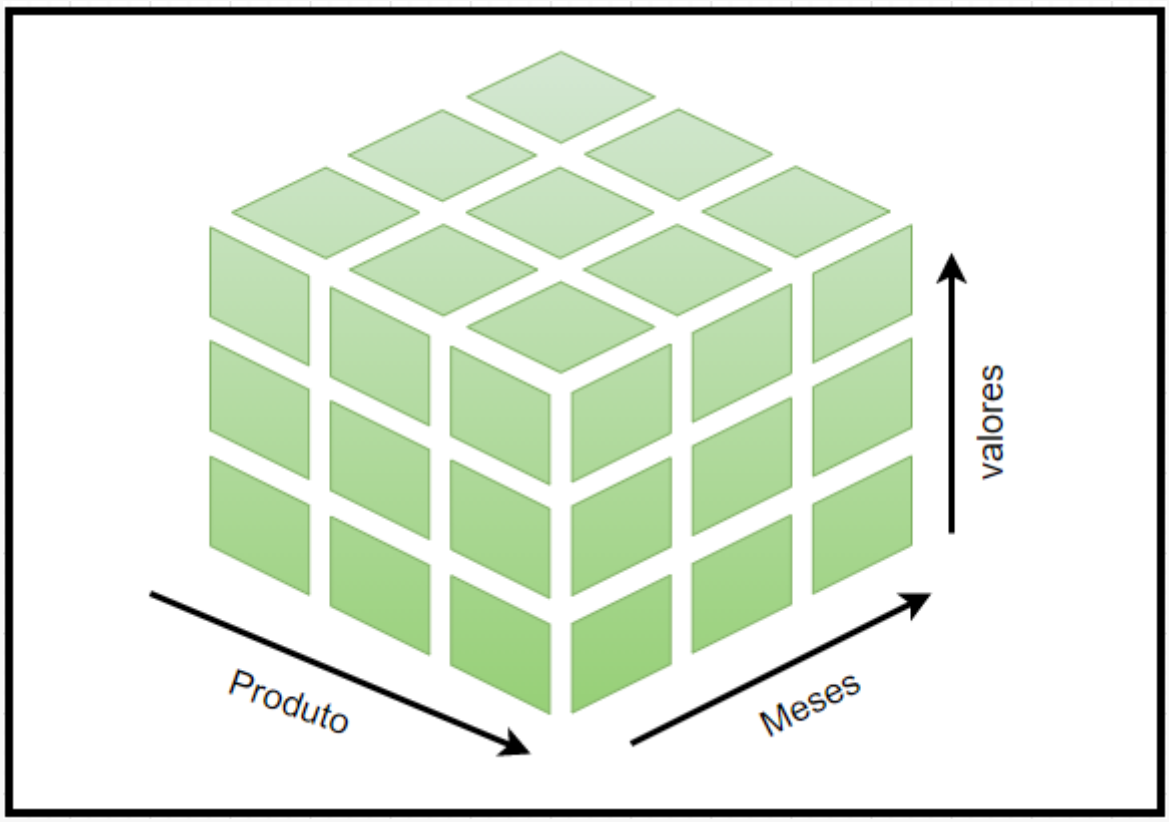
2.4. OLAP

O OLAP (*Online Analytical Processing*) é um conjunto de ferramentas e técnicas que permite realizar a exploração dos dados de um DW, utilizando os recursos de modelagem, visualização e análise de grandes conjuntos de dados. O OLAP ajuda a analisar de forma eficiente a quantidade de dados armazenados pelas organizações transformando-os em informação (JACOBSON; MISNER, 2007).

Segundo Barbieri (2001), o termo OLAP surgiu em 1993, caracterizando o conjunto de técnicas para tratar informações que serão

armazenadas em um DW. É um sistema que busca informações sumarizadas, e permite que essas sumarizações sejam apresentadas como suporte nas funções de derivação de dados complexos (COLAÇO, 2004). O OLAP fornece para organizações um método de acessar, visualizar e analisar dados corporativos com alta flexibilidade e performance. Para extrair informações válidas do DW, utilizamos técnicas de mineração de dados (DM – Data Mining). A Figura 7 ilustra a forma de representação de um cubo OLAP.

Figura 7 - Representação do Cubo OLAP.



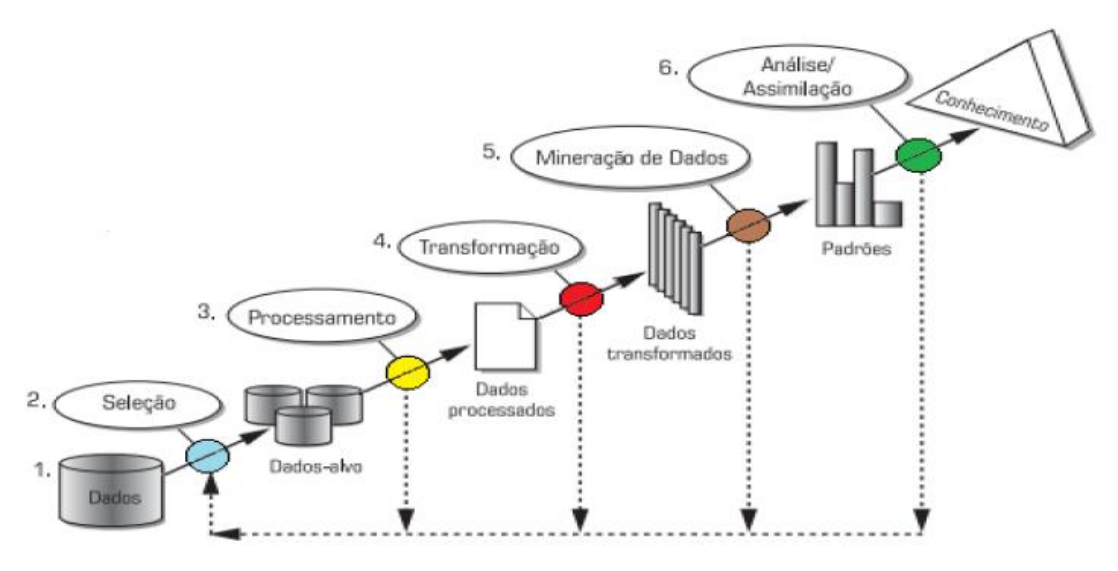
Podemos observar que através do cubo OLAP é possível ver diversas representações do dado. Desse modo, o usuário que estiver analisando o dado pode trabalhar a informação da forma que desejar. Assim, facilitando o entendimento e a exploração dos dados. Por exemplo, é possível analisar o faturamento de um determinado produto em um mês. Além disso, podemos realizar análises como média de preços, produtos mais vendidos, meses com maior faturamento etc.

2.5. KDD

O processo de descoberta de conhecimento é ilustrado pela Figura 8 que envolve diversas fases do KDD (*Knowledge Discovery Databases*) proposta por Fayyad et al (1996). O objetivo é extrair de grandes bases de dados, informações válidas e acionáveis, desconhecidas, úteis para tomada de decisão. De uma forma breve, o processo envolve três etapas iniciais: Seleção, pré-processamento e transformação, as quais compõem o que é

denominado de preparação dos dados. Após, vem a fase de mineração de dados e por fim, o conhecimento gerado deverá ser analisado, o que acontece na etapa de análise e assimilação dos resultados (COLAÇO JR., 2004).

Figura 8 - Etapas do processamento do Big Data.



Fonte: Colaço Jr. 2004.

- Seleção de dados: São identificadas as bases de dados a serem utilizadas nas descobertas de conhecimento, leva-se em consideração os objetivos do processo.

- Pré-processamento de dados: Como a informação pode vir de diversas bases distintas, podem surgir problemas de integração entre os dados. Isso deve ser resolvido nessa etapa. Por exemplo: suponha que que a informação sobre o sexo dos clientes de uma loja esteja armazenada em um banco como “M” e “F” e em outra como “H” e “M”. Neste caso o pré-processamento é feito corrigindo e atualizando os dados.

- Transformação de dados: O objetivo é transformar os dados já pré-processados, de modo a torná-los compatíveis com as entradas de diversos algoritmos de mineração existentes.

- Mineração de dados: Caracterizada pela escolha e aplicação do algoritmo e da técnica de mineração.

- Análise e assimilação dos resultados: O conhecimento gerado deve ser analisado nesta etapa. Na maneira a verificar se é realmente útil para a tomada de decisão. Caso a resposta não seja satisfatória, então deve-se repetir todo ou parte do processo de KDD.

2.6. Data Mining

O termo mineração de dados vem do inglês (*Data Mining* - DM) que tem o objetivo principal um processo mais amplo denominado descoberta de conhecimento em base de dados. A mineração de dados consiste em utilizar dados de estatísticas e de inteligência artificial bem estabelecidas que constroem modelos que predizem os padrões relevantes em um banco de dados. O DM identifica e interpreta padrões de dados que serão utilizados pelos gestores na tomada de decisão (COLAÇO JR., 2004). Embora KDD e Data Mining sejam frequentemente entendidos como sinônimos, é importante frisar que, enquanto o KDD compreende todas as etapas para a descoberta do conhecimento a partir da existência de dados, a mineração de dados é apenas e tão somente uma das etapas do processo.

A relação do DM com o Big Data é o desempenho de funções parecidas, porém o Big Data a realiza com uma base de dados muito maior. Essa diferença se observa ainda mais perante os resultados. O DM aponta relatórios com questões específicas já que se refere a uma questão mais pontual. O Big Data traz análises contínuas e de longo prazo e por isso é utilizado para traçar planos e previsões que cooperem com a estratégia organizacional.

2.7. Data Lake

Ao contrário dos Data Warehouses, que geralmente aceitam dados limpos, os Data Lakes armazenam dados em seu formato bruto. Os Data Lakes podem armazenar dados não estruturados e estruturados e são conhecidos por serem mais escalonáveis horizontalmente (em outras palavras, é fácil adicionar mais dados aos Data Lakes).

Um Data Lake consiste em duas partes: armazenamento e processamento. O armazenamento requer um repositório de armazenamento infinitamente escalável e tolerante a falhas projetado para lidar com grandes volumes de dados com formas, tamanhos e velocidades de ingestão variados. O processamento requer um mecanismo de processamento que possa operar com êxito os dados nessa escala (TEJADA, 2017).

O Data Lake foi desenvolvido para possuir vários mecanismos de processamento e ser o repositório centralizado de todos os dados gerados e coletados de toda a empresa. No Data Lake pode armazenar uma infinidade de dados, desde dados estruturados ou semiestruturados a dados completamente não estruturados. É possível armazenar com segurança qualquer tipo de dados, independentemente do volume ou formato, com capacidade ilimitada de escalonamento e fornece uma maneira mais rápida de analisar conjuntos de dados do que os métodos tradicionais (SINGH; AHMAD, 2016).

Os benefícios do Data Lake são:

- Capacidade de conter vários tipos de dados;
- São mais fácies de escalonar já que são baseados em nuvem;
- O armazenamento em nuvem o torna mais barato, permitindo que as empresas capturem e mantenham todos os dados

ainda que não saibam o que fazer com eles no momento da coleta.

Os desafios do Data Lake hoje são:

- A falta de treinamento da equipe pode gerar dificuldade em navegar pelo Data Lake;
- A velocidade da consulta pode ser afetada ocasionalmente pela falta de estrutura e o grande volume dos dados.

2.8. Bônus: Dicas de Boas Práticas

É muito comum quando estamos trabalhando com grande volume de dados, nos perder no meio de tantas informações. Pensando nisso, é importante termos uma organização dos dados. Existem diversas formas de padronizar os dados. Por exemplo, por tipo de sistema, assunto, tipo de tratamento realizado no dado, entre outros. Não importa o modo que é feito a padronização. Cada empresa pode possuir uma padronização que achar mais conveniente para o seu negócio. O importante é possuir algum padrão para que as pessoas ao verem um dado armazenado possam facilmente identificar do se trata, de onde ele vem e o que ele significa.

As Figuras 9, 10 e 11 ilustram algumas dicas de boas práticas para padronização de dados.

Figura 9 - Nomenclatura Padrão de Databases.

ORIGEM DOS DADOS	NOMENCLATURA INICIAL
API	api_<assunto>
BANCOS DE DADOS	db_<nome database>
ARQUIVOS FLAT FILE	ff_<assunto>

DESTINO DOS DADOS	NOME DO DATABASE
STAGE (ETL)	stage
MODELO DIMENSIONAL	dw
TABELAS FINAIS (NÃO DIMENSIONAIS)	flat_table
ÁREA TEMPORÁRIA	work
SANDBOX	sb_<squad ou área>

Figura 10 - Nomenclatura Padrão de Objetos.

OBJETO	NOMENCLATURA INICIAL
TABELA	TB_
VIEW	VW_
VIEW MATERIALIZADA	VM_
ÍNDICE	IDX_
PROCEDURE	PR_
FUNÇÃO	FN_
PACKAGE	PKG_
TRIGGER	TG_
SEQUENCE	SQ_
TABELA DIMENSÃO	DM_
TABELA FATO	FT_
TABELA AGREGADA	AGG_

Figura 11 - Nomenclatura Padrão de Colunas.

TIPO DE COLUNA	DATATYPE	NOMENCLATURA INICIAL
SURROGATE	Numérico	ID_
CÓDIGO SISTEMA ORIGEM	Numérico ou String	CD_
NOME	String	NM_
DESCRIÇÃO	String	DS_
DATA	Data	DT_
DATA E HORA	Data e hora	DT_
HORA	String	HR_
VALOR	Numérico	VL_
QUANTIDADE	Numérico	QT_
FLAG	Numérico	FL_
NÚMERO	Numérico ou String	NR_
HASH	Numérico ou String	CD_HASH_

> Capítulo 3

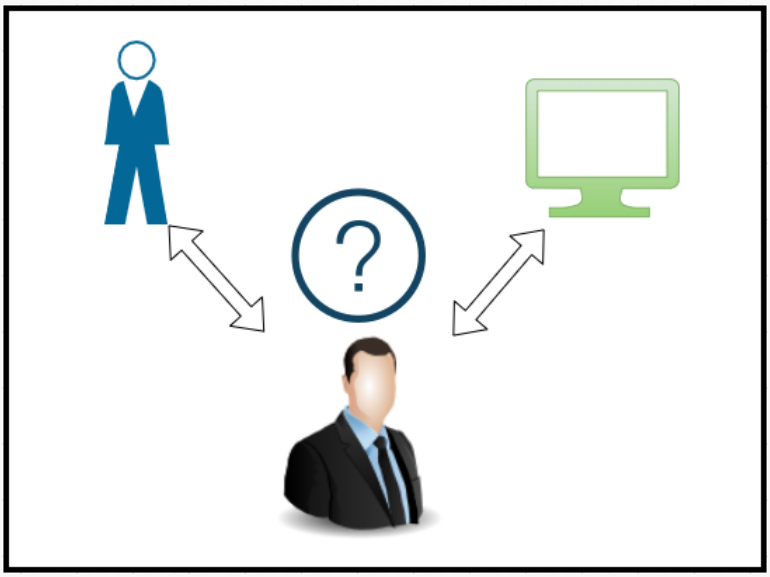
Capítulo 3. Técnicas para Trabalhar com Big Data

Após realizar a coleta e o processamento dos dados, é necessário utilizar técnicas de Big Data para extrair *insights*. Neste capítulo, vamos abordar as técnicas utilizadas para o processamento do Big Data.

3.1. Inteligência Artificial

A Inteligência Artificial (IA) é uma parte da ciência da computação onde os sistemas são planejados para a execução de tarefas que necessitam de inteligência humana e possui várias técnicas para realização da atividade. Para entendimento do comportamento de uma IA, vamos tomar como exemplo o teste de Turing. O teste de Turing testa a capacidade de uma máquina apresentar um comportamento inteligente equiparado a um ser humano. Podemos observar na Figura 12 um exemplo prático ilustrativo de um teste de Turing.

Figura 12 - Teste de Turing.



1. Um avaliador humano faz uma série de perguntas baseadas em texto para uma máquina e um humano, sem ver nenhum deles;

2. O humano e a máquina respondem à pergunta do avaliador;

3. Se o avaliador não conseguir diferenciar entre as respostas dadas foi de um humano ou de máquina, o computador é aprovado no teste de Turing. Isso significa que exibiu um comportamento semelhante ao humano ou inteligência artificial.

O Big Data é umas das principais fontes para a inteligência artificial, gerando dados para que cada dia mais ela crie sistemas que possuem a capacidade de auxiliar na vida cotidiana. Ao contrário do que se imaginava, a inteligência artificial não resultou em vários robôs realizando atividades humanas, mas tem operado de forma eficiente e um tanto silenciosa.

Com a grande coleta de dados do Big Data, é possível a criação de modelos que analisam e antecipam comportamentos e dinâmicas de sistemas complexos. Esses dados provêm não apenas da interação dos indivíduos na rede, mas também pelo rastro que eles deixam na internet sem saber.

O volume desses dados produz uma característica muito importante no processo da IA: descobrir quais dados são relevantes para a análise e utilizar ferramentas que são capazes de manipular e estudar essa quantidade exorbitante de dados. Entender esses elementos, suas origens e projetar as condições futuras permite um melhor planejamento estratégico.

Esse é o ponto onde a Inteligência Artificial entra. A quantidade dos dados somada a necessidade da análise de cada um pode ser um processo automatizado através de *Machine Learning* constante, que significa ter uma máquina capaz de aprender certa informação. O aprendizado de máquina utiliza códigos para fazer uma varredura em grandes quantidades de dados buscando padrões. Elevando a quantidade de vezes em que a máquina reproduz esse comportamento, ela será capaz de analisar grandes

quantidades muito mais rapidamente se comparado ao processo humano manual.

Aprender é uma atividade inerente ao ser humano, e quando há a tentativa e erro, produz-se outros resultados que podem auxiliar futuramente. O *Machine Learning* segue esse mesmo princípio, o que permite que os resultados se tornem sempre mais assertivos e específicos.

Se tratando de máquinas, o Big Data vai fornecer exatamente aquilo que é necessário para o bom desenvolvimento de seu aprendizado: dados não estruturados e contínuos. Isso replica a forma intuitiva do ser humano de produzir novos conhecimentos.

3.2. Machine Learning

O aprendizado de máquina (ML) é um subconjunto de inteligência artificial que funciona muito bem com dados estruturados. O objetivo por trás do aprendizado de máquina é que as máquinas aprendam padrões em seus dados sem que você os programe explicitamente para isso. Atualmente, o aprendizado de máquina não pode fornecer o tipo de IA que os filmes apresentam. Mesmo os melhores algoritmos não conseguem pensar, sentir, apresentar qualquer forma de autoconsciência ou exercer o livre arbítrio. O que o aprendizado de máquina pode fazer é realizar análises preditivas com muito mais rapidez do que qualquer ser humano. Como resultado, o ele pode ajudar os humanos a trabalhar com mais eficiência. O estado atual da IA, então, é o de fazer análise, mas os humanos ainda devem considerar as implicações dessa análise - tomando as decisões morais e éticas necessárias (MUELLER; MASSARON, 2016).

Existem 3 tipos de aprendizado de máquina. O aprendizado supervisionado, o não supervisionado e o por reforço. O aprendizado supervisionado é baseado na regressão básica e classificação. O humano fornece um banco de dados e ensina a máquina a reconhecer padrões e semelhanças através de rótulos. Por exemplo, a máquina pode reconhecer

um carro. A cor, tamanho e outras características podem variar. No entanto, ela aprende elementos chaves que identificam um carro.

Já no aprendizado não supervisionado, a máquina aprende com dados de teste que não foram rotulados, classificados ou categorizados previamente. Dessa forma, não existe supervisão humana. O aprendizado não supervisionado identifica semelhanças nos dados e reage com base na ausência ou presença das semelhanças em cada novo dado. A clusterização, que é uma técnica de aprendizado não supervisionado que permite dividir automaticamente o conjunto de dados em grupos de acordo com uma similaridade.

Aprendizado por reforço é o aprendizado baseado na experiência que a máquina tem e aprende a lidar com o que errou antes e procurar a abordagem correta. Podemos comparar o aprendizado por reforço de uma criança. Por exemplo, quando uma criança começa a engatinhar, ela tenta se levantar e cai várias vezes, e após muitas tentativas ela consegue uma forma de se levantar sem cair. Um outro exemplo são as recomendações de sites de entretenimento como o YouTube. Após assistir um vídeo, a plataforma irá te mostrar títulos semelhantes que acredita que você também irá gostar. No entanto, se você começa a assistir o recomendado e não o termina, a máquina entende que a recomendação não foi boa e irá tentar outra abordagem da próxima vez.

Para exemplificar ainda mais, vamos ilustrar um exemplo de aprendizado supervisionado para detecção de fraudes. De maneira geral e de alto nível, temos:

1. O analista cria regras para o que constitui fraude (por exemplo, uma conta com mais de 30 transações no mês, compras em diversos setores, saldo médio menor que R$200,00);

2. Essas regras são passadas para o algoritmo que recebe os dados que são rotulados como "fraude" ou "não fraude". Após isso, a máquina aprende o comportamento dos dados fraudulentos.

3. Com o apoio das regras, a máquina começa a prever as fraudes;

4. Ao final, é feita a validação do modelo previsto da máquina. Para isso, um analista investiga e verifica manualmente se as previsões do modelo preveem a fraude.

3.3. Deep Learning

Deep Learning (DL) ou Aprendizado profundo é um subconjunto do aprendizado de máquina que usa conjuntos de algoritmos modelados para simular computacionalmente o cérebro humano (também chamados de redes neurais artificiais). Esses algoritmos são muito mais complexos do que a maioria dos modelos de aprendizado de máquina e exigem muito mais tempo e esforço para serem construídos. Ao contrário do aprendizado de máquina, que se estabiliza após uma certa quantidade de dados, o aprendizado profundo continua a melhorar conforme o tamanho dos dados aumenta. Os algoritmos de aprendizado profundo possuem maior desempenho em conjuntos de dados complexos, como imagens, sequências e linguagem natural.

O aprendizado profundo é muito usado para tarefas que classificam imagens. Por exemplo, digamos que você queira construir um modelo de aprendizado de máquina para classificar se uma imagem contém um cachorro. Para isso, o algoritmo recebe como entrada centenas, milhares ou milhões de fotos. Algumas dessas fotos mostram cachorros e outras não. Com o tempo, o modelo aprende o que é e o que não é um cachorro. E com o passar do tempo o modelo vai poder identificar um cachorro com mais facilidade e rapidez em relação a outras imagens. É importante observar que, embora seja uma tarefa fácil para os humanos reconhecer um cachorro pelas suas características, uma máquina detectará coisas que não podemos -

coisas como padrões no pelo do cachorro ou o formato exato de seus olhos. Com isso, é capaz de tomar decisões rapidamente com base nessas informações.

Abaixo segue algumas das aplicações do *Deep Learning*:

- Processamento de Imagem;
- Processamento de Linguagem Natural;
- Previsão do mercado de ações;
- Veículos autônomos;
- Aplicações em Medicina:
  - Aplicações de Reconhecimento de Imagens;
  - Detecção de Câncer de Mama;
  - Doença de Alzheimer;
  - Diagnóstico Cardiovascular;
  - Detecção de câncer de pele;
  - Derrame Cerebral.

> Capítulo 4

Capítulo 4. Algoritmos utilizados no Big Data

Neste capítulo, vamos apresentar os principais algoritmos utilizados no processamento do Big Data.

4.1. Classificação dos algoritmos de *Machine Learning*

Como vimos na seção 3.2 do capítulo 3, os algoritmos de *Machine Learning* podem ser classificados em:

1. Aprendizado supervisionado;
2. Aprendizado por não supervisionado;
3. Aprendizado por reforço.

Neste capítulo, vamos apresentar os principais algoritmos para cada classificação.

Para os algoritmos de aprendizado supervisionados, temos basicamente duas classes: Classificação e Regressão.

Para os algoritmos de classificação, o objetivo é identificar a qual categoria pertence uma amostra do problema. Por exemplo, podemos classificar se uma transação é uma fraude ou não; se um e-mail é SPAM ou não; se uma mensagem em rede social possui sentimento positivo, negativo ou neutro; entre outros. Os principais algoritmos são árvores de decisão, *Naive Bayes*, redes neurais

Já para os algoritmos de regressão, a ideia é prever um valor de uma variável com base no valor de outra. Para isso, o modelo pode aprender uma função que prevê o preço das ações de um fundo imobiliário, uma demanda de venda, o tempo de desgaste de pneus em uma frota de carros, tempo de baixa de estoque de peças ou qualquer outro valor quantitativo. Os

algoritmos mais utilizados são regressão linear, regressão logística e as redes neurais que podem apresentar resultados com valores contínuos.

Os algoritmos de aprendizado não supervisionado podem ser divididos em duas classes: Associação e Clusterização.

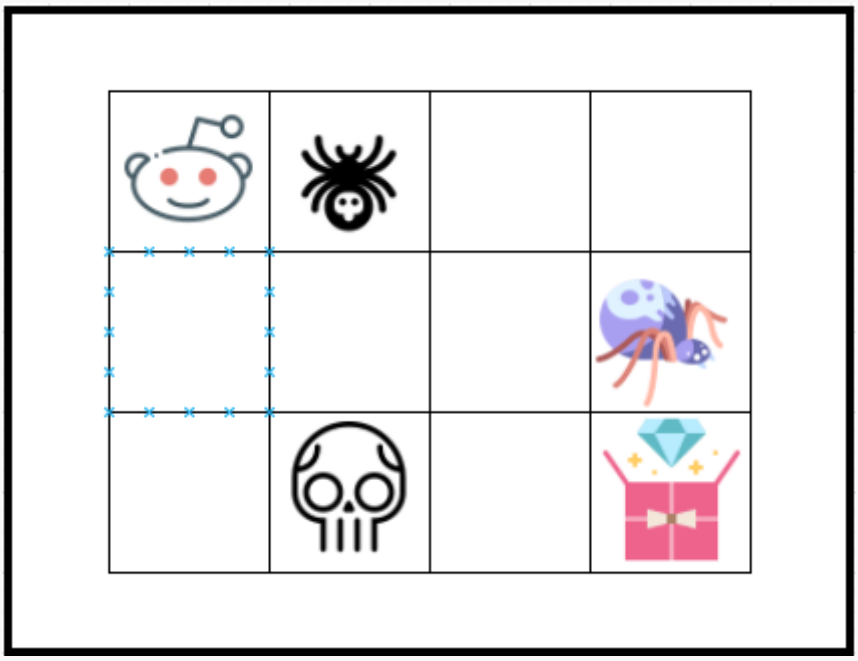
Os algoritmos de associação permitem o descobrimento de regras e correlação em uma base de dados, identificando conjuntos de itens que ocorrem juntos dentro de uma determinada frequência. Muito utilizado no setor de varejo, no qual os varejistas utilizam para analisar carrinhos de compras. Assim, descobrindo os itens frequentemente mais comprados em conjunto. Desta forma, criando novas estratégias de *marketing* e vendas. O Algoritmo mais utilizado nessa classe é o algoritmo de regras de associação.

Os algoritmos de clusterização ou agrupamento permitem que seja feito agrupamento de grupos com base nas semelhanças encontradas. É uma técnica que permite realizar a divisão de grupos em um conjunto de dados de forma automática, baseado em medidas de similaridade ou de distância. Existem vários métodos que permitem obter medidas de similaridade, podemos citar a similaridade de cosseno e a correlação de Pearson. Podemos citar os algoritmos baseado em particionamento, baseado em densidade, o hierárquico aglomerativo e hierárquico divisório

Os algoritmos de aprendizado por reforço permitem que o modelo aprenda executando ações e avaliando o resultado dessas ações.

Esse tipo de algoritmo é geralmente aplicado quando se conhece as regras, mas desconhece a melhor sequência de ações que devem ser executadas. Os algoritmos aprendem de forma interativa. A Figura 13 ilustra um exemplo de problema no qual queremos encontrar o melhor caminho para alcançar o diamante (recompensa).

Figura 13 - Exemplo algoritmo de aprendizado por reforço.



O agente deve encontrar o melhor caminho possível para alcançar a recompensa e quando encontrar um obstáculo, deve ser penalizado (pois ele deve escolher o caminho sem obstáculos). Com a Aprendizagem por reforço, podemos treinar o agente para encontrar o melhor caminho. Os algoritmos de aprendizado por reforço são muito aplicados no campo de estudo da robótica. Podemos citar os seguintes algoritmos: *Multi-Armed Bandits, Contextual Bandits* e *k-Armed Bandits*.

4.2. Aprendizado não supervisionado K-means

K-means é um dos algoritmos mais utilizados para realizar agrupamentos de dados numéricos em mineração de dados (DESAI et al. 2016), (KANUNGO et al. 2002). O objetivo do K-means é encontrar a melhor divisão de p dados em k grupos, de forma que a distância total entre os dados de um grupo e seu centroide, somados por todos os grupos, seja minimizada. Em sua forma mais comum descrita em Lloyd (1982) o algoritmo pode ser dividido em 4 principais passos:

1. Inserção do parâmetro de entrada com número de k grupos;

2. Definição do centroide;

3. Definição dos grupos em relação ao centroide

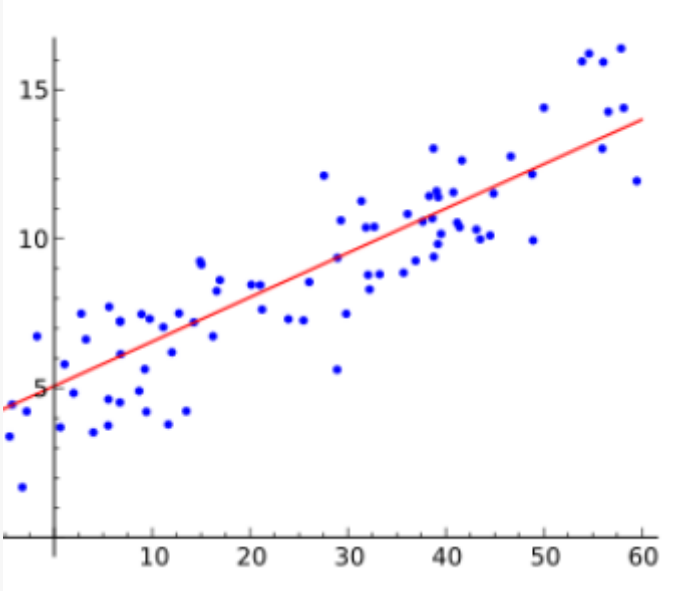
4. Associação dos elementos nos grupos.

A execução do K-means inicia com o recebimento do parâmetro de entrada que define o número de k grupos. Após isso, são selecionados k pontos aleatoriamente para representar os centroides de cada grupo. Em seguida, os elementos restantes são associados aos centroides mais próximos, essa associação é realizada através da distância euclidiana simples entre um elemento e o centroide. A cada novo passo, o algoritmo calcula novamente a média e define o novo centroide, realocando novamente os elementos nos grupos de maior similaridade. Esse processo de recalcular as médias dos grupos e realocar os elementos é repetido até que o critério de parada seja satisfeito.

4.3. Aprendizado supervisionado Regressão Linear

Regressão linear é um algoritmo de *Machine Learning* para aprendizado supervisionado. A análise de regressão linear é usada para prever o valor de uma variável com base no valor de outras variáveis. A variável cujo valor você deseja prever é chamada de variável *dependente* e a variável que é usada para prever o valor de outra variável é chamada de variável *independente*. A regressão, em geral, tem como objetivo tratar de um valor que não se consegue estimar inicialmente – por exemplo, qual o valor de um imóvel dado características como metragem, localização e estado de conservação. Ela permite gerar um modelo matemático através de uma reta que explique a relação linear entre variáveis. A regressão linear é chamada "linear" porque se considera que a relação da resposta às variáveis é uma função linear de alguns parâmetros. Essa forma de análise estima os coeficientes da equação linear, envolvendo uma ou mais variáveis independentes que melhor preveem o valor da variável dependente. A Imagem 14 ilustra um modelo de regressão linear.

Figura 14 - Modelo de regressão linear.



Fonte: Imagem ilustrativa

No caso mais simples, teremos a relação entre uma variável explicativa X e uma variável resposta Y. A equação que representa a relação entre duas variáveis pode ser vista abaixo.

$$y_i = \alpha + \beta X_i + \varepsilon_i$$

Onde:

yi: Variável explicada (dependente); representa o que o modelo tentará prever;

a: É uma constante, que representa a interceptação da reta com o eixo vertical;

β: Representa a inclinação (coeficiente angular) em relação à variável explicativa;

Xi: Variável explicativa (independente);

ε: Representa todos os fatores residuais mais os possíveis erros de medição. É um termo erro aleatório com média μ zero e variância σ2 constante.

4.3.1. Correlação Linear

Em pesquisas, frequentemente, procura-se verificar se existe relação entre duas ou mais variáveis, isto é, saber se as alterações sofridas por uma das variáveis são acompanhadas por alterações nas outras. Por exemplo, qual a relação entre o tempo de trabalho da pessoa e seu salário ou o consumo de bebidas com a renda mensal. A correlação permite verificar se duas variáveis independentes estão associadas uma com a outra. O termo correlação significa relação em dois sentidos (co + relação), e é usado em estatística para designar a força que mantém unidos dois conjuntos de valores. A verificação da existência e do grau de relação entre as variáveis é o objeto de estudo da correlação.

Existem graus de força entre a relação de duas variáveis. Para calcular essa força podemos utilizar o coeficiente de correlação de Pearson. Os graus de correlação vão de -1 a 1. A correlação pode ser positiva ou negativa. Se o valor for mais próximo de -1, dizemos que existe uma correlação forte negativa (quando uma variável aumenta a outra diminui). Por exemplo, o preço do dólar aumenta, diminui as compras de produtos internacionais.

Se o valor for mais próximo de 1, dizemos que existe uma correlação forte positiva (quando uma variável aumenta a outra também aumenta). Por exemplo, promoção no preço da carne, aumenta o volume de compra.

4.3.2. MSE - Mean squared error

Em estatística, o erro quadrático médio (MSE) ou desvio quadrático médio (MSD) de um estimador que mede a média dos quadrados dos erros é a diferença média quadrática entre os valores estimados e o valor real.

O MSE é uma medida da qualidade de um estimador. Como é derivado do quadrado da distância euclidiana, é sempre um valor positivo com o erro diminuindo à medida que o erro se aproxima de zero. A equação do MSE e dada por:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

Onde:

- n: número total de registros;
- Yi: valor real;
- ~Yi: valor previsto.

Na análise de regressão, a plotagem é uma maneira mais natural de visualizar a tendência geral de todos os dados. A média da distância de cada ponto até o modelo de regressão previsto pode ser calculada e mostrada como o erro quadrático médio. O quadrado é fundamental para reduzir a complexidade com sinais negativos. Para minimizar o MSE, o modelo poderia ser mais preciso, o que significaria que o modelo está mais próximo dos dados reais. Um exemplo de regressão linear usando esse método é o método dos mínimos quadrados - que avalia a adequação do modelo de regressão linear para modelar conjunto de dados bivariados, mas cuja limitação está relacionada à distribuição conhecida dos dados.

4.3.3. RMSE - root mean squared error

RMSE (root mean squared error) é a medida que calcula "a raiz quadrática média" dos erros entre valores observados (reais) e predições (hipóteses). Uma característica do RMSE é que os erros (reais - predições) são elevados ao quadrado antes de ter a média calculada. Portanto, pesos diferentes serão atribuídos à soma e, conforme os valores de erros das instâncias aumentam, o índice do RMSE aumenta consideravelmente. Ou seja, se houver um *outlier* no conjunto de dados, seu peso será maior para o cálculo do RMSE e, por conseguinte, prejudicará sua métrica deixando-a maior. A equação do RMSE e dada por:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Onde:

n: número total de registros;

Yj: valor real;

^Yi: valor previsto.

4.3.4. Erro Absoluto (MAE - mean absolut error)

MAE (*mean absolut erro*r) calcula o "erro absoluto médio" dos erros entre valores observados (reais) e predições (hipóteses).

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Onde:

n: número total de registros;

Yj: valor real;

^Yi: valor previsto.

> Capítulo 5

Capítulo 5. Computação Distribuída

Para processar Big Data, as organizações usam computação distribuída. Neste capítulo vamos abordar os fundamentos da computação distribuída.

5.1. Introdução a computação distribuída

Antigamente, por volta da década de 90, os sistemas computacionais eram escalados de forma vertical. Isso significa que, quando uma máquina não atendia às demandas e aos requisitos de uma aplicação, melhorava-se o seu hardware instalando memória RAM, CPU, disco rígido. No entanto, quando surgiu a popularização da internet, os requisitos das aplicações e o volume de dados produzidos aumentaram em uma escala mais elevada do que o modelo de escalonamento vertical poderia suportar. Para suprir essas limitações do modelo atual utilizado, a *Google* apresentou dois artigos que propuseram um novo paradigma baseado na escala horizontal de recursos. Surgiu, então, em 2003, o *Google File System* e, em 2006, o *Big Table*.

De maneira geral, se um sistema computacional chegava ao seu limite máximo de desempenho, era adicionado uma nova máquina que trabalhava em conjunto de forma paralela. As máquinas eram conectadas pela rede e compartilhavam seus recursos, a fim de suprir a necessidade da aplicação. As máquinas eram adicionadas de acordo com a necessidade e a demanda das aplicações. Ou seja, se uma aplicação precisasse de poder computacional mais elevado, era alocada a quantidade de máquinas necessárias para suprir aquela demanda. Assim, foram criados os primeiros clusters de computadores. Em 2006, a *Yahoo* fez uma engenharia reversa do software descrito pelo google e disponibilizou para a comunidade o primeiro *software* open source (*Apache Hadoop*) para essa nova categoria de sistemas. Desde então, diversas novas ferramentas foram criadas: *Apache*

HBase (*Powerset*, hoje, *Microsoft*) em 2008, *Apache Hive* (*Facebook*) e *Apache Spark* (*MIT*) em 2010, *Apache Kafka* (*LinkedIn*) em 2011, *Apache Airflow* (*Airbnb*) em 2015 etc. (PEREZ, 2021).

Podemos dizer que um sistema distribuído é um conjunto de computadores conectados em rede, que são coordenados por uma ou várias máquinas administradoras que utilizam softwares que permitem o compartilhamento de seus recursos para um único propósito. As aplicações são diversas, como quebrar um código, descobrir uma nova solução de um problema, criptografar dados sigilosos.

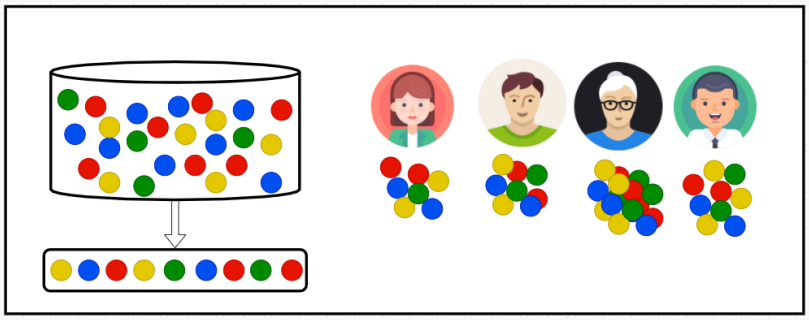
Ou seja, são vários computadores que se comportam como um só para fazer uma coisa de uma maneira rápida e eficiente, ou ter muito espaço de armazenamento. Os sistemas distribuídos são muito utilizados em "supercomputadores" que utilizam os recursos compartilhados para realizar predições climáticas, controle de doenças e epidemias, pesquisas científicas, simulações, entre outros.

5.1.2. Computação distribuída no Big Data

As aplicações Big Data fazem uso da computação distribuída como meio para criar soluções capazes de analisar volumes de dados, processar seus cálculos, identificar comportamentos e disponibilizar serviços especializados em seus domínios.

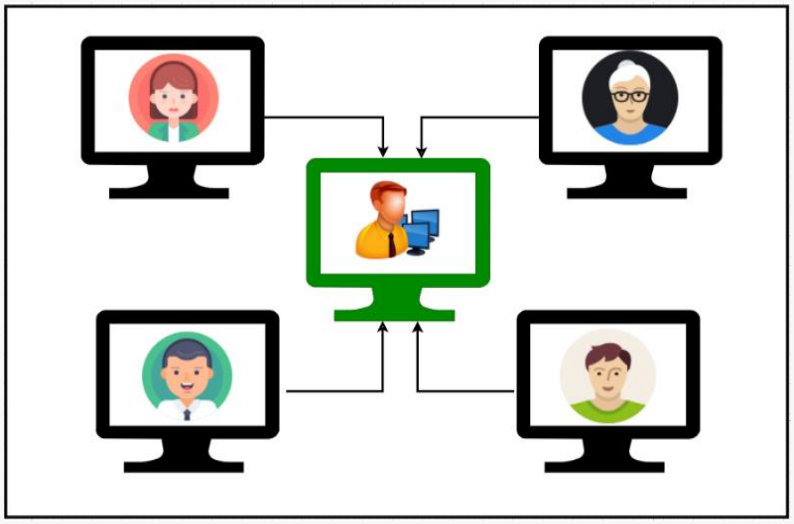
Para fixar o entendimento do assunto, crie o seguinte cenário: imagine que você possui a tarefa de separar e contar milhares de bolas coloridas dentro de uma piscina. As Figuras 15, 16 e 17 ilustram de maneira simples as etapas de um processamento distribuído em Big Data para essa tarefa.

Figura 15 - Distribuição dos dados.



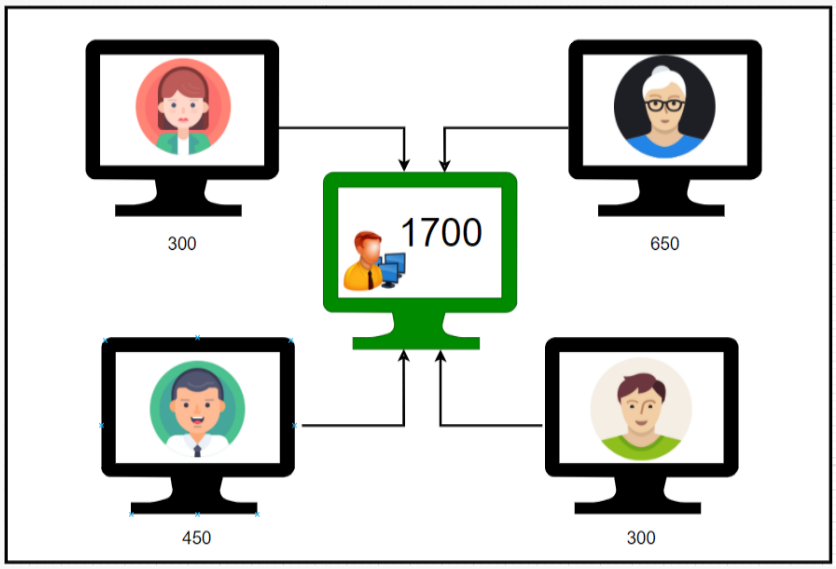
A Figura 15 ilustra a separação das bolas (dados) e a distribuição para as máquinas que irão processá-las. Cada máquina recebe uma quantidade de dados de acordo com a sua capacidade de processamento. A seguir, a Figura 16 ilustra as máquinas disponibilizadas, e estas compartilhando os recursos de processamento entre si. Percebam que cada máquina (cluster) está ligada a uma máquina central. Cada máquina vai processar os dados e, em seguida, enviar para a máquina central, que é responsável por agrupar e disponibilizar os dados processados.

Figura 16 - Compartilhamento de recursos.



Abaixo, a Figura 17 ilustra o agrupamento dos processamentos realizados por cada máquina. Cada máquina realizou a sua tarefa e, após os cálculos, envia para máquina central o resultado obtido. A máquina central, por sua vez, coleta todos os dados enviados pelas outras máquinas, agrupa-os e finaliza a tarefa.

Figura 17 - Centralização dos processos distribuídos.



Nesse exemplo simples você pode perceber que a computação distribuída nos permite processar o Big Data, porque os divide em partes mais gerenciáveis e distribui o trabalho entre os computadores que podem processar os dados.

Quando pensamos em termos de processamento de big data, há dois tipos de dados que processamos: dados em lote e *streaming*. Esses termos – lote e *streaming* – referem-se à maneira como obtemos nossos dados e à sua velocidade.

5.2. Processamento em lote

O processamento em lote, também conhecido por processamento em *batch*, são dados que temos em armazenamento e que processamos de uma vez ou em lote. O processamento em lote é usado com frequência quando se lida com grandes volumes de dados ou fontes de dados de sistemas legados, em que não é possível entregar dados em fluxos. Por definição, o processamento demanda que todos os dados necessários para o lote sejam carregados em algum meio de armazenamento, seja um banco de dados ou sistemas de arquivos. Por isso, as análises dos dados só são realizadas quando todos os dados são carregados.

O processamento em lotes é usado em uma variedade de cenários, de transformações de dados simples a um pipeline ETL (extração, transformação e carregamento) mais completo. Em um contexto de Big Data, o processamento em lotes pode operar em grandes conjuntos de dados, em que o cálculo leva um tempo significativo para ser concluído. O processamento em lotes geralmente leva a uma exploração mais interativa, fornece os dados prontos para modelagem para aprendizado de máquina ou grava os dados em um repositório de dados que é otimizado para análise e visualização. (MICROSOFT, 2018).

No nosso dia a dia, existem muitos processamentos em lote que não nos damos conta. Por exemplo, leituras de consumo de água, luz, cartões de crédito, entre outros. Na leitura do consumo de água, a empresa distribuidora não possui de forma automatizada o consumo gerado após a leitura do quadro de consumo, e sim somente após o envio dos dados coletados pelo funcionário que todos os dados serão processados.

Ademais, podemos citar as empresas de telecomunicações, as quais processam o uso do telefone celular a cada mês para gerar nossas contas telefônicas mensais. Para fazer isso, elas processam dados em lote, todas as ligações, as mensagens de texto e quaisquer cobranças adicionais que o cliente recebeu durante o ciclo de faturamento para gerar a fatura do cliente.

Vantagens:

- Acelera o processamento de informações em massa;
- Busca evitar a ociosidade do computador, não necessitando de supervisão ou interação do usuário;
- Permite o uso de diferentes prioridades para trabalhos interativos e não interativos;

- Executa apenas uma vez o programa para processar grandes quantias de dados, reduzindo a sobrecarga de sistema.

Desvantagem:

- Ele possui várias desvantagens, entre elas o usuário não consegue finalizar um processo durante a sua execução, de forma que seja necessário aguardar até a finalização do processo.

5.3 Processamento em Streaming

Os dados de *streaming* (“fluxo”) são dados que estão sendo produzidos continuamente por uma ou mais fontes e, portanto, devem ser processados de forma incremental à medida que chegam. Ou seja, eles podem ser processados, arquivados e analisados em tempo real. O stream também pode estar envolvido no processamento de grandes quantidades de dados, mas o processamento em lote funciona melhor quando você não precisa de análises em tempo real.

Como o processamento *stream* é responsável pelo processamento de dados em movimento e pelo rápido fornecimento de análise, ele gera resultados quase instantâneos usando plataformas de BI. O processamento em streaming está presente cada vez mais no nosso dia a dia. Podemos citar alguns exemplos de plataformas que utilizam:

Plataformas de áudio:

- Spotify;
- Deezer;
- Apple Music;
- YouTube Music;

- Amazon Music.

Plataformas de vídeo:

- YouTube;
- Netflix;
- Vimeo;
- Twitch.

O processamento em *streaming* também é utilizado em operações financeiras, como utilização de cartão de crédito e débito e operações de créditos de celulares. Por exemplo, no caso de pagamento em cartão de débito, o valor é descontado e atualizado na conta instantaneamente. Outro exemplo para o mundo real são os monitores cardíacos. Durante todo o dia, conforme você usa seu monitor cardíaco, ele recebe novos dados – dezenas de milhares de pontos de dados por dia, de acordo com o batimento do coração. Nesse sentido, cada vez que seu coração bate, novos dados são adicionados ao seu monitor cardíaco em tempo real. Se o seu monitor cardíaco exibe a média de seus batimentos cardíacos durante o dia, essa média deve ser constantemente atualizada com os novos números do fluxo de entrada de dados.

Vantagens:

- Processamento em tempo real;
- Não existe a necessidade de realizar download;
- Permite a inteligência operacional e análises de Big Data em tempo real;
- Processamento contínuo.

Desvantagens:

- Necessidade de ordenamento de dados;
- Consistência e durabilidade;
- Grande custo operacional.

A tabela da Figura 18 ilustra as principais diferenças entre o processamento em lote e o processamento em *Streaming*.

Figura 18 – Tabela de diferenças entre processamentos

	Processamento em Lotes	Processamento em Streaming
Análise	Dados analíticos complexos	Métricas mais simples, agregações e rotações
Escopo de dados	Consultas ou processamento de todos ou a marioria do conjunto de dados	Consultas ou processamento de dados a cada nova atualização
Performance	Latência de minutos a horas	Latências da ordem de segundos ou milissegundos
Tamanho dos dados	Grande lotes de dados	Registros individuais ou micro lotes

Fonte: Adaptado de Chambers e Zaharia (2018).

> Capítulo 6

Capítulo 6. Introdução aos frameworks e ferramentas do Big Data

Neste capítulo, vamos apresentar as principais ferramentas e *frameworks* utilizados para trabalhar com o Big Data.

6.1 Introdução às ferramentas utilizadas no Big Data

Existem diversas soluções disponíveis no mercado para o processamento de Big Data. A seguir, vamos apresentar algumas delas.

Coleta de dados:

- Import.io:
  - Com simples cliques é possível extrair todas as informações de uma página em um relatório completo, que poderá ser analisado por outros programas.
- Apache Chukwa:
  - Desenvolvido com base no *Hadoop*, essa ferramenta é *Open Source*, sendo bastante robusta para coletar, disponibilizar, monitorar e analisar os resultados da empresa.

Controle e armazenamento de dados:

- Apache Hadoop:
  - *Hadoop*, a primeira solução de uso amplo voltada à análise de grandes volumes de dados.
- Spark:

- O *Spark* foi criado para resolver uma limitação do *Hadoop* de trabalhar com dados diretamente na memória.
- Ferramenta de diferentes aplicações, incluindo armazenamento, integração de processos com dados embutidos em tempo real, entre outras.

- Cassandra:
  - Esse sistema de banco de dados permite o controle, compressão e transmissão de uma grande quantidade de informações, sem comprometer a performance do computador. Uma de suas características marcantes é a flexibilidade, uma vez que pode ser utilizado em computadores de baixo poder de processamento.

Tratamento e limpeza de dados

- Data Cleaner:
  - O programa transforma os arquivos em estruturas limpas, organizadas e prontas para serem lidas por softwares de visualização de dados.

- OpenRefine:
  - É uma ferramenta poderosa para trabalhar com dados confusos. Limpa-os, transformando-os em outros formatos e estendendo-os com serviços da web e dados externos.

Mineração de dados

- Oracle Data Mining:

- Suas funcionalidades incluem descoberta de padrões, predileções e alavancagem de dados. Permite a identificação do comportamento dos consumidores e traça precisamente seus perfis.

- RapidMiner:

  - O RapidMiner é uma plataforma de software para atividades de ciência de dados e fornece um ambiente integrado para preparação de dados, aprendizado de máquina, mineração de texto, análise preditiva, aprendizagem profunda, desenvolvimento de aplicações e prototipagem.

- Orange:

  - Uma ferramenta de código aberto, para novatos e experts, com recursos de *Machine Learning*, visualização de dados e workflow interativo.

- Knime:

  - O KNIME (*Konstanz Information Miner*) é uma plataforma gratuita de análise de dados, relatórios e integração de dados. O *KNIME* integra vários componentes para aprendizado de máquina e mineração de dados através de seu conceito modular de pipelining de dados.

- WEKA:

  - O *Weka* é um software livre largamente utilizado para mineração de dados e oferece uma lista ampla de algoritmos para análise de dados, podendo ser

instalado em qualquer computador com Windows ou Linux.

Análise Estatística

- Statwing:
  - Ferramenta que permite limpar as informações, procurar por dados relacionados, criar charts. Tudo isso é feito em minutos, ao invés de horas, como seria com as ferramentas comuns de análise.

Visualização de informações

- Tableau:
  - Sua plataforma de análise visual facilita ao usuário a exploração e o gerenciamento dos dados e agiliza a descoberta e o compartilhamento de informações que podem alavancar os resultados da sua empresa.
- Chartio:
  - Permite a combinação de dados e a criação de relatório diretamente de seu navegador, os arquivos podem ser exportados em PDF e enviados aos e-mails selecionados.

Integração de dados

- Pentaho:
  - Pentaho é um software de código aberto para inteligência empresarial, desenvolvido em Java. A solução cobre as tarefas de ETL, reporting, OLAP e mineração de dados.

6.2 Apache Hadoop

Apache Hadoop é uma estrutura que permite o processamento distribuído de grandes conjuntos de dados em *clusters* de computadores usando modelos de programação simples. Ele foi projetado para ser dimensionado de servidores únicos a milhares de máquinas, com cada uma oferecendo computação e armazenamento local. Em vez de depender de *hardware* para fornecer alta disponibilidade, a própria biblioteca foi projetada para detectar e tratar falhas na camada de aplicativo, oferecendo um serviço altamente disponível sobre um cluster de computadores, cada um dos quais sujeito a falhas. (APACHE HADOOP, 2021).

Quatro módulos compreendem o *framework* principal do *Hadoop* e funcionam coletivamente para formar o ecossistema *Hadoop*:

1. *Hadoop Distributed File System* (HDFS): como o principal componente do ecossistema *Hadoop*, o HDFS é um sistema de arquivos distribuídos que fornece acesso de alta capacidade aos dados do aplicativo, sem a necessidade de definir esquemas antecipadamente.

2. *Yet Another Resource Negotiator* (YARN): o YARN é uma plataforma de gerenciamento de recursos responsável por gerenciar recursos de computação em *clusters* e usá-los para programar os aplicativos dos usuários. Ele realiza programação e alocação de recursos em todo o sistema *Hadoop*.

3. *MapReduce*: o *MapReduce* é um modelo de programação para processamento de dados em grande escala. Usando algoritmos de computação distribuída e paralela, o *MapReduce* possibilita a transferência da lógica de processamento e ajuda a gravar aplicativos que transformam grandes conjuntos de dados em um conjunto gerenciável.

4. *Hadoop Common*: o *Hadoop* Common inclui as bibliotecas e utilitários usados e compartilhados por outros módulos do *Hadoop*.

Todos os módulos do *Hadoop* são projetados com a suposição fundamental de que as falhas de hardware de máquinas individuais ou conjunto de máquinas são comuns e devem ser tratadas automaticamente no software pelo framework. Os componentes *Apache Hadoop MapReduce* e HDFS derivaram originalmente dos documentos *Google MapReduce* e *Google File System* (GFS).

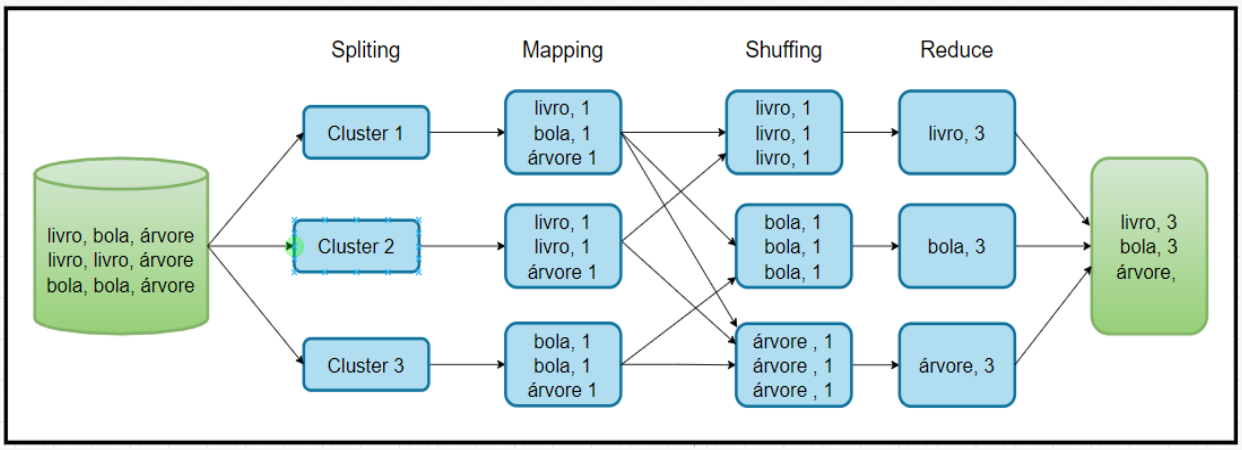
Além de HDFS, YARN e *MapReduce*, todo o ecossistema de código aberto *Hadoop* continua a crescer e inclui muitas ferramentas e aplicativos para ajudar a coletar, armazenar, processar, analisar e gerenciar Big Data. Eles incluem *Apache Pig, Apache Hive, Apache HBase, Apache Spark, Presto* e *Apache Zeppelin*.

6.2.1 Estrutura MapReduce

A estrutura *MapReduce* tem como objetivo realizar o processamento paralelo distribuído dos dados para os clusters, sendo processados e agregados ao resultado.

O *MapReduce* possui duas fases de processamento: o Map e o *Reduce*. A primeira fase, a fase de mapeamento, é responsável pelo processamento primário dos dados de entrada. Os dados são distribuídos entre os clusters e processados. Ao final dessa etapa, os dados processados são enviados para a função de *reduce*, em que é realizada a agregação dos resultados obtidos na etapa de *Map*. Com os dados já agregados, o próximo passo é consolidar todos os dados processados na função de *reduce*. A Imagem 19 ilustra as etapas da função *MapReduce*.

Figura 19 - Etapas da função MapReduce.



6.2.2 Benefícios do Apache Hadoop

Algumas das razões para se usar *Hadoop* é a sua capacidade de armazenar, gerenciar e analisar grandes quantidades de dados estruturados e não estruturados, de forma rápida, confiável, flexível e de baixo custo.

- Escalabilidade e desempenho – distribuídos tratamento de dados local para cada nó em um cluster *Hadoop* permite armazenar, gerenciar, processar e analisar dados em escala petabyte.

- Confiabilidade – clusters de computação de grande porte são propensos a falhas de nós individuais no cluster. *Hadoop* é fundamentalmente resistente – quando um nó falha de processamento é redirecionado para os nós restantes no cluster, e os dados são automaticamente replicados em preparação para falhas de nó futuras.

- Flexibilidade – ao contrário de sistemas de gerenciamento de banco de dados relacionais tradicionais, você não tem que ter esquemas estruturados criados antes de armazenar dados. Você pode armazenar dados em qualquer formato, incluindo formatos semiestruturados ou não estruturados e, em seguida, analisar e aplicar esquema para os dados quando ler.

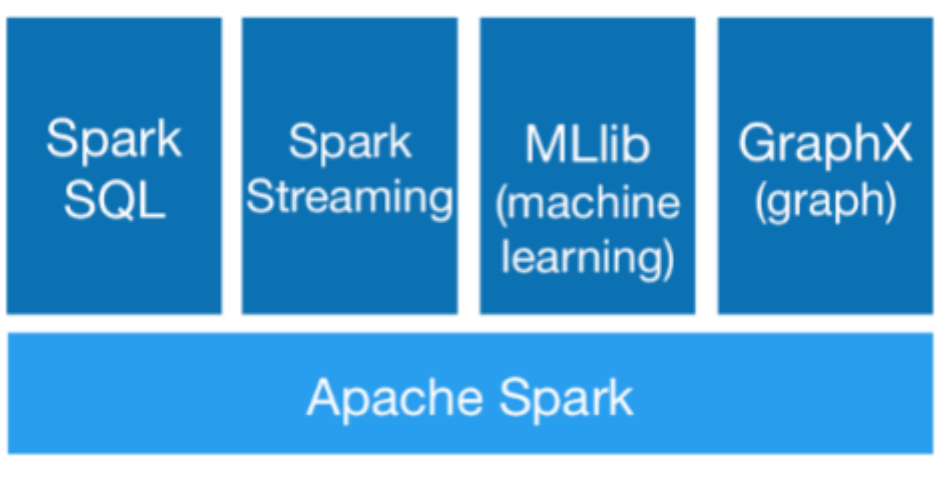
- Baixo custo – ao contrário de software proprietário, o *Hadoop* é open source e é executado em hardware *commodity* de baixo custo.

6.3 Apache Spark

O *Apache* é um framework open source para processamento de Big Data, construído com foco em velocidade, facilidade de uso e análises sofisticadas. Desde 2009, o *Apache Spark* tem sido desenvolvido pelo AMPLab da Universidade de Califórnia em Berkeley e, em 2010, seu código foi aberto como projeto da Fundação *Apache*. Desde então, o *Apache Spark* abrange um conjunto de desenvolvedores de mais de 300 empresas, com mais de 1200 desenvolvedores, que contribuíram para a construção do *Apache Spark*.

Apache Spark é um mecanismo de análise unificado para processamento de dados em grande escala. Ele fornece APIs de alto nível em Java, Scala, Python e R e um mecanismo otimizado que oferece suporte a gráficos de execução geral. Ele também oferece suporte a um rico conjunto de ferramentas de nível superior, incluindo *Spark* SQL para SQL e processamento de dados estruturados, MLlib para aprendizado de máquina, GraphX para processamento de gráfico e *Streams* estruturados para computação incremental e processamento de fluxo. A Figura 20 ilustra os componentes criados para o *Spark*.

Figura 20 - Componentes do *Apache Spark.*



Fonte: *Spark*.org.

Spark SQL é o módulo *Spark* para trabalhar com dados estruturados que oferece suporte a uma maneira comum de acessar uma variedade de fontes de dados. Ele permite consultar dados estruturados dentro de programas *Spark*, usando SQL ou uma API DataFrame familiar.

O *Spark* Streaming facilita a criação de soluções de streaming escalonáveis e tolerantes a falhas. Ele traz a API integrada à linguagem *Spark* para o processamento de stream, para que você possa escrever jobs de streaming da mesma forma que os jobs em lote. O *Spark* Streaming oferece suporte a Java, Scala e Python, e apresenta semânticas "exatamente uma vez" com estado, prontas para uso.

MLlib é a biblioteca de *Machine Learning* escalonável do *Spark* com ferramentas que tornam a ML prática escalonável e fácil. MLlib contém muitos algoritmos de aprendizado comuns, como classificação, regressão, recomendação e clustering. Também contém fluxos de trabalho e outros utilitários, incluindo transformações de recursos, construção de pipeline de ML, avaliação de modelo, álgebra linear distribuída e estatísticas.

GraphX é uma ferramenta que possui estruturas específicas para o armazenamento de grafos e oferece ainda uma ampla variedade de

algoritmos, tais como: PageRank, Componentes Conectados, Contagem de Triângulos, *Label Propagation*, Menor Caminho, dentre outros.

6.3.1 Arquitetura

A arquitetura de uma aplicação *Spark* é constituída por três partes principais:

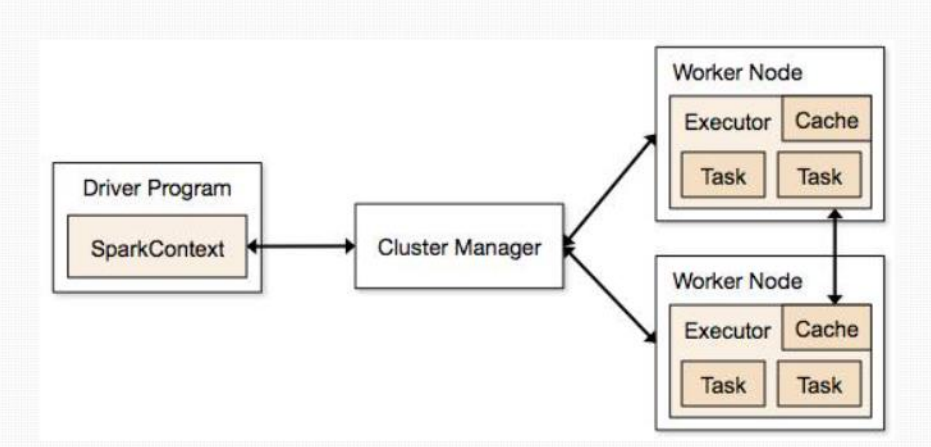
1. O *Driver Program* é a aplicação principal que gerencia a criação e define o processamento que será executado. Ele se comunica com os clusters para que os trabalhos possam ser distribuídos.

2. O *Cluster Manager* é um componente operacional responsável por administrar as máquinas que serão utilizadas. Gerencia os recursos do cluster, realiza balanceamento de carga, alocação de recursos e tolerância a falhas.

3. Os *Workers* são os componentes do cluster responsáveis por executar as tarefas que foram distribuídas pelo Cluster Manager.

A seguir, a Figura 21 ilustra a arquitetura do *Apache Spark.*

Figura 21 - Arquitetura do *Apache Spark.*



Fonte: *Spark*.org.

Além da arquitetura, é importante conhecer os principais componentes do modelo de programação do *Spark*. Existem três conceitos fundamentais que serão utilizados em todas as aplicações desenvolvidas:

1. *Resilient Distributed Datasets* (RDD): abstraem um conjunto de objetos distribuídos no cluster, geralmente, executados na memória principal. Estes podem estar armazenados em sistemas de arquivo tradicional, no HDFS e em alguns Banco de Dados NoSQL, como Cassandra e HBase. Ele é o objeto principal do modelo de programação do *Spark*, pois são nesses objetos que serão executados os processamentos dos dados.

2. Operações: representam transformações (como agrupamentos, filtros e mapeamentos entre os dados) ou ações (como contagens e persistências) que são realizadas em um RDD. Um programa *Spark* normalmente é definido como uma sequência de transformações ou ações que são realizadas em um conjunto de dados.

3. *Spark Context*: estabelece configurações de memória e processamento dos *Workers* Nodes. Além disso, é capaz de se conectar com os diferentes tipos de Cluster Manager (além do próprio *Spark Cluster Manager*) como *Apache* Mesos ou Yarn do *Hadoop*.

6.3.2 Benefícios do Apache Spark

Velocidade

O *Apache Spark* preferencialmente utiliza o processamento diretamente na memória RAM. Dessa forma, é possível executar cargas de trabalho 100 vezes mais rápido do que com o *Hadoop MapReduce*. O *Spark* atinge alto desempenho para dados em lote e de streaming usando um programador gráfico acíclico direcionado de última geração, um otimizador de consulta e um mecanismo de execução físico. No entanto, o *Spark* pode também utilizar o disco para esse fim, principalmente, quando faltam recursos de memória para que o processo possa ser corretamente concluído.

Fácil de usar

O *Apache Spark* oferece mais de 80 operadores de alto nível que facilitam a criação de apps paralelos. Você pode usá-lo interativamente a partir dos shells de Scala, Python, R e SQL para programar aplicativos rapidamente.

Generalidade

O *Apache Spark* capacita uma pilha de bibliotecas, incluindo SQL e DataFrames, MLlib para *Machine Learning*, GraphX e *Spark* Streaming. É possível combinar essas bibliotecas perfeitamente no mesmo aplicativo.

Inovações do framework de código aberto

O *Apache Spark* é apoiado por comunidades globais unidas pela introdução de novos conceitos e recursos de forma mais rápida e eficaz em relação a equipes internas que trabalham em soluções proprietárias. O poder coletivo de uma comunidade de código aberto oferece mais ideias, desenvolvimento mais rápido e solução imediata de problemas, o que possibilita um tempo de lançamento mais rápido.

Além disso, podemos citar que o *Apache Spark* possui tolerância a falhas, suporte a múltiplas linguagens, facilidade no desenvolvimento das aplicações, integração com SQL, modo de programação Python e Scala.

6.3.3 Diferenças entre Hadoop e Spark

A principal diferença entre o *Spark* e o *Hadoop* está na abordagem do processamento: o *Spark* pode fazer o processamento na memória, enquanto o *Hadoop MapReduce* precisa ler e gravar em um disco. Como resultado, a velocidade de processamento difere significativamente. O *Spark* pode ser até 100 vezes mais rápido. O *Spark* manipula a maioria de seus dados em memória, copiando da armazenagem física para a memória RAM. Isso reduz o tempo que é gasto para escrever e ler os dados do disco. O *Hadoop* escreve todos os blocos de dados de volta ao armazenamento físico após cada operação. Isso é feito para garantir a recuperação de falhas em caso de algo dar errado.

Em relação à tolerância a falhas, o *Hadoop* usa a replicação de todos os blocos de dados em várias cópias e realiza o armazenamento em disco após cada operação. O *Spark* usa o Conjunto de Dados Distribuídos Resiliente (RDD) que faz com que seus dados sejam armazenados em um conjunto de dados resilientes e distribuídos por todo o cluster.

Outra diferença entre o *Hadoop* e o *Spark* é que o *Spark* fornece uma variedade de APIs, que podem ser usadas com várias fontes de dados e idiomas. Além disso, eles são mais extensíveis que as APIs do *Hadoop*.

O *Spark* tem disponíveis ferramentas para processamento avançado de tarefas como aprendizado de máquina e aplicações de tempo real, que estão a frente do que é possível fazer com o *Hadoop* sozinho.

O *Hadoop* foi projetado para análise em lote com apenas dois operadores de mapeamento e de redução, enquanto o *Spark* possui mais operadores para trabalhar com processamento interativo.

> Capítulo 7

Capítulo 7. Capacitação e profissionais do Big Data

Neste capítulo, vamos conhecer os profissionais que compõem o time de Big Data e quais são os conhecimentos necessários para cada cargo.

7.1 Profissionais do Big Data

As equipes geralmente possuem quatro tipos de profissionais: os administradores de plataforma, os engenheiros de dados, os analistas de dados e os cientistas de dados. A seguir, vamos ver o que cada um faz.

Administradores de Plataforma: são aqueles que, como o próprio nome diz, gerenciam a plataforma. Eles oferecem suporte à infraestrutura de big data de uma organização, além de fazer um tipo de gestão para as equipes de desenvolvimento no que se refere às mudanças, configurações e atualizações para um sistema de big data. Frequentemente, eles também avaliam novas tecnologias que podem cooperar com a estrutura de big data já existente.

Algumas tarefas são:

- Definir e organizar a infraestrutura de big data;
- Atualizar e realizar manutenções quando necessárias;
- Realizar verificações de saúde;
- Acompanhar como a equipe está utilizando a plataforma;
- Implementar práticas recomendadas para gerenciamento de dados.

Engenheiro de Dados: é quem desenvolve, constrói, testa e mantém arquiteturas de dados, como bancos e sistemas de processamento de dados em grande escala. Uma vez criados esses reservatórios de dados, os

cientistas podem puxar conjuntos de informações que forem interessantes ao que estiverem procurando no momento.

Algumas das funções são:

- Projetar, construir, instalar, testar e manter sistemas de gerenciamento de dados altamente escaláveis;
- Construir algoritmos de alto desempenho;
- Pesquisar aquisição de dados e novos usos para os que já possuem;
- Integrar novas tecnologias de gerenciamento;
- Recomendar formas de melhorar a confiabilidade dos dados, juntamente à eficiência e à qualidade.

Analista de Dados: os analistas coletam os dados preparados pelos engenheiros para extrair insights. Eles geralmente apresentam os dados em formas de gráficos, diagramas e painéis, assim as partes interessadas podem tomar as decisões visualizando as informações necessárias. Geralmente, esses profissionais são bem versados nos conceitos de business intelligence e podem ser responsáveis por interpretar insights e comunicar de forma efetiva as descobertas realizadas.

Algumas funções são:

- Trabalhar com a equipe de TI, gestão ou cientistas de dados para alcançar os objetivos organizacionais;
- Coletar dados de fontes primárias e secundárias;
- Analisar e interpretar os resultados utilizando ferramentas estatísticas e técnicas convencionais;

- Fornecer relatórios de dados concisos e visualizações de dados claros para a gestão;
- Resolver problemas de códigos e questões relacionadas aos dados.

Cientista de Dados: os cientistas são grandes mineradores de dados, uma vez que recebem uma quantidade considerável de dados estruturados, ou não, e utilizam suas habilidades em matemática, estatística e programação para limpar, tratar e organizar os dados. Assim que os dados estão organizados, os cientistas utilizam da análise para gerar soluções para os problemas existentes e futuros.

Suas principais funções são:

- Realizar pesquisas e formular perguntas abertas aos dados;
- Extrair um grande número de dados de variadas fontes;
- Criar ferramentas que automatizam o trabalho;
- Comunicar os resultados obtidos e as previsões para as equipes que necessitam das informações;
- Recomendar mudanças aos procedimentos e estratégias existentes quando for o caso.

7.2 Skills dos profissionais do Big Data

Para desempenharem bem a função designada, esses profissionais necessitam de algumas competências profissionais específicas de cada cargo.

Os administradores de plataforma precisam usar as ferramentas de infraestrutura e serviços de monitoramento que os provedores de nuvem

geralmente oferecem para auxiliar na proteção e gerenciamento do sistema de big data.

Os engenheiros de dados necessitam de alguns conhecimentos específicos, como conhecer e utilizar linguagens de programação, como *Python* e *Scala*, os mecanismos de processamento de dados, como o *Apache Spark*, e ter diferentes soluções de armazenamento de dados.

Os analistas de dados precisam conhecer a linguagem de programação SQL e ferramentas como o *Tableau, Power BI, Looker.* Somado a isso, têm de saber sobre banco de dados relacionais e não relacionais, além de conhecimento sobre processamento de dados, matemática e estatística.

Os cientistas de dados necessitam de uma quantidade mais ampla e abrangente de conhecimentos para realizar sua função. Matemática, estatística e programação fazem parte desses conhecimentos, assim como *Machine* e *Deep* Learning, apresentação e visualização. Ademais, eles também necessitam de conhecimentos analíticos, como conhecimento da indústria e compreensão textual.

> Capítulo 8

Capítulo 8. Data Driven

Neste capítulo, vamos ver o conceito de *Data Driven* e conhecer algumas empresas que adotaram essa cultura e obtiveram êxito nos últimos anos.

8.1 Introdução ao *Data Driven*

A cultura de *Data Driven* consiste em adotar estratégias e tomar decisões baseadas na análise de informações, e não em intuições ou simples experiências.

Essa cultura não é como uma ferramenta, que pode ser utilizada em alguns momentos, mas uma metodologia bem estruturada que permite que as organizações tenham ideias mais precisas sobre seus negócios, e, assim, elas são capazes de aproveitar melhor as oportunidades.

Um dos pilares da cultura é realmente excluir quaisquer influências pessoais ou externas e basear as ações e estratégias nos dados que são apresentados. Dessa forma, o índice de assertividade se torna bastante elevado, embora o conceito possa soar um tanto impessoal.

Um dos objetivos do *Data Driven* é coletar dados de diversas fontes, tanto internas quanto externas, e cruzar as informações de forma a obter um panorama mais claro sobre o mercado e a própria instituição.

O *Data Driven* surgiu como um tipo de extensão da ciência de dados, utilizando os métodos científicos e os algoritmos, transformando dados em conhecimento.

Uma das principais diferenças entre empresas que aderem ao *Data Driven* e as que optam pelo modelo tradicional é o uso de dados de forma

integrada em seus processos e operações. Os dados geralmente ficam em nuvem, e não em servidores particulares. Dessa forma, todos os envolvidos possuem acesso às informações a qualquer instante.

O resultado fica ligado à inteligência coletiva, e não apenas na produtividade dos colaboradores de forma individual. Isso confere maior agilidade na rotina e maior propensão de avanço.

Por se tratar de uma mudança profunda de rotina, é primordial ter profissionais capacitados e especializados nesse assunto para que consigam trazer essa transformação para o cotidiano.

O *Chief Data Officer* (CDO) é um exemplo, posto que ele é o responsável por liderar as mudanças dentro da empresa e trazer um novo *mindset* aos colaboradores. Além disso, tem ao seu lado os cientistas de dados, que são profissionais que se relacionam diretamente com as informações, para sugerir ou indicar melhorias e resultados.

Os dois últimos pontos do *Data Driven* são fundamentais para que todo o resto ocorra bem: dados e tecnologia.

É de suma importância possuir dados organizados, acessíveis e integrados para que o processo caminhe como deve. Esse é o ponto que irá conceder aos profissionais aquilo que é necessário para extraírem o máximo de tudo que estiver disponível.

A tecnologia será a parte responsável por gerar soluções eficientes que irão sustentar toda a nova cultura organizacional. Com a tecnologia, será possível gerar ferramentas eficientes e pensadas para etapas e processos específicos dentro da organização, seja na gestão ou nas atividades de análise comuns, e isso gerará um negócio sustentável a longo prazo.

8.2 Empresas *Data Driven* e casos de sucesso

A cultura de *Data Driven* já vem sendo implementada por algumas instituições. Abaixo, vemos alguns cases em que as empresas implementaram essa cultura obtiveram muito sucesso. (MARR, 2016).

NETFLIX:

A empresa utiliza o *Data Driven* para nos oferecer exatamente aquilo que estamos procurando. A empresa de stream faz uma análise do comportamento de cada usuário e, em seguida, faz uma série de recomendações de acordo com as preferências de cada um. Se você entrar no perfil de uma adolescente, encontrará variados tipos de séries e filmes, mas, se entrar no perfil de um jovem adulto, poderá encontrar outros gêneros de filmes e documentários.

Adotar a cultura do *Data Driven* garantiu diferenciais competitivos a essa empresa, bem como a tornou um referencial no mercado. Esses dados são fornecidos conscientemente por cada usuário, respeitando todas as normas da lei de proteção aos dados.

SHELL:

Se você pensa que uma empresa petrolífera não tinha onde inovar, posto que lida com recursos fósseis, está completamente enganado. Não é uma novidade o fato de que os recursos fósseis utilizados para a produção de gases, energia e combustível estão cada vez mais difíceis de se encontrar, não só pela indisponibilidade iminente do recurso, mas também pela dificuldade atual na sua extração. Esses são desafios que as empresas petrolíferas estão encarando e buscando a melhor forma de sobressair no mercado.

Através do Big Data e da tecnologia que vem se desenvolvendo, a Shell tomou decisões importantes, principalmente no que diz respeito à tomada de decisões baseada nos dados. É por meio dessa tecnologia que a multinacional vem desenvolvendo uma estratégia que inclui captar ondas

sísmicas causadas pela atividade tectônica, as quais indicam onde há maior concentração de hidrocarbonetos.

Antes, essa técnica poderia obter qualquer resultado, mas, com o Big Data, é possível cruzar inúmeras informações de variadas fontes, trazendo maior assertividade no que se procura.

A tecnologia também favorece as máquinas utilizadas para extração e pesquisa, facilitando o trabalho. Tanto a Shell quanto qualquer outra companhia são extremamente cuidadosas quanto aos dados e à análise do mesmo, além de estarem mais confiantes em suas capacidades de prever as reservas, graças a análises avançadas do Big Data.

É por esses motivos que a Shell obteve grande sucesso adotando a cultura de *Data Driven*, pois, em um ramo de altos riscos e possibilidades, uma boa análise de dados e o gerenciamento destes torna-se um grande diferencial competitivo.

RALPH LAUREN:

A tecnologia avança para todos, inclusive para a moda. A empresa de moda americana Ralph Lauren, que comercia vestuário, itens de casa, acessórios e fragrâncias, aderiu ao *Data Driven* e utilizou o Big data para fornecer ao consumidor final um produto que superasse as expectativas quanto ao vestuário.

A verdade é que a tecnologia muda o comportamento do consumidor, dessa forma, faz-se necessário que o mercado ofereça ao seu cliente aquilo que é relevante para ele. Foi vendo esse desafio e pensando em como gerar um diferencial competitivo que a Ralph Lauren, juntamente com a empresa canadense OMsignal, desenvolveu a Polo Tech Shirt, que é uma camisa projetada para mapear as informações do usuário enquanto ele se exercita.

Basicamente, ela é um grande sensor que capta em tempo real a direção e o movimento, além de dados biométricos como a frequência cardíaca. Os dados da camisa são transmitidos para a nuvem e analisados usando algoritmos. O aplicativo, então, usa os insights a partir dessa análise para adaptar o treino de acordo com o usuário.

EXPERIAN:

A quarta empresa que iremos ver é a Experian, comumente conhecida no Brasil como "Serasa Experian". A companhia irlandesa atualmente emprega 17 mil colaboradores e tem sede em várias cidades ao redor do mundo. Essa empresa faz uso do Big Data para obter inúmeras melhorias e aumentar consideravelmente o índice de assertividade nas decisões. A Experian é uma empresa conhecida por fornecer referências de crédito, usadas por bancos e empresas de serviços financeiros para avaliar o risco ao decidir se deve emprestar dinheiro.

Eles também fornecem uma variedade de outros serviços baseados nos dados que coletaram, como proteção contra fraude e roubo de identidade. Recentemente, eles adicionaram serviços especializados de análise de dados voltados para ajudar clientes empresariais no mercado de automóveis, seguro saúde e pequenos negócios.

A Experian hospeda seu banco de dados de referência do consumidor de 30 petabytes em um cluster de computação seguro com *Linux* construído em torno da arquitetura *Hadoop*. O *Hadoop* é usado para armazenamento distribuído, e os núcleos do servidor também contribuem com poder de processamento para as operações analíticas – essencial para o processamento de dados de alto volume e alta velocidade, necessários para fornecer seus serviços quase em tempo real.

Referências

AUDY, J. L. N.; ANDRADE, G. K. C. *Fundamentos de sistemas de informação*. Porto Alegre: Bookman, 2005.

CIELEN, Davy; MEYSMAN, Arno; ALI, Mohamed. *Introducing Data Science*: Big Data, Machine Learning and More, Using Python tools. [*S. l.*: *s. n.*], 2016. Disponível em: <https://br1lib.org/book/2739951/aeeb06>. Acesso em: 16 fev. 2022.

COLAÇO JR., M. *Projetando sistemas de apoio à decisão baseados em Data Warehouse*. São Paulo: Axcel Books do Brasil, 2004.

DESAI, A.; SHAH, N.; DHODI, M. Student profiling to improve teaching and Learning: A data mining approach. In: *IEEE International Conference on Data Science and Engineering (ICDSE)*, 2016.

EMC EDUCATION SERVICES. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. 1. ed. Wiley, 2015. p. 410.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. (1996). Knowledge discovery and data mining: Towards a unifying framework. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96 Proceedings: 1996. p. 72-88.

INMON, WH. *What is a Data Warehouse.* [*S.I.*] Prism Tech Topic, 1995. Disponível em: <repository.binus.ac.id>. Acesso em: 05 ago. 2021.

JACOBSON, Reed; MISNER, Stacia; CONSULTING, Hitachi. *Microsoft SQL Server 2005*: Analysis Services. 1. ed. Bookman, 2007.

KANUNGO, T. at all. An efficient k-means clustering algorithm: Analysis and implementation. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence. Pattern Ana*l. Mach. Intell., 24(7): 881–892.

LLOYD, S. (1982). Least squares quantization in PCM. In: *IEEE Transactions on Information Theory*, 28(2):129–137.

MARR, Bernard. *Big Data In Practice*: How 45 Successful Companies Used Big Data Analytics To Deliver Extraordinary Results. 1. ed. West Sussex: John Wiley and Sons, 2016. p. 323.

MUELLER, John Paul; MASSARON, Luca. *Machine Learning*: For Dummies. 1. ed. John Wiley & Sons, 2016. p. 435

OLSEN, Wendy; BUENO, Dirceu da Silva Daniel. *Coleta de Dados*: Debates em Métodos Fundamentais em Pesquisa Social. 1. ed. Penso, 2015.

PEREZ, André. Sobre a História da Computação Distribuída e Clusters Kubernetes. *Medium*: Data Team Stone, 22 jan. 2021. Disponível em: https://medium.com/team-data-stone/sobre-a-história-da-computação-distribu%C3%ADda-e-clusters-kubernetes-3d0fe331db7. Acesso em: 16 fev. 2022.

SINGH, Ajit; AHMAD, Sultan. Architecture of Data Lake. International Journal of Scientific Research. In: *Computer Science Engineering and Information Technology*, 3 mar. 2016. Disponível em: https://www.researchgate.net/profile/Ajit-Singh-46/publication/331890045_Architecture_of_Data_Lake/links/6061ef85458515e8347d6ecc/Architecture-of-Data-Lake.pdf. Acesso em: 16 fev. 2022.

SUN, Zhaohao; ZOU, Huasheng; STRANG, Kenneth. *Big Data Analytics as a Service for Business Intelligence*. Lecture Notes in Computer Science, 26 nov. 2015.

TEJADA, Zoiner. *Mastering Azure Analytics*: Architecting in the Cloud with Azure Data Lake, HDInsight, and *Spark*. 1. ed. O'Reilly Media, 2017. p. 412.

TEJADA, Zoiner. Processamento em Lotes. *Microsoft Docs*, 12 fev. 2018. Disponível em: <https://docs.microsoft.com/pt-br/azure/architecture/data-guide/big-data/batch-processing>. Acesso em: 16 fev. 2022.

XU, Lida et al. Research on business intelligence in enterprise computing environment. In: *2007 IEEE International Conference on Systems, Man and Cybernetics*. 2007.