Fundamentos

Bootcamp Cientista de Dados

Davidson Ignacio Oliveira

2021

Fundamentos

Bootcamp Cientista de Dados

Davidson Ignacio Oliveira

Sumário

Capítulo 1. Introdução ao ecossistema

O curso aborda de forma geral diversos aspectos do processo de análise de dados. Serão apresentados exemplos de aplicações no contexto empresarial atual, envolvendo competências fundamentais para o profissional que almeja trabalhar ou assumir a liderança em projetos de análise de dados e tomada de decisão.

Antes de iniciar os estudos sobre as técnicas, processos e ferramentas de análise, precisamos entender alguns conceitos que fazem parte do ecossistema de Data Analytics.

É importante ressaltar também que por se tratar de assuntos relacionados à TI e por muitos desses conceitos terem surgido ou evoluído nos últimos anos, trazendo novidades para a área, grande parte do material disponível está em inglês. Durante o curso foi feito o esforço de traduzir o que é possível para o português, porém, alguns termos simplesmente perdem o sentido quando traduzidos, sendo globalmente entendidos na sua escrita original.

Big Data

O primeiro conceito a ser discutido é o de Big Data. Para o Grupo Gartner, que é uma importante empresa de consultoria em Tecnologia da Informação, com grande influência na área, Big Data pode ser definido da seguinte forma:

“São os ativos de informação de alto volume, alta velocidade e/ou alta variedade que demandam formas de processamento de informação inovadoras e efetivas em custo que permitem insights avançados, tomada de decisão e automação de processos.”

Edd Dumbil, em seu livro, apresenta outra definição: “Big Data é o dado que excede a capacidade de processamento convencional dos sistemas de bancos de dados”.

Para a McKinsey & Company112, outra companhia americana de consultoria empresarial, “Big Data é um termo utilizado para descrever um grande volume de dados, em grande velocidade e grande variedade; que requer novas tecnologias e técnicas para capturar, armazenar e analisar seu conteúdo; e é utilizado para abrilhantar a tomada de decisão, fornecendo introspecção e descobertas, e suportando e otimizando processos”.

Podemos ver que existem diversas definições na literatura. Mas um aspecto que sempre se destaca em qualquer uma delas é que no Big Data teremos sempre muito volume, velocidade e variedade de informações formando uma base de dados complexa para ser armazenada e processada com ferramentas convencionais

Porém, existem ainda muitas discussões sobre o limite do Big Data. Quando podemos considerar Big Data? O que pode ser considerado um grande volume de dados? Questões como essas continuam incentivando pesquisadores e autores a evoluir esse conceito a cada dia.

Outro fator que influencia a evolução dos conceitos de Big Data é a diversidade das fontes de dados. São tweets, posts no Instagram, informações de geolocalização e diversos outros tipos de dados que são relativamente novos. São dados que chegam cada vez menos estruturados para processamentos e análises. Além disso, muitos desses dados só fazem sentido se estiverem inseridos em um contexto, que por sua vez pode depender de mais dados, estruturados ou não. Isso significa que atualmente, várias ações do ser humano são convertidas em dados. Seja por meio de celulares, smart watches (relógios inteligentes) ou uso de redes sociais, estamos contribuindo a cada minuto com o aumento dessa massa de dados.

De acordo com um estudo realizado pela IBM, existe uma estimativa de que sejam gerados 2,3 trilhões de Gigabytes por dia. Carros modernos podem ter mais de 100 sensores para monitoramento completo do veículo, por exemplo, nível de óleo, gasolina, pressão dos pneus etc. Mais de 4 bilhões de horas de vídeos são assistidos por mês no YouTube. Um a cada três líderes não confia nas informações que usa para tomar as suas decisões, o que pode ser grave dependendo do impacto da decisão. Tudo isso faz com que os estudos sobre Big Data se tornem ainda mais

importantes, pois as pessoas precisam dos dados para tomarem as melhores decisões e serem cada vez mais competitivas.

Os V's do Big Data

Inicialmente, Big Data foi definido a partir de 3 V's. Posteriormente foi expandido para 5 V's, e hoje alguns autores já citam 7 ou 10 termos chave para a definição desse conceito. Veremos a seguir cada um dos V's do Big Data:

- Volume: é o conceito que mais se destaca no Big Data. Trata da quantidade de dados que temos disponíveis.
- Velocidade: retrata a rapidez com que os dados são gerados ou atualizados.
- Variedade: não temos no Big Data apenas dados estruturados. Temos também dados semiestruturados e não estruturados, que podem ser os mais comuns. São diversas fontes de dados entregando diversos tipos de dados diferentes como áudio, imagens, vídeos, dados de sensores etc.
- Veracidade: é um dos grandes desafios. Quanto mais crescem as outras características, maior é a tendência de desconfiança nos dados.
- Valor: talvez seja o aspecto mais importante. Mesmo se houver os demais conceitos, se não existir valor nos dados, se eles não servem para um propósito, podem até ser descartados.
- Visibilidade: é essencial que os dados tenham visibilidade, ou seja, sejam bem apresentados e disponibilizados. É algo que exige criatividade, pois o volume de dados requer uma ferramenta capaz de processar, filtrar, criar gráficos, tabelas e exibir tudo o que for necessário.
- Variabilidade: diferentemente de variedade, a variabilidade dos dados se associa com a inconsistência dos dados, por exemplo com a identificação de anomalias e outliers nos dados. Além da velocidade inconsistente que os dados podem ser criados.

- Vulnerabilidade: a vulnerabilidade está intimamente ligada às questões de segurança e privacidade dos dados. Os dados devem ser disponibilizados e visualizados por diversos grupos, criando a necessidade de um alto controle de acesso aos dados.

- Validade: similar à veracidade. Os dados precisam ser validados e é necessário compreender quão corretos e precisos eles são.

- Volatilidade: também é necessário compreender por quanto tempo os dados são relevantes e por quanto tempo precisam ser armazenados. Dados obsoletos podem prejudicar análises.

Assim, podemos perceber que geramos um volume cada vez maior de dados e que diversas empresas já se beneficiam da organização e análise desses dados. As ferramentas e técnicas analíticas precisam acompanhar essa evolução.

Machine Learning

Machine Learning, ou Aprendizado de Máquina em português, é um campo de estudo que já existe há vários anos. É uma subárea de Inteligência Artificial que não teve o devido destaque por muitos anos. Esse assunto despertou atenção novamente com a evolução da capacidade de processamento dos computadores e também com a necessidade de técnicas de análises e previsões a partir de grandes massas de dados.

De acordo com Arthur Samuel, “Machine Learning é um campo de estudo que dá aos computadores a habilidade de aprender sem terem sido programados para tal”. Ou seja, assim como seres humanos conseguem aprender coisas novas, os computadores são submetidos a técnicas de aprendizado para construção de modelos que conseguem aprender com os dados e fazem com que sejam capazes de resolver problemas. Em outras palavras, de acordo com Tom Mitchell, “Um programa de computador é dito para aprender com a experiência E com a relação a

alguma classe de tarefas T e medida de desempenho P, se o seu desempenho em tarefas em T, medida pelo P, melhora com a experiência E".

Aprender nesse caso pode significar classificar objetos, agrupar itens, responder perguntas sobre imagens etc. São várias as tarefas que podem ser realizadas através de Machine Learning. Para exemplificar:

- Tomada de decisão: o computador é capaz de analisar uma base histórica de dados, entender os padrões e decidir se vai aceitar ou rejeitar um empréstimo para um cliente de banco.
- Regressão: a partir de um conjunto de características de uma casa como o bairro, cidade, número de cômodos e comparações com outras casas similares, é possível prever o preço de venda dessa casa.
- Clustering (ou agrupamento): é possível também utilizar as técnicas de Machine Learning para identificar padrões nos dados e criar grupos similares, ou seja, uma base de clientes pode ser subdividida de acordo com características de sexo, idade, renda familiar etc.

Tipos de Aprendizagem

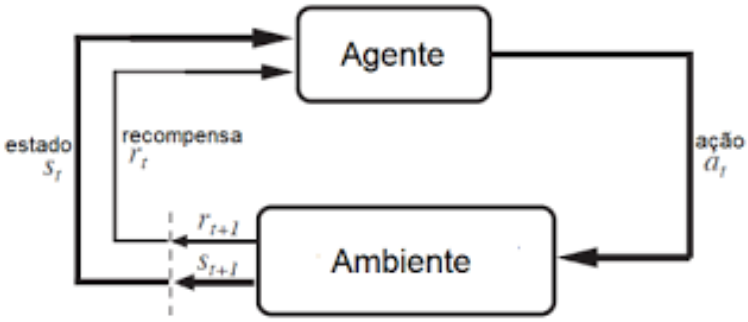
Para que o computador consiga aprender a resolver uma determinada tarefa é preciso instruí-lo da melhor forma possível. Isso pode ser feito por exemplo a partir de uma base de dados históricos. Porém existem casos em que não é possível coletar os dados históricos, por questões de tempo, esforço ou recursos. Para esses casos, existem técnicas de aprendizado específicas. Vejamos os detalhes de cada uma dessas técnicas:

- Aprendizagem supervisionada: é como um treino para o computador. Para esse tipo de aprendizagem existe previamente um conjunto de dados onde é possível conhecer qual é o resultado esperado para cada entrada do modelo. Com isso o sistema é capaz de avaliar se o resultado encontrado está próximo ao resultado correto.

- Aprendizagem não-supervisionada: nesse caso não existe um conjunto de treino. Consequentemente não existe um resultado específico esperado e não é possível prever os resultados do cruzamento das informações. Por isso é um tipo de aprendizado comumente utilizado para descobrir padrões entre os dados e resolver problemas de agrupamento.

- Aprendizagem por reforço: nesse tipo de aprendizagem existe um ambiente e um agente que interagem entre si através de percepções e ações. A cada iteração o agente recebe uma indicação do estado atual do ambiente e escolhe uma nova ação. A ação altera o estado do ambiente e o agente recebe um sinal de reforço. Ao final, o agente terá uma política de comportamento. Veja na Figura 1 o diagrama desse tipo de aprendizagem.

Figura 1 – Aprendizagem por Reforço.



Fonte: Adaptado de [Sutton 1999]

Ajustes do Modelo

É comum em sistemas de Machine Learning que um modelo construído seja específico demais ou que não consiga generalizar para outros casos. O ideal é que seu aprendizado seja bem equilibrado.

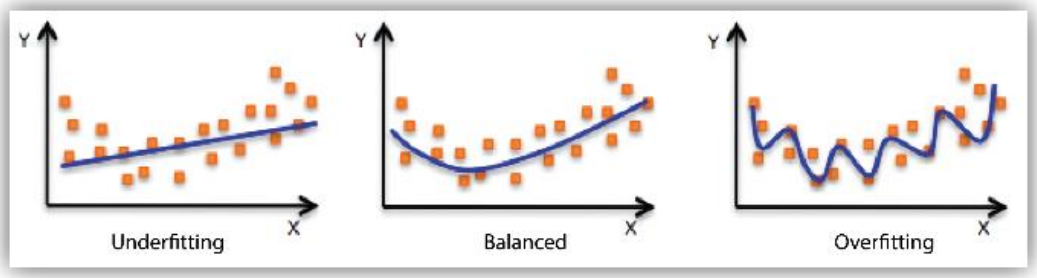
Quando um modelo é treinado pode acontecer de, na fase de treinamento, alcançar uma taxa de erro muito baixa, dando a impressão de que é um ótimo modelo.

Porém o desempenho é péssimo quando aplicado a um conjunto de teste. Isso provavelmente significa que é um caso de overfitting, ou seja, seu modelo não tem a capacidade de generalização. Nesse caso, o modelo memoriza os dados em vez de aprender.

Quando um modelo é treinado e na fase de treinamento a taxa de erro é relativamente alta e na fase de testes é mais alta ainda, provavelmente é um caso de *underfitting*, ou seja, o modelo é simples demais e acaba subestimando a realidade.

Veja na Figura 2 exemplos desses casos para um problema de classificação.

Figura 2 – Overfitting e Underfitting.

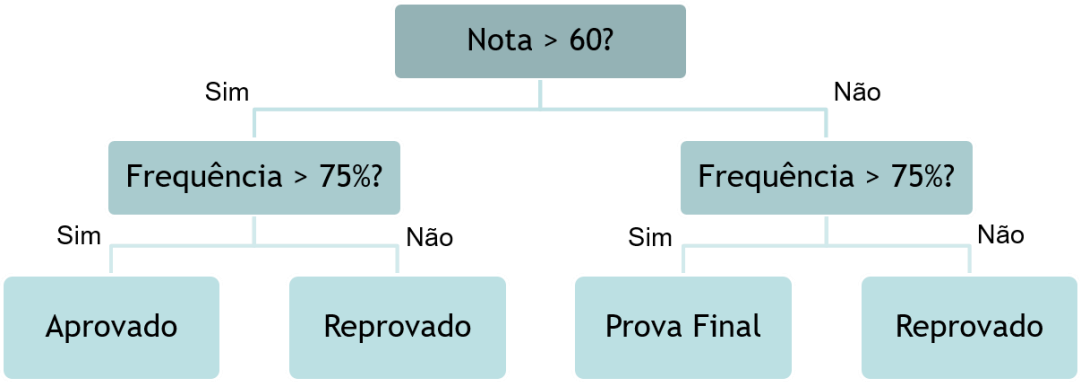


Fonte: AWS (2020)

Árvores de Decisão

Um dos tipos mais comuns de algoritmo de aprendizagem supervisionada para problemas de classificação é a árvore de decisão. Uma árvore de decisão representa um caminho de decisões ou classificações que devem ser tomadas a partir de um conjunto de dados até chegar em uma decisão final. Dessa forma, representam o mapeamento de possíveis resultados de uma série de escolhas relacionadas.

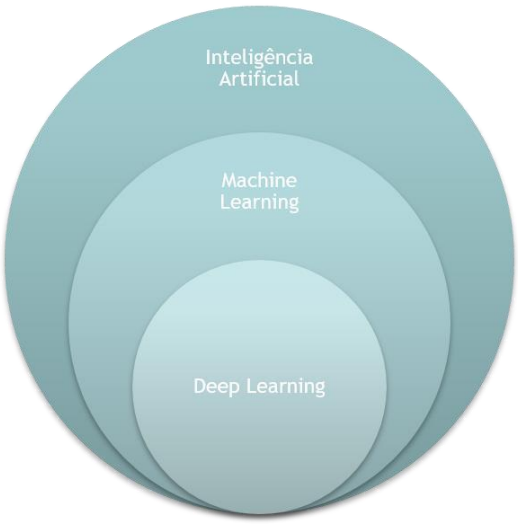
Figura 3 – Exemplo de uma Árvore de Decisão.



Deep Learning

Deep Learning, ou aprendizagem profunda em português, é uma subcategoria do aprendizado de máquina que se baseia em Redes Neurais Artificiais (RNA) para realizar o treinamento. Então, antes de continuar, precisamos entender basicamente o conceito de redes neurais. As RNA's são modelos computacionais inspirados no funcionamento dos neurônios do cérebro humano. O conhecimento é adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem. As forças de conexão entre neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido. Por não ser baseada em regras ou programas, a computação neural se constitui uma alternativa à computação algorítmica convencional. Técnicas de Deep Learning contribuem constantemente para a evolução da capacidade dos computadores de classificar e compreender os dados. Sua aplicação pode ser vista em carros autônomos, sistemas de recomendação, reconhecimento de voz, identificação de objetos etc. Sendo assim, é possível entender a relação entre os termos Inteligência Artificial, Machine Learning e Deep Learning, como mostrado na Figura 4.

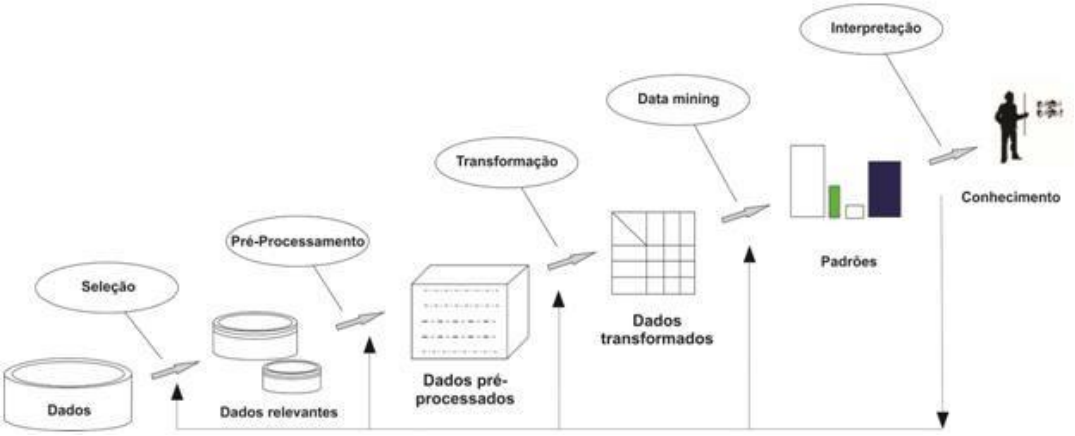
Figura 4 – Inteligência Artificial, Machine Learning, Deep Learning.



Data Mining

Com todo o avanço que existe na coleta e armazenamento de dados, o processo manual de análises e detecção de padrões indiscutivelmente se torna impraticável. Para extrair conhecimento de uma base de dados é preciso utilizar técnicas avançadas. Uma dessas técnicas é o Knowledge Discovery in Databases (KDD). De acordo com Fayyad, “KDD é um processo de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados. É uma tentativa de solucionar o problema causado pela chamada ‘Era da Informação’: a sobrecarga de dados”. Veja na Figura 5 o passo a passo do processo.

Figura 5 – Processo KDD.



Um dos principais passos é o Data Mining, ou Mineração de Dados em português. De acordo com David Hand, “Mineração de Dados é a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tanto úteis quanto compreensíveis ao dono dos dados”. Para isso são utilizadas técnicas de estatística, recuperação de informação, inteligência artificial, entre outros.

Para obter um bom resultado com a mineração de dados é fundamental conhecer a base de dados, saber se estão bem estruturados, conhecer previamente a relação entre alguns atributos. O processo de preparação dos dados é conhecido também como pré-processamento e contém quatro etapas, que são a limpeza, integração, transformação e redução. Durante a limpeza é preciso ficar atento aos registros incompletos, dados inconsistentes, definição de valores padrões. Na fase de integração será criada uma única base com dados das diversas fontes de dados. Para unificar as bases é necessário identificar quais são os atributos chave de cada uma. A transformação se encarrega da normalização e generalização dos dados. Finalmente, na fase de redução, o desafio é gerar uma base de dados menor sem perder a representatividade dos dados originais.

No processo de mineração de dados será necessário um treino para a criação do modelo. Utilizar todos os registros para esse processo pode ser inviável em alguns casos. Assim, são criados subconjuntos chamados amostras, que devem ter a maior representatividade possível:

- Conjunto de treino: Geralmente 70% do conjunto, mas pode variar dependendo do problema e das características dos dados. É o conjunto de dados que será usado para desenvolver o modelo.

- Conjunto de Testes: Geralmente 15% do conjunto. São registros diferentes dos que foram utilizados no treinamento. A base de testes estima a taxa de erro do modelo e quando comparado aos resultados de validação é possível perceber por exemplo se existe um *overfitting*.

- Conjunto de Validação: Quando se tem vários modelos criados a partir da base de treino, é possível utilizar a base de validação para encontrar o modelo com menor erro.

O caso mais comum de uso da mineração de dados é a detecção de padrões nos dados. Isso pode ser estendido a tarefas de descrição, classificação, regressão, predição, agrupamento e associação.

Para qualquer desses objetivos, existirão desafios no processo. As relações entre os atributos precisam ser muito bem definidas, caso contrário os resultados podem ser mal interpretados. Porém, nem sempre é possível conhecer essas relações e isso pode ser inclusive uma tarefa do modelo. Ou seja, o processo exige um alto conhecimento da base de dados para que os resultados sejam analisados corretamente.

Existem vários outros campos de estudo relacionados à Mineração de Dados como métodos genéticos e mineração em textos. Mas, como já mencionado, o objetivo é detectar padrões que permitam uma análise e descoberta de conhecimento.

Data Analytics

Data Analytics, ou Análise de Dados é o processo que torna possível a transformação de dados e informações em conhecimento para um propósito específico. É importante saber que uma análise de dados pode ocorrer mesmo sem a utilização de computadores. O importante é conseguir de fato transformar uma

informação em conhecimento. Nesse processo, geralmente temos primeiro um dado, que é o menor grão e a matéria prima da escala do conhecimento. Por exemplo, consultando um termômetro de um paciente, o médico vê o valor 36,5. Esse valor é o dado. A seguir temos uma informação, que é o dado organizado dentro de uma escala. Para o exemplo é o valor 36,5 °C. Ou seja, a partir desse momento já se sabe a unidade e que é uma temperatura. Então contextualizando essa informação, chega-se ao conhecimento. A contextualização nesse caso é que o paciente não está com febre. Finalmente é possível alcançar a sabedoria, que é um conjunto complexo de raciocínios a partir de um determinado conhecimento. Nesse caso, a sabedoria explicaria o que fazer caso o paciente estivesse com febre.

Para uma análise de dados é possível utilizar várias técnicas para lidar com Big Data, ou identificar padrões com técnicas de Mineração de Dados. Ou ainda usar Machine Learning para criar modelos de predição a partir dessa análise.

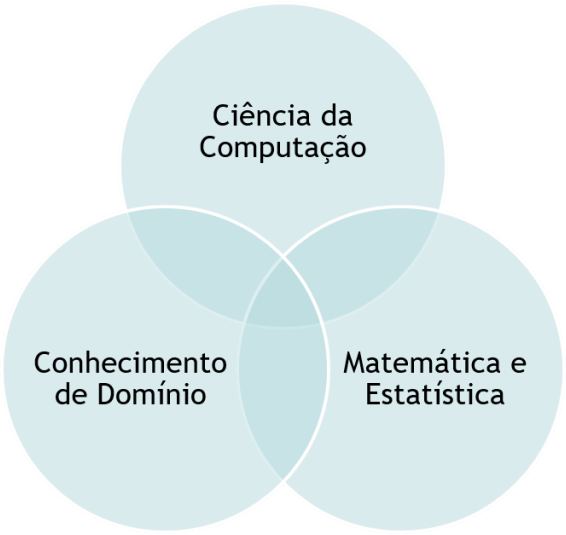
Existe um fluxo básico para análise de dados que veremos com mais detalhes no segundo capítulo. Esse fluxo começa com a origem dos dados. Podem ser dados da web, de sensores, PLC's etc. É preciso coletar os dados das diversas fontes disponíveis, e passar para a fase seguinte, de limpeza e transformação. Os dados raramente estão prontos para a análise. É preciso filtrar, verificar os valores e prezar pela qualidade da base de dados. Para uma boa análise ser feita, é preciso também que o analista saiba exatamente quais perguntas devem ser respondidas. Durante a análise é possível que algumas dessas perguntas mudem, ou apareçam novas. Mas inicialmente a análise deve ter um foco, uma linha de pensamento a ser seguida. Isso é fundamental para que no final da análise haja uma validação com estimativa do retorno do estudo realizado.

Data Science

Outro conceito que deve ser bem compreendido é o de Data Science, ou Ciência de Dados. É um conceito amplo que abrange todas as tarefas relacionadas à limpeza, preparação e análise de dados, além das técnicas utilizadas a fim de se

extrair dados e obter insights através de informações. Ou seja, se refere ao estudo de dados e informações características de um determinado assunto. Envolve diversas áreas de conhecimento como Matemática, Estatística, Computação além da compreensão do assunto específico da análise em questão. Veja na Figura 6 o diagrama da área de Data Science.

Figura 6 – Data Science.



Há alguns especialistas que dizem que “dados são o novo ouro” ou ainda “dados são o novo petróleo”. A intenção dessas frases é destacar o valor que os dados podem nos entregar quando são bem usados e analisados. A informação está cada vez mais disponível e dentro das empresas ninguém deveria mais tomar decisões sem se basear nos dados.

Business Intelligence

Business Intelligence, ou Inteligência de Negócio em português, é saber aplicar as estratégias de análise de dados ao negócio da empresa, melhorando o planejamento estratégico, previsões de mercado, orçamento etc. É a habilidade de transformar os dados em informação, e a informação em conhecimento, de forma que se possa otimizar o processo de tomada de decisões nos negócios.

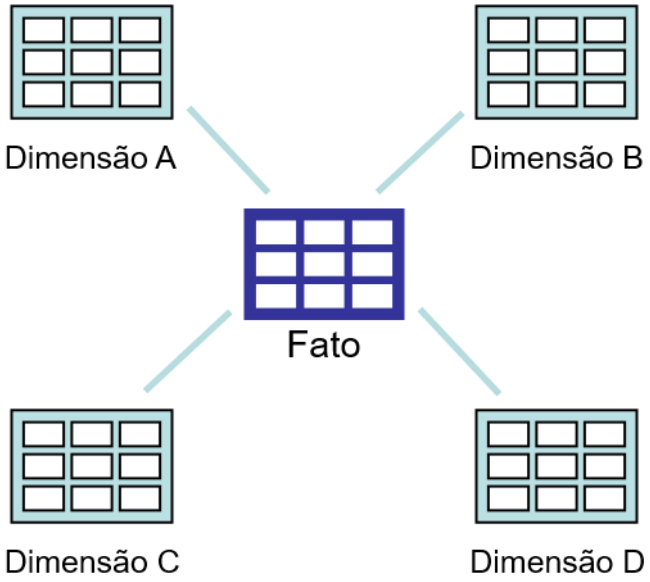
É comum algumas empresas criarem painéis e dashboards com informações gerais do processo e dizerem que estão aplicando BI em seu negócio. Porém BI vai além dos dashboards e de todo o ciclo de análise de dados.

Existem alguns conceitos que são muito presentes em projetos de BI, como ETL e OLAP. A sigla ETL significa Extract, Transform and Load, ou Extrair, Transformar e Carregar. É o processo de extração de dados de fontes externas, limpeza e transformação de acordo com regras de negócio e carga para um Data Warehouse. No segundo capítulo serão explicados com maiores detalhes os conceitos ETL e Data Warehouse.

OLAP (Online Analytical Processing)

A sigla OLAP significa Online Analytical Processing, ou Processamento Analítico Online. É um conceito de interface com o usuário que fornece a capacidade de manipular e analisar um grande volume de dados sob múltiplas perspectivas. Isso proporciona um ambiente criativo para o usuário, estimulando novas ideias e insights, e permite a análise profunda em diversos ângulos. O conceito de tabela Fato contribui para a arquitetura de uma solução OLAP. Uma base de dados de BI reúne informações de diversas fontes, e é comum que o Data Warehouse tenha uma arquitetura multidimensional. Com isso são criadas diversas tabelas que são chamadas “Dimensões” e uma tabela principal chamada “Fato”. Cada tabela Dimensão possui características que qualificam de alguma forma as informações da tabela Fato. Por exemplo, caso seja criada uma base de dados para a área de vendas de uma loja de roupas, podemos criar dimensões como clientes, produtos, fornecedores e vendedores. Veja na Figura 7 um diagrama ilustrativo de uma base de dados multidimensional. Com essas dimensões é possível dizer que foi construído um cubo do Fato vendas.

Figura 7 – Base de dados multidimensional.



Quando se implementa OLAP, é necessário atender algumas funções básicas como visualização multidimensional dos dados, exploração, rotação e permitir vários modos de visualização. Para que essas funções sejam atendidas, várias operações são disponibilizadas para o usuário. Veja algumas delas:

- Slice: selecionar dados de uma única dimensão.
- Dice: extrair um subcubo da seleção de duas ou mais dimensões.
- Pivot ou Rotation: visualizar dados por uma nova perspectiva.

Existem diversas arquiteturas diferentes de armazenamento para o OLAP e a solução pode ser classificada de acordo com essa característica. As três principais são:

- ROLAP: os dados são armazenados de forma relacional.
- MOLAP: os dados são armazenados de forma multidimensional.
- HOLAP: híbrido. Uma combinação dos métodos ROLAP e MOLAP.

Ainda existem outros tipos como DOLAP (Desktop OLAP), WOLAP (Web OLAP), XOLAP (eXtended OLAP). Cada um desses tipos tem uma aplicação diferente e pode ter um tempo de consulta diferente. Cada caso deve ser estudado para que seja encontrado o tipo ideal para o projeto.

Self Service BI

Existe uma grande parte de qualquer projeto de BI que depende exclusivamente de uma equipe de TI qualificada e à disposição para coleta, armazenamento e processamento dos dados. Esse processo pode se tornar lento em algumas organizações, por exemplo por questões burocráticas. No mercado atual o tempo certo da análise faz toda a diferença. A base do conceito de Self Service BI é que o poder deve ser dado ao usuário do negócio. Isso significa que a própria área de negócios deve ter disponível todas as informações necessárias, sem o contato direto com a TI, possibilitando uma tomada de decisão muito mais ágil.

Para que um projeto de Self Service BI seja implantado, é necessário que sejam discutidos vários tópicos internamente. Talvez um dos mais importantes seja a privacidade dos dados. Como não haverá mais uma equipe intermediária sempre que novos dados ou organização dos dados forem solicitados, a base de consulta deverá ter muito mais informações disponíveis. Novas estratégias de acesso e segurança podem ser estabelecidas para neutralizar esse problema.

Outro ponto a ser observado é que na maioria dos casos não bastará apenas disponibilizar os dados e a ferramenta de visualização. Os usuários de negócio em geral conhecem bastante dos seus clientes, produtos e mercado de atuação, mas conhecem pouco sobre as ferramentas e processos relativos ao processamento de dados. Por essa razão talvez sejam necessários treinamentos que capacitem os usuários a realizarem todas as tarefas necessárias.

K-Means na Prática

O K-Means é um algoritmo de Machine Learning baseado em aprendizado não supervisionado que tem o objetivo de tentar encontrar similaridades entre os dados e os agrupa conforme a quantidade de clusters (K).

De forma interativa, atribui os pontos de dados ao grupo que representa a menor distância. O algoritmo gera K (ou menos) clusters.

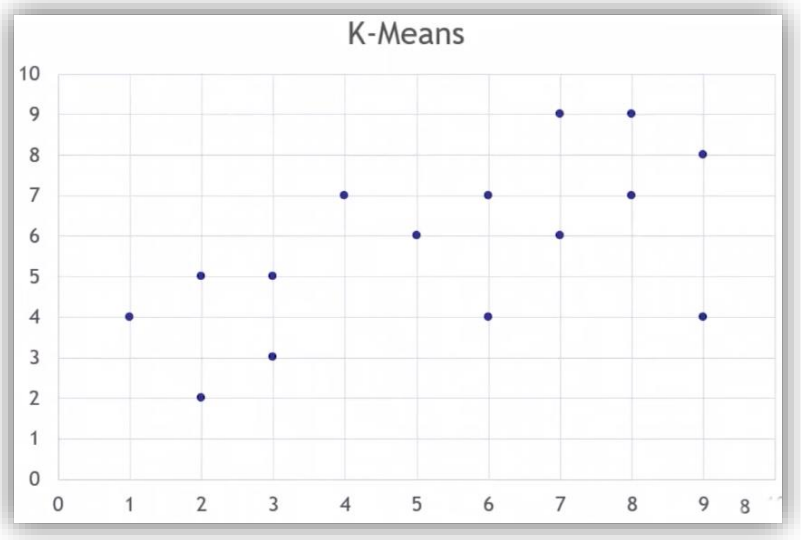
O K-Means é um dos algoritmos mais populares para resolver problemas de clusterização. A seguir os principais usos:

- Agrupamento de clientes/usuários similares.
- Segmentação de mercado.
- Agrupamento de produtos semelhantes.
- Agrupamento de usuários em redes sociais.
- Agrupamento de notícias, documentos.
- Agrupamento de pacientes para identificar situações de risco.

O passo a passo dele consiste em:

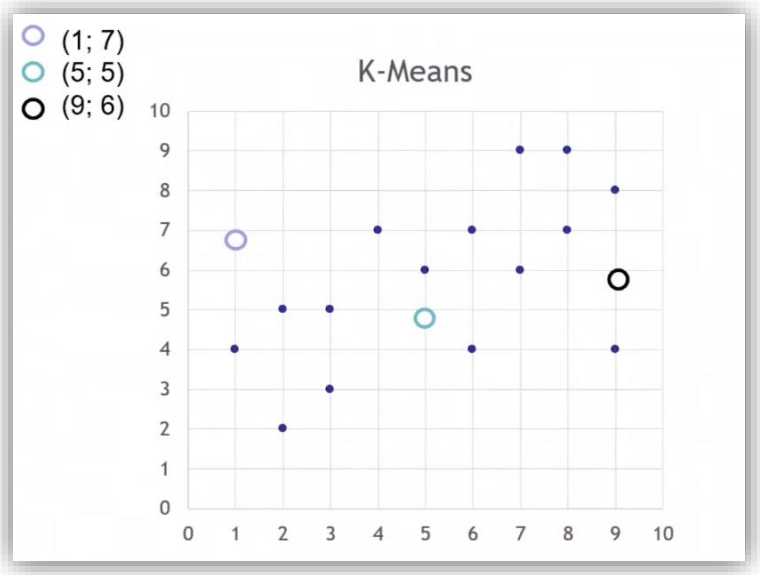
1. Inicializar os centroides aleatoriamente (necessário saber o valor de k antes de começar)
2. Para cada ponto, calcula a distância para cada centroide e associa ao que está mais próximo.
3. Calcula a média de todos os pontos relacionados a um centroide e define um novo centroide.
4. Volta e recalcula as instâncias, pois pode ser que o centroide mudou de cluster.

Figura 8 – Início Kmeans.



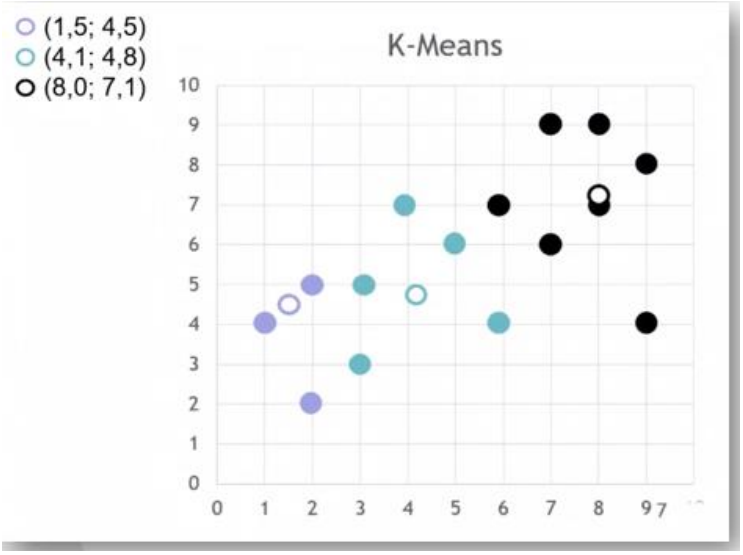
Considerando um conjunto de dados pequeno, como o descrito na figura 8, o K-means executará as interações buscando agrupar os pontos criando os clusters.

Figura 9 – Inicialização dos centroides.



Depois de inicializar os centroides, o K-means começa a fazer as iterações baseadas num cálculo matemático que leva em consideração as somas dos quadrados das distâncias. Ele faz esse cálculo para cada ponto de dados, estabelecendo ao final uma coordenada que aplica a clusterização em todos os pontos determinados pelo algoritmo.

Figura 10 – Clusterização final e coordenadas dos centroides.



Como o KMeans calcula a distância das observações até o centro do agrupamento que ela pertence, o ideal é que essa distância seja a menor viável. Matematicamente falando, nós estamos buscando uma quantidade de agrupamentos em que a soma dos quadrados intra-clusters seja a menor possível, sendo zero o resultado ótimo.

O número ideal de clusters pode ser escolhido por meio da curva do cotovelo, o método WCSS (Within Cluster Sum of Squares). O KMeans da biblioteca scikit-learn no Python já calcula o wcss. A distância entre os dados pode ser feita pela distância euclidiana, distância de Manhattan, entre outras. A condição ideal para parar as iterações é a própria quantidade de iterações e a estabilidade dos clusters.

Uma observação importante é a atenção na inicialização dos centroides. Tudo pode mudar se eles forem inicializados de maneira inadequada.

Capítulo 2. Processo completo de Data Analytics

Antes de começar qualquer projeto é preciso fazer um bom planejamento, montar um cronograma, avaliar os riscos etc. Com projetos de análise de dados não é diferente. Existe um ciclo de análise que pode acontecer várias vezes para um mesmo projeto. Veja na Figura 8 o ciclo de vida da análise de dados.

Figura 10 – Ciclo de vida para análises de dados.

Como podemos ver, o primeiro passo é entender o problema. É recomendado que seja investido um bom tempo nessa etapa para que todos os envolvidos possam compreender bem qual é o resultado esperado.

Para essa primeira etapa, uma dica é utilizar a técnica dos 5 W's, onde são definidas várias perguntas importantes para direcionar a análise. As perguntas básicas são: Why? Who? What? Where? When? Em português: Porque? Quem? O que? Onde? Quando? Veja um exemplo:

1. Por que é importante essa análise para o negócio?
2. Quem iremos analisar? Nossos compradores? Fornecedores?
3. O que iremos analisar? Comportamento de compra?

4. A análise estará voltada para o contexto nacional ou internacional?

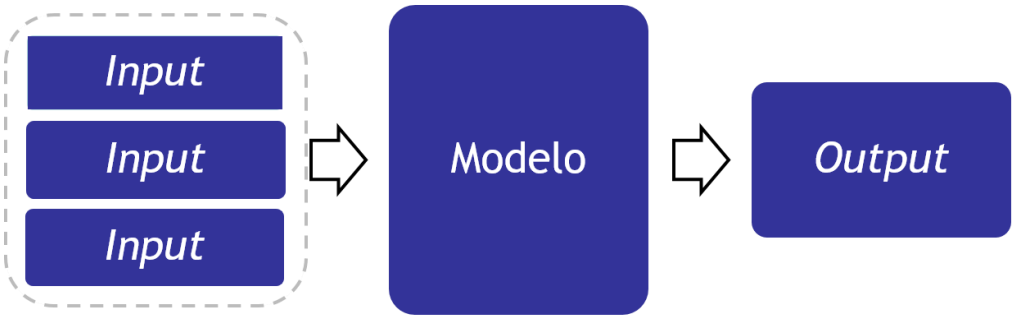
5. Qual período será considerado para as análises?

Todas essas perguntas podem ser utilizadas para definir a meta ou o objetivo final da análise.

As etapas seguintes são de coleta e preparação dos dados. Nessa etapa devemos analisar qual é a fonte de dados e definir se a coleta será por exemplo contínua, periódica ou ocasional. Sabendo disso, um arquiteto de dados deve ser responsável por modelar, estruturar e armazenar esses dados. Vários detalhes dessa etapa serão estudados no tópico ETL dessa apostila.

É chegada então a fase de processamento dos dados e criação do modelo. Para essa etapa podem ser utilizados todos os recursos disponíveis como Data Mining, Machine Learning, Deep Learning, Modelos estatísticos etc. O importante é que seja gerado um modelo, que pode ser uma função, um modelo numérico, visual, estatístico e assim por diante. Esse modelo deve suportar novas entradas, diferentes das que já estavam disponíveis na sua construção, para que sejam geradas novas saídas. Esses novos resultados do modelo é que serão o apoio necessário para uma tomada de decisão, um insight ou mesmo uma previsão de mercado. Veja na Figura 9 o funcionamento básico de um modelo.

Figura 9 – Modelo gerado a partir de um processamento.



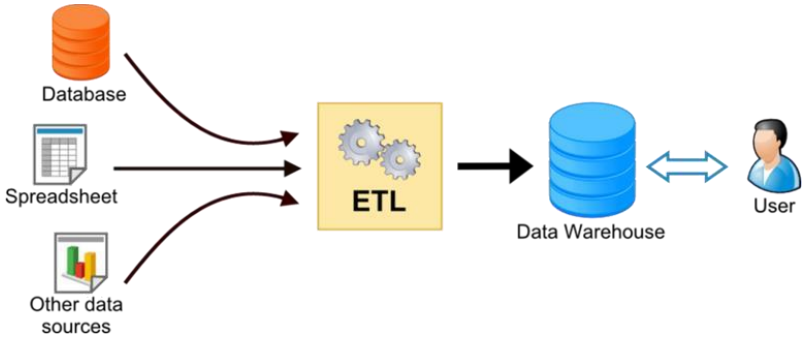
Com o modelo pronto, é hora de apresentar os resultados e coletar feedbacks. São vários os desafios para visualização de dados, e esses serão discutidos em

detalhes no tópico Visualização de Dados deste capítulo. É fundamental que a apresentação dos seus dados responda a todas as perguntas elaboradas no início da análise. Assim será possível avaliar se a meta foi cumprida. Talvez seja necessária uma nova iteração, isso quer dizer executar todo o ciclo novamente. Em uma nova iteração podem ser feitas novas perguntas, coletados dados diferentes, escolhidos novos modelos etc. Essa nova iteração pode acontecer logo em seguida ou após algum tempo (semanas ou meses). Isso porque o modelo pode se degradar com o tempo. Dificilmente teremos a garantia de que o modelo obtido permanecerá com alta precisão eternamente.

ETL – A etapa chave

ETL é o processo que tem como objetivo trabalhar com toda a parte de extração de dados de fontes externas, a transformação de acordo com as necessidades dos negócios e a carga para Data Warehouse e/ou Data Mart. Se essa fase não for executada corretamente poderá trazer consequências graves para as análises. Ou seja, o processo de ETL tem o poder de inutilizar ou potencializar a análise dos dados. Por isso é uma das fases mais críticas e existem estudos que indicam que o ETL e as ferramentas de limpeza de dados consomem cerca de 70% dos recursos de desenvolvimento e manutenção de um projeto de Análise de Dados. Veja na Figura 10 o esquema de um projeto ETL.

Figura 10 – Esquema ETL.



Para a execução de um projeto ETL deve-se avaliar bem o negócio e ter uma documentação que manifeste todos os requisitos da análise, pois serão fundamentais para o processo. Nessa fase também será estudada a viabilidade de coleta dos dados. Por questões de segurança, por questões técnicas ou outras pode ser que os dados não sejam disponibilizados e o escopo da análise seja alterado.

O processo de ETL se divide em três partes. A primeira delas é a “Extract”. Essa é a fase em que os dados são extraídos dos OLTPs e conduzidos para a “Staging Area”, onde são convertidos para um único formato. No caso, a sigla OLTP significa Online Transaction Processing, ou Processamento de Transações em Tempo Real e representam os sistemas que se encarregam de registrar todas as transações contidas em uma determinada operação organizacional. Geralmente possuem bom desempenho em manipulação de dados operacionais, mas são ineficientes para análises gerenciais. A chamada Staging Area é uma área onde são colocados os dados após a extração a partir dos sistemas de origem. É importante ressaltar que raramente as staging area são normalizadas. Elas existem para que se faça um processamento intermediário e reduza a sobrecarga de acessos aos sistemas fontes. Por isso é dedicada apenas para a fase ETL e não é disponibilizada para os usuários finais. Ou seja, relatórios não podem acessar seus dados.

Na segunda parte, chamada “Transform” realizamos os devidos ajustes, podendo assim melhorar a qualidade dos dados e consolidar dados de duas ou mais fontes. Existem vários tipos de ajustes, como por exemplo:

- Tradução de valores codificados.
- Derivação de um novo valor calculado.
- Resumo de várias linhas de dados.
- Geração de valores de chaves substitutas.
- Correção de diferenças de unidades de medidas e precisão.

Enfim, a terceira parte, chamada de “Load”, consiste em fisicamente estruturar e carregar os dados para dentro da camada de apresentação seguindo o

modelo dimensional. Antes de carregar os dados o banco de dados precisa ser modelado. A ferramenta escolhida para armazenamento depende da modelagem, do tipo de dados, da quantidade disponível, entre uma série de outros aspectos.

As etapas de extração e carga de dados são consideradas obrigatórias, pois sempre será necessário transferir o dado para uma base onde será realizada a análise. Porém a etapa de transformação é opcional, ou seja, é possível que os dados já estejam organizados de forma que não demande um ajuste extra.

Armazenamento dos Dados

Data Warehouse

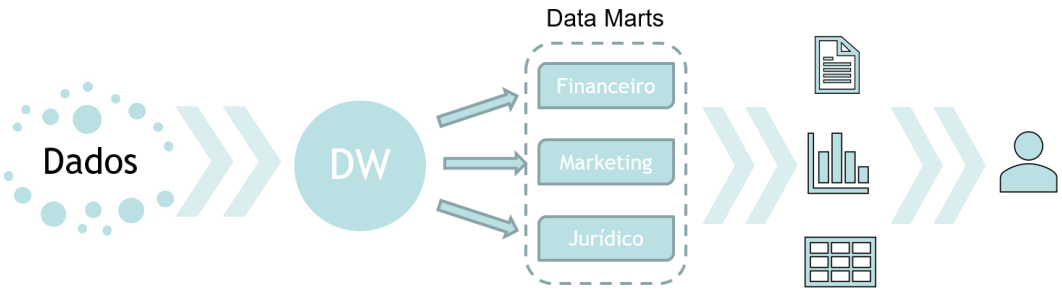
Quando se fala em análise de dados, constantemente se fala também em uma base de dados que seja capaz de armazenar, estruturar e processar consultas de forma eficiente para alcançar os resultados. O tipo de base de dados mais comum é o Data Warehouse (DW). De acordo com William H. Inmon, “um data warehouse é um conjunto de dados baseado em assuntos, integrado, não volátil e variável em relação ao tempo, de apoio às decisões gerenciais”. Analisando cada detalhe dessa definição podemos perceber que um DW é orientado a assuntos, pois por exemplo, aborda detalhes específicos de vendas de produtos a diferentes tipos de clientes, ou atendimentos e diagnósticos de pacientes. Também é integrado, pois possui diferentes nomenclaturas, formatos e estruturas das fontes de dados que precisam ser acomodadas em um único esquema para prover uma visão unificada e consistente da informação. Um DW é considerado não volátil, pois os dados de um data warehouse não são modificados como em sistemas transacionais (exceto para correções), mas somente carregados e acessados para leituras, com atualizações apenas periódicas. Por fim, é variável em relação ao tempo, pois possui um histórico de dados de um período de tempo superior ao usual em bancos de dados transacionais, o que permite analisar tendências e mudanças. Veja na Figura 11 o diagrama de representação de um DW.

Figura 11 – Data Warehouse.

Data Mart

Um Data Warehouse pode possuir informações de diversas áreas. O que pode ter suas vantagens e desvantagens. Considerando que várias análises são realizadas dentro de um contexto mais específico, pode ser viável a criação de Data Marts. Um Data Mart refere-se a cada uma das partes de um Data Warehouse corporativo. É um subconjunto do DW que contém os dados para um setor específico da empresa, ou seja, corresponde às necessidades de informações de uma determinada comunidade de usuários. Veja na Figura 12 o diagrama de representação de um Data Mart.

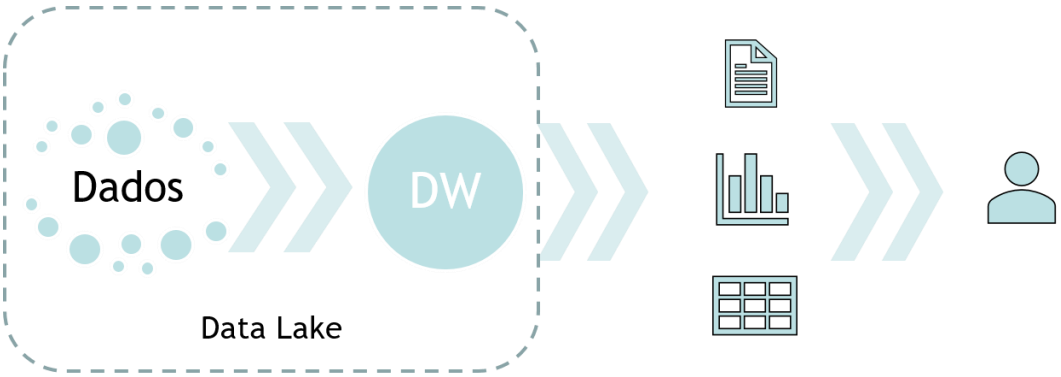
Figura 12 – Data Mart.



Data Lake

Diferentemente do Data Mart, que é como uma especialização do Data Warehouse, o Data Lake preza por um único repositório, com os dados brutos e que estejam disponíveis para qualquer pessoa que precise realizar uma análise. De acordo com James Dixon, “são os dados em grandes volumes e em seu estado natural, vindos de todos os tipos de fontes, onde os usuários podem “mergulhar” e tirar amostras. Um lago cheio de dados”. Esse lago cheio de dados pode novamente levantar discussões importantes sobre segurança e privacidade de dados. Veja na Figura 13 o diagrama de representação de um Data Lake.

Figura 13 – Data Lake.



Importante observar que existem algumas diferenças fundamentais que podem influenciar bastante na decisão de qual estrutura de dados será adotada. Por exemplo, o processamento dos dados em um Data Warehouse ocorre no momento da escrita dos dados. Já no Data Lake, ocorre no momento de leitura. Se o Data Lake promove uma flexibilidade maior com relação à disponibilização e configuração, existe um preço que pode ser um tempo maior se comparado ao DW quando são realizadas as consultas na base de dados.

Os Quatro Tipos de análises

A etapa do processo que cria mais expectativa é a análise de dados. É nessa etapa que os dados são traduzidos em informações úteis para o usuário. E são várias

as análises que podem ser feitas quando os dados já estão prontos, pré-processados e filtrados. Existem quatro principais tipos de análises que podem executadas, e cada tipo de análise tem seu próprio escopo e sua própria finalidade. Talvez seja possível inclusive conhecer a maturidade em Data Analytics de uma determinada empresa baseado em quais análises fazem parte de sua realidade.

Descritiva

A análise descritiva utiliza os dados históricos para identificar o que aconteceu. São análises que possuem um alto valor significativo por serem de fácil consumo e por gerar uma visão clara da situação atual da empresa no presente momento.

Seu objetivo é conhecer os dados e identificar os padrões. Esse tipo de modelo permite que sejam compreendidos eventos em tempo real e auxilia em tomadas de decisões imediatas. Logo, o tipo de pergunta que será respondida por essa análise é: “O que aconteceu?” ou “O que está acontecendo?”.

Diagnóstica

É uma análise que vai além da explicação do que aconteceu. Essa análise tem objetivo de avaliar os impactos das ações tomadas. Ou seja, o foco está na relação entre os eventos e as consequências. Para muitas aplicações não basta apenas entender o que aconteceu, é preciso entender e explicar o que foi detectado, além de identificar quais fatores influenciaram o resultado atual. Por esse motivo, funciona bem em conjunto com análises preditivas. A principal pergunta que será respondida por essa análise é: “Por que aconteceu?”.

Preditiva

Talvez seja o tipo de análise mais conhecido no mercado. O objetivo é analisar dados relevantes ao longo do tempo, buscar padrões comportamentais, e

prever como será o comportamento no futuro, dadas as condições atuais. É indicada para quem precisa prever a probabilidade de ocorrência de algum evento.

Como já mencionado, o valor da análise aumenta quando utilizada em conjunto com a análise diagnóstica, pois o processo de entender o passado e compreender os impactos de uma determinada ação ou evento, auxilia uma determinada previsão realizada por essa análise. É um processamento que geralmente utiliza algoritmos de regressão, classificação e agrupamento. Além disso demandam um volume significativo de dados de boa qualidade e exigem um maior grau de sofisticação. Isso quer dizer que é necessária uma boa base histórica para um bom resultado. A principal pergunta que será respondida por essa análise é: “O que vai acontecer?”.

Prescritiva

De acordo com Gartner, apenas 3% das empresas fazem uso dessa análise. É um número extremamente baixo, mas é possível entender o motivo. É considerada a análise mais complexa, que demanda técnicas de Data Science e Analytics e ainda conhecimento específico do negócio e do mercado em questão. Essa é a análise que verifica quais serão as consequências de eventos futuros. Ou seja, permite que o usuário entenda o que pode acontecer caso tome alguma ação. Por isso, também conhecida como “Análise de Recomendação”. Em outras palavras, a análise preditiva identifica tendências futuras, a prescritiva traça as possíveis consequências de cada ação. A principal pergunta que será respondida por essa análise é: “O que fazer se for acontecer?”.

Streaming de dados

Uma questão fundamental de qualquer projeto de Análise de Dados ou BI é a periodicidade de coleta de dados. Dados coletados periodicamente proporcionam um controle muito maior ao analista sobre a quantidade de dados que se deseja processar, analisar e visualizar. Esse tipo de análise é chamado de análise em batch,

ou em lote. Porém é notório que existem diversas aplicações onde um processamento em batch não é suficiente. São diversos exemplos que podem ser citados como monitoramento de segurança, automação industrial, alguns data lakes e dispositivos *IoT.*

Em várias dessas aplicações, os dados são gerados em tempo real e em fluxo contínuo. É o chamado Streaming de Dados. Em função disso, existe a necessidade de processamento em tempo real. Nem sempre as técnicas utilizadas para processamento em lote conseguem atender às necessidades de streamings. Sistemas dessa natureza não podem ser tratados da mesma forma que sistemas em batch. O controle do usuário sobre a quantidade e periodicidade dos dados diminui, então essas aplicações que requerem processamento em tempo real estão empurrando os limites da infraestrutura de processamento de dados tradicional.

Com esse alto volume de dados e a necessidade de baixa latência no processamento, novas técnicas precisam estar disponíveis para o analista responsável. Uma delas é o chamado processamento “In-Stream”. Esse tipo de processamento assume que o sistema deve ser capaz de processar os dados sem a necessidade de gravar o dado em disco. É indicado para as aplicações em tempo real, onde a baixa latência é um requerimento básico, ou seja, à medida que os dados chegam, são processados e analisados.

IoT

Atualmente temos um cenário onde a maioria dos dados é gerada por seres humanos. Seja por uma ação ou por um clique, ainda é considerável a atuação humana na geração dos dados. Contudo, o cenário mostra uma tendência clara de mudança com o crescimento da Internet of Things ou Internet das Coisas (IoT). O termo se refere a grupos de dispositivos digitais que coletam e/ou transmitem dados pela internet. São televisões, sensores, lâmpadas, automóveis, celulares, objetos de todo tipo que possuem a capacidade de conectar à internet. Veja na Figura 14 exemplos de dispositivos IoT.

Figura 14 – Internet of Things.

A evolução da IoT traz impacto para diversas áreas. Com a produção em larga escala, o custo dos dispositivos será cada vez menor. Isso possibilitará o monitoramento em detalhes de qualquer processo, pessoa ou objeto. Será possível por exemplo monitorar ainda mais hardware e software de sistemas computacionais; monitorar de forma mais efetiva as redes sociais; utilizar todos os dados fornecidos através de celulares (geolocalização, preferências etc.); calibrar com mais precisão sistemas de recomendação e marketing; entre outras inúmeras possibilidades.

A consequência disso é um fluxo de dados enorme, como novos dados a cada segundo e uma grande necessidade de processamento em tempo real. Ou seja, a IoT talvez seja o maior contribuinte para as pesquisas relacionadas a streaming de dados. Novas técnicas devem surgir. Novas ferramentas serão necessárias. O conceito de Big Data será estendido.

CEP

O termo Complex Event Processing (CEP), ou Processamento de Eventos Complexos em português define um padrão arquitetural de software para

processamento de fluxos contínuos de grandes volumes de eventos em tempo real. É uma tecnologia que fornece o controle ao usuário para que possa programar uma ferramenta de análise de eventos para a observação de padrões e relacionamentos entre esses eventos. É possível utilizar a temporalidade, causalidade, dependência e composição. Com as regras criadas, o usuário consegue identificar os eventos relevantes dentro de uma série de eventos de menor valor. Uma grande vantagem desse criterioso filtro de eventos é o gerenciamento de alertas em tempo real. Apenas para os eventos pertinentes é que serão enviados e-mails, mensagens, notificações etc.

Existem diversas ferramentas disponíveis no mercado para implementar soluções de CEP. Geralmente as regras para descrição dos padrões desejados são definidas em uma linguagem de consulta sobre os fluxos de eventos, de forma similar ao SQL. Para exemplificar, considere uma indústria com monitoramento em tempo real do consumo de energia elétrica em seus equipamentos, com leitura a cada minuto. Obviamente o gestor da área tem interesse em ser informado caso o consumo seja muito acima do limite para um determinado dispositivo. Porém, é comum também a ocorrência de oscilações na rede elétrica e por consequência na leitura desse consumo. Caso o gestor seja notificado por cada leitura acima do limite pode acontecer dele receber diversas notificações irrelevantes por erros de leitura. Para esse caso podem ser criadas regras como por exemplo: Caso um medidor apresente 5 leituras consecutivas acima do limite, crie um evento de atenção. Caso esse evento de atenção aconteça por 10 vezes em um período de 24 horas, notifique o gestor responsável.

Processamento em tempo real exige técnicas e ferramentas diferentes do processamento em lote. O CEP é uma excelente alternativa para tratar o fluxo de eventos gerado pela IoT e evita que usuários se percam nessa abundância de informações disponíveis. Ou melhor, permite manter o foco naquilo que é importante.

De nada adianta uma análise perfeita sem uma boa apresentação do resultado. A visualização de dados é uma forma acessível e prática de ver os dados e entender diversos aspectos como exceções, tendências e padrões. É essencial para analisar as informações e tomar decisões. A escolha correta da visualização pode influenciar plenamente a interpretação. Isso porque nossos olhos são atraídos por cores e padrões. Ao invés de ler valores individualmente, como em tabelas ou texto, através de representações visuais podemos perceber e compreender inúmeros valores de uma só vez. São diversos tipos de recursos que podemos utilizar como gráficos, tabelas, Mapas, Infográficos, Painéis e assim por diante. A visualização de dados é uma maneira simples e rápida de transmitir conceitos de modo universal.

Existe um conceito extremamente importante que deve ser conhecido e utilizado sempre que possível, com a devida prudência. É o chamado “Processamento pré-atentivo”. O nosso cérebro consegue identificar alguns detalhes antes mesmo da nossa atenção consciente. Isso significa que, quando visualizamos algo, antes de percebermos a imagem como um todo, nosso cérebro já dedicou uma atenção especial a alguns elementos ou objetos. É um processamento rápido e paralelo, sobre o qual temos pouco controle. Isso é um artifício que determina quais objetos são importantes e pode facilitar a obtenção de insights por parte do usuário.

Para exemplificar, veja a Figura 15 e tente contar quantos algarismos “5” existem. Logo após tente fazer o mesmo com a Figura 16.

Figura 15 – Números sem processamento pré-atentivo.

12768679489326456584791209193021483490386
24814001480912808401209475283758237503407
67465748572308402394083590235803275904376
49679024376043765096730964036753067034760
37603760934706734096709347609430697039462
09765902347306047307603476034076034650967

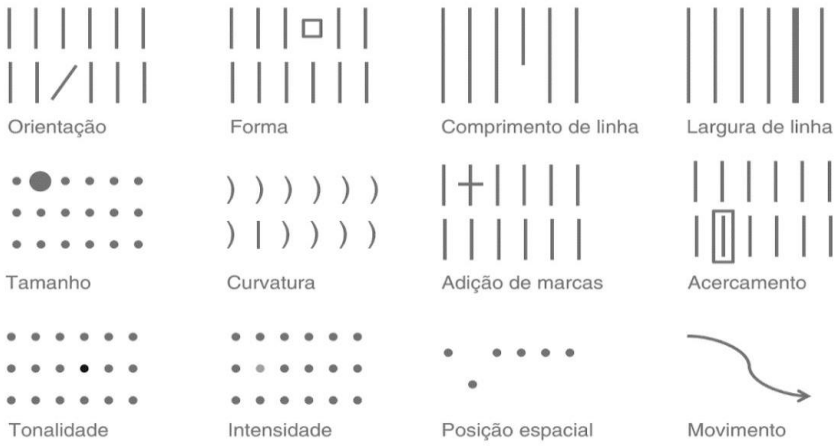
Figura 16 – Números com processamento pré-atentivo.

12768679489326456584791209193021483490386
24814001480912808401209475283758237503407
67465748572308402394083590235803275904376
49679024376043765096730964036753067034760
37603760934706734096709347609430697039462
09765902347306047307603476034076034650967

Com certeza a figura 16 colaborou para a realização dessa simples tarefa. Pelo fato de atribuir uma cor diferente aos valores que realmente importam, o cérebro humano os identifica com muito mais facilidade. Porém deve-se ficar atento para que os destaques sejam coerentes. Por exemplo, caso um outro número que não fosse o 5 também estivesse com a cor diferente, a maioria dos leitores poderia se confundir na contagem.

Diversos recursos pré-atentivos podem ser utilizados. Veja na Figura 17 alguns exemplos.

Figura 17 – Recursos pré-atentivos.



Outro detalhe significativo que deve ser observado ao criar uma visualização de dados é a tipografia, ou seja, a fonte do texto que será usado. Há uma imensidão de opções disponíveis. Um detalhe bem conhecido é a serifa. Fontes serifadas possuem pequenos “pés” ou linhas no acabamento das suas extremidades. As fontes

não serifadas não possuem esse detalhe. É comum haver discussões sobre por exemplo o melhor tipo de fonte para cada caso, ou então qual tipo de fonte dá mais legibilidade ao texto. Veja na Figura 18 exemplos de fontes com e sem serifa, favorecendo ou não a legibilidade.

Figura 18 – Exemplos de fontes.

Boa Legibilidade Com serifa	Boa Legibilidade Sem serifa	Difícil Legibilidade Com serifa	Difícil Legibilidade Sem serifa
Times New Roman	Arial	STENCIL	Britannic Bold
Palatino Linotype	Verdana	Baskerville Old Face	Papyrus
Courier New	Tahoma	*Monotype Corsiva*	PT Sans Narrow

Outro tópico a ser tratado é a identidade visual da apresentação dos dados. Dificilmente um resultado de análise terá apenas um gráfico ou uma tabela. O comum é que seja construído um conjunto de elementos em um ou mais painéis ou dashboards. Essa composição de objetos, figuras, textos e tabelas deve seguir um padrão que seja identificável, por exemplo, com um conjunto de cores.

Capítulo 3. Frameworks e Ferramentas

Nesse capítulo serão apresentadas diversas ferramentas conhecidas e consolidadas no mercado, além de conceitos arquiteturais de aplicações Big Data.

Computação em nuvem

Computação em Nuvem, ou Cloud Computing em inglês, é um paradigma de infraestrutura de computação, com disponibilização através da internet de servidores que podem ser reconfigurados dinamicamente com relação aos seus recursos de memória, armazenamento e processamento, ou seja, com alta escalabilidade.

É um conceito recente que ganhou força com a evolução da internet e da capacidade de recursos computacionais disponíveis para criação de data centers. Junto com a flexibilidade de oferta de serviços da computação em nuvem aparecem diferentes desafios. Por exemplo, uma questão bastante comum é a segurança e privacidade dos dados. Todos os dados coletados podem ser disponibilizados através da nuvem e acessados de qualquer lugar. O que pode ser sensacional por um lado, pode ser desastroso por outro. Isso porque não haverá mais problemas de disponibilização e acesso, porém aumenta a chance de vazamento de dados críticos e dificulta o controle de acessos indevidos. Além disso, como todos os dados estarão online, outro desafio é o envio de tantos bytes para a nuvem. A infraestrutura disponível para envio e acesso deve suportar um fluxo cada vez maior de informações.

Uma vantagem da utilização de computação em nuvem é que não há necessidade de comprar hardware ou software para a implantação e utilização de suas aplicações. Ou seja, você paga somente pelos serviços que usa. Com isso há uma redução de custos. Outra vantagem é que os fornecedores de computação em nuvem já possuem a infraestrutura de data center pronta para entregar alta disponibilidade de seus servidores. São definidas estratégias de replicação e sincronização de dados para que nenhuma informação se perca caso ocorra alguma falha de hardware. Além disso, em um cenário de projeto de análise de dados, é

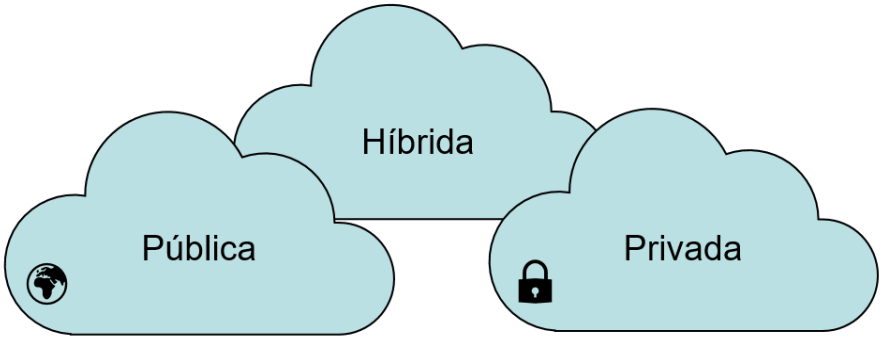
comum ter períodos de estudos e análises que não demandam processamento computacional. Com a computação em nuvem você paga por computação adicional somente quando necessário.

Modelos de Implantação

Existem vários tipos de nuvem que se diferenciam pela forma de disponibilização dos recursos. São as nuvens públicas, privadas e híbridas.

É importante deixar claro que não há necessidade de uma arquitetura diferente, ou modificações nos acessos. Pode existir uma diferenciação nos tipos de serviços oferecidos. Veja na Figura 21 os modelos de implantação.

Figura 21 – Modelo de Implantação.



- **Nuvem pública:** é o tipo mais comum. Nessa nuvem os recursos computacionais são compartilhados com outros usuários. Isso significa que o mesmo processador será usado para diferentes aplicações de diferentes usuários. Geralmente o acesso é feito por um navegador da Web. Por compartilhar recursos, é a opção mais economicamente viável, pois reduz significativamente os custos de manutenção.

- **Nuvem privada:** são servidores exclusivos para uma única empresa / organização. Por ser exclusivo, pode estar localizada fisicamente no datacenter local da empresa. Além disso, provê uma maior flexibilidade na configuração, pois não é compartilhado e o usuário tem a liberdade de

configurar conforme necessário. Também oferecem maior segurança e por isso são comumente usadas por órgãos governamentais, instituições financeiras e outras empresas com operações e/ou dados críticos.

- **Nuvem híbrida:** é a composição de duas ou mais nuvens. Com esse tipo de nuvem o usuário tem a possibilidade de alternar entre o modelo público e o privado conforme a necessidade do negócio. Com isso tende a oferecer o melhor custo-benefício. Assim como a nuvem privada, é indicada para empresas que possuem uma boa infraestrutura interna. Porém nesse caso também necessitam da nuvem pública

Tipologia

A computação em nuvem oferece serviços de diversos tipos. Os três principais são Software as a Service (SaaS), Infrastructure as a Service (IaaS) e Platform as a Service (PaaS).

- **SaaS:** a aplicação funciona diretamente na nuvem. Não há necessidade de instalação na máquina do cliente. É o tipo mais utilizado por atingir o maior número de usuários. Por exemplo: E-mail, calendário, Google Docs, Office 365.

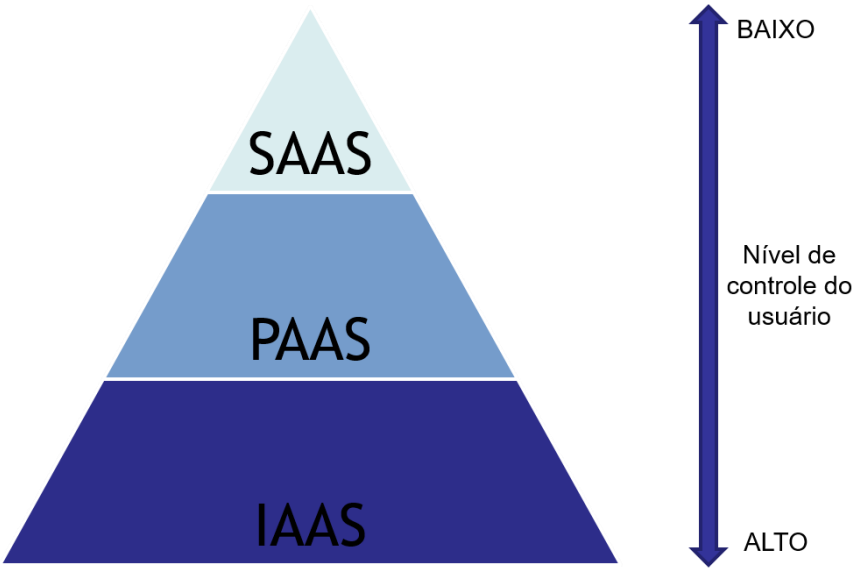
- **IaaS:** disponibiliza os recursos necessários para que o próprio usuário faça a implantação, configuração e utilização de suas aplicações. O usuário geralmente é um administrador de rede, ou analista de infraestrutura e tem a liberdade de provisionar o ambiente da forma que achar necessário. É comum para hospedagem de sites e também para empresas que precisam de um servidor remoto, com alta disponibilidade para implantar sua aplicação. Para isso é possível comprar servidores por exemplo na Azure, AWS e Google Cloud.

- **PaaS:** o provedor fornece uma plataforma que é usada para desenvolver e disponibilizar aplicações. Não é um software pronto para uso final, mas é um

arcabouço para o desenvolvimento. Por isso é comum que os usuários sejam programadores e desenvolvedores de software. É um meio termo entre SaaS e IaaS. Um exemplo é o Google App Engine, que permite a criação e disponibilização de aplicativos.

Cada tipo de oferta de computação em nuvem tem a sua aplicação. É importante perceber que o nível de controle do usuário pode aumentar ou diminuir conforme o serviço. Veja na Figura 22 a relação entre os serviços.

Figura 22 – Tipologias para Computação em Nuvem.



Ainda existem diversos outros tipos como CaaS (Comunicação), HaaS (Hardware), SECaaS (Segurança) ou BDaaS (Big Data). Independentemente do tipo de oferta, sempre teremos um serviço útil sendo oferecido ao usuário para que ele possa usufruir das vantagens de servidores acessíveis de qualquer lugar através da internet.

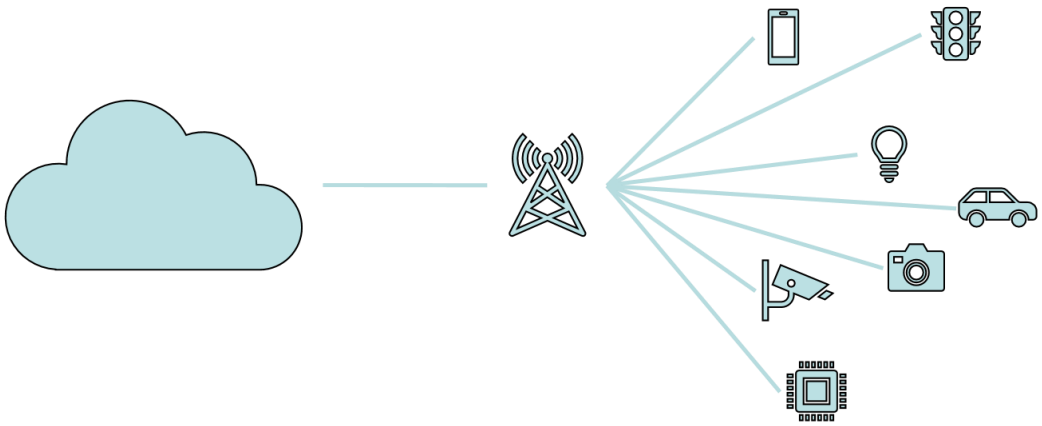
Edge Computing

O conceito de IoT apresenta dispositivos que possuem a capacidade de se conectar à internet para enviar ou receber dados. Para a Edge Computing, ou Computação em Borda em português, esses dispositivos são ainda mais qualificados.

Nesse contexto, os dispositivos são capazes de realizar análises e processamentos avançados. É um modelo distribuído que pode evitar a necessidade do envio de dados para uma nuvem remota ou um sistema centralizado para realizar o processamento. Dessa forma facilita também processamentos de dados em tempo real. É uma grande tendência para o futuro da análise de dados. Existem alguns especialistas que dizem sobre o fim da Cloud Computing. Na verdade Edge e Cloud são complementares. Cada um tem suas vantagens e também suas limitações. Por exemplo, ainda existe dificuldade de realizar processamentos pesados em dispositivos de borda. Seja por limitações de recursos ou por disponibilização de dados.

Um exemplo de aplicação que depende muito da análise distribuída é o desenvolvimento de veículos autônomos. Não é viável construir um carro que envie todas as informações dos seus sensores e câmeras para um servidor e depois decida a ação a ser tomada. Veja na Figura 23 outros exemplos de dispositivos para Edge Computing.

Figura 23 – Dispositivos Edge Computing.



Todas essas vantagens mencionadas mostram que Computação em Nuvem e BI são uma ótima combinação. A evolução da Edge Computing tende a agregar cada vez mais valor às análises. Com isso os dados se tornam mais acessíveis, é fornecida uma alta disponibilidade de dados e servidores, há uma redução de custo com infraestrutura de TI e o analista pode focar no que é realmente necessário.

Processamento paralelo e distribuído

Com a intenção de utilizar cada vez melhor os recursos computacionais disponíveis, pode-se utilizar técnicas de processamento paralelo e/ou distribuído. Essas técnicas utilizam melhor o poder de processamento e por isso podem apresentar um melhor desempenho.

Entretanto, não é qualquer tipo de problema que pode ser resolvido com esses recursos. Para que os ganhos sejam reais, é preciso que os problemas tenham as seguintes características:

- Podem ser particionados em subproblemas ou unidades de trabalho que podem ser resolvidas simultaneamente;
- Podem executar múltiplas instruções a qualquer momento no decorrer da resolução do problema;
- Podem ser resolvidos em menor unidade de tempo com múltiplos recursos computacionais do que com um único recurso computacional.

O processamento paralelo é diferente do processamento distribuído em diversos aspectos. Pode-se dizer que Programação Paralela é a prática de dividir uma determinada tarefa em tarefas menores que possam ser executadas de forma simultânea e independente. Porém há apenas uma máquina e todas essas subtarefas são executadas no mesmo servidor. Veja nas Figuras 24 e 25 o particionamento de um problema.

Figura 24 – Particionamento para execução sequencial.

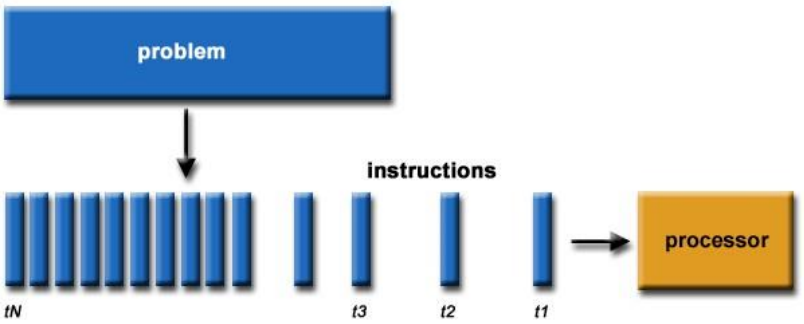
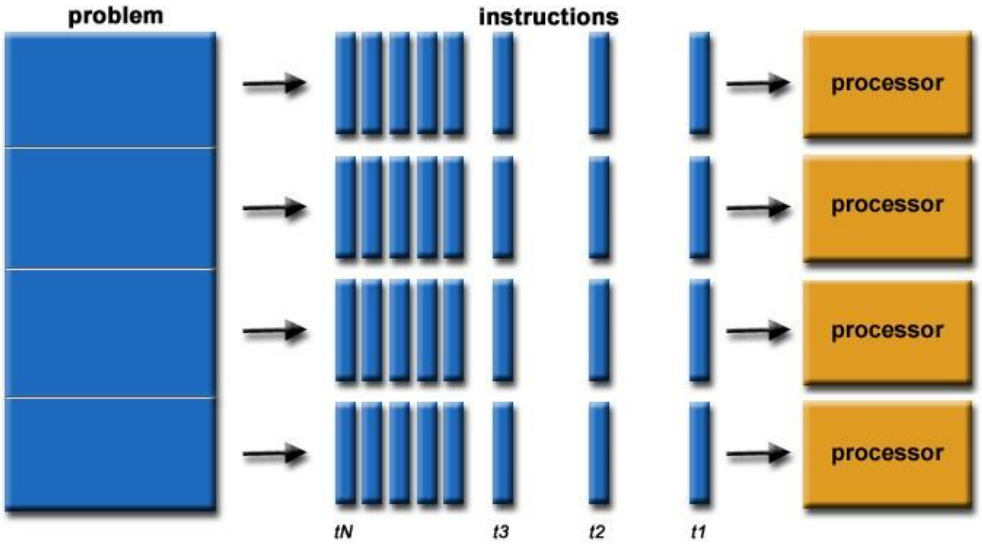


Figura 25 – Particionamento para execução em paralelo.



Ao particionar um problema em tarefas menores é preciso ficar atento ao nível de decomposição e à granularidade obtida. Se forem criadas muito mais subtarefas do que a quantidade de processadores disponíveis ocorrerá uma alta concorrência por recursos e a execução pode ficar até mesmo mais lenta que a execução sequencial, que ocorre quando se utiliza apenas um processador. Isso acontece devido ao evento chamado “troca de contexto” do processador. Cada vez que uma subtarefa termina a execução, antes de iniciar a próxima o processador deve enviar os dados para os registradores, coletar os dados para a nova tarefa, capturar o conjunto de instruções e só então começar o processamento.

Para avaliar se o algoritmo paralelo possui um bom desempenho, deve-se comparar o tempo de execução em paralelo com o tempo de execução sequencial.

Existe uma métrica chamada SpeedUp, que é a razão entre esses dois valores:

$$S(p) = \frac{T(1)}{T(p)}$$

Sendo:

$T(1)$ = Tempo de execução com *um* processador.

$T(p)$ = Tempo de execução com *p* processadores.

A partir do SpeedUp é possível determinar a eficiência do algoritmo paralelo desenvolvido, que é a razão entre o desempenho e os recursos computacionais disponíveis. Em outras palavras, mede o grau de aproveitamento dos recursos:

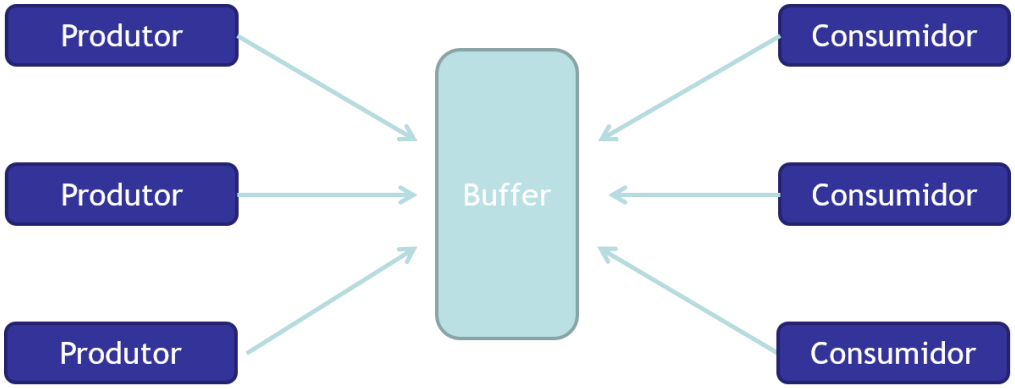
$$E(p) = \frac{S(p)}{p} = \frac{T(1)}{p \times T(p)}$$

Sendo $S(p)$ o SpeedUp para p processadores.

Existem diversas outras técnicas e métricas para avaliação de desempenho de algoritmos paralelos como redundância, grau de utilização, qualidade etc.

Uma arquitetura comum quando se trabalha com programação paralela é a chamada produtor consumidor. Nessa estrutura, dois processos compartilham um *buffer* de informações. O produtor é responsável por inserir os dados no *buffer* e o consumidor remove esses dados para utilizar como quiser. É comum ser utilizado quando uma tarefa A deve esperar completar a tarefa B antes de continuar sua execução. Veja na Figura 26 o esquema de funcionamento.

Figura 26 – Produtor Consumidor.



A programação distribuída é a habilidade de propagar o processamento de uma determinada tarefa através de múltiplas máquinas físicas ou virtuais interligadas por serviços de rede. Isso significa que na prática teremos diversos servidores em um *cluster* trabalhando em conjunto para resolver uma tarefa.

Uma grande vantagem é a escalabilidade que pode ser oferecida para o processamento dos dados. Mas também insere uma complexidade extra de recursos de rede que devem ser analisados. Outro desafio é a replicação e sincronização dos dados. Como são vários servidores trabalhando em conjunto, é preciso definir uma estratégia para que todos tenham um acesso rápido aos dados. E caso algum dado seja alterado, todos os outros servidores do cluster devem perceber essa alteração.

Para a programação paralela ou distribuída, é um desafio desenvolver, gerenciar e manter o sistema, se comparado com programação sequencial. É preciso modelar bem o problema para facilitar o controle concorrente de acesso aos dados e recursos. Além disso, deve-se criar estratégias para que o sistema se comporte bem caso aconteçam falhas de máquinas ou de rede.

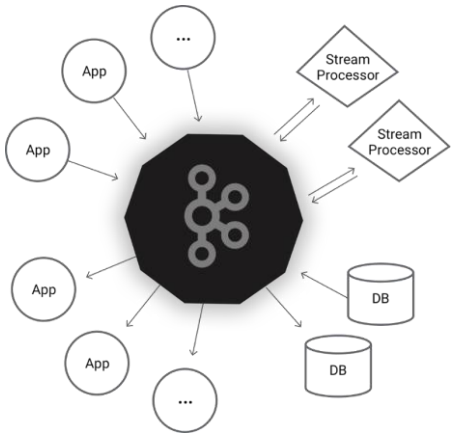
Apache Kafka

O Apache Kafka possui em seu próprio site uma definição simples e objetiva. “É uma plataforma distribuída de mensagens e streaming”. Essa ferramenta é bastante utilizada atualmente por facilitar integrações com diversas fontes diferentes de dados e suportar o fluxo contínuo. Assim, é possível publicar, armazenar, processar e consumir uma grande quantidade de dados. No Kafka todos os eventos podem ser resumidos em mensagens, através de tópicos. Uma mensagem pode ser um simples texto, valor numérico, e-mail, ou uma estrutura de dados mais complexa e estruturada como JSON. O Kafka permite que sejam criados esquemas de mensagens para controlar o tipo de dados nos fluxos. Com esse recurso o analista consegue melhorar a qualidade dos dados, remover valores inválidos e bloquear mensagens fora do padrão.

Para o funcionamento básico do Kafka, você produz uma mensagem que é anexada a um tópico. Em seguida você consome essa mensagem. Esse esquema de funcionamento é semelhante ao que foi estudado no tópico Processamento Paralelo e Distribuído.

A mensagem pode vir de diversas fontes como outras aplicações ou bancos de dados. E da mesma forma, o destino também pode ser várias aplicações ou bancos de dados. Veja na Figura 27 o diagrama básico do Apache Kafka.

Figura 27 – Apache Kafka.



Para o desenvolvimento de aplicações com o Apache Kafka, são disponibilizadas quatro principais API's:

- **Producer API:** permite publicar os dados em um ou mais tópicos
- **Consumer API:** permite a inscrição em um ou mais tópicos e consome os dados.
- **Streams API:** permite que uma aplicação seja um "processador de fluxo". Essa API lê de um tópico, realiza o processamento e escreve o resultado em outro tópico.
- **Connector API:** permite criar produtores e consumidores para tópicos do Kafka. Os conectores são reutilizáveis.

Para entender melhor o que são os tópicos, pode-se considerar que são filas onde os dados são registrados. Em outras palavras, são conjuntos de dados onde as mensagens são publicadas. Dessa forma é possível categorizar os grupos de mensagens. Cada tópico pode agrupar diversas mensagens e pode ter um ou mais consumidores. Cada consumidor terá acesso a todas as mensagens, mas não poderá alterar o estado da mensagem.

O produtor é o responsável por criar uma mensagem e enviar os dados para um tópico específico. Uma vez que uma mensagem é produzida em um tópico o próprio Kafka organiza a mensagem para garantir sempre a ordem das mensagens produzidas

O consumidor "assina" um tópico para que possa ter acesso aos dados armazenados. Como já mencionado, vários consumidores podem assinar o mesmo tópico, em grupo ou individualmente. Um grupo de consumidores permite que a leitura de dados seja realizada de forma mais rápida e dinâmica.

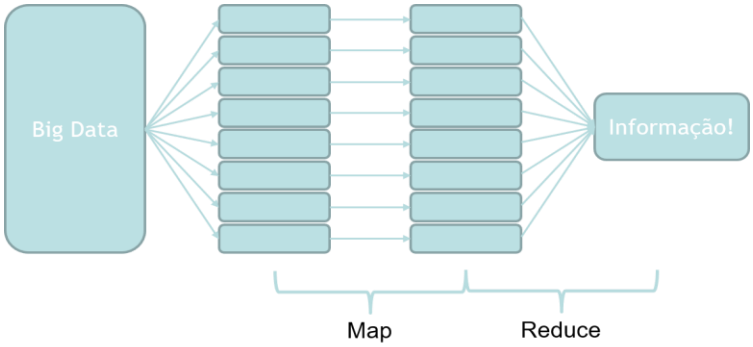
Novamente citando os dizeres do site do Kafka, "se você quer mover e transformar um grande volume de dados em tempo real entre diferentes sistemas, então Apache Kafka pode ser exatamente o que você precisa".

O Apache Hadoop é um framework para desenvolvimento de aplicações que necessitam de armazenamento e processamento distribuído de grandes conjuntos de dados. É uma ferramenta que conquistou espaço no mercado e mesmo com o surgimento de concorrentes sua popularidade continua em alta. Uma grande inovação que o Hadoop trouxe foi a implementação de um sistema de arquivos próprio, chamado Hadoop Distributed File System (HDFS). Esse sistema de arquivos foi projetado para abranger grandes clusters de servidores e escalar até centenas de petabytes e milhares de servidores.

Além de gratuito, o Hadoop com sua própria estrutura de armazenamento de dados pode ser usado em hardware de baixo custo. Com isso pode-se ter um ambiente com escalabilidade e flexibilidade sem pagar o preço de uma grande infraestrutura.

No Hadoop os dados são tratados como pares chave / valor (Key / Value) e processados por meio de duas funções principais, Map e Reduce. Essa estratégia pode tirar uma vantagem da localidade de dados. Veja na Figura 28 o esquema de funcionamento do Map Reduce.

Figura 28 – Map Reduce.



Com essa técnica, o processamento ocorre próximo ao armazenamento em cada nó no cluster, a fim de reduzir a distância que deve ser transmitido. Na função Map, os dados são separados em pares, distribuídos e processados. Esse processamento gera dados intermediários, que não estão mais em sua forma original,

mas estão prontos para serem enviados para a fase seguinte de redução. Na função Reduce os dados são agregados em conjuntos de dados, gerando a informação, ou o resultado do processamento.

Dessa forma o Hadoop tem a capacidade de processar grandes quantidades de dados estruturados e não estruturados armazenados no HDFS. Pode ser uma ótima solução para empresas que buscam reduzir seus custos de infraestrutura de TI e ainda capitalizar os benefícios do Big Data.

Apache Spark

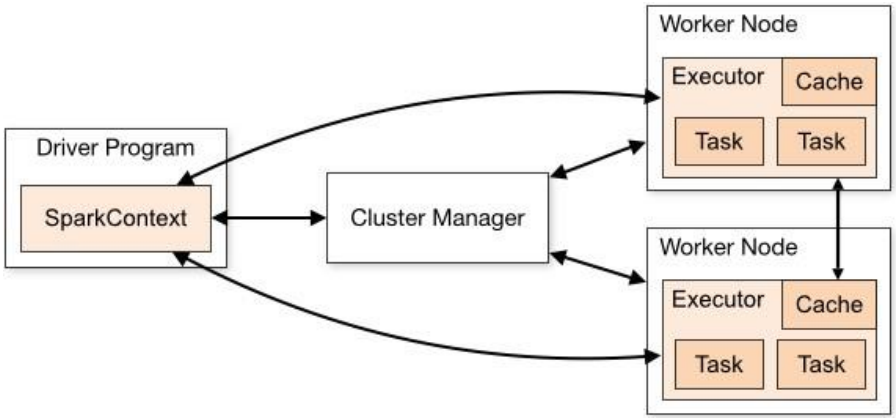
O Apache Spark é um projeto open source de um framework que estende o modelo de programação Map Reduce popularizado pelo Apache Hadoop. Sua principal vantagem, que traz muita competitividade é que o processamento é realizado em memória. Seu principal objetivo é ser veloz, tanto no processamento de queries quanto de algoritmos. É atualmente uma das ferramentas com maior potencial em Data Science e vem ganhando muita popularidade.

O próprio Spark possui ainda alguns recursos extras para auxiliar no desenvolvimento de aplicações que são:

- Spark Streamming: para processamento em tempo real;
- GraphX: ferramenta com algoritmos e técnicas para processamento sobre grafos;
- SparkSQL: permite que sejam utilizadas consultas SQL sobre os dados no Spark;
- MLlib: biblioteca com diversos algoritmos de aprendizado de máquina (Machine Learning)

O funcionamento do Spark pode ser resumido em um Driver Program, Cluster Manager e diversos Worker Node. Veja na Figura 29 a arquitetura do Apache Spark.

Figura 29 – Apache Spark.



O Driver Program é a aplicação principal. É responsável por controlar toda a execução e orquestração das tarefas. O Cluster Manager é responsável por administrar e gerenciar as máquinas que serão utilizadas como Worker. Ou seja, esse componente só é necessário se o Spark for executado de forma distribuída. Os Worker Nodes são os componentes que realmente executarão as tarefas enviadas pelo Driver Program.

O Spark tem a característica de usar muita memória RAM para o processamento e armazenamento de dados intermediários das iterações do algoritmo. Caso esse recurso não seja um problema o Spark é uma alternativa viável ao Hadoop. Porém ambas as ferramentas ainda prometem perdurar por bastante tempo, evoluindo a cada dia como solução para uma grande quantidade de aplicações de diversos tamanhos e características.

O Spark foi concebido com o principal objetivo de ser veloz, tanto no processamento de queries quanto de algoritmos, além de processamento em memória e eficiente recuperação de falha.

Bancos de dados relacionais e não relacionais (NoSQL)

Os bancos de dados relacionais são uma coleção de dados organizada de forma que um computador possa armazená-los e recuperá-los de maneira eficiente. Para isso os dados são organizados em tabelas.

Para um usuário usufruir dos recursos de um banco de dados são utilizados Sistemas Gerenciadores de Bancos de Dados (SGBD). Os SGBD's são o software para facilitar a manipulação das informações de um banco de dados. Eles possibilitam todo o controle de acesso de usuário, criação de views, visualização dos dados nas tabelas, geração de diagramas de relação etc. Os SGBD's também são responsáveis por gerenciar as transações de um banco de dados.

Cada tabela possui chave para identificação única dos registros. Através dessas chaves é possível criar relações entre as diversas tabelas. Essa é uma característica fundamental dos bancos relacionais e uma das mais importantes por garantir a integridade e consistência dos dados através das chaves estrangeiras.

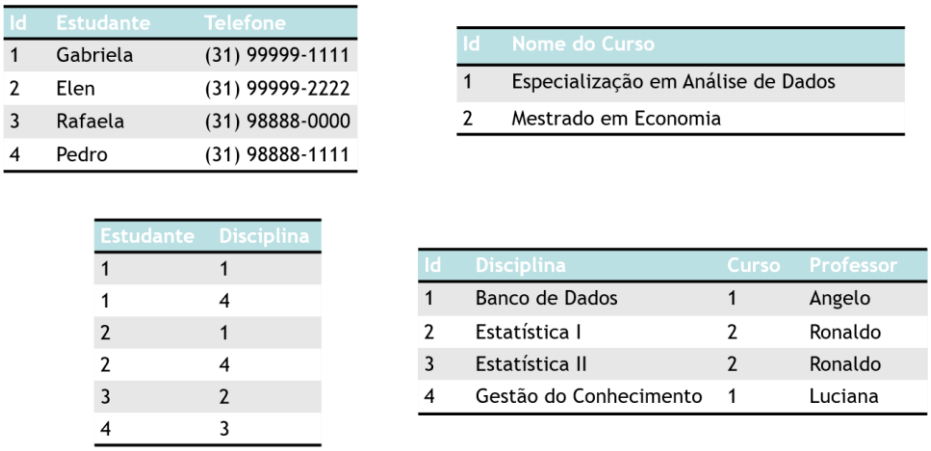
Para consultar os bancos relacionais, é utilizada o SQL (structured query language), que é uma linguagem baseada em álgebra relacional. Por isso foram herdadas por exemplo várias funções relativas a conjuntos.

Veja nas Figuras 30 e 31 dois exemplos de bancos de dados para uma mesma aplicação. É um cadastro de pessoas com seus respectivos telefones, além de informações de matrículas em disciplinas de um determinado curso. Na primeira figura podemos perceber que várias informações aparecem mais de uma vez, mostrando que a modelagem pode não ser a ideal para garantir a integridade dos dados.

Figura 30 – Exemplo de modelagem incorreta de um banco de dados.

Estudante	Telefone	Curso	Disciplina	Professor
Gabriela	(31) 99999-1111	Especialização em Análise de Dados	Banco de Dados	Angelo
Gabriela	(31) 99999-1111	Especialização em Análise de Dados	Gestão do Conhecimento	Luciana
Elen	(31) 99999-2222	Especialização em Análise de Dados	Banco de Dados	Angelo
Elen	(31) 99999-2222	Especialização em Análise de Dados	Gestão do Conhecimento	Luciana
Rafaela	(31) 98888-0000	Mestrado em Economia	Estatística I	Ronaldo
Pedro	(31) 98888-1111	Mestrado em Economia	Estatística II	Ronaldo

Figura 31 – Exemplo de modelagem correta de um banco de dados.



Id	Estudante	Telefone
1	Gabriela	(31) 99999-1111
2	Elen	(31) 99999-2222
3	Rafaela	(31) 98888-0000
4	Pedro	(31) 98888-1111

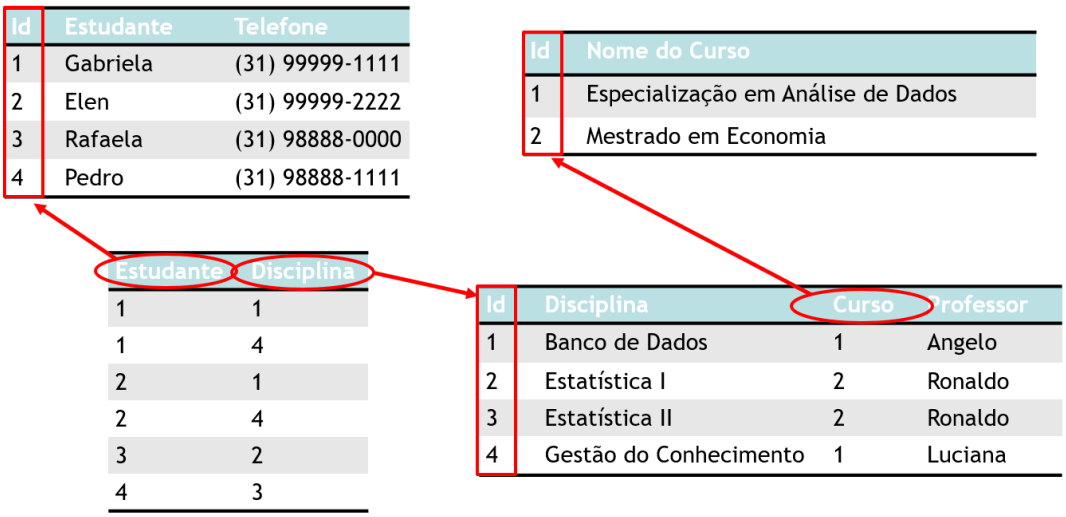
Id	Nome do Curso
1	Especialização em Análise de Dados
2	Mestrado em Economia

Estudante	Disciplina
1	1
1	4
2	1
2	4
3	2
4	3

Id	Disciplina	Curso	Professor
1	Banco de Dados	1	Angelo
2	Estatística I	2	Ronaldo
3	Estatística II	2	Ronaldo
4	Gestão do Conhecimento	1	Luciana

Na segunda figura dessa comparação, as informações não se repetem. É uma modelagem mais complexa, que garante a integridade dos dados. Nesse caso é dito que o banco de dados está normalizado. É possível destacar as relações dessa modelagem, como mostrado na Figura 32.

Figura 32 – Destaque para as relações do banco de dados.



Id	Estudante	Telefone
1	Gabriela	(31) 99999-1111
2	Elen	(31) 99999-2222
3	Rafaela	(31) 98888-0000
4	Pedro	(31) 98888-1111

Id	Nome do Curso
1	Especialização em Análise de Dados
2	Mestrado em Economia

Estudante	Disciplina
1	1
1	4
2	1
2	4
3	2
4	3

Id	Disciplina	Curso	Professor
1	Banco de Dados	1	Angelo
2	Estatística I	2	Ronaldo
3	Estatística II	2	Ronaldo
4	Gestão do Conhecimento	1	Luciana

Na Figura 32 é possível perceber que existem vários conceitos importantes quando se fala em banco de dados relacional. O primeiro deles é entidade. Nesse caso são quatro entidades. Uma para os estudantes, outra para os cursos, outra para as disciplinas e uma quarta exclusivamente para a relação entre estudante e

disciplina. Nesse caso, esse tipo de tabela exclusiva para representar uma relação é chamada de tabela de ligação.

O segundo conceito importante a ser percebido é que cada entidade possui diversos atributos, que são as colunas das tabelas. Por exemplo, a tabela dos estudantes possui três atributos. Um para identificação única do registro, o nome e o telefone. O terceiro conceito é o destaque da imagem, que são as relações entre as tabelas.

Os bancos de dados relacionais são tradicionais e atendem uma extensa variedade de aplicações. Porém, não foram projetados para tratar grandes quantidades de dados (Big Data). Além disso, existe uma dificuldade em se utilizar bancos de dados relacionais para tratar dados não-estruturados ou semiestruturados. Daí surgem os bancos de dados NoSQL.

A denominação NoSQL significa Not Only SQL, ou seja, não é somente SQL. Na prática os dados possuem uma liberdade de modelagem e estrutura muito maior. Essa estrutura depende da ferramenta que se é utilizada. Geralmente os bancos de dados NoSQL podem ser orientados a documentos, colunas, chave-valor ou grafos. A seguir serão apresentados exemplos dessas ferramentas.

MongoDB

O MongoDB é atualmente um dos principais bancos de dados NoSQL. É orientado a documentos, e armazena JSON em sua estrutura. Os documentos são organizados em coleções. A Figura 33 mostra o símbolo do MongoDB.

Figura 33 – MongoDB.

Fazendo uma analogia ao banco de dados relacional, as coleções seriam as tabelas, cada documento uma linha da tabela e cada informação dentro do JSON uma coluna, ou atributo. Documentos muito complexos prejudicam a performance do banco.

Um dos seus diferenciais é que funciona bem em cluster, ou seja, com vários servidores ao mesmo tempo. Com isso consegue oferecer uma alta disponibilidade dos dados. Além disso, possui conectores para Spark, facilitando a integração com ferramentas já populares no mercado. Um ponto de atenção é que costuma ocupar bastante espaço do disco rígido.

Cassandra

O Apache Cassandra é um banco de dados NoSQL que utiliza o armazenamento em colunas. A divisão dos seus dados se dá em KeySpace, família de colunas e a coluna. Cada uma dessas colunas possui o nome do campo, o valor e um timestamp. Timestamp é um instante, com data e hora. A Figura 34 mostra o símbolo do banco de dados Cassandra.

Figura 34 – Apache Cassandra.

No Cassandra não existe o conceito de transação como em bancos de dados relacionais. Por isso é importante a utilização do timestamp no momento de persistir os dados em disco. Também não existem os conceitos de relacionamentos e constraints. Foi popularizado por ser utilizado por grandes empresas como Twitter e Facebook. Para consultar os dados armazenados deve ser utilizada uma linguagem específica chamada Cassandra Query Language, ou CQL.

Neo4j

Este é um exemplo de um banco de dados NoSQL orientado a grafos. Possui em sua estrutura nós, arestas e propriedades. Para utilizar bancos de dados desse tipo deve-se encontrar uma aplicação onde a topologia dos dados é mais importante que os dados. A Figura 35 mostra o símbolo do Neo4j.

Figura 35 – Neo4j.

Importante destacar que nesse banco de dados é possível ter uma visualização do grafo construído e armazenado. Isso auxilia em análises e obtenção de insights dos relacionamentos entre os registros. É possível também agrupar nós que o usuário entende que pertencem a um mesmo domínio.

Couchbase

O banco de dados Couchbase é um exemplo orientado a chave-valor e documentos. Isso significa que existem vários documentos, que possuem em seu conteúdo dados JSON, e que são indexados através de chaves únicas. Os documentos são organizados em *buckets*. Veja na Figura 36 o símbolo do Couchbase.

Figura 36 – Couchbase.

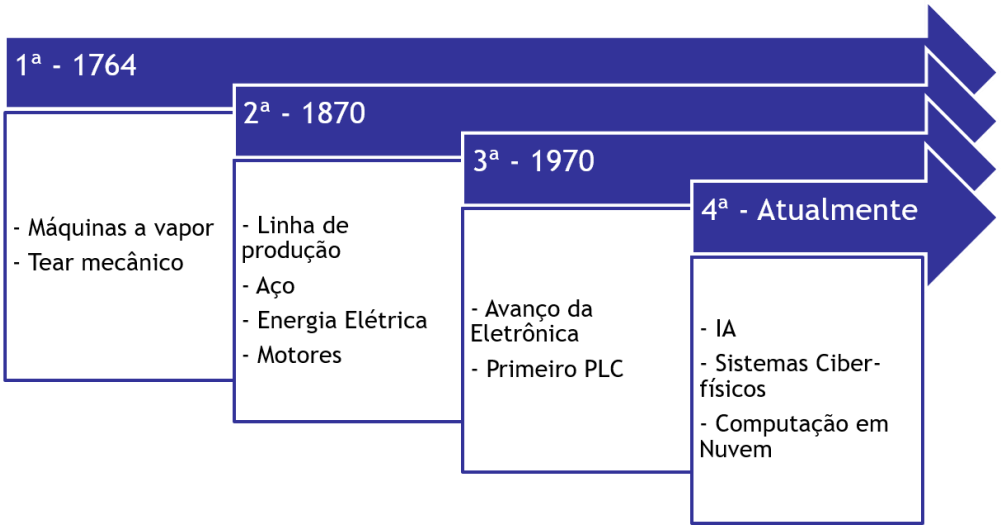
Esse banco de dados trouxe para sua estrutura um recurso já conhecido e discutido nesse curso, que é o conceito de Map Reduce. É possível utilizar os dados armazenados nos documentos para criar funções de mapeamento e redução e criar novas visualizações para os dados. Sua interface é bastante intuitiva, o que facilita que o administrador realize suas operações de gerenciamento do banco. Funciona muito bem em cluster, com várias estratégias de sincronização e replicação de dados entre os nós. Para consultar os dados armazenados é utilizada a linguagem N1QL que é um super conjunto da linguagem SQL.

Capítulo 4. Indústria 4.0

Cada revolução industrial trouxe uma evolução disruptiva para a indústria e para a sociedade. O cenário atual, chamado de “Era da Informação” trouxe um ambiente onde as possibilidades de digitalização, coleta de dados, processamentos e análises estão por todos os lados. Essa é a chamada indústria 4.0.

Veja na Figura 37 os principais marcos de cada revolução industrial. Em resumo a primeira revolução foi baseada em evoluções da mecânica. A segunda, da elétrica. A terceira da automação. E a quarta de um conjunto de fatores que englobam IoT, Robótica, Big Data, e todos esses fenômenos já discutidos.

Figura 37 – Indústria 4.0.



Pode-se dizer que a indústria 4.0 engloba as principais inovações tecnológicas dos campos de automação, controle e tecnologia da informação, aplicadas aos processos de manufatura. Isso permite uma evolução natural do monitoramento de operações em tempo real, com ações mais rápidas e correções mais efetivas nas linhas de produção. Com a digitalização, as empresas podem criar plantas virtuais, modelos de simulação de equipamentos, prever falhas e otimizar processos de produção. Ocorre ainda uma descentralização de decisões, com maior autonomia das máquinas diante do contexto entregue pelos sensores.

Com a indústria 4.0 a tendência é que haja uma conectividade cada vez maior entre áreas. As análises de dados devem cruzar as informações e os benefícios serão para todos. Com todo esse controle virtual sobre o processo, automaticamente haverá uma redução de custos de forma geral. Isso porque será mais fácil prever falhas desastrosas e os ajustes serão muito mais rápidos nas linhas de produção. Com agilidade nos ajustes, é possível personalizar produtos e atender cada vez melhor os clientes.

Esse cenário contribui também para o surgimento de novos modelos de negócio. Vários são os exemplos recentes de empresas inovadoras em mercados antigos. Por exemplo Uber, Airbnb ou Netflix.

Para acompanhar a evolução da indústria, a mão de obra deve se adaptar aos novos processos. A tendência é de redução de funções repetitivas e braçais, dando oportunidade para as pessoas dedicarem mais tempo às tarefas que exigem raciocínio e inteligência. Ou seja, o contexto exige novas habilidades e qualificações.

Assim como as três primeiras revoluções, a quarta é um caminho sem volta. Cada passo traz um novo desafio e uma nova oportunidade.

Cultura Data Driven

Cada empresa possui uma maturidade diferente com relação ao uso de Analytics em sua rotina e em seus processos. Várias empresas ainda estão descobrindo o verdadeiro valor dos dados e buscando alcançar o chamado *data driven*, ou orientação pelos dados. Para a evolução e utilização correta das informações disponíveis é necessário que se crie uma abordagem dos problemas de maneira objetiva. Definir qual é a meta e entender qual é o problema são os primeiros passos de empresas orientadas a dados. Com isso devem ser enaltecidos os líderes que tomam suas decisões baseadas em dados e fatos. Não devem ser tolerados líderes que utilizam apenas o achismo e intuições.

É comum que não existam muitos dados prontos e disponíveis. Por certas vezes a criatividade fará diferença na coleta e uso dos dados. As empresas precisam pensar fora da caixa. Mesmo quando já existe a possibilidade de disponibilização dos dados, ainda podem existir barreiras contra a chamada democratização dos dados. Esse conceito defende que os dados precisam estar disponíveis para qualquer pessoa. Todo mundo tem acesso a algum dado. Dessa forma, existirá uma transparência que em conjunto com as pessoas e ferramentas certas alcançará a maturidade que uma empresa precisa. Isso mostra como as pessoas são fundamentais nesse processo.

São as pessoas que tem a capacidade de deter o conhecimento do negócio, analisar os dados e tomar as decisões corretas. Por isso não podem faltar treinamentos e capacitações para que se incentive e desenvolva cada vez mais uma mentalidade orientada para análises.

Outro fator de influência na cultura é a comunicação, ou sensibilização de todos com relação à situação da empresa apresentada pelos dados. Para isso não basta apenas gerar relatórios, dashboards e alertas. É preciso um diálogo sincero que dê às pessoas o senso crítico de análise. Isso trará certamente o engajamento necessário para que as pessoas possam ouvir, debater e discutir diferentes pontos de vista, sempre com o objetivo de guiar a tomada de decisão, para ser a melhor possível.

Quando se fala em cultura orientada a dados, há muita discussão sobre a melhor ferramenta para digitalizar, visualizar, armazenar ou processar os dados. Cada caso deve ser analisado. Talvez até exista uma ferramenta mais indicada para o objetivo de uma determinada empresa. Porém a ferramenta ideal é aquela que resolve o problema.

Alguns erros comuns ainda acontecem frequentemente na busca dessa cultura. Talvez o mais frequente seja tomar uma decisão e logo depois buscar dados para apoiar essa escolha. O caminho correto é o inverso. Deve-se primeiramente buscar os dados, analisar e depois tomar a decisão correta.

Outra falha comum é não começar a cultivar a cultura por subestimar os dados que já estão disponíveis ou então achar que a análise é complexa demais. Muitas organizações já estão percorrendo esse caminho, os que ficarem de fora sofrerão as consequências. Além do erro comum de subestimar os dados, pode acontecer também a utilização de dados obsoletos, que já não fazem mais sentido. A coleta deve ser constante e a atualização das análises e modelos devem acompanhar na medida do possível esse progresso.

São várias as orientações de consultorias e empresas *data driven* para alcançar maturidade em Analytics. Por exemplo, é interessante deixar claro para todo o time a importância dos dados, mostrando a influência positiva do seu bom uso. Os gestores devem também buscar evoluir a capacidade analítica de sua empresa. Talvez um passo importante para isso seja automatizar a coleta de dados. O que por sua vez trará um volume muito maior e um novo desafio: organizar e estruturar os dados. Como já mencionado, as pessoas são peças fundamentais no processo, então os líderes devem buscar profissionais *data driven*, ou seja, profissionais que já possuem a mentalidade analítica, mesmo que não saibam disso.

Existem algumas perguntas que podem ser feitas para reflexão e para tentar constatar o nível de maturidade da cultura de uma organização:

- As decisões são tomadas com base em dados?
- O time sabe o que significa data driven?
- São utilizadas ferramentas que possibilitam sempre capturar novos dados e melhorar sua análise?
- Os gestores justificam suas decisões com base em dados?
- Estão sendo contratados profissionais data driven?

Diversos aspectos influenciarão a maturidade e cultura organizacional. É importante que os líderes percebam que a cultura é construída pelas pessoas. Com o time certo ficará mais fácil seguir na direção certa.

O Cientista de Dados moderno

Uma das profissões mais desejadas do século é o Cientista de Dados. Muitas empresas já estão criando áreas internas dedicadas à análise de dados e procurando no mercado esse tipo de profissional. As principais atividades desempenhadas por um cientista de dados são:

- Coletar e transformar os dados.
- Programar em diversas linguagens (Python, R etc.).
- Realizar análises estatísticas.
- Construir modelos (Machine Learning, Data Mining etc.).
- Ser uma “ponte” entre áreas internas da empresa.
- Detectar tendências.

Facilmente é possível perceber que não é trivial encontrar um profissional que possua todas essas habilidades. São tarefas que exigem conhecimento técnico profundo, ou *hard skills*, como Matemática, Estatística, Ciência da Computação e todos aqueles tópicos discutidos sobre Ciência de Dados. Além disso, um Cientista de Dados também deve possuir *soft skills* como comunicação, liderança, conhecimento do negócio, criatividade, pensamento analítico, capacidade de resolução de problemas e assim por diante.

Apesar do desafio de desempenhar tal função, muitas pessoas estão correndo atrás da formação necessária para se tornarem Cientistas de Dados. O ponto de atenção nesse caso é achar que o aprendizado é fácil e rápido. Em nenhuma profissão á assim. Para uma área tão multidisciplinar a situação é mais crítica ainda. Requer tempo, experiência e dedicação. Por isso não adianta aprender muitos conceitos ao mesmo tempo. É preferível focar em qualidade e não em quantidade. Começar por problemas simples ajudará a fixar os conhecimentos, construir experiências e motivar a dar um passo de cada vez. Outro cuidado é não focar apenas

na programação. Não basta aprender Python ou R. A função de Cientista de Dados exigirá muito mais que um bom programa de computador.

Devido à complexidade da função e dificuldade de encontrar pessoas prontas para o desempenho completo, são criados diversos papéis dentro de projetos de BI ou de análises de dados como por exemplo:

- Gerente de BI.
- Projetista de ETL.
- Analista Programador ETL.
- Analista Programador OLAP.
- Cientista de Dados.
- Engenheiro de Dados.

Esse grupo de papéis e pessoas trabalhando em conjunto traz desafios para a gestão do projeto, mas pode trazer muito mais benefícios também. Dando autonomia na medida certa, a diversidade de formações e pensamentos alimentará discussões com pontos de vistas completamente diferentes. Claro que o gestor deve se preocupar com a falta de alinhamento e padronização. Mas definindo bem o papel de cada pessoa, será mais fácil explorar o potencial individual de cada membro da equipe. O importante é manter a equipe com foco no projeto e no resultado.

Big Cases – Data Analytics

Diversas empresas já estão colhendo os frutos da utilização dos dados a seu favor. Nesse capítulo veremos alguns exemplos de cases de sucesso envolvendo Big Data, Machine Learning e outros recursos. Os cases são de empresas bem conhecidas no mercado, que dispensam apresentações. Foram selecionadas organizações de diversos nichos de mercado para mostrar que qualquer área pode se beneficiar da análise de dados.

Liverpool FC

O Liverpool Football Club é um clube de futebol profissional de Liverpool, Inglaterra, que compete na Premier League, a primeira divisão do futebol inglês. Internamente, o clube conquistou dezenove títulos da Liga, sete FA Cup, um recorde de oito Copas da Liga e quinze FA Community Shields. Nas competições internacionais, o clube conquistou seis Copas da Europa, mais do que qualquer outro clube inglês, três Copas da UEFA, quatro Supertaças da UEFA (também recordes da Inglaterra) e uma Copa do Mundo de Clubes da FIFA.

Dentro e fora de campo, o trabalho com o time de Analytics em conjunto com o time de Pesquisa contribuiu diretamente para as conquistas recentes do clube. Inclusive a contratação do técnico Jurgen Klopp, do meio campo Keyta e do atacante artilheiro Mohamed Salah, por exemplo, foram frutos de cruzamentos de bases e sugestão do time de dados.

O trabalho de análise realizado no Liverpool se tornou um modelo recrutamento de sucesso em todo o ecossistema do futebol mundial.

Copa do Mundo

Após as derrotas consecutivas nas copas de 2002, 2006 e 2010, a seleção alemã resolveu fazer algo diferente. Um grupo de 50 alunos de uma universidade começou a realizar pesquisas e análises dos jogos de futebol. Eram analisados vários fatores como número de toques, tempo médio de posse de bola, velocidade de movimento, mapas de calor e distância percorrida. Todos esses dados eram coletados através de oito câmeras posicionadas estrategicamente. Em apenas 10 minutos, 10 jogadores com três bolas geram mais de sete milhões de dados.

Além da atuação de cada jogador, eram analisadas também estratégias de jogo, tanto da seleção alemã, quanto dos seus adversários. Dessa forma pode-se criar técnicas diferentes para encaixar cada tipo de jogo, de acordo com o oponente.

O resultado da análise e das ações tomadas foi que o tempo médio de posse de bola por jogador caiu de 3,4 segundos em 2010 para 1,1 segundo em 2014. Essa e outras melhorias fizeram com que a seleção alemã fosse a grande campeã da copa do mundo e ainda deixou a seleção brasileira com o trauma histórico do famoso 7x1.

Stranger Things

A série Stranger Things se tornou um fenômeno de produção da Netflix, que é uma empresa que fornece serviço de streaming de vídeos pela internet e está presente em mais de 130 países e possui mais de 75 milhões de assinantes. Alguns dados divulgados pela empresa mostram que existem usuários que chegam a consumir 45 GB de dados por mês. Com todo esse ambiente disponível para coleta de dados, é possível realizar diversos processamentos.

A Netflix soube aproveitar bem a situação e desde que nasceu se preocupou em conhecer o comportamento do seu consumidor. Isso significa que para cada usuário, é conhecido o horário de preferência para assistir os vídeos, qual o gênero favorito, entre outros detalhes do perfil.

A série possui elementos que foram frutos de cruzamentos de dados que mapearam a receita da audiência. Isso fez com que a direção da série pudesse trazer referências para todas as gerações, personagens, cenários, enredos e histórias que geram identificação em qualquer tipo de público.

Mas a empresa não parou por aí. A maioria dos vídeos selecionados para visualização hoje foi a partir de recomendações da plataforma. Um sistema de recomendação tão eficiente como esse é referência para o mercado e está em constante evolução.

Para conhecer cada vez melhor seus clientes a empresa criou uma competição e ofereceu US$ 1 milhão para quem desenvolvesse um algoritmo capaz de prever um comportamento específico dos seus clientes e provasse que sua solução era mais eficiente do que o modelo que a companhia utilizava. Soluções como

essa só são possíveis através da coleta de cada ação do usuário, com cruzamento de perfil, preferências e hábitos.

Walmart

A Walmart Inc. é uma empresa multinacional americana de varejo que opera uma rede de hipermercados, lojas de departamentos de descontos e mercearias, com sede em Bentonville, Arkansas. A empresa foi fundada por Sam Walton em 1962 e incorporada em 31 de outubro de 1969.

Análises no DW apontaram certa relação entre fraldas e cervejas, mas variáveis eram aparentemente desconexas. No entanto, a decisão de aproximação física dos produtos foi tomada e isso gerou um estouro de vendas, tornando o case um modelo de analytics e tomada de decisão orientada por dados.

Airbnb

A Airbnb é um aplicativo para aluguel de imóveis que chega a hospedar 2,5 milhões de pessoas em uma noite. Com todos esses dados a plataforma consegue cruzar os dados dos hóspedes, anfitriões, com datas comemorativas e prever exatamente a procura para qualquer época do ano.

Possui um sistema de recomendação baseado no perfil do usuário que consegue avaliar o histórico do hóspede, o tipo de imóvel de preferência, quais os preços indicados, entre outros.

Um fato interessante é que ao cadastrar um novo imóvel na plataforma, existe um algoritmo de Machine Learning que analisa todas as características e sugere um preço para o aluguel. Existe um estudo que mostra que caso o anfitrião aceite a sugestão de valores da empresa, a possibilidade de alugar o seu espaço cresce em quatro vezes.

Netshoes

A Netshoes é um site para comércio eletrônico de produtos esportivos. Com o rápido crescimento da plataforma, surgiram diversos desafios para promoções, precificação e personalização de produtos. Com a contratação de serviços de consultoria e análises de dados várias soluções internas foram apresentadas para a melhoria dos processos.

Um dos resultados é o monitoramento ativo das ações dos usuários, gerando um perfil de uso e preferências. Com isso a Netshoes sabe quando um usuário realizou última compra, qual o gasto médio por usuário, quanto repete a compra etc. Com um ano de uso de uma nova plataforma houve um aumento de 40% na receita, sendo que o principal fator foi a busca mais precisa dentro do site.

Outro número importante é que as buscas sem resultados caíram 80%, ou seja, o mecanismo está mais genérico e consegue atender melhor os usuários exibindo produtos cada vez mais relevantes. Isso também faz com que o tempo médio de navegação diminua. No caso da Netshoes esse tempo caiu de 5 para 3 minutos em média. Esse é o tempo gasto entre a escolha do produto e a conclusão do pedido.

Como essa ação está mais rápida, o cliente fica mais satisfeito e a chance de retornar ao site para realizar outras compras é muito maior.

Referências

AWS. Página institucional. Disponível em: <https://docs.aws.amazon.com/pt_br/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html>. Acesso em: 02 mar. 2021.

APACHE CASSANDRA. *Página institucional*. Disponível em: <http://cassandra.apache.org/>. Acesso em: 02 mar. 2021.

APACHE COUCHBASE. *Página institucional*. Disponível em: <http://couchdb.apache.org/>. Acesso em: 02 mar. 2021.

APACHE HADOOP. *Página institucional*. Disponível em: <https://hadoop.apache.org/>. Acesso em: 02 mar. 2021.

APACHE KAFKA. *Página institucional*. Disponível em: <https://kafka.apache.org/>. Acesso em: 02 mar. 2021.

APACHE SPARK. *Página institucional*. Disponível em: <https://spark.apache.org/>. Acesso em: 02 mar. 2021.

BARBIERI, C. *Business Intelligence*: Modelagem e Qualidade. Rio de Janeiro: Elsevier, 2011.

CAMILO, C.O.; SILVA, J.C. *Mineração de Dados*: Conceitos, Tarefas, Métodos e Ferramentas. Universidade Federal de Goiás, 2009. Disponível em: <http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf>. Acesso em: 02 mar. 2021.

DUMBILL, E. *Planning for Big Data*. Sebastopol: O'Reilly Media, 2012.

FAYYAD, U.M., et al. *From Data Mining to Knowledge Discovery in Databases*. AI

MAGAZINE. England, 1996. Disponível em:

<https://www.aaai.org/ojs/index.php/aimagazine/article/view/1230>. Acesso em: 02 mar. 2021.

GARTNER. *Página institucional*. Disponível em: <https://www.gartner.com>. Acesso em: 02 mar. 2021.

HAMSTRA, Mark; ZAHARIA, Matei; KARAU, Holden. *Learning Spark: Lighting-Fast Big Data Analysis*. 1. ed. O'Reilly Media, 2015.

HAND, D.J; MANIILA, H. SMYTH, P. *Principles of Data Mining by David Hand*. A Massachusetts: Bradford Book, 2001.

HEUMANN, Christian; SCHOMAKER, Michael; SHALABH, Sinha. *Introduction to Statistics and Data Analysis:* With Exercises, Solutions and Applications in R. 1. ed.

Springer, 2019.

IBM. Página institucional. *Infographics & Animations*: The Four V's of Big Data. Disponível em: <https://www.ibmbigdatahub.com/infographic/four-vs-big-data>.

Acesso em: 02 mar. 2021.

INMON, W.H. *Building the Data Warehouse*. Wiley: 4 edition, 2005.

JAMES, Gareth, et al. *An Introduction to Statistical Learning:* With Applications in R: 103. 1. ed. Springer, 2014.

KAIROSDB. *Página institucional*. Disponível em <https://kairosdb.github.io/>. Acesso em: 02 mar. 2021.

KNAFLIC, C.N. *Storytelling Com Dados*: Um Guia Sobre Visualização de Dados Para Profissionais de Negócio. Rio de Janeiro: Alta Books, 2017.

MINELLI, Michael; CHAMBERS, Michele; DHIRAJ Ambiga. *Big Data, Big Analytics:* Emerging Business Intelligence Analytic Trends for Todays Business. Wiley CIO, 2013.

MONGODB. *Página institucional*. Disponível em: <https://www.mongodb.com/>.

Acesso em: 02 mar. 2021.

NEO4J. Página institucional. Disponível em <https://neo4j.com/>. Acesso em: 02 mar. 2021.

REED, Jeff. *Data Analytics:* Applicable Data Analysis to Advance Any Business Using the Power of Data Driven Analytics. 1. ed. New York: Makron Books, 2018.

SUTTON, R. S., 1999. Reinforcement Learning. The MIT Encyclopedia of Cognitive Sciences, MIT Press.