



Fundamentos de Data Analytics

Capítulo 1. Conceitos e fundamentos

Prof. Angelo Assis



Aula 1.1. Big Data

Nesta aula

- ☐ O que é Big Data.
- ☐ Os V's do Big Data.

“Big Data 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”

IT Glossary – Gartner

- 40 zetabytes de dados serão criados até 2020;
- 6 bilhões de pessoas possuem celulares hoje;
- Estima-se que 2,3 trilhões de gigabytes são criados diariamente;
- Em 60 segundos ocorrem 695 mil atualizações de status no Facebook e 168 milhões de e-mails são enviados.

Fonte: IBM (2018).

Os V's do Big Data

- Volume.
- Velocidade.
- Variedade.

Os V's do Big Data

- Volume
- Velocidade
- Variedade
- Veracidade
- Valor

Os V's do Big Data

- Volume
- Velocidade
- Variedade
- Veracidade
- Valor
- Visibilidade
- Variabilidade

Os V's do Big Data

- Volume
- Velocidade
- Variedade
- Veracidade
- Valor
- Visibilidade
- Variabilidade
- Vulnerabilidade
- Validade
- Volatilidade

Os V's do Big Data

- Volume
- Velocidade
- Variedade
- Veracidade
- Valor
- Visibilidade
- Variabilidade
- Vulnerabilidade
- Validade
- Volatilidade

V ... ?

- ☑ Geramos um volume cada vez maior de dados;
- ☑ Diversas empresas já se beneficiam da organização e análise desses dados;
- ☑ Sua definição ainda é discutida frequentemente.

Próxima aula

☐ Machine Learning.



Aula 1.2. Machine Learning

- ☐ Machine Learning.
- ☐ Tipos de aprendizado.
- ☐ Deep Learning.

“Machine Learning é um campo de estudo que dá aos computadores a habilidade de aprender sem terem sido programados para tal”

Arthur Samuel

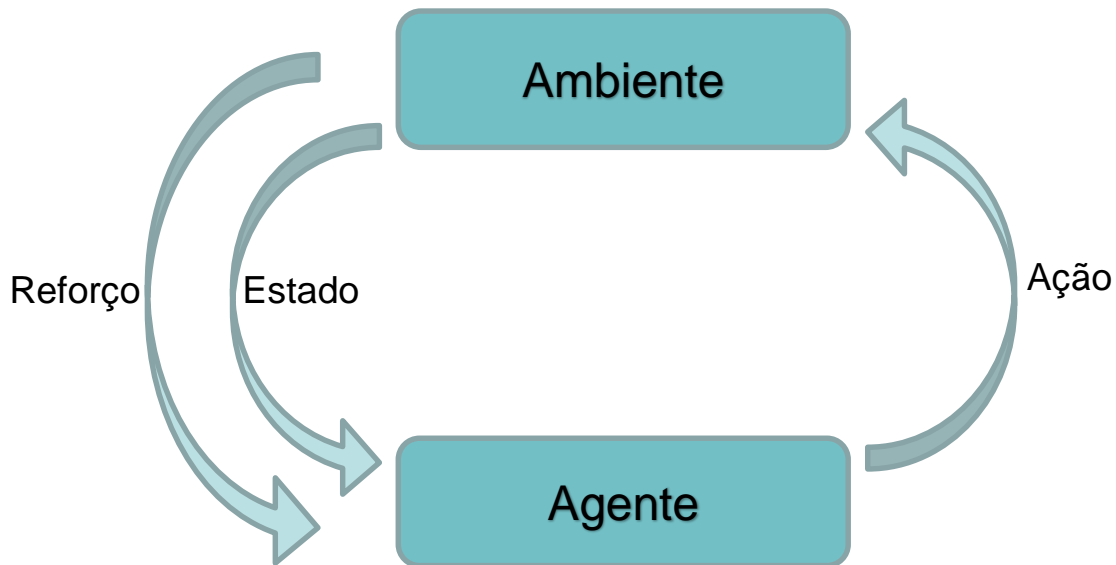
“Um programa de computador é dito para aprender com a experiência E com a relação a alguma classe de tarefas T e medida de desempenho P , se o seu desempenho em tarefas em T , medida pelo P , melhora com a experiência E ”

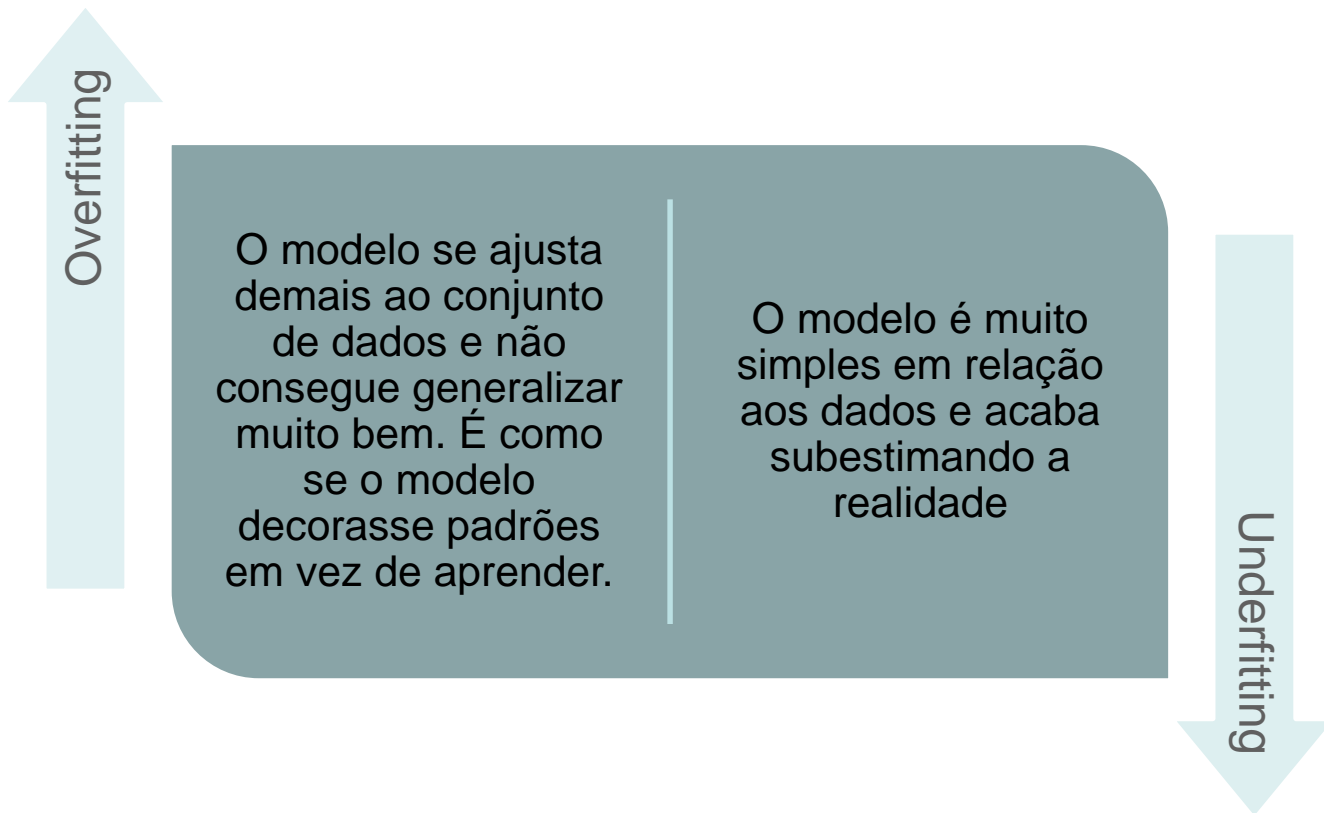
Tom Mitchell

- Exemplos de tarefas que Machine learning pode ajudar a solucionar:
 - **Tomada de decisão:** Auxílio na tomada de decisão com base nos dados;
 - **Regressão:** Tentativa de prever um resultado numérico com base em uma ou mais variáveis de entrada;
 - **Clustering:** Agrupamento de objetos com base nas semelhanças em seus dados;
 - **Deteccção de anomalias:** Observação de mudanças no padrão dos dados.

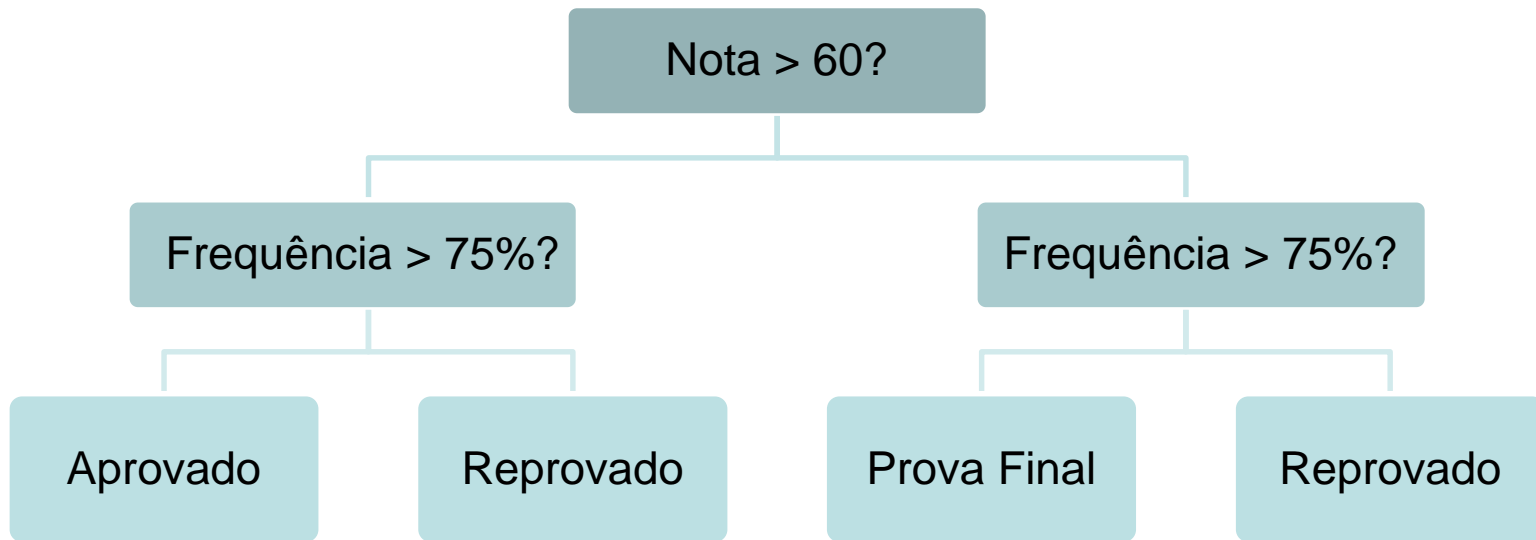
- Supervisionada:
 - Existe um conjunto prévio de dados (input e output);
 - Utilizado para problemas de “regressão” e “classificação”.
- Não supervisionada:
 - Não existe um resultado específico esperado. Não é possível prever os resultados do cruzamento das informações;
 - Descobre padrões entre os dados (agrupamento/clustering).

- Aprendizagem por reforço:
 - Nesse ambiente existe um agente que interage através de percepções e ações;
 - A cada passo o agente recebe uma indicação do estado atual do ambiente e escolhe uma 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.



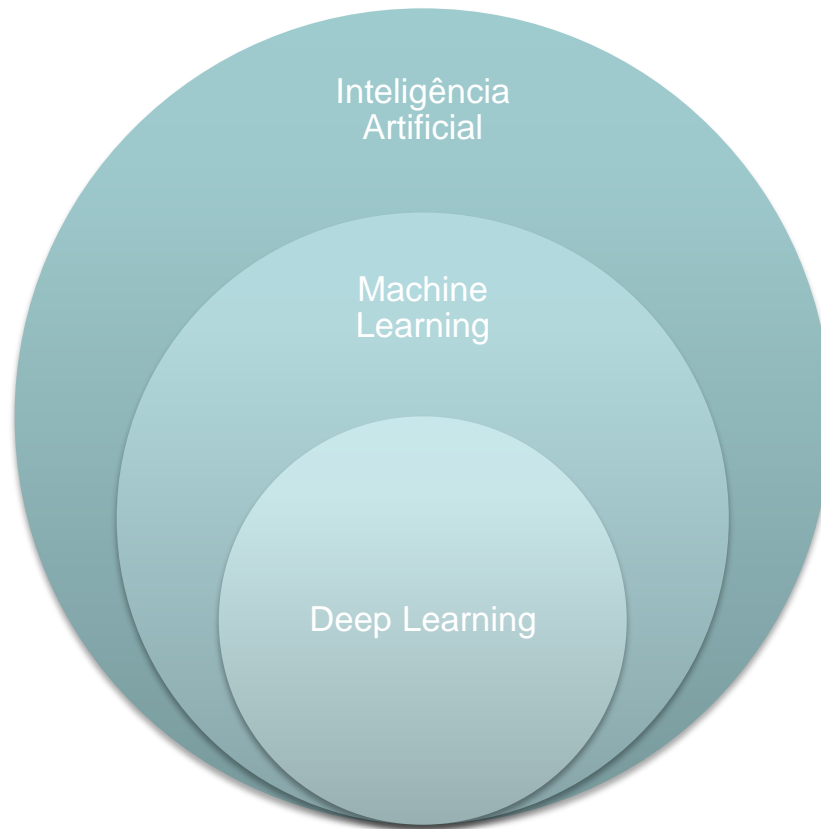


- Representam o mapeamento de possíveis resultados de uma série de escolhas relacionadas;
- Podem ter uma representação gráfica na forma de fluxogramas.



- Subcategoria do Aprendizado de Máquina que se baseia em Redes Neurais Artificiais para realizar o treinamento.
- Uma rede neural é um processador paralelamente distribuído constituído de unidades de processamento simples, que têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para o uso.

- Combina avanços no poder da computação e tipos especiais de redes neurais para que as máquinas “aprendam” padrões complicados em quantidades exponenciais de dados.
- Utilizado em carros autônomos, sistemas de recomendação, reconhecimento de voz, reconhecimento facial, identificação de objetos.



Próxima aula

☐ Data Mining.



Aula 1.3. Data Mining

- ☐ KDD.
- ☐ Data Mining.
- ☐ Qualidade dos Dados.
- ☐ Limitações e desafios.

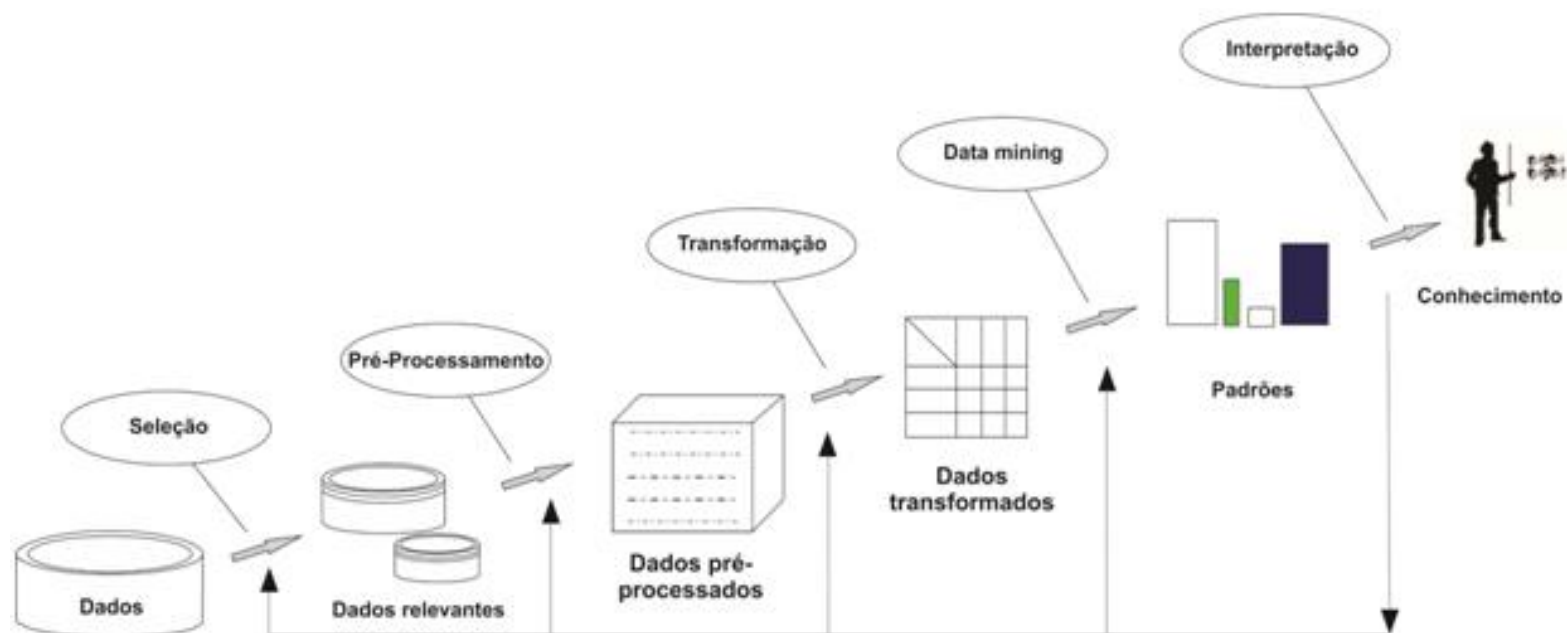
- Não há dúvidas! Há um avanço na coleta e no armazenamento de dados!
- O processo manual torna-se impraticável.
- Então, o que fazer com os dados armazenados?

- Descoberta de Conhecimento em Bases de Dados:

“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”

Usama M. Fayyad

Processo KDD



“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”

David Hand

- É interdisciplinar, e utiliza técnicas de estatística, recuperação de informação, inteligência artificial, reconhecimento de padrões, entre outros.

- Conhecer o tipo dos dados com o qual se irá trabalhar também é fundamental para a escolha do método mais adequado.
- Podem ser quantitativos ou qualitativos.
- Atenção à qualidade dos dados!
 - Valores em branco ou nulo, valores viciados, variáveis duplicadas.

- Limpeza:
 - Registros incompletos, dados inconsistentes, valores padrões.
- Integração:
 - De diversas fontes para repositório único e consistente.
- Transformação:
 - Suavização, agrupamento, generalização, normalização.
- Redução:
 - Menor massa de dados, com a mesma representatividade dos dados originais.

1. Conjunto de Treinamento (Training Set)
2. Conjunto de Testes (Test Set)
3. Conjunto de Validação (Validation Set)

- Tarefas comuns para um processo de Mineração de Dados:
 - Descrição;
 - Classificação;
 - Regressão;
 - Predição;
 - Agrupamento;
 - Associação.

- As relações entre os atributos precisam ser muito bem definidas, caso contrário, os resultados podem ser mal interpretados;
- Exige um alto conhecimento do usuário com relação aos dados;
- Técnicas para lidar com base de dados cada vez maiores;
- A velocidade com que os dados mudam faz com que os modelos gerem resultados inválidos;
- Baixa qualidade dos dados.

- ☑ Data Mining é um processo para gerar informações a partir de um conjunto de dados.
- ☑ A qualidade dos dados é fundamental para o processo.
- ☑ Não há “receita de bolo”.

■ Próxima aula

☐ Data Analytics.

☐ Data Science.



Aula 1.4. Data Analytics e Data Science

- ☐ Data Analytics.

- ☐ Data Science.

- É o processo que torna possível a transformação de dados e informações em conhecimento para um propósito específico.
- Data Analytics pode ocorrer mesmo sem a utilização de computadores.

- **Dado:** o menor grão e a matéria prima da escala do conhecimento
 - 36,5
- **Informação:** dados de maneira organizada dentro de uma escala
 - 36,5 °C
- **Conhecimento:** informação contextualizada
 - Não tem febre
- **Sabedoria:** conjunto complexo de raciocínios
 - O que fazer se estivesse com febre?



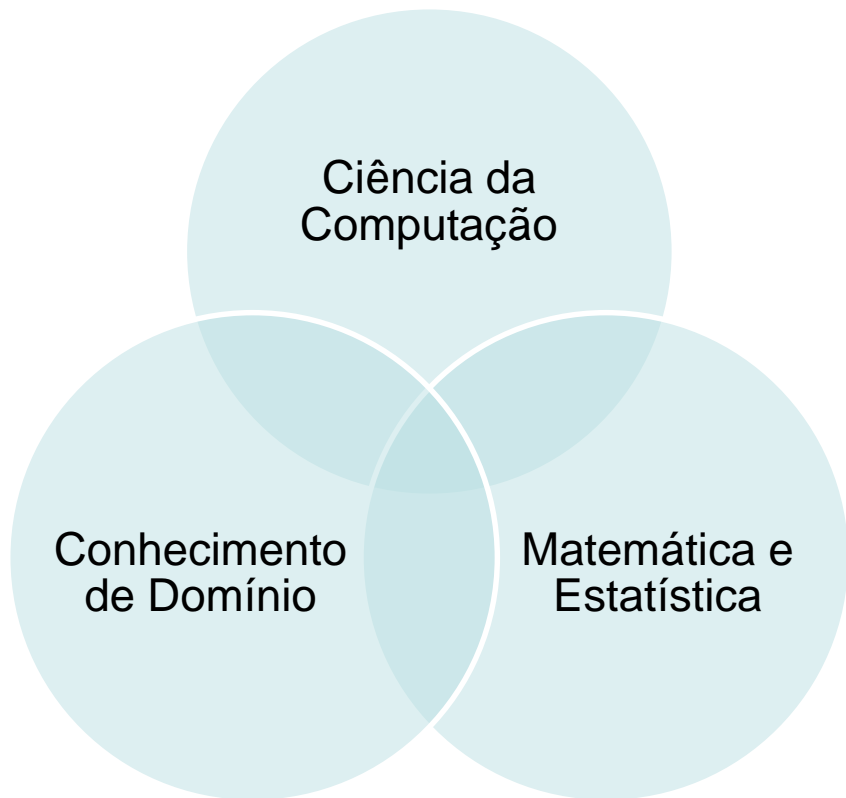
- Origem dos Dados:
 - Excel, Banco de Dados.
- Limpeza e transformação:
 - Raramente o dado está pronto.
- Propósito da Análise:
 - Definir quais perguntas devem ser respondidas.
- Validação:
 - Qual o retorno?

“Se refere a todo o estudo de dados e informações de um determinado assunto a fim de se extrair dados e obter insights. Abrange a limpeza, preparação e análise de dados, envolvendo diversas áreas de conhecimento como Matemática, Estatística, Computação além da compreensão do assunto específico da análise em questão.”

“Dados são o novo ouro”

Tudo pode ser coletado, armazenado e analisado.





- ☑ Não é necessário uma infraestrutura excepcional para iniciar um projeto de análise de dados.
- ☑ Data Science é multidisciplinar.

■ Próxima aula

- ❑ Business Intelligence.



Aula 1.5. Business Intelligence

☐ Business Intelligence.

- É a Análise de Dados aplicada a um ramo de negócio específico com o objetivo de converter os dados em insights que os líderes e gestores possam usar para tomada de decisão.
- **O BI mede o desempenho passado para prever o futuro.**

- Não se resume ao uso de softwares como Tableau, Power BI, Pentaho, Qlik, etc.
- Não é apenas montar Dashboard.
- Pode enfrentar barreira cultural.

- Vai além de todo o ciclo da Análise de Dados.
- Coleta e organização + análise e visualização + compartilhamento e monitoramento + suporte a tomada de decisão + decisões com base em evidências + medir o desempenho passado + planejar o futuro.

- ETL – Extract, Transform and Load:
 - 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*.

- OLAP - Online Analytical Processing:
 - É um conceito de interface com o usuário que proporciona a capacidade de manipular e analisar um grande volume de dados em diversos ângulos.

- Métodos de armazenamento:
 - ROLAP (OLAP Relacional);
 - MOLAP (OLAP Multidimensional);
 - HOLAP (OLAP Híbrido);
 - DOLAP (OLAP Desktop);
 - WOLAP (Web OLAP).

- O poder para o usuário de negócio!
- Permite que usuários sem especialização técnica consigam elaborar seus próprios relatórios e dashboards de maneira rápida e assertiva.
- Não basta apenas fornecer as ferramentas para os usuários.

- ☑ Permite que as empresas operem e criem estratégias de forma mais inteligente.
 - Entender o passado e prever o futuro!
- ☑ Empresas que sabem usar os dados a seu favor possuem um alto diferencial competitivo no mercado.

- ☐ O que é o algoritmo K-Means.
- ☐ Passo a passo do K-Means.
- ☐ Demonstração.



Aula 1.6. Algoritmo K-Means na prática

- ☐ O que é o algoritmo K-Means.
- ☐ Passo a passo do K-Means.
- ☐ Demonstração.

O que é o K-Means?

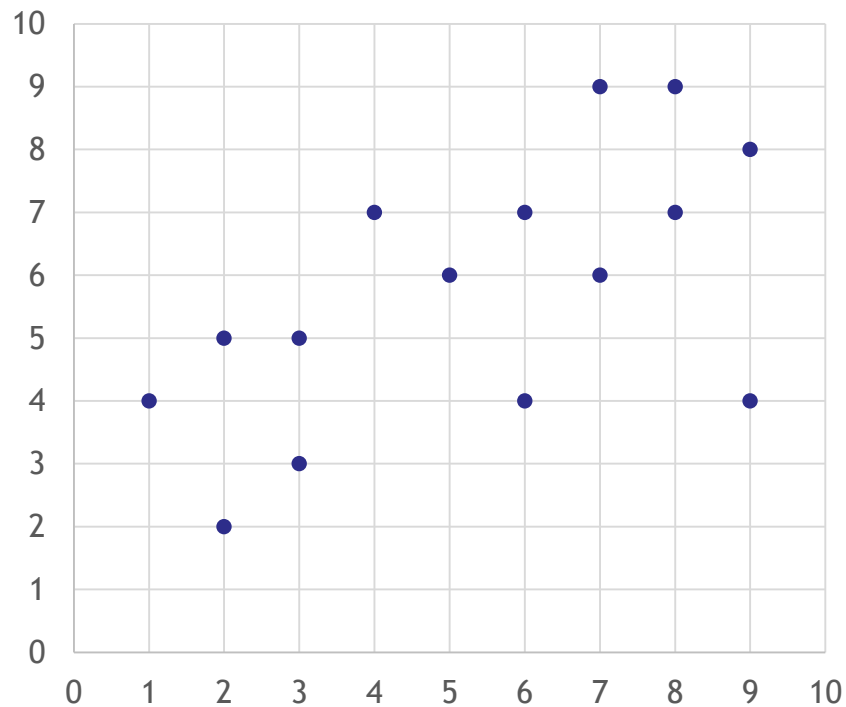
- K-Means é um algoritmo de Machine Learning baseado em aprendizado não supervisionado
- O objetivo é encontrar similaridades entre os dados e agrupá-los conforme o número de clusters (K).
- O algoritmo, de forma iterativa, atribui os pontos de dados ao grupo que representa a menor distância.
- O algoritmo gera K (ou menos) clusters

- 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
- E diversas outras aplicações...

1. Inicializar os centroides aleatoriamente (precisamos saber o valor de k antes de começar);
2. Para cada ponto, calcular a distância para cada centroide e associar ao que estiver mais próximo;
3. Calcular a média de todos os pontos relacionados a um centroide e definir um novo centroide

1. Inicializar os centroides aleatoriamente (precisamos saber o valor de k antes de começar);
2. Para cada ponto, calcular a distância para cada centroide e associar ao que estiver mais próximo;
3. Calcular a média de todos os pontos relacionados a um centroide e definir um novo centroide

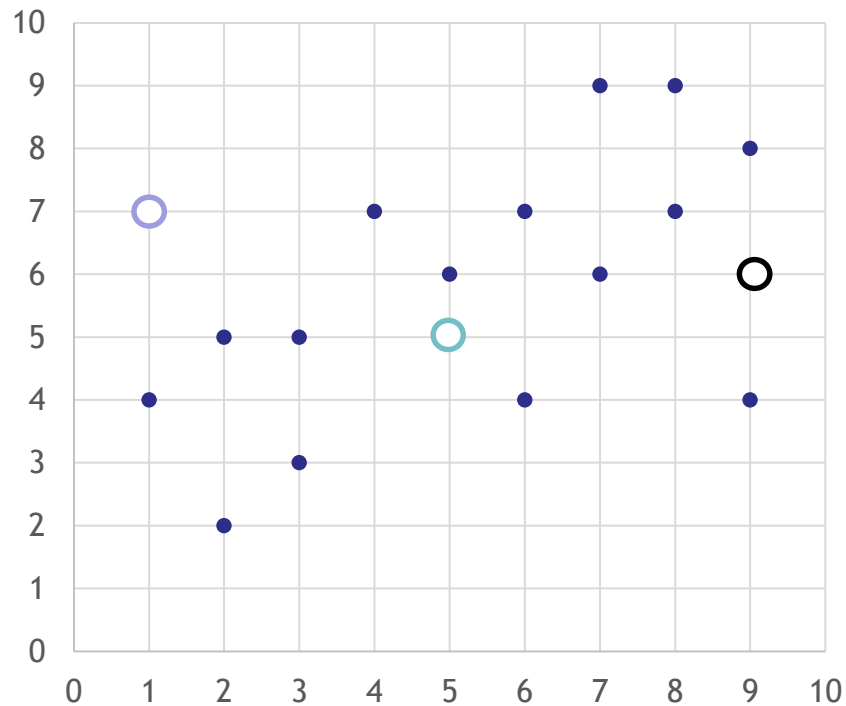
K-Means



Passo a passo

K-Means

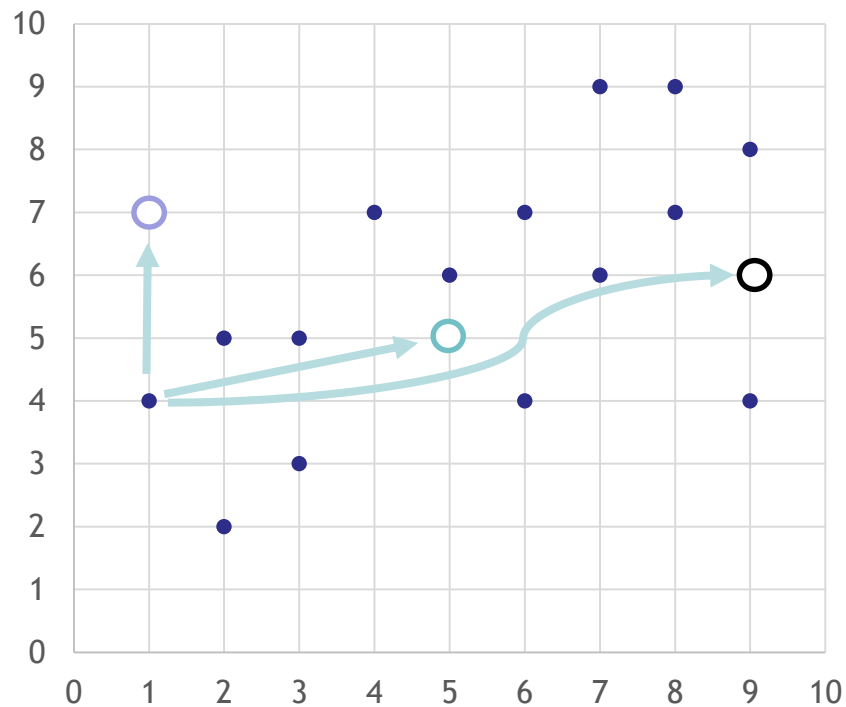
- (1; 7)
- (5; 5)
- (9; 6)



Passo a passo

K-Means

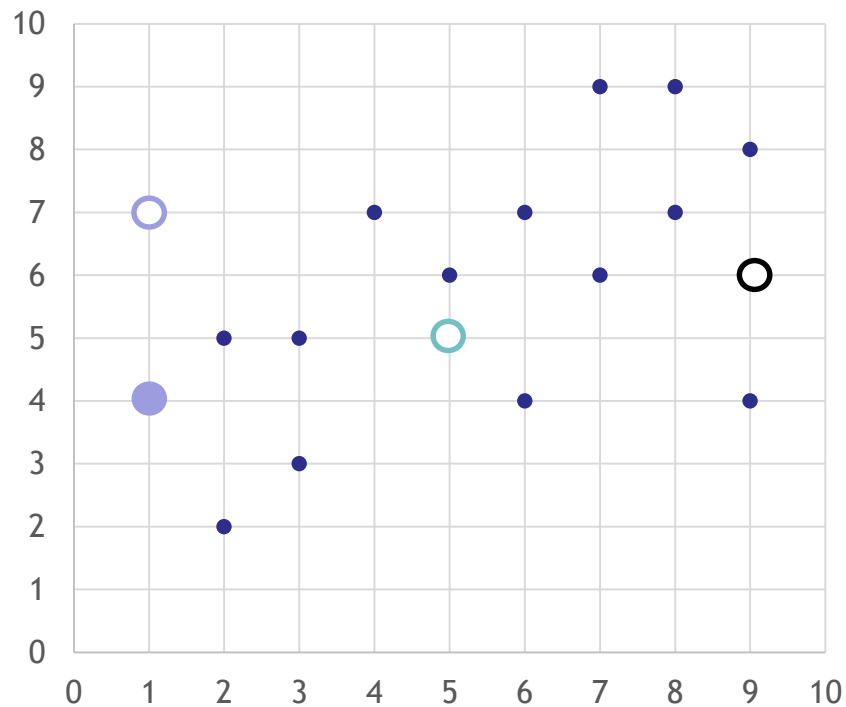
- (1; 7)
- (5; 5)
- (9; 6)



Passo a passo

K-Means

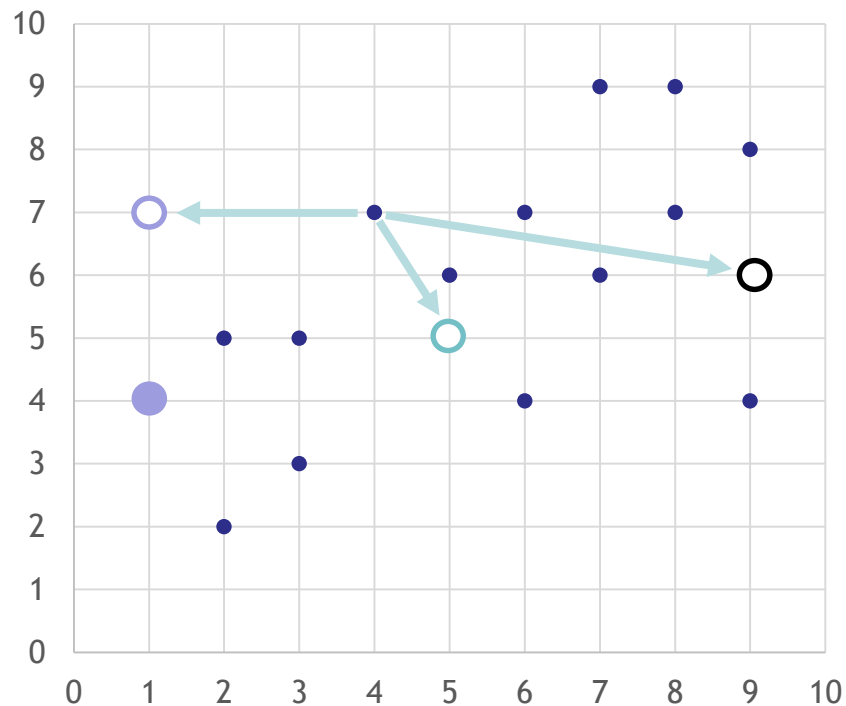
- (1; 7)
- (5; 5)
- (9; 6)



Passo a passo

K-Means

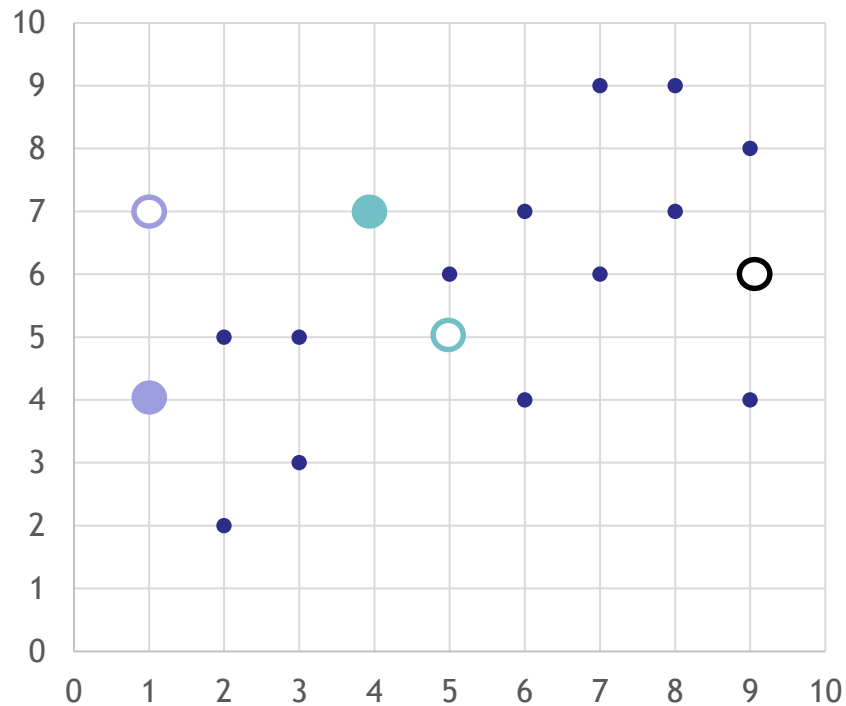
- (1; 7)
- (5; 5)
- (9; 6)



Passo a passo

K-Means

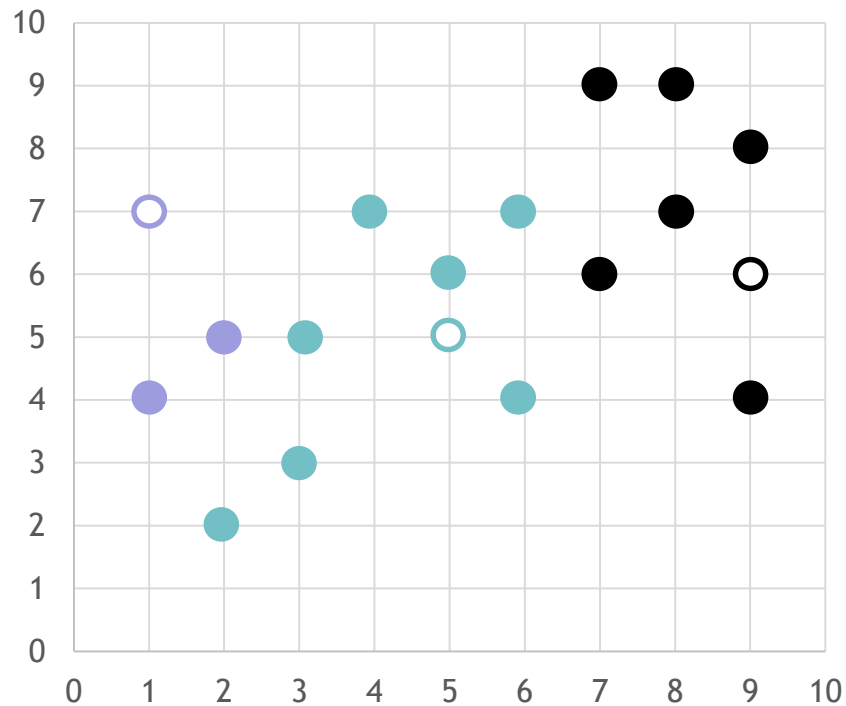
- (1; 7)
- (5; 5)
- (9; 6)



Passo a passo

K-Means

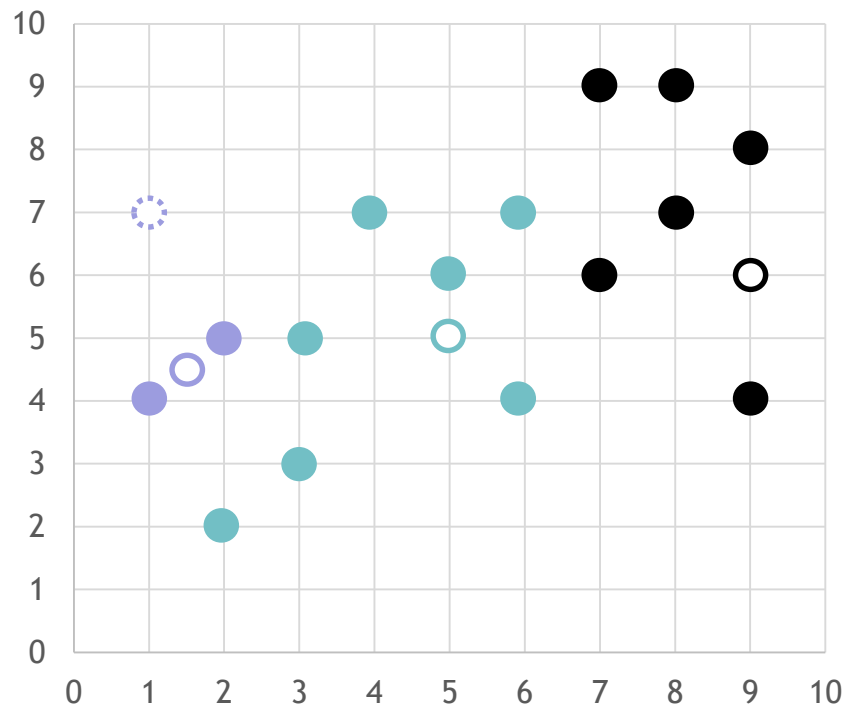
- (1; 7)
- (5; 5)
- (9; 6)



Passo a passo

K-Means

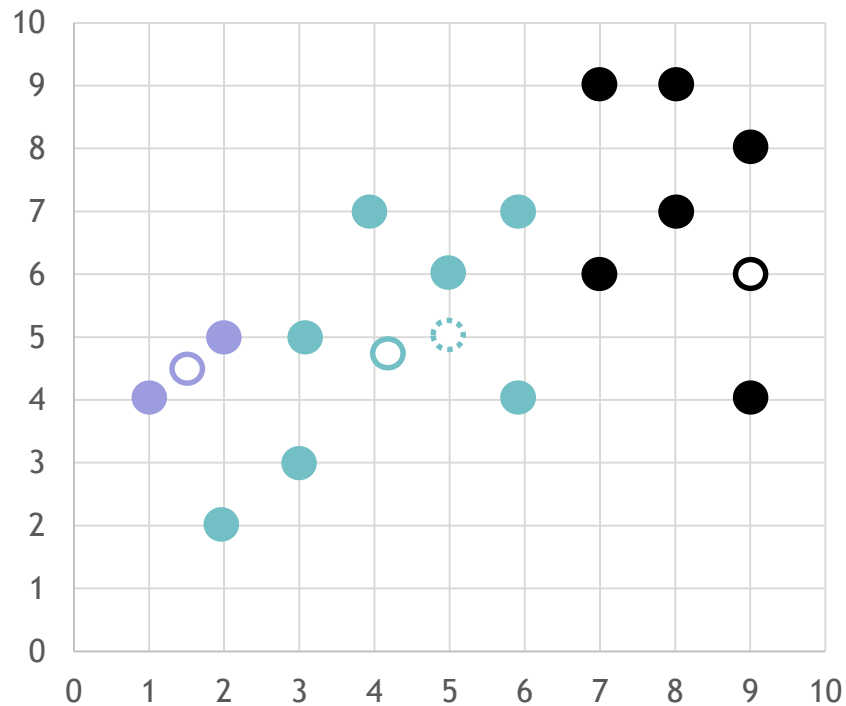
- (1,5; 4,5)
- (5; 5)
- (9; 6)



Passo a passo

K-Means

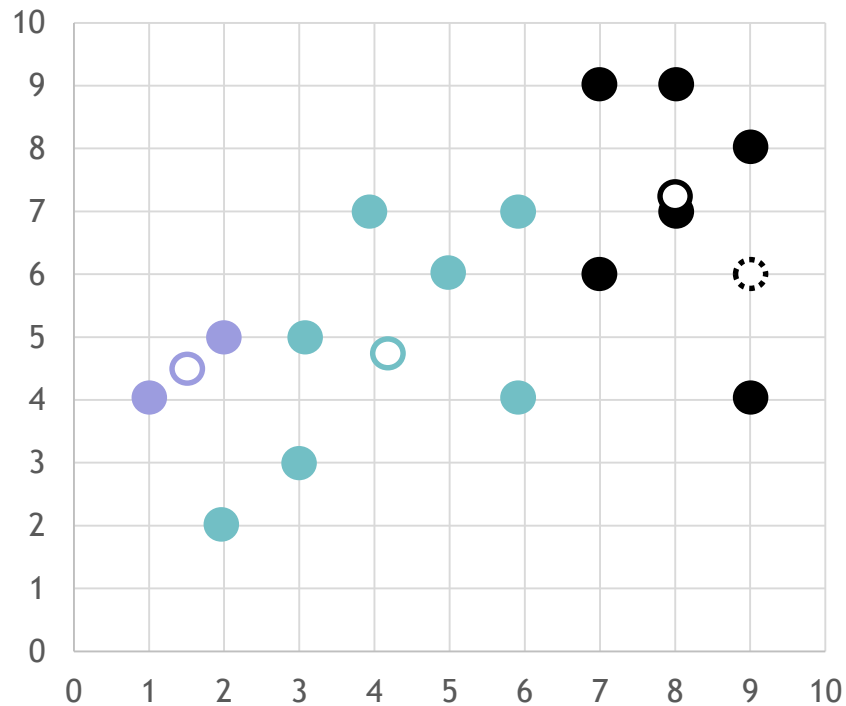
- (1,5; 4,5)
- (4,1; 4,8)
- (9; 6)



Passo a passo

K-Means

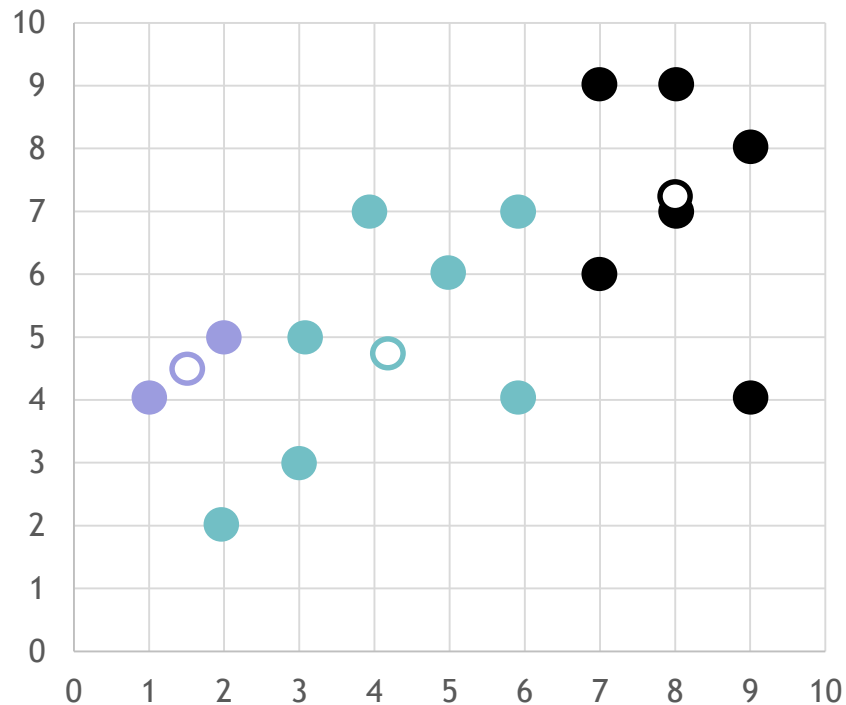
- (1,5; 4,5)
- (4,1; 4,8)
- (8,0; 7,1)



Passo a passo

K-Means

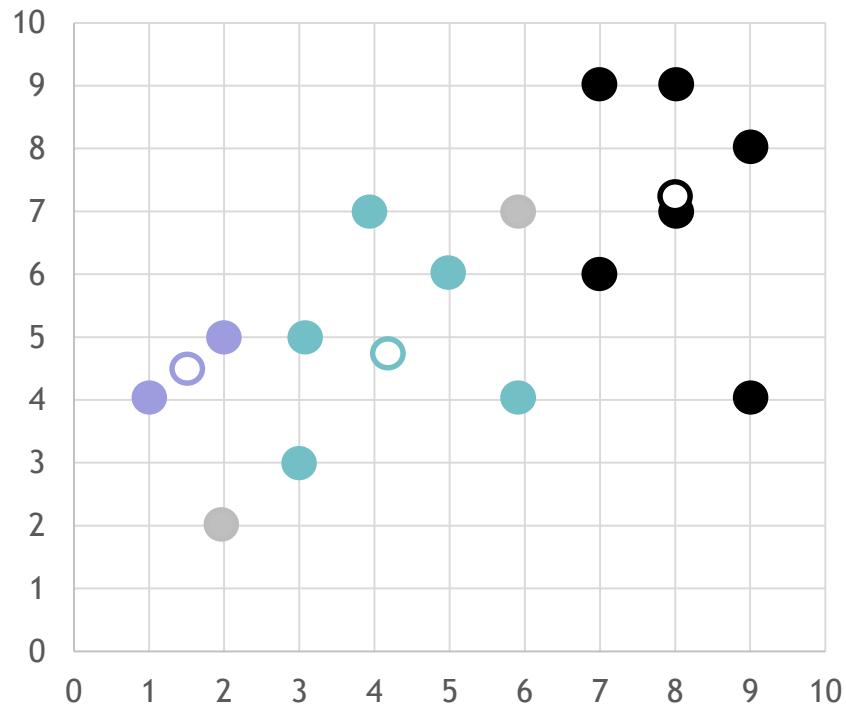
- (1,5; 4,5)
- (4,1; 4,8)
- (8,0; 7,1)



Passo a passo

K-Means

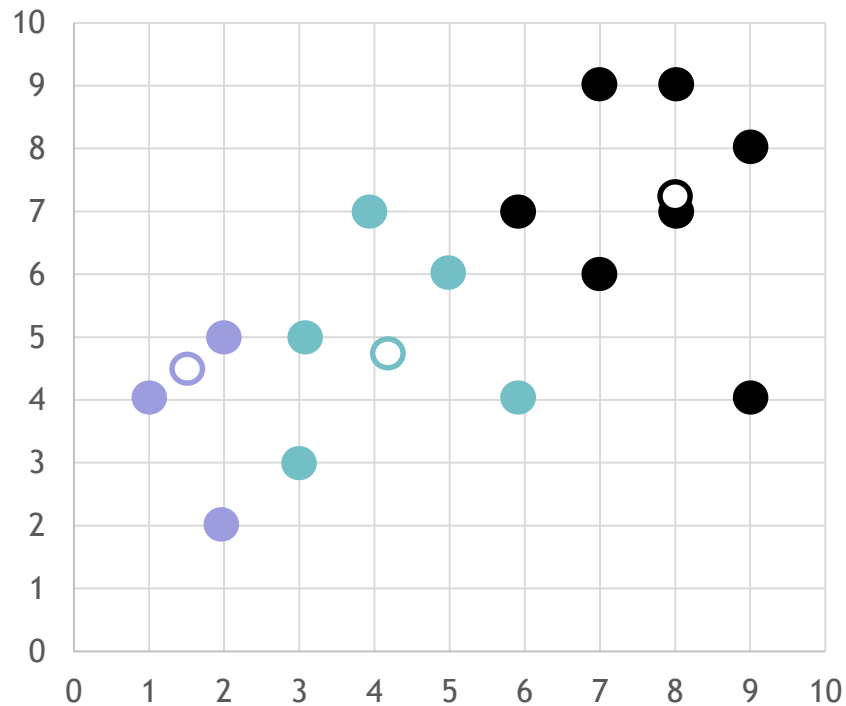
- (1,5; 4,5)
- (4,1; 4,8)
- (8,0; 7,1)



Passo a passo

K-Means

- (1,5; 4,5)
- (4,1; 4,8)
- (8,0; 7,1)



- Como escolher o número de clusters?
 - Within Cluster Sum of Squares (WCSS)
- Como calcular a distância entre os dados?
 - Distância euclidiana, distância de Manhattan
- Qual é a condição de parada ideal?
 - Número de iterações, estabilidade dos clusters
- Atenção à inicialização dos centroides

- Google Colab

☐ Capítulo 2:

- Processos e Técnicas para Análise de Dados.



Fundamentos de Data Analytics

Capítulo 2. Processos e técnicas para Análise de Dados

Prof. Angelo Assis



Aula 2.1. Etapas do Processo

- ☐ Ciclo de análise de dados.
- ☐ Coleta e preparação dos dados.
- ☐ Visualização e feedback.



- Técnica dos 5 W's

Pergunta		Exemplo
Why?	Porque?	Porque é importante essa análise para o negócio?
Who?	Quem?	Quem iremos analisar? Nossos compradores? Fornecedores?
What?	O que?	O que iremos analisar? Comportamento de compra?
Where?	Onde	A análise estará voltada para o contexto nacional ou internacional?
When?	Quando	Qual período será considerado para as análises?

- Métricas SMART:
 - **E**specíficas
 - **M**ensuráveis
 - **A**lcançáveis
 - **R**elevantes
 - Com limite de **T**empo



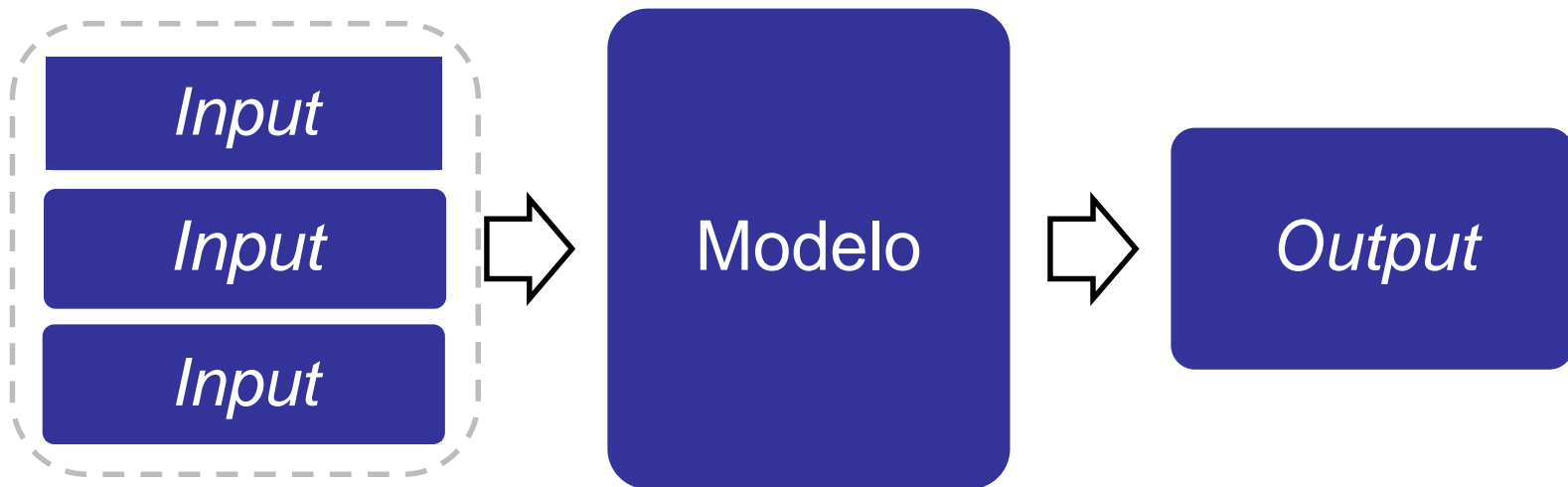
- Disponibilidade dos dados.
- Armazenamento dos dados.
- Estrutura dos dados.
- Qualidade dos dados.

- Disponibilidade dos dados.
- Armazenamento dos dados.
- Estrutura dos dados.
- Qualidade dos dados.

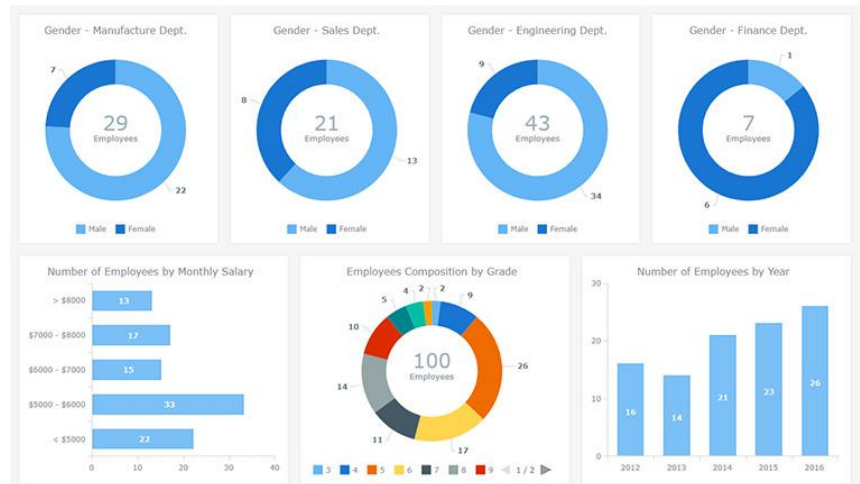
ETL + DW

- Data Mining.
- Machine Learning.
- Deep Learning.
- Modelos Estatísticos.

- Deve suportar novas entradas, para ajudar a tomar uma decisão ou prever um resultado.



- Os dados respondem às perguntas?
- Como apresentar os dados?
 - Use e abuse do *storytelling*
- Será preciso uma nova iteração?
- O modelo é válido por quanto tempo?



- ☑ Antes de começar qualquer análise precisamos definir o resultado esperado.
- ☑ Nosso modelo pode se degradar com o tempo.
- ☑ De nada adianta uma análise perfeita sem uma boa apresentação do resultado.

□ ETL

- Extract, Transform and Load.



Aula 2.2. ETL (Extract, Transform and Load)

☐ ETL.

☐ Extract.

☐ Transform.

☐ Load.

E_{xtract}



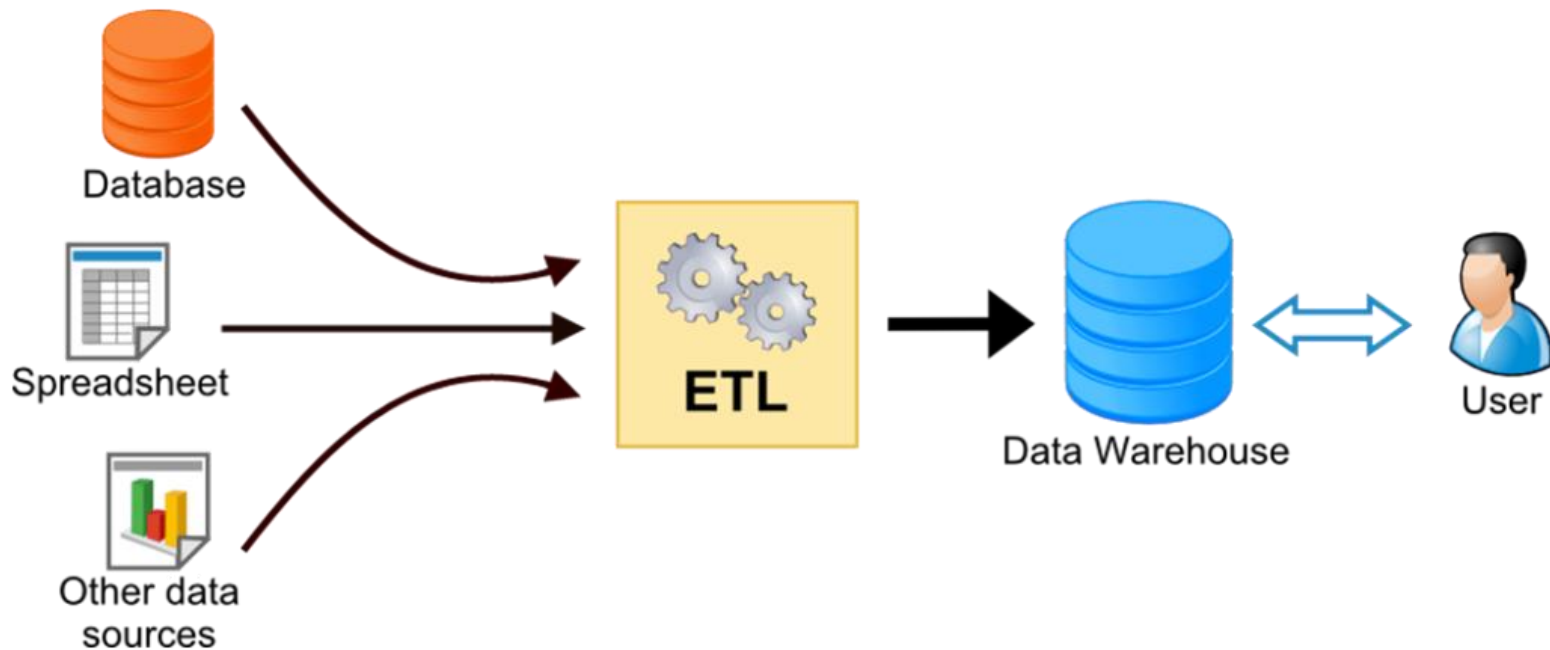
T_{ransform}



L_{oad}



- 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.



- É uma das fases mais críticas.
- 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.

- Fase em que os dados são extraídos dos *OLTPs* e conduzidos para a *staging area*, onde são convertidos para um único formato.

- Fase em que os dados são extraídos dos *OLTPs* e conduzidos para a *staging area*, onde são convertidos para um único formato.
- OLTP.
- Staging Area.

- ***Online Transaction Processing* ou Processamento de Transações em Tempo Real:** são 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.

- Área onde os dados são colocados após a extração a partir dos sistemas de origem.
- Raramente as *staging area* são normalizadas.
- Reduz a sobrecarga de acessos aos sistemas fontes.
- **Dedicada para a fase ETL e não disponível para os usuários finais.**
 - Relatórios não podem acessar os dados da *staging area*.
 - Somente processos ETL podem ler e escrever na *staging área*.

- É nesta etapa que realizamos os devidos ajustes, podendo assim melhorar a qualidade dos dados e consolidar dados de duas ou mais fontes.

- Tradução de valores codificados e codificação de valores de forma livre:
 - Ex: Mapear “Masculino”, “1” e “Sr.” para M.
- Derivação de um novo valor calculado:
 - $\text{montante_vendas} = \text{qtde} * \text{preço_unitário}$.
- Resumo de várias linhas de dados:
 - total de vendas para cada loja e para cada região.
- Geração de valores de chaves substitutas:
 - surrogate Keys.

- Consiste em fisicamente estruturar e carregar os dados para dentro da camada de apresentação seguindo o modelo dimensional.

- ☑ ETL tem grande impacto na Análise de Dados.
- ☑ Extração e carga são obrigatórios.
- ☑ Transformação é “opcional”.

Próxima aula

☐ Data Warehouse.

☐ Data Lake.



Aula 2.3. Data Warehouse e Data Lake

☐ Data Warehouse.

☐ Data Lake.

☐ Data Mart.

“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”

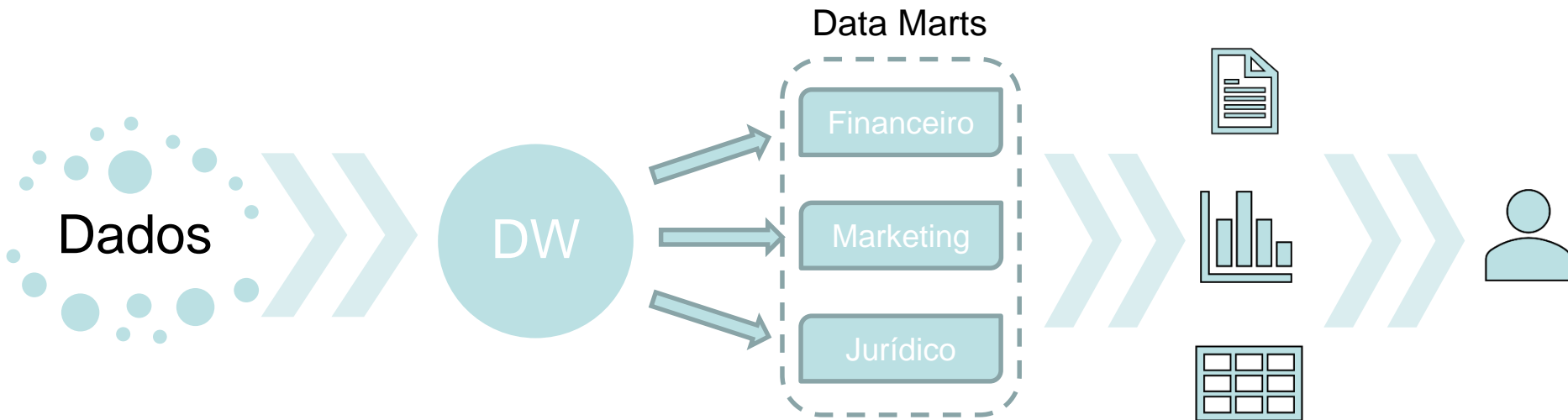
William H. (Bill) Inmon

- **Orientado a assuntos:** por exemplo, vendas de produtos a diferentes tipos de clientes, atendimentos e diagnósticos de pacientes, rendimento de estudantes.
- **Integrado:** diferentes nomenclaturas, formatos e estruturas das fontes de dados precisam ser acomodadas em um único esquema para prover uma visão unificada e consistente da informação.

- **Não volátil:** 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..
- **Variável em relação ao tempo:** o histórico dos dados por um período de tempo superior ao usual em BDs transacionais permite analisar tendências e mudanças.



- 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.

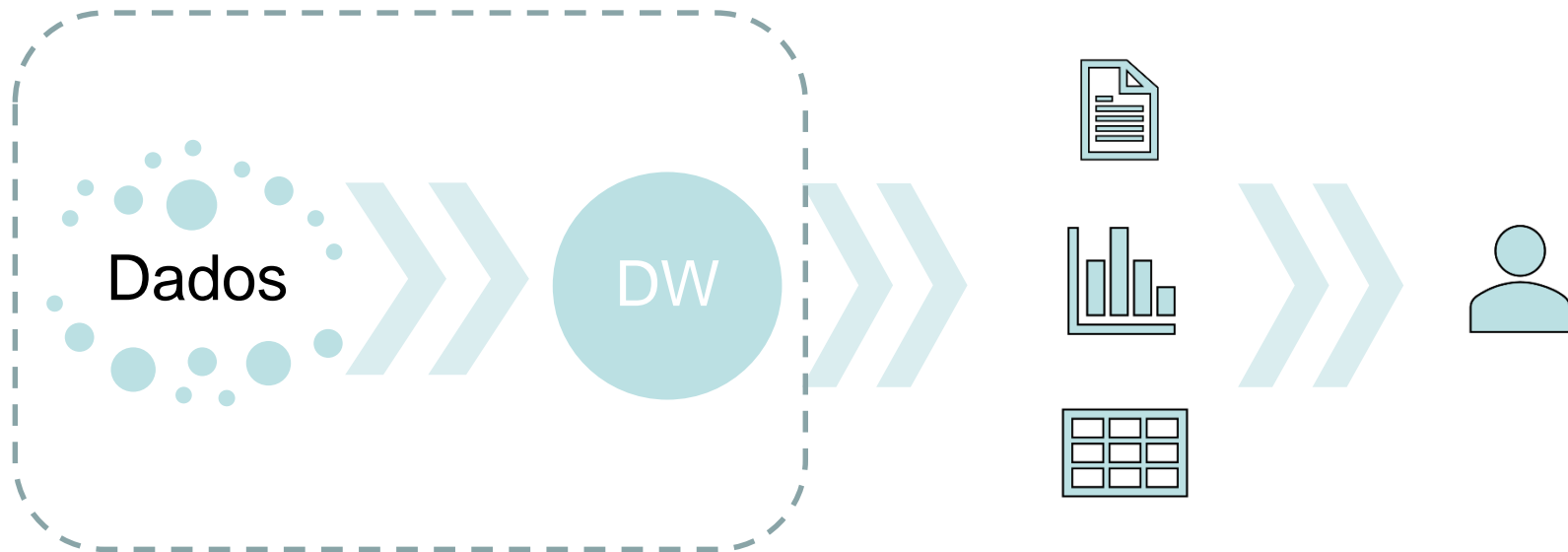


“são os dados em grandes volumes e em seu estado natural, vindos de todos os tipos de fontes, onde os usuários podem “mergulhar” e tirar amostras. Um lago cheio de dados”

James Dixon

- Seria um Data Warehouse repaginado?
- Dados brutos → Data Swamp (pântano de dados).
- Como lidar com segurança e privacidade?





	Data Warehouse	Data Lake
Dados	Estruturados Processados	Estruturados / Semi-estruturados / Não estruturados Não processados (em estado bruto)
Processamento	Ocorre no momento da escrita	Ocorre no momento de leitura
Armazenamento	Alto custo para grande volume de dados	Projetado para baixo custo, independente do volume de dados
Agilidade	Pouco ágil, configuração fixa	Bastante ágil, pode ser configurado e reconfigurado conforme necessário
Segurança	Estratégias maduras	Precisa aperfeiçoar o modelo de segurança e acesso aos dados

Próxima aula

☐ Tipos de Análises.



Aula 2.4. Tipos de Análises

- ☐ Análise Descritiva.
- ☐ Análise Diagnóstica.
- ☐ Análise Preditiva
- ☐ Análise Prescritiva.

- Foco no passado, identificação de padrões.
- Possuem um valor significativo e são de fácil consumo.
- Ajuda a tomar decisões imediatas.

- O objetivo é entender e explicar o que foi detectado.
- Compreender as causas de um evento ou analisar o impacto e alcance de uma ação tomada.
- Identificar quais fatores influenciaram o resultado atual.
- Funciona bem em conjunto com análises preditivas.

- Tipo de análise mais conhecido.
- 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.
- Utiliza algoritmos de regressão, classificação e agrupamento.
- Demandam um volume significativo de dados de boa qualidade e exigem um maior grau de sofisticação.

- Também conhecida como “Análise de Recomendação”.
- Busca trazer informações das consequências de acontecimento previsto.
- A análise preditiva identifica tendências futuras, a prescritiva traça as possíveis consequências de cada ação.

- ☑ Cada tipo de análise tem seu próprio escopo e sua própria finalidade.
- ☑ É possível conhecer a maturidade em Data Analytics de uma determinada empresa baseado em quais análises fazem parte de sua realidade.

- ☑ Análise Descritiva:
 - O que aconteceu / está acontecendo?
- ☑ Análise Diagnóstica:
 - Por que aconteceu?
- ☑ Análise Preditiva:
 - O que vai acontecer?
- ☑ Análise Prescritiva:
 - O que fazer se for acontecer?

Próxima aula

☐ Streaming.



Aula 2.5. Streaming

☐ Streaming.

☐ IoT.

☐ CEP.

- Dados são gerados em tempo real e em fluxo contínuo;
- Existe a necessidade de processamento em tempo real!
- Sistemas desta natureza não podem ser tratados da mesma forma que sistemas em *batch*.

- Alto volume de dados e baixa latência no processamento.
- Para conseguir a baixa latência, um sistema precisa ser capaz de processar os dados sem a necessidade de gravar o dado em disco.
- Para aplicações em tempo real, onde a baixa latência é um requerimento básico, o processamento deve ser feito “In-Stream”, ou seja, à medida que os dados vão chegando, vão sendo processados e analisados em memória.

- Internet of Things ou Internet das Coisas (IoT) se refere à grupos de dispositivos digitais que coletam e/ou transmitem dados pela internet.



- Monitoramento de Operações:
 - Monitoramento de infraestrutura, hardware / software.
- Web Analytics:
 - Sistemas de recomendação / Marketing.
- Mídias Sociais:
 - Twitter, Facebook, Instagram, Youtube, etc.
 - Desafio: Dados não estruturados.
- Mobile:
 - Localização, preferências, etc

- CEP é um **padrão arquitetural de software para processamento de fluxos contínuos de grandes volumes de eventos em tempo real**, correlacionando-os com objetivo de identificar padrões de ocorrências e assim apurar eventos relevantes.
- As regras para descrição dos padrões desejados são definidas em uma linguagem de consulta sobre os fluxos de eventos, de forma similar ao SQL.

- ☑ Existe uma tendência de crescimento no volume e velocidade dos dados.
- ☑ Os métodos de Análise de Dados devem evoluir para atender às novas demandas.

☐ Visualização de Dados.



Aula 2.6. Visualização de dados

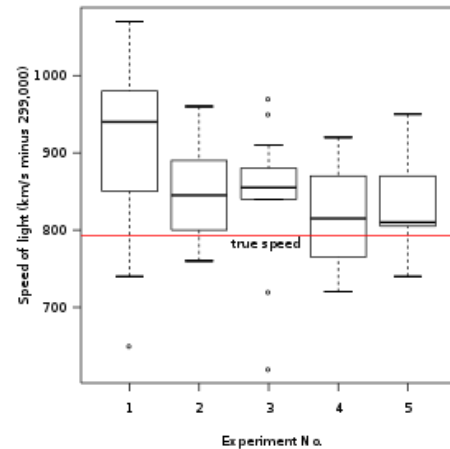
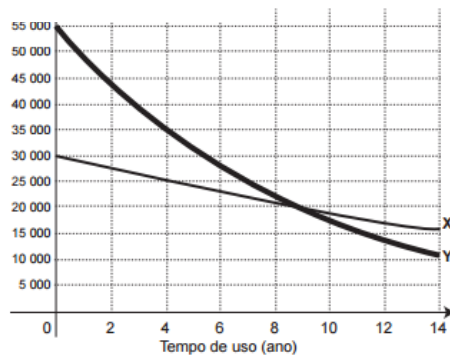
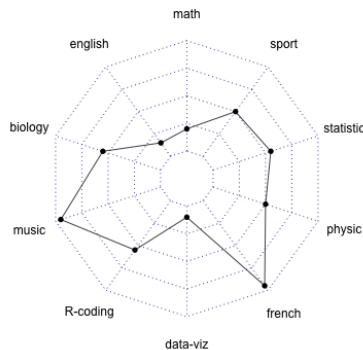
- ☐ Visualização de dados.
- ☐ Tipos comuns de visualização.
- ☐ Data Storytelling.

- A visualização de dados é uma forma acessível de ver e entender exceções, tendências e padrões nos dados. É essencial para analisar as informações e tomar decisões;
- 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.

- Gráficos;
- Tabelas;
- Mapas;
- Infográficos;
- Painéis.

Tipos comuns

- Gráficos
- Tabelas
- Mapas
- Infográficos
- Painéis



Tipos comuns

- Gráficos
- Tabelas
- Mapas
- Infográficos
- Painéis

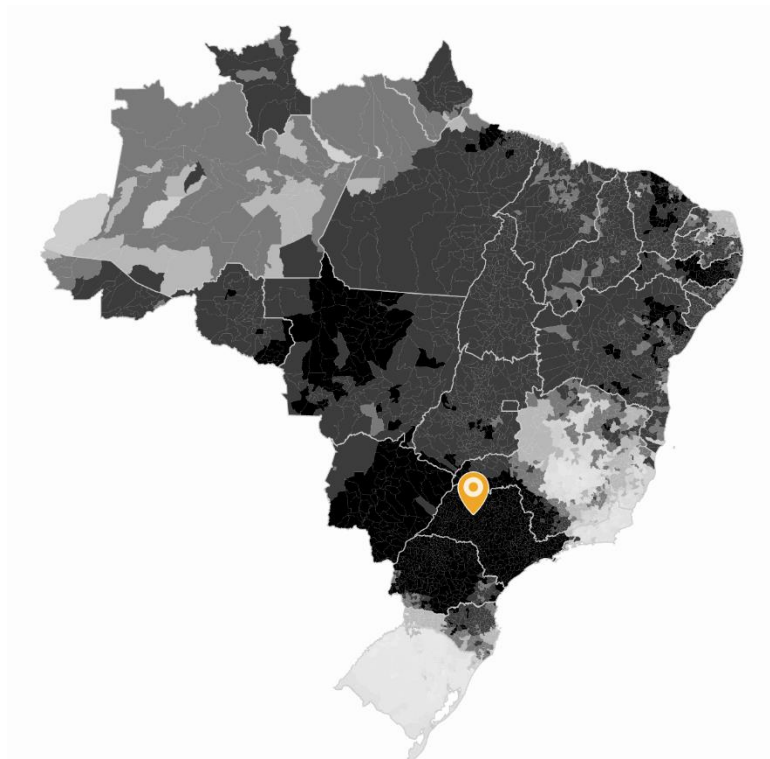
Transaction Amount	Fee Amount	Transaction Type	Service Agent	Transaction Date	Customer Type	Transaction Status
827.25	6.62	Mortgage Loans	Tom Oliver	9/18/2016	Business	Processed
574.38	10.34	Check Cashing	Peter Jacobs	9/22/2016	Business	Pending
53.13	0.21	Loan Settlement	Mark Brown	9/5/2016	Business	Processed
270.75	0.54	Check Cashing	Peter Jacobs	9/15/2016	Business	Pending
617	9.87	Mortgage Loans	Gustavo Sanchez	9/28/2016	Business	Processed
58.75	0.12	Check Cashing	Mark Brown	9/6/2016	Business	Archived

Top 10 cities with the most Global 500 companies

Rank	City	Country	Number of companies	Revenues (\$ mn)	Average Revenue
1	Tokyo	Japan	51	\$2,237,560	\$43,874
2	Paris	France	27	\$1,399,172	\$51,821
3	Beijing	China	26	\$1,361,407	\$52,362
4	New York	United States	18	\$869,150	\$48,286
5	London	United Kingdom	15	\$994,772	\$66,318
6	Seoul	South Korea	11	\$519,351	\$47,214
7	Madrid	Spain	9	\$434,393	\$48,266
8	Munich	Germany	7	\$485,386	\$69,341
8	Moscow	Russia	7	\$380,530	\$54,361
8	Osaka	Japan	7	\$291,492	\$41,642
8	Zürich	Switzerland	7	\$242,595	\$34,656
8	Toronto	Canada	7	\$195,510	\$27,930

Tipos comuns

- Gráficos
- Tabelas
- Mapas
- Infográficos
- Painéis



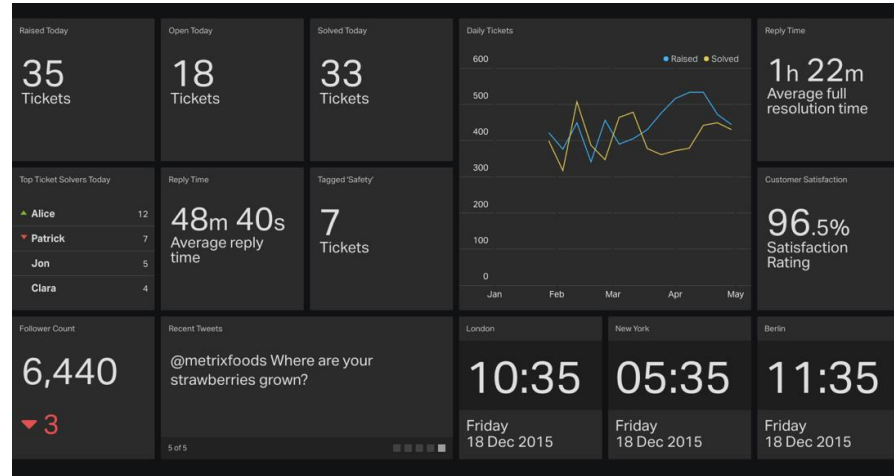
Tipos comuns

- Gráficos
- Tabelas
- Mapas
- Infográficos
- Painéis



Tipos comuns

- Gráficos
- Tabelas
- Mapas
- Infográficos
- Painéis



- Determina quais objetos são oferecidos à nossa atenção.
- Ocorre antes da atenção consciente.

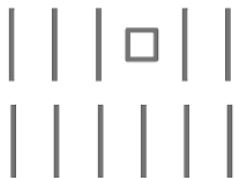
12768679489326456584791209193021483490386
24814001480912808401209475283758237503407
67465748572308402394083590235803275904376
49679024376043765096730964036753067034760
37603760934706734096709347609430697039462
09765902347306047307603476034076034650967

12768679489326456584791209193021483490386
24814001480912808401209475283758237503407
67465748572308402394083590235803275904376
49679024376043765096730964036753067034760
37603760934706734096709347609430697039462
09765902347306047307603476034076034650967

Processamento pré-atentivo



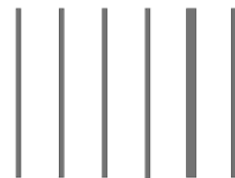
Orientação



Forma



Comprimento de linha



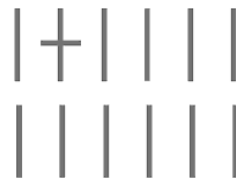
Largura de linha



Tamanho



Curvatura



Adição de marcas



Acercamento



Tonalidade



Intensidade

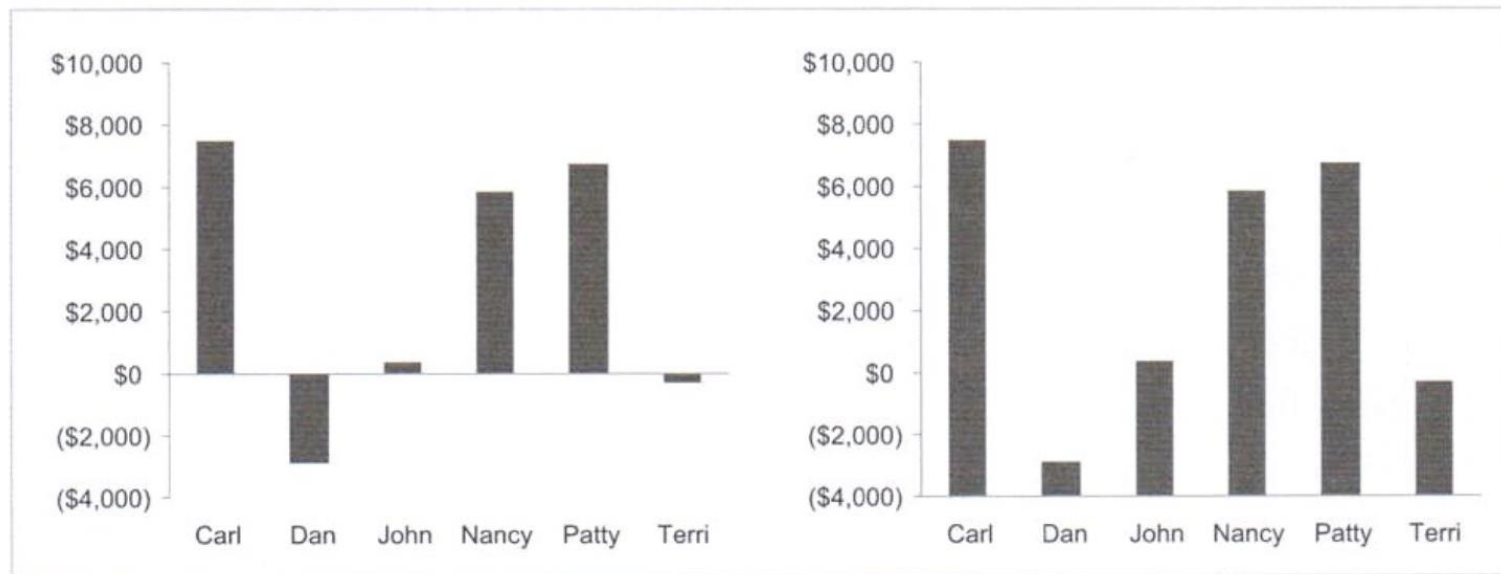


Posição espacial



Movimento

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	<i>Monotype Corsiva</i>	PT Sans Narrow

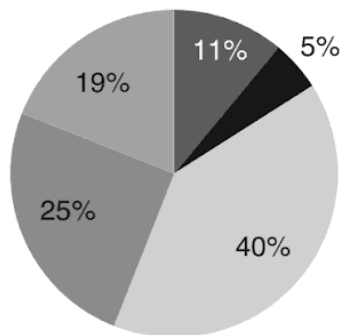


- É o processo que permite organizar as informações e ideias para abordar problemas, tomar decisões, e adquirir conhecimento. O foco passa a ser a experiência do público-alvo na busca por respostas aos problemas encontrados;
- Aplicar Design Thinking em Visualização de Dados significa focar no problema que deve ser resolvido!

- “É a arte de contar história”.
- Técnica fundamental para um Analista de Dados que precisa apresentar seus resultados e/ou sua linha de raciocínio para outras pessoas, tendo elas conhecimento técnico ou não.

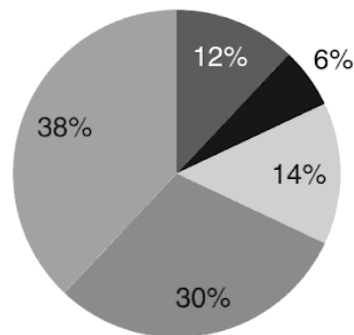
ANTES: Como você se sente
em relação à ciência?

■ Entediado ■ Não muito
interessado ■ Bem ■ Um pouco
interessado ■ Muito
interessado



DEPOIS: Como você se sente
em relação à ciência?

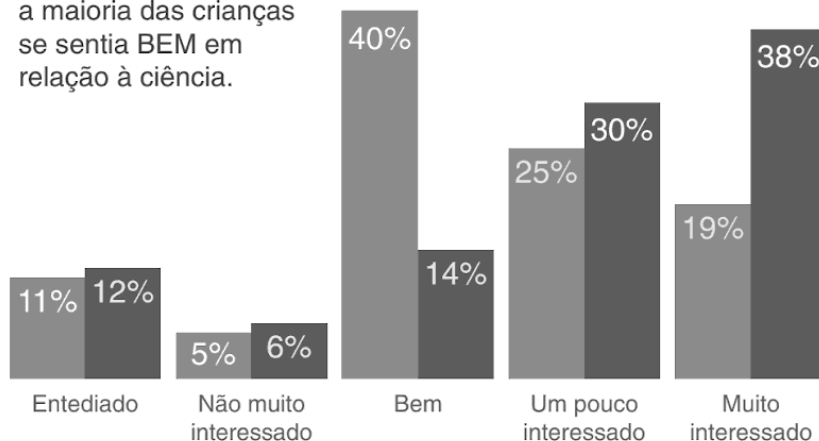
■ Entediado ■ Não muito
interessado ■ Bem ■ Um pouco
interessado ■ Muito
interessado



O programa-piloto foi um sucesso

Como você se sente em relação à ciência?

ANTES do programa,
a maioria das crianças
se sentia **BEM** em
relação à ciência.



DEPOIS
do programa,
mais crianças tinham
algum interesse e
muito interesse
em ciências.

Baseado na avaliação de 100 alunos, realizada antes e depois do programa-piloto (taxa de respostas de 100% nas duas avaliações)

1. Entenda o contexto.
2. Escolha uma apresentação visual adequada.
3. Elimine a saturação.
4. Foque a atenção onde você deseja.
5. Pense como um designer.
6. Conte uma história.

- ☑ A visualização de dados é uma maneira simples e rápida de transmitir conceitos de modo universal
- ☑ Data Storytelling é o ato de você explicar o que você fez, como fez e por quê fez, tudo isso de forma que mantenha seu leitor ou ouvinte engajado.

☐ Capítulo 3:

- Frameworks e Ferramentas.



Fundamentos de Data Analytics

Capítulo 3. Frameworks e Ferramentas

Prof. Angelo Assis



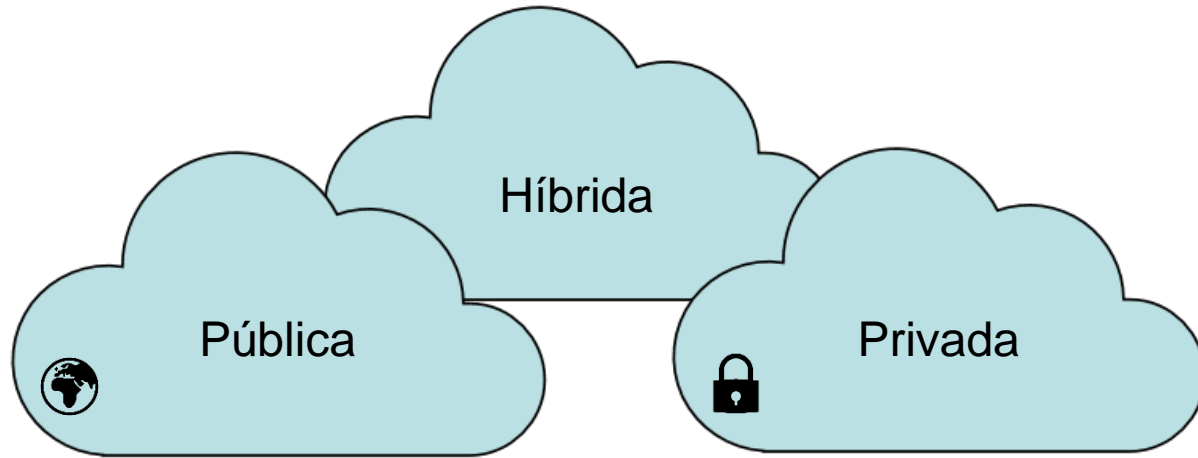
Aula 3.1. Computação em Nuvem

Nesta aula

- ☐ Computação em Nuvem.
- ☐ Modelos de implantação.
- ☐ Tipologias.

- 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.

■ Nuvem Pública, Privada e Híbrida



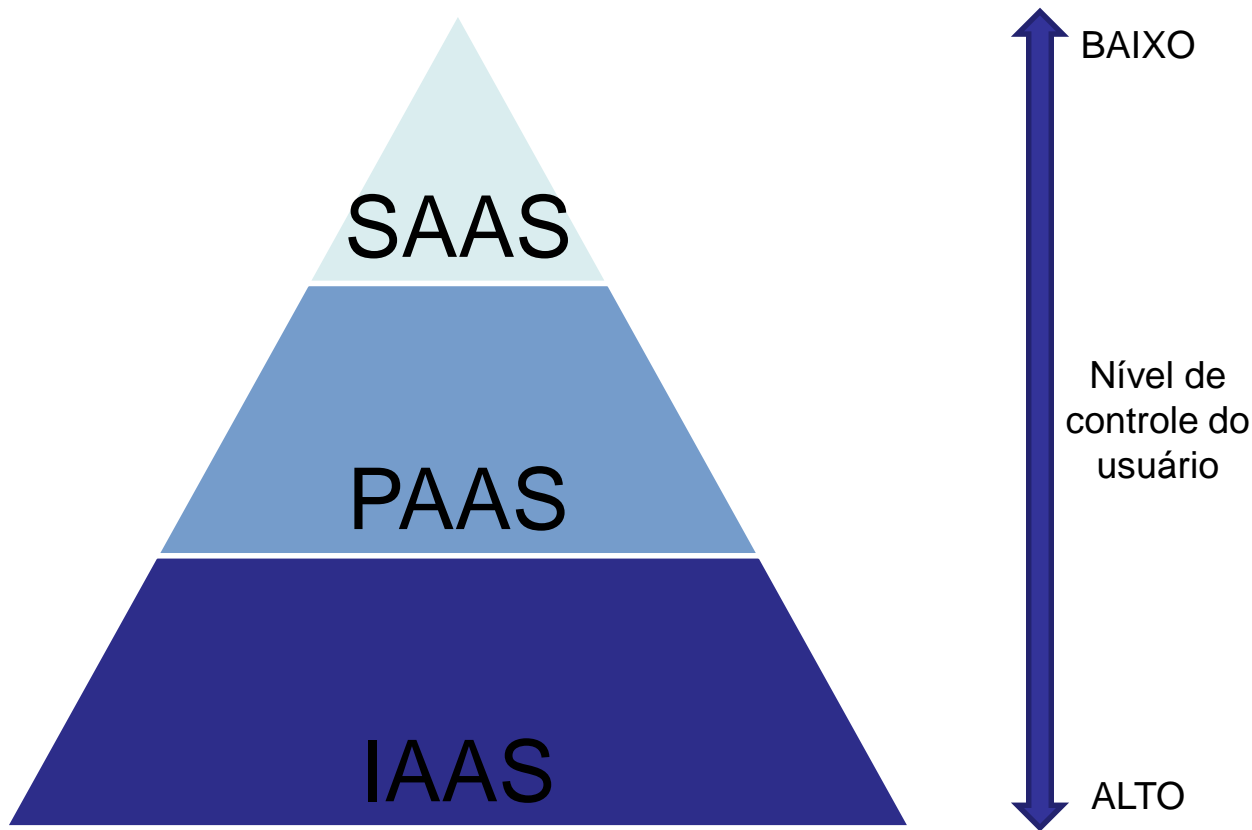
- É o tipo mais comum.
- Compartilha recursos com outros “usuários” da nuvem.
- Acesso feito por um navegador da Web – Internet.
- Reduz custos.

- Exclusivas para uma única empresa/organização.
- Pode estar localizada fisicamente no datacenter local da sua organização.
- Maior flexibilidade.
- Maior segurança.

- É a composição de duas ou mais nuvens.
- Possibilidade de alternar entre o modelo público e o privado conforme a necessidade do negócio.
- Custo-benefício.
- Indicado para empresas que possuem uma boa infraestrutura interna e também necessitam da nuvem pública.

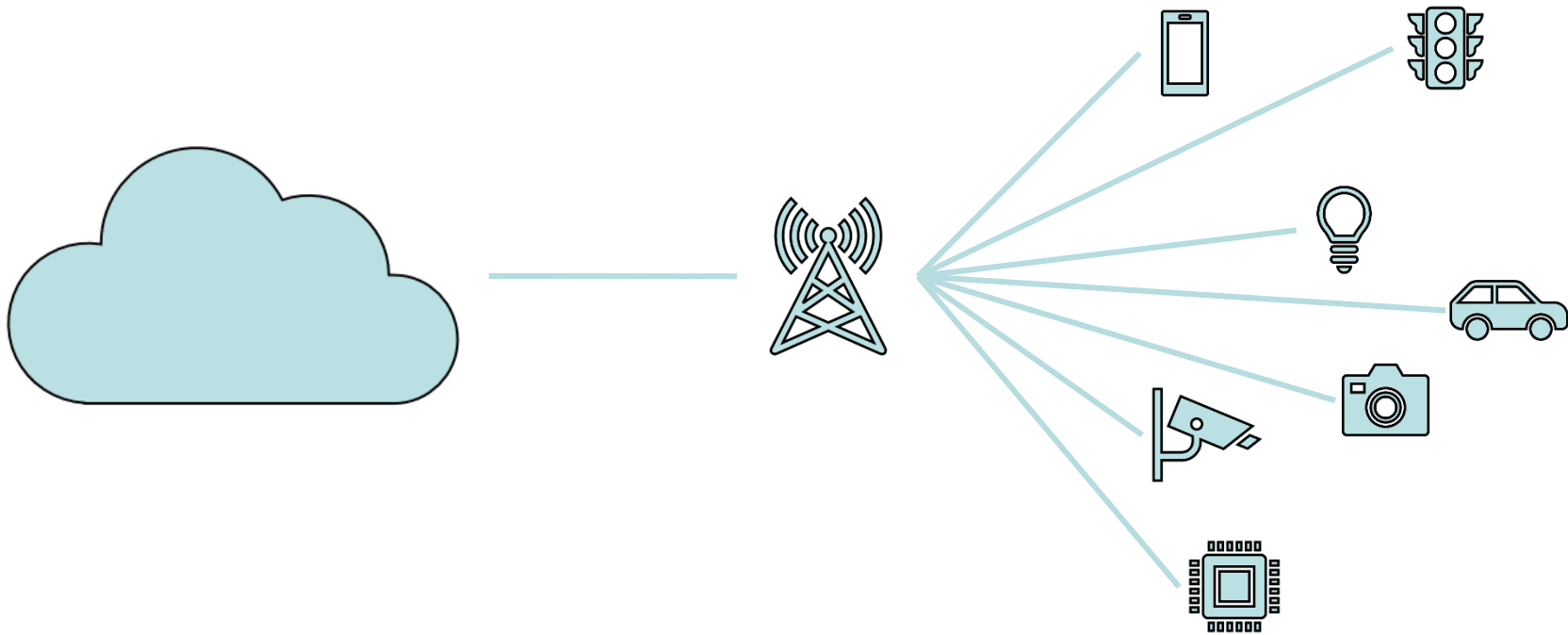
- SaaS:
 - A aplicação funciona diretamente na nuvem. Não há necessidade de instalação na máquina do cliente.
- IaaS:
 - Disponibiliza os recursos necessários para que o usuário faça a implantação, configuração e utilização suas aplicações.
- PaaS:
 - O provedor fornece uma plataforma que é usada para desenvolver e disponibilizar aplicações.

Tipos de Oferta de Computação em Nuvem



- “Dispositivos que carregam a habilidade para realizar processamentos e análises avançados”
- Só foi possível com a evolução dos dispositivos.
- Permite processamento em tempo real.
- Cloud Computing é um modelo mais centralizado.

Edge Computing



- ☑ Computação em Nuvem e BI são uma ótima combinação.
- ☑ Redução de custo com infraestrutura de TI.
- ☑ Dados acessíveis.
- ☑ Escalabilidade.
- ☑ Alta disponibilidade.
- ☑ Atenção ao tráfego de dados.
- ☑ Pode enfrentar barreiras culturais.

☐ Processamento Paralelo e Distribuído.



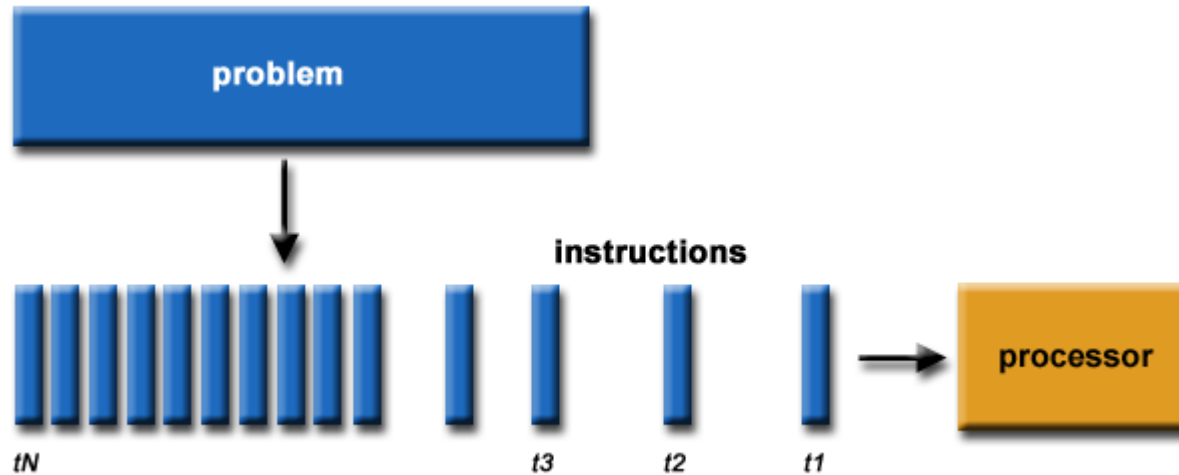
Aula 3.2. Processamento Paralelo e Distribuído

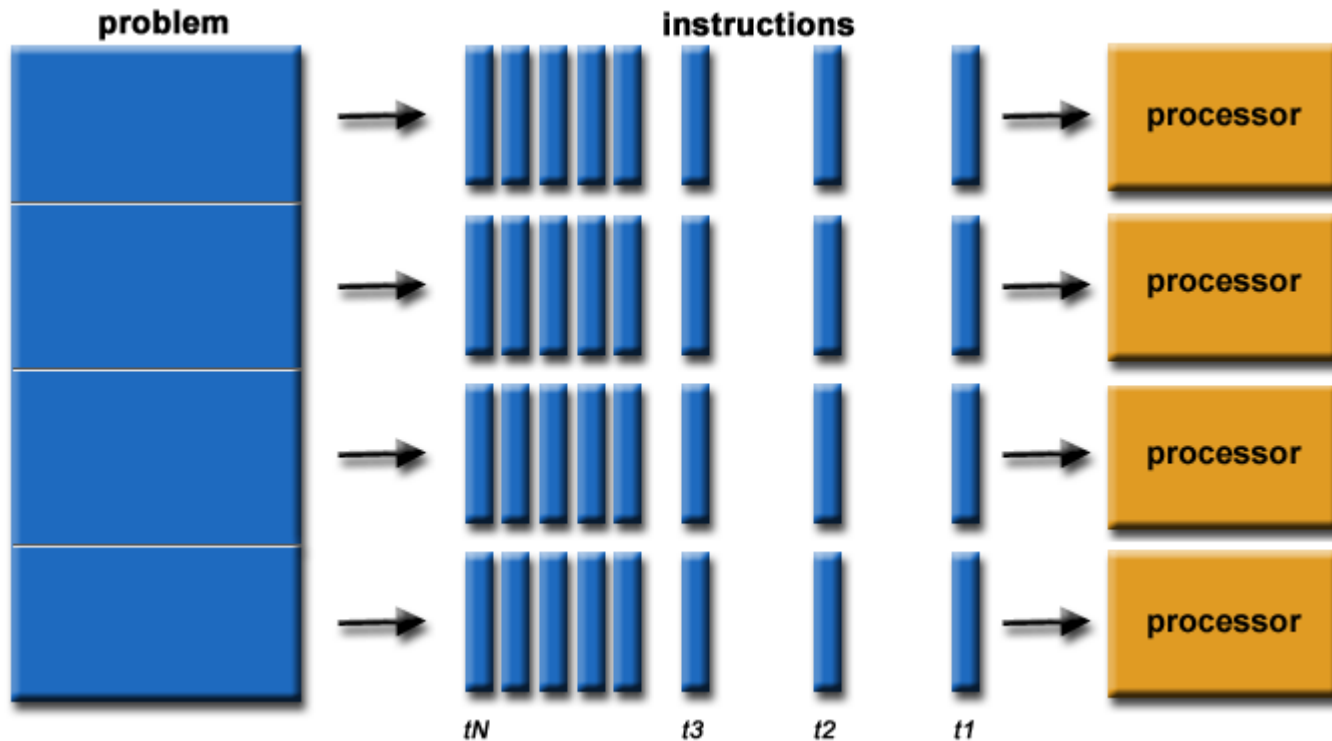
- ☐ Processamento Paralelo.
- ☐ Processamento Distribuído.
- ☐ Desafios.

- Utilizam melhor o poder de processamento.
- Apresentam um melhor desempenho.
- Permitem compartilhar dados e recursos.
- Atendem um maior número de usuários.

- 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.

“Programação Paralela é a prática de dividir uma determinada tarefa em tarefas menores que possam ser executadas de forma simultânea e independente.”



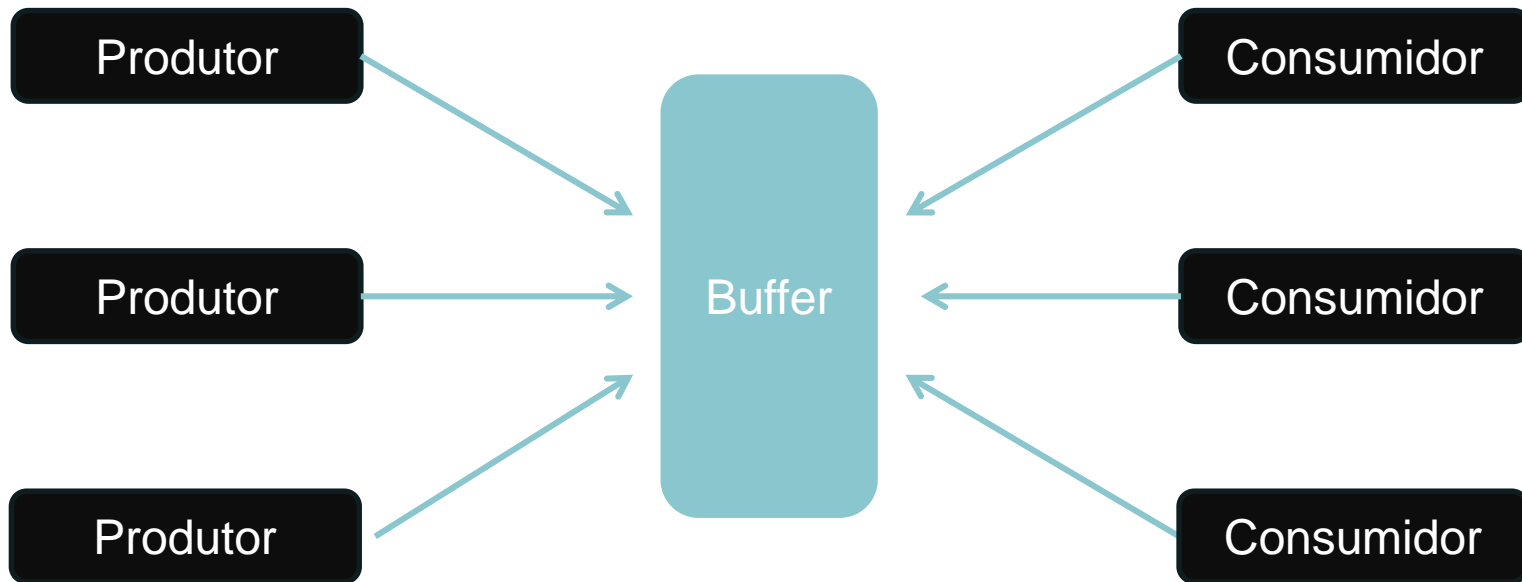


- SpeedUp: é a razão entre o tempo de execução sequencial e o tempo de execução paralelo.

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

- $T(1)$ = Tempo de execução com um processador
- $T(p)$ = Tempo de execução com p processadores

Arquitetura Produtor - Consumidor



“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.”

- Comunicação entre servidores:
 - Latência vs Largura de Banda.
- Escalabilidade.
- Sincronização.
- Replicação de dados.

- Desenvolver, gerenciar e manter o sistema.
- Controlar o acesso concorrente a dados e a recursos compartilhados
- Evitar que falhas de máquinas ou da rede comprometam o funcionamento do sistema
- Garantir a segurança do sistema e o sigilo dos dados trocados entre máquinas
- Lidar com a heterogeneidade do ambiente

	Paralelo	Distribuído
Acoplamento	Forte	Fraco
Previsibilidade	Mais previsível	Mais sensível a falhas (rede)
Tempo de comunicação entre processos	Desprezível	Depende da rede
Controle	Único	Independente

Próxima aula

☐ Apache Kafka.



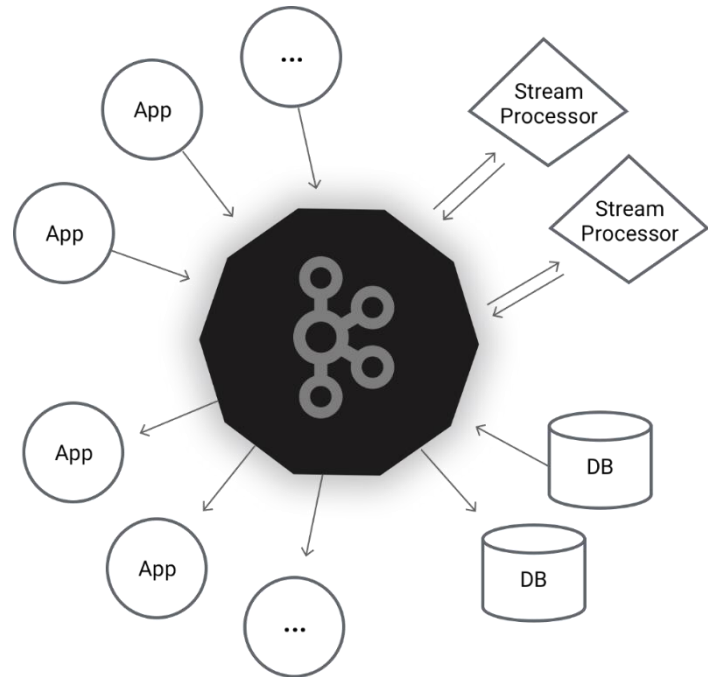
Aula 3.3. Apache Kafka

- ❑ Apache Kafka.

“Apache Kafka é uma plataforma distribuída de mensagens e streaming.”

Apache Kafka

- Você produz uma mensagem.
- Essa mensagem é anexada em um tópico.
- Você então consome essa mensagem



- **Producer API:**
 - Permite publicar os dados em um ou mais tópicos.
- **Consumer API:**
 - Permite a inscrição em um ou mais tópicos e processe os dados.
- **Streams API:**
 - Permite que uma aplicação seja um “processador de fluxo”.
- **Connector API:**
 - Permite criar produtores e consumidores para tópicos do Kafka.

- Tópicos são conjuntos de dados. É uma fila onde um registro (mensagem) é publicado.
- Cada tópico pode agrupar diversas mensagens e pode ter um ou mais consumidores.
- O consumidor não pode alterar o estado de uma mensagem.

- É responsável por produzir e enviar uma mensagem para um tópico específico.
- Uma vez que uma mensagem é produzida em um tópico o próprio Kafka organiza a mensagem e garante sempre a ordem das mensagens produzidas.

- Os consumidores “assinam” um tópico que possam ler os dados armazenados.
- Vários consumidores podem assinar o mesmo tópico, podendo ser em grupo ou individualmente.

- ☑ “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.”

Próxima aula

☐ Apache Hadoop.

☐ Apache Spark.



Aula 3.4. Hadoop e Spark

☐ Apache Hadoop.

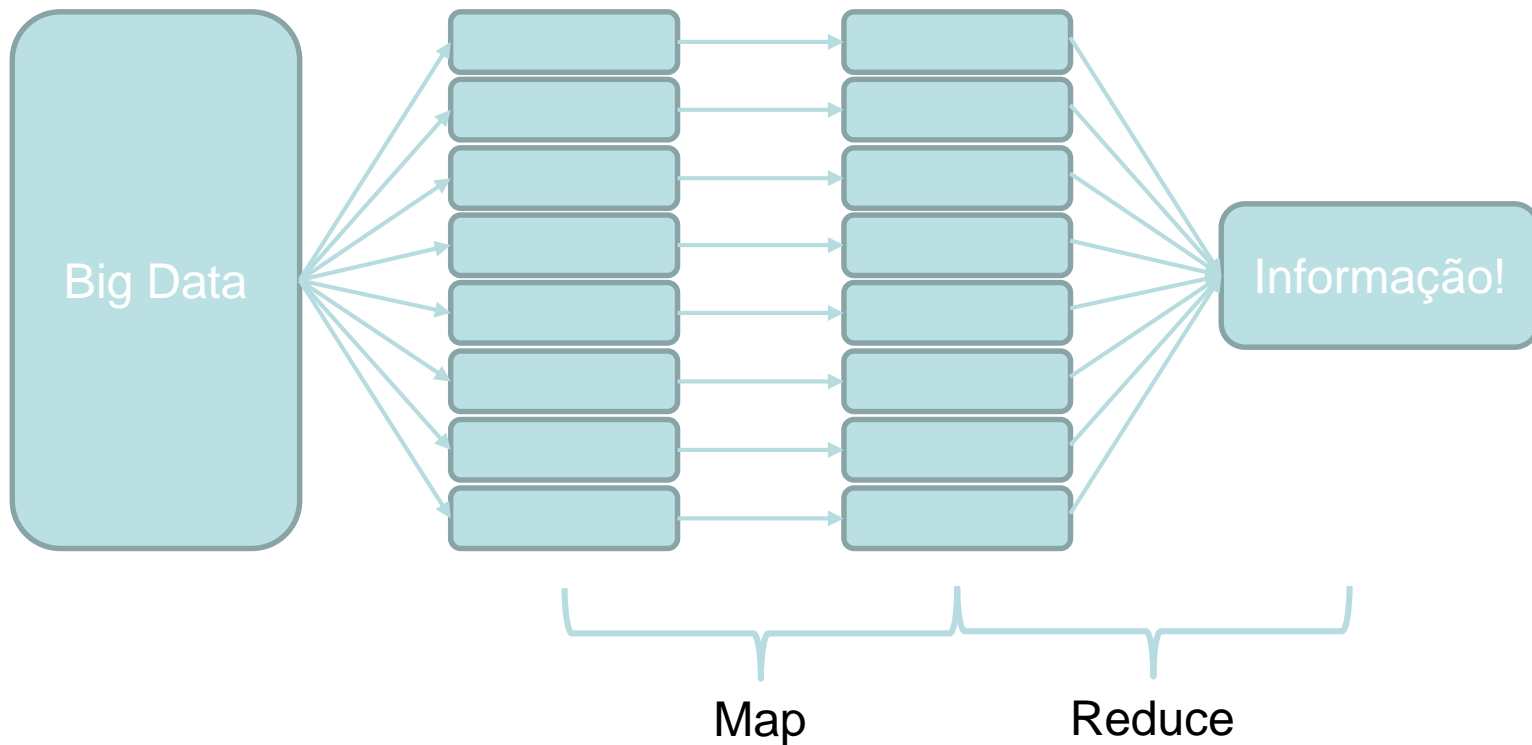
☐ Apache Spark.

“É um framework para desenvolvimento de aplicações que necessitam de armazenamento e processamento distribuído de grandes conjuntos de dados.”

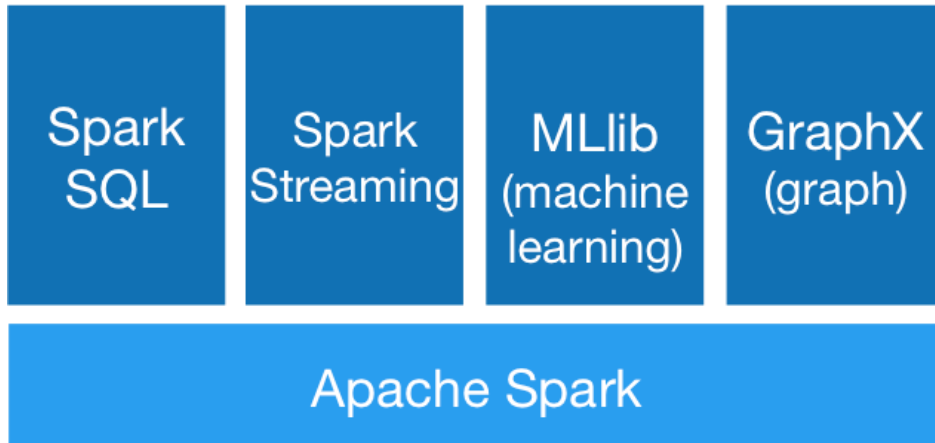
- Hadoop Distributed File System (HDFS): Projetado para abranger grandes clusters de servidores e escalar até centenas de petabytes e milhares de servidores.

- No Hadoop os dados são tratados como **pares chave / valor** (*Key / Value*) e processados por meio de duas funções principais, **Map e Reduce**.
- Map Reduce pode tirar vantagem da localidade de dados:
 - O processamento ocorre próximo ao armazenamento em cada nó no cluster, a fim de reduzir a distância que deve ser transmitido.

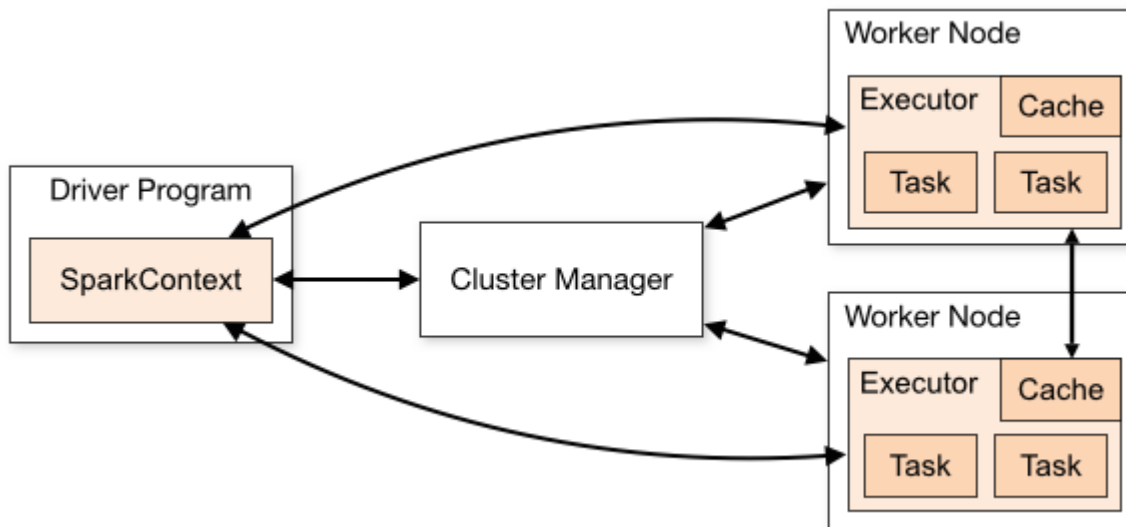
Modelo Map Reduce



- Processamento em memória!
- Estende o modelo de programação MapReduce popularizado pelo Apache Hadoop.



Apache Spark



- ☑ Spark vs Hadoop.
- ☑ Spark necessita de muita memória RAM.
- ☑ Hadoop continua sendo muito utilizado.

☐ Bancos de Dados Relacionais e Não-Relacionais (NoSQL).



Aula 3.5. Bancos Relacionais e Não-Relacionais (NoSQL)

☐ Bancos de Dados Relacionais.

☐ MongoDB.

☐ Cassandra.

☐ Neo4J.

☐ Couchbase.

- Os dados são organizados em tabelas.
- As tabelas possuem “chaves” para identificação única dos registros.
- Consultas SQL.
- É possível criar índices para otimizar a leitura.
- Atenção à “Integridade dos 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

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

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

- NoSQL = Not Only SQL.
- Os dados são modelados de forma diferente, não necessariamente em tabelas.
- Provê escalabilidade.
- Flexibilidade na modelagem.

- Um dos principais bancos NoSQL.
- É orientado a **documentos**.
- Armazena **JSON**.
- Organiza os documentos em **coleções**.
- Permite criação de índices.
- Possui conector para Spark e BI.



- Armazenamento em colunas:
 - KeySpace, família de colunas e a coluna.
 - Cada coluna possui o nome do campo, o seu valor e o timestamp.
- Não existe o conceito de transação.
- Também não existem relacionamentos e constraints.
- Consultas podem ser realizadas com o CQL.



- Orientado a grafos:
 - Nós, arestas, propriedades.
- Topologia dos dados é mais importante que os dados.
- Consultas realizadas com *Cypher*.



- Chave-valor e documento:
 - JSON.
- É possível criar views com o conceito de MapReduce.
- Possui interface gráfica intuitiva.
- Dados são armazenados em buckets.
- Funciona bem em cluster.
- Consultas realizadas com N1QL.



- ☑ Bancos de Dados Relacionais possuem mais integridade.
- ☑ NoSQL não surgiu para substituir o SQL.
- ☑ NoSQL:
 - Documentos, chave-valor, grafos, colunas.
 - Indicado para necessidades maiores de armazenamento.

■ Próxima aula

☐ Couchbase.



Aula 3.6. Couchbase

Instalação do Couchbase

The screenshot shows a web browser window with the Couchbase website. The address bar shows 'couchbase.com/get-started'. The navigation bar includes links for PRODUCTS, SOLUTIONS, CUSTOMERS, RESOURCES, and COMPANY, along with a search icon and a 'Downloads' button. The main content area is titled 'Download Couchbase Server' and features a 'Collapse' button. Below this, there are tabs for 'ENTERPRISE' and 'COMMUNITY', with a 'Compare >' link. The 'COMMUNITY' tab is selected, showing 'Couchbase Server 6.5.1 Community'. The text describes the release as the first maintenance release in the 6.5.x series, highlighting support for bounded polygons in geospatial search queries. To the right, there is a 'Windows' dropdown menu and a 'Download' button, with a 'Release notes' link below. A vertical 'Contact Us' button is visible on the right side of the page.

Get Started, Learn NoSQL | Couch x +

couchbase.com/get-started

Anônima

Couchbase NoSQL

PRODUCTS SOLUTIONS CUSTOMERS RESOURCES COMPANY

Downloads

Server

MODIE

DOWNLOAD COUCHBASE SERVER

Download Couchbase Server

Collapse

ENTERPRISE COMMUNITY Compare >

Couchbase Server 6.5.1 Community

Couchbase Server 6.5.1, released in April 2020, is the first maintenance release in the 6.5.x series for Couchbase Server. This release adds support for bounded polygons in geospatial search queries in addition to improvements and important bug fixes in various components.


Windows Download

Release notes

Contact Us



Instalação do Couchbase

 Couchbase > New Cluster

Cluster Name


Create Admin Username

Create Password

Confirm Password

[< Back](#)

[Next: Accept Terms](#)

 Couchbase > New Cluster

Terms and Conditions Community Edition

Last updated February 28, 2020. Replaces all prior versions.

Couchbase, Inc.
Community Edition License Agreement

This Couchbase Community Edition License Agreement between you and Couchbase, Inc., governs your use of the community edition of Couchbase's software accompanying this agreement, including but not limited to Couchbase Server Community, Couchbase Synch Gateway Community and Couchbase Lite Community, and any Couchbase services or updates for that software, in addition

☒ I accept the [terms & conditions](#)

[< Back](#) [Finish With Defaults](#) [Configure Disk, Memory, Services](#)

Instalação do Couchbase



Couchbase > New Cluster > Configure

Host Name / IP Address

Fully-qualified domain name

127.0.0.1

Data Disk Path

Path cannot be changed after setup

c:/Program Files/Couchbase/Server/var/lib/couchbase/data

Free: 88 GB

Indexes Disk Path

Path cannot be changed after setup

c:/Program Files/Couchbase/Server/var/lib/couchbase/data

Free: 88 GB

Service Memory Quotas

Per service / per node

☒ Data

9096

MB

☒ Query

☒ Query

☒ Index

512

MB

☒ Search

512

MB

TOTAL QUOTA 10120 MB

RAM Available 16270MB Max Allowed Quota 15246MB

Index Storage Setting

☒ Standard Global Secondary

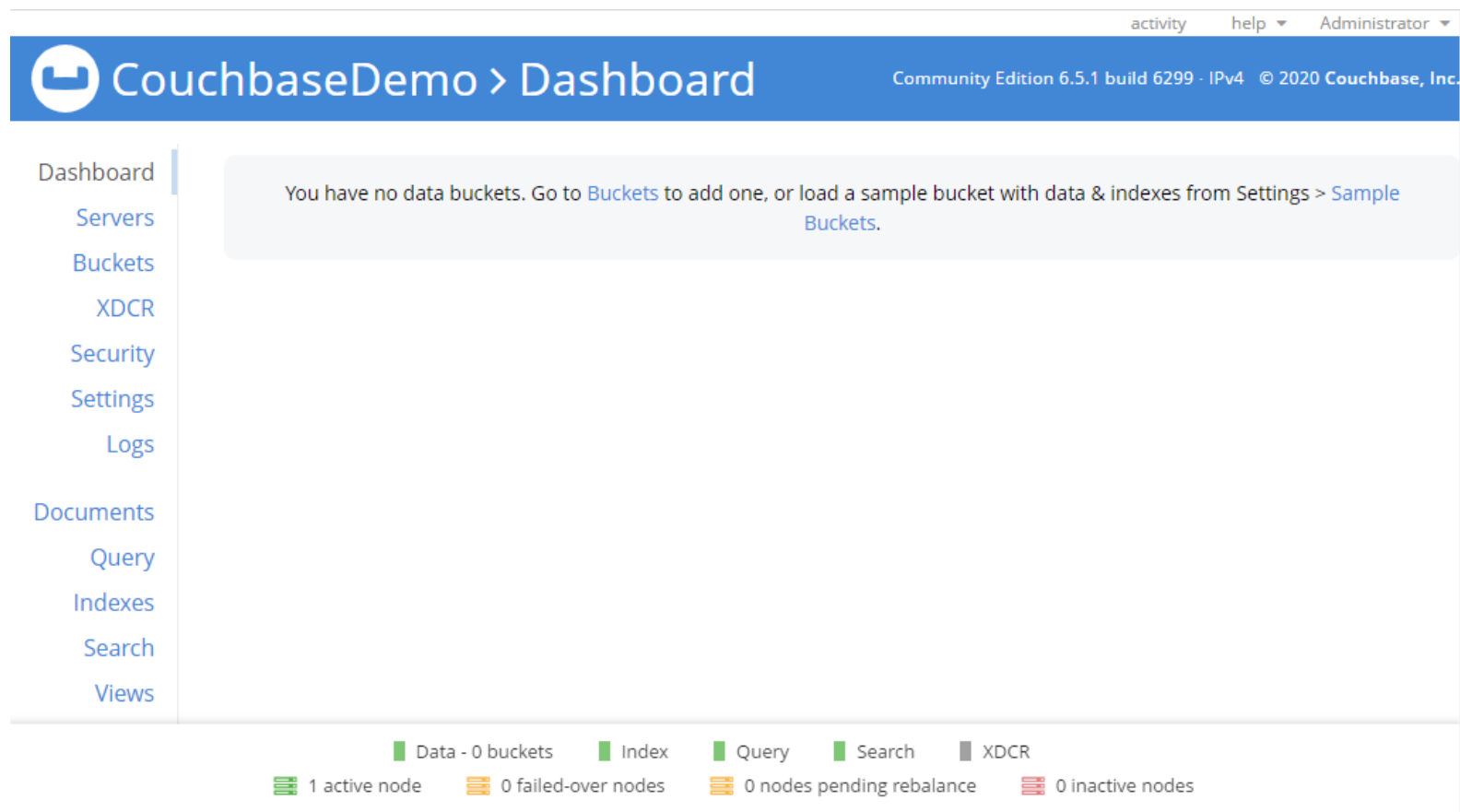
☐ Memory-Optimized ⓘ

☒ Share usage information with Couchbase and get software update notifications. ⓘ

[< Back](#)

[Save & Finish](#)

Instalação do Couchbase



The screenshot shows the Couchbase Demo Dashboard interface. At the top right, there are links for 'activity', 'help', and 'Administrator'. The main header is a blue bar with the Couchbase logo and the text 'CouchbaseDemo > Dashboard'. To the right of the header, it says 'Community Edition 6.5.1 build 6299 - IPv4 © 2020 Couchbase, Inc.'. On the left side, there is a vertical navigation menu with links: 'Dashboard', 'Servers', 'Buckets', 'XDCR', 'Security', 'Settings', 'Logs', 'Documents', 'Query', 'Indexes', 'Search', and 'Views'. The main content area has a light blue background with a message: 'You have no data buckets. Go to [Buckets](#) to add one, or load a sample bucket with data & indexes from Settings > [Sample Buckets](#).' At the bottom, there is a status bar with several indicators: 'Data - 0 buckets', 'Index', 'Query', 'Search', 'XDCR', '1 active node', '0 failed-over nodes', '0 nodes pending rebalance', and '0 inactive nodes'.

activity help Administrator

CouchbaseDemo > Dashboard Community Edition 6.5.1 build 6299 - IPv4 © 2020 Couchbase, Inc.

Dashboard

Servers

Buckets

XDCR

Security

Settings

Logs

Documents

Query

Indexes

Search

Views

You have no data buckets. Go to [Buckets](#) to add one, or load a sample bucket with data & indexes from Settings > [Sample Buckets](#).


Data - 0 buckets Index Query Search XDCR

1 active node 0 failed-over nodes 0 nodes pending rebalance 0 inactive nodes

Instalação do Couchbase

IGTi

activityhelp ▼Administrator ▼

 CouchbaseDemo > Buckets

ADD BUCKET

Dashboard

Servers

Buckets

XDCR

Security

Settings

Logs

Documents


Query

Indexes

Search

Views

You have no data buckets. Use "ADD BUCKET" above to create one, or load a sample bucket with data & indexes.

 CouchbaseDemo > Settings

GeneralAuto-CompactionEmail AlertsSample Buckets ▾

Dashboard
Servers
Buckets
XDCR
Security
Settings
Logs
Documents
Query
Indexes
Search
Views

Sample Buckets

Sample buckets contain example data, views, and indexes for your experimentation.

Sample buckets — like all buckets in Couchbase Server 5.0+ — can only be accessed by a user with privileges for that bucket.

Available Samples

☒ beer-sample

☒ gamesim-sample

☒ travel-sample

Load Sample Data

Installed Samples

none

- Demonstração

Próxima aula

☐ Capítulo 4:

- Indústria 4.0.



Fundamentos de Data Analytics

Capítulo 4. Indústria 4.0

Prof. Angelo Assis

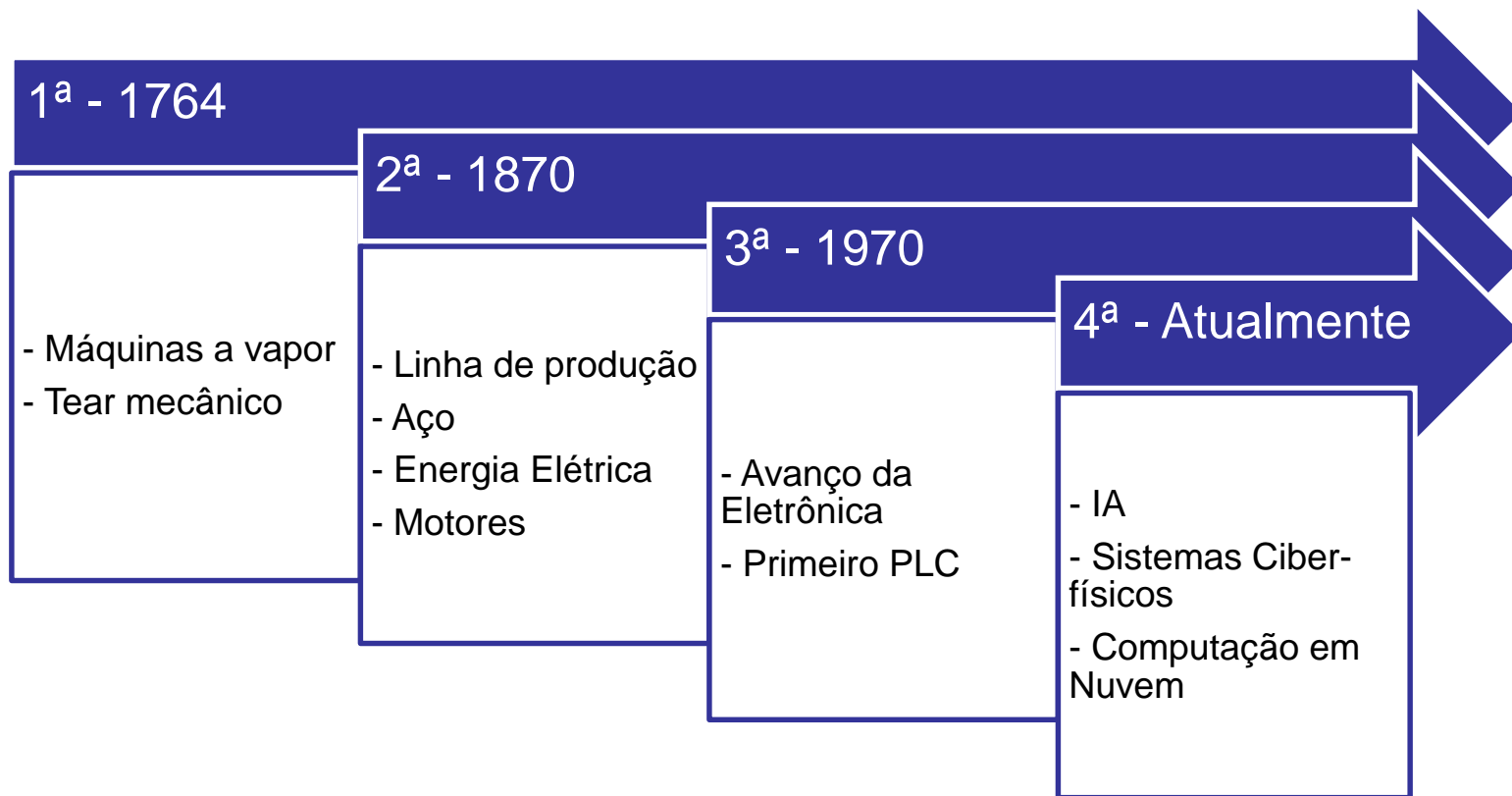


Aula 4.1. Indústria 4.0

Nesta aula

☐ Indústria 4.0.

☐ Cultura Orientada a Dados.



- É um novo conceito de indústria.
- 4ª Revolução industrial.
- Engloba as principais inovações tecnológicas dos campos de automação, controle e tecnologia da informação, aplicadas aos processos de manufatura.

- Operação em tempo real.
- Virtualização.
- Descentralização.
- Orientação a Serviço.
- Modularidade.

- Conectividade.
- Redução de custos.
- Ajustes muito mais rápidos na linha de produção.
- Personalização dos produtos.
- Novos modelos de negócio.

Os profissionais da Indústria 4.0

- Novas habilidades e qualificações.
- Redução de funções repetitivas e braçais.
- Outras funções devem surgir.



- ☑ É um caminho sem volta.
- ☑ As empresas exigirão novos perfis.

Próxima aula

- ☐ Cultura orientada a dados.



Aula 4.2. Cultura orientada a dados

- ☐ Cultura de Data Analytics.
- ☐ Como engajar.
- ☐ Empresas orientadas a dados.

- Abordagem dos problemas de maneira objetiva.
- Criatividade na coleta e uso dos dados.
- Todo mundo tem acesso a algum dado.
- Pessoas em primeiro lugar.
- Mentalidade orientada para as análises.



- Democratização dos dados.
- Comunicação.
- Engajamento.
- Ferramentas.
- Tomada de decisão.



- Tomar uma decisão e só depois buscar os dados.
- Confiar nos achismos e heurísticas.
- Achar que é complexo demais.
- Subestimar os dados que já estão disponíveis.
- Trabalhar com dados obsoletos.
- Alimentar uma cultura de individualismo.



O que fazer

- Deixar claro para o time a importância dos dados.
- Buscar evoluir a capacidade analítica de sua empresa.
- Automatizar a coleta de dados. Organizar os dados.
- Pragmatismo na escolha da tecnologia.
- Contratar profissionais *data driven*.



Como saber se sua empresa já é data driven?

- As decisões são tomadas com base em dados?
- Seu time sabe o que significa data driven?
- Você utiliza ferramentas que possibilitam sempre capturar novos dados e melhorar sua análise?
- Seus gestores justificam sua decisões com base em dados?
- Você está contratando profissionais data driven?

- ☑ A cultura organizacional é construída pelas pessoas.
- ☑ Escolha bem as tecnologias, ferramentas e pessoas.
- ☑ Cada empresa tem uma maturidade diferente.
 - Assim como cada área da empresa.

Próxima aula

☐ O Cientista de Dados.



Aula 4.3. O Cientista de Dados

- ☐ Atividades do Cientista de Dados.
- ☐ Habilidades necessárias.
- ☐ Equipe multidisciplinar.

- 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.

Hard Skills

- Matemática / Estatística
- Ciência da Computação
 - Inteligência Artificial
 - Lógica de Programação
 - Linguagens de Programação
 - Banco de Dados



Soft Skills

- Comunicação e Liderança
- Conhecimento do Negócio
- Criatividade
- Pensamento analítico
- Resolução de problemas
- Gestão de Projetos

- Achar que o aprendizado é fácil e rápido.
- Aprender muitos conceitos ao mesmo tempo.
- Começar por problemas muito complexos.
- Focar apenas na programação.

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

- O sucesso começa com a diversidade.
- Cuidado! Falta de alinhamento e padronização no processo.
- Explore os potenciais individuais.
- Foco no projeto e no resultado e não na função.
- Autonomia para a equipe e as pessoas.



- ☑ Ser um cientista de dados vai muito além de programação e estatística.
- ☑ O caminho é longo.
- ☑ Equipes multidisciplinares podem ser uma ótima solução.

■ Próxima aula

□ PIMS.



Aula 4.4. PIMS

□ PIMS.

- Sistema de gerenciamento de informações de processo.
- Coleta de dados de chão de fábrica e sistemas corporativos.
- Armazenamento de dados históricos.
- Eficiente compressão dos dados.
- Proporciona uma visão unificada do processo:
 - Tempo real e histórico.



YOKOGAWA



OSIsoft®



SIEMENS

- ☑ PIMS é essencial para indústrias.
- ☑ Possibilita automatização na coleta dos dados.
- ☑ Centraliza as informações.

■ Próxima aula

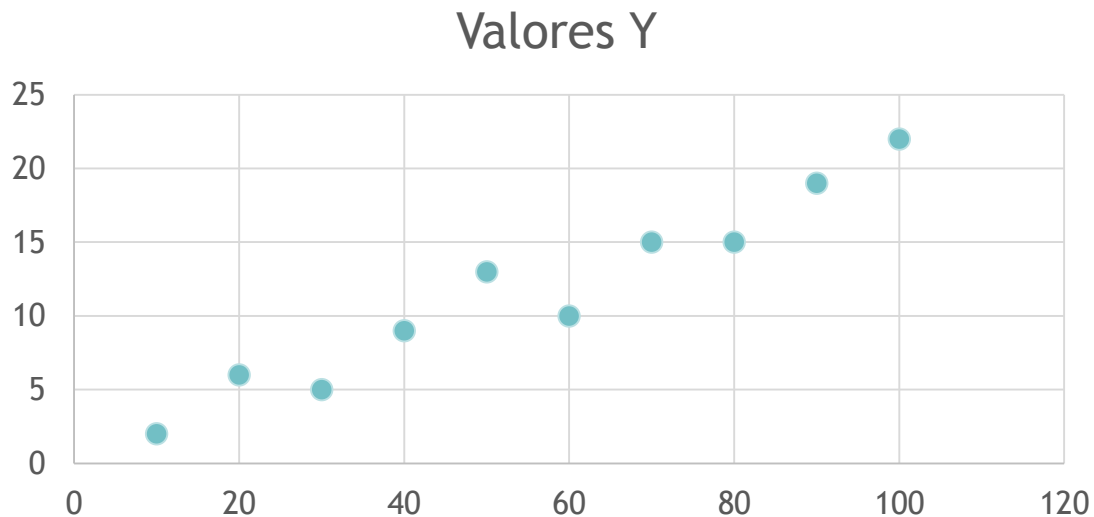
□ Regressão e Correlação.



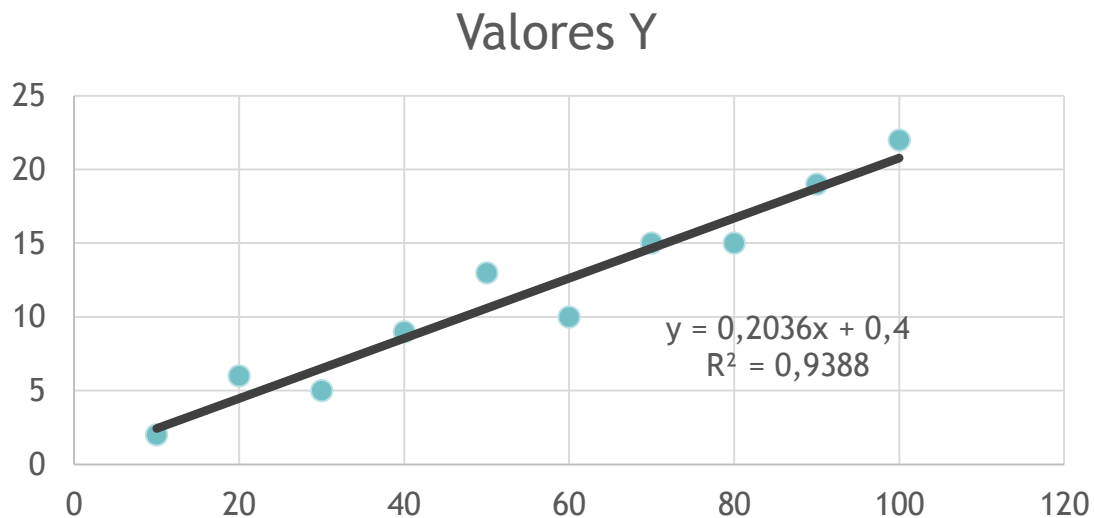
Aula 4.5. Regressão e Correlação

- Duas técnicas relacionadas, com objetivo de estimar o grau de relação existente entre duas variáveis.
- **Regressão:** Uma equação matemática que descreve o relacionamento entre variáveis.
 - Pode ser linear, polinomial, logística, etc
- **Correlação:** O grau de relacionamento entre duas variáveis
 - Varia de -1 a 1

- A regressão linear utiliza os pontos de dados para encontrar a melhor linha de ajuste para modelar essa relação.



- A regressão linear utiliza os pontos de dados para encontrar a melhor linha de ajuste para modelar essa relação.



- Demonstração com Google Colab

□ Capítulo 5:

– Cases.



Fundamentos de Data Analytics

Capítulo 5. Cases Big Data e Analytics

Prof. Angelo Assis



Aula 5.1. Cases Nestlé e Netflix

Nesta aula

☐ Nestlé.

☐ Netflix.

- Empresa do setor de alimentos e bebidas.
- Vídeo do Greenpeace → Kit Kat.
- Atingiu 1,5 milhões de pessoas!



Nestlé



- Empresa do setor de alimentos e bebidas.
- Vídeo do Greenpeace → Kit Kat.
- Atingiu 1,5 milhões de pessoas!
- Comunicado oficial + Facebook.
- Monitoramento de redes sociais.
- Time de aceleração digital.
- Reputação 16ª → 12ª.



Nestlé



- Serviço de streaming de vídeos.
- 130 países, 75 milhões de assinantes.
- Comportamento do consumidor!
- Competição:
 - US\$ 1 milhão para o melhor algoritmo.

The word "NETFLIX" in a large, bold, red, sans-serif font.

- ☑ Atenção ao uso das redes sociais.
- ☑ Conheça seus clientes e usuários.
- ☑ Melhore sempre!

■ Próxima aula

□ Cases.



Aula 5.2. Cases McDonald's e Tinder

☐ McDonald's.

☐ Tinder.

- Maior rede de restaurantes do mundo.
- 75 sanduíches vendidos a cada segundo.
- 34 mil restaurantes, 62 milhões de pessoas, 118 países.



- Maior rede de restaurantes do mundo.
- 75 sanduíches vendidos a cada segundo.
- 34 mil restaurantes, 62 milhões de pessoas, 118 países.
- Otimização do tráfego de drive-thru.



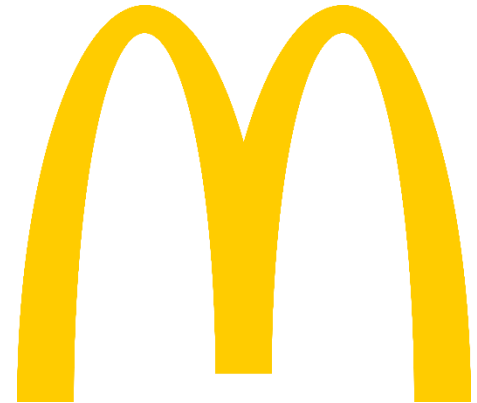
- Maior rede de restaurantes do mundo.
- 75 sanduíches vendidos a cada segundo.
- 34 mil restaurantes, 62 milhões de pessoas, 118 países.
- Otimização do tráfego de drive-thru.
- Novas opções de menu.



- Maior rede de restaurantes do mundo.
- 75 sanduíches vendidos a cada segundo.
- 34 mil restaurantes, 62 milhões de pessoas, 118 países.
- Otimização do tráfego de drive-thru.
- Novas opções de menu.
- Controle de qualidade.



- Maior rede de restaurantes do mundo.
- 75 sanduíches vendidos a cada segundo.
- 34 mil restaurantes, 62 milhões de pessoas, 118 países.
- Otimização do tráfego de drive-thru.
- Novas opções de menu.
- Controle de qualidade
- Monitoramento de redes sociais.



- Aplicativo de relacionamentos.
- 50 milhões de usuários.
- Geolocalização.
- Perfil de usuário, interesses em comum.
- Monitoramento de chat.
- Limitações no *swipe*.



- ☑ Mesmo nas pequenas opções que fazemos no dia a dia, podemos estar contribuindo para grandes mudanças no futuro.

■ Próxima aula

□ Cases.



Aula 5.3. Cases

Nesta aula

☐ Copa do Mundo.

☐ Airbnb.

☐ Netshoes.

- Derrota nas copas de 2002 e 2006.
- 2010: Grupo de 50 alunos iniciaram pesquisas:
 - Número de toques;
 - Tempo médio de posse;
 - Velocidades de movimento;
 - Mapas de calor;
 - Distância percorrida.
- 2014: redução do tempo médio de posse de bola:
 - 3,4 segundos em 2010 para 1,1 segundo em 2014



- Plataforma de aluguel de imóveis.
- Chega a 2,5 milhões de hóspedes em uma noite.
- Cruzamento de dados para previsões de épocas mais procuradas.
- Machine Learning para sugestão de preços.
- Sistema de recomendação.



- Comércio eletrônico de produtos esportivos.
- Desafios em promoções, precificação e customização de produtos.
- Perfil do usuário:
 - Quando usuário realizou última compra, gasto médio por usuário, quanto repete a compra, etc.

NETSHOES

- Comércio eletrônico de produtos esportivos.
- Desafios em promoções, precificação e customização de produtos.
- Perfil do usuário:
 - Quando usuário realizou última compra, gasto médio por usuário, quanto repete a compra, etc.
- Aumento de 40% na receita.
- Buscas sem resultados caiu 80%.
- Tempo médio de navegação caiu de 5 para 3 minutos.

NETSHOES

- ☑ Qualquer área pode e deve usar os dados a seu favor.
- ☑ Precisamos amadurecer a cultura de análise de dados.