BIG DATA

**Storytelling** é a capacidade de contar histórias de maneira relevante, onde os recursos audiovisuais são utilizados juntamente com as palavras. É um método que promove o seu negócio sem que haja a necessidade de fazer uma venda direta. Em outras palavras, o storytelling tem um caráter muito mais persuasivo do que invasivo.

**ETL (extração, transformação e carga de dados)**

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

**OLTP (online transaction processing)** processamento de transação em tempo real

* Sistemas que se encarregam de registrar todas as transações em uma determinada operação organizacional
* Geralmente possuem um bom desempenho em manipulação de dados operacionais, mas, são ineficientes para análises gerenciais.

**STAGING AREA**

* Área onde dos dados são colocados após a extração a partir dos sistemas de origem
* Raramente são normalizadas
* Reduza a sobrecarga de acessos aos sistemas fontes
* dedicada para a fase e ETL e não disponível para usuários finais

-Relatórios não podem acessar os dados da staging area.

-Somente processos ETL podem ler e escrever na staging area

**TRASFORM**

* É nessa 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:

- montante\_vendas = qtd \* preço\_unitario

* Resumo de várias linhas de dados:

- Total de vendas para cada loja e para cada região

* Geração de valores de chaves substitutas

- surrogate Keys

**LOAD**

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

**Conclusão**

**ETL** tem grande impacto na análise de dados

Extração e carga são obrigatórias

Transformação é opcional

**Data Warehouse**

Conjunto de dados baseado em assuntos, integrado, não volátil e variável em relação ao tempo, de apoio às decisões gerenciais

**Orientado à assuntos:** Ex: 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 consistente da informação

**Não volátil** os dados de uma data Warehouse não são modificados como em sistemas transacionais (exceto para correções), mas somente carregados de acessados para leituras, com atualizações apenas periódicas.

**Variável 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 uma 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.

**Data Lake**

* São os dados em grandes volumes em seu estado natural, vindos de todos os tipos de fontes, onde os usuários podem ‘mergulhar ‘e tirar amostras. Um Lago cheio de dados.

**Tipos de análises**

1. **Análise Descritiva**

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

1. **Análise Diagnóstica**

* O objetivo é entender 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.

1. **Análise Preditiva**

* tipo de análise mais conhecido.
* 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 de classificação e agrupamento.
* Demandam um volume significativo de dados de boa qualidade e exigem um maior grau de sofisticação.

1. **Análise Prescritiva**

* Também conhecida como “análise de recomendação”.
* Busca trazer informações das consequências de acontecimentos previsto**.**
* Análise preditiva identifica tendências futuras**,** a prescritiva traça as possíveis consequências de cada ação.

**Conclusão**

* Cada tipo de análise em seu próprio escopo de sua própria finalidade.
* É possível conhecer a maturidade em Data Analytics de uma determinada empresa baseada em quais análises fazem parte da sua realidade.

**STREAMING**

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

**Desafios**

* Alto volume de darmos embaixo latente 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.

**STREAMING E IoT**

* Monitoramento das 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.

**Processamento de Eventos Complexos (CEP)**

* 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 é assim apurar eventos relevantes.
* As regras para a descrição dos padrões desejados são definidos em uma linguagem de consulta sobre os fluxos de eventos, de forma similar ao SQL.

**Conclusão**

* 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**

* A visualização de dados é uma forma acessível de ver e entender exceções, tendências de padrões nos dados. É essencial para analisar as informações de tomar decisões;
* Nossos olhos são atraídos por cores é 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.

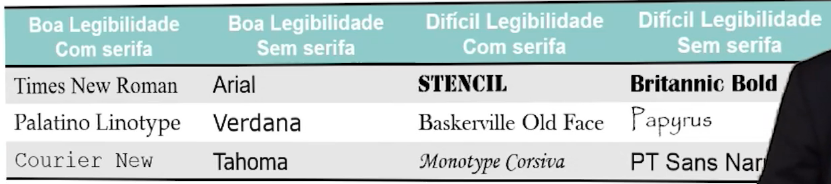
**TIPOS COMUNS**

* Gráficos;
* Tabelas;
* Mapas;
* Infográficos;
* Painéis;

**Processamento pré atentivo**

* Determina quais objetos são oferecidos na nossa atenção o final
* Ocorre antes da atenção consciente.

**FONTES**

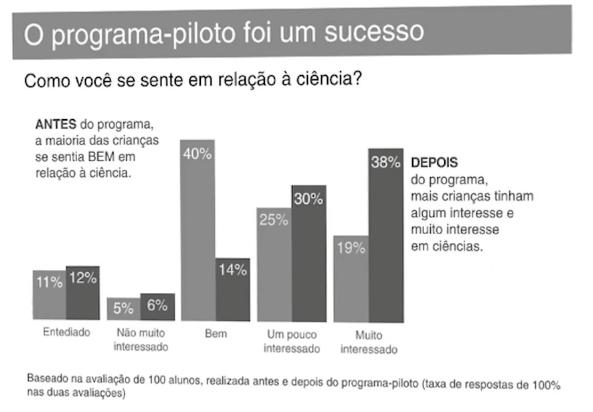


**Design Thinking**

* 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

**DATA STORYLLING**

* 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.

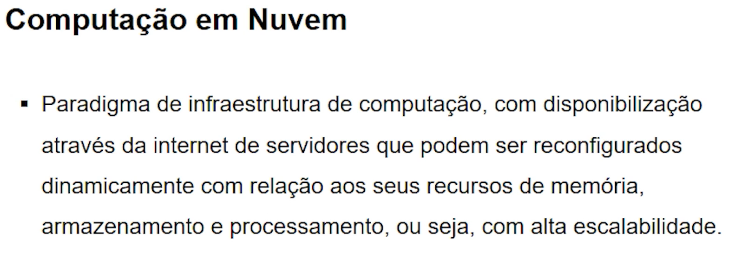
****

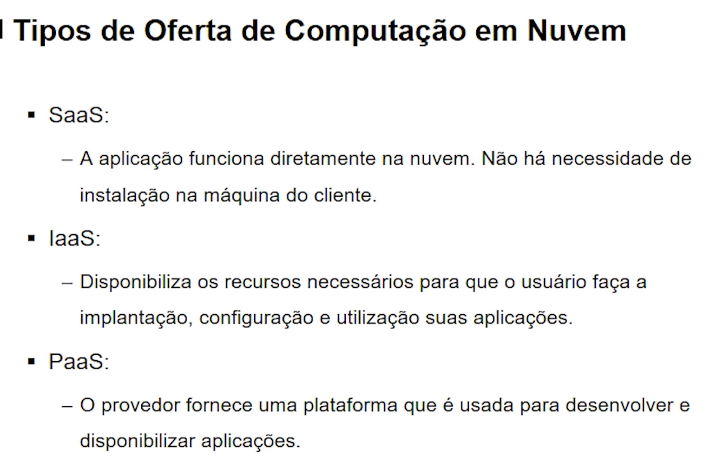
1. Entenda o contexto
2. Escolha uma apresentação visual adequada
3. Elimine saturação
4. Foque a atenção onde você deseja
5. Pense como um designer
6. Conte uma história

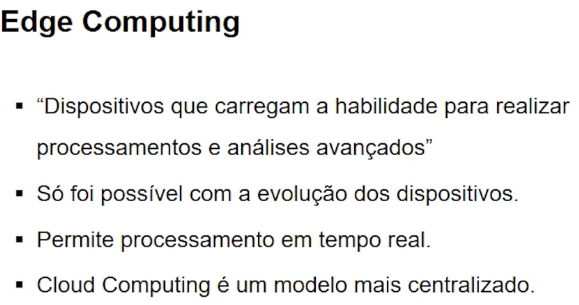
**CONCLUSÃO**

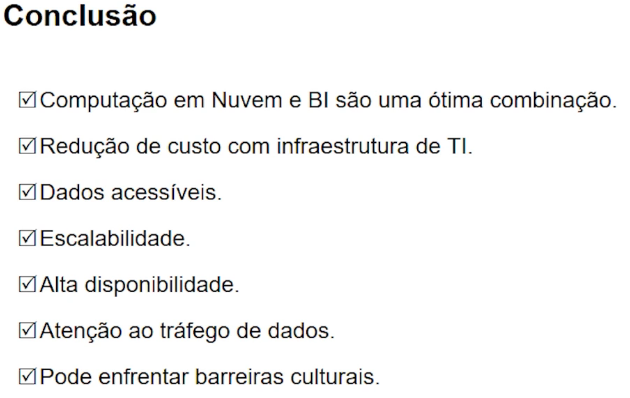
* A visualização de dados é uma maneira simples e rápidas de transmitir conceitos de modo universal
* Data Storytelling é o ato de você explicar o que você fez, com o fez e por que fez, tudo isso de forma que mantenha seu ouvinte engajado

**CLOUD COMPUTING**

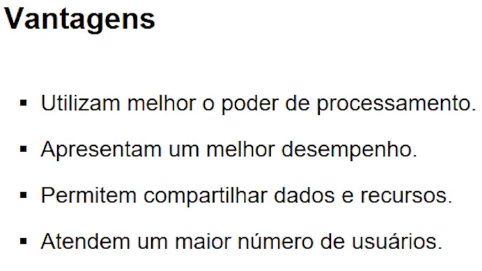
****

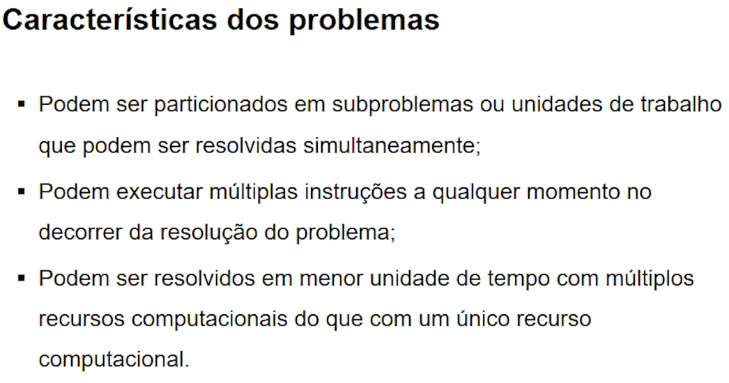
****

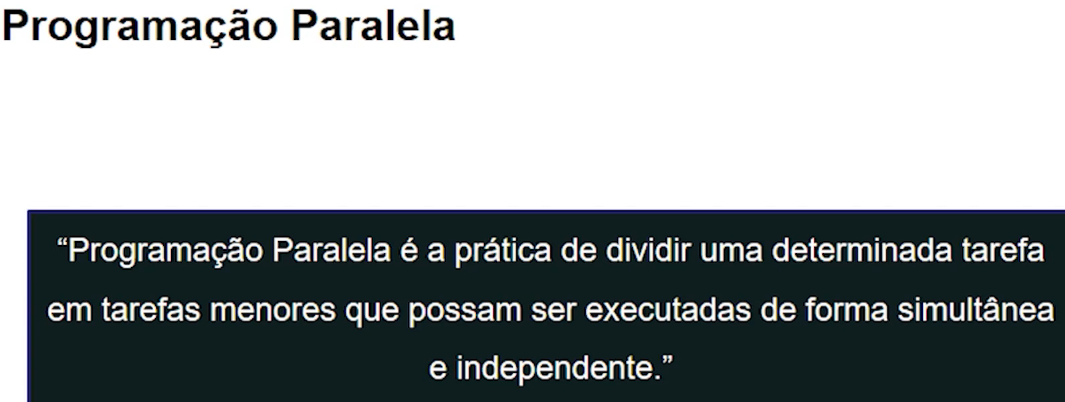
****

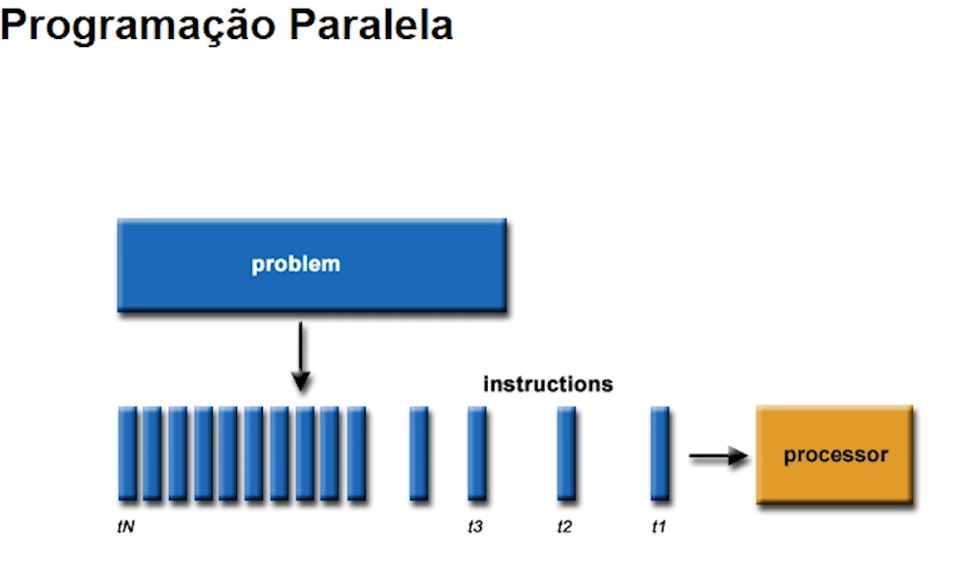
****

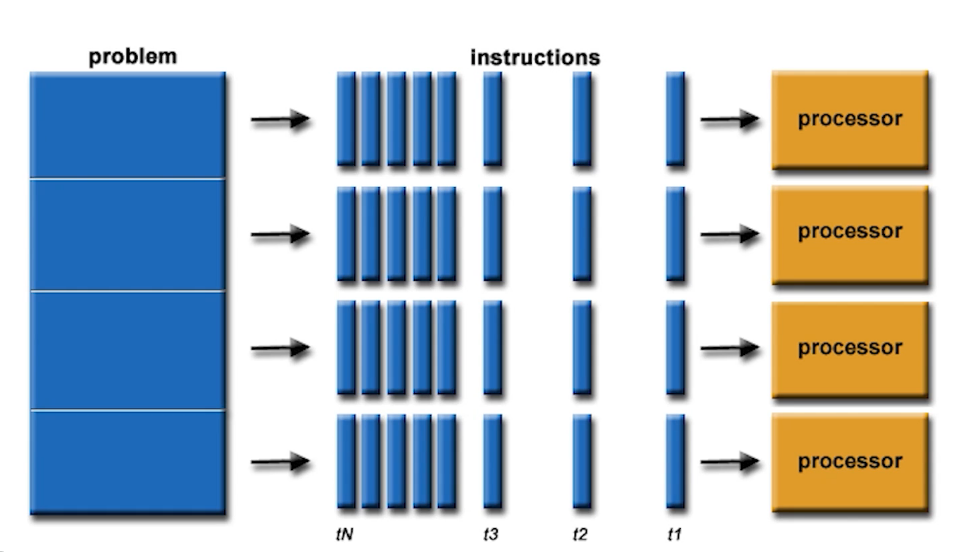
**PROCESSAMENTO PARALELO E DISTRIBUIDO**

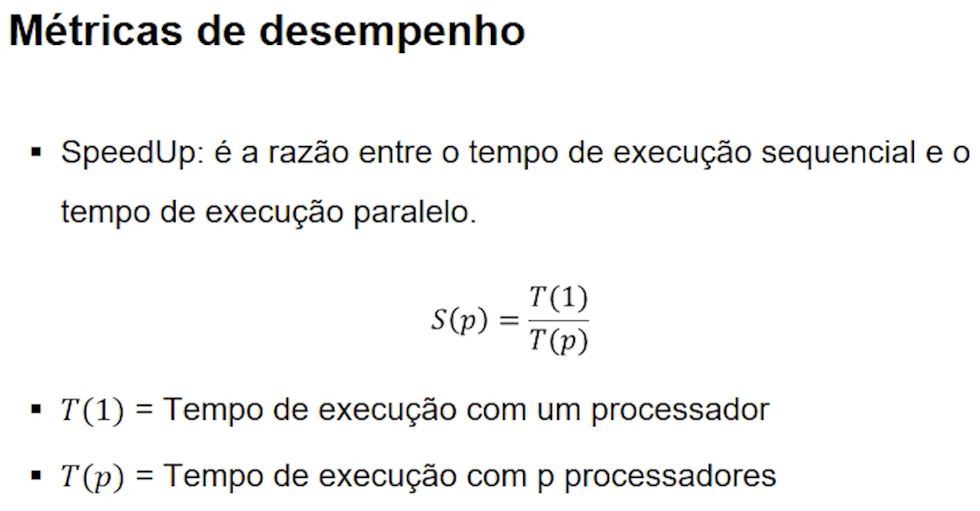


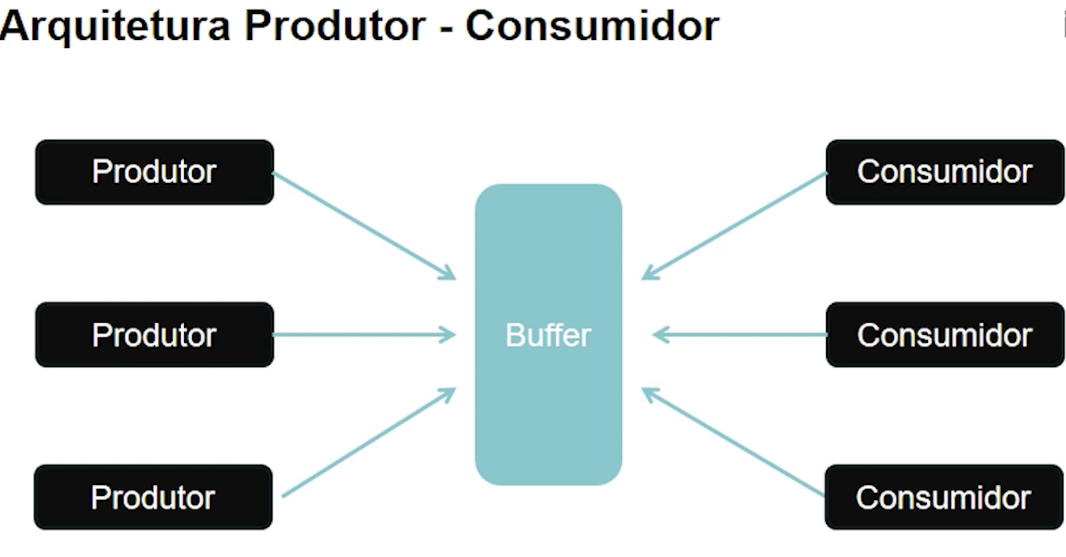


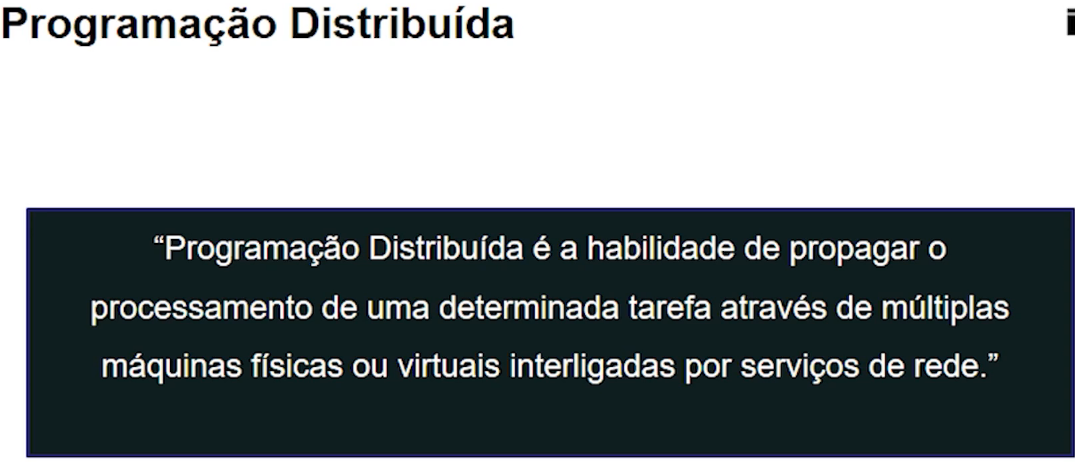


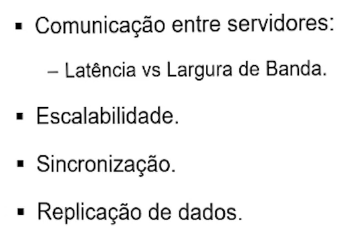


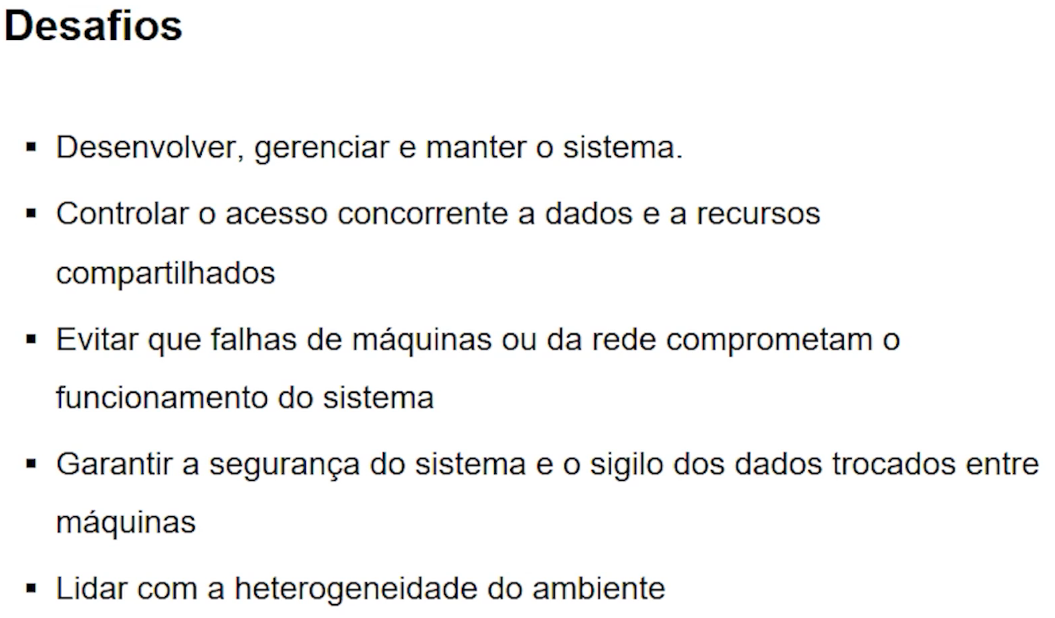


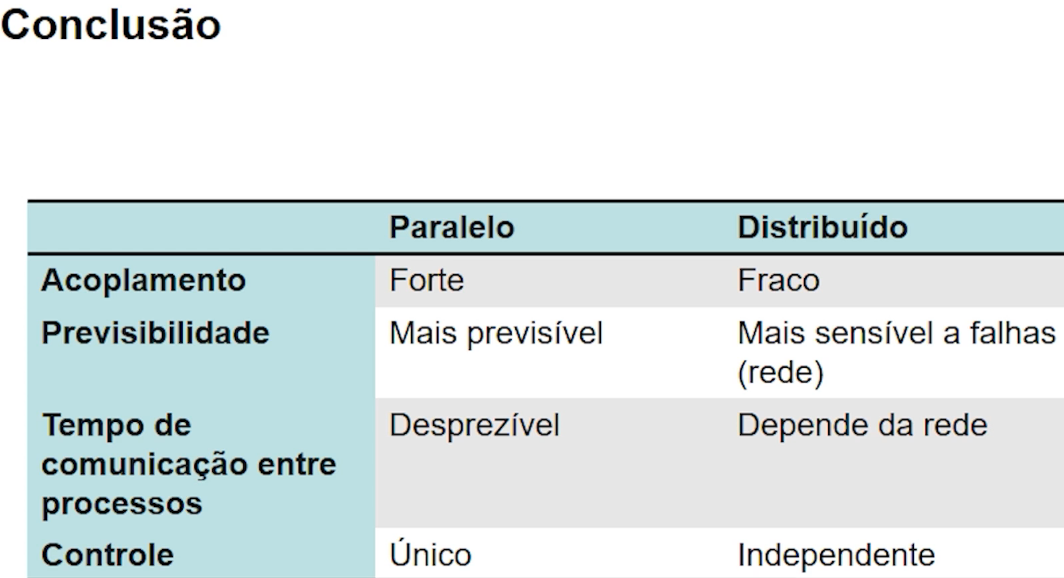




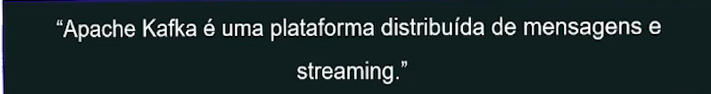


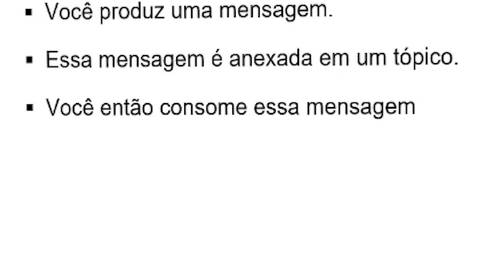


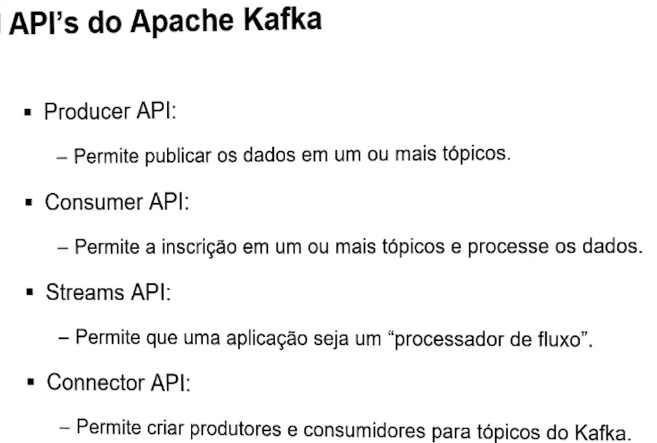


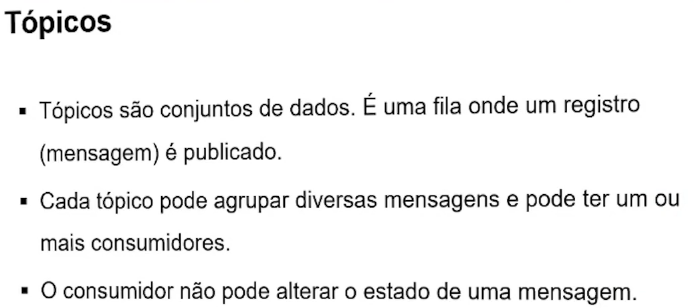


**Apache kafka e Hadoop spark**

****

****

****

****

