escola britânica de artes criativas & tecnologia

Profissão: Analista de dados

escola britânica de artes criativas & tecnologia

BIG DATA I - PROCESSAMENTO

GUIA DA AULA 1

Compreenda o Big Data

- **Big data**
- **3vs: Volume, variedade e velocidade**
- **Computação distribuída e paralela**

Acompanhe aqui os temas que serão tratados na videoaula

Big data

Big data é um termo que geralmente representa um conjunto de dados muito grande, complexo e de difícil processamento. Apesar do apelo comercial, o termo pouco contribui para a definição de um problema com o volume de dados, pois:

- tamanho é relativo, um problema de processamento em um computador pode não ser para outro que contenha mais recursos (RAM, CPU etc.);
- tamanho não é o único desafio moderno no processamento de dados: a variedade do tipo dos dados e a velocidade com que são produzidos somam-se a essa complexidade.

3vs: Volume, variedade e velocidade

A criação da *internet* (1990s) e a sua democratização através da adoção em massa de computadores pessoais (2000s), *smartphones* (2010s) e dispositivos da *internet* das coisas (2010s), trouxe diversos novos desafios para o ecossistema de dados em três frentes: volume, variedade e velocidade.

- **Volume**: os recursos de um sistema computacional não são suficientes para processar um determinado volume de dados em uma determinada janela de tempo / tamanho dos dados representa frações da memória do computador (décimos etc.). Via de regra, arquivos com mais de 100 MB de tamanho são problemáticos para serem processados em computadores tradicionais.

Exemplo: um arquivo de texto (txt) de *log* de acesso de usuários a um *website* facilmente atinge 100 MB de tamanho em poucos dias. Com esse arquivo é possível responder perguntas de negócio como:

- Qual é o período do dia/semana/mês/ano com mais acessos?
- Qual o tempo médio da etapa de *login*?
- Qual a taxa de erro de acesso por dia/semana/mês?

- **Variedade**: as fontes de dados modernas armazenam e disponibilizam dados em diversos formatos. Somaram-se aos tradicionais bancos de dados relacionais (SQL) diferentes formatos de arquivos (csv, json, txt, html, pdf, jpeg, png etc.), bases de dados não relacionais (NoSQL ou dados semi/não estruturados), APIs (json) etc.

Exemplo: os *sites* dos tribunais de justiça dos estados publicam diariamente o andamento dos processos judiciais que tramitam na segunda instância em arquivos do tipo `pdf`. Como fazer para extrair e armazenar estes arquivos diariamente? Como extrair o número e o *status* do processo do documento?

- **Velocidade**: processamento de dados em lote (*batch*) já não atende mais às necessidades do negócio. Dispositivos permanecem conectados às redes de computadores (*internet, internet* móvel etc.) o tempo todo, logo, continuamente produzindo dados.

Exemplo: um *e-commerce* registra os *clicks* de um usuário enquanto este navega pelo seu *website*. Com este dados e com o histórico do usuário, seria possível disponibilizar um cupom de desconto para que o usuário não deixe o *website* sem finalizar uma compra? Qual o melhor momento para enviar o cupom?

Computação distribuída e paralela

A estratégia para lidar com o aumento da demanda por recursos para processamento de dados sempre foi a de melhoria do *hardware* de um mesmo computador: mais memória, mais velocidade de processamento etc. Contudo, após os anos 2000s, a demanda cresceu em um ritmo muito mais acelerado se comparado a capacidade de melhoria de *hardware*. E dessa necessidade nasceu uma nova arquitetura de computadores e um novo paradigma de computação: *clusters* de computadores (múltiplos computadores) e computação distribuída e paralela, respectivamente.

- **Arquitetura**

Um *cluster* é um conjunto de computadores (mesmas configurações, idealmente mesmo *hardware* etc.) conectados em uma rede privada. Um gerenciador de *cluster* (*cluster manager*) é uma aplicação que orquestra as atividades de armazenamento e processamento de dados distribuído e paralelo, abstraindo a complexidade para usuários e aplicações. Os gerenciadores de *cluster* mais utilizados são o Apache Hadoop e o Kubernetes.

Nota: Computadores de um *cluster* são conhecidos como nós.

- **Armazenamento**

Dados são armazenados em arquivos (`csv` , `txt` etc.) e são "quebrados" em blocos (128 MB geralmente), distribuídos e replicados (três vezes geralmente) nos nós. O gerenciador de *cluster* mantém um mapa da distribuição dos blocos.

- **Processamento**

Existem algumas maneiras de processar dados distribuídos. Uma das maneiras mais eficientes é enviar a operação de processamento (agregações como soma, por exemplo) para o nó em que o dado está armazenado, realizar o processamento localizado e coletar apenas os resultados.

Nota: Operações de junção (*joins*) costumam ser caras (em termos de tempo de processamento e consumo de memória), pois blocos inteiros de dados devem trafegar pela rede de um nó para o outro.