Tecnologia da Informação

RESUMO

Big Data

CONCEITOS DE BIG DATA

Big Data é um termo amplamente utilizado na atualidade para nomear **conjuntos de dados muito grandes ou complexos**, que os **aplicativ de processamento de dados tradicionais ainda não conseguem lidar**.

O **Big Data** pode ser definido genericamente como a **captura, gerenciamento e a análise de dados que vão além dos dados tipica ente estruturados**, que podem ser consultados e pesquisados através de b cos de dados relacionais.

O **objetivo do Big Data** é **propiciar dados e informações que possa er analisados visando subsidiar tomadas de decisão**.

O **Big Data** tanto pode ser encarado como o **grande volume de ados estruturados e não estruturados que são gerados a cada segu lo** ou também como as **tecnologias que são utilizadas para lidar con ste grande volume de dados**. Esta segunda acepção é, por vezes, chama de **Big Data Analytics.**

TIPOS DE ANÁLISE

- **Análise preditiva**: Análise de possibilidades futuras com base em dados passados.
- **Análise prescritiva**: Traça as possíveis consequências de cada ação.
- **Análise descritiva**: Compreensão em tempo real dos acontecimentos.
- **Análise diagnóstica**: Compreensão das possibilidades fornecidas por uma base de dados.

PREMISSAS DE BIG DATA

Premissas

- **Volume**
  - Análise de grandes volumes de dados
  - Guarda os dados em diferentes localidades e juntando-os através de software
- **Velocidade**
  - Respostas com velocidade e em tempo hábil
  - Analisar dados no instante em que são criados, sem ter de armazená-los em bancos de dados
- **Variedade**
  - Diferentes formatos de informação
  - Fontes estruturadas, semi-estruturadas e a grande maioria em fontes não estruturadas.
- **Veracidade**
  - Informações Verdadeiras
  - Proveniência ou à confiabilidade da fonte de dados.
- **Valor**
  - Informações devem agregar valor ao negócio
  - Custos não devem ser superiores aos benefícios.
- **Variabilidade**
  - (1) número de inconsistências nos dados
  - (2) multiplicidade de dimensões de dados de diferentes fontes de dados (complexidade)
  - (3) velocidade inconstante na carga dos dados
- **Validade**
  - Precisão e correção dos dados para o uso pretendido
- **Vulnerabilidade**
  - Novas preocupações de segurança
- **Volatilidade**
  - Estabelecer regras para a atualidade e a disponibilidade de dados
- **Visualização**
  - Maneiras diferentes de representar dados
- **Viscosidade**
  - Resistência ao fluxo no volume de dados
- **Viralidade**
  - Rapidez com que as informações são dispersas
- **Governança**
  - Governança para as novas fontes de dados e a maneira como os dados serão usados.
- **Pessoas**
  - Pessoas com aptidões relevantes disponíveis e compromisso de patrocinadores.

PADRÕES DE BIG DATA

Padrões de consumo:

Visualização	Descoberta ad-hoc	Aumentar os armazenamentos de dados tradicionais	Notificação	Iniciar resposta automatizada
Abordagem tradiocional nem sempre fornece a melhor visualização.	Criação de relatórios padrão adequados para todas as necessidades de negócios não é viável.	**Ampliar o escopo de dados** disponível para a analítica atual para incluir dados que residem dentro e fora dos limites organizacionais.	Notificações para **indicar eventos**.	**Acionar ou iniciar outros processos** de negócios ou transações.
O objetivo é fazer com que seja **mais fácil consumir os dados de forma intuitiva**.	Capacidade de **enviar consultas ad hoc (feitas sob demanda)** ao procurar por informações específicas.		Lidar com o volume antecipado de notificações a serem enviadas de maneira oportuna.	

Padrões de acesso:

Padrão de acesso à web e mídia social

- A web e a mídia socialsão **úteis em praticamente todas as análises**, mas são necessários mecanismos de acesso diferentes para obter esses dados.

Padrão de dados gerados por dispositivos

- Inclui **dados de sensores**.
- Os dados são **detectados a partir das origens** de dados.

Padrão de dados de warehouse, operacionais e transacionais

- É possível **armazenar os dados de warehouse, operacionais e transacionais** existentes para evitar a limpeza ou o arquivamento deles ou para reduzir a carga no armazenamento tradicional quando os dados são acessados por outros consumidores.

Padrões de processamento:

Análise de dados históricos	•**Analisar as tendências históricas para um determinado período, conjunto de períodos e produtos e compará-las** aos dados atuais disponíveis.
Analítica avançada	•**Correlacionar diferentes conjuntos** de dados em muitos contextos. •Inclui previsões, decisões, processos inferenciais, simulações, informações contextuais e resoluções da entidade.
Pré-processamento de dados brutos	•Para executar a análise em quaisquer dados, eles devem estar em **algum tipo de formato estruturado**.
Análise ad-hoc	•Pelo fato de as fontes e formatos dos dados não serem fixos e exigirem mecanismos diferentes para recuperá-los e processá-los.

Padrões de armazenamento:

Dados não estruturados e distribuídos	Dados estruturados e distribuídos	Armazenamento de dados tradicionais	Armazenamento em nuvem
•Dados não estruturados devem ser armazenados como estão.	•Dados convertidos devem ser armazenados para evitar a frequente conversão de dados brutos para dados estruturados.	•Usar o data warehouse, o sistema relacional e outros armazenamentos de conteúdo existentes.	•Uso de soluções de nuvem que fornecem o gerenciamento de sistemas, manutenção e armazenamento de big data

HADOOP

O **Hadoop** é um framework de código aberto, implementado em **utilizado para o processamento e armazenamento em larga escala alta demanda de dados**, **utilizando máquinas comuns**. Os serv Hadoop fornecem armazenamento, processamento, acesso, governança, segurança e operações de dados.

O Hadoop é formado por vários subprojetos.

Subprojetos para gerenciamento

O **Chukwa** é o **sistema especialista em coleta e análise de logs em sistemas de larga escala**.

O **ZooKeeper** é o arcabouço criado pelo Yahoo! em 2007 com o objetivo de fornecer um **serviço de coordenação para aplicações distribuídas de alto desempenho.**

Subprojetos para acesso aos dados

O **Pig** é uma **linguagem de alto nível orientada a fluxo de dados e u arcabouço de execução para computação paralela**.

O **Hive** fornece uma infraestrutura que permita utilizar Hive QL, uma linguagem de consulta similar a SQL bem como demais conceitos de dados relacionais tais como tabelas, colunas e linhas, para **facilitar as análises complexas feitas nos dados não relacionais de uma aplicação Hadoop**. Existe também uma definição que trata o **Hive** como um **datawarehouse distribuído** que facilita o uso de grandes conjuntos de dados. Nesse caso, ele seria enquadrado como um subprojeto para armazenamento dos dados.

O **Avro** é o **sistema de seriação de dados baseado em schemas**.

Subprojetos para processamento de dados

Modelo de programação e um **arcabou especializado no processamento de conjuntos de dados distribuídos** em u aglomerado computacional (cluster).

Programador **não precisa se preocupar com o paralelismo**.

MapReduce

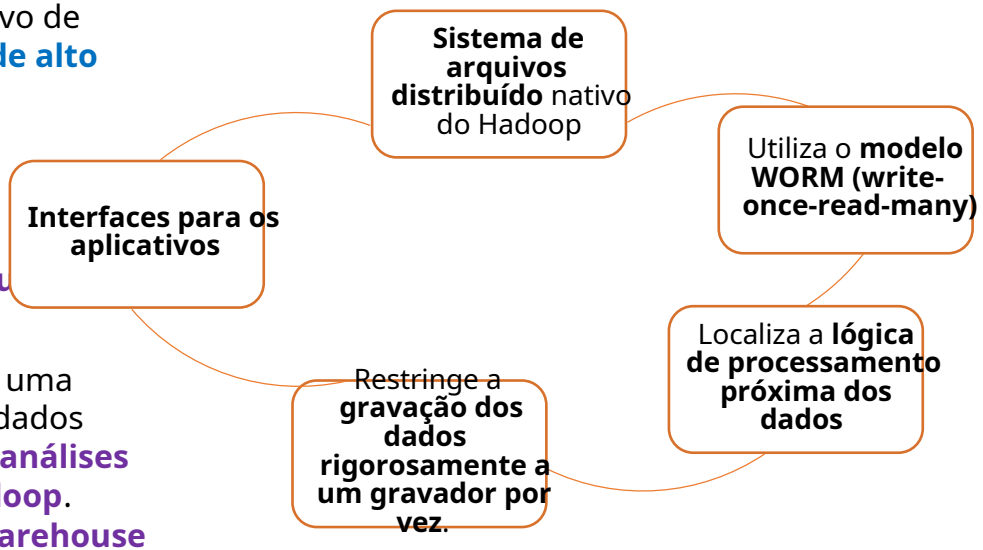
Map: fase de mapeamento, **procesame primário dos dados**.

Reduce: **geração do resultado** final.

A leitura é realizada de **arquivos com pares chave/valor** (geralmente .csv)

Subprojetos para armazenamento de dados

O **HDFS:**



O Hbase é um **banco de dados distribuído e escalável que dá suporte a armazenamento estruturado e otimizado para grandes tabelas**.

Sqoop (SQL to Hadoop)

O **Sqoop (SQL to Hadoop)** é um aplicativo de interface de linha de comando para **transferência de dados entre bancos de dados relacionais e Hadoop**.

Flume

O **Flume** é um software distribuído, confiável e disponível para **coletar, agregar e mover com eficiência grandes quantidades de dados de log**.

Componentes do Hadoop

NameNode

- **Gerenciar os arquivos armazenados** no HDFS.

SecondaryNameNode

- **Auxiliar o NameNode** a manter seu serviço.
- Ser uma **alternativa de recuperação** no caso de uma falha do NameNode.

DataNode

- Efetivamente realizam o **armazenamento dos dados**.

JobTracker

- Função de **gerenciamento sobre o plano de execução das tarefas** a serem processadas pelo MapReduce.

TaskTracker

- Responsável pela **execução de tarefas MapReduce**.