

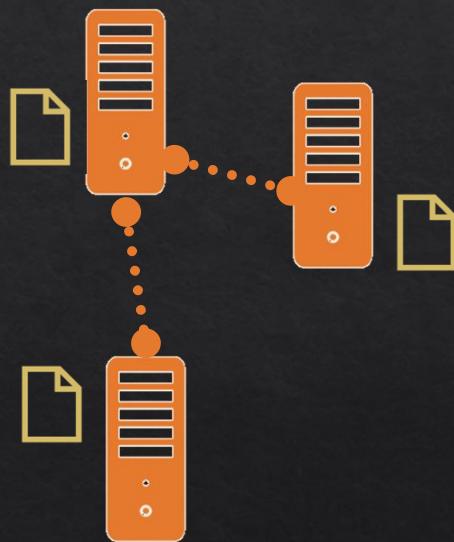
Cluster



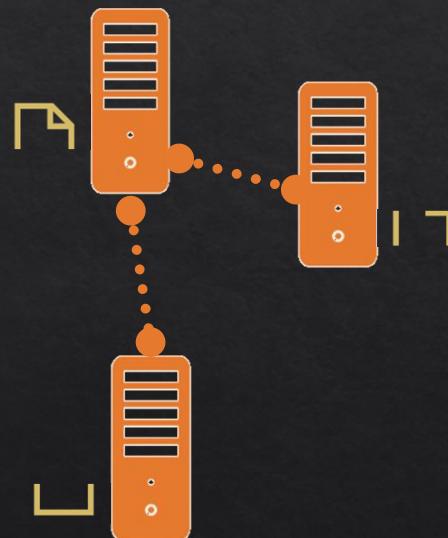
- Computação Distribuída
- Cada computador é chamado nó
- Computadores operando em conjunto, conectados, com o mesmo objetivo (processar dados) - Dividir para Conquistar
- Podem estar dispersos fisicamente, inclusive em diferentes continentes
- Pode estar no mesmo rack
- Podem ser virtualizados
- Estrutura Master/Slave
 - Master: coordena, distribui, agenda
 - Slave: armazena, processa

Replicação VS Partição

◊ Replicação: dados são copiados

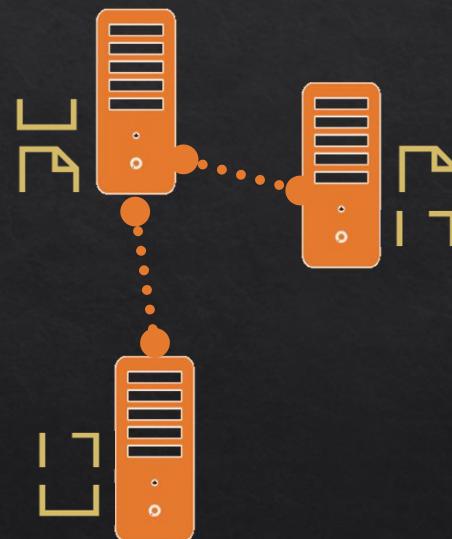


◊ Partição: dados dão divididos



Replicação e Partição

- ❖ Dados são divididos e copiados



Tipos de Partição

- ❖ Centralizada:
 - ❖ Serviço central gerencia os dados nas partições. Possui todos os metadados.
 - ❖ Pode ser um gargalo pro sistema
- ❖ Range:
 - ❖ Divide em intervalos
 - ❖ Pode ficar desbalanceado
- ❖ Hash:
 - ❖ Usa uma Hash Table para distribuir os dados de forma balanceada entre os nós

Mutation

- ❖ Suporte ao sistema em alterar dados
- ❖ Em sistemas distribuídos isso pode não ser tão simples
- ❖ Alguns formatos de arquivo, como colunas, o processo também pode ser complexo

Data Warehouse Versus Data Lake

Data Warehouse (Clássico)

- Implementado a partir dos anos 90
- Estruturado
- Dimensional

Data Lake

- Big Data
- Todo tipo de informação
- Hadoop/HDFS

Batch VS Streaming VS Interativo



Batch: dados são coletados (extraídos, tratados, armazenados) e processados. Pode levar horas, semanas, meses

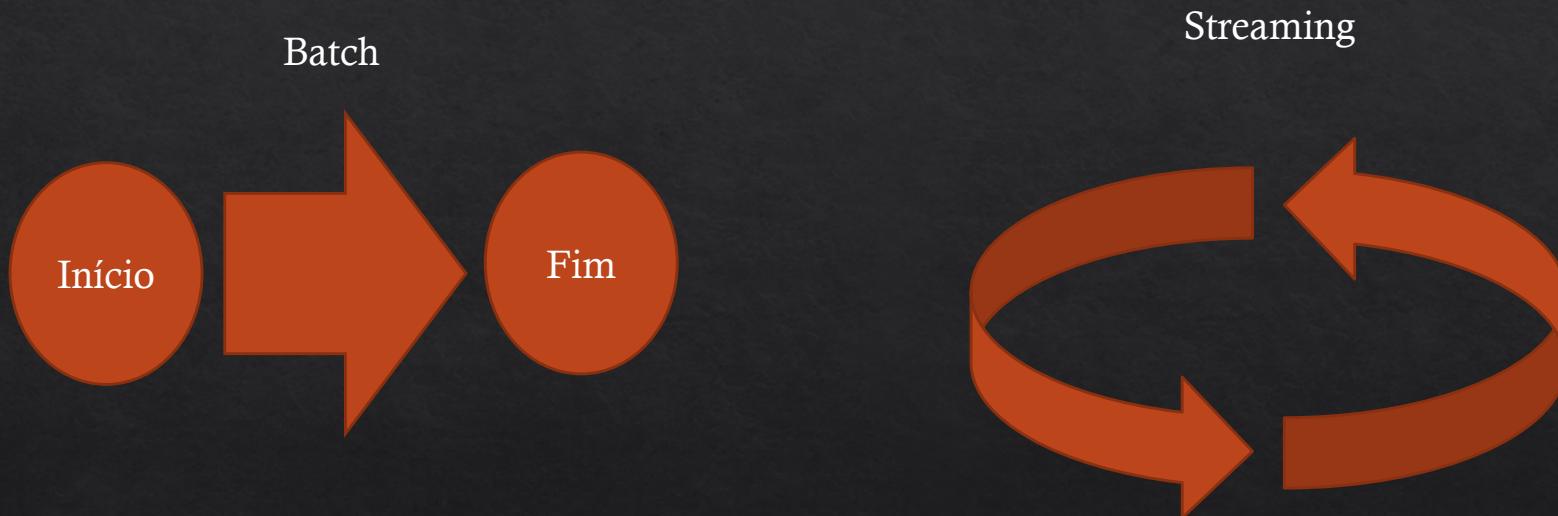


Streaming: dados processados em tempo real, a medida que são produzidos. Saída em tempo real ou próximo ao tempo real.
Processamento constante (Fluxo)



Interativo (Ad Hoc): Interface onde processamento é requerido e o resultado é analisado

Batch VS Streaming



Batch VS Streaming VS Interativo

Batch:

Sistema de busca de produtos de varejo on-line: Milhões de termos pesquisados por dia
Diariamente, os dados são processados para criar um ranking das palavras mais buscadas

Streaming:

Sistema de busca de produtos de varejo on-line: Avaliação de transação de pagamento é fraudulenta
As transações são processadas a medida que são executadas

Interativo (Ad Hoc):

Rodar uma consulta para ver as vendas do dia de determinado produto

Tipos de Streaming

- ❖ Native: dados são processados assim que chegam, sem esperar por demais dados
- ❖ Micro-baching: dados são agrupados e processados em grupo (delay)

Nativo



Micro-baching



Streaming



Event: dado



Producer: gerador



Subscriber: consome

Latência

- ❖ Intervalo de tempo entre a produção da informação e seu processamento
 - ❖ Em batch: alta latência
 - ❖ Streaming: baixa latência

Real Time e Near Real Time

- ❖ Real Time: Em tempo real
- ❖ Near Real Time: Próximo ao tempo real
- ❖ Normalmente quando fala-se Real Time, refere-se a Near Real Time

Processamento Distribuído

- ❖ Um processamento em batch / streaming / memória pode ou não ser distribuído



Commodity Hardware

- ❖ Commodity Hardware:
 - ❖ Barato
 - ❖ Compatível
 - ❖ Amplamente Disponível
 - ❖ Intercambiável

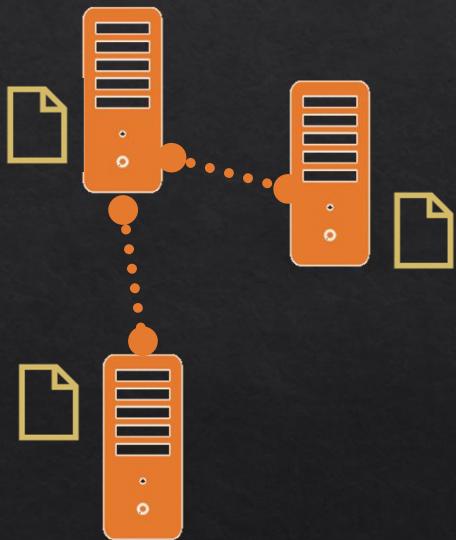


Big Data

- ❖ Uma das características de sistemas resilientes - tolerantes a falhas, é ser capaz de continuar operando mesmo sob alguma falha

Sistemas Distribuídos

- Replicação: dados são copiados



- ◊ A replicação dos dados é um destes aspectos
- ◊ Falhas em:
 - ◊ Rede
 - ◊ Software
 - ◊ Etc.
 - ◊ **Disco Rígido**

Por que?

Se o cluster é composto por um volume significativo de discos, certamente haverão falhas!

Falhas em Disco Rígidos

- ❖ Média de Falhas em Discos Rígidos: 4.81%
- ❖ Se você tiver 1000 HDs, em 1 ano:
 - ❖ 48 discos falharão!
 - ❖ 4 por mês
 - ❖ 1 por semana!

Fonte:

<https://www.backblaze.com/blog/hard-drive-reliability-q3-2015/>

