

BIG DATA

...PODEMOS DIZER QUE SÃO DADOS GRANDES?

Deyvirson Mendonça

SUMÁRIO

- ▶ Definição
- ▶ Os Vs do Big Data
- ▶ Desafios
- ▶ Hadoop
- ▶ BD e Big Data

PARA COMEÇAR... O QUE SÃO DADOS?

- ▶ Informações diversas, que podem ser organizadas de maneira diferentes.
- ▶ Data Science
- ▶ Trata os dados utilizando estatística, matemática e mais uma porção de disciplinas.

ONDE ESTÃO E DE ONDE VEM ESSES DADOS?

- ▶ Quatro devices por pessoa (~) conectado à internet **E enviando dados.**
- ▶ Em apenas um minuto, 4 milhões de vídeos são vistos no Youtube.
- ▶ 29 milhões de mensagens são trocadas no WhatsApp.
- ▶ Mais de 80 horas de Netflix são assistidas.
 - ▶ É muito Dado!!!

ONDE ESTÃO E DE ONDE VEM ESSES DADOS?



O QUE É BIG DATA?

- ▶ Você não vai achar aqui a resposta!
- ▶ Não existe uma resposta exata.
 - ▶ MAAAAS....
- ▶ Problema com grande volume de dados

BIG DATA SÃO VÁRIOS V

- ▶ Costuma-se dizer que são 3 Vs
 - ▶ Volume, Velocidade, Variedade
- ▶ 5 Vs
 - ▶ Veracidade, Valor, Viabilidade
- ▶ 7 Vs
- ▶ 9 Vs

PILARES DO BIG DATA

- ▶ Quando você tem um problema que envolve:
 - ▶ Volume
 - ▶ Velocidade
 - ▶ Variedade
 - ▶ Veracidade
 - ▶ Valor de Dados
- ▶ Você precisa trabalhar com Big Data!!!

OS 5 VS

- ▶ **Volume** - Organizações coletam dados de uma grande variedade de fontes.
- ▶ **Velocidade** - Os dados fluem em uma velocidade sem precedentes e devem ser tratados em tempo hábil.
- ▶ **Variedade** - Os dados são gerados em todos os tipos de formatos.

OS 5 VS

- ▶ **Veracidade** - Além da velocidade e variedade de dados cada vez maiores, os fluxos de dados podem ser altamente inconsistentes com picos periódicos.
- ▶ **Valor de Dados** - Grandes dados para valor socioeconômico.

OU SEJA...

- ▶ Muito volume de Dados?
 - ▶ Big Data!
- ▶ Dados que precisam ser transmitidos em velocidade muito rápida?
 - ▶ Big Data!
- ▶ Dados com muita variedade?
 - ▶ Big Data!

ACHANDO UM PROBLEMA DE BIG DATA

- ▶ Problema computacional onde:
 - ▶ Difícil lidar com dados
 - ▶ Tamanhos dos dados são proibitivos para se tratar utilizando abordagens tradicionais
 - ▶ Sim, isso é um problema de Big Data

ONDE ESTÃO E DE ONDE VEM ESSES DADOS?

- ▶ Da internet, onde basicamente são:
 - ▶ Redes Sociais
 - ▶ Sensores de IoT
 - ▶ Netflix
 - ▶ YouTube

FAZENDAS INTELIGENTES E IOT

- ▶ Sensores no solo que coletam informações de:
 - ▶ Umidade, Acidez, pH.
- ▶ Tempo de Crescimento de determinada planta.
- ▶ Tratores Inteligentes.
 - ▶ TUDO ISSO É BIG DATA!

E O QUE FAZER COM TODAS ESSAS INFORMAÇÕES?

- ▶ Se trabalhar com dados não estruturados, isso é Big Data!
- ▶ Mas o que danado são dados não estruturados?
 - ▶ Sabe os reactions do facebook? Curtir, adorar, rir
 - ▶ Tipos de músicas que o usuário escuta.
- ▶ São dados que não seguem um formato.

O QUE FAZER PARA LIDAR COM TODOS ESSES DADOS?

- ▶ Cloud passa a ser **obrigatório**
- ▶ Trabalhar com plataforma de processamento massivo paralelo (PMP)
- ▶ Aqui que aparece o Excel!

HADOOP

O QUE É O HADOOP?

- ▶ Plataforma de computação distribuída voltada para clusters e processamento de grande massa de dados.
- ▶ Armazena e Processa
- ▶ Componentes (HDFS e Map-Reduce)

hadoop overview

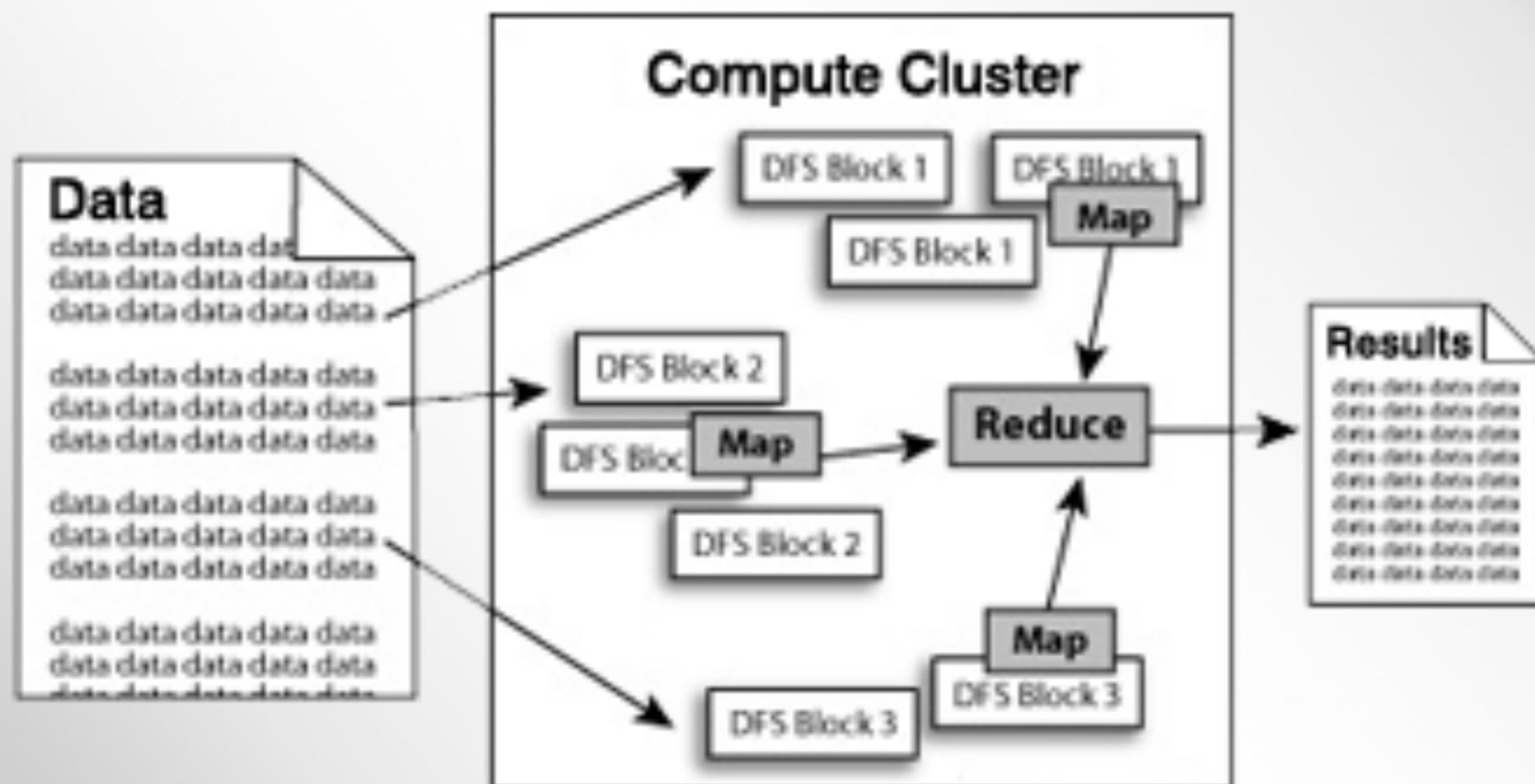


image courtesy of the
Apache Software Foundation

HDFS (HADOOP DISTRIBUTED FILE SYSTEM)

- ▶ Sistema de Arquivo Escalonável
- ▶ Baseado no GFS (Google File System)
- ▶ Sistema de arquivo distribuído são necessários , uma vez que os dados se tornam grande demais para serem armazenados em apenas uma máquina.

MAP-REDUCE

- ▶ É um termo em que cada tarefa é especificada em termos de funções de mapeamento e redução.
- ▶ As tarefas rodam paralelamente nos clusters.

BATCH E STREAMING

STREAMING

- ▶ Utilizado para resposta imediata.
- ▶ Waze
- ▶ Processamento mais caro
- ▶ Ferramentas utilizadas (Cloudera, Spark Streaming, Kineses)

BATCH

- ▶ Mais barato
- ▶ Utilizado em processamento em lotes
- ▶ Imposto de Renda

BANCO DE DADOS

BD NO BIG DATA

- ▶ Quando falamos de Big Data, as ferramentas de BD precisam ser adequadas.
- ▶ Banco de dados distribuídos > BD relacional
- ▶ Escalar Geograficamente
- ▶ Vários grupos de dados são tratados em diversas máquinas em diversos lugares. (NoSQL)
- ▶ Cassandra, MongoDB

PONTO FORTE DO NOSQL

- ▶ Extremamente rápido
- ▶ Distribuidos
- ▶ Tratar quantidade absurda de dados
- ▶ Netflix Utiliza Cassandra.
- ▶ Graças a isso é possível Escalar!

O QUE AS EMPRESAS GANHAM COM BIG DATA?

- ▶ Eu!
- ▶ Você!
- ▶ Todo o mundo!
- ▶ Ofertas personalizadas
- ▶ Netflix recomenda filmes.

REFERÊNCIAS

- ▶ <https://www.concrete.com.br/2017/06/02/uma-pequena-introducao-a-big-data/>
- ▶ <http://www.devmedia.com.br/big-data-e-hadoop-descubra-o-que-e/30163>
- ▶ <https://imasters.com.br/tecnologia/redes-e-servidores/big-data-e-hadoop-o-que-e-tudo-isso/?trace=1519021197>
- ▶ <https://arxiv.org/find/all/1/all:+AND+big+data/0/1/0/all/0/1>
- ▶ <https://www.infoq.com/bigdata/>
- ▶ <http://hadoop.apache.org/>

DÚVIDAS?