Apontamentos do Curso Data Science

BIG DATA FUNDAMENTOS

Eduardo Alves Fonseca Queirós
GitHub: code36u4r60

Entidade Formadora: Data Science Academy
e-mail: suporte@datascienceacademy.com.br
Website: datascienceacademy.com.br/blog

Índice

1 – Big Data Fundamentos

1.1 O que é Big Data

Cerca de 90% de todos os dados gerados no planeta, foram gerados nos últimos 2 anos.

Aproximadamente 80% dos dados são não-estruturados ou estão em diferentes formatos, o que dificulta a análise.

Modelos de análise de dados estruturados, possuem limitações quando precisam tratar grandes volumes de dados.

Muitas empresas não sabem que dados precisam se analisados.

Muitas empresas nem mesmo sabem que os dados estão disponíveis.

Dados preciosos são descartados por falta de conhecimento ou ferramentas de tratamento.

É caro manter e organizar grandes volumes de dados não-estruturados.

Estamos em um período de transformação no modo em que dirigimos nossos negócios e, principalmente, as nossas vidas.

Neste exato momento, uma verdadeira enxurrada de dados, ou 2.5 quintilhões de bytes por dia, é gerada para nortear indivíduos, empresas e governos – e está dobrando a cada dois anos.

Toda vez que fazemos uma compra, uma ligação ou interagimos nas redes sociais, estamos produzindo esses dados.

E com a recente conectividade em objetos, tal como relógios, carros e até geladeiras, as informações capturadas se tornam massivas e podem ser cruzadas para criar roadmaps cada vez mais elaborados, apontando e, até prevendo, o comportamento de empresas e clientes.

Entre 2005 e 2020, o universo digital irá crescer de 130 3exabytes para 40.00 exabytes ou 40 trilhões de gigabytes.

Em 2020, haverá 5.200 gigabytes para cada homem, mulher e criança no planeta.

Dados são a matéria prima dos negócios.

A revolução não está nas máquinas que calculam os dados e sim nos dados em si e na maneira que são utilizados.

O Big Data nos dá uma visão clara do que é granular.

No mundo do Big Data, por sua vez, não temos de nos fixar na casualidade. Podemos descobrir padrões e correlações nos dados que nos propiciem novas e valiosas ideias.

1.2 O que é Big Data

1.2.1 Mas afinal, o que é Big Data?

> *Big Data é uma coleção de conjuntos de dados, grandes e complexos, que não podem ser processados por bancos de dados ou aplicações de processamentos tradicionais.*

Capacidade de uma sociedade de obter informações de maneiras novas a fim de gerar ideias úteis e bens e serviços de valos significativo.

O Google estima que a humanidade criou nos últimos 5 anos, o equivalente a 300 Exabytes de dados ou seja: 300.000.000.000.000.000.000 bytes de dados

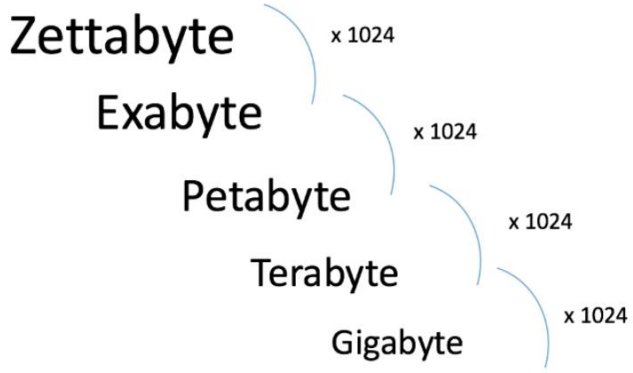
Muitos dos dados gerados, possuem um tempo de vida curto e se não analisados, perdem a utilidade.

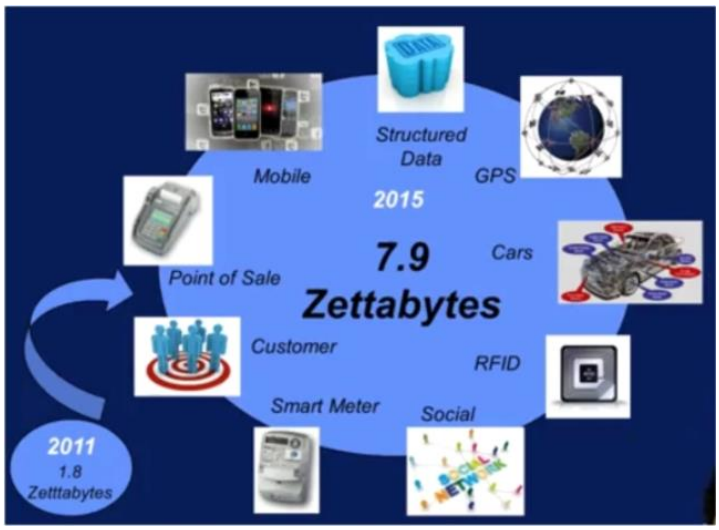
Dados são transformados em informações que precisam ser colocadas em contexto para que possam fazer sentido.

É caro integrar grandes volumes de dados não estruturados.

Dados potencialmente valiosos em sistemas ERP, CRM ou SCM são descartados ou perdidos apenas porque ninguém presta atenção a eles.

1.2.2 Qual é o tamanho do Big Data?





1.3 A importância do Big Data

1.3.1 E por que Big Data tem se tornado tão importante?

Porque surgiram tecnologias que permitem processar esta grande quantidade de dados de forma eficiente e com baixo custo.

Os dados podem ser analisados em seu formato nativo, seja ele estruturado, não estruturado ou streaming (fluxo constante de dados).

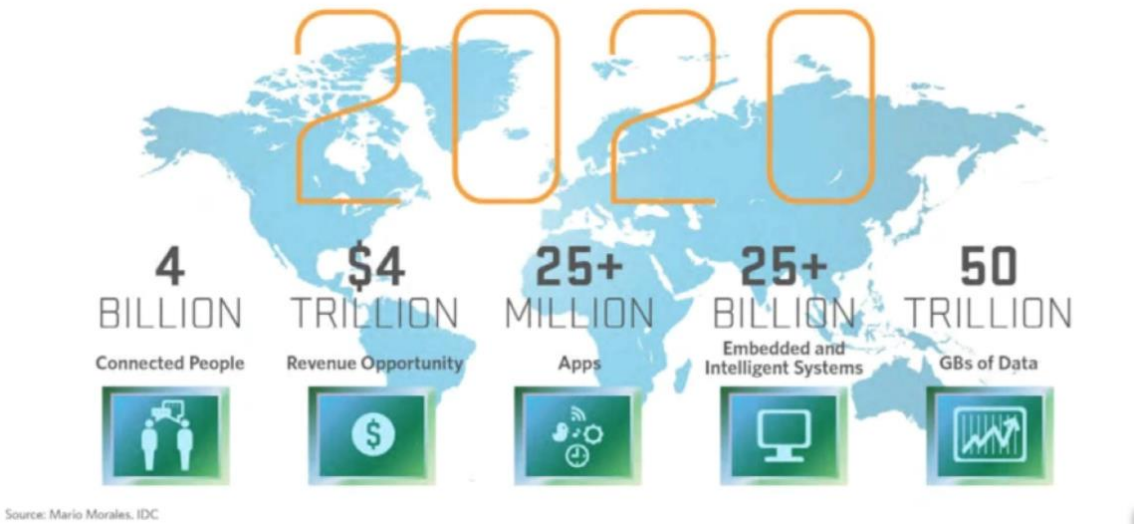
Dados podem ser capturados em tempo real.

Dados podem ser transformados em insights de negócios.

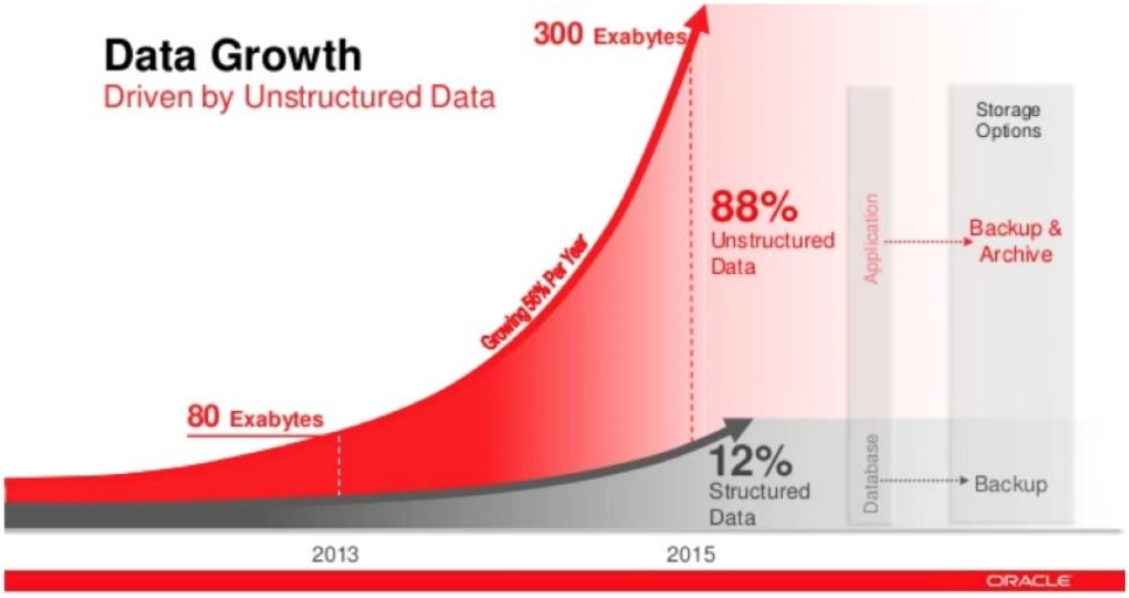
1.3.2 Desafios trazido pelo Big Data

- Encontrar profissionais habilitados em Big Data e Hadoop
- Compreender a plataforma e ferramentas para Big Data
- Coletar, armazenar e analisar dados de diferentes fontes, em diferentes formatos e gerados em deferentes velocidades.
- Migrar do sistema tradicional de coleta e armazenamento de dados, para uma estrutura de Big Data

E você acha que já temos muitos dados atualmente?
Espere para ver o que a Internet das Coisas vai fazer com o volume atual de dados!



Dados criados pelas IOT (Internet das coisas)



Estimativa da Oracle em relação a quantidade de dados

1.3.3 E como iniciar projetos de Big Data?

- Comece por compreender o valor de retorno sobre o investimento
- Não ignore os dados de todos os departamentos da empresa
- Big Data não é apenas sobre tecnologia. É sobre mudança de paradigma.
- *Não construa paredes contrua pontes*
- Não inicie um projeto de Big Data, sem antes entender o ROI (Retorno sobre o Investimento)

Até 2018, haverá um deficit de 140 a 190 mil profissionais com habilidades em análise de dados e mais de 1,5 milhões de gerentes e analista que saibam usar Big Data de forna efetiva para tomada de Decisões.

- McKinsey Global Institute "Big Data Report 2015"

1.4 Os 4V´s do Big Data

- Volume – Tamanho dos dados
- Variedade – Formato dos dados
- Velocidade – Geração dos dados
- Veracidade – Confiabilidade dos dados

Ver imagem: http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg

1.4.1 Volume

- Espera-se que 40 zettabytes de dados sejam criados até 2020 no mundo;
- Cerca de 2.5 quintillionbytes de dados são criados por dia;
- Existem atualmente cerca de 6 bilhões de telefones móveis no planeta;
- Cada empresa americana armazena cerca de 100 terá bytes de dados.

1.4.2 Variedade

- 150 exabytes é a estimativa de dados que foram gerados especificamente para tratamento de casos de doença em todo o mundo no ano de 2011;
- Mais de 4 bilhões de horas por mês são usadas para assistir vídeos no YouTube;
- 30 bilhões de imagens são publicadas por mês no Facebook;
- 200 milhões de usuários ativos por mês, publicam 400 milhões de tweets por dia.

1.4.3 Velocidade

- 1 terabyte de informação é criada durante uma única sessão da bolsa de valores Americana, a New York Stock Exchange (NYSE);
- Aproximadamente 100 sensores estão instalados nos carros modernos para monitorar nível de combustível, pressão dos pneus e muitos outros aspetos do veículo;
- 18.9 bilhões de conexões de rede existirão até 2016.

1.4.4 Veracidade

- Atualmente, 1 em cada 3 gestores tem experimentado problemas relacionados a veracidade dos dados para tomar decisões de negócios.
- Além disso, estima-se que 3.1 trilhões de dólares por ano sejam desperdiçados devido a problemas de qualidade dos dados.

1.4.5 O Big Data traz um oceano de oportunidades!

- Processar de forma eficiente e com baixo custo grandes volumes de dados
  - Transformar 12 TB de tweets gerados cada dia em produtos de análise de sentimento
- Responder ao aumento da velocidade de geração dos dados
  - Investigar 5 milhões de eventos de trade nas bolsas de valores a fim de identificar fraudes
- Coletar e analisar dados de diferentes formatos e fontes
  - Monitorar milhares de vídeos de segurança a fim de identificar pontos perigosos em uma cidade

1.5 Quiz

1) Qual dos seguintes se referem a tamanho dos dados em Big Data?
    a. Variedade
    b. Valor
    c. Volume
    d. Velocidade

Reposta certa é a C.

2) Quais os 4Vs do Big Data?
    a. Volume, Variedade, Velocidade e Veracidade
    b. Volume, Valor, Volatilidade, Velocidade
    c. Volume, Setorização, Volatilização e Valorização
    d. Volume, Variedade, Valor e Volatilidade

Resposta certa é a A.

3) Qual a estimativa de Volume do Big Data criado diariamente?
    a. Cerca de 2.5 terabytes de dados são criados por dia
    b. Cerca de 2.5 millionbytes de dados são criados por dia
    c. Cerca de 2.5 quintillionbytes de dados são criados por dia
    d. Cerca de 2.5 megabytes de dados são criados por dia

Resposta certa é a C.

4) Aproximadamente 10 sensores estão instalados nos carros modernos para monitorar nível de combustível, pressão dos pneus e muitos outros aspetos do veículo.
    a. Verdadeiro
    b. Falso

Resposta certa é a B. Aproximadamente 100 sensores estão instalados nos carros modernos para monitorar nível de combustível, pressão dos pneus e muitos outros aspetos do veículo.

5) Qual das seguintes, está correta, no tocante as oportunidades do Big Data?
    a. Processar de forma eficiente e com baixo custo grandes volume de dados.
    b. Responder ao aumento da velocidade de geração dos dados.
    c. Coletar e analisar dados de diferentes formatos e fontes.
    d. Todas estão corretas.

Resposta certa é a D.

2 – Introdução ao Hadoop

2.1 Introdução ao Hadoop

Apache Hadoop é um software open source para armazenamento e processamento em larga escala de grandes conjuntos de dados (Big Data), e clusters e hardware de baixo custo.

Temos visto o aumento crescente da capacidade de armazenamento dos discos rígidos.

Mas a velocidade de leitura e escrita dos discos rígidos não tem crescido na mesma proporção.

Leitura e escrita paralela e simultânea em diversos discos rígidos, requer tecnologia avançada.

Hadoop é um sistema de armazenamento compartilhado, distribuído e altamente confiável para processamento de grandes volumes de dados através de clusters de computadores.

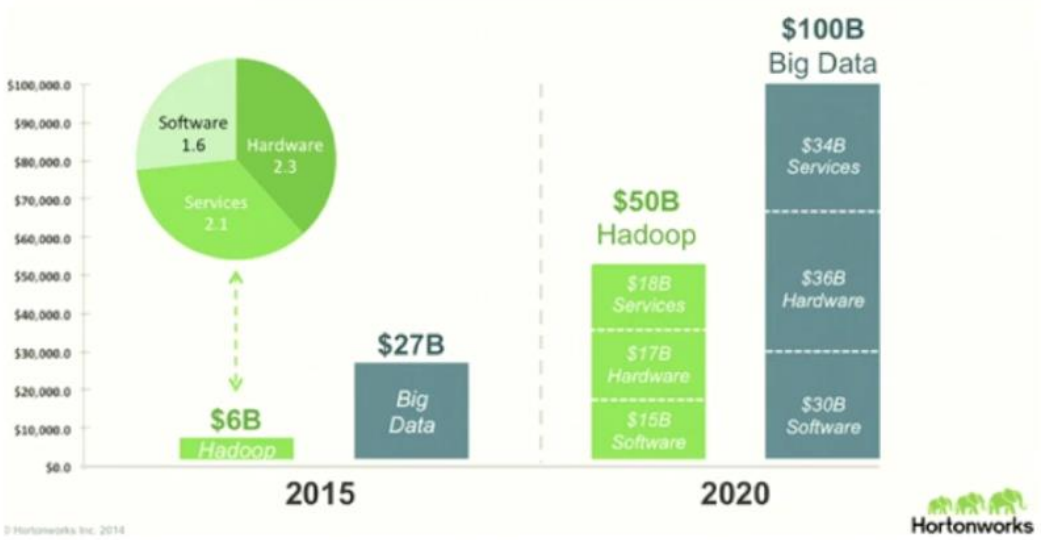
Em outras palavras Hadoop é um framework que facilita o funcionamento de diversos computadores, com o objetivo de analisar grandes volumes de dados.

O projeto Apache Hadoop é composto de 3 módulos principais:

- Hadoop Distributed File System (HDFS)
- Hadoop Yarn
- Hadoop MapReduce

Hadoop is for problems too BIG for traditional systems to handle

Pesquisas tem mostrado que o crescimento o Hadoop tem sido vertiginoso:



Hadoop é um framework gratuito, baseado em linguagem de programação Java, que suporta o processamento de grandes conjuntos de dados em ambientes de computação distribuída (através diversos computadores simultaneamente).

Ele é baseado no Google File System (GFS)

Hadoop permite executar aplicações em sistemas distribuídos através de diversos computadores (nodes), envolvendo petabytes de dados.

Hadoop utiliza o HDFS (Hadoop Distributed File System), que permite rápida transferência de dados entre os nodes. A segurança do Hadoop é feita com o Kerberos.

Hadoop é usado quando problemas muito grandes (Big) precisam de solução.

Hadoop tem um baixo custo, não apenas por ser livre, mas por permitir o uso de hardware simples, computadores de baixo custo agrupados em cluster.

Uma das principais características do Hadoop é a confiabilidade e sua capacidade de se recuperar de falhas automaticamente.

2.2 Componentes Principais do Hadoop

O Apache Hadoop é composto de 2 componentes principais:

- Hadoop HDFS
- Hadoop MapReduce

De uma forma bem simples, podemos dizer:

- HDFS – armazenamento distribuído
- MapReduce – computação distribuída

Porque é que o Hadoop está se tornando o padrão nos projetos de Big Data?

- Baixo Custo
- Escalável
- Tolerante a Falhas
- Flexível
- Livre

2.2.1 Hadoop HDFS

- Tolerância a falhas a recuperação automática
- Portabilidade entre hardware e sistemas operacionais heterogêneos
- Escalabilidade para armazenar e processar grandes quantidades de dados
- Confiabilidade, através da manutenção de várias cópias de dados

2.2.2 Hadoop MapReduce

- Flexibilidade – processa todos os dados independente do tipo e formato, seja estruturado ou não-estruturado
- Confiabilidade – permite que os jobs sejam executados em paralelo e em caso de falhas de um job, outro não são afetados
- Acessibilidade – suporte a diversas linguagens de programação como Java, C++, Python, Apache Pig

2.3 HDFS

2.3.1 Introdução

- Foi desenvolvido utilizando o projeto do sistema de arquivos distribuídos (DFS). Ele é executado em hardware commodity (baixo custo). Ao contrário de outros sistemas distribuídos, HDFS é altamente tolerante a falha.

- DFS (Distributed File System) – foi criado para gestão de armazenamento em uma rede de computadores.
- HDFS é otimizado para armazenar grandes arquivos.
- HDFS foi pensado para executar em clusters de computadores de baixo custo.
- HDFS foi pensado para ser ótimo em performance do tipo WORM (Write Once, Read Many Times), que é um eficiente padrão de processamento de dados.
- HDFS foi pensado considerando o tempo de leitura de um conjunto de dados inteiro e não apenas o primeiro registo.

HDFS cluster possui 2 tipos de nodes:

- Namenode (master node)
- Datanode (worker node)

2.3.2 Namenode

- Gerência a estrutura do filesystem
- Gerência os metadados de todos os arquivos e diretórios dentro da estrutura

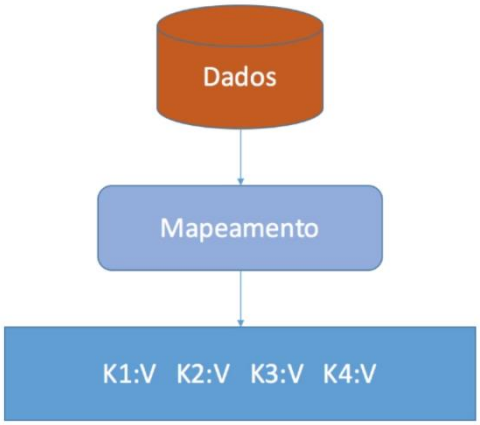
2.3.3 Datanode

- Armazena e busca blocos de dados quando solicitado pelo cliente ou Namenode
- Reporta periodicamente para o Namenode com a lista de blocos que foram armazenados.

2.4 MapReduce

2.4.1 Introdução

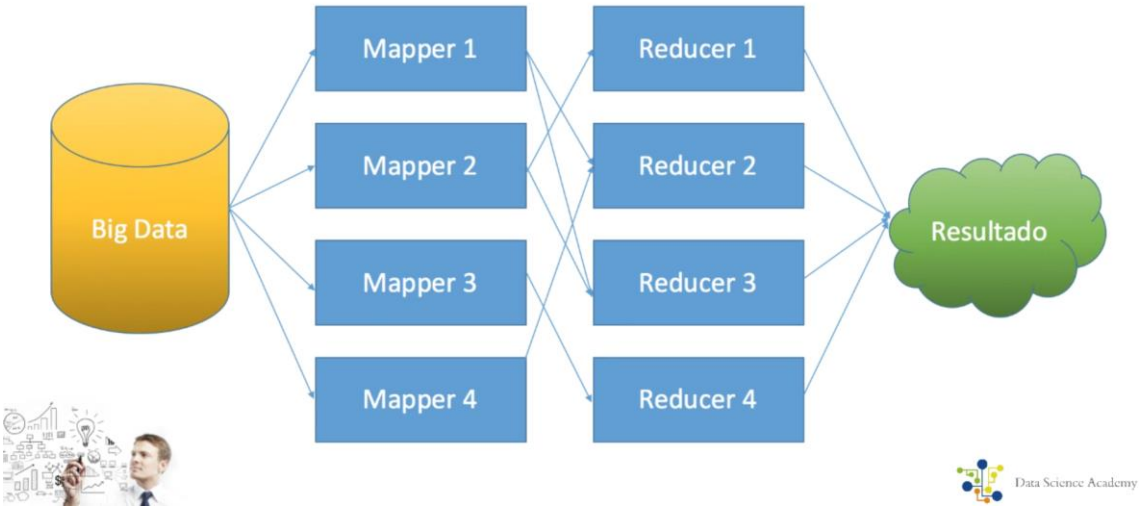
- MapReduce é um modelo de programação para processamento e geração de grandes conjuntos de dados.
- MapReduce transforma o problema de análise em um processo computacional que usa conjuntos de chaves e valores.
- MapReduce foi desenvolvido para tarefas que consomem minutos ou horas em computadores conectados em rede de alta velocidade gerenciados por um único master.
- MapReduce usa um tipo de análise de dados por força bruta. Todo o conjunto de dados é processado em cada query.
- Modelo de processamento em batch.



A Função de mapeamento, converte dados em pares de chave (K) / valor (V)

K = Key
V = Value

A principal vantagem do MapReduce é a sua escalabilidade de processamento de dados ao longo de varios nodes de computação.



2.5 Seek Time x Transfer Rate

- MapReduce permite a execução de queries ad-hoc em todo o conjunto de dados em um tempo escalável.
- Muitos sistemas distribuídos combinam dados de múltiplas fontes (o que é bem complicado), mas MapReduce faz isso de forma eficiente e efetiva.
- O segredo da performance do MapReduce, esta no balanceamento entre seeking e transfer: reduzir operações de seeking e usar de forma efetiva as operações de transfer.

> **Seek time** – é o delay para encontrar um arquivo.
>
> **Tranfer rate** – é a velocidade para encontrar o arquivo.

Transfer rates tem melhorado significativamente (é bem mais veloz que **Seek times**)

- O **MapReduce** é bom para atualizar todo (ou a maior parte) de um grande conjunto de dados.
- **RDBMS** (Relational Database Management System) são ótimos para atualizar pequenas porções de grandes bancos de dados.
- **RDBMS** utiliza o tradicional B-Tree, que é altamente dependente de operações de seek.
- **MapReduce** utiliza operações de Sort e Merge para recriar o banco de dados, o que é mais dependente de operações e transfer.

O MapReduce se baseia em operações de transfer o que deixa o acesso aos dados muito mais veloz.

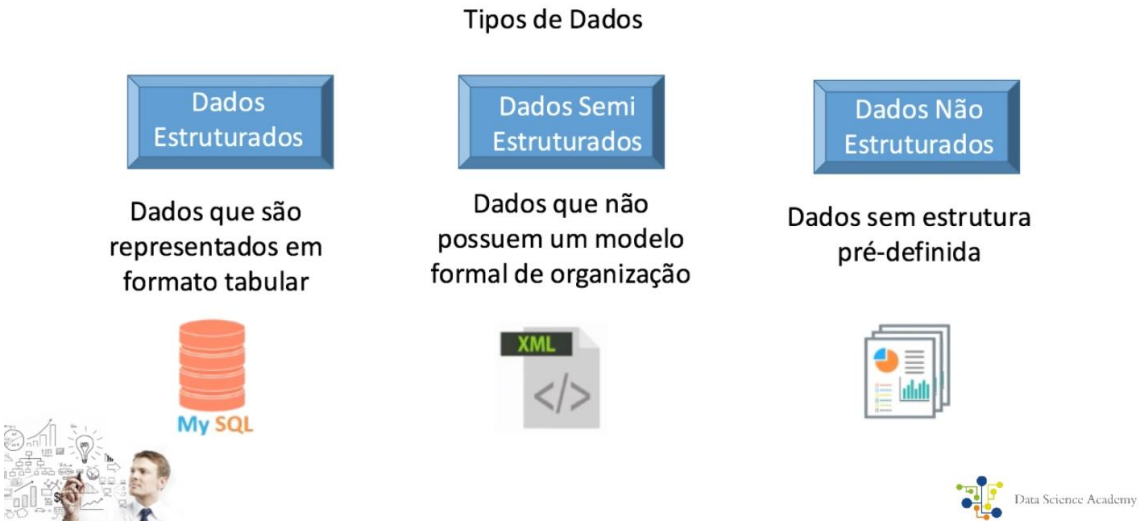
MapReduce x RDBMS

	RDBMS*	MapReduce
Tamanho dos dados	Gigabytes (10^9)	Petabytes (10^12)
Acesso	Interativo e Batch	Batch
Updates	Leitura e Escrita diversas vezes	WORM (Write Once, Read Many Times)
Estrutura de Dados	Esquema estático	Esquema dinâmico
Integridade	Alta	Baixa
Escalabilidade	Não-linear	Linear

* RDBMS = Relational Database Management System

Data Science Academy

2.6 Tipos de dados



MapReduce é muito efetivo com dados semi ou não estruturados!

MapReduce interpreta dados durante as sessões de processamento de dados. Ele não utiliza propriedades intrínsecas. Os parâmetros usados para selecionar os dados, são definidos pela pessoa que está fazendo a análise.

2.7 Quiz

1) Quais os principais módulos do Apache Hadoop?
    a. Hadoop File System
    b. Hadoop Yarn
    c. Hadoop MapReduce
    d. Todas estão corretas

Resposta certa é a D.

2) Apache Hadoop é um software proprietário para armazenamento e processamento em larga escala de grandes conjuntos de dados (Big Data), em clusters de hardware de alto custo.
    a. Verdeiro
    b. Falso

Resposta correta é a B. Apache Hadoop é um software open source para armazenamento e processamento em larga escala de grandes conjuntos de dados (Big Data), em clusters de hardware de baixo custo.

3) O Hadoop é um framework gratuito baseado em qual linguagem de programação?
    a. Java
    b. PostGree
    c. C++
    d. Cobol

Resposta correta é a A.

4) Uma das principais características do Hadoop é a confiabilidade e sua capacidade de se recuperar de falhas automaticamente.
    a. Verdadeiro
    b. Falso

Resposta correta é a A.

5) HDFS cluster possui dois tipos de nodes?
    a. Namenode e Datanode
    b. Bytenode e Teranode
    c. Marter Node e Circle Node
    d. Nanosystem e Datanode

Resposta correta é a A.

6) Usa um tipo de análise de dados por força bruta. Todo o conjunto de dados é processado por cada query.
    a. Hive
    b. Mapreduce
    c. HDFS
    d. Cassandra

A resposta correta é B.

3 – Arquitetura Hadoop

3.1 Arquitetura

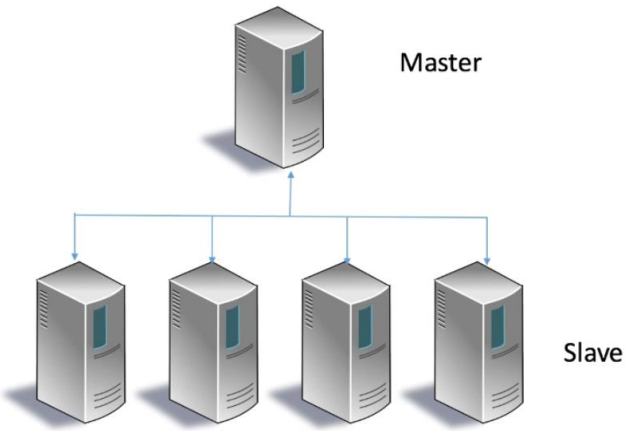
	Hadoop	*RDBMS*
Modelo de Computação	Conceito de Jobs Cada Job é uma unidade de trabalho Não há controlo de concorrência	Conceito de transações Uma transação é uma unidade de trabalho Controlo de concorrência
Modelo de dados	Qualquer tipo de dado pode ser usado Dados em qualquer formato Modelo de apenas leitura	Dados estruturados com controle de esquema Modelo de leitura / escrita
Modelo de Custo	Máquinas de custo mais baixo podem ser usadas	Servidores de maior custo são necessários
Tolerância a falhas	Simples, mas eficiente mecanismo de tolerância a falha	Falhas são raras de ocorrer Mecanismos de recuperação

O Apache Hadoop é composto de 2 componentes principais:

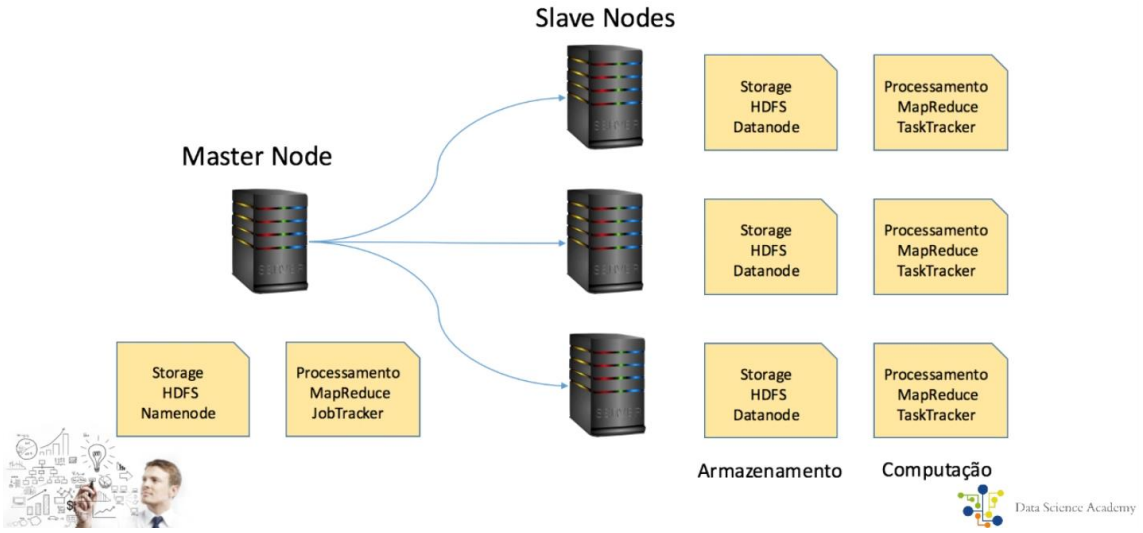
- Hadoop HDFS
- Hadoop MapReduce

Cluster Hadoop possui 2 tipos de nodes:

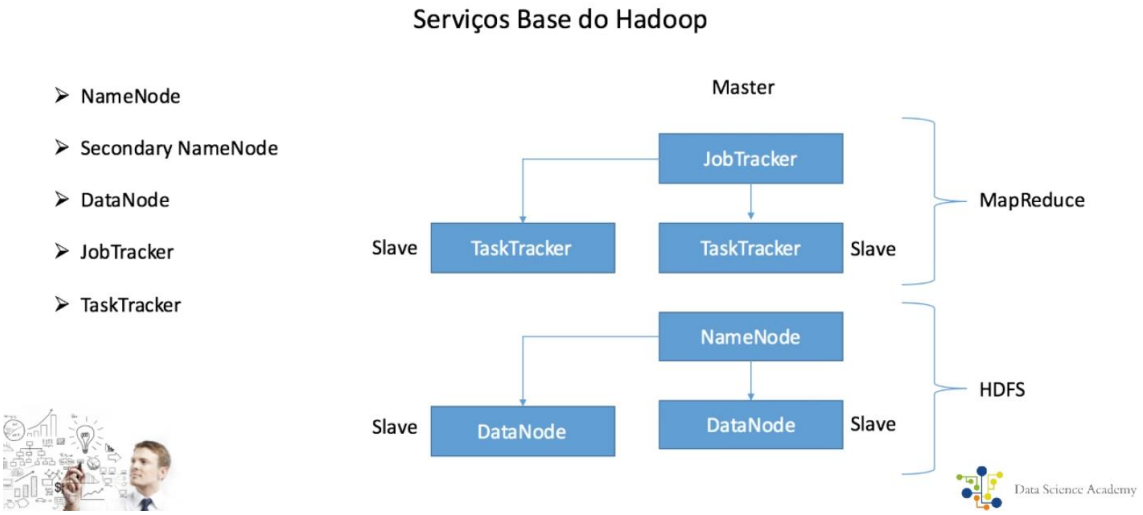
Master node
Worker (slave) node

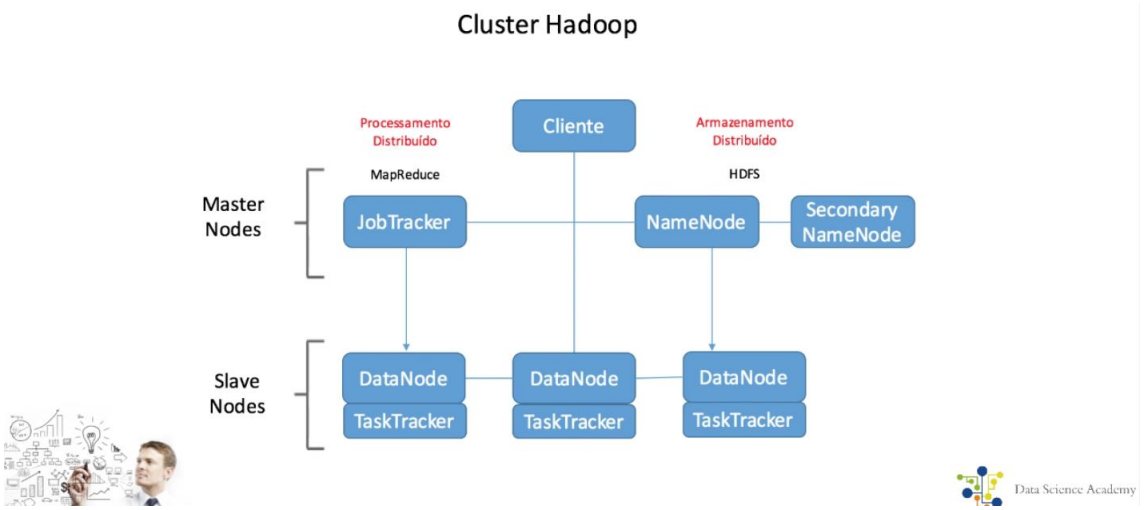


Data Science Academy

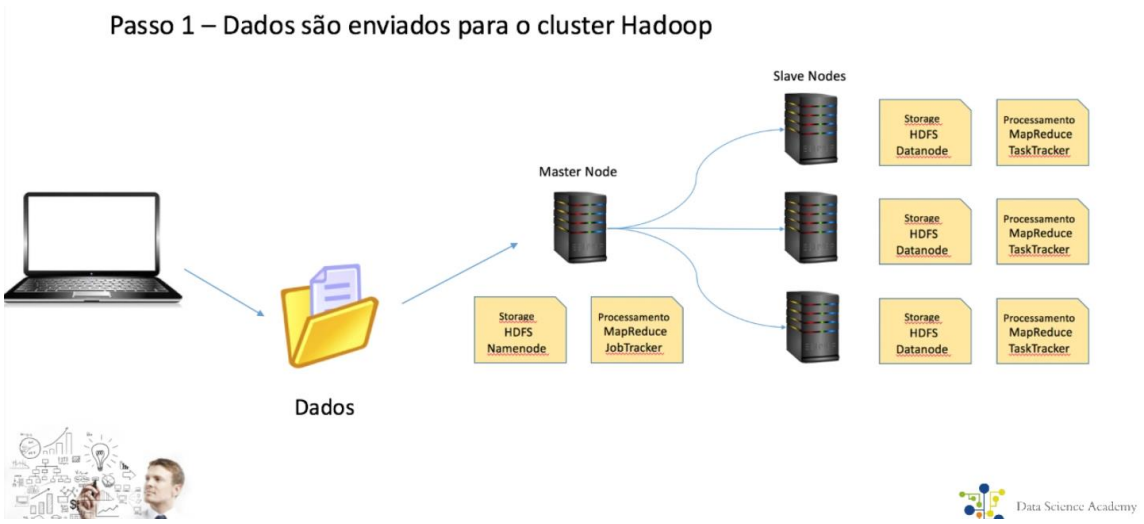


USER:Group	Daemons
hdfs:hadoop	NameNode, Secondary NameNode, JournalNode, DataNode
yarn:hadoop	ResourceManager, NodeManager
mapred:hadoop	MapReduce, JobHistory, Server





Todos clusters Hadoop basicamente funciona com 2 passos:

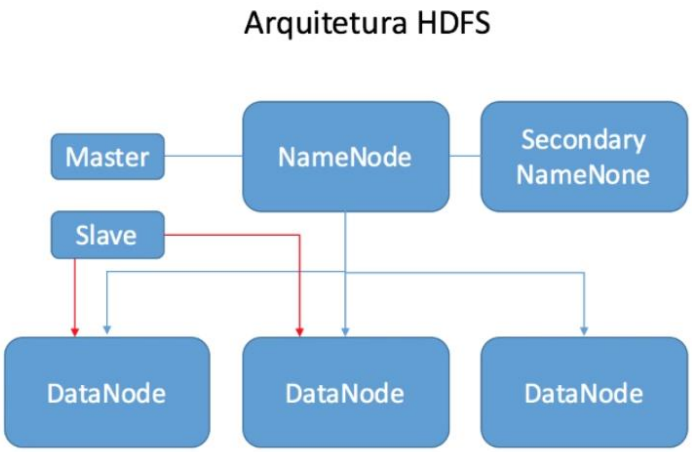


3.2 Modos de Configuração do Hadoop

Hadoop suporta 3 modos de configuração:

- Modo Standalone – Todos os serviços Hadoop são executados em uma única JVM, no mesmo servidor.
- Pseudo distribuído – Serviços individuais do Hadoop são atribuídos a JVM's individuais, no mesmo servidor.
- Totalmente distribuído – Serviços individuais do Hadoop são executados em JVM's individuais, mas através de cluster.

3.3 HDFS – Hadoop Distributed File System



O Hadoop Distributed File System (HDFS) é um Sistema de arquivos distribuído projetado para executar em hardwares simples (computadores básicos)

Ele tem muitas semelhanças com sistemas de arquivos distribuídos existentes. No entanto, as diferenças de outros sistemas de arquivos distribuídos são significativas.

HDFS é altamente tolerante a falhas e é projetado para ser implementado em hardware de baixo custo.

3.3.1 Funcionamento do processo de uma arquitetura HDFS

1) Os serviços NameNode e SecondaryNode, constitue os serviços Master. Os serviços DataNode são os slaves.
2) O serviço Master é responsável por aceitar os Jobs das aplicações clientes e garantir que os dados requeridos para a operação sejam carregados e segregados em pedaços de blocos de dados.
3) O HDFS permite que os dados sejam armazenados em arquivos. Um arquivo é dividido em um ou mais blocos que são armazenados e replicados pelo DataNodes. Os blocos de dados são então distribuídos para o sistema de DataNodes dentro do cluster. Isso garante que as réplicas de dados sejam mantidas.
4) As réplicas de cada bloco de dados são distribuídas em computadores em todo o cluster para permitir o acesso de dados confiável e de forma rápida.

3.4 Cluster HDFS

Cluster Single-Node	Cluster Multi-Node
Hadoop é instalado em um único computador (chamado node)	Hadoop é instalado em diversos nodes.
São usados para processamento mais simples, bem como operações triviais de MapReduce e HDFS	São usados para computação complexa, normalmente envolvendo aplicações de Analytics

3.5 Processamento MapReduce

MapReduce foi projetado para usar computação paralela distribuída em Big Data e transformar os dados em pedaços menores.

MapReduce funciona através de 2 operações:

Mapeamento e Redução.

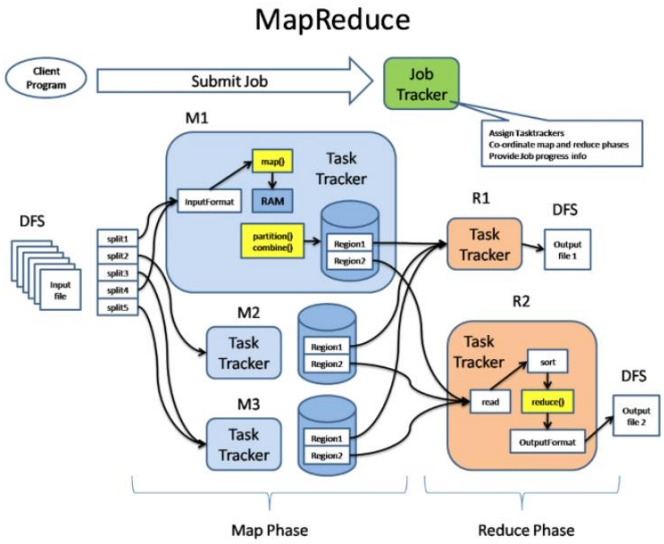
No processo de **mapeamento** (Map), os dados são separados em pares (key-value pairs), transformados e filtrados. Então os dados são distribuídos para os nodes e processados.

No processo de **redução** (Reduce), os dados são gerados em conjuntos de dados (datasets) menores. Os dados resultantes do processo de redução são transformados em um formato padrão de chave-valor (key-value), onde a chave (key) funciona como o indenrtificador do registo e o valor (value) é o dado (conteúdo) que é identificado pela chave.

3.5.1 Processo de MapReduce

Todo o processo se inicia com a requisição feita pelo cliente e o job submetido. O Job Tracker se encarrega de coordenar como o job será distribuído.

- **Mapeamento dos dados** – os dados de entrada são primeiramente distribuídos em pares key-value e divididos em fragmentos, que são então atribuídos a tarefas de mapeamento.
- **Redução dos dados** – cada operação de redução dos dados tem um fragmento atribuído.



3.5.2 MapReduce em Tempo Real.

MapReduce vem sendo largamente utilizado em aplicações real-time. Alguns exemplos:

- Classificação Bayesiana para operações de data mining.
- Operações de search engine, como indexação de keywords, rendering e page rank.
- Análise de Gaussian para localização de objetos astronômicos.
- Web Semática e Web 3.0

3.6 Cache Distribuído

Distributed Cache ou Cache Distribuído, é uma funcionalidade do Hadoop que permite cache dos arquivos usados pelas aplicações.

Isso permite ganhos consideráveis de performance quando tarefas de map e reduce precisam acessar dados em comum. Permite ainda, que um node do cluster acesse os arquivos no filesystem local, ao invés de solicitar o arquivo em outro node.

É possível fazer o cache de arquivos zip e tar.gz.

Uma vez que você armazena um arquivo em cache para o seu trabalho, a estrutura Hadoop irá torna-lo disponível em cada node (em sistema de arquivos, não em memória) onde as tarefas de mapeamento / redução estão em execução.

3.7 Segurança

O Hadoop utiliza o Kerberos, um mecanismo de autenticação usado por exemplo no sistema de diretórios dos servidores Windows e também no sistema operacional Linux.

Por padrão Hadoop é executado no modo não-seguro em que não é necessária a autenticação real. Após se configurado, o Hadoop é executado em modo de segurança e cada usuário e serviço precisa ser autenticado pelo Kerberos, a fim de utilizar os serviços do Hadoop.

3.8 Quiz

1. Pseudo Distribuído são serviços Hadoop executados em uma única JVM, no mesmo servidor.
    a. Verdadeiro
    b. Falso

Resposta: Falso. Pseudo Distribuído são serviços individuais do Hadoop são atribuídos a JVM's individuais, no mesmo servidor.

2. Quais os 3 modos de configuração suportados pelo Hadoop?
    a. Modo Standalone, PseudoDistribuído e Totalmente Distribuído
    b. Modo em Batch, PseudoDistribuído e Totalmente Distribuído
    c. Modo Standalone, Bath Distribuído e Totalmente Distribuído
    d. Modo em Bath, Distribuído e Totalmente Distribuído

Resposta: A. Modo Standalone, PseudoDistribuído e Totalmente Distribuído

3. Os serviços NameNode e SecondaryNode, constituem os serviços Master.Os serviços DataNode são os slaves.
    a. Veradeiro
    b. Falso

Resposta: Verdadeiro

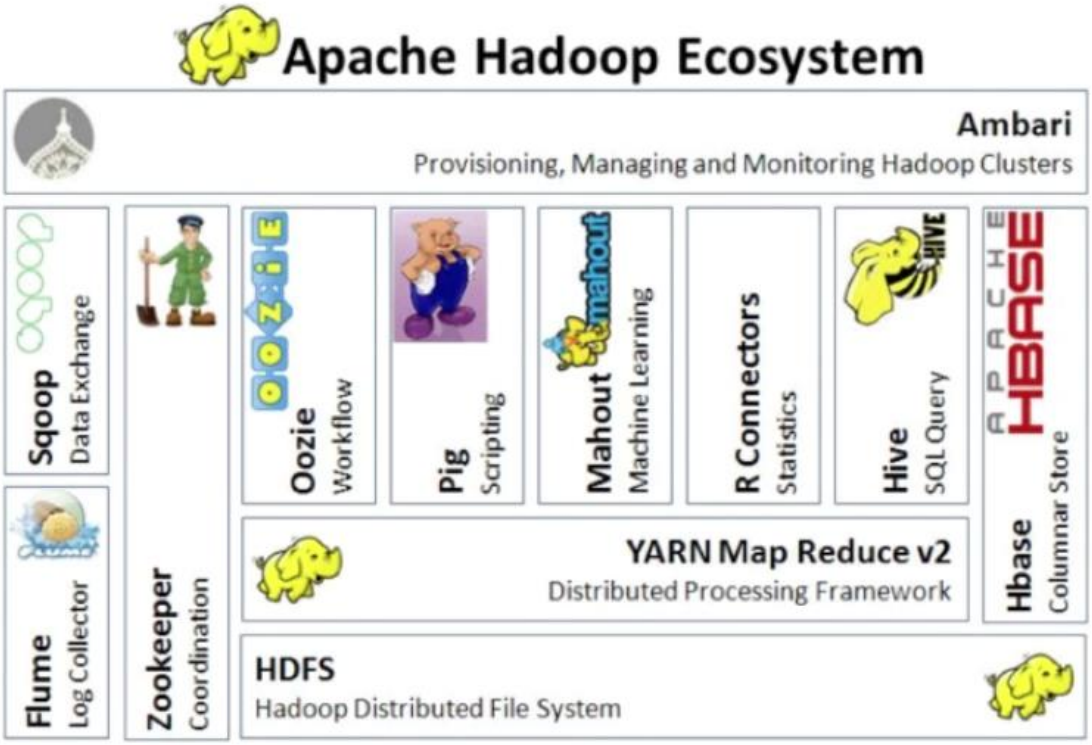
4. O HDFS permite que os dados sejam armazenados em arquivos. Um arquivo é dividido em um ou mais blocos que são armazenados e replicados pelos DataNodes. Os blocos de dados são então distribuídos para o sistema de DataNodes dentro do cluster. Isso garante que as réplicas de dados sejam mantidas.
    a. Verdadeiro
    b. Falso

Resposta: Veradeiro

5. Quais as características de um Cluster Single-Node?
    a. O Hadoop é instalado em um único computador.
    b. O Hadoop é instalado em diversos nodes.
    c. São usados para processamento simples, boas operações triviais de Mapreduce e HDFS.
    d. São usados para computação complexa, normalmente envolvendo aplicações de Analytics.
    e. Somente a e c estão corretas.

Resposta: E. Somente A e C estão corretas.

4 – Ecossistema Hadoop



Pense no ecossistema como as apps do sistema operacional iOS ou Android.

Os aplicativos servem para aprimorar a capacidade do SO.

Mesmo raciocínio pode ser aplicado para os componentes do ecossistema Hadoop.

4.1 ZooKeeper

Apache ZooKeeper é uma solução open-source de alta performance, para coordenação de serviços em aplicações distribuídas. Ele é uma espécie de guardião do Zoo!

ZooKeeper é um serviço de coordenação distribuída para gerenciar grandes conjuntos de hosts (Clusters).

Coordenação e gestão de um serviço em um ambiente distribuído é um processo complicado. ZooKepper resolve este problema com a sua arquitetura simples.

ZooKeeper permite que os desenvolvedores se concentrem na lógica do aplicativo principal sem se preocupar com a natureza distribuída do aplicativo.

O framework ZooKeeper foi originalmente construído no “Yahoo!” para aceder aos seus aplicativos de uma forma fácil e robusta.

Mais tarde, Apache ZooKeeper se tornou um padrão para a organização de serviços do Hadoop, HBase e outras estruturas distribuídas.

Por exemplo, o HBase usa ZooKeeper para acompanhar o estado de dados distribuídos através do Cluster.

ZooKeeper proporciona um ponto comum de acesso a uma ampla variedade de objetos utilizados em ambientes de Cluster

4.2 Oozie

Apache Oozie é um sistema de agendamento de workflow usado para gerenciar principalmente os Jobs de MapReduce.

Oozie é integrado com o restante dos componentes do ecossistema Hadoop para apoiar vários tipos de trabalhos do Hadoop (como Java Map-Reduce, streaming Map-Reduce, Pig, Hive e Sqoop), bem como Jobs específicos do sistema (como programas Java e scripts shell).

Oozie é um sistema de processamento de fluxo de trabalho que permite aos usuários definir uma séria de jobs escritos em deferentes linguagens – como Map Reduce, Pig e Hive – e então inteligentemente liga-los um ao outro.

Oozie permite aos usuários especificar, por exemplo, que uma determinada consulta só pode ser iniciada após os jobs anteriores que acedam os mesmos dados sejam concluídos.

Oozie é um sistema versátil que pode ser usado para configurar e automatizar até memso o mais complicado workflow de processamento de dados.

Lembre-se que estamos a falar de processamento de Big Data, em Clusters que podem chegar a milhares de nodes.

4.3 Hive

4.3.1 Apache Hive

Apache Hive é um Data Warehouse que funciona com Hadoop e MapReduce.

Hive é um sistema de armazenamento de dados para Hadoop que facilita a agregação dos dados para relatórios e análise de grandes conjuntos de dados (Big Data).

Hive permite consultas sobre os dados usando uma linguagem SQL-like, chamada HiveQL (HQL).

Provê capacidade de tolerância a falha para armazenamento de dados e depende do MapReduce para execução.

Ele permite conexões JDBC / ODBC, por isso é facilmente integrado com outras ferramentas de inteligência de negócios como Tableau, Microstrategy, Microsoft Power BI entre outras.

Hive é orientado a bach e possui alta latência para execução de queries. Assim como o Pig, gera jobs Map Reduce que executam no cluster Hadoop.

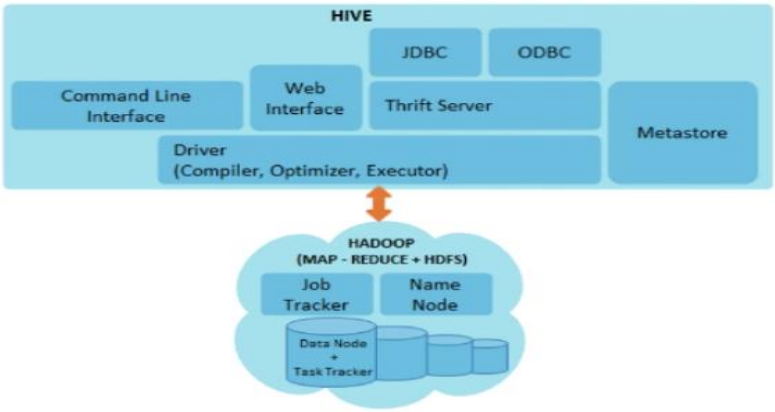
Foi desenvolvido pelo Facebook.

Hive é um sistema para gestão de query de dados não estruturados, em formato estruturado.

Hive utiliza:

- MapReduce para execução
- HDFS para armazenamento e pesquisa de dados

Arquitetura e Componentes Hive



4.3.2 Hive Query Language – HQL

Hive Query Language (HQL) é a linguagem de queries para o engine Hive.

HQL suporta os conceitos básicos da linguagem SQL, como:

- Clásula From
- ANSI Join (somente equi-join)
- Insert
- Group-by
- Sampling

Exemplos:

hive> select * from tb_folha_pagamento;

hive> show tables;

hive> describe tb_folha_pagamento;

4.4 Sqoop

Scoop é um projeto do ecossistema do Apache Hadoop, cuja responsabilidade é importar e exportar dados de bancos de dados relacionais.

Sqoop significa SQL-to-Hadoop

Basicamente, o Sqoop permite mover os dados de bancos tradicionais como Microsoft SQL Server ou Oracle, para o Hadoop.

É possível importar tabelas individuais ou bancos de dados inteiros para HDFS e o desenvolvedor pode determinar que colunas ou linhas serão importadas.

Ferramenta desenvolvida para transferir dados do Hadoop para RDBMS e vice-versa.

Transforma os dados no Hadoop, sem necessidade de desenvolvimento adicional.

Ele também gera classes Java através das quais você pode facilmente interagir com os dados importados.

Utiliza conexão JDBC para conectar com os bancos de dados relacionais.

Pode criar diretamente tabelas no Hive e suporta importação incremental.

Exemplo:

Listando tabelas de um banco MySQL com Sqoop:

```
sqoop list-tables -username dsacademy -password dasacademybr \ --connect
jdbc:mysql://dbname
```

4.5 Pig

Apache Pig é uma ferramenta que é utilizada para analisar grandes conjuntos de dados que representam fluxos de dados.

Pig é geralmente usado com Hadoop; podemos realizar todas as operações de manipulação de dados no Hadoop usando Apache Pig.

Para escrever programas de análise de dados, Pig oferece uma linguagem de alto nível conhecido como Pig Latin. Esta linguagem fornece vários operadores que os programadores podem usar para criar as suas próprias funções para leitura, escrita e processamento de dados.

Para analisar dados usando Apache Pig, os programadores precisam escrever scripts usando linguagem Pig Latin. Todos esses scripts são convertidos internamente para tarefas de mapeamento e redução. Apache Pig tem um componente conhecido como Pig engine que aceita os scripts Pig Latin como entrada e converte esses scripts em Jobs MapReduce.

Componentes do pig:

- Pig Latin Script Language:
    - Linguagem procedimental de fluxo de dados
    - Contém sintaxe e comandos que podem ser aplicados para implementar lógica de negócios.

- Runtime engine:
    - Compilador que produz sequências de programas MapReduce
    - Utiliza HDFS para armazenar e buscar dados
    - Usado para interagir com sistemas Hadoop
    - Valida e compila script de operações em sequências de Jobs MapReduce

Pig X SQL

Pig	SQL
Linguagem de script usada para interagir com o HDFS	Linguagem de query usada para interagir com bancos de dados
Passo a passo	Bloco único
Avaliação não imediata	Avaliação imediata
Permite resultados intermediários	Requer que um join seja executado 2 vezes ou materializado como um resultado intermediário

4.6 HBase

Apache HBase é um banco de dados orientado a coluna construído sobre o sistema de arquivos Hadoop.

HBase é o banco de dados oficial do Hadoop.

HBase é um modelo de dados que é semelhante ao Big Table do Google projetado para fornecer acesso aleatório rápido a grandes quantidades de dados.

Ele aproveita a tolerância a falhas fornecida pelo sistema de arquivos Hadoop (HDFS). É uma parte do ecossistema Hadoop que fornece em tempo real acesso aleatório de leitura / gravação aos dados do HDFS.

Pode-se armazenar os dados em HDFS quer diretamente quer através do HBase.

- O HBase é um tipo de banco de dados NoSQL e utiliza o modelo key-value (chave-valor).
- Cada valor é identificado por uma chave.
- Chaves e valores são do tipo byte-array.
- Valores são armazenados por ordem de acordo com a chave.
- Os valore podem ser facilmente acedidos por suas respetivas chaves.
- No HBase, as tabelas não possuem schemas.

O objetivo do HBase é armazenar tabelas realmente grandes, com bilhões de dados.

Arquitetura HBase

HBase possui 2 tipos de Nodes: Master e RegionServer

Master	RegionServer
Somente um node Master pode ser executado. A alta disponibilidade é mantida pelo ZooKeeper	Um ou mais podem existir
Responsável pela gestão de operações de cluster, como assignment, load balancing e splitting	Responsável por armazenar as tabelas, realizar leituras e buffers de escrita
Não faz parte de operações de read/write	O cliente comunica com o RegionServer para processar operações de leitura/escrita

Subconjuntos de dados de tabelas, são chamadas de regiões no HBase. O Node Master deteta o status dos RegionServers e atribui regiões a eles.

HBase x RDBMS

HBase	RDBMS
Particionamento automático	Particionamento manual, realizado pelo administrador
Pode ser escalado de forma linear e automática com novos nodes	Pode ser escalado verticalmente com a adição de mais hardware
Utiliza hardware commodity	Requer hardware mais robustos e portanto, mais caros
Possui tolerância a falha	Tolerância a falha pode estar presente ou não
Com MapReduce, alavanca processos batch	Precisa de muitas threads ou processos para processamento

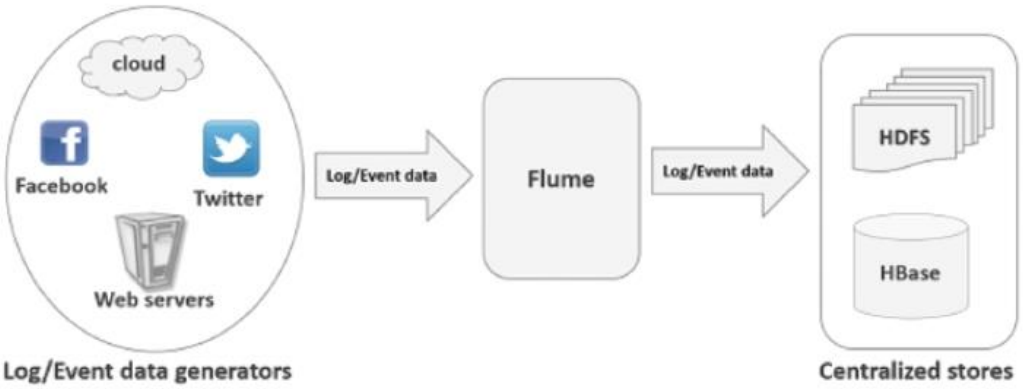
4.7 Flume

Apache Flume é um serviço que basicamente permite enviar dados diretamente para o HDFS.

Foi desenvolvido pela Cloudera e permite mover grandes quantidades de dados.

Basicamente, o Apache Flume é um serviço que funciona em ambiente distribuído para coletar, agregar e mover grandes quantidades de dados de forma eficiente.

Ele possui uma arquitetura simples e flexível baseada em streaming (fluxo constante) de dados.



O modelo de dados do Flume, permite que ele seja usado em aplicações analíticas online.

O Flume também pode ser usado em infraestruturas de TI. Agentes dão instalados em servidores web, servidores de aplicações ou aplicativos mobile, para coletar e integrar os dados com Hadoop, para análise online em tempo real.

4.8 Mahout

Apache Mahout é uma biblioteca open-source de algoritmos de aprendizado de máquina, escalável e com foco em clustering, classificação e sistemas de recomendação.

O Mahout é dedicado ao Machine Learning.

O Mahout permite a utilização dos principais algoritmos de clustering, testes de regressão e modelagem estatística e os implementa usando um modelo MapReduce.

E quando utilizar o Mahout?

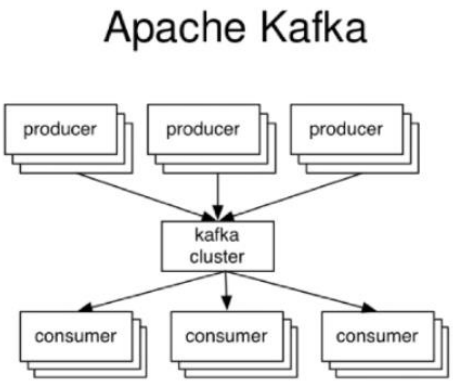
- Você pode utilizar algoritmos de Machine Learning com alta performance?
- Sua solução precisa ser open-source e livre?
- Você possui um grande conjunto de dados (Big Data) e pretende utilizar ferramentas de análise como R, Matlab e Octave?
- Seu processamento de dados será feito usando um modelo batch (você não precisa utilizar dados gerados em tempo real)?
- Você precisa de uma biblioteca madura e disponível no mercado há alguns anos que já tenha sido testada e validada?

4.9 Kafka

O Apache Kafka foi desenvolvido pelo LinkedIn e posteriormente liberado como um projeto open-source, em 2011.

O Apache Kafka é um sistema para gerenciamento de fluxos de dados em tempo real, gerados a partir de web site, aplicações e sensores.

Essencialmente, o Kafka age como uma espécie de “sistema nervoso central”, que coleta dados de alto volume como por exemplo a atividade de usuários (clicks em um web site), logs, cotações de ações etc… e torna estes dados disponíveis como um fluxo em tempo real para o consumo por outras aplicações.



O Apache Kafka foi desenvolvido com um propósito específico em mente: servir como um repositório central de fluxos de dados.

Mas por que fazer isso?

Havia duas motivações.

- Integração dos dados
- Baixa latência

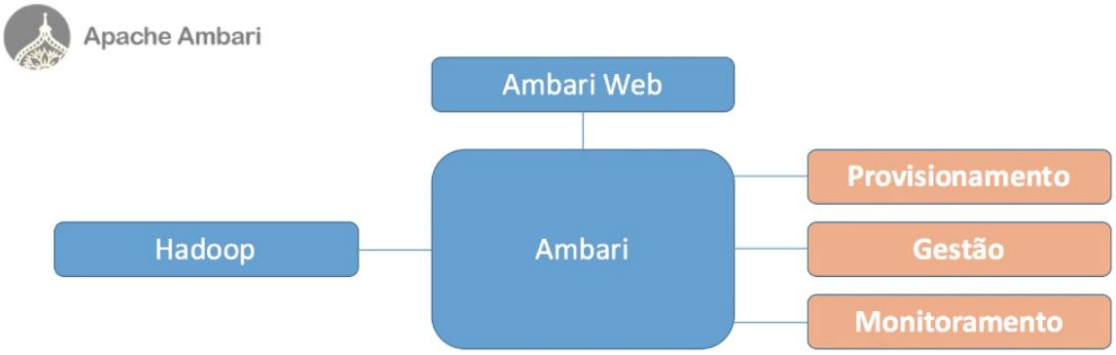
O Apache Kafka está ajudando a mudar a forma como os dados são usados dentro das empresas.

Não faz mais sentido falar apenas em dados armazenados em tabelas, com linhas e colunas.

O volume de dados agora é tão grande, que os dados precisam ser vistos como o que realmente são: um fluxo constante, que precisa ser analisado em tempo real.

4.10 Ambari

Apache Ambari é um framework para provisionamento, getão e monitoramento de clusters Apache Hadoop.



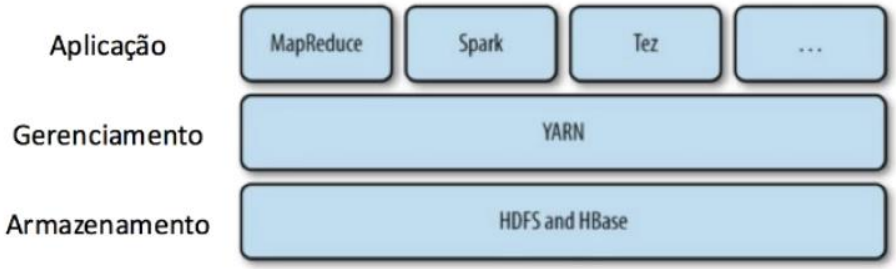
Apache Ambari é um conjunto de ferramentas para administrar e monitorar cluster Hadoop, que foi desenvolvido pela equipe de engenheiros de Hortonworks.

4.11 YARN

O Apache YARN é um sistema gerenciador de Cluster Hadoop.

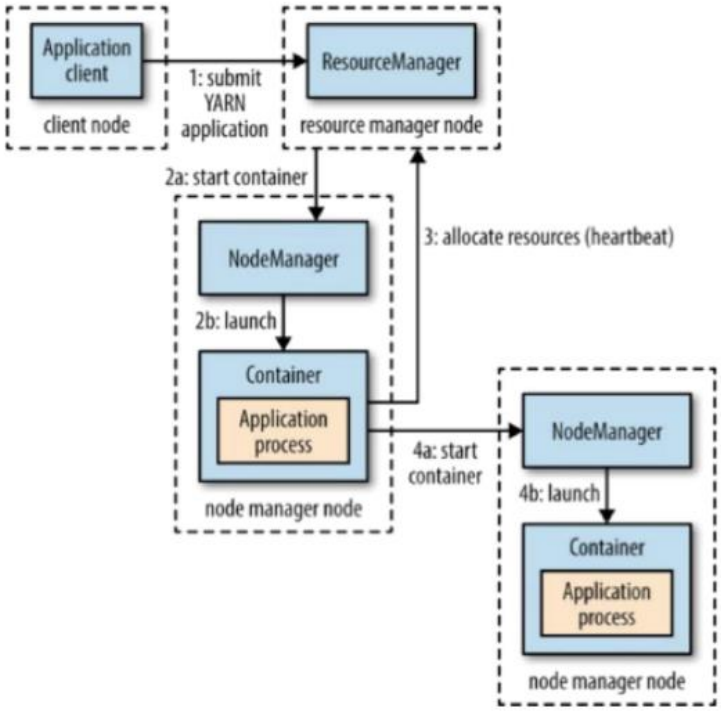
YARN (Yet Another Resource Negotiator).

O YARN foi introduzido na versão 2.0 do Hadoop para melhorar a implementação do MApReduce, mas ele suporta outros paradigmas de computação distribuída.



O YARN funciona através de 2 serviços:

- Resource Manager (um por cluster)
- Node Manager (que é executado em todos os nodes do cluster)



Uma das principais funções do YARN é garantir que os algoritmos de processamento dos dados distribuídos, utilizem de forma eficiente os recursos do Cluster.

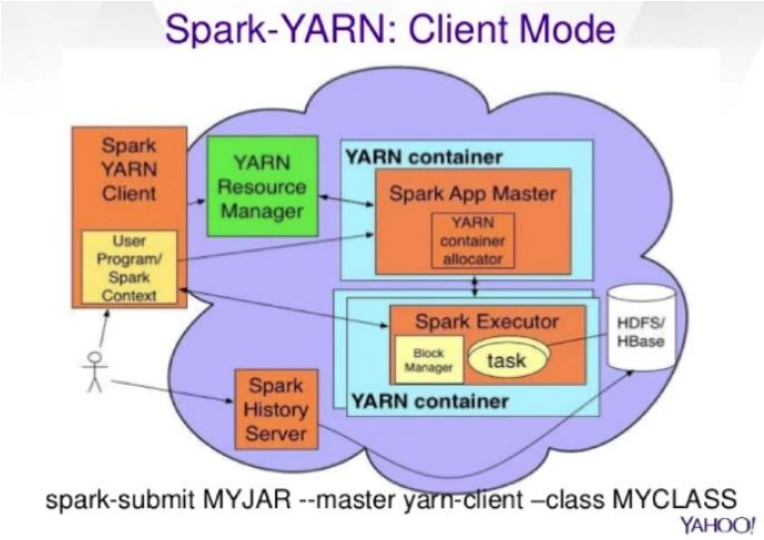
Também é possível utilizar o Spark sobre o YARN, que é o método mais conveniente de usar o Spark, quando existe um Cluster Hadoop.

Existem 2 modos de executar o Spark com YARN:

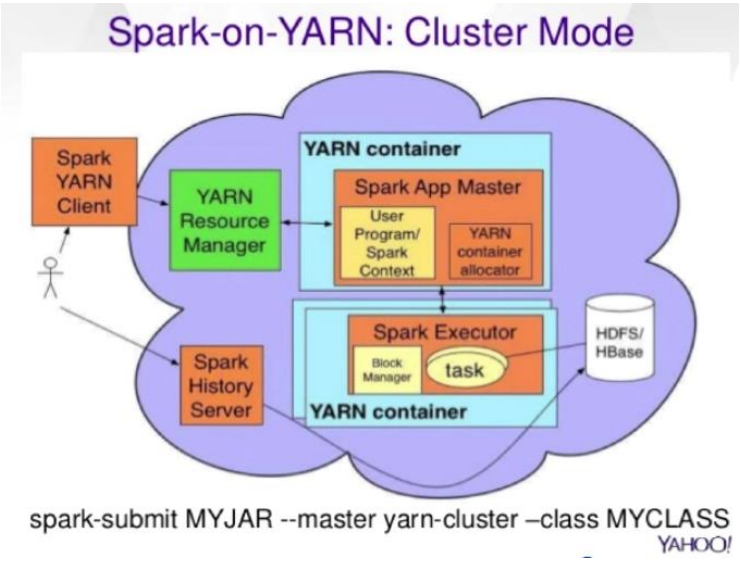
- YARN Client Mode
- YARN Cluster Mode

O YARN Client Mode é utilizado quando o programa possui um componente interativo, como o spark-shell ou pyspark.

O Client Mode é também importante quando se está construindo programas Spark, pois o debug é imediatamente visível.



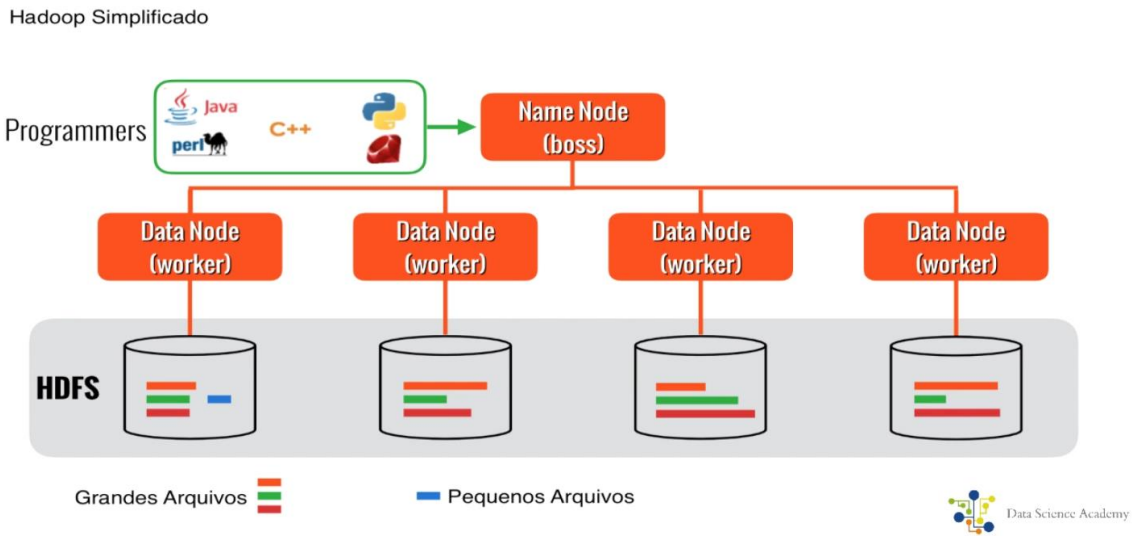
O YARN Cluster Mode é indicado para os jobs em ambiente de produção, pois toda a aplicação será executada em Cluster.



Característica	Descrição
Compatibilidade	Aplicações MapReduce desenvolvidas para o Hadoop versão 1.0, podem usar o YARN para execução com versões mais novas do Hadoop, sem mudar os processos existentes
Escalabilidade	O Resource Manager do YARN tem o foco em gerenciar o cluster, à medida que novos nodes são adicionados, expandindo o cluster para milhares de nodes e e petabytes de dados
Utilização do Cluster	O YARN promove a alocação dinâmica de recursos do cluster, melhorando sua utilização e agindo de forma muito mais eficiente que as regras estáticas do MapReduce

4.12 HDFS

- HDFS é um filesystem desenvolvido em Java e baseado no Google File System.
- Permite armazenar grande quantidades de dados em hardware de baixo custo.
- Foi criado para trabalhar com pouca quantidade de grandes arquivos de dados e não com muita quantidade de pequenos arquivos.
- Não é otimizado para operações de leitura randómica, como RDBMS's.
- WORM (Write Once Read Many Times).
- Os arquivos são gerados em bolocos de 64 a 128 MB.
- Os blocos são replicados através dos datanodes, com um fator de replicação padrão, igual a 3 (cada bloco é replicado 3 vezes).
- Os blocos replicados são armazenados em diferentes máquinas.
- O Namenode mantém um "mapa" de como os blocos compõem cada arquivo.
- O Namenode precisa estar disponível para que o Cluster Hadoop possa ser acessado.
- O Namenode tem os metadados gravados em Memória e periodicamente os grava em disco.



4.13 MapReduce

- MApReduce é um modelo de programação para processamento de grandes volumes de dados, tipicamente usado para computação distribuída em clusters.
- Jobs de Mapper e Reducer realizam as tarefas.
- Quando uma tarefa tenta processar um conjunto de dados e falha por 4 vezes, a tarefa é cancelada e o job falha.
- Todos os dados recebem a forma de pares chave-valor (key-value).

4.14 Quiz

1) O framework ZooKeeper foi originalmente construído no "Yahoo!" para acessar seus aplicativos de uma forma fácil e robusta. Mais tarde, Apache ZooKeeper se tornou um padrão para a organização de serviços do Hadoop, HBase e outras estruturas distribuídas. Por exemplo, HBase usa ZooKeeper para acompanhar o estado de dados distribuídos.
    a. Veradeiro
    b. Falso

Resposta: Veradeiro

2) Qual o nome do sistema de agendamento de workflow usado para gerenciar os Jobs de MapReducep?
    a. Hadoop File System
    b. Apache Oozie
    c. Zookeeper
    d. Todas estão corretas.

Resposta: B

3) Sobre o Hive é incorreto afirmar?
    a. Apache Hive é um Data Warehouse que funciona com Hadoop e MapReduce.
    b. Utiliza o Hadoop para armazenar e distribuir grandes conjuntos de dados.
    c. Provê uma linguagem baseada em SQL, chamada HiveQL (HQL). Por conta desta interface SQL, Hive é uma escolha natural para soluções de analytics com Hadoop.
    d. Hive é orientado a sistema distribuído e possui baixa latência para execução de queries

Resposta: D

4) Sobre o Sqoop é correto afirmar:
    a. é um projeto do ecossistema do Apache Hadoop, cuja responsabilidade é importar e exportar dados de bancos de dados relacionais.
    b. Sqoop permite mover os dados de bancos tradicionais como Microsoft SQL Server ou Oracle, para o Hadoop.
    c. Transforma os dados no Hadoop como MapReduce ou Hive, sem necessidade de desenvolvimento adicional.
    d. Todas estão corretas.

Resposta: D

5) Possui 2 tipos de Nodes chamados: Master e Region Server
    a. Apache Hive
    b. Apache Hbase
    c. Apache Pig
    d. Apache Flume

Resposta: B

6) O Mahout e dedicado ao Machine Learning. Este componente trabalha com clustering, que é uma das principais técnicas de Machine learning.
    a. Verdadeiro
    b. Falso

Resposta: A

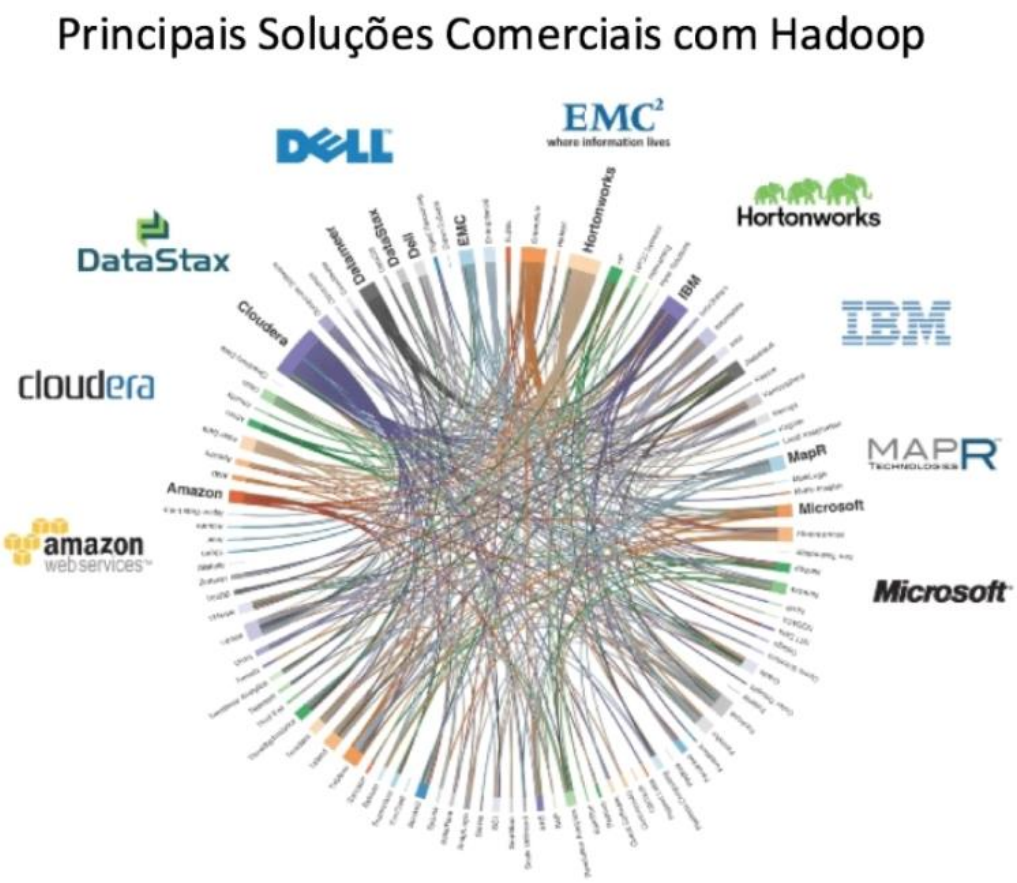
5 – Soluções Comerciais com Hadoop

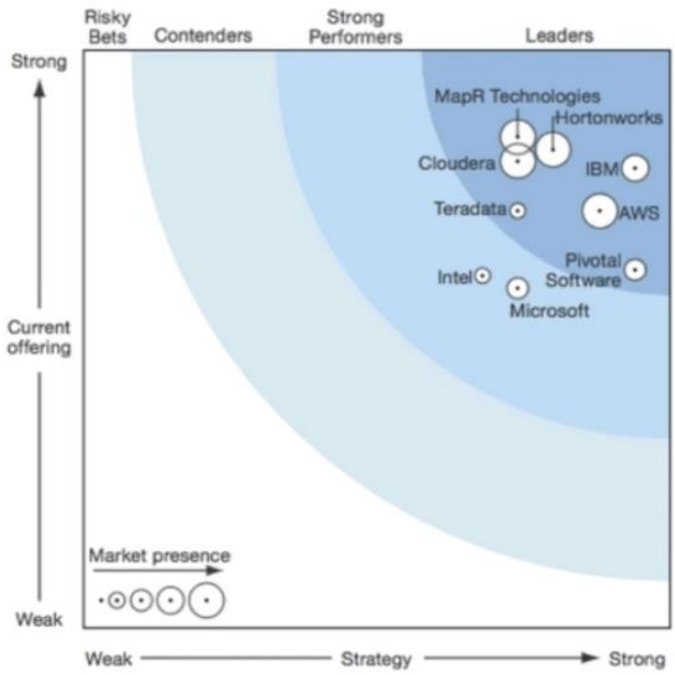
5.1 Por que usar soluções comerciais com Hadoop?

Você pode estar se perguntando:

Se o Hadoop é livre, porque eu usaria soluções comerciais do software?

- Suporte – as principais soluções comerciais do Hadoop oferecem suporte, guias, assistência e melhores práticas.
- Confiança – sempre que um bug é detetado, as soluções comerciais prontamente atualizam o software.
- Pacote completo – as soluções oferecem pacotes completos, com tudo que é necessário para uma infraestrutura de BigData.





5.2 Amazon Web Services Elastic MapReduce Hadoop

A distribuição Hadoop da Amazon, foi uma das primeiras distribuições comerciais do Hadoop.

AWS Elastic MapReduce é uma plataforma de análise de dados bem organizada e construída sobre a arquitetura HDFS.

Com foco principal em consultas de mapeamento / redução o AWS EMR explora ferramentas de Hadoop, fornecendo uma plataforma de infraestruturas escalável e segura para seus usuários.

Amazon Web Services EMR está entre uma das distribuições comerciais do Hadoop com a maior participação no mercado global.

Site: https://aws.amazon.com/pt/emr/

5.3 Cloudera

Cloudera Hadoop ocupa o topo na lista de grandes fornecedores de dados Hadoop, possui uma plataforma confiável para uso comercial desde 2008.

Cloudera, fundada por um grupo de engenheiros do Yahoo, Google e Facebook, está focada em fornecer soluções empresariais do Hadoop.

Cloudera Hadoop possui cerca de 350 clientes, incluído o Exército dos EUA, AllState e Monsanto.

Alguns deles com implementações de 1000 nós em cluster Hadoop para análise de dados de cerca de um petabyte.

Cloudera utiliza produtos 100% open-source.

- Apache Hadoop
- Apache Pig
- Apache Hive

- Apache HBase
- Apache Sqoop

Cloudera possui um sistema amigável de gestão, chamado Cloudera Manager, para gestão de dados e que possui suporte técnico.

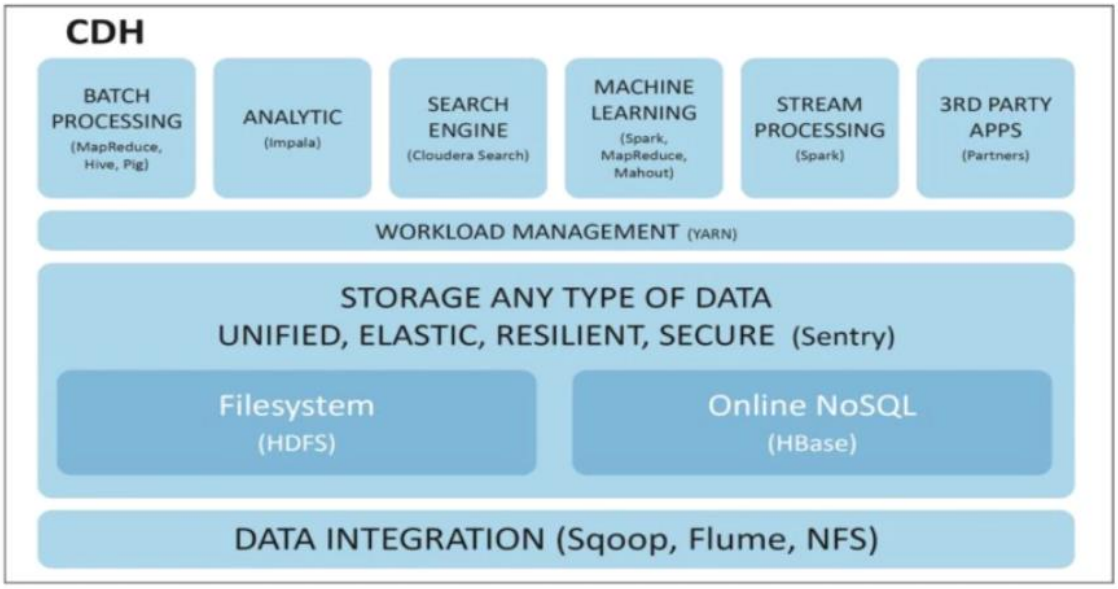


Image source: cloudera.com

Site: https://www.cloudera.com/

5.4 Hortonworks

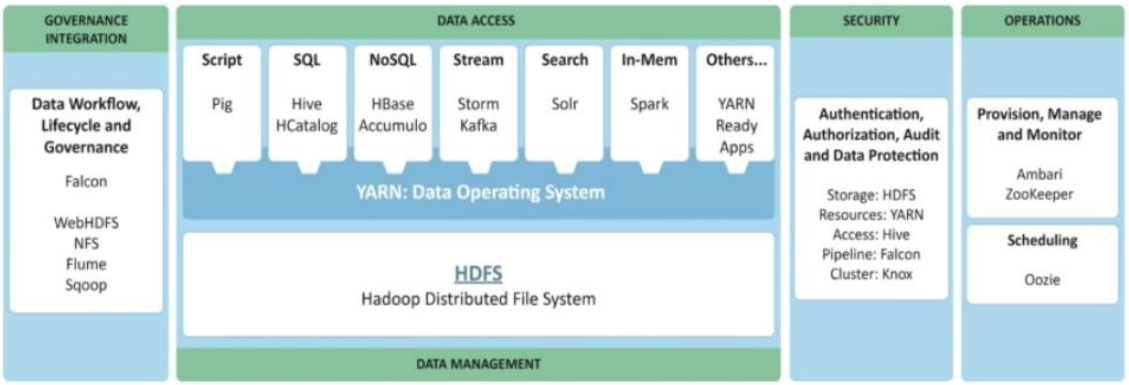
Hortonworks Data Platform (HDP) é uma suíte de funcionalidades essenciais para implementação do Hadoop, que pode ser usado para qualquer plataforma tecnológica de dados.

O principal objetivo da Hortonworks é conduzir todas as suas inovações através da plataforma de dados abertos Hadoop e construir um ecossistema de parceiros que acelere o processo de adoção do Hadoop entre empresas.

Apache Ambari é um exemplo de console de gerenciamento cluster do Hadoop desenvolvido pelo fornecedor Hortonworks para a gestão e monitoramento de clusters Hadoop.

A Horteworks Hadoop tem atraído mais de 60 novos clientes a cada trimestre com algumas contas gigantes como Samsung, Spotify, Bloomberg e eBay.

A Hortemworks tem atraído fortes parceiras de engenharia como RedHat, Microsoft, SAP e Teradata.



Download available on http://hortonworks.com/hdp/downloads/

Site: https://hortonworks.com/

5.5 MapR

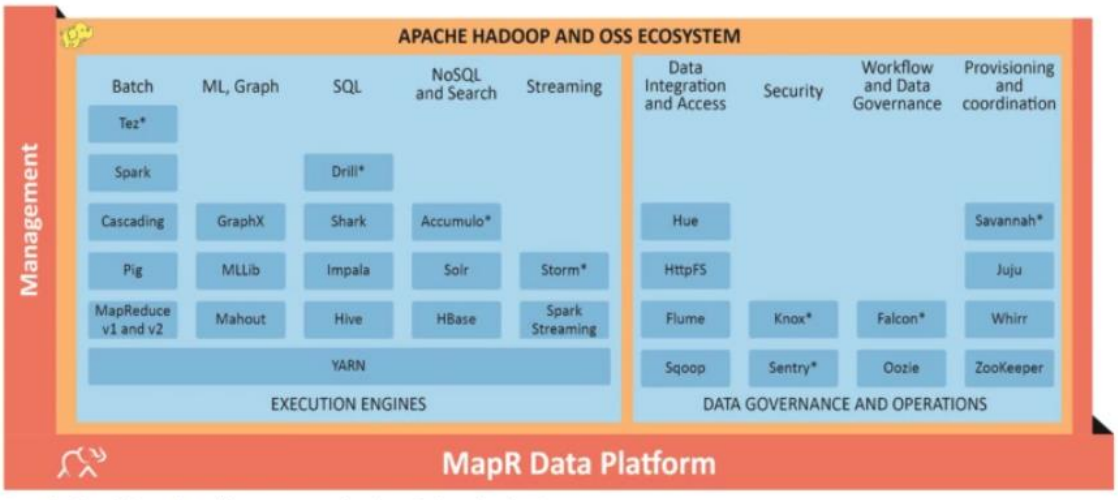
MapR Data Platform suporta mais de 20 projetos open-source.

MapR foi reconhecida amplamente por suas distribuições avançadas em Hadoop, no relatório de Gartner “Super Vendedores e, Infraestrutura da Informação e Bug-Data, 2012”

MapR foi projetado tendo em mente as operações de TI em Data Centers.

O MapR permite a utilização de aplicações baseadas em Hadoop e Spark, para atender às necessidades críticas de negócios, que operam 24x7.

O MapR suporta amplamente processamento de dados em bacth ou streaming de dados em tempos real.



Download available on: https://www.mapr.com/products/hadoop-download

Site: https://mapr.com/

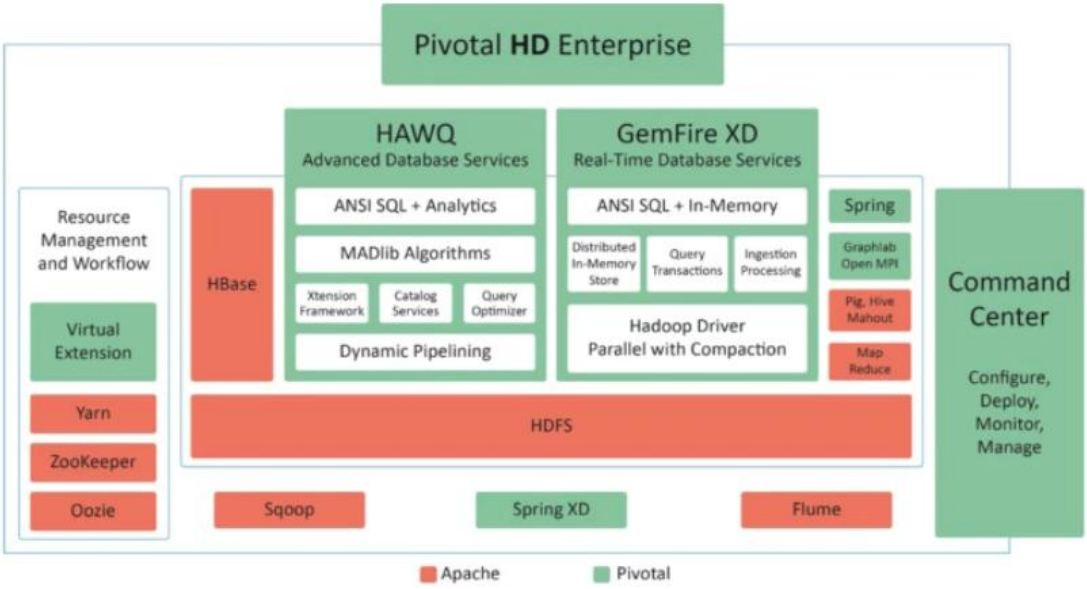
5.6 Pivotal HD

Pivotal HD é uma distribuição comercial do Hadoop. Ele consiste em um conjunto de ferramentas que visam acelerar projetos de análise de dados e expandir as funcionalidades do Hadoop.

Possui capacidade de análise em tempo real e decisões de processos de negócio podem ser tomadas quase que imediatamente a análise de dados.

Pivotal Big Data suite fornece um motor SQL nativo para o Hadoop.

Possui ainda suporte para processamento de Big Data em memória, o que acelera o processamento de dados.



Site: https://pivotal.io/

5.7 Microsoft Azure HDInsight

Azure HDInsight é uma distribuição Apache Hadoop distribuída em Cloud.

O Azure HDInsight consegue lidar com quantidades de dados de terabytes até petabytes, permitindo a inclusão de nodes sob demanda.

Por ser 100% Apache Hadoop, HDInsight pode processar dados semiestruturados ou não-estruturados, tais como cliques em páginas web, posts em mídia social, logs de servidores, dados de sensores, etc…

O HDInsight também possui extensões para programação em C#, Java e .NET, que podem ser usadas para criar, configurar, submeter e monitorar jobs Hadoop.

Por ser integrado com Excel, o HDInsight permite visualizar e analisar dados do Hadoop de forma que seja familiar aos usuários finais.

Site: https://azure.microsoft.com/pt-pt/

5.8 Quiz

1) Se o Hadoop é livre, porque eu usaria soluções comerciais do software?
    a. as principais soluções comerciais do Hadoop oferecem suporte, guias, assistência e melhores práticas
    b. sempre que o um bug é detectado, as soluções comerciais prontamente atualizam o software.
    c. as soluções oferecem pacotes completos, com tudo que é necessário para uma infraestrutura de BigData
    d. Todas estão corretas.

Resposta: D

2) Amazon Web Services EMR está entre uma das distribuições comerciais do Hadoop com a maior participação no mercado global.
    a. Verdadeiro
    b. Falso

Resposta: A

3) Fazem parte da solução comercial Cloudera?
    a. Hadoop, Java, Hive, Hbase e Sqoop
    b. Hadoop, Java, HTML, Hbase e Sqoop
    c. Hadoop, C++, Hive, xbase e Sqoop
    d. Hadoop, Pig, Hive, Hbase e Sqoop

Resposta: D

4) O principal objetivo da Hortonworks é conduzir todas as suas inovações através da plataforma de dados abertos Hadoop e construir um ecossistema de parceiros que diminui o processo de adoção Hadoop entre empresas.
    a. Verdadeiro
    b. Falso

Resposta: B. O principal objetivo da Hortonworks é conduzir todas as suas inovações através da plataforma de dados abertos Hadoop e construir um ecossistema de parceiros que acelera o processo de adoção Hadoop entre empresas.

5) Pivotal HD é uma distribuição comercial do Hadoop. Ele consiste em um conjunto de ferramentas que visam acelerar projetos de análise de dados e expandir as funcionalidades do Hadoop.
    a. Verdadeiro
    b. Falso

Resposta: A

6 – Introdução ao Apache Spark

6.1 O que é Apache Spark?

Apache Spark é um engine rápido e de uso geral para processamento de dados em larga escala.

É significantemente mais veloz que o Hadoop MapReduce e vem ganhando popularidade.

Utiliza o Hadoop (HDFS) como base, mas pode ser usado com Cassandra, HBase e MongoDB.

Pode ser usado com linguagens Python, R e Scala.

Usado por empresas como Clobo.com, Yelp, Washington Post, Yahoo e Twiter.

6.1.1 Algumas das suas características

- Velocidade – sua velocidade de execução pode ser até 100x mais rápido que o Hadoop MapReduce em memória e 10x em disco.
- Facilidade de uso – aplicações podem ser escritas em Java, Scala e Python.
- Generalidade – Cobina SQL Streaming e análise complexa, além do uso de ferramentas de alto nível como Sparl SQL, MLlib para Machine Learning, Graph X e Spark Streaming.
- Integração com Hadoop – Executa sobre o YARN cluster manager e permite leitura e escrita de dados no HDFS.

Spark é um projeto open-source, mantido por uma comunidade de desenvolvedores que foi criada em 2009 na Universidade da Califórnia, Berkeley.

Os desenvolvedores estavam trabalhando com Hadoop MapReduce e perceberam ineficiência na execução de computação iterativa.

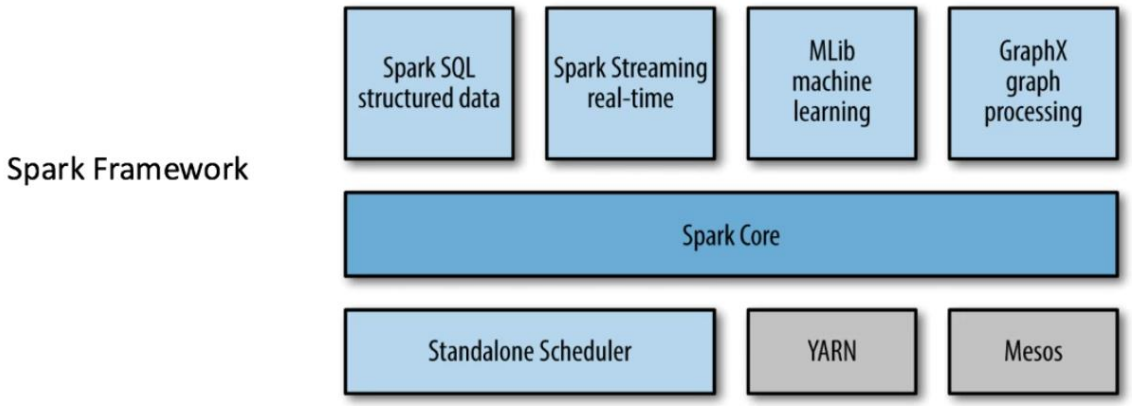
Em pouco tempo, Apache Spark tem se tornado o mecanismo de processamento de Big data para a próxima geração e está sendo aplicado em todo o mercado de dados mais rápido do que nunca.

6.1.2 Benefícios do Apache Spark

O Apache Spark oferece basicamente 3 principais benefícios:

1) Facilidade de uso – é possível desenvolver API's de alto nível em Java, Python e R, que permitem focar apenas no conteúdo a ser computado, sem se preocupar com configurações de baixo nível e extremamente técnicas.
2) Velocidade – Spark é veloz, permitindo uso iterativo e processamento rápido de algoritmos complexos. Velocidade é uma característica especialmente importante no processamento de grandes conjuntos de dados e pode fazer a diferença entre analisar os dados de forma interativa ou ficar aguardando vários minutos pelo fim de cada processamento. Com Spark, o processamento é feito em memória.
3) Uso geral – Spark permite a utilização de diferentes tipos de computação, como processamento de linguagem SQL (SQL Spark), processamento de texto, Machine Learning (MLlib) e processamento gráfico (Graph X). Estas características fazem do Spark uma exelente opção para projetos de Big Data.

6.2 Spark Framework



Spark Core - Contém as funcionalidades básicas do Spark, incluindo componentes para agendamento de tarefas, gestão de memória, recuperação de falha e sistemas de armazenamento.

Resilient Distributed Datasets(RDD's)

Spark SQL - Spark SQL é um pacote para tarefas com dados estruturados. Ele permite realizar queries nos dados através de linguagem SQL e HQL (Apache Hive Query Language – a variação do SQL desenvolvida pela Apache), além de suportar diversas fontes de dados com Hive e JSON.

Spark Streaming - Esse é um componente do framework Spark para processamento de streams de dados em tempo real.

MLlib - A biblioteca MLlib é uma funcionalidade para Machine Learning.

GraphX - O GraphX é uma biblioteca para manipulação de gráficos e computação em paralelo.

O resultado de um projeto de Big Data, pode ser a criação de um sistema de análise de dados em tempo real, que pode ser tornar o componente de uma aplicação de negócio.

6.3 Spark x Hadoop

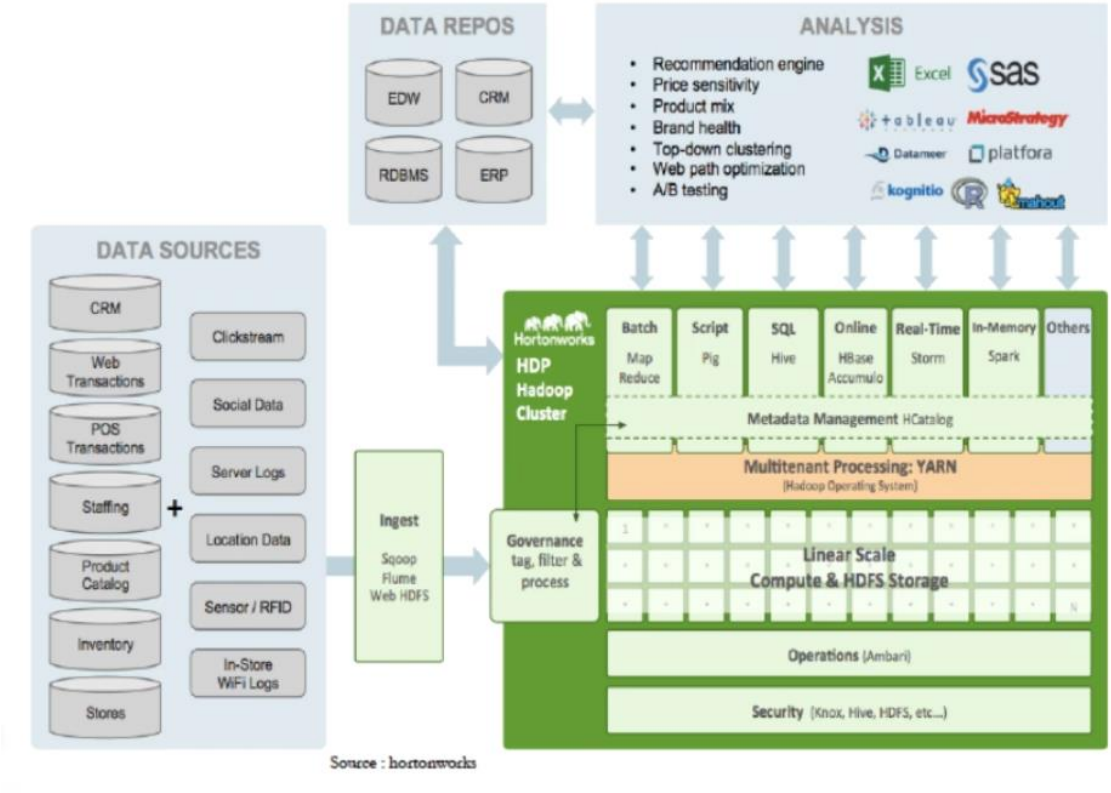
Quando se trata de Hadoop e Spark, duas perguntas são frequentes:

1) Já estou usando Hadoop, devo usar o Spark?
2) Estou pensando em usar Hadoop, devo desistir e usar Spark?

Vamos investigar as diferenças entre Hadoop e Spark e responder a estas perguntas!

O Hadoop é a plataforma original do Big Data, que tem sido usado e testado no mercado. Permite trabalhar com Petabytes de dados, habilitando a análise de quantidades massivas de dados.

O Hadoop possui um ecossistema bem definido que permite estender suas funções, como no caso da utilização do Pig, Hive e HBase.



A verdade é que criaram o Hadoop para processar grandes volumes de dados em batch. O Big Data.

Mas e se o volume de dados não for tão grande assim?

E se o volume de dados estiver em streaming, ou seja, fluxo contínuo de dados?

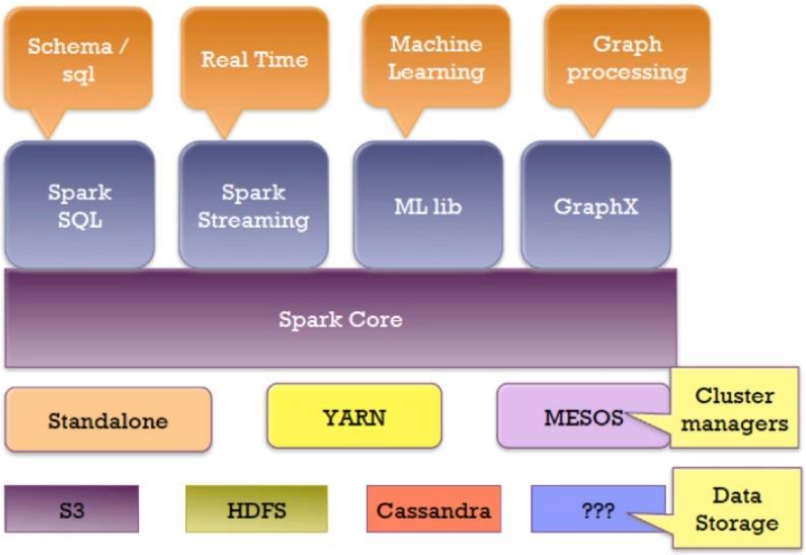
O Hadoop MapReduce possui limitações e não atende a alguns requisitos cada vez mais importantes:

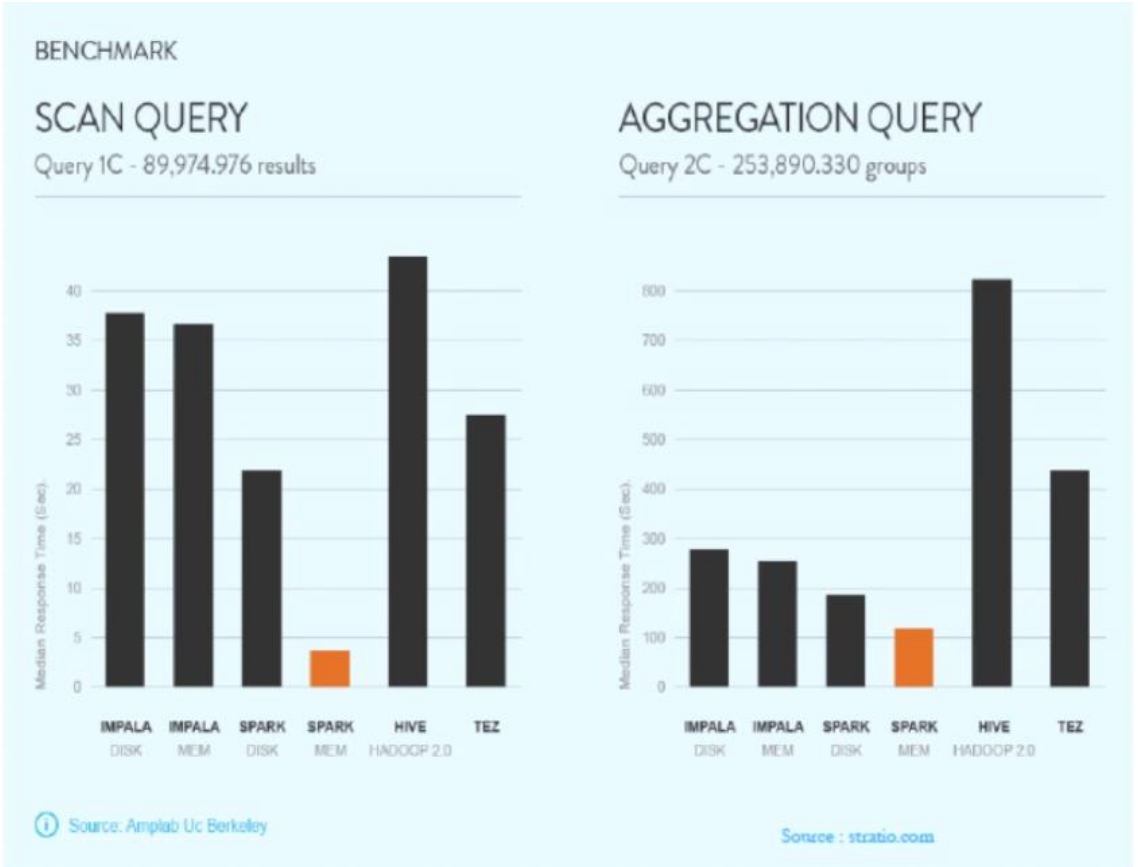
- Programação iterativa (Machine Learning, Algoritmos, etc…)
- E streaming de dados (possui alta latência)

O Spark é engine de computação em cluster.

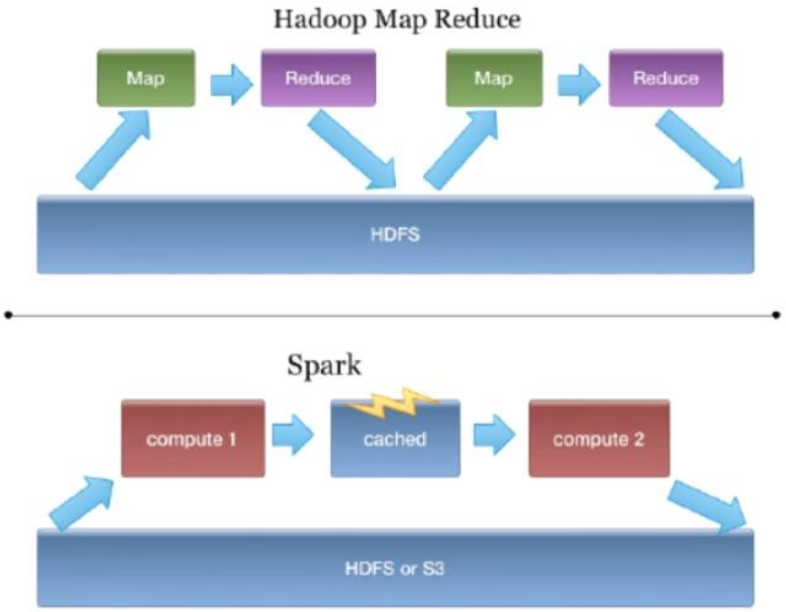
- Veloz – em memória os dados são processados até 100x mais rápido que no MapReduce.
- Propósito geral – SQL, Streaming, Machine Learning
- Compatibilidade – Hadoop, Mesos, Yarn, Standalone, HDFS, S3, Cassandra, HBase
- Mais fácil e simples

É a primeira plataforma de Big Data a integrar batch, streaming e computação interativa em único framework.





Hadoop	Spark
Armazenamento distribuído + Computação distribuída	Somente computação distribuída
Framework MapReduce	Computação genérica
Normalmente processa dados em disco (HDFS)	Em disco / **Em memória**
Não é ideal para trabalho iterativo	Excelente para trabalhos iterativos (Machine Learning)
Processo batch	Até 10x mais rápido para dados em disco Até 100x mais rápido para dados em memória
Basicamente Java	Suporta Java, Python, Scala
Não possui um shell unificado	Shell para exploração ad-hoc

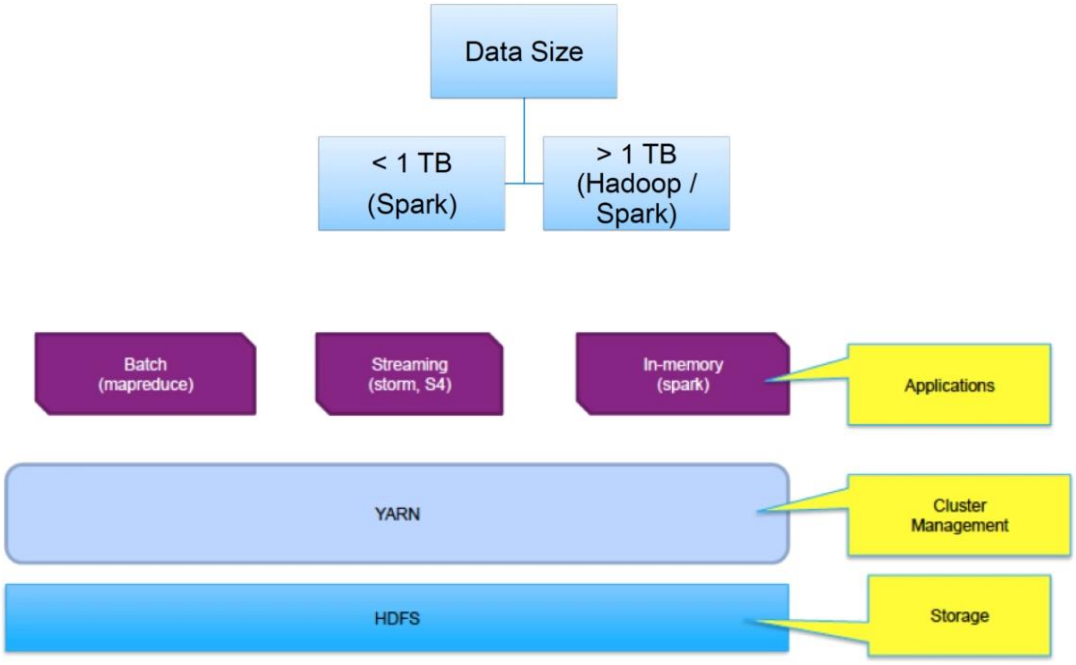


Então o Spark vai substituir o Hadoop? Não

- Spark executa sobre o HDFS / YARN
- Pode acessar o HDFS
- Usa YARN para gerenciamento do cluster
- Spark é realmente bom quando os dados podem ser processados em memória. Mas e quando não podem (por exemplo, gigantescos volumes de dados)?

	Hadoop	Spark
Processamento batch	Hadoop MapReduce (Java, Pig, Hive)	Spark RDD (Java, Python, Scala)
Query SQL	Hadoop: Hive	Spark SQL
Processamento Stream / Processamento em Tempo Real	Storm, Kafka	Spark Streaming
Machine Learning	Mahout	Spark ML Lib
Algoritmos iterativos	Lento	Muito rápido (em memória)
Workflow ETL	Pig, Flume	Pig com Spark ou Mix de Spark SQL e programação RDD
Volume de Dados	Volume gigante (Petabytes)	Volume médio (Gigabytes / Terabytes)

Rocomendação para o uso das duas tecnologias:



6.4 Apache Storm

6.4.1 Apresentação

O Apache Storm se tornou o padrão para processamento em tempo real distribuído e permite processar grandes quantidades de dados.

O Apache Storm foi desenvolvido em Java.

Foi criado para processar grandes quantidades de dados em ambientes tolerantes a falhas e escaláveis.

Basicamente, o Storm é um framework para dados streaming (fluxo contínuo de dados) e possui uma lata taxa de ingestão de dados.

A gestão do estado do cluster, é feita através do Zookeper.

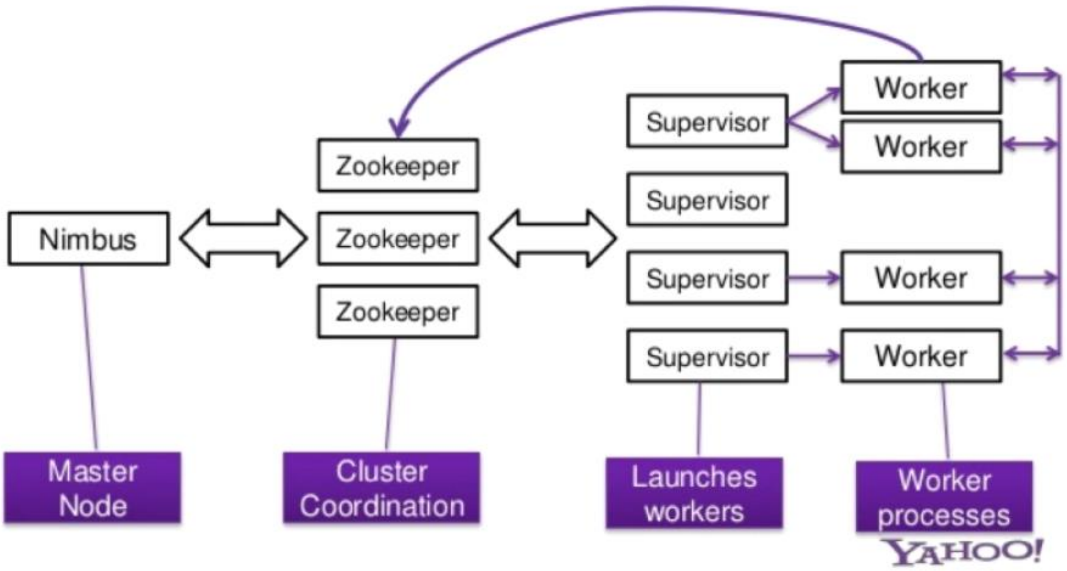
O Storm é simples e você pode executar todos os tipos de manipulação de dados em tempo real, em paralelo.

O Apache Storm é um dos líderes em Real-Time Analytics.

6.4.2 Principais benefícios de se utilizar o Storm:

- Storm é open-source, roubusto e amigável (fácil utilização)
- Tolerante a falhas, flexível, confiável e suporta diversas linguagens de programação.
- Processa dados em tempo-real
- Storm é incrivelmente veloz

6.4.3 Arquitetura Storm



Master Node – No Master Node encontramos o serviço Nimbus, que é responsável pela atribuição de tarefas aos Supervisors.

Coordenação do Cluster – O Zookeeper faz a coordenação do funcionamento do cluster.

Supervisor – Os supervisors são responsáveis por 1 ou mais workers e sua função e garantir que os workers executem os jobs.

Worker Node – Os workers nodes, executam as tarefas (jobs).

Esta arquitetura garante uma das principais características do Storm: **No single-point** de falha.

6.4.4 Hadoop vs Storm

O Storm realiza todas as operações, executa persistência, enquanto o Hadoop é bom em tudo, exceto computação de dados em tempo real.

Hadoop	Storm
Processamento em batch	Processamento de streams em tempo real
Arquitetura Master/Slave com ou sem o Zookeeper	Arquitetura Master/Slave com o Zookeeper
O HDFS utiliza o MapReduce para processar grandes quantidades de dados em minutos ou horas	Processa streams de dados e milhares de mensagens podem ser processadas por segundo em um cluster

6.4.5 Spark vs Storm

A diferença principal entre Spark e Storm, é que o Spark realiza computação paralela de dados, enquanto o Storm realiza computação paralela de tarefas. No mais, são bem parecidos e tem como objetivo o processamento de streaming de dados.

Spark	Storm
Linguagem de programação Java, Scala	Linguagem de programação Java, Clojure, Scala
Fonte de streams no HDFS	Fonte de streams no Spout
Gestão de Recursos com YARN, Mesos	Gestão de Recursos com YARN, Mesos

Qual framework utilizar afinal?

Situação	Framework
Baixa Latência	Storm consegue obter melhor latência que o Spark
Baixo custo de desenvolvimento	Com Spark, o mesmo código pode ser usado para processamento em batch e processamento de streams. No Storm, isso não é possível
Tolerância a falhas	Ambos são tolerantes a falhas

6.5 Quiz

1) O projeto Spark contém diversos componentes integrados. Basicamente, Spark é um engine de computação, responsável por agendar, distribuir e monitorar aplicações de diversas tarefas de processamento através de diferentes servidores em cluster.
    a. Verdadeiro
    b. Falso

Resposta: A

2) Contém as funcionalidades básicas do Spark, incluindo componentes para agendamento de tarefas, gestão de memória, recuperação de falha e sistemas de armazenamento.
    a. Spark SQL
    b. Spark Core
    c. Spark Streaming
    d. Todas estão corretas

Resposta: C

3) O Apache Spark é compatível com?
    a. SQL, Java e C++, cobol
    b. Memoría distribuida, Mapreduce e NoSql
    c. Hadoop, Mesos, Yarn, Standalone, HDFS, Cassandra e HBase
    d. Todas estão erradas

Resposta: C

4) Sobre o Apache Spark é incorreto afirmar:
    a. O Hadoop utiliza Framework Mapreduce e o Spark computação genérica.
    b. O Hadoop realiza armazenamento distribuído e computação distribuída e o spark utiliza somente computação distribuída.
    c. O Hadoop não é ideal para trabalho iterativo enquanto o Spark é excelente para trabalhos iterativos.
    d. O Hadoop processa dados em memória e o Spark somente em disco.

Resposta: D

7 – Bancos de Dados NoSQL

7.1 O que são Bancos de Dados NoSQL?

Bancos de dados tradicionais RDBMS (Relational Database Management Systems) não foram projetados para tratar grandes quantidades de dados (Big Data).

Bancos de dados tradicionais foram projetados somente para tratar conjuntos de dados que possam ser armazenados em linhas e colunas e portanto, possam ser consultados através do uso de queries utilizando linguagem SQL (Structured Query Language).

Bancos de Dados relacionais não são capazes de tratar dados não-estruturados ou semiestruturados.

Ou seja, Banco de Dados relacionais simplesmente não possuem funcionalidades necessárias para atender os requisitos do Big Data, dados gerados em grandes volumes e alta velocidade.

Esta lacuna está sendo preenchida por Bancos de Dados NoSQL.

Bancos de dados NoSQL, são bancos de dados distribuídos e não-relacionais, que foram projetados para atender os requerimentos de Big Data.

Bancos de Dados NoSQL oferecem uma arquitetura muito mais escalável e eficiente que os bancos relacionais e facilitam as consultas no-sql de dados semiestruturados ou não-estruturados.

Existe alguma discussão sobre o significado de NoSQL.

Alguns afirmam que a sigla significa *Not Only SQL*, enquanto outros afirmam que significa *Non-SQL*. Não há um consenso sobre isso. Mas pense sobre NoSQL como uma classe de banco de dados não-relacionais que não se enquadram na classificação de bancos de dados relacionais (RDBMS), que utilizam linguagem SQL.

Embora o modelo relacional e a Structured Query Language (SQL) foram por décadas para armazenamento de dados, é fato que os bancos de dados relacionais não são mais os vencedores quando se trata de flexibilidade e escalabilidade.

Isto tornou-se verdadeiro especialmente com o advento das redes sociais online e Internet das Coisas.

A este respeito, NoSQL surgiu como um paradigma não-tradicional para lidar com grandes volumes de dados e para resolver os desafios colocados pela chegada de implementações de Big Data.

Atualmente, bancos de dados NoSQl como MongoDB, Cassandra e CouchDB introduzem novas características e funcionalidades, trazendo ainda mais inovação e resultados surpreendentes.

Bancos de Dados NoSQL oferecem 4 categorias principais de bancos de dados:

- Graph databases
- Document databases
- Key-values stores
- Column family stores

7.1.1 Graph Databases

Essa categoria de Bancos de dados NoSQL, geralmente são aderentes a cenários de rede social on-line, onde os nós representam as entidades e os laços representam as interconexões entre eles.

Desta forma, é possível atravessar o gráfico seguindo as relações. Esta categorias têm sido usada para lidar com problemas relacionados a sistemas de recomendações e listas de controle de acesso, fazendo uso da sua capacidade de lidar com dados altamente interligados.

7.1.2 Document Databases

Esta categoria de Bancos de Dados NoSQL permite o armazenamento de milhões de documentos.

Por exemplo, você pode armazenar detalhes sobre um emprego, junto com o currículo dele (como um documento) e então pesquisar sobre potenciais candidatos a uma vaga, usando um campo específico, como telefone ou conhecimento em uma tecnologia.

7.1.3 Key-Value Store

Nesta categoria, os dados são armazenados no formato key-value (chave-valor) e os valores (dados) são identificados pelas chaves.

É possível armazenar bilhões de registos de forma eficiente e o processo de escrita é bem rápido. Os dados podem ser então pesquisados através das chaves associadas.

7.1.4 Column Family Store

Também chamados de banco de dados orientados a coluna, os dados são organizados em grupos de colunas e tanto o armazenamento, quanto as pesquisas de dados são baseadas em chaves.

HBase e Hypertable são os exemplos mais comuns desta categoria.

7.1.5 Principais Bancos de Dados NoSQL

- Graph
    - Neo4J
    - FlockDB
    - GraphDB
    - ArangoDB

- Key-value
    - Oracle NoSQL DB
    - MencacheDB
    - Redis
    - Voldemort

- Doucument
    - MongoDB
    - CounchDB
    - RavenDB
    - Terrastore

- Column
    - HBase
    - Cassandra *
    - Hypertable
    - Accumulo

*Cassandra é hibrido, Column e Key-value

Para uma lista completa de Bancos de Dados NoSQL visite: http://nosql-database.org/

Como NoSQL oferece funcionalidades nativas para cada uma destas categorias, ele se torna uma alternativa eficiente para armazenamento e consulta para a maioria dos dados não-relacionais.

Esta adaptabilidade e eficiência, tem transformado os bancos de dados NoSQL em uma excelente solução para tratar Big Data e superar os problemas relacionados ao processamento de grandes volumes de dados.

E por que usar banco de dados NoSQL?

- Representação de dados sem esquema.
- Tempo de desenvolvimento menor
- Velocidade
- Escalabilidade

7.2 MongoDB

https://www.mongodb.com/

MongoDB é um banco de dados orientado a documento, uma das categorias de dados NoSQL.

Um banco de dados NoSQL orientado a documento, substitui o conceito de "linha" como em bancos de dados relacionais, por um modelo mais flexível, o "documento".

O MongoDB é open-source e um dos líderes no segmento de banco de dados NoSQL. Ele foi desenvolvido em linguagem C++.

Algumas das principais características do MongoDB:

- Indexação – suporta índices secundários, permitindo construção de queries mais velozes.
- Agregação – permite a construção de agregações complexas de dados, otimizando o desempenho.
- Tipos de dados especiais – suporta coleções **time-to-live** para dados que expiram em um determinado tempo, como sessões por exemplo.
- Armazenamento – suporta o armazenamento de grandes quantidades de dados.

Algumas características presentes em bancos de dados relacionais, não estão presentes no MongoDB, como alguns tipos de join e transações multi-linha.

Comparação entre MongoDB e Base de dados Relacionais

MongoDB	RDBMS
Database	Database
Collection	Table
Document	Tuple/Row
Field	Column
Embedded Documents	Table Join
Primary Key	Primary Key

Onde usar o MongoDB?

- Big Data
- Gestão de Conteúdo
- Infraestrutura Social e Mobile
- Gestão de Dados de Usuários
- Data Hub

7.3 Apache Cassandra

http://cassandra.apache.org/

Apache Cassandra é um banco de dados NoSQL, livremente distribuído, de lata performance, extremamente escalável e tolerante a falha.

Ele foi concebido com a premissa que falhas de sistema ou de hardware sempre ocorrem.

Foi inicialmente desenvolvido pelo Facebook, como um combinação do BigTable (Google) and Dynamo Data Store (Amazon).

O Cassandra é usado para armazenar gigantescas quantidades de dados (Big Data), de forma rápida.

O Cassandra também funciona bem quando se faz necessário a pesquisa de dados de forma indexada.

É voltado para trabalhar em clusters, sendo totalmente escalável. Novos node podem ser adicionados, à medida que os dados crescem.

É ainda uma excelente solução quando se necessita de alta performance para leitura e escrita.

Algumas empresas / websites que usam o Cassandra: eBay, GitHub, GoDaddy, Instagram, Netflix, Reddit, CERN, Comcast, entre outros.

7.4 CouchDB

http://couchdb.apache.org/

CounchDB é um banco de dados totalmente voltado para a web.

No CounchDB os dados são armazenados em documentos JSON (Java Script Object Notation), que consiste em campos que podem ser strings, números, datas, listas ordenadas e mapas associativos.

O CouchDb suporta aplicativos web e mobile.

O CounchDB é distribuído em pares com um server m cliente, que podem ter cópias independentes do mesmo banco de dados.

Apache CouchDB foi o banco de dados que deu o pontapé inicial do movimento NoSQL.

Ele foi construído a partir do zero com alto desempenho e tolerância a falhas em mente.

ScouchDB permite aos usuários armazenar, reproduzir, sincronizar e processar grandes quantidades de dados (Big Data), distribuídos em dispositivos móveis, servidores, Data Centers e regiões geográficas distintas em qualquer configuração de implementação, incluído ambiente em nuvem (Cloud).

7.5 Quiz

1. Bancos de Dados tradicionais foram projetados somente para tratar conjuntos de dados que possam ser armazenados em linhas e colunas e portanto, possam ser consultados através do uso de queries utilizando linguagem SQL (Structured Query Language).
    a. Verdadeiro
    b. Falso

Resposta: A

2. Bancos de Dados relacionais não possuem funcionalidades necessárias para atender os requisitos do Big Data, dados gerados em grande volume e alta velocidade.
    a. Verdadeiro
    b. Falso

Resposta: A

3. A respeito de Banco de Dados NoSql marque a opção incorreta?
    a. Bancos de Dados NoSQL, são bancos de dados distribuídos e não-relacionais, que foram projetados para atender os requerimentos de Big Data.
    b. Bancos de Dados NoSQL oferecem uma arquitetura mais escalável e eficiente que os bancos relacionais.
    c. Bancos de Dados NoSQL facilitam consultas no-sql de dados semi-estruturados ou não-estruturados.
    d. O modelo relacional e a Structured Query Language (SQL) foram por décadas o padrão para armazenamento de dados e agora são os mais adequados para aplicação em Big Data.

Resposta: D

4. Quais as quatro categorias principais dos Bancos de Dados NoSQL?
    a) Graph databases, Document Oraclesystem, Key-values, Column family stores.
    b) Graph Xbase, Document databases, Key-values stores, Column stores database.
    c) Graph databases, Document databases, Key-values stores, Column family stores.
    d) Graph Xbase, Document Oraclesystem, Key-values stores, Column family stores.

Resposta: C

5. Qual tipo de indexação é suportada pelo MongoDB?
    a. Índices secundários
    b. Índices primários
    c. Índices externos.
    d. Índices internos.

Resposta: A

8 – Como as Empresas estão a utilizar o Big Data

8.1 Big Data no Ambiente Corporativo

Caesars Entertainment – A companhia de entretenimento em cassinos está usando o ambiente Hadoop para identificar diferentes segmentos de consumidor e criar campanhas de marketing específicas para cada um deles.

O novo ambiente reduziu o tempo de processamento de 6 horas para 45 minutos para posições-chave. Isso permitiu à Caesars promover uma análise de dados mais rápido e exata, aprimorando a experiência de consumidor e fazendo com que a segurança atendesse os requisitos do setor de pagamentos com cartões.

A empresa agora processa mais de 3 milhões de registos por hora.

http://caesarscorporate.com/

Cerner – A empesa de tecnologia para setor de saúde contruiu um hub de dados corporativos no CDH (Cloudera Distribuution), para criar uma visão mais compreensível de qualquer paciente, condição ou tendência.

A tecnologia ajuda a Cerner e seus clientes a monitorarem mais de 1 milhão de pacientes diariamente.

Entre outras coisas, ela colabora na determinação mais exata da probabilidade de um paciente estar com infeção em sua corrente sanguínea.

https://www.cerner.com/

eHarmony – O site namoro online recentemente atualizou seu ambiente na nuvem, usando o CDH para analisar um volume massivo e variado de dados.

A tecnologia ajuda a eHarmony a disponibilizar novas combinações a milhões de pessoas diariamente.

O novo ambiente cloud acomoda análises mais complexas, criando resultados mais personalizados e aumentando a chance de sucesso nos relacionamentos.

http://www.eharmony.com/

MasterCard – A empresa foi a primeira a implementar a distribuição CDH do Hadoop após receber certificação PCI completa.

A companhia usou os servidores Intel para integrar conjuntos de dados a outros ambientes já certificados.

A MasterCard incentiva seus clientes a adotarem o sistema através do seu braço de serviços profissionais, o MasterCard Advisors.

https://www.mastercard.us/en-us.html

FarmLogs – A companhia de software para gerenciamento de produção agrícolas usa analytics em tempo real rodando nos processadores Intel Xeon E5 para fornecer dados sobre colheitas, condições de plantio e estado da vegetação para 20% das fazendas americanas.

A tecnologia ajuda os fazendeiros a aumentarem a produtividade de seus acres. ~

https://farmlogs.com/

Nippon Paint – Uma das maiores fornecedoras de tinta da Ásia usa os processadores Intel Xeon E7 v2 (rodando no software SAP HANA de analytics in-memory) para compreender o comportamento de clientes, otimizar a sua cadeia de suprimentos e melhorar a suas campanhas de marketing.

A Nippon Paint agora testa um novo sistema baseado no Hadoop para usufruir das ferramentas de alto desempenho e processar Big Data.

http://www.nipponpaint.com/

Outras empresas usando Hadoop:

Empresa	Especificações Técnicas	Utilização
Facebook	Mais de 12 TB de storage	Hadoop é utilizado em soluções de relatórios e Machine Learning
Twitter	--	Hadoop é usado desde 2010 para o processamento de logs e tweets
LinkedIn	4100 nodes Hadoop	Todos os dados do LinkedIn passam através de um cluster Hadoop
Yahoo!	4500 nodes Hadoop e mais de 1 TB de storage	Usado no portal do Yahoo
Ebay	4000 nodes Hadoop	Um dos maiores clusters Hadoop que se tem notícia, usado para processar as mais de 300 milhões de pesquisas feitas pelos usuários

Data Science Academy

Empresa	Especificações Técnicas	Utilização
Accenture	De acordo com a demanda do cliente	Projetos de Big Data na área financeira, telecom e varejo
Ning	--	Plataforma de Rede Social, utiliza o Hadoop para relatórios e Big Data Analytics
Spotify	690 nodes em cluster Hadoop, totalizando 38 TB de memória RAM e 28 PB de storage	Usa Hadoop para geração de conteúdo e agregação de dados
Yahoo!	4500 nodes Hadoop e mais de 1 TB de storage	Usado no portal do Yahoo
Fox	70 nodes Hadoop	Usado para análise de logs e Machine Learning

Para cada 100 vagas com exigência de conhecimentos em Big Data, existem apenas 2 profissionais qualificados!

Até 2018, serão criadas mais de 200 mil vagas em Big Data e mais da metade ficará sem ser preenchida, por falta de profissionais qualificados!

“Em uma pesquisa com 3000 empresas globais, mais de 83% dos pesquisados identificaram análise de negócios a partir de Big Data como uma prioridade” – IBM

Não é nenhum segredo que o Hadoop e o Apache Spark são as tecnologias mais quentes no mercado de Big Data, mas o que é menos frequentemente notado é que ambos são open-source.

Os clientes apreciam o open source por permitir “experimentar antes de comprar”, mas também já começam a ver o mundo open source evoluindo mais rapidamente do que o mundo proprietário por causa do compartilhamento entre os desenvolvedores.

Todo o ecossistema Hadoop está se movendo mais rápido do que aconteceria caso dependesse de um único fornecedor.

Por tudo isso, organizações como a Forrester acreditam que o Hadoop é uma plataforma que precisa de ser usada em grandes empresas, formando a pedra angular de qualquer futura plataforma flexível de gestão de dados.

Se sua empresa tem dados estruturados, semiestruturados ou não estruturados, há espaço relevante para o Hadoop.

E há duas grandes razões para isso: as empresas têm muito mais dados para gerir e o Hadoop é uma grande plataforma, especialmente por permitir combinar dados antigos legados com novos dados não estruturados.

Quando um novo produto é lançado, a empresa pode usar dados de uma variedade de fontes para determinar a demanda, avaliar os preços dos concorrentes e desenvolver a sua própria estrutura de preços para maximizar vendas e lucros.

Por exemplo, utilizando os dados recolhidos a partir de medias sociais, histórico de navegadores, fóruns e informações demográficas, a empresa pode determinar se o próximo brinquedo será um sucesso de venda ou se ficará pegando poeira nas prateleiras.

Dados de Geolocalização em aplicativos móveis são uma maneira poderosa e eficaz para maximizar o potencial de vendas da empresa.

Fazer a venda online de um tablet é algo normal em diversos sites de produtos eletrônicos. Mas e se além do tablet, o site oferecer (através de um sistema de recomendação baseado nos cliques de outros clientes), produtos associados, tais como teclado ou mouse sem fio, carregador para carro ou até mesmo um protetor a prova d’água. O que seria uma venda simples, pode se transformar em uma venda muito maior.

Estas recomendações não são vistas apenas como tentativas de vender mais. São vistas como um serviço valioso ao cliente, que será lembrado em adquirir outros acessórios necessários para o tablet.

Muitas startups acreditam que, utilizando Big Data, derrubarão líderes de mercado como Cisco, Google ou Apple. Elas acreditam que conectarão suas ferramentas de análise de dados a bolas de cristal e descobrirão segredos que magicamente a catapultarão a posição de grandes vencedores.

Pois saiba que isso raramente acontece, se é que acontece. Histórias de sucesso envolvendo Big Data tipicamente começam com pequenas perguntas:

- Qual é o melhor quarteirão para instalar uma nova loja?
- Como podemos tornar a escolha das localizações em um processo sistemático?
- O que fazer para o time de vendas convencer os clientes em ligações telefônicas?
- Como mudar a abordagem nas ofertas de varejo, em tempo real, para alinhá-las as preferências dos consumidores?

8.2 Netflix

Está presente em mais de 40 países somada a 50 milhões de assinantes e altos números de audiência. Uma conta simples, mas que coloca a **Netflix** como o carro-chefe dentre os concorrentes que oferecem serviços de TV por internet – disponibilizando séries de sucesso, filmes e novelas.

O completar seus 18 anos de existência, a empresa, fundada no estado da Califórnia, foi avaliada em mais de US$25 bilhões, segundo a Forbes e continua conquistando novos mercados.

Qual seria a "fórmula mágica" da Netflix?

"Não é magia, é tecnologia". Big Data é a tal fórmula para chegar a resultados tão certeiros. Desde o momento em streaming tornou-se a forma primária de levar conteúdo aos assinantes, foi necessário mensurar dados como "os dias em que filmes são assistidos", "tempo gasto na escolha de filmes" e "quão frequente o playback era interrompido" – tanto pelo usuário como por limitações da rede.

Neste contexto, os colaboradores da Netflix são motivados a descobriram novas informações diariamente. Dados são utilizados, inclusive, em títulos, cores, capas, ou seja, em todos os aspetos do negócio.

A visualização de dados é de suma importância para a empresa. Disso não há dúvidas, já que algoritmos, insights e a resolução de questões do próprio negócio são todas abordagens construídas no dia a dia.

Entretanto, apesar do forte uso de Big Data pela Netflix, a real motivação está na predição do que os consumidores irão gostar de assistir, o que, irá entretê-los. É por isso que os sistemas de recomendação existentes na interface também dependem de Big Data Analytics.

8.3 AirBnB

O Airbdnb precisou de um bom tempo para contruir bases sólidas e isso ocorreu quando descobriu que o seu principal obstáculo era prevenir-se que pessoas escolhessem ficar em hotéis no lugar de contratar seus serviços.

Rile Newman, líder de Analytics e Cientista de Dados da empresa, conduziu um processo de regressão para determinar as características mais impactantes no fechamento de uma reserva. Ele descobriu algo que hoje soa trivial: apartamentos cujas fotos não eram bonitas não eram alugados / reservados. Simples assim!

Com base na descoberta, o Airbnb passou aenviar fotógrafos profissionais a vários apartamentos para refazer imagens. Os resultados foram surpreendentes, com ganhos nos números de reservas e na confiança dos usuários / locatários.

8.4 Starbucks

Em tempos de internet, muitas empresas direcionam negócios fortemente para a estratégia puramente digital, ignorando que, o e-commerce ainda corresponde apenas 17% das vendas de varejo. Trocando em miúdos: grande parte do dinheiro ainda passa na frente da vitrine e entra em espaços físicos.

No passado, donos dessas empresas direcionariam o investimento a áreas que parecem ser uma boa aposta medindo o fluxo de tráfego, o número de pedestres por hora ou comparando os empreendimentos existentes na região.

A Starbucks confia em análise de dados para guiar o processo de abertura de cafeterias, indo tão longe quanto a construção de um plano de mercado e aplicações para desenvolvimento de lojas em um sistema chamado Atlas.

A melhor maneira de explicar o Atlas é como uma ferramenta de análise de grandes volumes de dados que possui, acima, uma camada de software de mapas e informações geográficas. Com ele, a rede de cafeteria consegue avaliar um volume elevado de variáveis que podem contribuir com os sucessos das lojas de variáveis que podem contribuir com o sucesso das lojas, visualizando-as nos mapas e procurando pontos similares em outras localidades.

Aprender com dados e mapas não garante o sucesso dos esforços, mas o processo assegura redução drástica dos riscos associados ao lançamento de uma nova loja.

8.5 Atenção ao Usar Big Data

5 pontos de atenção que devem ser observados quando usado Big Data:

- Selecionar as fontes erradas.
- Não definir um objetivo
- Ignorar a qualidade dos dados
- Não categorizar os dados
- Não criar uma cultura orientada a dados.

8.6 Quiz

1. O setor financeiro está alavancando o uso do Big Data para transformar seus processos, suas organizações e em breve toda a indústria por meio de análise de clientes através da construção de novos produtos e serviços, gestão de risco nas carteiras de crédito, detecção e prevenção de fraudes, etc. Empresas financeiras usando Big Data tendem a gerar sólidos resultados comerciais, em particular no âmbito do cliente.
    a. Verdadeiro
    b. Falso

Resposta: A

2. Quais as vantagens que a Empresa Caesars adquiriu utilizando Big Data?
    a. Reduziu o tempo de processamento de operações chaves.
    b. Promoveu análise de dados mais rápida e exata.
    c. Aprimorou a experiência do consumidor.
    d. Todas estão corretas.

Resposta: D

3. Que empresa utilizou Big Data para criar resultados personalizados e aumentar a chance de sucesso entre relacionamentos amorosos?
    a. Alma entrelaçada
    b. Conheço você
    c. eHarmony
    d. Nenhuma acima

Resposta: C

4. Quais dessas empresas utilizam dados do Big Data:
    a. Facebook
    b. Twitter
    c. Linkedin
    d. Yahoo
    e. Todas acima

Resposta: E

5. Sobre Hadoop e Spark é incorreto afirmar?
    a. São tecnologias proprietárias.
    b. Ambas são open-source.
    c. São tecnologias que estão em voga no mercado de Big Data.
    d. Permitem compartilhar conhecimento entre desenvolvedores.

Resposta: A

UFA! Chegamos ao final do curso!