Ambiente Livre
Tecnologia

Hadoop

Treinamento Hadoop – Big Data Open Source - Fundamental.

Instrutor: Marcio Junior Vieira.
marcio@ambientelivre.com.br

O que é Hadoop

- O Apache Hadoop é um projeto de software open-source escrito em Java. Escalável, confiável e com processamento distribuído.
- Filesystem Distribuído
- Inspirado Originalmente pelo GFS e MapReduce da Google (Modelo de programação MapReduce)
- Utiliza-se de Hardware Comum (Commodity cluster computing)
- Framework para computação distribuída
- infraestrutura confiável capaz de lidar com falhas (hardware, software, rede)

O Hadoop Inclui:

- **MapReduce** é um sistema de arquivos distribuído, um framework para grandes clusters.
- **Master/Slave**
- **JobTracker** organiza todas as tarefas e coordena o fluxo de dados entre os TaskTrackers
- **TaskTracker** manipula todos os worker no node
- **Worker Task** executa as operações de map ou reduce
- Integra-se com HDFS com os dados localmente

Características

- Grandes volumes de dados (Peta Bytes)
- Processamento dos dados
- Tolerância a falha
- Distribuição do serviço
- Escrito em Java, OpenSource
- Roda em hardware comum
- Linux, Mac OS/X, Windows e Solaris

Características

- Um sistema escalável e confiável para armazenamento compartilhado e análises.
- Ele automaticamente trata da replicação de dados e da falhas em cada nó.
- Ele faz o trabalho duro - desenvolvedor pode se concentrar em processamento da lógica de dados
- Permite que os aplicativos usem petabytes de dados em paralelo

Motivações Atuais

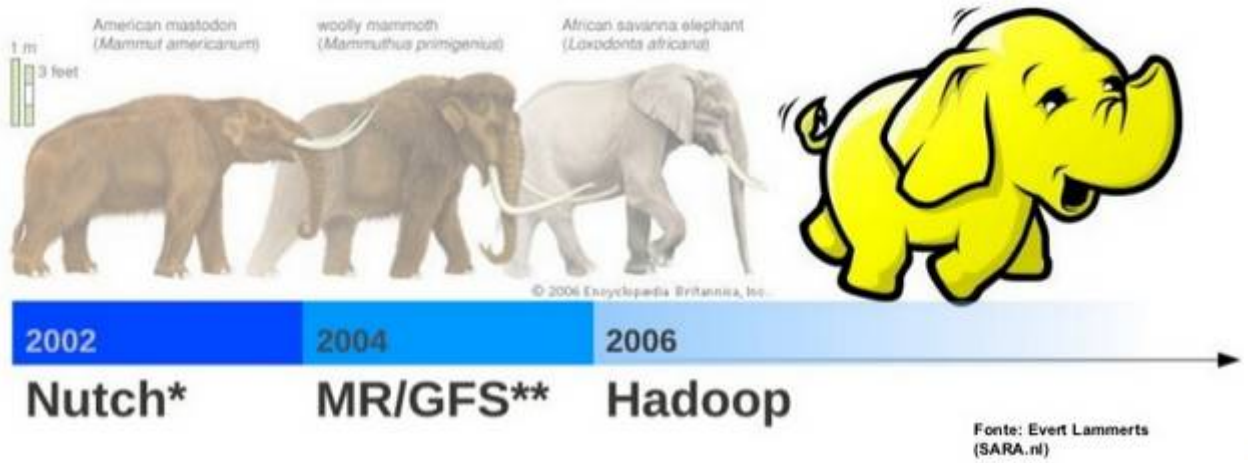
- Grande quantidade (massiva) de dados
- Dados não cabem em uma máquina
- Demoram muito para processar de forma serial
- Máquinas individuais falham
- Computação nas nuvens
- Escalabilidade de aplicações
- Computação sob demanda

Por que usar Hadoop

- **Baixo custo**
  - Uso de hardware comum
  - Compartilhar recursos com vários projetos
  - Fornecer escala quando necessário
- **infraestrutura confiável**
  - capaz de lidar com falhas - hardware, software, networking (A falha é esperada, e não uma exceção)
  - Transparente para as aplicações

Histórico

- Hadoop tem o seu início com o projeto Nutch.
- Era a construção de um search engine web que tinha problemas para ser executado em um grupo de computadores .
- Com o lançamento do Paper do Google eles começaram a reescrever o Nutch e surgiu o Hadoop.



Google Paper

- Paper publicado pela Google.

  MapReduce: Simplied Data Processing on Large Clusters

- http://research.google.com/archive/mapreduce.html
- PDF: http://research.google.com/archive/mapreduce-osdi04.pdf
- HTML: http://rese apreduce-osdi04-slides/ind

Histórico

- **2004** - Versão inicial do que é hoje Hadoop Distributed Filesystem e Map-Reduce implementado por Doug Cutting e Mike Cafarella.
- **Dez/2005** - Nutch portado para o novo framework. Hadoop é executado de forma confiável em 20 nós.
- **Jan/2006** - Doug é contratado pelo Yahoo!
- **Fev/2006** - Apache projeto começou oficialmente a apoiar o desenvolvimento standalone do MapReduce e HDFS.
- **Fev/2006** - adoção do Hadoop pela equipe Yahoo!
- **Abr/2006** - Sort Benchmark (10 GB / node) executado em 188 nós em 47,9 horas.

Histórico

- **Mai/2006** - Yahoo! criar um cluster Hadoop pesquisa 300 nós.
- **Mai/2006** - Sort Benchmark é executado em 500 nós em 42 horas.
- **Out/2006** – Criado Cluster de Pesquisa atinge 600 nós.
- **Dez/2006** -Sort Benchmark executado em 20 nós em 1,8 horas, 100 nós em 3,3 horas, 500 nós em 5,2 horas, 900 nós em 7,8 horas.
- **2008** - Cloudera é Fundada (www.cloudera.com)
- **2011** - 24 Engenheiros do Yahoo juntan-se e fundam a Hortonworks (http://hortonworks.com)
- **01/2014** – Mídias especializadas do Brasil anunciam cientista de dados como primeiro cargo mais cobiçado do ano entre 8.

A origem do Nome

- Segundo Doug Cutting, criador do Hadoop

  “O nome que meu filho deu a um elefante amarelo de pelúcia. Curto, relativamente fácil de soletrar e pronunciar, sem sentido, e não usado em outro lugar: esses são os meus critérios de nomeação. Crianças são bons em gerar tal”

- Sub-projetos tem recebido nomes de animais também como PIG.

Eco Sistema Hadoop

Spark

The Apache Software Foundation
http://www.apache.org/

AVRO

hadoop HDFS

HBASE

hadoop

mahout

core

oozie

Ganglia

hadoop MapReduce

chukwa

HIVE

Ecosistema Hadoop

- **HDFS** - Um sistema de arquivos distribuído que funciona em grandes aglomerados de máquinas de commodities.

- **PIG** - Uma linguagem de fluxo de dados e ambiente de execução para explorar grandes conjuntos de dados.É executado no HDFS e grupos MapReduce.

- **Hive** - Um armazém de dados (datawarehouse) distribuídos. Gerencia os dados armazenados no HDFS e fornece uma linguagem de consulta baseada em SQL para consultar os dados.

Eco sistema Hadoop

- **HBase** – É um banco de dados orientada por colunas distribuída. HBase usa o HDFS por sua subjacente armazenamento e suporta os cálculos de estilo lote usando MapReduce e ponto consultas (leituras aleatórias).
- **ZooKeeper** – É um serviço de coordenação altamente disponível e distribuído. Fornece primitivas tais como bloqueios distribuídos que podem ser usados para a construção de aplicações distribuídas.

Eco sistema Hadoop

- **Sqoop** – É uma ferramenta para a movimentação eficiente de dados entre bancos de dados relacionais e HDFS.
- **Hadoop-Core** – É um conjunto de componentes e interfaces para sistemas de arquivos distribuídos e geral de I/O (serialização, Java RPC, estruturas de dados persistentes).

Eco sistema Hadoop

- **Mahout -** Aprendizagem por máquina escalável, de fácil uso comercial para a construção de aplicativos inteligentes.

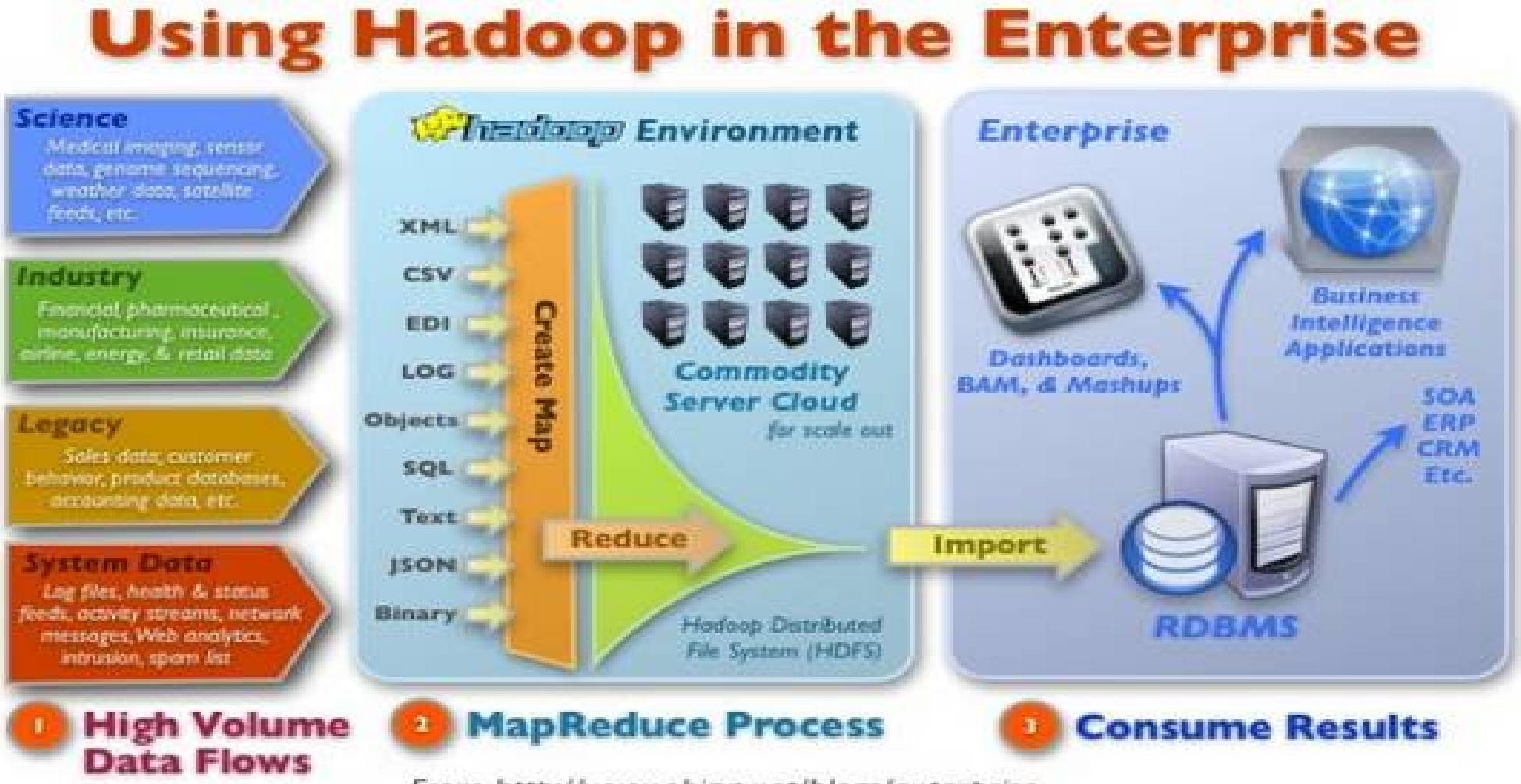
- **Avro** - Um sistema de serialização para eficiência, cross-language RPC, e os dados persistentes armazenamento

- **MapReduce** – É um modelo de processamento distribuído de dados e ambiente de execução que corre em grande clusters de máquinas de commodities.

Possibilidades de Uso

- DataWareHouse
- Business Intelligence
- Aplicações analíticas
- Mídias Sociais
- Sugestão de Compras
- Analise preditiva
- Compras Coletivas
- Recomendações

Possibilidades de Uso



From http://www.ebizq.net/blogs/enterprise

Empresa Usando Hadoop

- Amazon
- Facebook
- Google
- IBM
- Yahoo
- Buscapé
- Linkedin
- Joost
- Last.fm
- New York Times
- PowerSet
- Veoh
- Twitter
- Ebay

Facebook e Hadoop

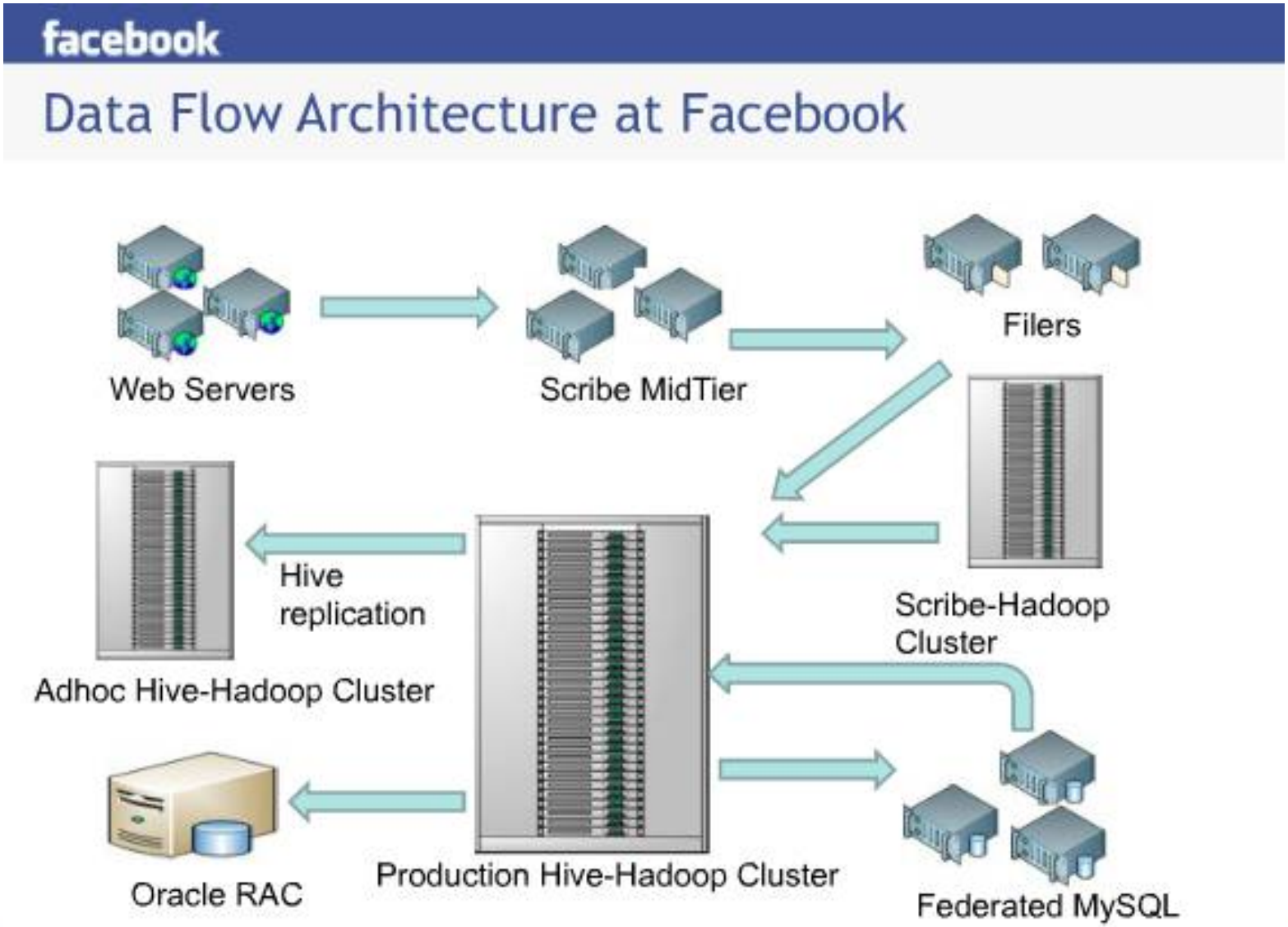
- Hadoop Hive, e HBase para armazenamento de dados e serviço de aplicação em tempo real.
- Seus clusters de armazenamento de dados são petabytes em tamanho, com milhares de nós, e eles usam grupos HBase-driven, em tempo real, em separado para mensagens e análises em tempo real

Cluster Facebook

- Cluster em produção
- 4800 cores, 600 máquina, 16GB por máquina – Abril/2009
- 8000 cores, 1000 máquinas, 32 GB por máquina – julho/2009
- 8400 cores, X? máquina, 32GB por máquina, Jan/2010
- 2 níveis de rede hierarquica, 40 máquinas por rack
- Total do tamanho do cluster 12.5 PB em 2010
- 12 TB novos dados, e 135TB de dados processados por dia.

  Fonte: https://www.slideshare.net/royans/facebooks-petabyte-scale-data-warehouse-using-hive-and-hadoop

Data Flow - Facebook



Yahoo e Hadoop

- Usa Hadoop para análise de dados, aprendizado de máquina, ranking de busca, antispam e-mail, a otimização de anúncios. Combinados, tem mais de 40 mil servidores que executam Hadoop com 170 PB de armazenamento.

Twitter e Hadoop

- Twitter usa Hadoop, Pig e HBase para análise de dados, visualização, análise de gráfico social e aprendizagem de máquina. Twitter LZO comprime todos os seus dados.

Outras empresa usando...

- Samsung, Rackspace, JP Morgan, Groupon, AOL e StumbleUpon são algumas das outras organizações que também são fortemente investidos em Hadoop
- A Microsoft também está começando a trabalhar com Hortonworks para garantir que Hadoop funciona em sua plataforma.
- Todas as empresas da lista de Fortune 500 de alguma forma usam o Hadoop.
- http://wiki.apache.org/hadoop/PoweredBy

Distribuições Hadoop

- **Open Source**
  Apache
- **Comercial Open Source**
  - Cloudera
  - HortonWorks
  - MapR
  - AWS MapReduce
  - Microsoft HDInsight (beta)

Ambiente Livre
Tecnologia

Big Data no Brasil e com Hadoop