

# BIG DATA



# **Disciplina: Aplicações de Big Data com Hadoop**

## **Tema da Aula: Introdução ao Hadoop e HDFS**

### **Coordenação:**

Prof. Dr. Adolpho Walter  
Pimazzi Canton

Profa. Dra. Alessandra de  
Ávila Montini

**Profa. Rosangela de Fátima Pereira**

**Junho de 2016**

# Curriculum

## Formação

- Mestrado em Engenharia de Computação pela Escola Politécnica da Universidade de São Paulo (Poli-USP) ([em andamento](#))
- Especialização em Tecnologia Java pela Universidade Tecnológica Federal do Paraná (UTFPR) (2011)
- Tecnologia em Análise e Desenvolvimento de Sistemas pela UTFPR ([2011](#))
- Bacharelado em Administração de Empresas pela Universidade Estadual do Norte do Paraná (UENP) (2007)

## Experiência

- Professora de Big Data Analytics em empresas e programas de MBA - FIA ([2013 - atual](#))
- Pesquisadora no Laboratório de Arquitetura e Redes de Computadores (LARC) – USP ([2013 - atual](#))
- Professora de cursos de engenharia na UTFPR ([2011 -2012](#))
- Analista de sistemas na BSI Tecnologia ([2009-2010](#))

LinkedIn: <https://br.linkedin.com/pub/rosangela-de-fatima-pereira/68/a10/b56>

Apaixonada por Big Data!

# Conteúdo da Aula

- História do Hadoop
- Características do Hadoop
- HDFS

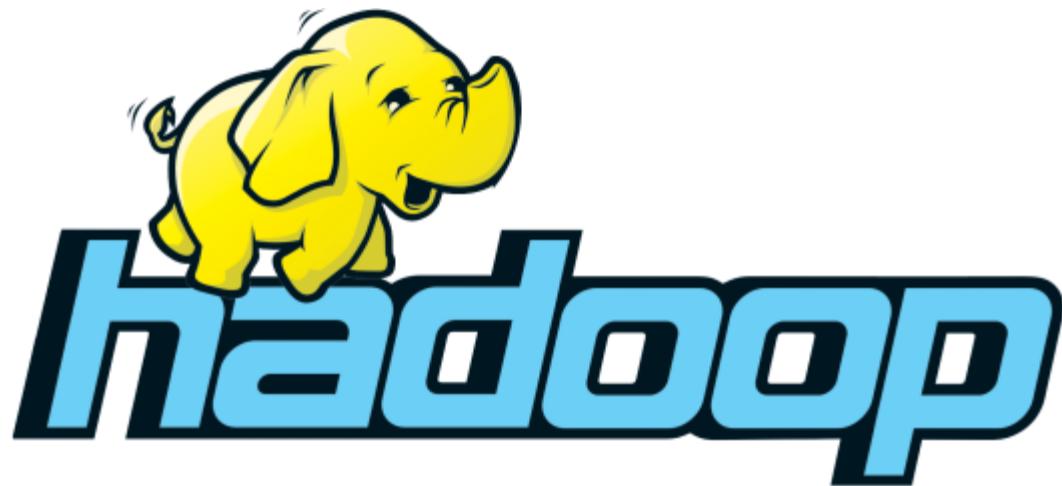
# Objetivo da Aula

Introduzir o aluno ao arcabouço Hadoop, apresentando suas características e vantagens, apresentando também a arquitetura e funcionamento do sistema de arquivos do Hadoop

# Big Data

## Necessidade

Novas tecnologias capazes de oferecer  
escalabilidade, disponibilidade, flexibilidade e  
desempenho para o processamento de grande  
volume de dados (Big Data)



# Hadoop - Características

Arcabouço de software **open source** que permite a execução de aplicações utilizando milhares de máquinas

# Hadoop - Características

Arcabouço de software **open source** que permite a execução de aplicações utilizando milhares de máquinas

Oferece recursos de armazenamento, gerenciamento e processamento **distribuído** de dados (SO de Big Data)

# Hadoop - Características

Arcabouço de software **open source** que permite a execução de aplicações utilizando milhares de máquinas

Oferece recursos de armazenamento, gerenciamento e processamento **distribuído** de dados (**SO de Big Data**)

Projetado para **processamento em lote** de grandes conjuntos de dados

# Hadoop - Características

Arcabouço de software **open source** que permite a execução de aplicações utilizando milhares de máquinas

Oferece recursos de armazenamento, gerenciamento e processamento **distribuído** de dados (**SO de Big Data**)

Projetado para **processamento em lote** de grandes conjuntos de dados

Um dos **pioneiros** da geração de tecnologias de Big Data

# Hadoop - História



- Motor de busca Web
- Projeto open source da Apache
- Muitas tarefas para implementar
- Escalabilidade limitada (4 máquinas)
- Criado por Doug Cutting e Mike Cafarella



# Hadoop - História

Obrigada Google!

## The Google File System

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung  
Google\*

### ABSTRACT

We have designed and implemented the Google File System, a scalable distributed file system for large distributed data-intensive applications. It provides fault tolerance while running on inexpensive commodity hardware, and it delivers high aggregate performance to a large number of clients.

While sharing many of the same goals as previous distributed file systems, our design has been driven by observations of our application workloads and technological environment, both current and anticipated, that reflect a marked departure from some earlier file system assumptions. This has led us to reexamine traditional choices and explore radically different design points.

### 1. INTRODUCTION

We have designed and implemented the Google File System (GFS) to meet the rapidly growing demands of Google's data processing needs. GFS shares many of the same goals as previous distributed file systems such as performance, scalability, reliability, and availability. However, its design has been driven by key observations of our application workloads and technological environment, both current and anticipated, that reflect a marked departure from some earlier file system design assumptions. We have reexamined traditional choices and explored radically different points in the design space.

First, component failures are the norm rather than the

## MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat  
jeff@google.com, sanjay@google.com  
Google, Inc.

### Abstract

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. Many

given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures conspire to obscure the original simple computa-

2003

2004

# Hadoop - História



- Motor de busca Web
- Projeto open source da Apache
- Sistema distribuído
- Tarefas automatizadas
- Melhoria na escalabilidade (20 máquinas)

Para aumentar a escalabilidade (execução em milhares de máquinas) era necessário mais recursos e desenvolvedores

# Hadoop - História

Como indexar  
toda a Web?

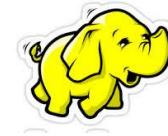
YAHOO!

# Hadoop - História



Como indexar  
toda a Web?

Yahoo! contrata Doug Cutting



Sistema distribuído do Nutch passa a ser um  
outro projeto Apache independente

Em 2006, o novo projeto passa a se chamar  
Hadoop

Em 2009 Yahoo! executa 100 terabytes de  
dados em mais de 3 mil nós

# Hadoop - Aplicações

Primeiramente utilizado por empresas do Vale do Silício



# Hadoop - Aplicações

Atualmente utilizado por diferentes áreas



Telecomunicação

Bioinformática



Agricultura

Varejo



Entretenimento

Manufatura



# Hadoop - Aplicações

[Globo.com](#)

## veja também

**Ministro defende importância do Reintegra e do PSI**

Declarações são do ministro do Desenvolvimento, Armando Monteiro Neto. 'Não faz sentido retirar um instrumento como o Reintegra', afirmou ele.

09/02/2015



'Câmbio não pode ser usado para combater a inflação', defende ministro

02/02/2015



Governo anunciará em março novo plano para incentivar exportações

02/02/2015

## Recomendadas para você



Galaxy A8, suposto top 'finíssimo' da Samsung, aparece em vídeo



Pesquisa desenvolve chip potente que promete ser ecologicamente correto



Facebook Lite: confira todas as funções disponíveis no app 'levinho'



Google Chrome não abre no PC? Três programas podem estar atrapalhando

# Hadoop - Aplicações

## China Mobile



# Hadoop - Aplicações



## China Mobile

**Desafio:** armazenar milhões de registros de chamadas móveis e fornecer acesso em tempo real aos registros das chamadas e informações de faturamento para os clientes

**Requisitos:** escalabilidade, qualidade do serviço e custo

**Solução:** HBase foi usado para armazenar bilhões de linhas de detalhes dos registros de chamadas. 30TB de dados é mensalmente adicionado à base de dados

**Cluster:** mais de 100 nós

Fonte: <http://www.intel.com/content/dam/www/public/us/en/documents/articles/intel-distribution-for-apache-hadoop-mobile-case-study.pdf>



# Hadoop - Aplicações

NetApp



# Hadoop - Aplicações



**NetApp:** coleta dados de diagnóstico de seus sistemas de armazenamento implantados nas instalações dos clientes. Esta informação é usada para analisar a saúde dos sistemas NetApp

**Desafio:** coletar mais de 600.000 transações de dados semanais, consistindo de registros não estruturados e informação de diagnóstico dos sistemas

**Solução:** um sistema Hadoop capta os dados e permite o processamento paralelo dos dados

**Cluster / tamanho de dados:** mais de 30 nós; 7TB de dados / mês



# Hadoop - Aplicações

Demais empresas que utilizam Hadoop



**NOKIA**



**box**



Serasa Experian



**JOHN DEERE**



**Spotify**

**MONSANTO**



**ebay**

**Tail Target**



**SAMSUNG**  
SAMSUNG ELECTRONICS

**NATIONAL CANCER INSTITUTE**

# Hadoop - Distribuições

Essas empresas utilizam diferentes distribuições Hadoop



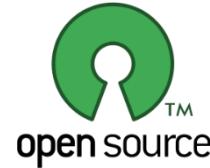
# Hadoop – Motivação

Por que as empresas estão  
escolhendo o Hadoop?

# Hadoop – Motivação

Por que as empresas estão escolhendo o Hadoop?

Redução de custo



# Hadoop – Motivação

Por que as empresas estão escolhendo o Hadoop?

Flexibilidade



# Hadoop – Motivação

Por que as empresas estão escolhendo o Hadoop?

Escalabilidade



# Hadoop – Motivação

Por que as empresas estão escolhendo o Hadoop?

Novas análises

Recomendação de serviços

Motor de busca

Personalização de conteúdo

Análise de DNA

Análise de influência

Segmentação de clientes

Detecção de fraude

Detecção de anomalias

Análise de risco

Análise de sentimento

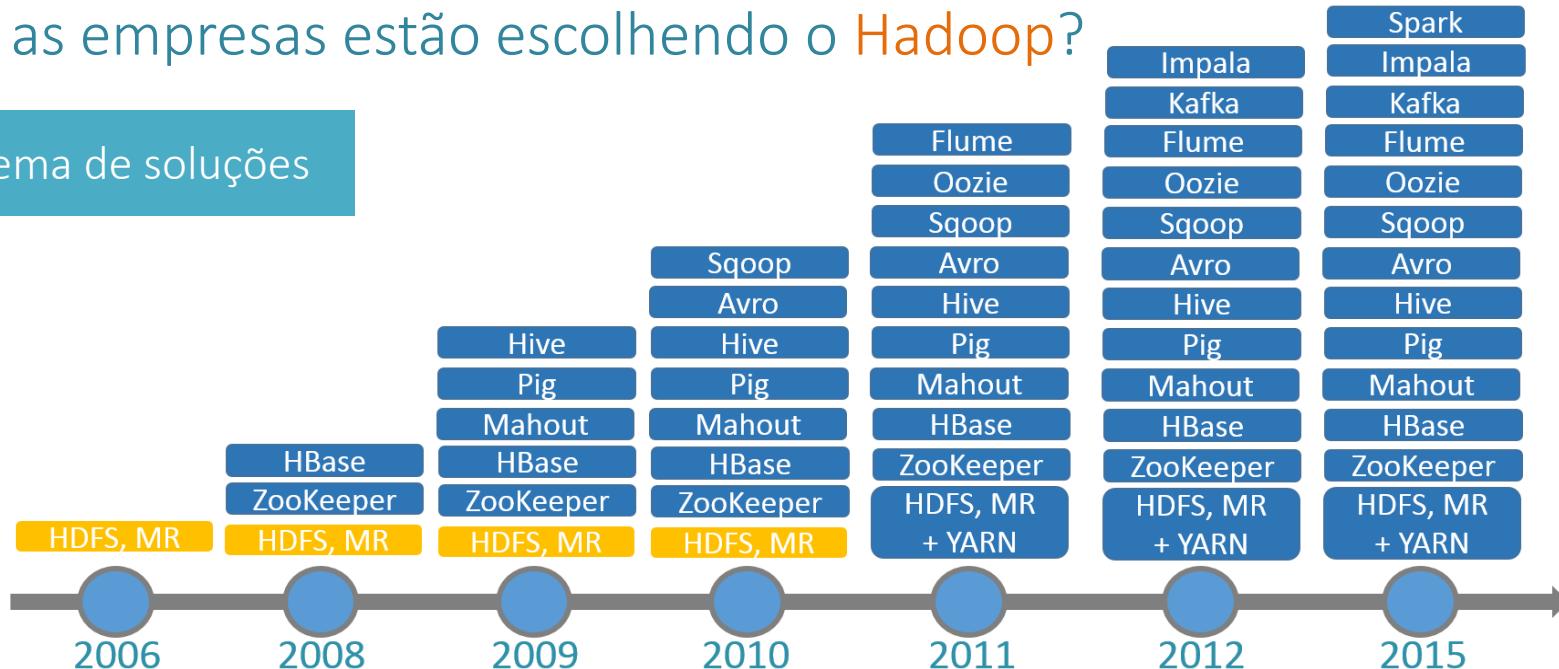
Publicidade personalizada

Manutenção preditiva

# Hadoop – Motivação

Por que as empresas estão escolhendo o Hadoop?

Ecossistema de soluções



# Hadoop – Motivação

Como Hadoop oferece esses  
benefícios?

# Hadoop

Hadoop possui dois componentes principais





# HDFS

## HADOOP DISTRIBUTED FILE SYSTEM - HDFS

- Sistema de arquivos distribuído

# HDFS

## HADOOP DISTRIBUTED FILE SYSTEM - HDFS

- Sistema de arquivos distribuído
- Executado em um sistema de arquivos nativo

# HDFS

## HADOOP DISTRIBUTED FILE SYSTEM - HDFS

- Sistema de arquivos distribuído
- Executado em um sistema de arquivos nativo
- Otimizado para processamento de grande volume de dados (alta taxa de transferência)

# HDFS

## HADOOP DISTRIBUTED FILE SYSTEM - HDFS

- Sistema de arquivos distribuído
- Executado em um sistema de arquivos nativo
- Otimizado para processamento de grande volume de dados (alta taxa de transferência)
- Abstrai questões de armazenamento distribuído dos dados

# HDFS

## HADOOP DISTRIBUTED FILE SYSTEM - HDFS

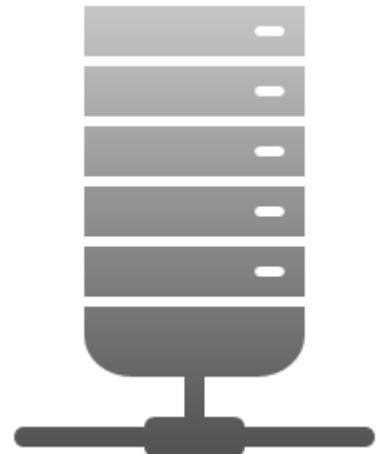
- Sistema de arquivos distribuído
- Executado em um sistema de arquivos nativo
- Otimizado para processamento de grande volume de dados (alta taxa de transferência)
- Abstrai questões de armazenamento distribuído dos dados
- Escalável e tolerante a falhas

# HDFS

## ABORDAGEM TRADICIONAL



TRANSFERÊNCIA DOS DADOS PARA  
LOCAL DE PROCESSAMENTO

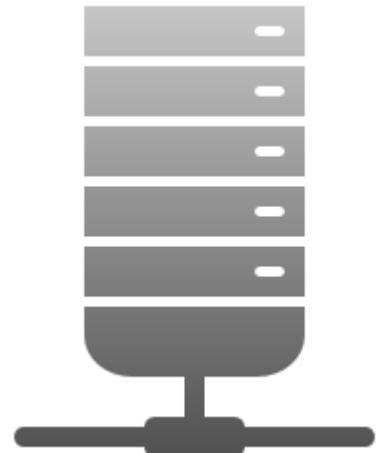


# HDFS

## ABORDAGEM TRADICIONAL



Como mover 500 GB, 1 TB, 2 TB....?



# HDFS

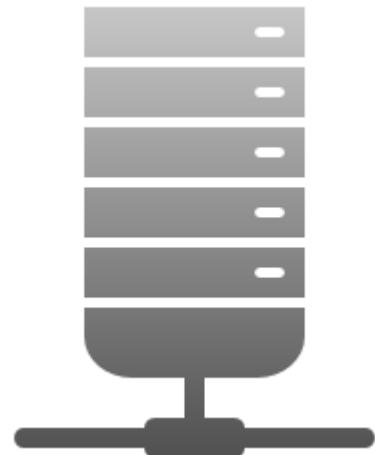
## ABORDAGEM DO HDFS



TRANSFERÊNCIA DE TAREFAS PARA  
ONDE DADOS ESTÃO  
ARMAZENADOS



*“Write once, read many”*



# HDFS

HDFS possui 3 tipos de processos

NameNode

- Gerencia o *namespace* do sistema de arquivos do Hadoop

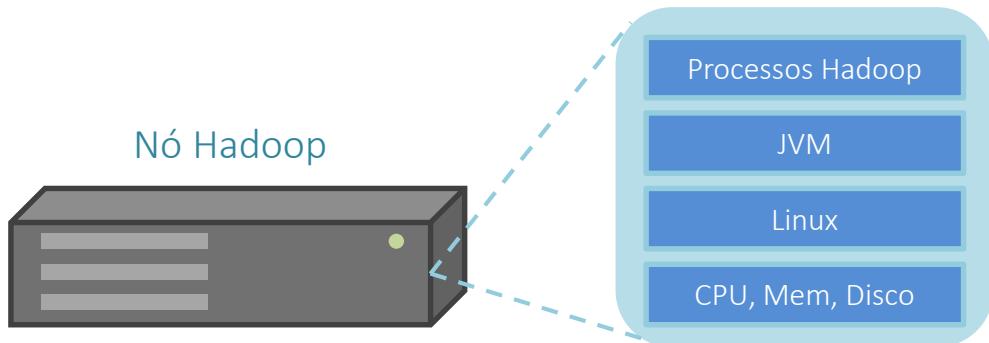
DataNode

- Armazena os blocos de dados em um nó

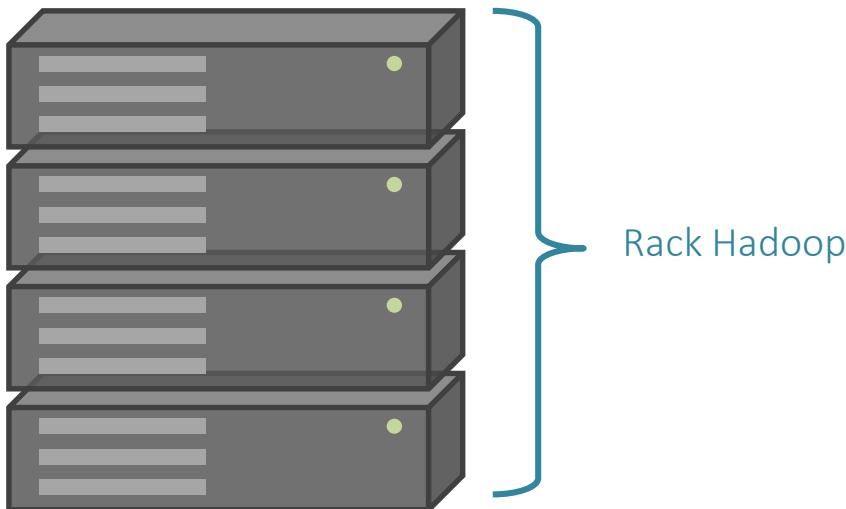
SecondaryNameNode

- Oferece tarefas de ponto de verificação e manutenção do NameNode

# HDFS

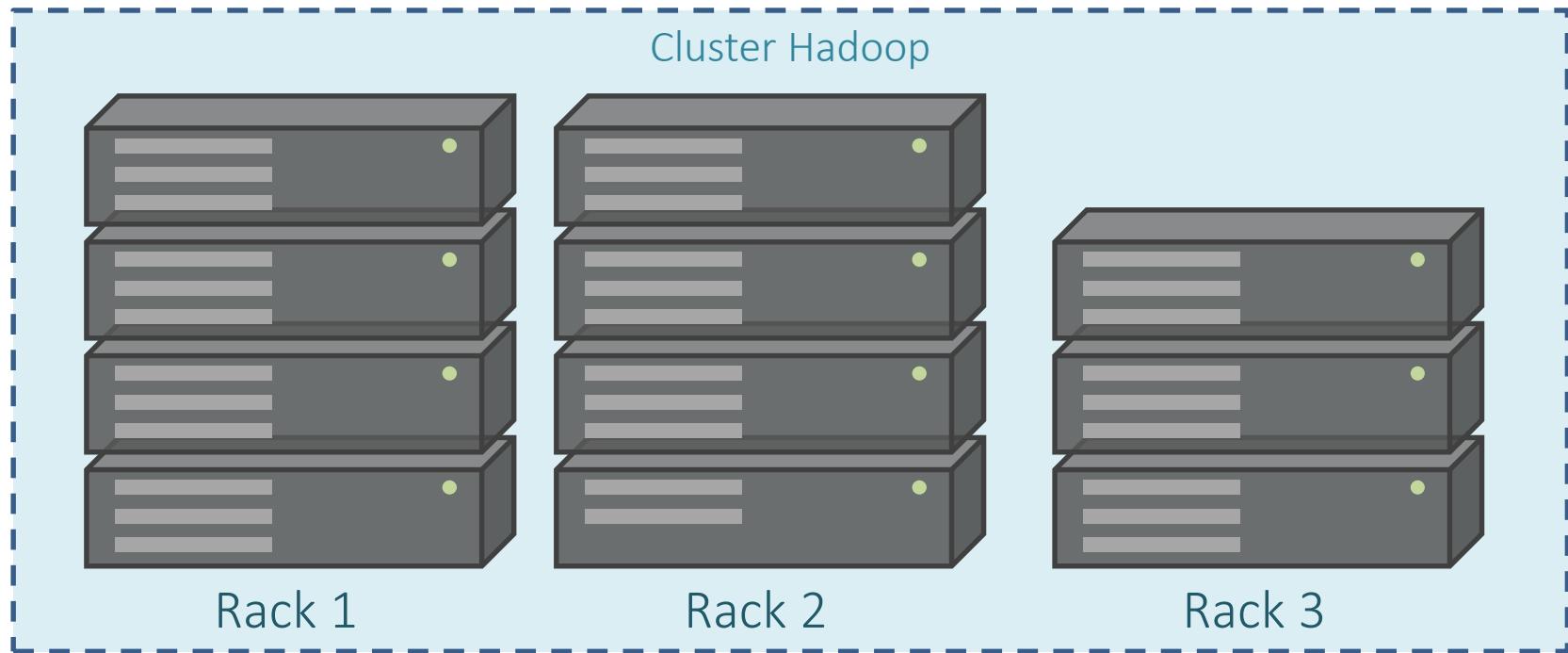


# HDFS

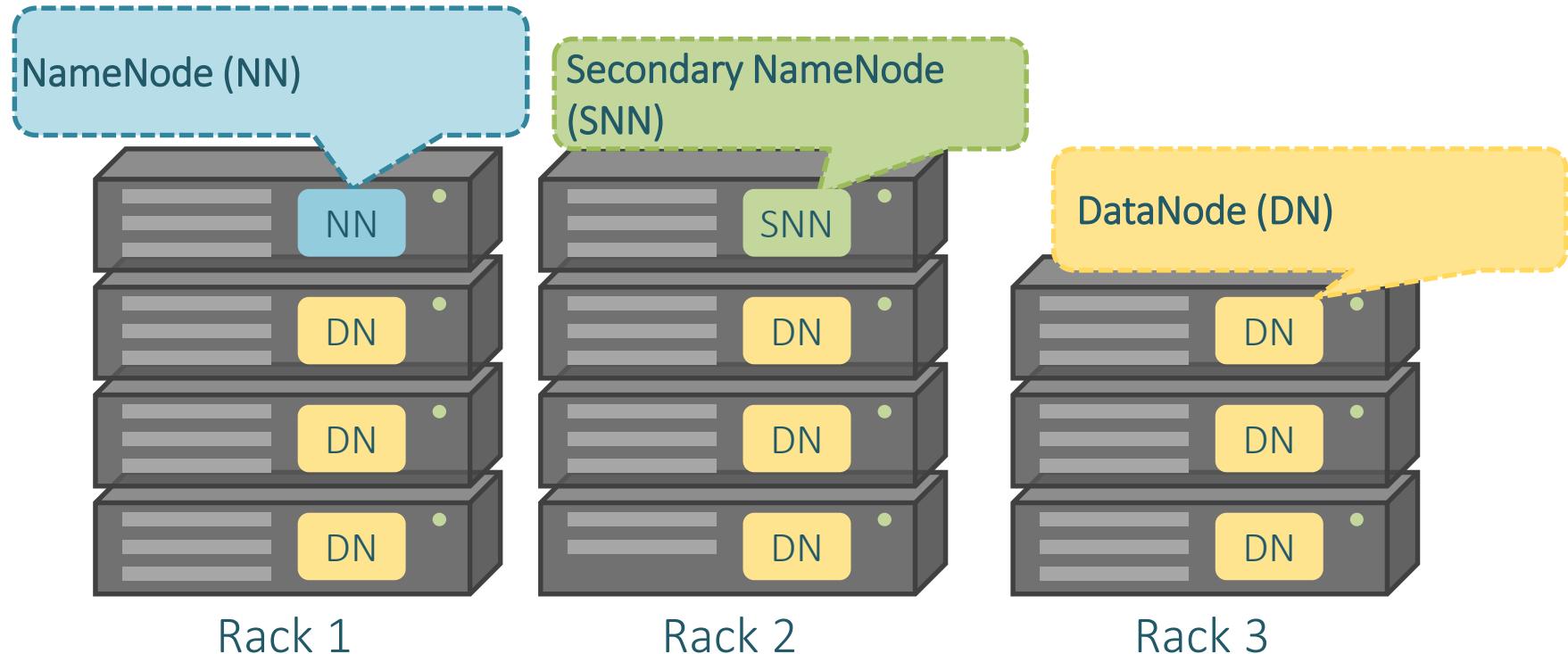


Rack Hadoop

# HDFS

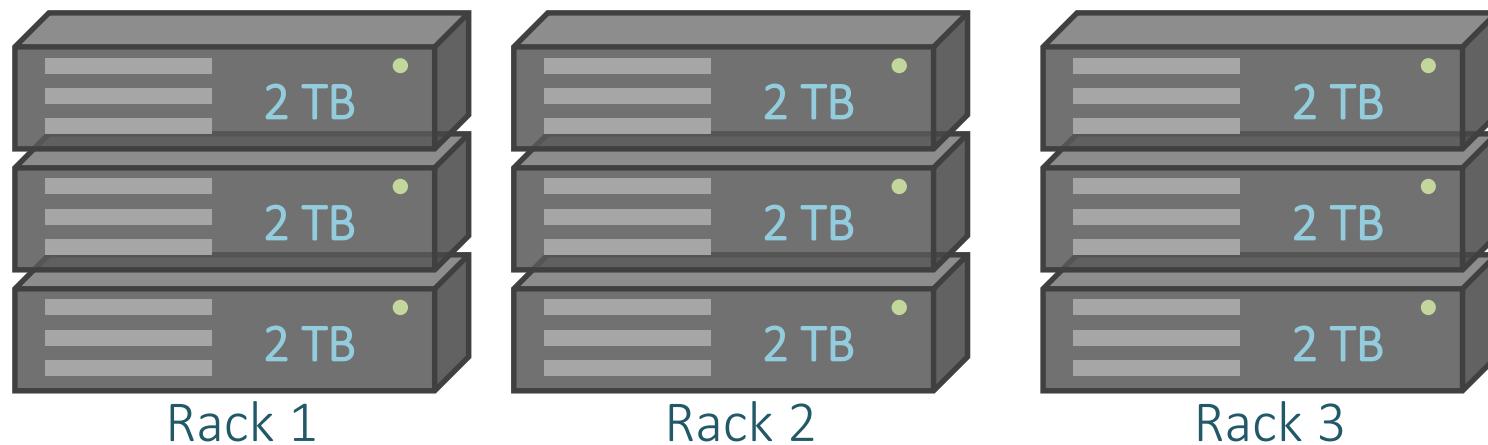


# HDFS



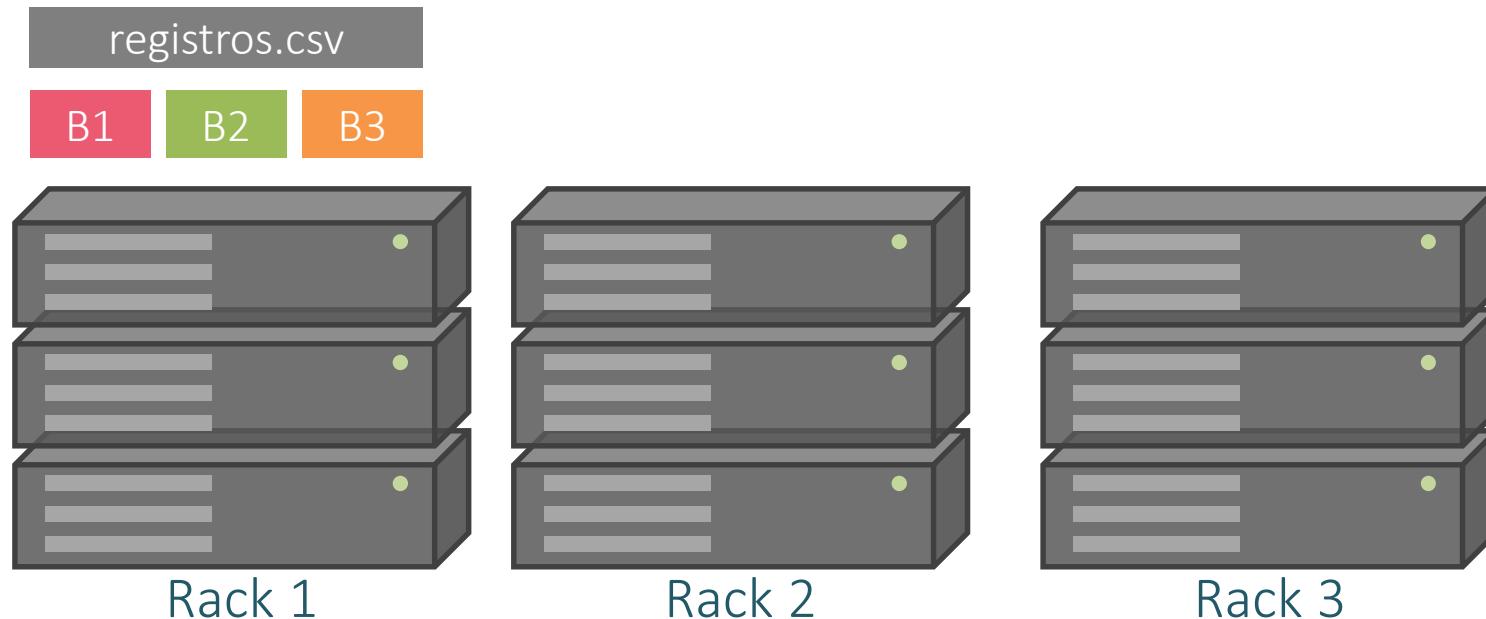
# HDFS

Otimizado para ler/armazenar grandes arquivos em um cluster



# HDFS

Arquivos são divididos em blocos de 64 MB (tamanho default)

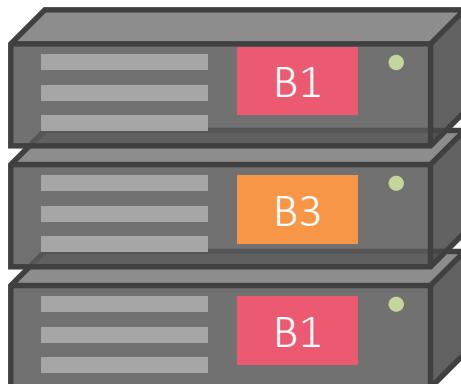


# HDFS

Blocos são replicados para tolerância a falhas (default - 3 réplicas)



Rack 1



Rack 2



Rack 3

# HDFS

## Opções para carregamento de dados para o HDFS

- Linha de comando: **File System (fs) shell**
  - `hadoop fs –ls`
  - `hadoop fs –mkdir`
  - `hadoop fs –put`
  - `hadoop fs -get`

# HDFS

## Opções para carregamento de dados para o HDFS

- Java API
  - Classe FileSystem

```
InputStream in = null;
try {
    in = new URL("hdfs://host/path").openStream();
    // process in
} finally {
    IOUtils.closeStream(in);
}
```

# HDFS

## Opções para carregamento de dados para o HDFS

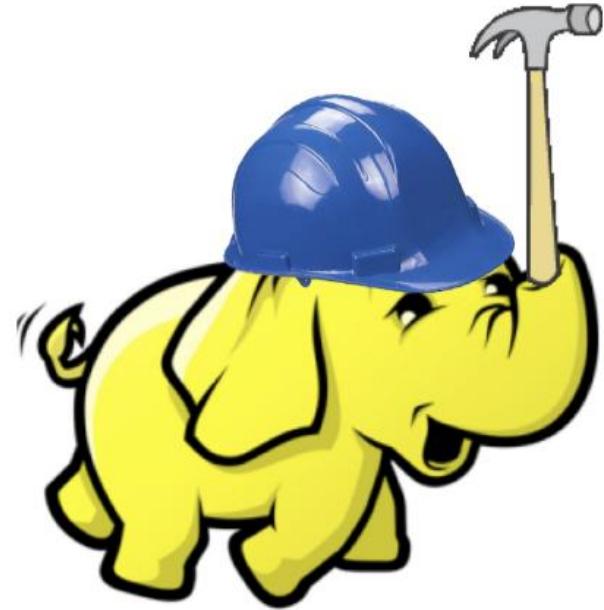
- Ecossistema Hadoop
  - Sqoop
  - Flume
  - Hue
  - ...



THE IMPALA WEB UI



# HDFS na prática



# Configurando o Hadoop

## Modos de execução

- Local (standalone)
  - Executado como um único processo java
  - Recomendado para depuração de código
- Pseudo-distribuído
  - Todos os componentes Hadoop são executados em uma única máquina
  - Cada componente é executado em um processo java separado
- Completamente distribuído
  - Cluster Hadoop utilizando múltiplas máquinas

# HDFS na Prática

Software: [VMware](#)

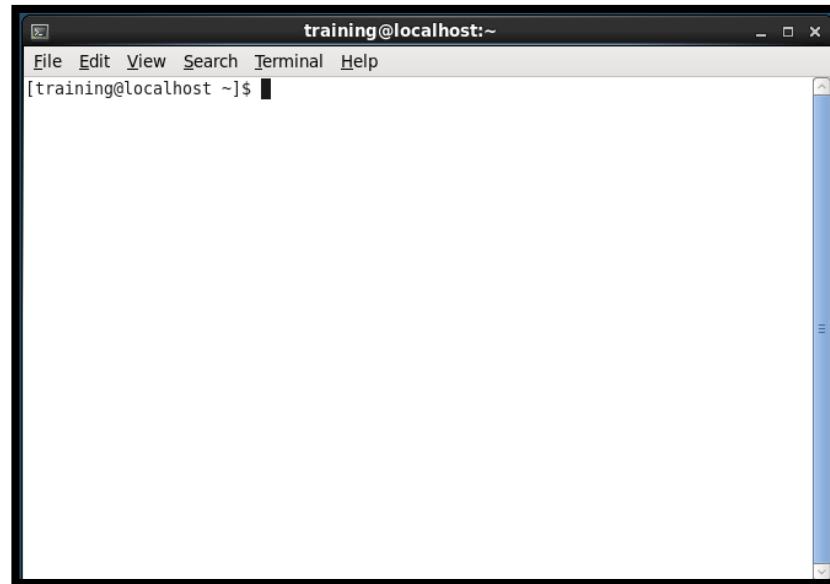
Máquina virtual: Cloudera-training-analyst



**cloudera**

# HDFS na Prática

Abrir um terminal



# HDFS na Prática

1. Comandos FS Shell
2. Importando dados do HDFS para MySQL usando Sqoop

# HDFS na Prática – FS Shell

## Objetivo:

- Submeter uma base de dados de compras de clientes para o HDFS
- Manipular dados armazenados no HDFS
- Permitir a análise de dados em ambiente distribuído e escalável

# HDFS na Prática – FS Shell

## Base de dados

- Dados de registros de compras em um e-commerce
- 1 milhão de registros
- Campos: **data, horário, local, categoria, valor, forma**

## Exemplo:

2015-01-01; 09:00:00; Sao Paulo; Livros; 214.05; Visa

2015-01-01; 09:01:00; Osasco; Artesanato; 88.25; Visa

# HDFS na Prática – FS Shell

## Verificar processos

```
[training@localhost ~]$ sudo jps
8452 JobTracker
1951 TaskTracker
19241 Jps
2015 DataNode
8849 NameNode
```

# HDFS na Prática – FS Shell

## Verificar versão do Hadoop

```
training@localhost ~]$ hadoop version
Hadoop 2.5.0-cdh5.2.0
Subversion http://github.com/cloudera/hadoop -r
e1f20a08bde76a33b79df026d00a0c91b2298387
Compiled by jenkins on 2014-10-11T21:00Z
Compiled with protoc 2.5.0
From source with checksum 309bccd135b199bdfdd6df5f3f4153d
This command was run using /usr/lib/hadoop/hadoop-common-2.5.0-cdh5.2.0.jar
```

# HDFS na Prática – FS Shell

## Criar um diretório no HDFS

```
[training@localhost ~]$ hadoop fs -mkdir input
```

# HDFS na Prática – FS Shell

## Enviar base de dados para o HDFS

```
[training@localhost ~]$ hadoop fs -put ~/bases/compras.txt input
```

# HDFS na Prática – FS Shell

## Listar arquivos armazenados no HDFS

```
[training@localhost ~]$ hadoop fs -ls
Found 1 items drwxrwxrwx - training supergroup 0 2016-03-29 09:12 input
```

# HDFS na Prática – FS Shell

## Listar arquivos armazenados no HDFS

```
[training@localhost ~]$ hadoop fs -ls input
Found 1 items-rw-rw-rw- 1 training supergroup 53576655 2016-03-29 09:12
input/compras.txt
```

# HDFS na Prática – FS Shell

Contar número de linhas no arquivo compras.txt

```
[training@localhost ~]$ hadoop fs -cat input/compras.txt | wc -l  
1000000
```

# HDFS na Prática – FS Shell

## Visualizar conteúdo de arquivo no HDFS

```
[training@localhost ~]$ hadoop fs -cat input/compras.txt
```

```
2015-01-01;09:00:00;Sao Paulo;Roupas masculinas;214.05;Amex
2015-01-01;09:00:00;Rio de Janeiro;Roupas femininas;153.57;Visa
2015-01-01;09:00:00;Curitiba;Musica;66.08;Dinheiro
...
...
```

```
[training@localhost ~]$ hadoop fs -tail input/compras.txt
```

```
i;Video Games;271.24;Visa
2015-03-30;11:04:00;Campinas;Computadores;20.96;Amex
2015-03-30;11:04:00;Santos;Saude e beleza;28.85;Dinheiro
2015-03-30;11:04:00;Barra Mansa;Roupas femininas;400.96;Visa
...
...
```

# HDFS na Prática – FS Shell

## Copiar arquivo do HDFS para arquivo local

```
[training@localhost ~]$ hadoop fs -get input/compras.txt  
/home/training/bases/compras_cp.txt
```

```
[training@localhost ~]$ ls /home/training/bases  
compras_cp.txt compras.txt
```

# HDFS na Prática – FS Shell

## Copiar um arquivo para outro diretório (ou outro cluster)

```
[training@localhost ~]$ hadoop distcp /user/training/input  
/user/training/input2
```

```
16/03/29 09:48:07 INFO tools.DistCp: srcPaths=[/user/training/input]  
16/03/29 09:48:07 INFO tools.DistCp: destPath=/user/training/input2  
16/03/29 09:48:09 INFO tools.DistCp: /user/training/input2 does not exist.  
16/03/29 09:48:10 INFO tools.DistCp: sourcePathsCount=2  
16/03/29 09:48:10 INFO tools.DistCp: filesToCopyCount=1  
16/03/29 09:48:10 INFO tools.DistCp: bytesToCopyCount=48.3 M  
16/03/29 09:48:10 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement  
Tool for the same.  
16/03/29 09:48:10 INFO mapred.JobClient: Running job: job_201602181111_0004  
16/03/29 09:48:11 INFO mapred.JobClient: map 0% reduce 0%  
...
```

# HDFS na Prática – FS Shell

## Removendo um arquivo

```
[training@localhost ~]$ hadoop fs -rm input2/compras.txt
16/03/29 09:52:46 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval =
0 minutes, Emptier interval = 0 minutes.Deleted input2/compras.txt
```

```
[training@localhost ~]$ hadoop fs -ls
Found 3 items
drwxrwxrwx - training supergroup          0 2016-03-29 09:48 _distcp_logs_12isje
drwxrwxrwx - training supergroup          0 2016-03-29 09:12 input
drwxrwxrwx - training supergroup          0 2016-03-29 09:52 input2
```

# HDFS na Prática – FS Shell

## Removendo um diretório

```
[training@localhost ~]$ hadoop fs -rm -r input2
16/03/29 09:53:14 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval =
0 minutes, Emptier interval = 0 minutes. Deleted input2
```

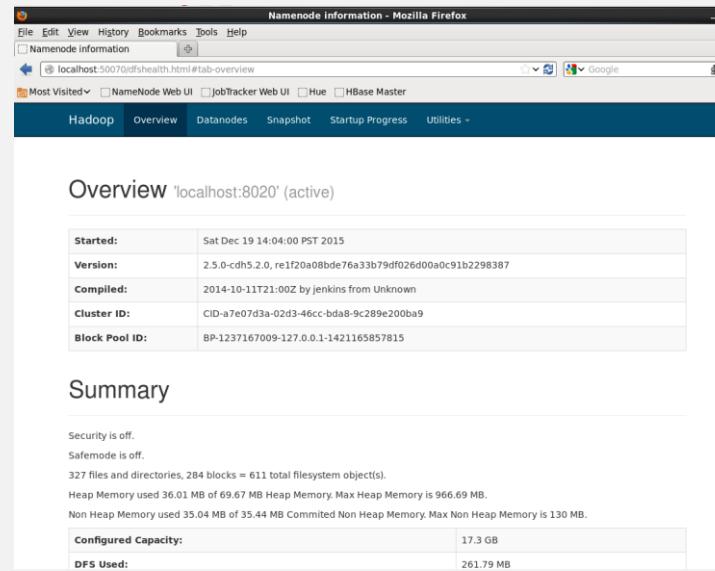
```
[training@localhost ~]$ hadoop fs -ls
Found 2 items
drwxrwxrwx - training supergroup      0 2016-03-29 09:48 _distcp_logs_12isje
drwxrwxrwx - training supergroup      0 2016-03-29 09:12 input
```

# HDFS na Prática – FS Shell

## Acessar a interface WEB do HDFS

Abrir o navegador **Firefox**

Selecionar a opção **NameNode**



Started: Sat Dec 19 14:04:00 PST 2015  
Version: 2.5.0-cdh5.2.0, re1f20a08bde76a33b79df026d00a0c91b2298387  
Compiled: 2014-10-11T21:00Z by jenkins from Unknown  
Cluster ID: CID-a7e07d3a-02d3-46cc-bda8-9c289e200ba9  
Block Pool ID: BP-1237167009-127.0.0.1-1421165857815

Summary

Configured Capacity: 17.3 GB  
DFS Used: 261.79 MB

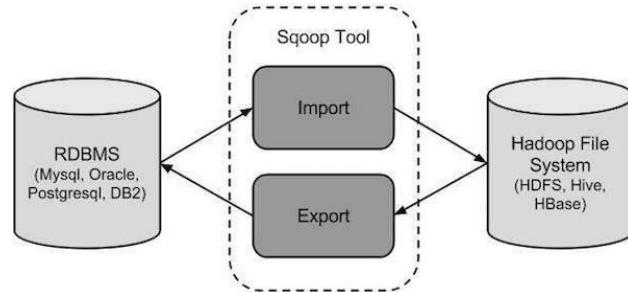
# HDFS na Prática

1. Comandos FS Shell
2. Exportar dados do HDFS para MySQL usando Sqoop

# HDFS na Prática – Sqoop

## Sqoop

- Ferramenta projetada para transferir dados entre o HDFS e bancos de dados relacionais
- Facilita a manipulação de dados estruturados e não estruturados



# HDFS na Prática – Soop

## Acessar interface MySQL

```
[training@localhost ~]$ mysql -uroot
```

Welcome to the MySQL monitor. Commands end with ; or \g.

Your MySQL connection id is 21

Server version: 5.1.69 Source distribution

Copyright (c) 2000, 2013, Oracle and/or its affiliates.

All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its

affiliates. Other names may be trademarks of their respective owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

# HDFS na Prática – Soop

## Criar banco de dados

```
mysql> create database compras;
```

# HDFS na Prática – Soop

## Criar banco de dados

```
mysql> use compras;  
  
mysql> create table compras_2015 (  
        data      DATE,  
        horario   TIME,  
        local     VARCHAR(50),  
        categoria VARCHAR(50),  
        valor     VARCHAR(20),  
        forma     VARCHAR(30));  
  
mysql> exit
```

# HDFS na Prática – Sqoop

## Exportar dados do HDFS para a tabela do MySQL

```
[training@localhost ~]$ sqoop export  
  -m 1  
  --connect jdbc:mysql://localhost:3306/compras  
  --username root  
  --table compras_2015  
  --export-dir /user/training/input/  
  --input-fields-terminated-by ';' '  
  --mysql-delimiters
```

# HDFS na Prática – Sqoop

## Exportar dados do HDFS para a tabela do MySQL (continuação)

```
16/03/29 10:58:05 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.2.0
16/03/29 10:58:05 INFO manager.SqlManager: Using default fetchSize of 1000
16/03/29 10:58:05 INFO tool.CodeGenTool: Beginning code generation
16/03/29 10:58:05 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `compras_2015` AS t LIMIT 1
16/03/29 10:58:05 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `compras_2015` AS t LIMIT 1
16/03/29 10:58:08 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-
training/compile/dd529748aa611292ff0ac42f07600349/compras_2015.jar
16/03/29 10:58:08 INFO mapreduce.ExportJobBase: Beginning export of compras_2015
...
16/03/29 10:58:51 INFO mapred.JobClient: Total committed heap usage (bytes)=31850496
16/03/29 10:58:51 INFO mapreduce.ExportJobBase: Transferred 51.1698 MB in 40.3806 seconds (1.2672 MB/sec)
16/03/29 10:58:51 INFO mapreduce.ExportJobBase: Exported 1000000 records.
```

# HDFS na Prática – Soop

## Acessar interface MySQL

```
[training@localhost ~]$ mysql -uroot
```

Welcome to the MySQL monitor. Commands end with ; or \g.

Your MySQL connection id is 21

Server version: 5.1.69 Source distribution

Copyright (c) 2000, 2013, Oracle and/or its affiliates.

All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its

affiliates. Other names may be trademarks of their respective owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

# HDFS na Prática – Soop

## Acessar base de dados compras

```
mysql> use compras
```

Reading table information for completion of table and column names

You can turn off this feature to get a quicker startup with –A

Database changed

# HDFS na Prática – Soop

## Verificar quantidade de registros

```
mysql> select count(*) from compras_2015;  
+-----+  
| count(*) |  
+-----+  
| 1000000 |  
+-----+  
1 row in set (0.00 sec)
```

# HDFS na Prática – Soop

## Verificar registros da tabela

```
mysql> select * from compras_2015 LIMIT 20;
```

| data       | horario  | local          | categoria         | valor  | forma      |
|------------|----------|----------------|-------------------|--------|------------|
| 2015-01-01 | 09:00:00 | Sao Paulo      | Roupas masculinas | 214.05 | Amex       |
| 2015-01-01 | 09:00:00 | Rio de Janeiro | Roupas femininas  | 153.57 | Visa       |
| 2015-01-01 | 09:00:00 | Curitiba       | Musica            | 66.08  | Dinheiro   |
| 2015-01-01 | 09:00:00 | Belo Horizonte | Pet               | 493.51 | Visa       |
| 2015-01-01 | 09:00:00 | Aracaju        | Roupas infantis   | 235.63 | MasterCard |
| 2015-01-01 | 09:00:00 | Salvador       | Roupas masculinas | 247.18 | MasterCard |
| 2015-01-01 | 09:00:00 | Campinas       | Cameras           | 379.6  | Visa       |
| 2015-01-01 | 09:00:00 | Sao Paulo      | Eletronicos       | 296.8  | Dinheiro   |
| 2015-01-01 | 09:00:00 | Londrina       | Brinquedos        | 25.38  | Visa       |
| 2015-01-01 | 09:00:00 | Rio de Janeiro | Brinquedos        | 213.88 | Visa       |
| 2015-01-01 | 09:00:00 | Sao Paulo      | Video Games       | 53.26  | Visa       |
| 2015-01-01 | 09:00:00 | Sao Paulo      | Video Games       | 39.75  | Dinheiro   |
| 2015-01-01 | 09:00:00 | Campinas       | Cameras           | 469.63 | MasterCard |
| 2015-01-01 | 09:00:00 | Ourinhos       | DVds              | 290.82 | MasterCard |
| 2015-01-01 | 09:00:00 | Rio de Janeiro | Musica            | 260.65 | Visa       |
| 2015-01-01 | 09:00:00 | Brasilia       | Jardim            | 136.9  | Visa       |
| 2015-01-01 | 09:00:00 | Porto Alegre   | Roupas femininas  | 483.82 | Visa       |
| 2015-01-01 | 09:00:00 | Sao Paulo      | Roupas femininas  | 215.82 | Dinheiro   |
| 2015-01-01 | 09:00:01 | Maringa        | Cameras           | 418.94 | Amex       |
| 2015-01-01 | 09:00:00 | Florianopolis  | Roupas infantis   | 309.16 | Visa       |

20 rows in set (0.00 sec)

# Resumo

- Hadoop é um framework para armazenamento e processamento de grande volume de dados
- HDFS é o componente do Hadoop responsável pelo armazenamento distribuído dos dados
- O carregamento de dados no HDFS pode ser feito usando FS Shell e Sqoop
- Sqoop permite a transferência de dados entre o HDFS e bancos de dados relacionais

# Sugestão de leitura

## **Apache Hadoop: conceitos teóricos e práticos, evolução e novas possibilidades**

Alfredo Goldman, Fabio Kon, Francisco Pereira Junior,  
Ivanilton Polato, Rosangela de Fátima Pereira

XXXI JAI

<http://www.ime.usp.br/~ipolato/JAI2012-Hadoop.pdf>

# Sugestão de Leitura

OR TEAM et al. **Big data now: current perspectives from O'Reilly Radar**. O'Reilly Media, 2014.

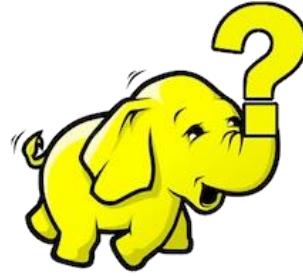
<http://www.oreilly.com/data/free/big-data-now-2014-edition.csp>

DUMBILL, Edd. **Planning for big data**. " O'Reilly Media, Inc.", 2012.

<http://www.oreilly.com/data/free/planning-for-big-data.csp>

# Perguntas

[rpereira@larc.usp.br](mailto:rpereira@larc.usp.br)



# Referências Bibliográficas

- WHITE, Tom. *Hadoop: The definitive guide.* " O'Reilly Media, Inc.", 2015.
  - <http://hadoop.apache.org/>
  - BORTHAKUR, Dhruba. *HDFS architecture guide.* HADOOP APACHE PROJECT [http://hadoop.apache.org/common/docs/current/hdfs\\_design.html](http://hadoop.apache.org/common/docs/current/hdfs_design.html), 2008.
  - OR TEAM et al. **Big data now: current perspectives from O'Reilly Radar.** O'Reilly Media, 2014.
  - DUMBILL, Edd. **Planning for big data.** " O'Reilly Media, Inc.", 2012.
  - OWENS, Jonathan R.; FEMIANO, Brian; LENTZ, Jon. **Hadoop Real World Solutions Cookbook.** Packt Publishing Ltd, 2013.