DIGITAL INNOVATION ONE

UNIVERSITY everis

Aceleração Global Dev #4 everis

Monitoramento de clusters Hadoop de alto nível com HDFS e Yarn

Rodrigo Garcia
Big Data Projects Team Lead

Objetivos da Aula

1. Entender o conceito de Big Data, escalabilidade horizontal e cluster

2. HDFS: Conceito de replicação e principais comandos

3. YARN: Monitoramento de clusters

DIGITAL INNOVATION ONE

Requisitos Básicos

UNIVERSITY everis

- ✓ Linux básico
- ✓ Noções de Shellscript
- ✓ Noções de processamento clusterizado

DIGITAL INNOVATION ONE

Parte 1: Big Data

Monitoramento de clusters Hadoop de alto nível com HDFS e Yarn

Conceito de Big Data

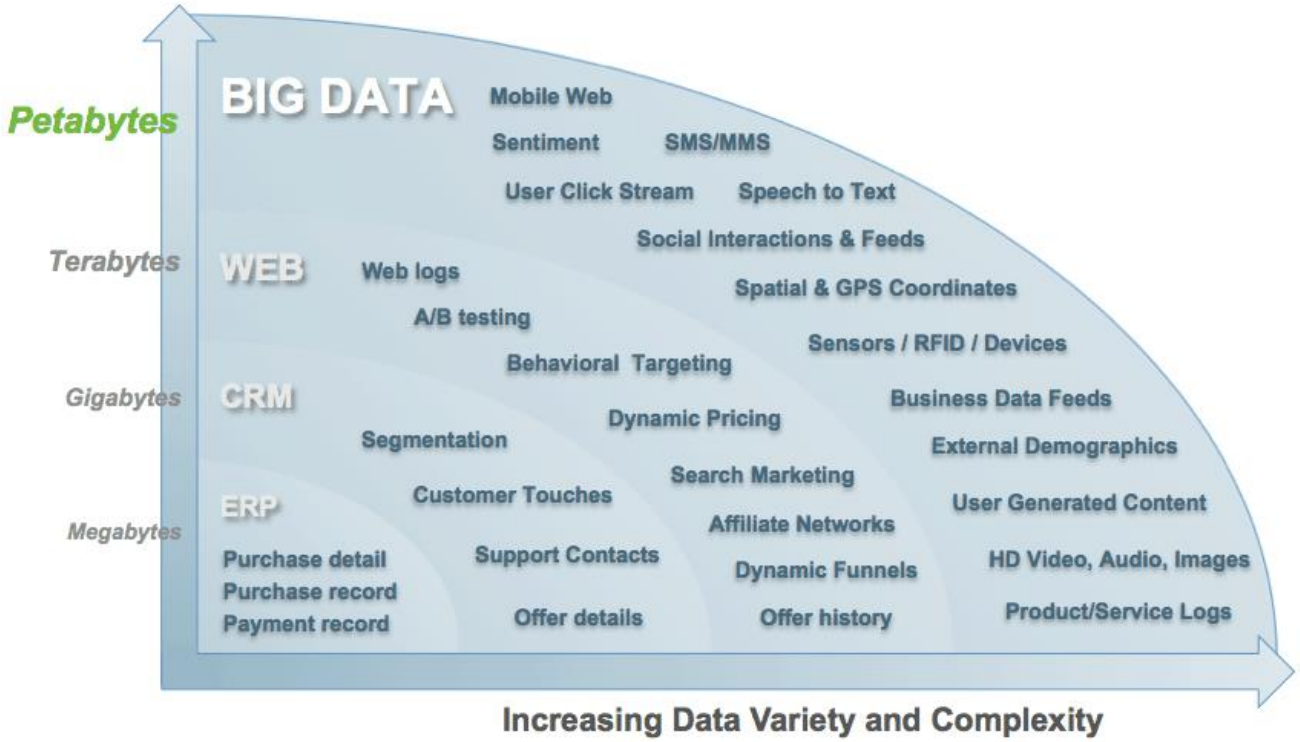
"**Big Data** é um processo de **análise e interpretação** de um **grande volume de dados** armazenados remotamente. (...) O Big Data pode **integrar** qualquer dado coletado sobre um **assunto ou uma empresa**, como os registros de compra e venda e mesmo os canais de interação não digital (telemarketing e call center). **Onde há um registro feito, a tecnologia o alcança.**"

FIA, 2018

DIGITAL INNOVATION ONE

Conceito de Big Data

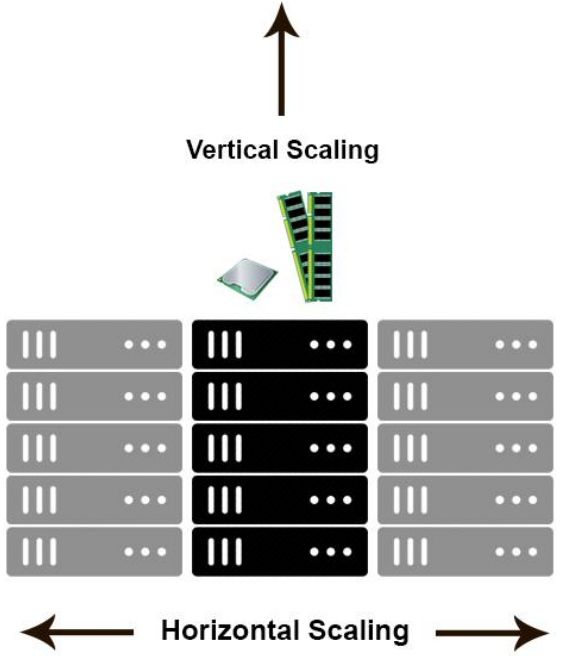
UNIVERSITY everis



Escalabilidade Horizontal

Processamento **tradicional**:
escalabilidade vertical

Processamento **distribuído**:
escalabilidade horizontal



DIGITAL INNOVATION ONE

Cluster

UNIVERSITY everis

O que é um cluster?

É um grupo de computadores que trabalham juntos. Provê armazenamento, processamento e gerenciamento de recursos

DIGITAL INNOVATION ONE

Cluster

UNIVERSITY everis

O que é um nó?
Computador individual no cluster.
O nó master (driver) gerencia a distribuição de trabalho para os nós workers.

O que é um daemon?
É um programa (serviço) rodando em um nó.
Cada um tem sua função no cluster.

DIGITAL
INNOVATION
ONE

Parte 2: HDFS

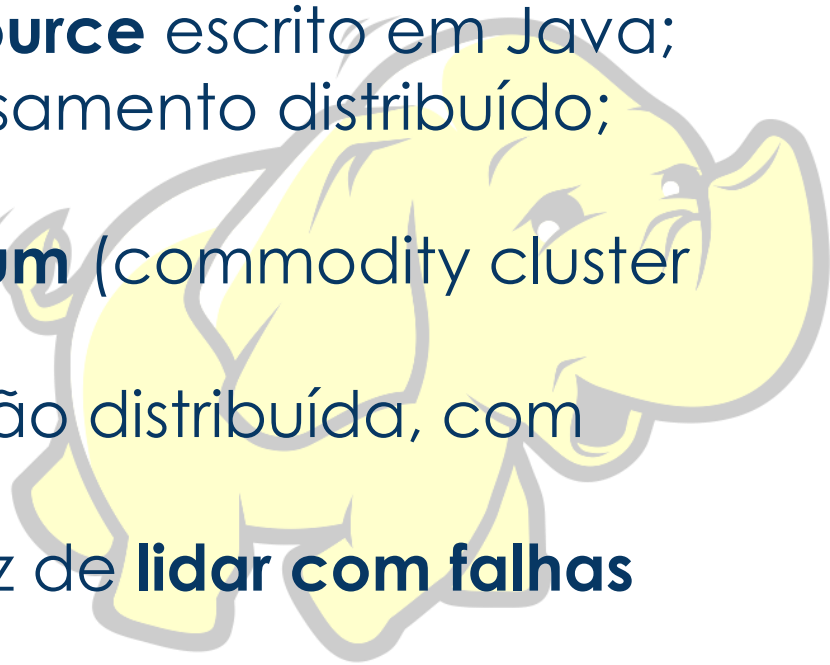
Monitoramento de clusters Hadoop de alto nível com HDFS e Yarn

DIGITAL INNOVATION ONE

Wait, Hadoop what?

UNIVERSITY everis

- Projeto de software **open-source** escrito em Java;
- Escalável, confiável, processamento distribuído;
- "S.O de Big Data";
- Pode utilizar **hardware comum** (commodity cluster computing);
- Framework para computação distribuída, com filesystem distribuído (**HDFS**);
Infraestrutura confiável capaz de **lidar com falhas** (hardware, software, rede).



DIGITAL INNOVATION ONE

Wait, Hadoop what?

UNIVERSITY everis

Distros

Open Source

Apache Hadoop

Commercial Open Source

Cloudera (+Hortonworks)

MapR

AWS ElasticMapReduce

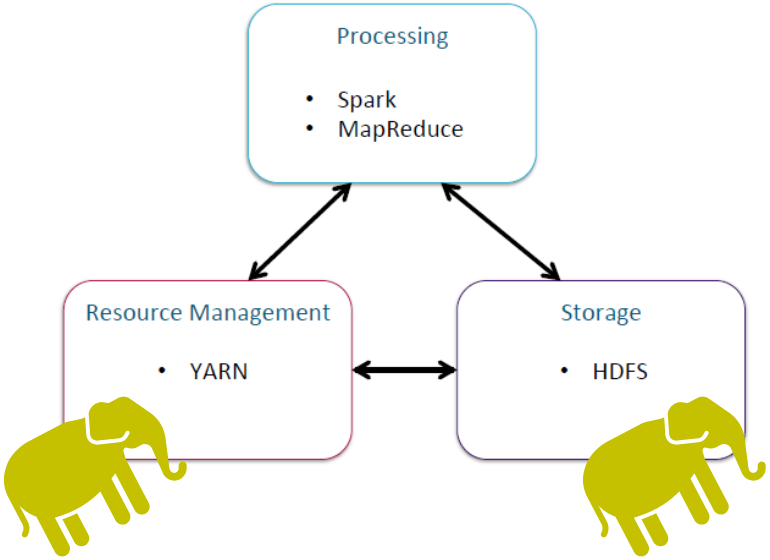
Microsoft HDInsight

DIGITAL INNOVATION ONE

Wait, Hadoop what?

UNIVERSITY everis

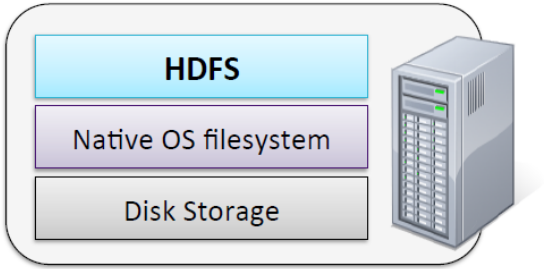
Core Hadoop



HDFS

Hadoop Distributed File System

- Baseado no **Google FS**;
- Escalável e tolerante a falhas;
- Arquivos Texto, sequence file, Parquet, AVRO, ORC...
- Tamanho mínimo de um bloco (default: 128MB);
- Fator de replicação (default: 3).

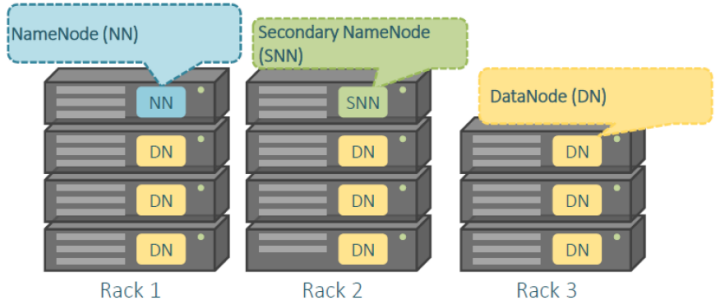


HDFS

NameNode
Gerencia o namespace
Se o Namenode para o cluster fica inacessível

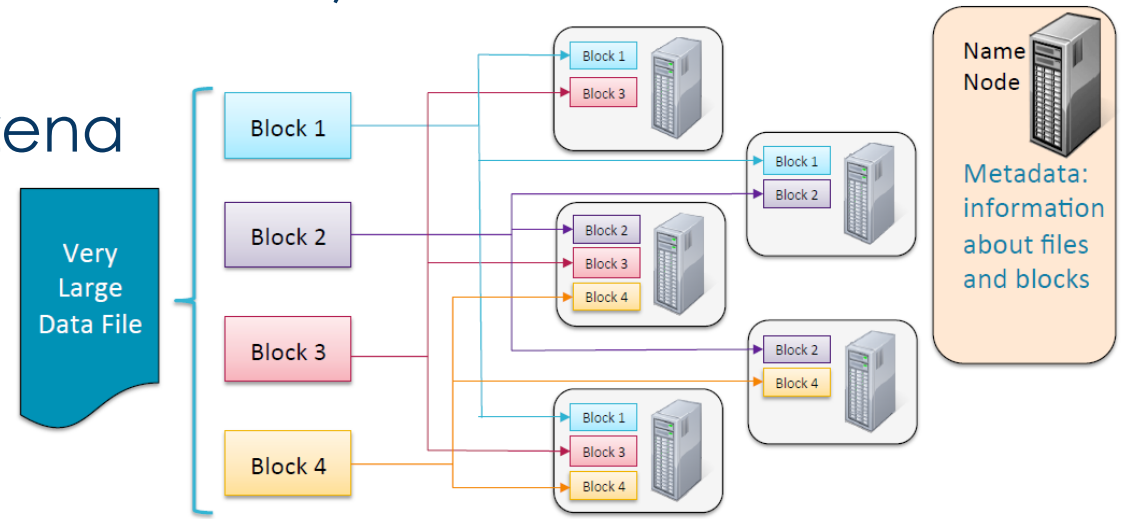
DataNode
Armazena os blocos de arquivos

Secondary NameNode
Oferece tarefas de ponto de verificação e manutenção do Namenode



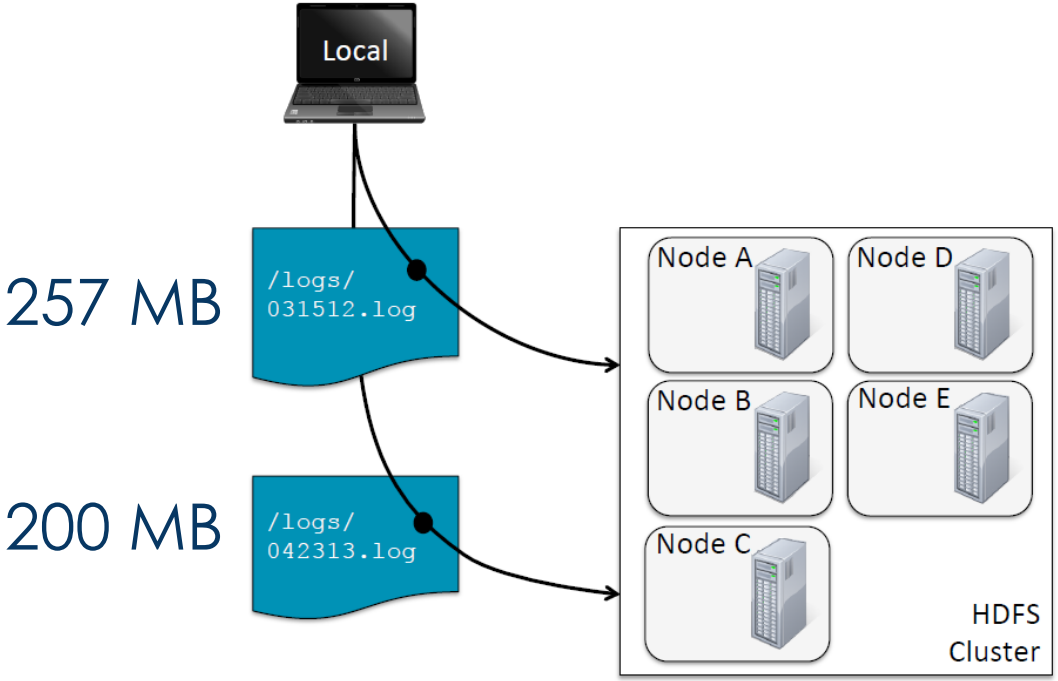
HDFS

- Dados separados em **blocos**;
- **Replicado** em 3;
- **Namenode** armazena os metadados



HDFS

Exemplo:

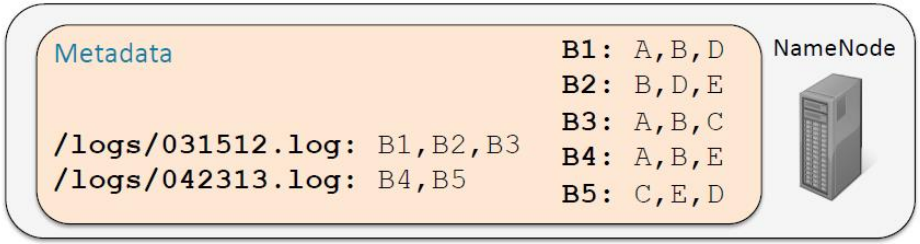


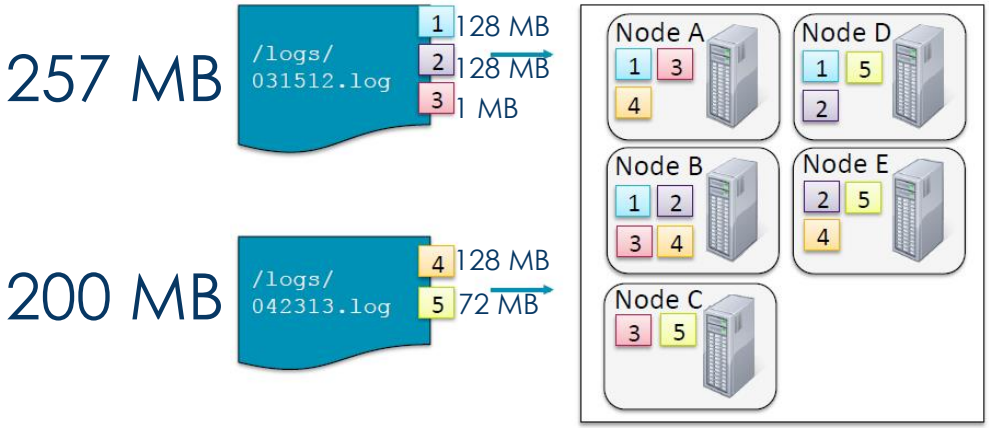
DIGITAL INNOVATION ONE

HDFS

UNIVERSITY everis

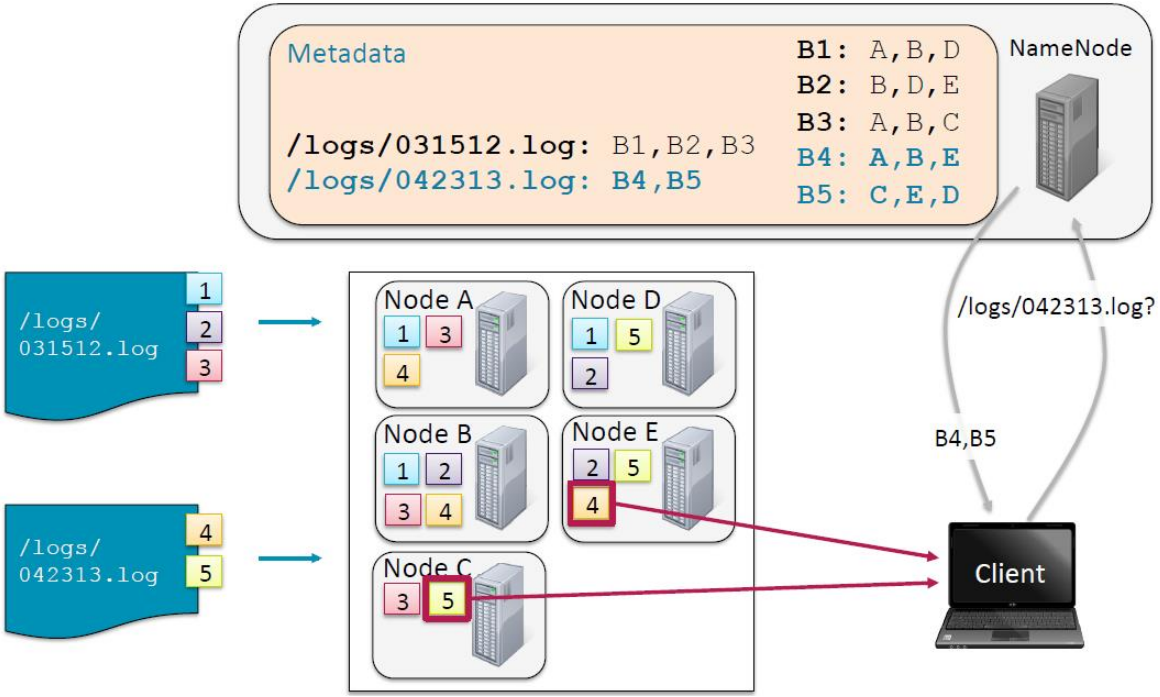
Exemplo:





HDFS

Exemplo:



DIGITAL INNOVATION ONE

HDFS

UNIVERSITY everis

PUT / GET

- Copiar arquivo HDFS para local

```
$ hdfs dfs -get /tmp/file_teste.txt
```

- Ingestão manual

```
$ hdfs dfs -put file_teste.txt /user/everis-bigdata/
```

DIGITAL INNOVATION ONE

HDFS

UNIVERSITY everis

Live Demo

```
$ sudo -u hdfs hdfs dfs -chmod -R 777 /tmp
$ hdfs dfs -ls -h /
$ hdfs dfs -cat /tmp/file_teste.txt |head -10
$ hdfs dfs -rm /tmp/file_teste.txt
$ hdfs dfs -mkdir /tmp/delete
$ hdfs dfs –cp /tmp/file_teste.txt /tmp/delete/
$ hdfs dfs –touchz /tmp/delete/empty_file
$ hdfs dfs -rm -R /tmp/delete
$ hdfs dfs -du -h /user
$ hdfs fsck /tmp/ -files -blocks
```

DIGITAL INNOVATION ONE

Parte 3: YARN

Monitoramento de clusters Hadoop de alto nível com HDFS e Yarn

DIGITAL INNOVATION ONE

YARN

UNIVERSITY everis

Yet Another Resource Negotiator

- Gerenciamento de recursos;
- Gerenciamento e monitoramento de Jobs;
- Recursos dos nós são alocados somente quando requisitado (via container).

DIGITAL INNOVATION ONE

YARN

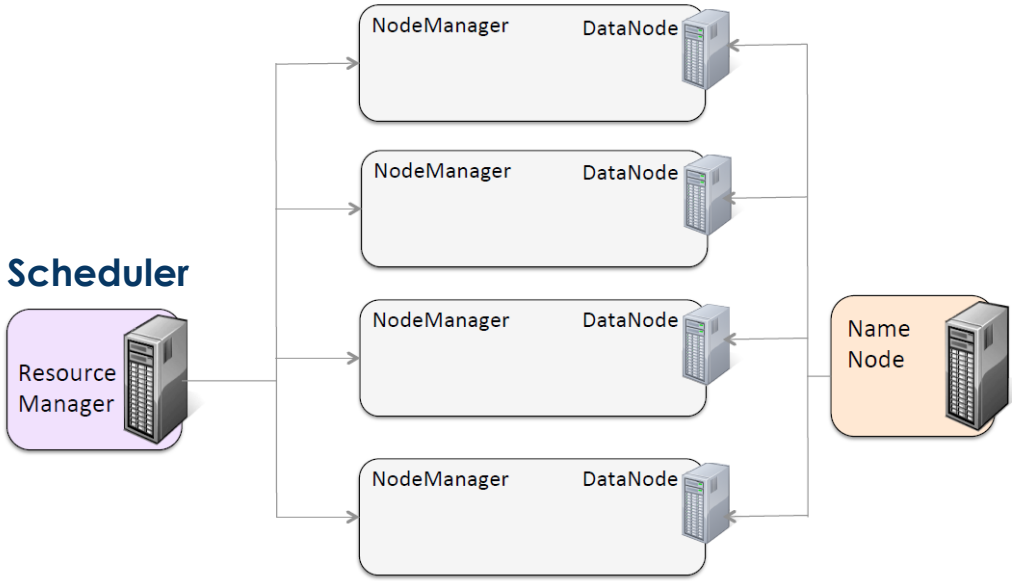
UNIVERSITY everis

Componentes

Application: um job submetido ao Hadoop;
Application Master: gerencia a execução e o escalonamento das tarefas (1 por aplicação);
Container: unidade de alocação de recursos (ex. c1 = 1 GB RAM, 2 CPU);
Resource Manager: gerenciador global de recursos;
Node Manager: gerencia o ciclo de vida e monitora os recursos do Container.

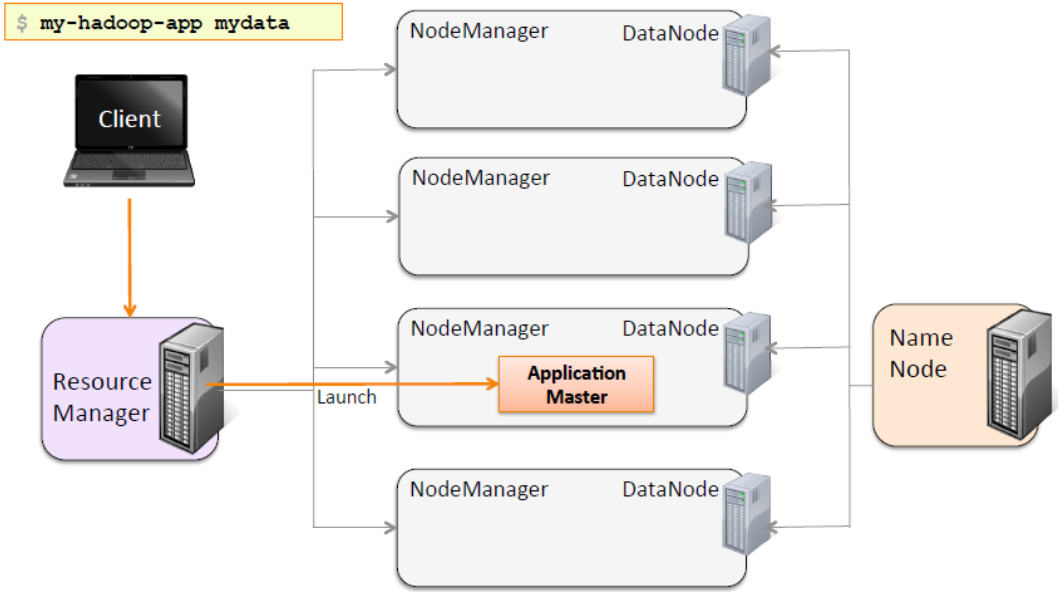
YARN

Execução de aplicação



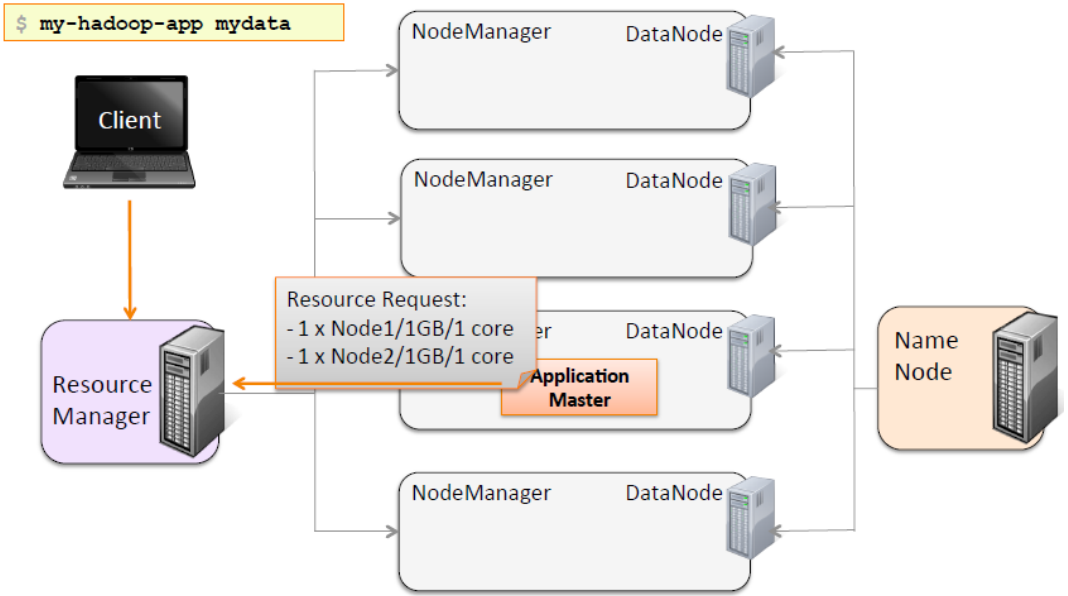
YARN

Execução de aplicação



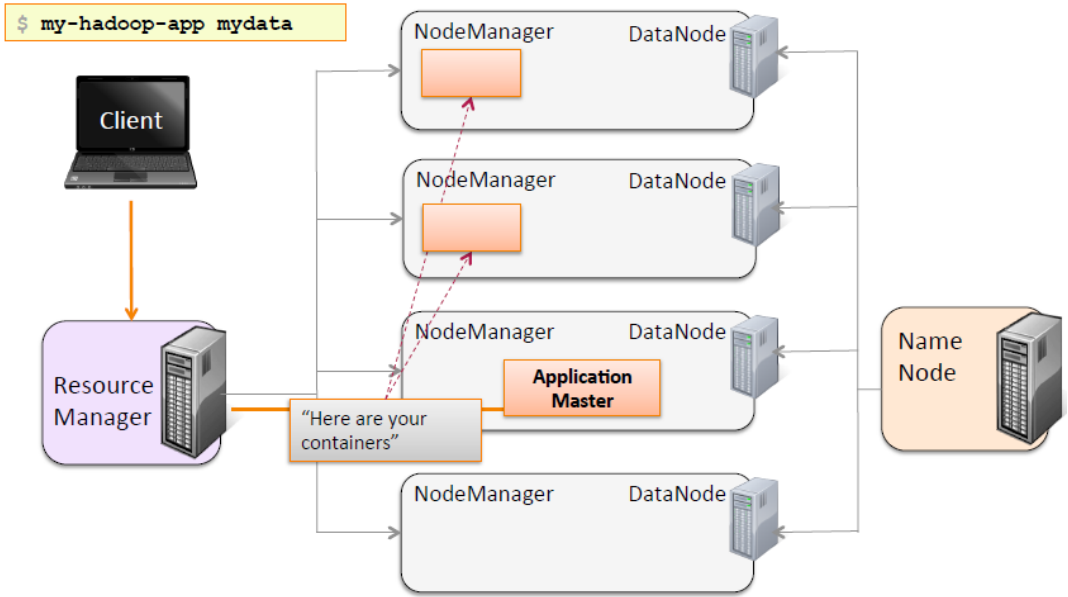
YARN

Execução de aplicação



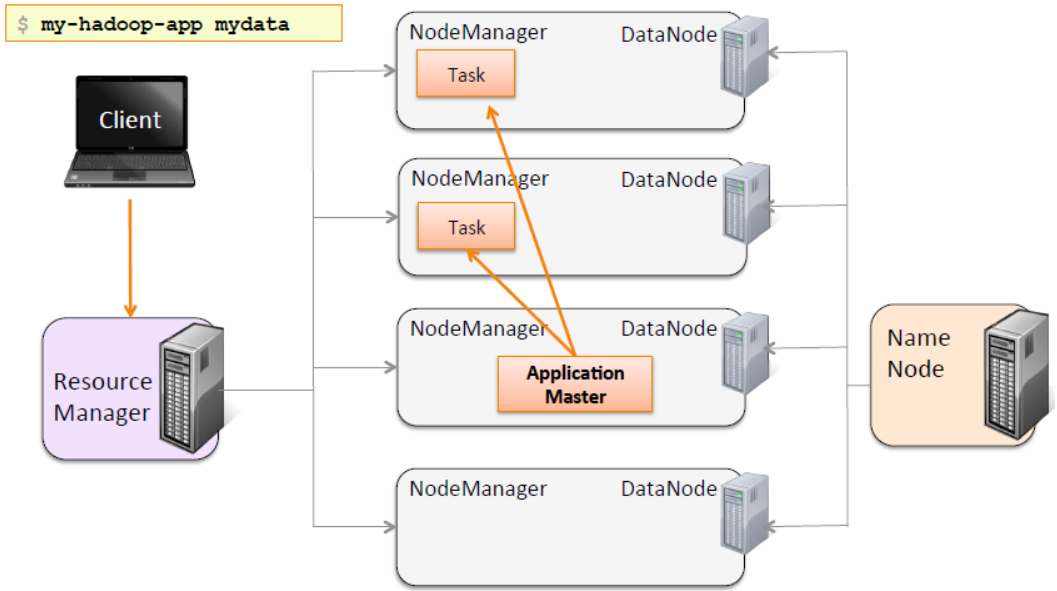
YARN

Execução de aplicação



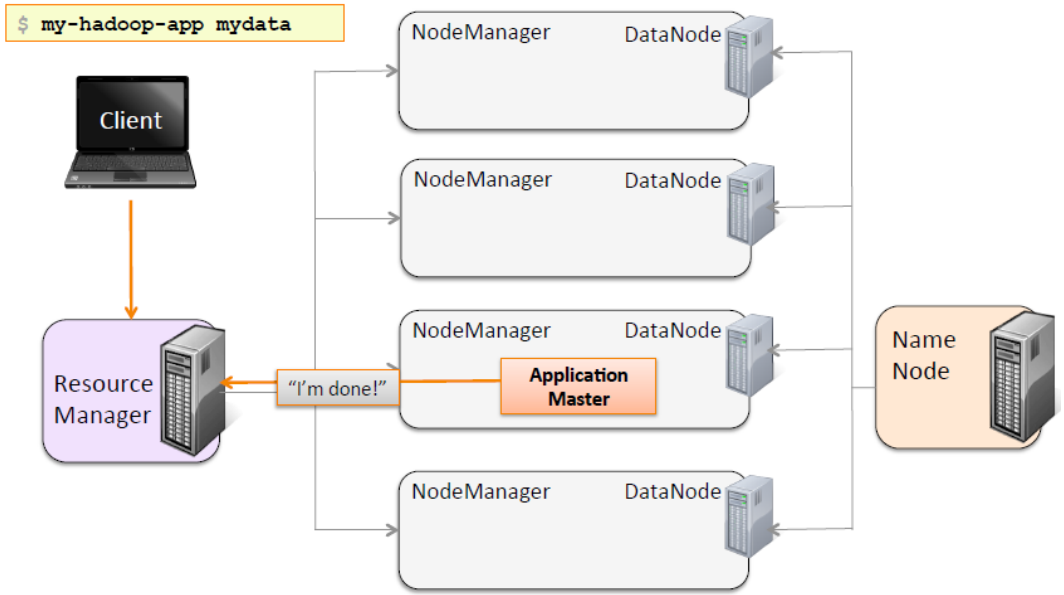
YARN

Execução de aplicação



YARN

Execução de aplicação



DIGITAL INNOVATION ONE

YARN

UNIVERSITY everis

Live Demo

```
$ sudo service hadoop-hdfs-namenode start
$ sudo service hadoop-hdfs-secondarynamenode start
$ sudo service hadoop-hdfs-datanode start
$ sudo service hadoop-mapreduce-historyserver start
$ sudo service hadoop-yarn-resourcemanager start
$ sudo service hadoop-yarn-nodemanager start
```

DIGITAL INNOVATION ONE

YARN

UNIVERSITY everis

Live Demo

```
$ sudo sed -i 's|hdfs://|hdfs://bigdata-srv:8020/|g'
/etc/hadoop/conf/yarn-site.xml

$ sudo –u hdfs yarn jar /usr/lib/hadoop-mapreduce/hadoop-
mapreduce-examples.jar wordcount /tmp/file_teste.txt
/tmp/wc_output
```

YARN

Log

```
$ sudo -u hdfs yarn logs -applicationId
application_1611089476809_0001 |more

$ sudo -u hdfs yarn logs -applicationId
application_1611089476809_0001 > wordcount.log
```

DIGITAL INNOVATION ONE

Resumão

UNIVERSITY everis

HDFS é a camada de armazenamento do Hadoop;
- Divide os dados em blocos e os distribui pelo cluster;
- Os workers rodam o daemon DataNode e o master o daemon NameNode;

MapReduce foi o primeiro framework de computação distribuída utilizado com o HDFS;
- Levou o processamento aos servidores onde o dado está armazenado.

DIGITAL INNOVATION ONE

Resumão

UNIVERSITY everis

YARN gerencia os recursos no cluster
- Trabalha com o HDFS para executar as tarefas quando o dado é armazenado;
- Os workers rodam o daemon "NodeManager" e o master o daemon "ResourceManager";
- É possível monitorar os jobs através da porta 8088.

DIGITAL
INNOVATION
ONE

Dúvidas?

Monitoramento de clusters Hadoop de alto nível com HDFS e Yarn

DIGITAL INNOVATION ONE

Referências úteis

UNIVERSITY everis

https://repositorio.ufscar.br/bitstream/handle/ufscar/534/5351.pdf?sequence=1&isAllowed=y

https://alissonmachado.com.br/hadoop-cluster/

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

https://azure.microsoft.com/pt-br/services/hdinsight/