

Hadoop

Hortonworks Data Platform (HDP)

Sandbox

Victor Manuel Mondragon M. Phd.

Sesión 8 - Lab

Noviembre 30 de 2020

- 1. Conceptos ecosistema Hadoop**
- 2. Archivos Distribuidos HDFS**
- 3. Procesamiento Distribuido con MapReduce**
- 4. Herramientas Hadoop para procesar Datos**

HDFS: Sistema de Archivos Distribuido

MapReduce: Framework de procesamiento distribuido.

En 2008 como proyecto de código abierto a la fundación Apache.

¿ Que es Hadoop?

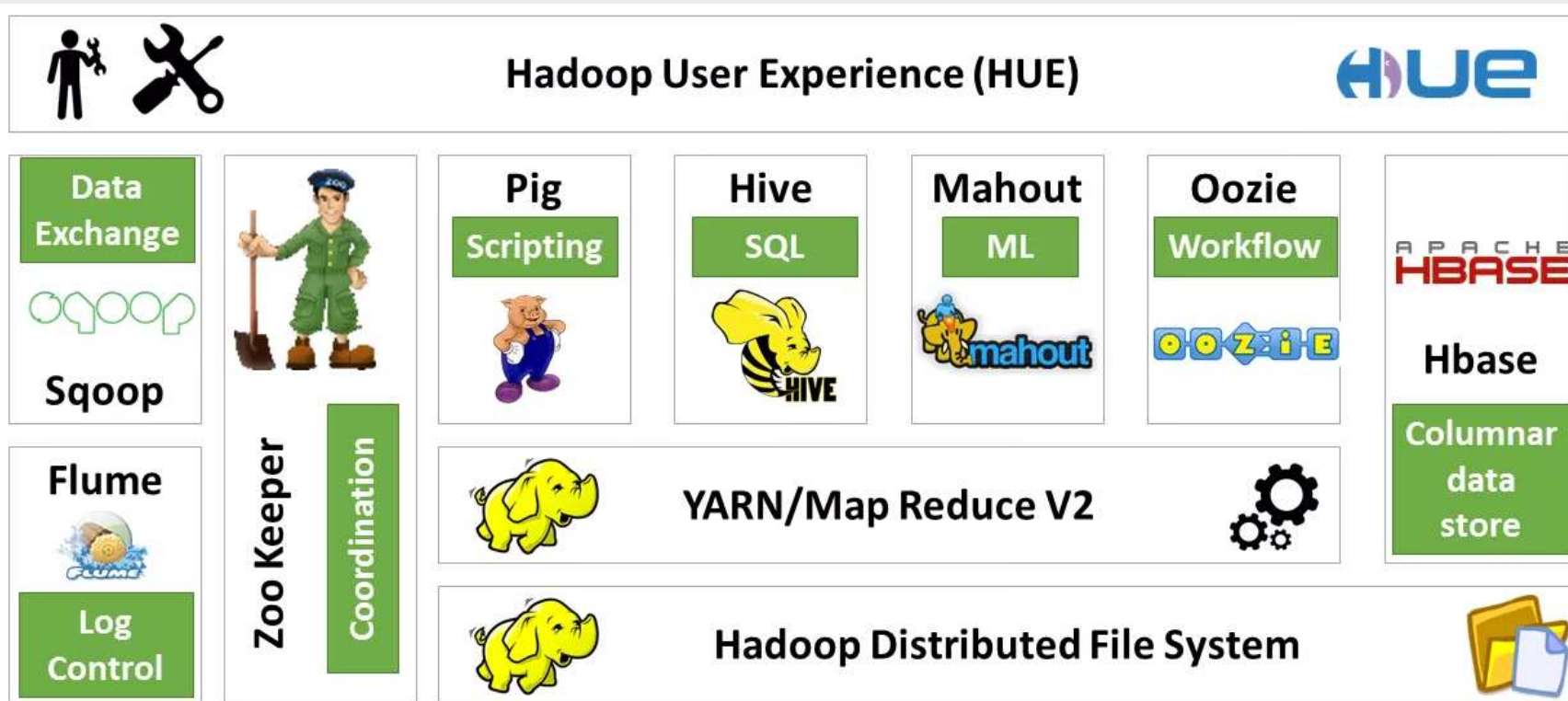
Software de código abierto que sirve para almacenar y procesar datos de forma distribuida.

¿Por qué utilizar Hadoop?

- Económico
- Escalabilidad Horizontal (mas ordenadores al Clúster)
 - Crece indefinidamente
- Transparencia en el Clúster
 - Igual en 2 nodos o 200
- Redundante.
 - Cada nodo esta replicado en tres(3) veces por defecto
 - Se puede escoger el nivel de replicación.
- Código Abierto.
 - Contribuciones
- Guarda estructurada o no estructurada

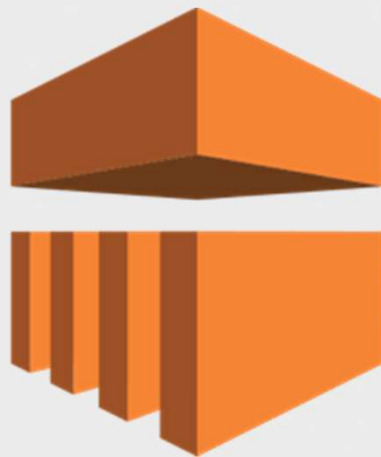
Conceptos Ecosistema Hadoop II

Ecosistema de Big Data



¿Qué es Amazon EMR?

Amazon EMR es una plataforma de clúster administrada que simplifica la ejecución de los marcos de trabajo de Big Data, tales como [Apache Hadoop](#) y [Apache Spark](#) en AWS para procesar y analizar grandes cantidades de datos.



amazon
EMR

Arquitectura de Amazon EMR

Librerías

Pig



EMR



EMR



Motores de
Procesamiento



EMR



EMR

Samza

Gestores de
aplicación de
Recursos

EMR

YARN



MESOS

Ingesta y
almacenamiento
de datos



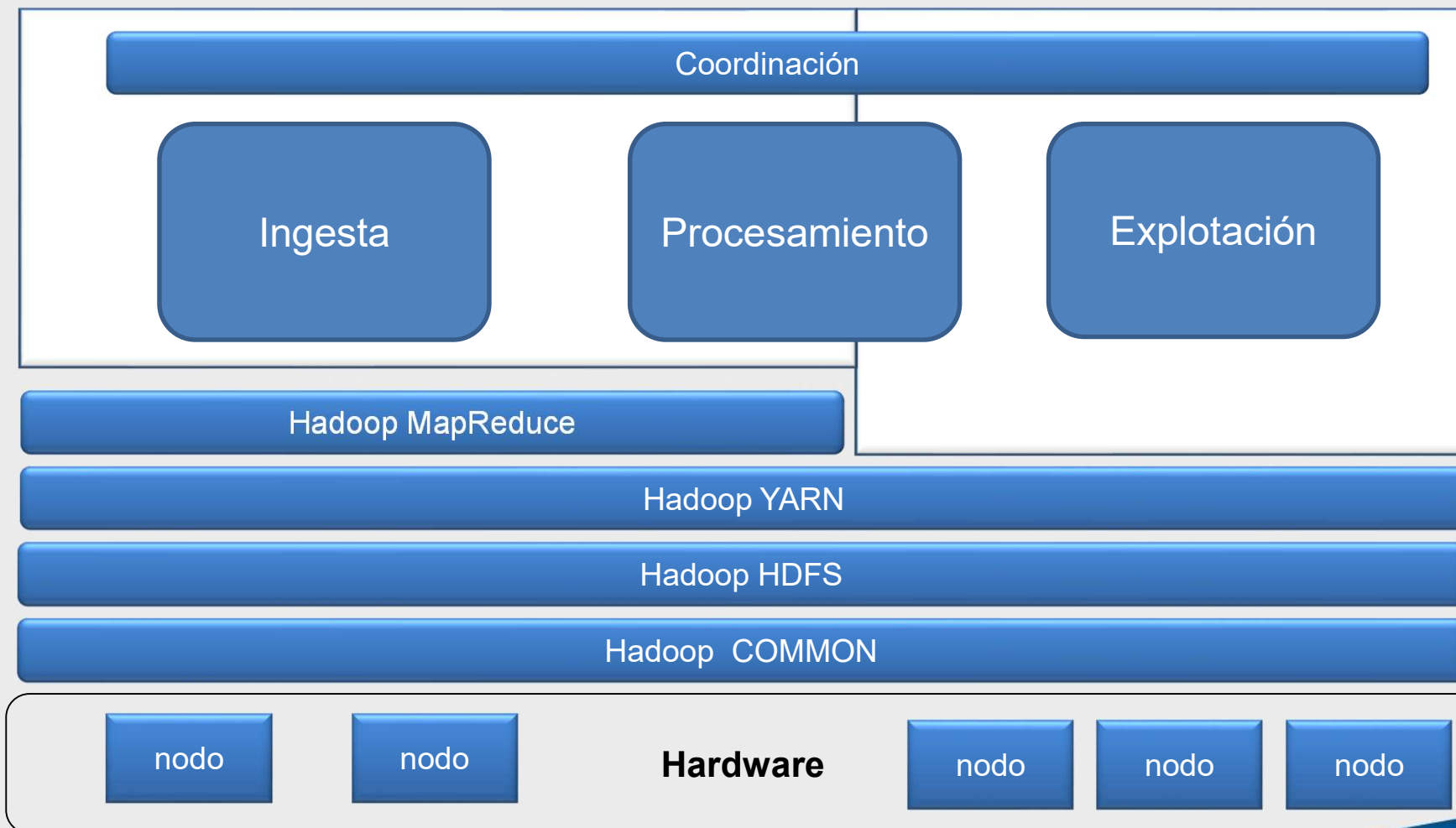
EMR



EMR

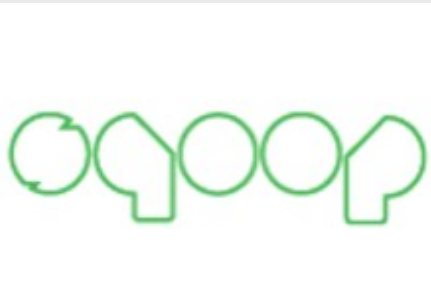


Ecosistema Hadoop



Herramientas del Ecosistema

Ingesta, Procesamiento y Explotación

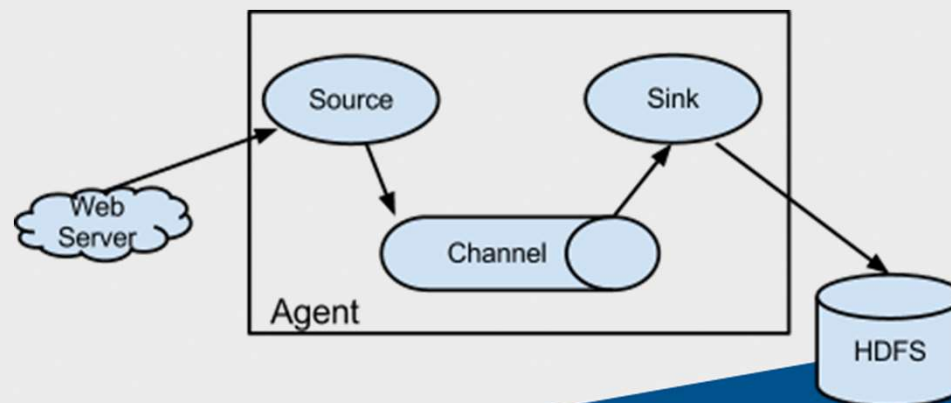


Sqoop es una herramienta cuya principal funcionalidad es **importar** datos entre bases de datos relacionales o Data Warehouse y Hadoop.

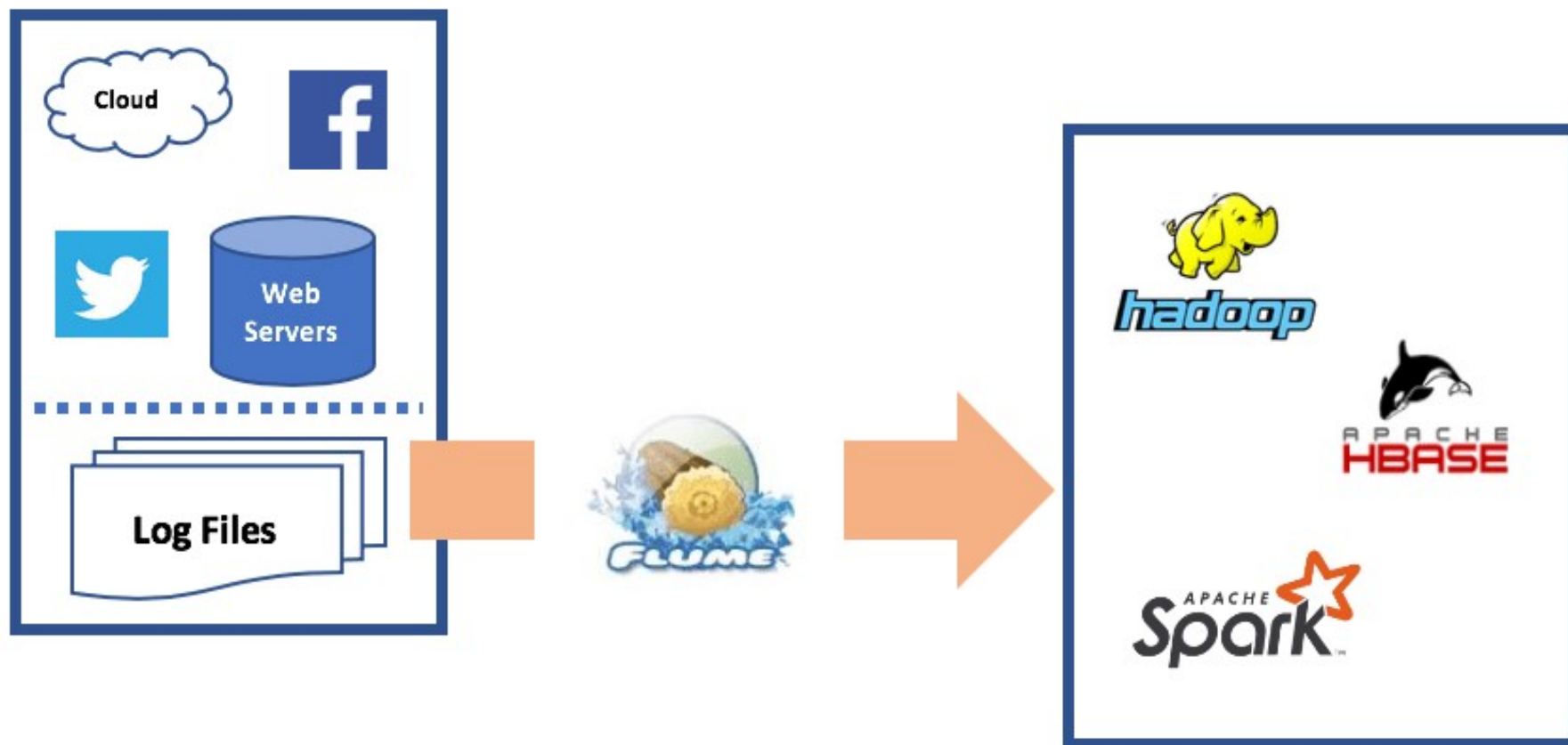


Flume es un servicio distribuido que permite ingestar datos en tiempo real.

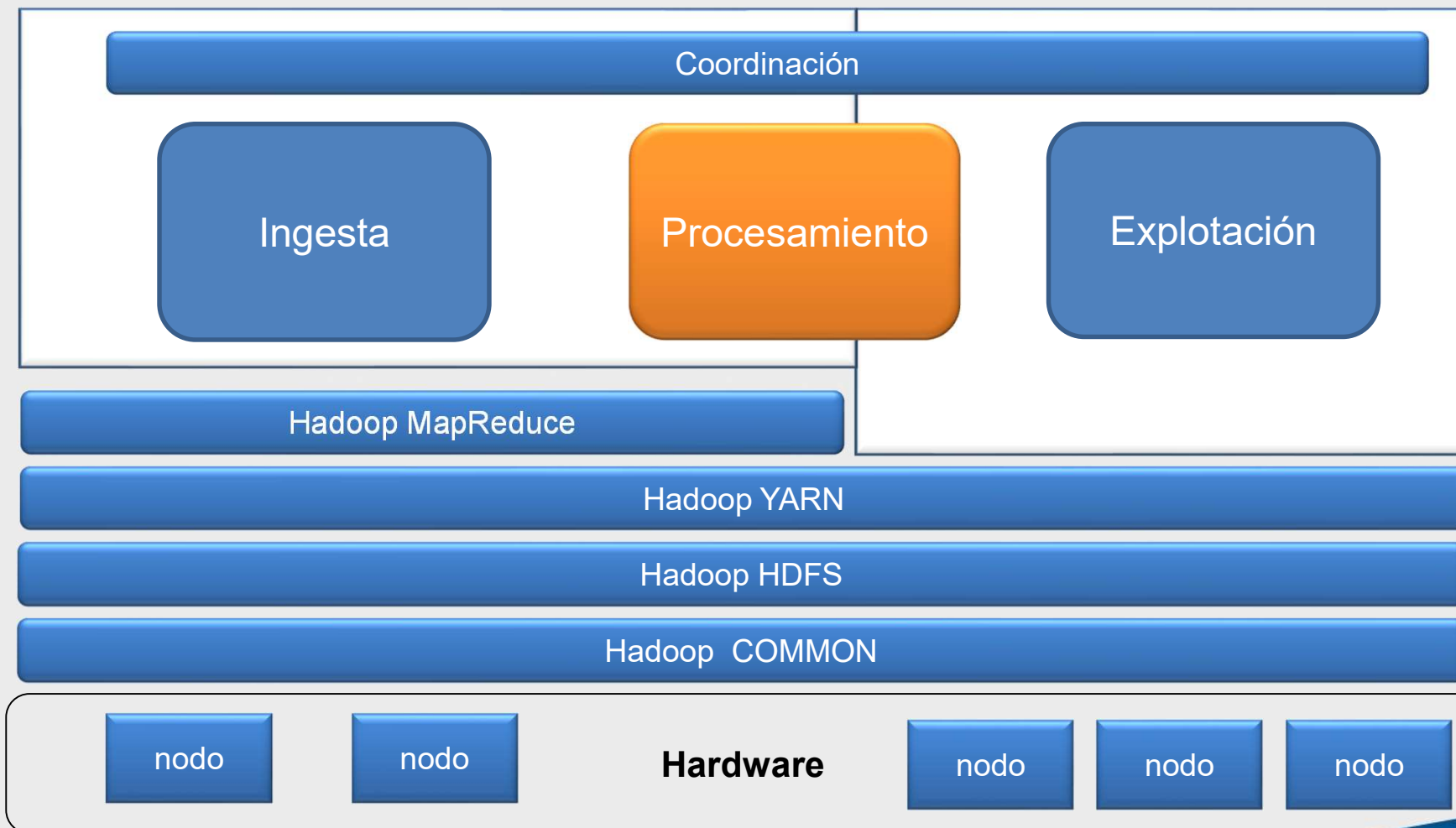
Herramienta de recopilación de datos de transmisión altamente confiable, distribuida y configurable.



Ingesta – Ejemplo Flume

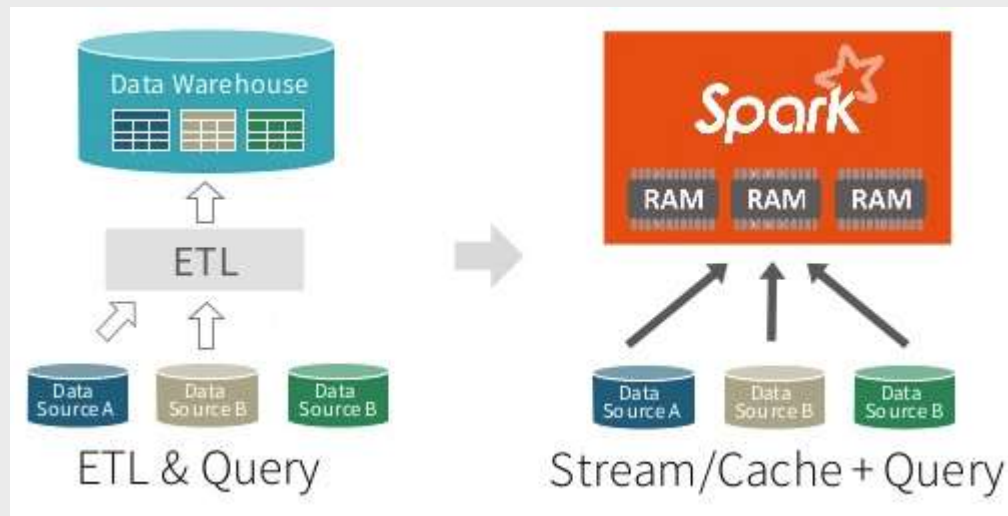


Ecosistema Hadoop



Procesamiento - SPARK

SPARK es una herramienta de procesamiento de datos que hace un uso intensivo de la memoria por lo que sus cálculos suelen ser mucho más rápidos que los demás medios

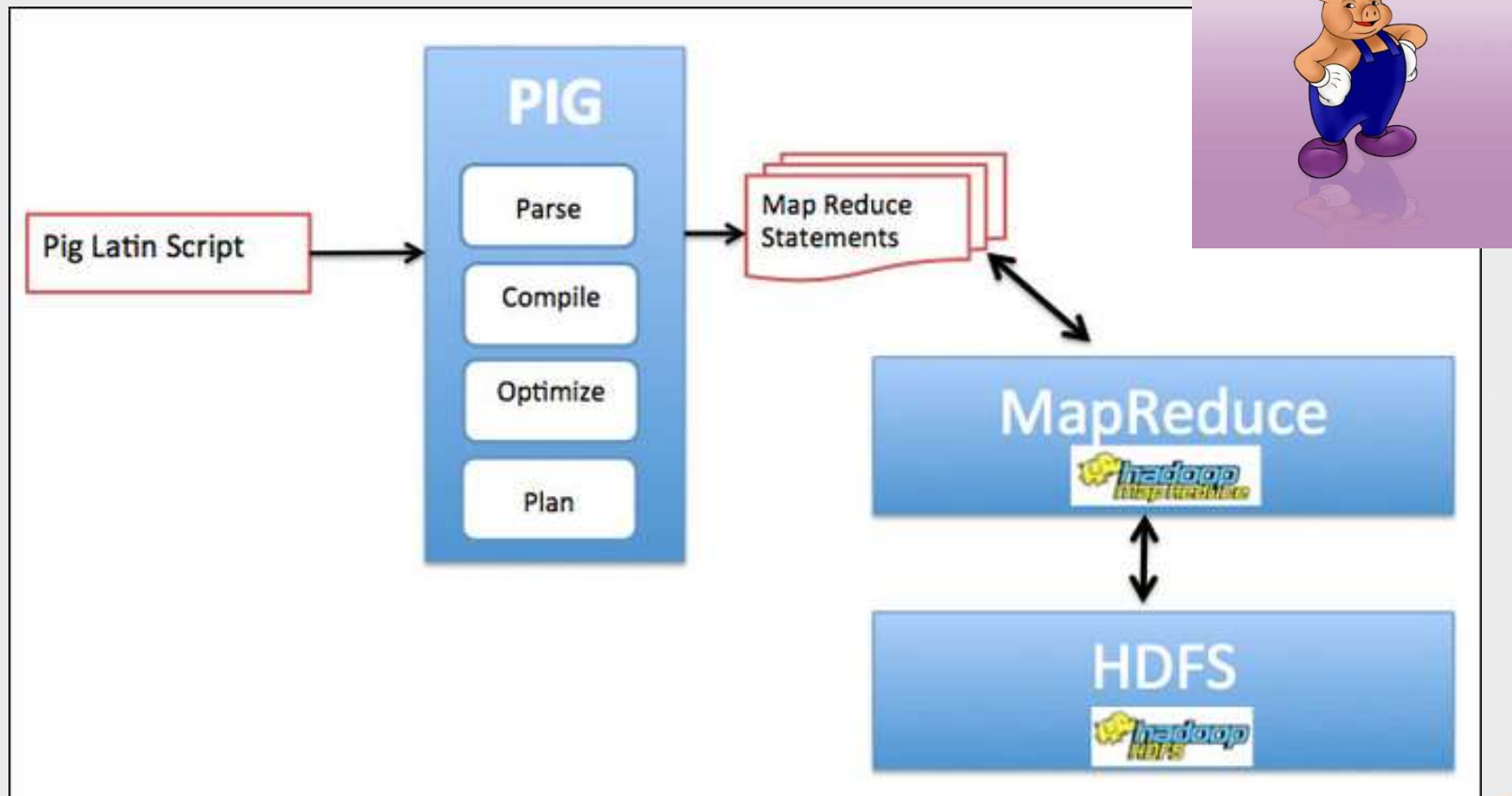


SPARK es un motor ultrarrápido para el almacenamiento, procesamiento y análisis de grandes volúmenes de datos.

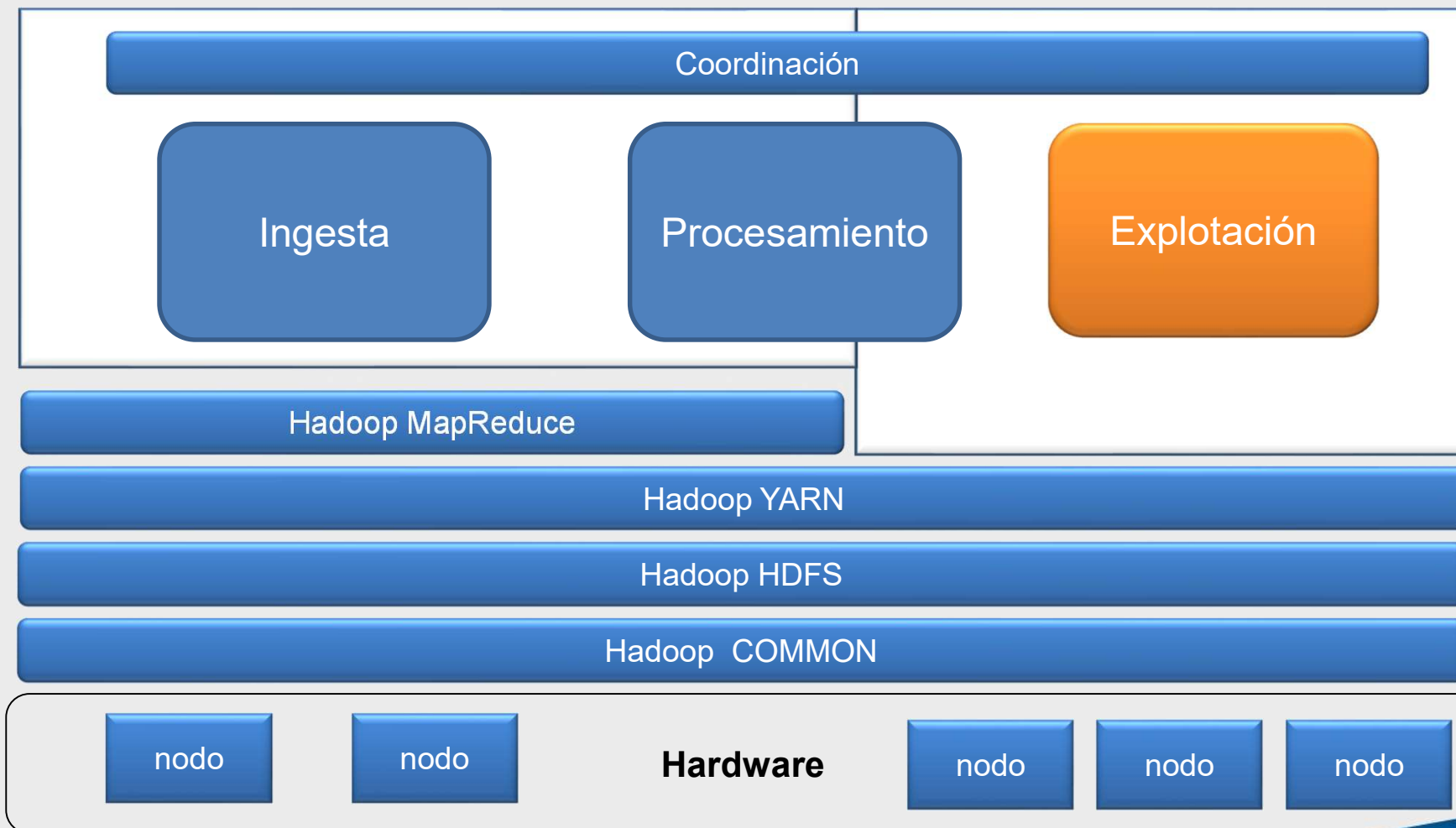
100 veces mas rápido que MapReduce

Procesamiento - Pig

Pig es una herramienta permite escribir scripts rápidos para hacer limpieza y procesamiento de datos.



Ecosistema Hadoop



Explotación entendemos todo aquello que tiene que ver con el uso final de la información procesada.

Impala: permite realizar consultas en un lenguaje muy similar al **SQL** de las bases de datos tradicionales las consultas en Impala suelen ser muy **rápidas** por lo que es posible trabajar con esta herramienta de manera **interactiva**.

Hive: permite realizar consultas en un lenguaje muy similar al **SQL** de las bases de datos tradicionales, Consultas de **procesamiento**, son mas lentas pero robustas.

HBASE: una base de datos no SQL orientada a columnas, suele utilizarse para servir los datos aplicaciones externas ya que permite leer y escribir información con latencias muy bajas

Laboratorio Azure Hortonworks Data Platform (HDP) Sandbox

Caso practico
Diseñar un almacén de datos Hadoop

Crear una cuenta Azure

<https://azure.microsoft.com/en-us/free/students/>

<https://azure.microsoft.com/en-us/free/>

<https://medium.com/@sukhmanjawa/100-credits-with-azure-for-students-without-credit-card-94a09986e0b4>

Azure - Hortonworks Data Platform

← → ↻ 🔒 portal.azure.com

☰ Microsoft Azure

🔍 Search resources, services, and docs (G+)

Home > Hortonworks Data Platform (HDP) Sandbox

Hortonworks Data Platform (HDP) Sandbox

Hortonworks



Hortonworks Data Platform (HDP) Sandbox

♡ Save for later

Hortonworks

Create

Start with a pre-set configuration

Want to deploy programmatically? [Get started](#)

Overview Plans

About To Deploy?

For a step-by-step guide on how to deploy the Hortonworks Sandbox on Azure, visit: [Deploying Hortonworks Sandbox on Microsoft Azure](#).

Already Set Up and Looking to Learn?

There are a series of tutorials to get you going with HDP fast. To learn more about the HDP Sandbox check out: [Learning the Ropes of the Hortonworks HDP Sandbox](#). To get started using Hadoop to store, process and query data try this HDP 2.6 tutorial series: [Hello HDP an introduction to Hadoop](#)

Have Questions?

For all your Hadoop and Big Data questions, and to get answers directly from the pros fast, visit: [Hortonworks Community Connection](#)

Learn More

- [Browse: Big Data Tutorials](#)
- Tutorial: [Deploying Hortonworks Sandbox on Microsoft Azure](#)
- Tutorial: [Learning the Ropes of the Hortonworks Sandbox](#)

Create a virtual machine

Create a virtual machine

Basics Disks Networking Management Advanced Tags Review + create

Create a virtual machine that runs Linux or Windows. Select an image from Azure marketplace or use your own customized image.
Complete the Basics tab then Review + create to provision a virtual machine with default parameters or review each tab for full customization.

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * ⓘ

Azure para estudiantes

Resource group * ⓘ

[Create new](#)

Instance details

Virtual machine name * ⓘ

Region * ⓘ

Availability options ⓘ

Image * ⓘ

Azure Spot instance ⓘ

A resource group is a container that holds related resources for an Azure solution.

Name *

recurso_tadeo2

OK

Cancel

Create a virtual machine

TADEOBIGDATA ✓

(US) East US 2 ✓

No infrastructure redundancy required ✓

Hortonworks Sandbox with HDP 2.6.4 ✓

[Browse all public and private images](#)

☐ Yes ☒ No

Standard B2s

2 vcpus, 4 GiB memory (\$30.37/month)

[Change size](#)

☒ Password ☐ SSH public key

vmondragon ✓

..... ✓

..... ✓

☒ Password ☐ SSH public key

vmondragon ✓

..... ✓

..... ✓

ts are accessible from the public internet. You can specify more limited or granular

☐ None ☒ Allow selected ports

HTTP (80), HTTPS (443), SSH (22) ^

- ☒ HTTP (80)
- ☒ HTTPS (443)
- ☒ SSH (22)

Create a virtual machine

Create a virtual machine

Basics Disks **Networking** Management Advanced Tags Review + create

Define network connectivity for your virtual machine by configuring network interface card (NIC) settings. You can control ports, inbound and outbound connectivity with security group rules, or place behind an existing load balancing solution. [Learn more](#)

Network interface

When creating a virtual machine, a network interface will be created for you.

Virtual network * ⓘ

(new) recurso_tadeo2-vnet
[Create new](#)

Subnet * ⓘ

(new) default (10.0.1.0/24)
[Create new](#)

Public IP ⓘ

(new) TADEOBIGDATA-ip
[Create new](#)

NIC network security group ⓘ

☐ None ☒ Basic ☐ Advanced

Public inbound ports * ⓘ

☐ None ☒ Allow selected ports

Select inbound ports *

HTTP (80), HTTPS (443), SSH (22)
[Create new](#)

⚠ This will allow all IP addresses to access your virtual machine. This is only recommended for testing. Use the Advanced controls in the Networking tab to create rules to limit inbound traffic to known IP addresses.

Name *

TADEOBIGDATA-ip

SKU ⓘ

☐ Basic ☒ Standard



Assignment ⓘ

☒ Static

Availability zone

Zone-redundant

Create a virtual machine

 Connect  Start  Restart  Stop  Capture  Delete  Refresh

Resource group ([change](#)) : [recurso_tadeo2](#)

Status : Running

Location : East US 2

Subscription ([change](#)) : [Azure para estudiantes](#)

Subscription ID : b7c53e1b-bc30-47d9-81a9-55582ae3e612

Computer name : TADEOBIGDATA

Operating system : Linux (centos 7.4.1708)

Size : Standard B2s (2 vcpus, 4 GiB memory)

Tags ([change](#)) : [Click here to add tags](#)



Azure Spot : N/A

Public IP address : [52.177.9.36](#)

Private IP address : 10.0.1.4

Public IP address (IPv6) : -

Private IP address (IPv6) : -

Virtual network/subnet : [recurso_tadeo2-vnet/default](#)

DNS name : [Configure](#)



Reference & email address



Preferred phone number *



Create

< Previous

Next >

[Download a template for automation](#)

Create a virtual machine

[Connect](#)
[Start](#)
[Restart](#)
[Stop](#)
[Capture](#)
[Delete](#)
[Refresh](#)

Resource group (change)	: recurso_tadeo2	Azure Spot	: N/A
Status	: Running	Public IP address	: 52.177.9.36
Location	: East US 2	Private IP address	: 10.0.1.4
Subscription (change)	: Azure para estudiantes	Public IP address (IPv6)	: -
Subscription ID	: b7c53e1b-bc30-47d9-81a9-55582ae3e612	Private IP address (IPv6)	: -
Computer name	: TADEOBIGDATA	Virtual network/subnet	: recurso_tadeo2-vnet/default
Operating system	: Linux (centos 7.4.1708)	DNS name	: Configure
Size	: Standard B2s (2 vcpus, 4 GiB memory)		
Tags (change)	: Click here to add tags		

Network Interface: [tadeobigdata713](#)
[Effective security rules](#)
[Topology](#)

Virtual network/subnet: [recurso_tadeo2-vnet/default](#)
 NIC Public IP: [52.177.9.36](#)
 NIC Private IP: [10.0.1.4](#)
 Accelerated networking: **Disabled**

[Inbound port rules](#)
[Outbound port rules](#)
[Application security groups](#)
[Load balancing](#)

Network security group [TADEOBIGDATA-nsg](#) (attached to network interface: [tadeobigdata713](#))
 Impacts 0 subnets, 1 network interfaces

[Add inbound port rule](#)

Priority	Name	Port	Protocol	Source	Destination	Action	
300	SSH	22	TCP	Any	Any	Allow	...
320	HTTP	80	TCP	Any	Any	Allow	...
340	HTTPS	443	TCP	Any	Any	Allow	...
65000	AllowVnetInBound	Any	Any	VirtualNetwork	VirtualNetwork	Allow	...
65001	AllowAzureLoadBalancerInBound	Any	Any	AzureLoadBalancer	Any	Allow	...
65500	DenyAllInBound	Any	Any	Any	Any	Deny	...

Agregar puestos de acceso 8080

Attach network interface Detach network interface

IP configuration ⓘ
ipconfig1 (Primary) ▼

Network Interface: tadeobigdata713 [Effective security rules](#) [Topology](#)
Virtual network/subnet: [recurso_tadeo2-vnet/default](#) NIC Public IP: **52.177.9.36** NIC Private IP: **10.0.1.4** Acceleration: Off

Inbound port rules Outbound port rules Application security groups Load balancing

Network security group **TADEOBIGDATA-nsg** (attached to network interface: [tadeobigdata713](#))
Impacts 0 subnets, 1 network interfaces

Priority	Name	Port	Protocol
300	SSH	22	TCP
320	HTTP	80	TCP
340	HTTPS	443	TCP
65000	AllowVnetInBound	Any	Any
65001	AllowAzureLoadBalancerInBound	Any	Any
65500	DenyAllInBound	Any	Any

Basic

Source * ⓘ
Any ▼

Source port ranges * ⓘ
*

Destination * ⓘ
Any ▼

Destination port ranges * ⓘ
8080

Protocol *
Any TCP UDP ICMP

Action *
Allow Deny

Priority * ⓘ
350

Name *
Port_8080

Description
PUERTO DE ACCESO AMBARI ✓

Add

Agregar puestos de acceso 4200

Attach network interface Detach network interface

IP configuration ⓘ
ipconfig1 (Primary) ▼

Network Interface: [tadeobigdata713](#) [Effective security rules](#) [Topology](#)
Virtual network/subnet: [recurso_tadeo2-vnet/default](#) NIC Public IP: **52.177.9.36** NIC Private IP: **10.0.1.4** Acceleration: **On**

Inbound port rules Outbound port rules Application security groups Load balancing

Network security group **TADEOBIGDATA-nsg** (attached to network interface: [tadeobigdata713](#))
Impacts 0 subnets, 1 network interfaces

Priority	Name	Port	Protocol
300	SSH	22	TCP
320	HTTP	80	TCP
340	HTTPS	443	TCP
65000	AllowVnetInBound	Any	Any
65001	AllowAzureLoadBalancerInBound	Any	Any
65500	DenyAllInBound	Any	Any

Source * ⓘ
Any ▼

Source port ranges * ⓘ
*

Destination * ⓘ
Any ▼

Destination port ranges * ⓘ
4200 ✓

Protocol *
Any TCP UDP ICMP

Action *
Allow Deny

Priority * ⓘ
350

Name *
Port_4200 ✓

Description
Aceso a Consola ✓

Add

Puertos de Acceso en la Red

Attach network interface
Detach network interface

IP configuration ⓘ
ipconfig1 (Primary)

Network Interface: [inteligencia894](#)
[Effective security rules](#)
[Topology](#)

Virtual network/subnet: [Recurso_Tadeo-vnet/default](#)
NIC Public IP: **52.167.219.23**
NIC Private IP: **10.0.0.5**
Accelerated networking: **Disabled**

[Inbound port rules](#)
[Outbound port rules](#)
[Application security groups](#)
[Load balancing](#)

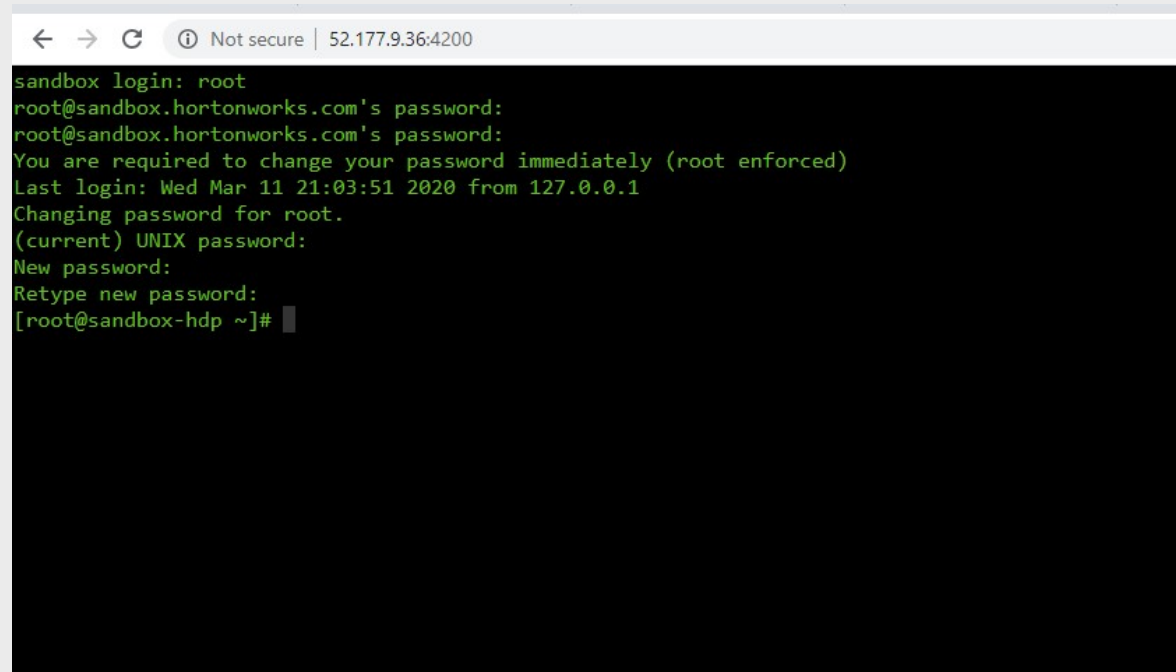
Network security group [INTELIGENCIA-nsg](#) (attached to network interface: [inteligencia894](#))
Impacts 0 subnets, 1 network interfaces

Add inbound port rule

Priority	Name	Port	Protocol	Source	Destination	Action	
300	SSH	22	TCP	Any	Any	Allow	...
310	Port_80	80	Any	Any	Any	Allow	...
320	Port_443	443	Any	Any	Any	Allow	...
330	Port_4220	4200	Any	Any	Any	Allow	...
340	Port_8080	8080	Any	Any	Any	Allow	...
65000	AllowVnetInBound	Any	Any	VirtualNetwork	VirtualNetwork	Allow	...
65001	AllowAzureLoadBalancerInBound	Any	Any	AzureLoadBalancer	Any	Allow	...
65500	DenyAllInBound	Any	Any	Any	Any	Deny	...

<http://52.177.9.36:4200>

La primera vez, ingrese user: root y passwor: hadoop



```
← → ↻ ⓘ Not secure | 52.177.9.36:4200
sandbox login: root
root@sandbox.hortonworks.com's password:
root@sandbox.hortonworks.com's password:
You are required to change your password immediately (root enforced)
Last login: Wed Mar 11 21:03:51 2020 from 127.0.0.1
Changing password for root.
(current) UNIX password:
New password:
Retype new password:
[root@sandbox-hdp ~]#
```

<http://52.177.9.36:4200>

- La primera vez, ingrese user: **root** y passwor: **hadoop**
- Ejecutar el comando: **ambari-admin-password-reset**
- **Cambiar la interfaz de acceso AMBARI WEB**

```
← → ↻ ⓘ Not secure | 52.177.9.36:4200
[root@sandbox-hdp ~]# ambari-admin-password-reset
Please set the password for admin:
Please retype the password for admin:

The admin password has been set.
Restarting ambari-server to make the password change effective...

Using python /usr/bin/python
Restarting ambari-server
Waiting for server stop...
Ambari Server stopped
Ambari Server running with administrator privileges.
Organizing resource files at /var/lib/ambari-server/resources...
```

<http://52.177.9.36:8080>

52.177.9.36:8080/#/login



Ambari

Sign in

Username

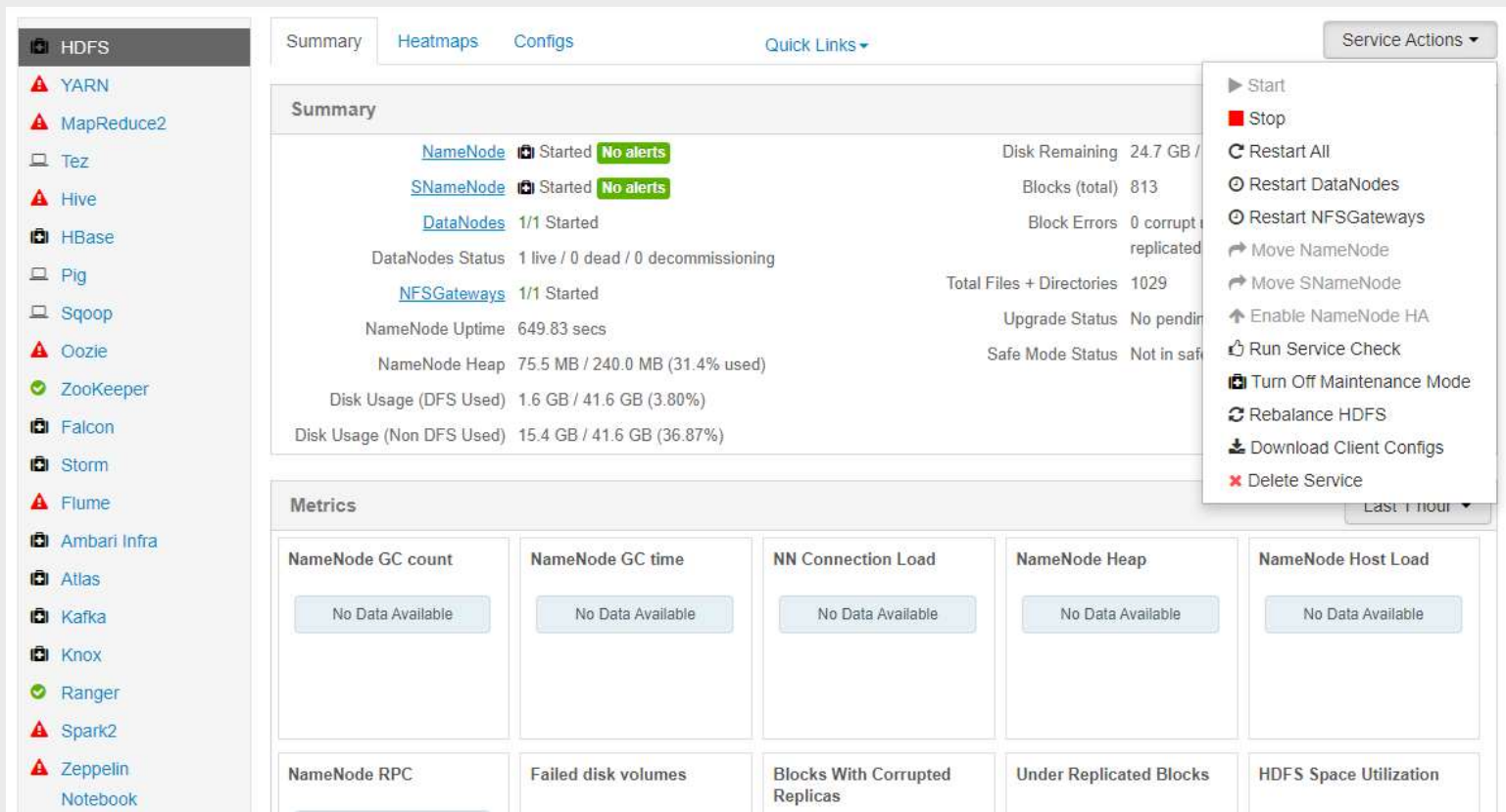
admin

Password

.....

Sign in

Activar HDFS – Modo mantenimiento OFF



The screenshot shows the Ambari Sandbox interface for the HDFS service. The left sidebar lists various services, with HDFS selected. The main panel displays the 'Summary' tab, showing the status of NameNode, SNameNode, DataNodes, and NFSGateways. A 'Service Actions' dropdown menu is open, highlighting the 'Turn Off Maintenance Mode' option.

Summary

- [NameNode](#) Started **No alerts** Disk Remaining 24.7 GB / 41.6 GB
- [SNameNode](#) Started **No alerts** Blocks (total) 813
- [DataNodes](#) 1/1 Started Block Errors 0 corrupted / 0 replicated
- DataNodes Status 1 live / 0 dead / 0 decommissioning Total Files + Directories 1029
- [NFSGateways](#) 1/1 Started Upgrade Status No pending upgrades
- NameNode Uptime 649.83 secs Safe Mode Status Not in safe mode
- NameNode Heap 75.5 MB / 240.0 MB (31.4% used)
- Disk Usage (DFS Used) 1.6 GB / 41.6 GB (3.80%)
- Disk Usage (Non DFS Used) 15.4 GB / 41.6 GB (36.87%)

Metrics













NameNode GC count	NameNode GC time	NN Connection Load	NameNode Heap	NameNode Host Load
No Data Available	No Data Available	No Data Available	No Data Available	No Data Available

NameNode RPC	Failed disk volumes	Blocks With Corrupted Replicas	Under Replicated Blocks	HDFS Space Utilization
No Data Available	No Data Available	No Data Available	No Data Available	No Data Available

Service Actions

- Start
- Stop
- Restart All
- Restart DataNodes
- Restart NFSGateways
- Move NameNode
- Move SNameNode
- Enable NameNode HA
- Run Service Check
- Turn Off Maintenance Mode**
- Rebalance HDFS
- Download Client Configs
- Delete Service

1 Background Operation Running

Operations	Start Time	Duration	Show: All (10)
 Refresh YARN Capacity Scheduler 	Today 23:14	12.54 secs	<div><div></div></div> 35% ▶
 Start All Services	Today 23:04	209.79 secs	<div><div></div></div> 100% ▶
 Start HDFS	Today 23:00	221.71 secs	<div><div></div></div> 100% ▶
 Stop All Services	Thu Feb 01 2018 12:54	78.95 secs	<div><div></div></div> 100% ▶
 Stop required services	Thu Feb 01 2018 12:54	7.52 secs	<div><div></div></div> 100% ▶
 Stop required services	Thu Feb 01 2018 12:53	63.83 secs	<div><div></div></div> 100% ▶
 Stop required services	Thu Feb 01 2018 12:53	4.32 secs	<div><div></div></div> 100% ▶
 Stop required services	Thu Feb 01 2018 12:52	15.07 secs	<div><div></div></div> 100% ▶
 Stop required services	Thu Feb 01 2018 12:52	12.85 secs	<div><div></div></div> 100% ▶
 Stop required services	Thu Feb 01 2018 12:52	12.69 secs	<div><div></div></div> 100% ▶
 Stop required services	Thu Feb 01 2018 12:52	1.94 secs	<div><div></div></div> 100% ▶

[Show more...](#)

☐ Do not show this dialog again when starting a background operation

OK

Laboratorio Azure Hortonworks Data Platform (HDP) Sandbox

almacenamiento , procesamiento y
consulta de información
estructurada en Hadoop

Objetivo:

Realizar acciones de almacenamiento , procesamiento y consulta de información estructurada en Hadoop.

Tecnologías:

HDFS y MapReduce

Sqoop, Pig, Hive, **Impala**

Aplicaciones:

Migrar el almacén de datos (data warehouse) a Hadoop.

- Llevar toda la información dispersa a una plataforma Big Data.
- Reducir el tiempo necesario para generar informes requeridos en la organización.

Actividades

- ☐ Hadoop: Dimensionar un clúster
- ☐ HDFS: Ingesta con las bases de datos a Hadoop.
- ☐ Hive e Impala: Generar informes mediante consultas SQL

Necesidades del negocio

Se requiere desde el negocio, tener un informe diario que permita analizar la evolución de las ventas y establecer estrategias basadas en los productos y minoristas mas relevantes.



Toma de decisiones

1. Definir una estructura de directorios en HDFS
2. Importar los datos con Sqoop
3. Procesar con MapReduce y Pig
4. Planificar con Oozie
5. Consultar con Hive e Impala

Origen de los datos –Data WareHouse

Fuente de información de la empresa : Brazilian E-Commerce

<https://www.kaggle.com/olistbr/brazilian-ecommerce/>

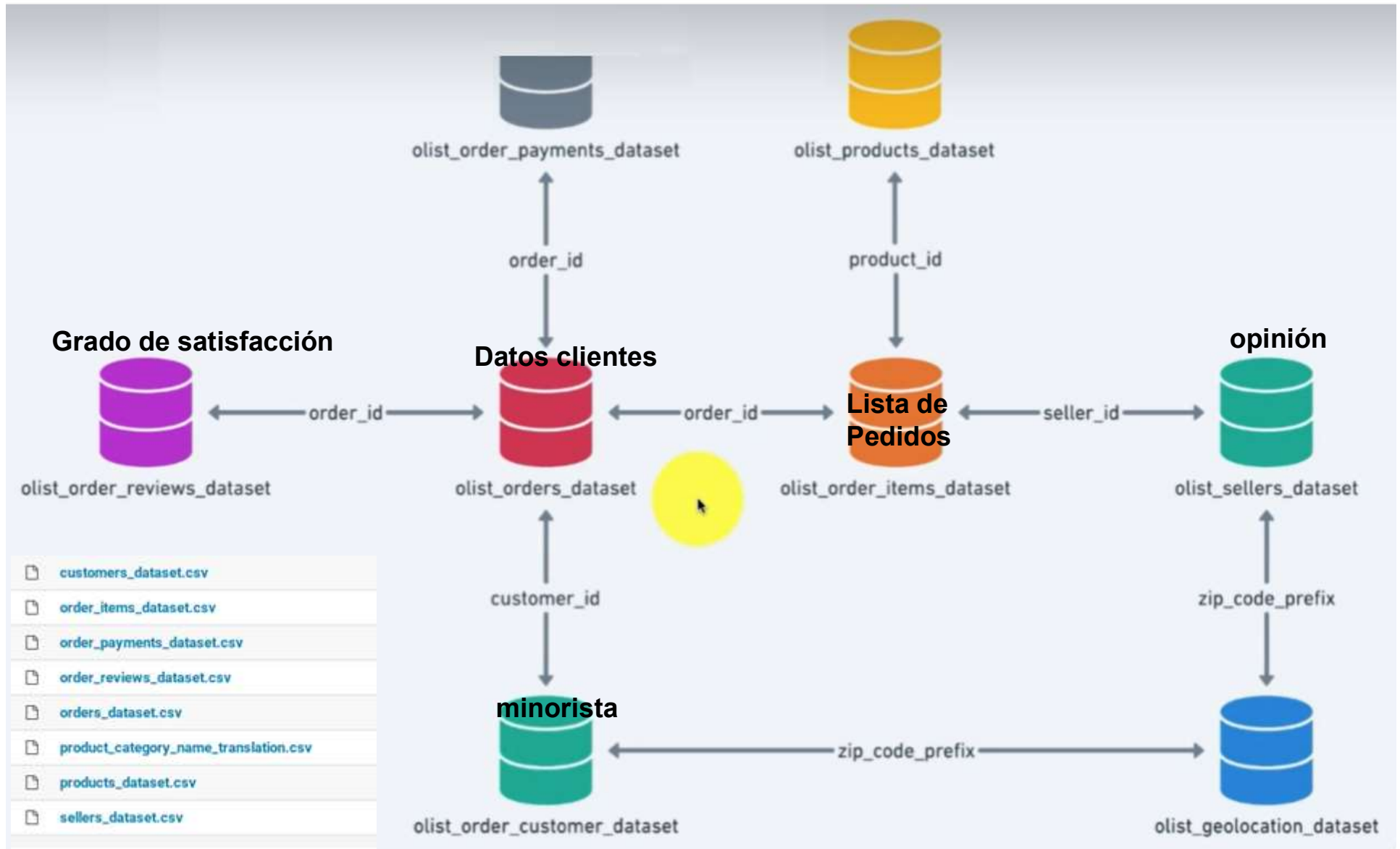
<https://cutt.ly/FtsxMg7>

(hadoop_course-master.zip)

```
cloudera@quickstart:~/workspace/hadoop_course-master/datawarehouse/data
File Edit View Search Terminal Help
[cloudera@quickstart data]$ pwd
/home/cloudera/workspace/hadoop_course-master/datawarehouse/data
[cloudera@quickstart data]$ ls -al
total 63372
drwxrwxr-x 2 cloudera cloudera 4096 Jan 3 15:25 .
drwxrwxr-x 5 cloudera cloudera 4096 Jan 3 15:25 ..
-rw-rw-r-- 1 cloudera cloudera 9033957 Jan 3 15:25 customers_dataset.csv
-rw-rw-r-- 1 cloudera cloudera 15438671 Jan 3 15:25 order_items_dataset.csv
-rw-rw-r-- 1 cloudera cloudera 5777138 Jan 3 15:25 order_payments_dataset.csv
-rw-rw-r-- 1 cloudera cloudera 14409007 Jan 3 15:25 order_reviews_dataset.csv
-rw-rw-r-- 1 cloudera cloudera 17654914 Jan 3 15:25 orders_dataset.csv
-rw-rw-r-- 1 cloudera cloudera 2613 Jan 3 15:25 product_category_name_translation.csv
-rw-rw-r-- 1 cloudera cloudera 2379446 Jan 3 15:25 products_dataset.csv
-rw-rw-r-- 1 cloudera cloudera 174703 Jan 3 15:25 sellers_dataset.csv
[cloudera@quickstart data]$
```

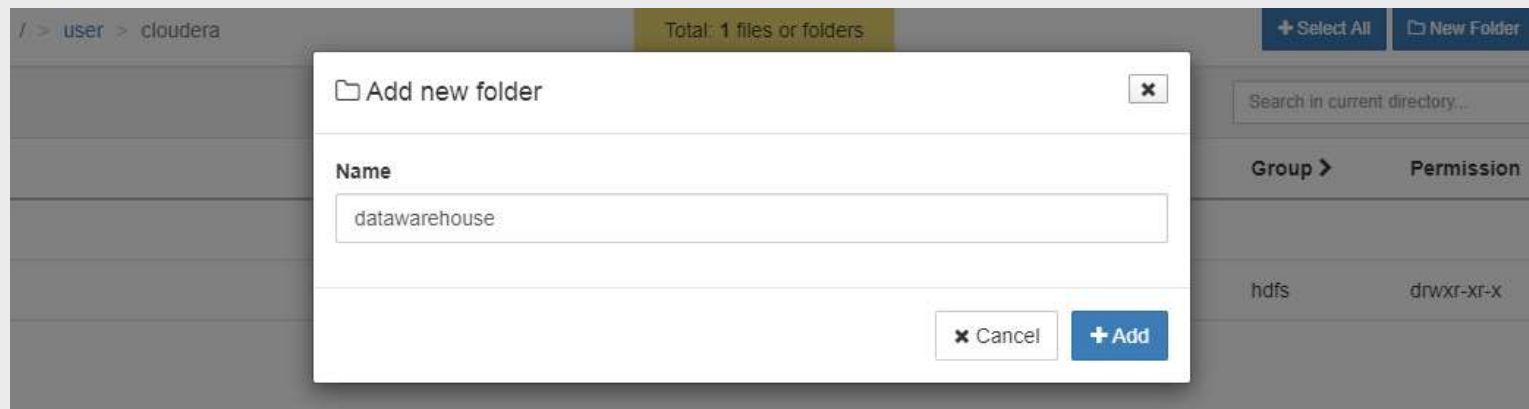
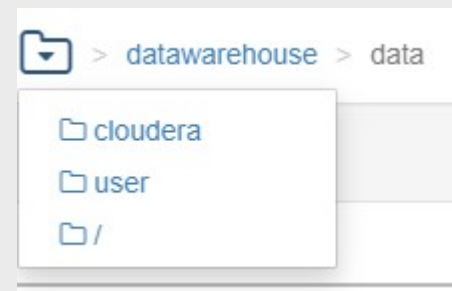
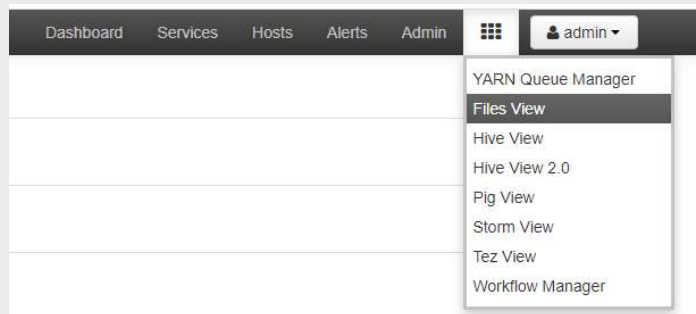
Explorar archivos de datos HDFS

Modelo Relacional

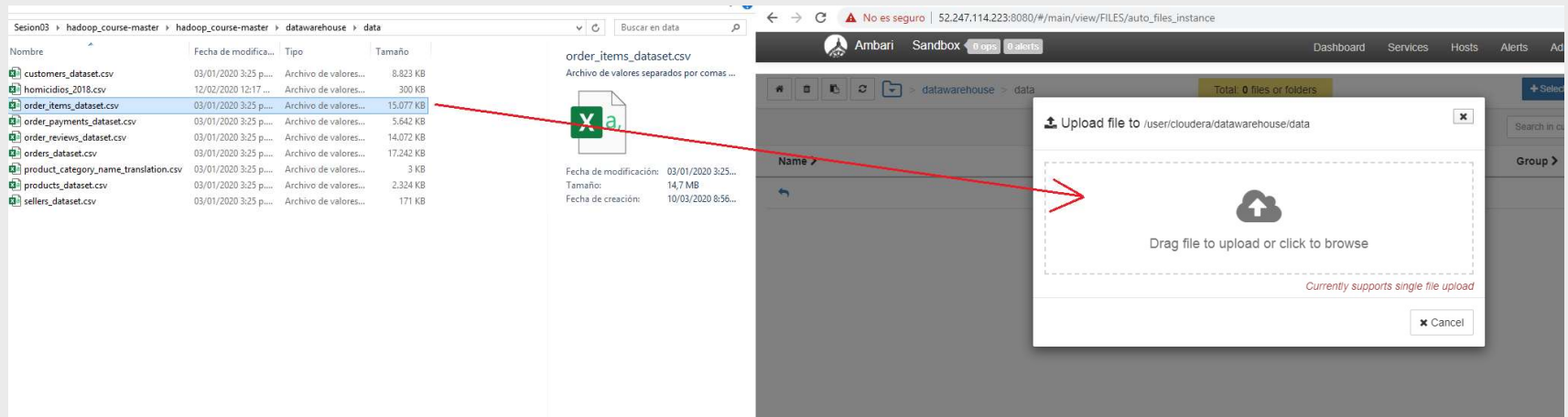
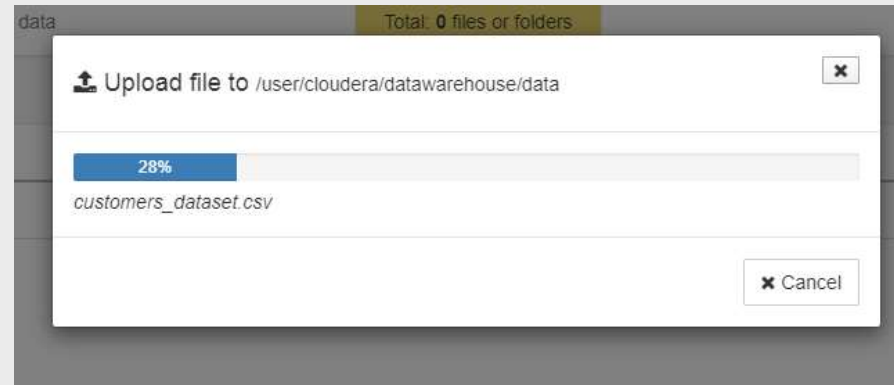
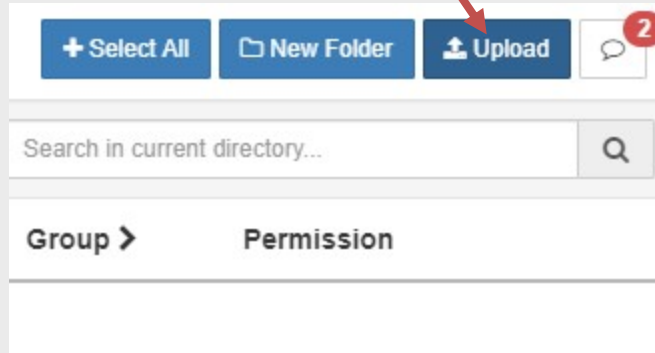


Crear la estructura de directorio














Crear la ruta, estructura de directorio



Cargar los archivos desde FILE



Cargar los archivos desde FILE

<div>      </div> <div>datawarehouse > data</div> <div>Total: 8 files or folders</div> <div> <div>+ Select All</div> <div>New Folder</div> <div>Upload</div> </div>					
<div>Search in current directory...</div>					
Name >	Size >	Last Modified >	Owner >	Group >	Permission
<div>↶</div>					
 customers_dataset.csv	8.6 MB	2020-03-10 21:13	admin	hdfs	-rw-r--r--
 order_items_dataset.csv	14.7 MB	2020-03-10 21:14	admin	hdfs	-rw-r--r--
 order_payments_dataset.csv	5.5 MB	2020-03-10 21:14	admin	hdfs	-rw-r--r--
 order_reviews_dataset.csv	13.7 MB	2020-03-10 21:15	admin	hdfs	-rw-r--r--
 orders_dataset.csv	16.8 MB	2020-03-10 21:15	admin	hdfs	-rw-r--r--
 product_category_name_translation.csv	2.6 kB	2020-03-10 21:16	admin	hdfs	-rw-r--r--
 products_dataset.csv	2.3 MB	2020-03-10 21:16	admin	hdfs	-rw-r--r--
 sellers_dataset.csv	170.6 kB	2020-03-10 21:16	admin	hdfs	-rw-r--r--

Verificar que se encuentra en HDFS

```
[root@sandbox-hdp conf]# hadoop fs -ls /user/cloudera/datawarehouse/data
Found 8 items
-rwxrwxrwx   1 admin hdfs    9033957 2020-03-11 02:13 /user/cloudera/datawarehouse/data/customers_dataset.csv
-rwxrwxrwx   1 admin hdfs   15438671 2020-03-11 02:14 /user/cloudera/datawarehouse/data/order_items_dataset.csv
-rwxrwxrwx   1 admin hdfs    5777138 2020-03-11 02:14 /user/cloudera/datawarehouse/data/order_payments_dataset.csv
-rwxrwxrwx   1 admin hdfs   14409007 2020-03-11 02:15 /user/cloudera/datawarehouse/data/order_reviews_dataset.csv
-rwxrwxrwx   1 admin hdfs   17654914 2020-03-11 02:15 /user/cloudera/datawarehouse/data/orders_dataset.csv
-rw-r--r--   1 admin hdfs     2613 2020-03-11 02:16 /user/cloudera/datawarehouse/data/product_category_name_translation.csv
-rw-r--r--   1 admin hdfs   2379446 2020-03-11 02:16 /user/cloudera/datawarehouse/data/products_dataset.csv
-rw-r--r--   1 admin hdfs    174703 2020-03-11 02:16 /user/cloudera/datawarehouse/data/sellers_dataset.csv
[root@sandbox-hdp conf]# hadoop fs -ls /user/cloudera/datawarehouse/data
```

Crear una Base en HIVE

Hive

Query

Saved Queries

History

UDFs

Upload Table

Database Explorer

brazilian

Search tables...

Databases

default

foodmart

xademo

brazilian

Query Editor

Worksheet *

1 CREATE DATABASE brazilian;

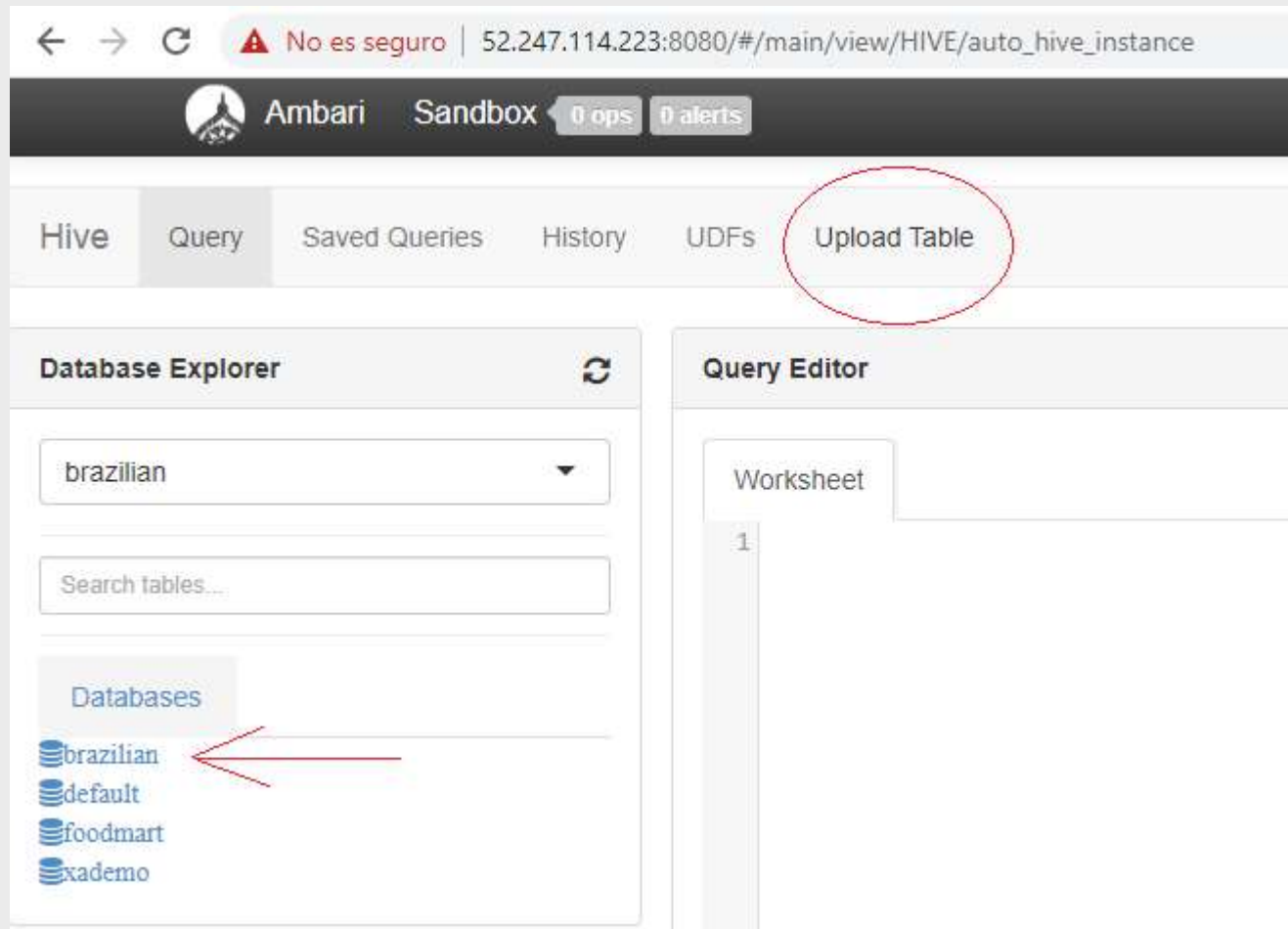
Execute

Explain

Upload

Save as...

Cargar las tablas en la BD – HIVE HDFS



The screenshot displays the Ambari web interface for managing Hadoop clusters. The top navigation bar includes the Ambari logo, the name 'Sandbox', and status indicators for '0 ops' and '0 alerts'. The main navigation tabs are 'Hive', 'Query', 'Saved Queries', 'History', 'UDFs', and 'Upload Table', with 'Upload Table' circled in red. The 'Database Explorer' panel on the left shows a dropdown menu set to 'brazilian' and a search bar labeled 'Search tables...'. Below this, a list of databases is shown: 'brazilian', 'default', 'foodmart', and 'xademo', with a red arrow pointing to 'brazilian'. The 'Query Editor' panel on the right shows a 'Worksheet' tab with a line number '1'.

Cargar las tablas en la BD – HIVE HDFS

Hive

Query

Saved Queries

History

UDFs

Upload Table

Upload from Local



File type

CSV



Upload from HDFS



HDFS Path

/user/cloudera/datawarehouse/data/order_payments

 Ambari

Sandbox

0 ops

0 alerts

Dashboard

Services

Hosts

Alerts

Admin

admin

Hive

Query

Saved Queries

History

UDFs

Upload Table

Upload from Local



File type

CSV



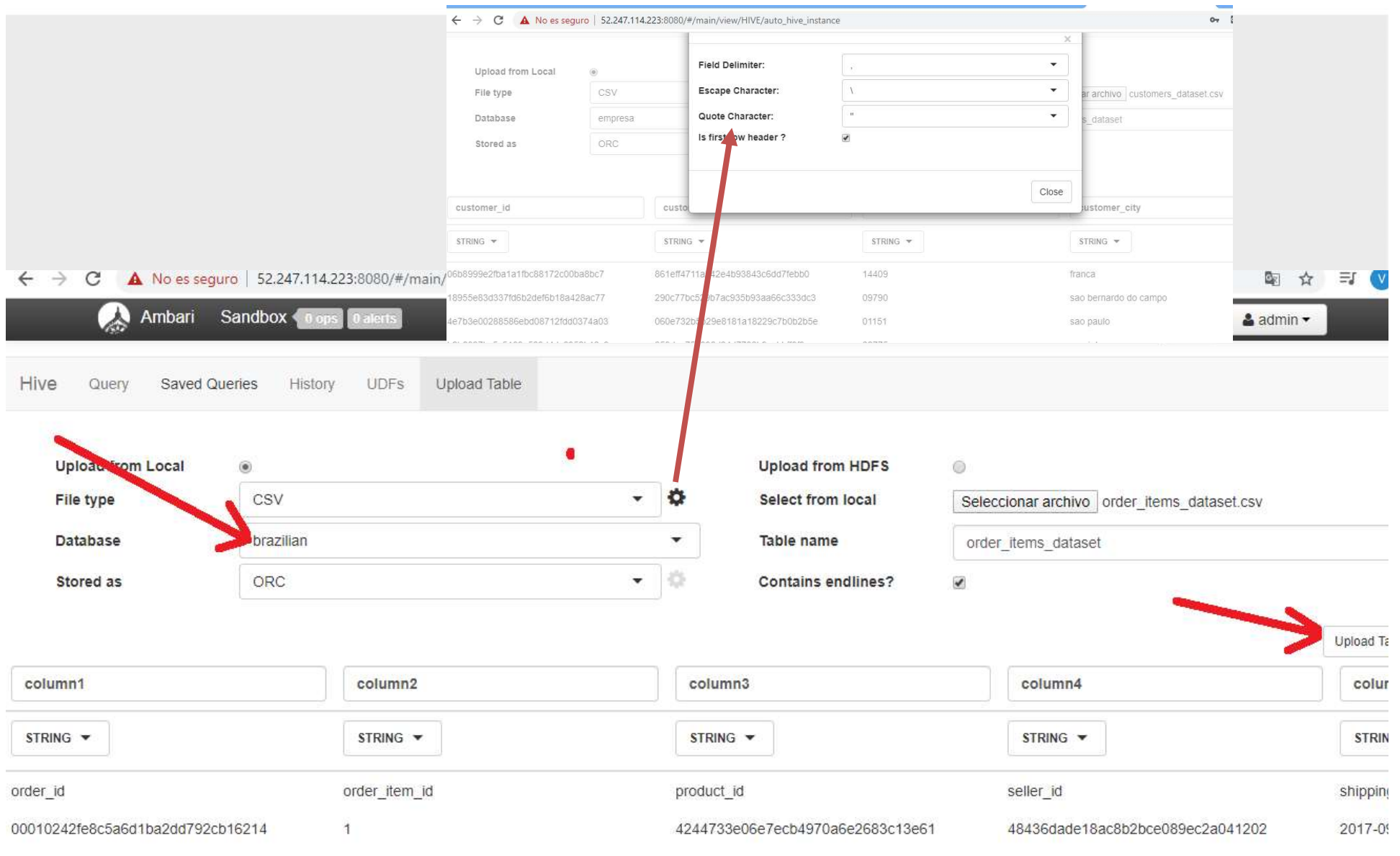
Upload from HDFS



Select from local

Seleccionar archivo order_items_dataset.csv

Cargar Tabla a DB HDFS



The screenshot shows the Ambari interface for uploading a table to a database. The 'Upload Table' tab is selected. The 'Upload from Local' section is active, showing the following configuration:

- File type: CSV
- Database: brazilian
- Stored as: ORC

A modal window is open for configuring the upload, showing the following settings:

- Field Delimiter: ,
- Escape Character: \
- Quote Character: "
- Is first row header?: ☒

The 'Upload from HDFS' section is also visible, showing the following configuration:

- Select from local: Seleccionar archivo order_items_dataset.csv
- Table name: order_items_dataset
- Contains endlines?: ☒

The 'Upload Table' button is located at the bottom right of the interface.

column1	column2	column3	column4	column5
order_id	order_item_id	product_id	seller_id	shipping_date
00010242fe8c5a6d1ba2dd792cb16214	1	4244733e06e7ecb4970a6e2683c13e61	48436dade18ac8b2bce089ec2a041202	2017-01-01

Cargar Tabla a DB HDSF

The screenshot shows a web interface for uploading a table to a database. A modal dialog titled 'Upload Progress' is displayed in the center, listing the following steps:

- Successfully created Actual table.
- Successfully created Temporary table.
- Successfully uploaded file.
- Waiting for insertion of rows from temporary table to actual table.

The background interface includes a tabbed menu with 'Query', 'Saved Queries', 'History', 'UDFs', and 'Upload Table'. The 'Upload Table' tab is active, showing a form with the following fields:

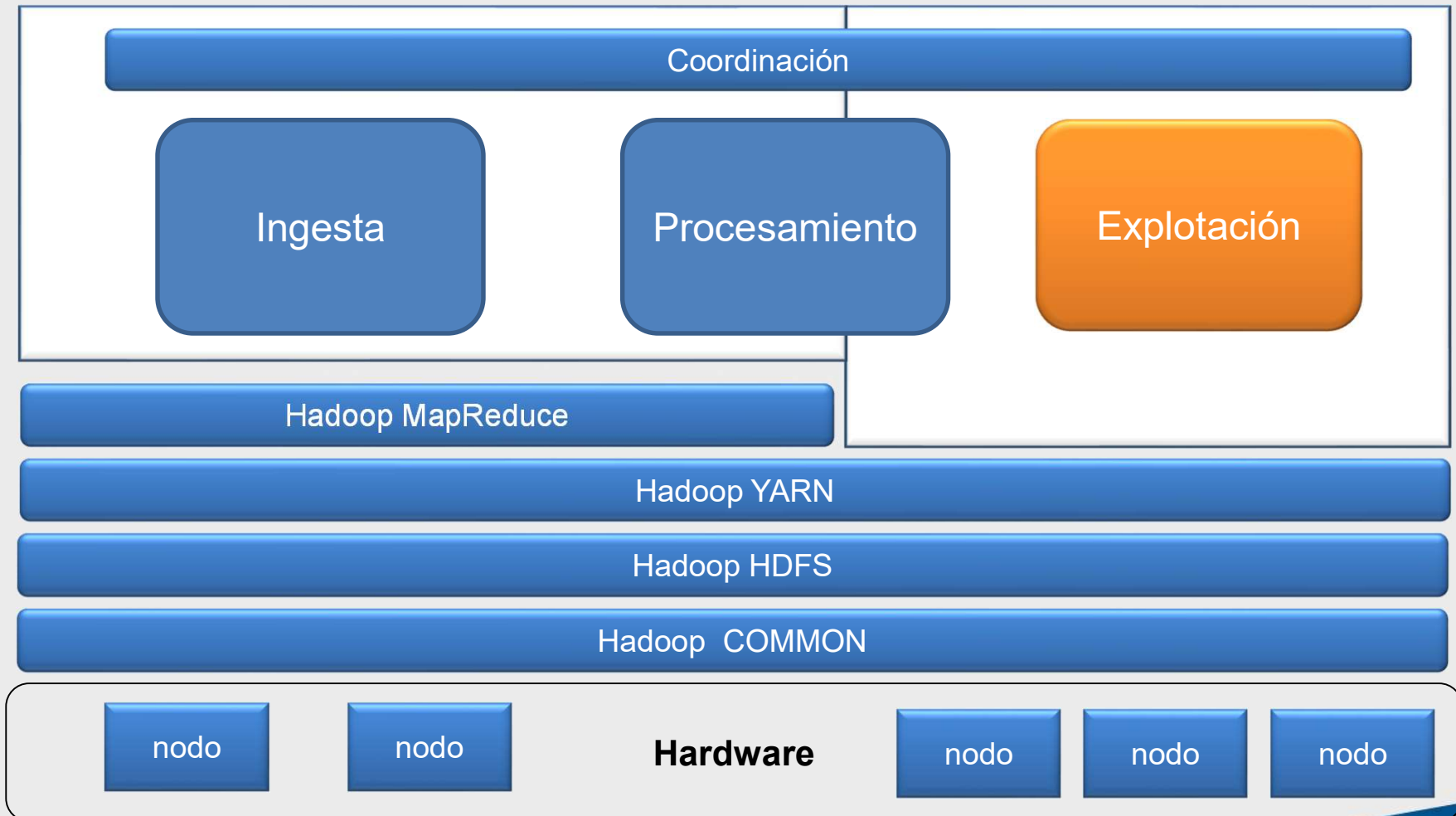
- Upload from Local**: A radio button is selected.
- Type**: A dropdown menu with 'CSV' selected.
- Base**: A dropdown menu with 'default' selected.
- Format**: A dropdown menu with 'ORC' selected.
- Contains endlines?**: A checkbox that is checked.
- File Name**: A text input field containing 'order_items_dataset.csv'.
- Table Name**: A text input field containing 'ms_dataset'.
- Columns**: A table with 4 columns and 2 rows of data types.

column1	column2	column3	column4
STRING	STRING	STRING	STRING

Cargar las tablas por las dos modalidades (10 Minutos)

DROP TABLE IF EXISTS employee;

Practica de Explotación - Hadoop



TOP 10 de ventas en la empresa

-- HIVE Categorías con mayor número de ventas----

```
SELECT products_dataset.product_category_name,  
COUNT(order_items_dataset.product_id) num_ventas  
FROM products_dataset JOIN order_items_dataset  
ON (products_dataset.product_id = order_items_dataset.product_id)  
GROUP BY products_dataset.product_category_name  
ORDER BY num_ventas DESC  
LIMIT 10;
```

Hive - Consultas para el Negocio

Query Process Results (Status: SUCCEEDED)

Logs

Results

Filter columns...

products_dataset.product_category_name **num_ventas**

cama_mesa_banho 11115

beleza_saude 9670

esporte_lazer 8641

moveis_decoracao 8334

informatica_acessorios 7827

utilidades_domesticas 6964

relogios_presentes 5991

telefonica 4545

ferramentas_jardim 4347

automotivo 4235

TOP 10 de ventas en la empresa

-- Categorías con mayor número de ventas

```
SELECT p.product_category_name categorias , COUNT(o.product_id) num_ventas  
FROM order_items_dataset AS o INNER JOIN products_dataset AS p  
ON o.product_id = p.product_id  
GROUP BY categorias  
ORDER BY num_ventas DESC  
LIMIT 10 ;
```

-- Ingresos en función del mes

```
SELECT MONTH(orders_dataset.order_purchase_timestamp)
mes, ROUND(sum(order_items_dataset.price)) ingresos
FROM orders_dataset JOIN order_items_dataset ON
(order_items_dataset.order_id=orders_dataset.order_id)
GROUP BY orders_dataset.order_purchase_timestamp
ORDER BY ingresos DESC;
```

Query Process Results (Status: SUCCEEDED)

Logs

Results

Filter columns...

mes ingresos

9	13440.0
7	7160.0
2	6735.0
7	6729.0
5	6499.0
11	5935.0
4	4799.0
4	4690.0
7	4600.0
6	4590.0

-- Ingresos en función del mes

-- Ingresos en funcion del mes

```
SELECT month(orders.order_purchase_timestamp) mes,  
round(sum(order_items.price)) ingresos  
FROM orders_dataset as orders  
join order_items_dataset as order_items on  
orders.order_id=order_items.order_id  
GROUP BY mes  
ORDER BY ingresos DESC;
```


-- Estados que generan el mayor numero de pedidos

```
SELECT customers.customer_state, count(*) num_orders
FROM orders_dataset as orders
JOIN customers_dataset as customers on
orders.customer_id=customers.customer_id
GROUP BY customers.customer_state
ORDER BY num_orders DESC
LIMIT 10;
```

Hive – Informe estados - resultado

customers.customer_state	num_orders
SP	41746
RJ	12852
MG	11635
RS	5466
PR	5045
SC	3637
BA	3380
DF	2140
ES	2033
GO	2020

Para las consultas realizadas para Impala contenidas en el archivo:

**hadoop_course-master\datawarehouse\queries \
consultas.sql**

Realizar para HIVE, teniendo en cuenta que las diferencias de sintaxis entre IMPALA y HIVE.