

EL ENTORNO: CLOUDERA MANAGER

Procesamiento de Datos

IES Clara del Rey

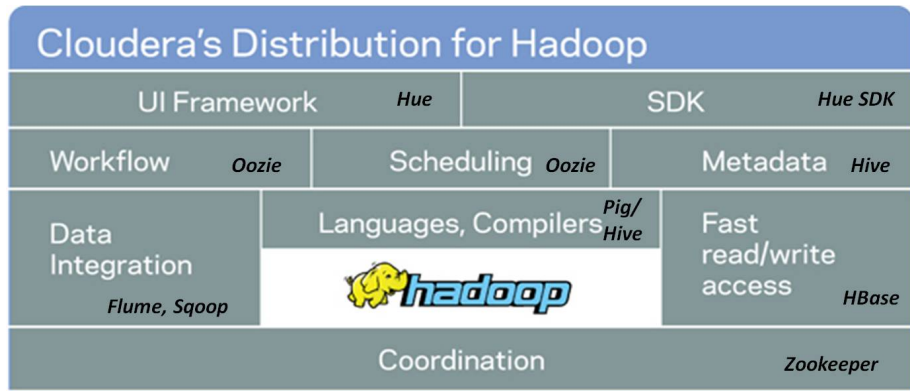


Figure 1: Arquitectura de CDH

Máquina virtual. Requisitos

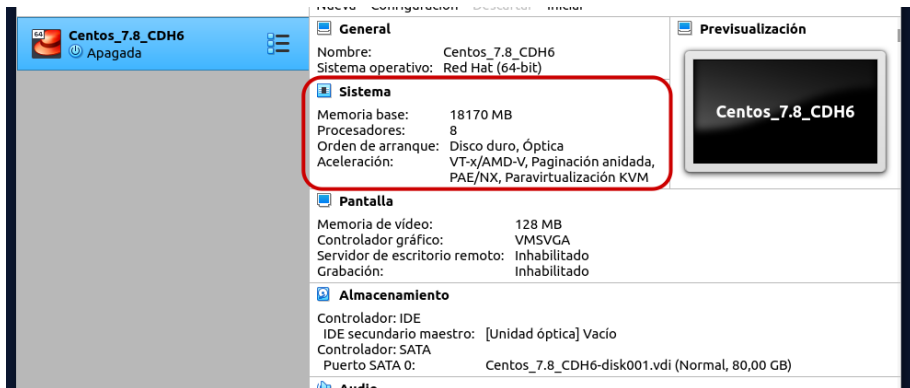


Figure 2: Procesador y memoria de la MV

Visión general de Cloudera Manager

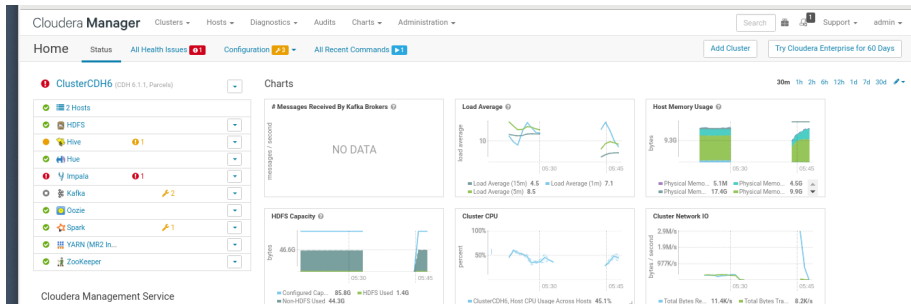


Figure 3: Servicios instalados con su estado

Visión general de Cloudera Manager (cont)

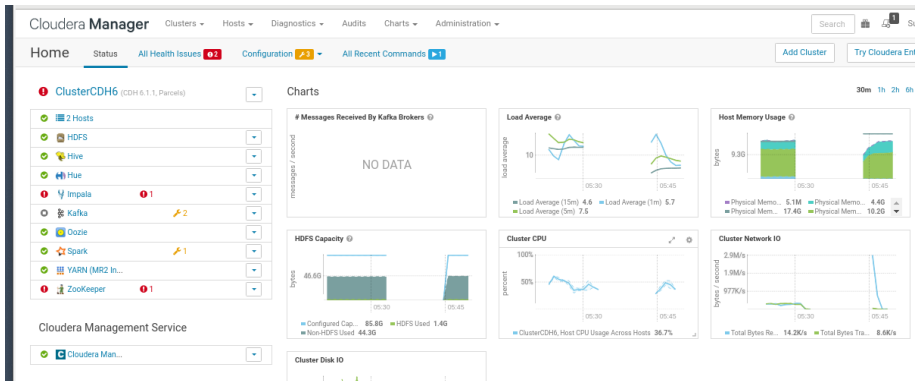


Figure 4: Servicios en diferentes estados

Reinicio de Cloudera Manager

- ✓ Oozie
- ✓ Spark 🔑 1
- ✓ YARN (MR2 In...
- ✓ ZooKeeper

Cloudera Management Service

- ✓ Cloudera Man...

Cloudera Management Service Actions

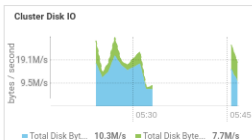
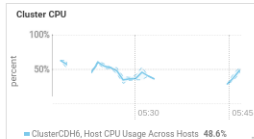
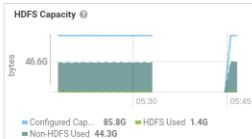
- Start
- Stop
- Restart

- Instances

- Configuration

- Add Role Instances

- Rename



Reinicio del Cluster de Cloudera

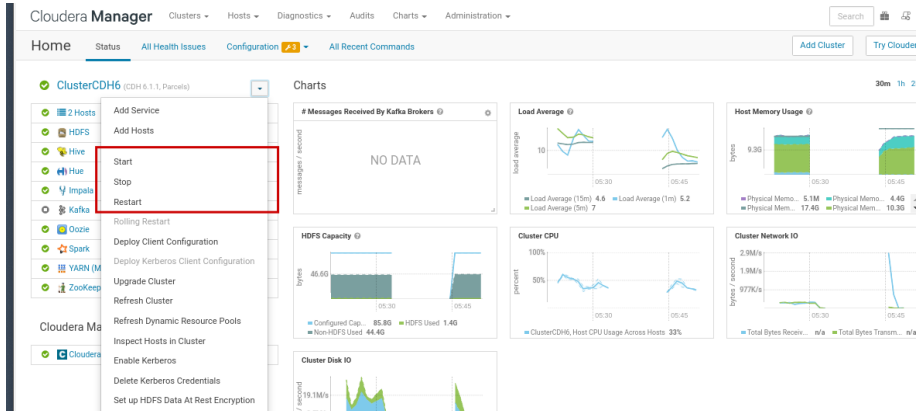


Figure 5: Parada y arranque del Cluster

Reinicio del Cluster de Cloudera

Status ☐ Running Context [ClusterCDH6](#) Mar 4, 5:51:43 PM [Abort](#)

✓ Completed 1 of 2 step(s).

☒ Show All Steps ☐ Show Only Failed Steps ☐ Show Only Running Steps

Execute command Stop on cluster ClusterCDH6	ClusterCDH6	Mar 4, 5:51:43 PM	29.25s
Execute command Start on cluster ClusterCDH6	ClusterCDH6	Mar 4, 5:52:13 PM	Abort
Execute command Start on service ZooKeeper	ZooKeeper	Mar 4, 5:52:13 PM	24.43s
Execute command Start concurrently on 2 services Successfully completed 2 steps.		Mar 4, 5:52:37 PM	33.59s
Execute command Start on service Kafka	Kafka	Mar 4, 5:52:37 PM	25.08s
Execute command Start on service HDFS	HDFS	Mar 4, 5:52:38 PM	33.52s
Execute command Start on service YARN (MR2 Included)	YARN (MR2 Included)	Mar 4, 5:53:11 PM	Abort
Starting 3 roles on service 0/3 start commands completed.		Mar 4, 5:53:11 PM	Abort
Execute command Start on service Spark			
Execute command Start on service Hive			
Execute command Start concurrently on 2 services			

[Abort](#) [Close](#)

Figure 6: Proceso de arranque de los servicios

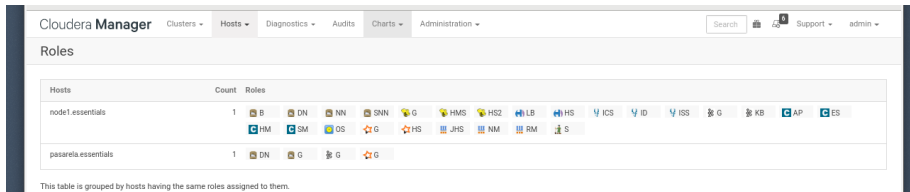
Hosts del Cluster de Cloudera

The screenshot shows the Cloudera Manager interface. The top navigation bar includes 'Clusters', 'Hosts', 'Diagnostics', 'Audits', 'Charts', and 'Administration'. The 'Hosts' menu is highlighted with a red box. Below the navigation bar, the 'All Hosts' page is displayed. On the left, there is a 'Filters' sidebar with 'STATUS' (Good Health) and 'CLUSTERS' (CORES, COMMISSION STATE, LAST HEARTBEAT). The main content area shows a table of hosts with the following columns: Status, Name, IP, Roles, Commission State, Last Heartbeat, Load Average, Disk Usage, Physical Memory, and Swap Space. Two hosts are listed:

Status	Name	IP	Roles	Commission State	Last Heartbeat	Load Average	Disk Usage	Physical Memory	Swap Space
Good Health	node1.essentials	172.18.0.2	25 Role(s)	Commissioned	14.79s ago	4.07 5.74 4.89	28.7 GiB / 47.7 GiB	11.1 GiB / 17.4 GiB	0 B / 8 GiB
Good Health	pasarela.essentials	172.18.0.1	4 Role(s)	Commissioned	6.28s ago	3.76 5.62 4.86	36.1 GiB / 72 GiB	11.1 GiB / 17.4 GiB	0 B / 8 GiB

Figure 7: Propiedades de los hosts del cluster

Roles de cada host de Cloudera



The screenshot shows the Cloudera Manager interface with the 'Roles' tab selected. The table lists roles assigned to hosts, grouped by host. The first group, 'node1.essentials', has a count of 1 and lists roles: B, DN, NN, SMN, G, HMS, HS2, LB, HS, ICS, ID, ISS, G, KB, AP, and ES. The second group, 'pasarela.essentials', has a count of 1 and lists roles: DN, G, G, and G.

Hosts	Count	Roles
node1.essentials	1	B DN NN SMN G HMS HS2 LB HS ICS ID ISS G KB AP ES HM SM OS G HS JHS NM RM S
pasarela.essentials	1	DN G G G

This table is grouped by hosts having the same roles assigned to them.

Figure 8: Roles instalados en cada host

Agregar roles o hosts al Cluster de Cloudera

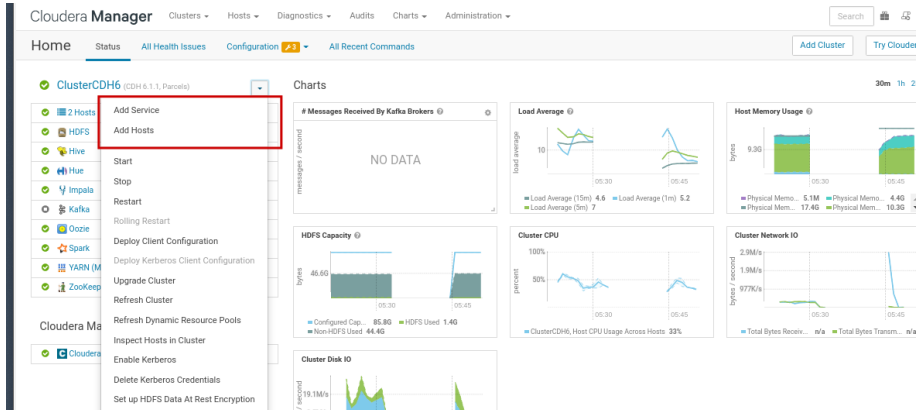


Figure 9: Add Service / Add Hosts

Añadir un servicio a Cloudera Manager

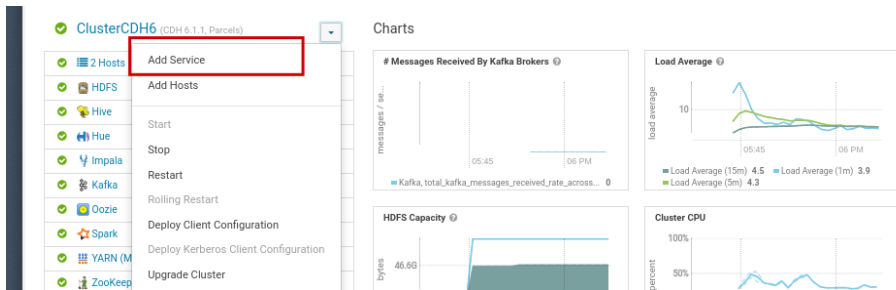


Figure 10: Add Service

Servicios disponibles en Cloudera (I)

Add Service to ClusterCDH6

Select the type of service you want to add.























Service Type	Description
<input type="radio"/>  ADLS Connector	The ADLS Connector service provides key management for accessing Azure Data Lake Stores from CDH services.
<input type="radio"/>  Accumulo	The Apache Accumulo sorted, distributed key/value store is a robust, scalable, high performance data storage and retrieval system. This service only works with releases meant to run on top of CDH6.
<input type="radio"/>  Flume	Flume collects and aggregates data from almost any source into a persistent store such as HDFS.
<input type="radio"/>  HBase	Apache HBase provides random, real-time, read/write access to large data sets (requires HDFS and ZooKeeper).
<input type="radio"/>  HDFS	Apache Hadoop Distributed File System (HDFS) is the primary storage system used by Hadoop applications. HDFS creates multiple replicas of data blocks and distributes them on compute hosts throughout a cluster to enable reliable, extremely rapid computations.
<input type="radio"/>  Hive	Hive is a data warehouse system that offers a SQL-like language called HiveQL.
<input type="radio"/>  Hue	Hue is a graphical user interface to work with the Cloudera Distribution Including Apache Hadoop (requires HDFS, MapReduce, and Hive).
<input type="radio"/>  Impala	Impala provides a real-time SQL query interface for data stored in HDFS and HBase. Impala requires the Hive service and shares the Hive Metastore with Hue.

Figure 11: Lista de servicios disponibles

Servicios disponibles en Cloudera (II)

<input type="radio"/>	 Impala	Impala provides a real-time SQL query interface for data stored in HDFS and HBase. Impala requires the Hive service and shares the Hive Metastore with Hue.
<input type="radio"/>	 Isilon	EMC Isilon is a distributed filesystem.
<input type="radio"/>	 Java KeyStore KMS	The Hadoop Key Management Service with file-based Java KeyStore. Maintains a single copy of keys, using simple password-based protection. Requires CDH 6.0+. Not recommended for production use.
<input type="radio"/>	 Kafka	Apache Kafka is publish-subscribe messaging rethought as a distributed commit log.
<input type="radio"/>	 Key-Value Store Indexer	Key-Value Store Indexer listens for changes in data inside tables contained in HBase and indexes them using Solr.
<input type="radio"/>	 Kudu	Kudu is a true column store for the Hadoop ecosystem.
<input type="radio"/>	 Oozie	Oozie is a workflow coordination service to manage data processing jobs on your cluster.
<input type="radio"/>	 S3 Connector	The S3 Connector Service securely provides a single set of AWS credentials to Impala and Hue. This enables Hue administrators to browse the S3 filesystem and define Impala tables backed by S3 data authorized to that AWS identity, and also enables Impala users to query S3-backed tables without directly providing AWS credentials, subject to having the proper permissions defined via Sentry. The S3 Connector only supports the S3A protocol.
<input type="radio"/>	 Sentry	Sentry service stores authorization policy metadata and provides clients concurrent and secure access to this metadata.
<input type="radio"/>	 Solr	Solr is a distributed service for indexing and searching data stored in HDFS.
<input type="radio"/>	 Spark	Apache Spark is an open source cluster computing system. This service runs Spark as an application on YARN.
<input type="radio"/>	 Sqoop 1 Client	Configuration and connector management for Sqoop 1.
<input type="radio"/>	 YARN (MR2 Included)	Apache Hadoop MapReduce 2.0 (MRv2), or YARN, is a data computation framework that supports MapReduce applications (requires HDFS).
<input type="radio"/>	 ZooKeeper	Apache ZooKeeper is a centralized service for maintaining and synchronizing configuration data.

[Back](#)[Continue](#)

Figure 12: Lista de servicios disponibles

Menú de gestión del Cluster

The screenshot displays the Cloudera Manager web interface for a cluster named 'ClusterCDH6' (CDH 6.1.1, Parcels). A dropdown menu is open, showing various services and management options. The 'Status' tab is selected, showing a summary of the cluster's health. The 'Health Tests' section indicates 'Show 7 Good'. The 'Status Summary' table lists the following components and their health:

Component	Health
Balancer	1 None
DataNode	2 Good Health
Gateway	1 None
NameNode	1 Good Health
SecondaryNameNode	1 Good Health
Hosts	2 Good Health

The dropdown menu includes the following options:

- HDFS
- Hive
- Hue
- Impala
- Kafka
- Oozie
- Spark
- YARN (MR2 Included)
- ZooKeeper
- Hosts
- Roles
- Host Templates
- Parcels
- Impala Queries
- YARN Applications
- Dynamic Resource Pool Configuration
- Static Service Pools
- Cloudera Management Service

The main dashboard also features several charts and sections:

- Health**: A bar chart showing the percentage of nodes in different health states (bad, disabled, concerning, good) over time.
- Alerts Across DataNodes**: A line chart showing the number of alerts across data nodes over time.
- Average Disk Flush Time Across DataNodes**: A line chart showing the average disk flush time across data nodes over time.
- Status Summary**: A table summarizing the health of various components.

Figure 13: Servicios y opciones del Cluster

Menú para los servicios instalados en el Cluster

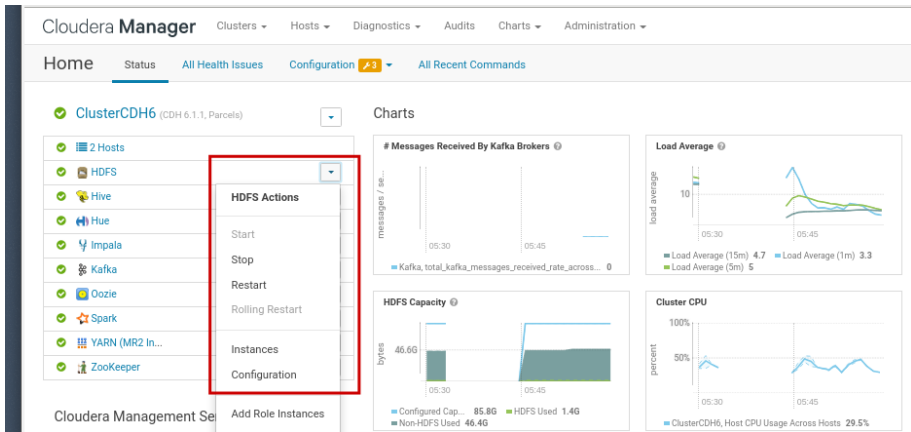


Figure 14: Parada y arranque de cada servicio

Instancias de los servicios del Cluster

ClusterCDH6

HDFS Actions

Mar 4, 6:01 PM CET

Status Instances Configuration Commands Charts Library Cache Statistics Audits NameNode Web UI Quick Links

Show 1 Suppressed Warning(s)

Search

Filters

- STATUS
 - None 2
 - Good Health 4
- COMMISSION STATE
- MAINTENANCE MODE
- RACK
- ROLE GROUP
- ROLE TYPE
- STATE

Actions for Selected ▾ Migrate Roles Add Role Instances Role Groups

<input type="checkbox"/>	Role Type	State	Host	Commission State	Role Group
<input type="checkbox"/>	Balancer	N/A	node1.essentials	Commissioned	Balancer Default Group
<input type="checkbox"/>	DataNode	Started	node1.essentials	Commissioned	DataNode Default Group
<input type="checkbox"/>	DataNode	Started	pasarela.essentials	Commissioned	DataNode Default Group
<input type="checkbox"/>	Gateway	N/A	pasarela.essentials	Commissioned	Gateway Default Group
<input type="checkbox"/>	NameNode (Active)	Started	node1.essentials	Commissioned	NameNode Default Group
<input type="checkbox"/>	SecondaryNameNode	Started	node1.essentials	Commissioned	SecondaryNameNode Default Group

Figure 15: Instancias de HDFS

Configuraciones de los servicios del Cluster

The screenshot displays the Cloudera Manager interface for the 'ClusterCDH6' instance. The 'HDFS' service is selected, and the 'Configuration' tab is active. On the left, a 'Filters' sidebar lists various configuration categories and their counts. The main area shows a list of services and their configurations. The 'HDFS Block Size' configuration is highlighted with a red box, showing a value of 128 MiB. Other visible configurations include 'ZooKeeper Service', 'KMS Service', 'Object Store Service', 'Default Umask', and 'Enable WebHDFS'.

Service	Configuration	Value
ZooKeeper Service	HDFS (Service-Wide)	ZooKeeper
KMS Service	HDFS (Service-Wide)	none
Object Store Service	HDFS (Service-Wide)	none
HDFS Block Size	dfs.blocksize	128 MiB
Default Umask	fs.permissions.umask-mode	022
Enable WebHDFS	dfs.webhdfs.enabled	<input checked="" type="checkbox"/> HDFS (Service-Wide)

Figure 16: Opción del tamaño de bloque de HDFS

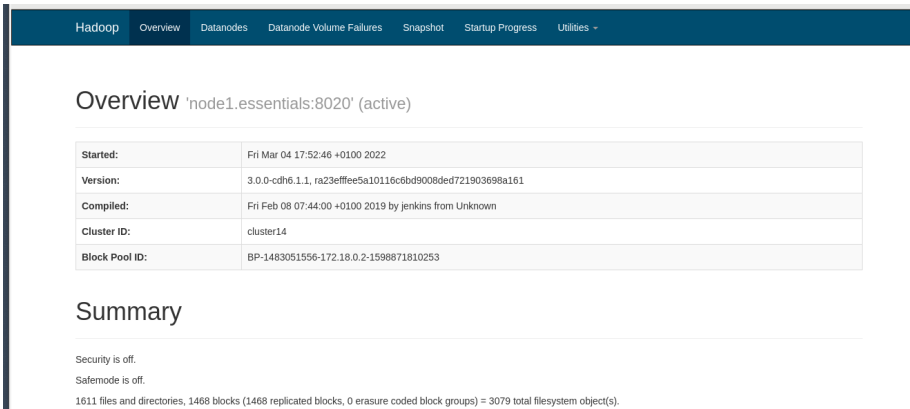


Figure 17: Interfaz de HDFS

HUE: File Browser

The screenshot displays the HUE File Browser interface. The top navigation bar features a 'Query' dropdown and a search bar. A red box highlights the 'File Browser' tab. Another red box highlights the user profile 'root'. The main content area shows a breadcrumb path '/ user / root' and a table of files and directories.

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	.		hdfs	supergroup	drwxrwxrwx	September 09, 2021 08:49 AM
<input type="checkbox"/>	.		root	root	drwxrwxrwx	April 05, 2021 03:34 AM
<input type="checkbox"/>	.Trash		root	root	drwxrwxrwx	September 13, 2021 03:00 AM
<input type="checkbox"/>	.staging		root	root	drwxrwxrwx	April 05, 2021 03:36 AM
<input type="checkbox"/>	2015_11_18		root	supergroup	drwxrwxrwx	September 01, 2020 03:28 AM
<input type="checkbox"/>	2015_11_19		root	supergroup	drwxrwxrwx	September 01, 2020 03:28 AM

Figure 18: Directorios de HDFS

HUE: Editor de consultas

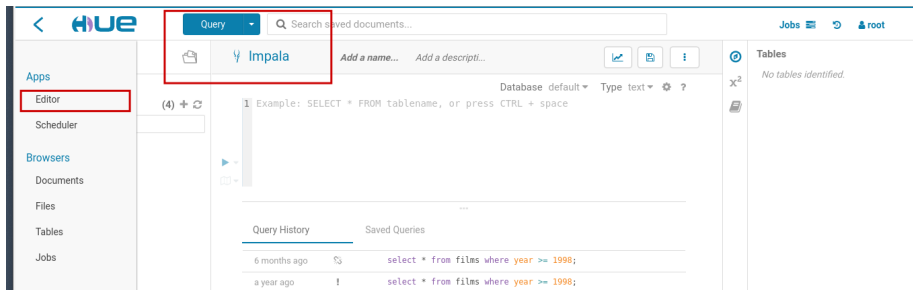


Figure 19: Consultas para Impala con HUE

Empecemos. . . .

cloudera

