

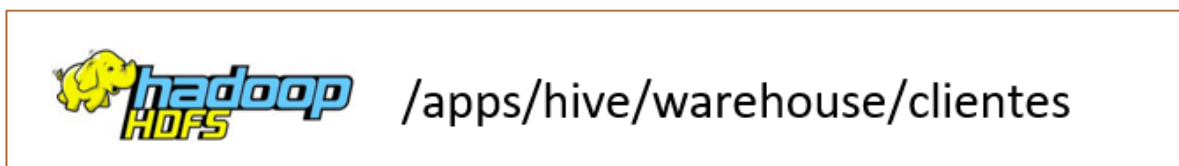
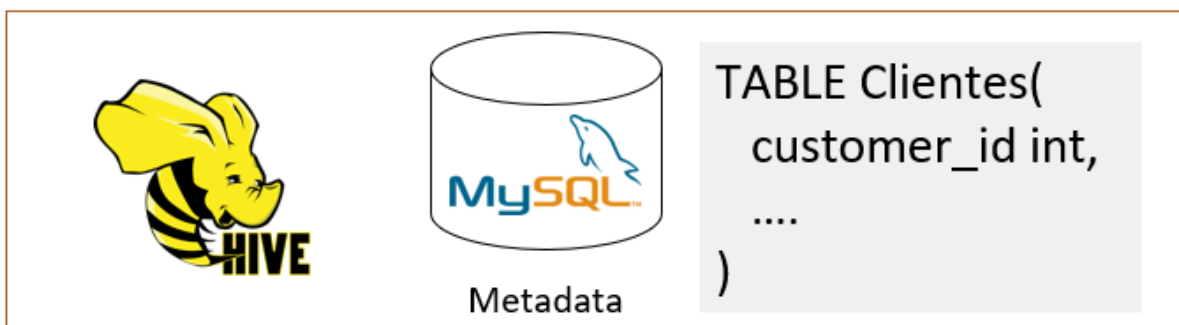
Minería de Datos II

Clase 6 - Hive

Permite crear infraestructuras de tipo de data warehouse sobre Hadoop para realizar análisis de grandes volúmenes de datos

Asigna una estructura tabular (metadata) a los datos en bruto almacenados en HDFS

```
SELECT * FROM clientes;
```



- HiveQL (Hive Query Language)

Hive utiliza un subconjunto de comandos SQL.

Data Definition Language

<https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL>

Data Manipulation Language

<https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DML>

Importante: las operaciones de UPDATE y DELETE no están habilitadas por defecto.

- Tipos de Tablas

MANAGED	EXTERNAL
Hacen referencia a un path dentro de HDFS que es administrado por Hive	Generan metadata para un path de HDFS que no es administrado por Hive
El valor por defecto se especifica en el parámetro <code>hive.metastore.warehouse.dir</code> y típicamente es <code>/user/hive/warehouse/</code>	Debemos agregar la palabra clave <code>EXTERNAL</code> y especificar el path de HDFS en la sección <code>LOCATION</code>
En caso de realizar una operación de tipo <code>DROP TABLE</code> , Hive eliminaría la metadata de la tabla y los datos	En caso de realizar una operación de tipo <code>DROP TABLE</code> , Hive eliminaría la metadata de la tabla pero no los datos

-Tipos de Dato

Hive, además de los tipos de datos comunes a todos los motores de bases de datos relacionales, ofrece una nueva categoría de tipos de datos complejos

Complex Types

`ARRAY<data_type>`

`MAP<primitive_type, data_type>`

`STRUCT<col_name : data_type, ...>`

	Name	Type
0	id	int
1	lastname	string
2	firstname	string
3	dob	date
4	newsletter	boolean
5	contacts	map<string,string>
6	orders	array<string>
7	site	string

-Formatos de Almacenamiento

Hive permite leer y escribir datos en diferentes formatos de archivos.

Habitualmente se utilizan 2 formatos:

- CSV para los datos en bruto
- Parquet para los datos procesados



-Particiones

El particionamiento es una forma de dividir una tabla en partes relacionadas en función de los valores de columnas particulares (por ej. fecha, la ciudad y el departamento).

Cada tabla puede tener una o más claves de partición para identificar una partición particular.

Esta forma de almacenar los datos permite realizar consultas más eficientes.

```
/user/hive/warehouse/logs
├── dt=2001-01-01/
│   ├── country=GB/
│   │   ├── file1
│   │   └── file2
│   └── country=US/
│       └── file3
└── dt=2001-01-02/
    ├── country=GB/
    │   └── file4
    └── country=US/
        ├── file5
        └── file6
```

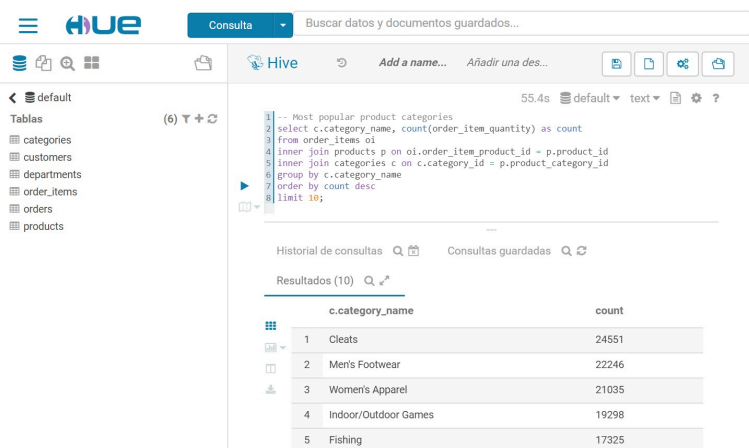
-Ejemplo Hive

```
CREATE EXTERNAL TABLE page_view_stg(viewTime INT, userid BIGINT,
    page_url STRING, referrer_url STRING,
    ip STRING COMMENT 'IP Address of the User',
    country STRING COMMENT 'country of origination')
COMMENT 'This is the staging page view table'
ROW FORMAT DELIMITED FIELDS TERMINATED BY '44' LINES TERMINATED BY '12'
STORED AS TEXTFILE
LOCATION '/user/data/staging/page_view';
```

```
hadoop dfs -put /tmp/pv_2008-06-08.txt /user/data/staging/page_view
```

```
FROM page_view_stg pvs
INSERT OVERWRITE TABLE page_view PARTITION(dt='2008-06-08', country='US')
SELECT pvs.viewTime, pvs.userid, pvs.page_url, pvs.referrer_url, null, null, pvs.ip
WHERE pvs.country = 'US';
```

-Hue (Hadoop User Experience)



The screenshot shows the Hue web interface. On the left is a sidebar with a navigation menu including 'default', 'Tablas', 'categories', 'customers', 'departments', 'order_items', 'orders', and 'products'. The main area displays a SQL query in a text editor with a syntax highlighter. Below the query, there's a section for 'Historial de consultas' and 'Consultas guardadas'. The bottom part shows the 'Resultados (10)' of the query, which is a table with two columns: 'c.category_name' and 'count'. The results are as follows:

c.category_name	count
1 Cleats	24551
2 Men's Footwear	22246
3 Women's Apparel	21035
4 Indoor/Outdoor Games	19298
5 Fishing	17325

Es una interfaz web que permite ejecutar consultas SQL hacia diferentes motores de bases de datos, principalmente relacionados a Big Data.

Bases de datos soportadas
[Conectores](#)

Entorno de prueba gratuito
[Demo](#)

-Ranger:

Proporciona una gestión de seguridad y control de acceso basada en políticas para entornos de big data. Su objetivo principal es proporcionar una capa de seguridad unificada y centralizada para proteger los datos almacenados y procesados:

Policy Details :

Policy Type **Row Level Filter**

Policy Name * rowFilter: cust.customer table **enabled**

Hive Database * **cust**

Hive Table * **customer**

Audit Logging **YES**

Description Restrict employees to access only country-specific customer records

Row Filter Conditions :

Select Group	Select User	Access Types	Row Level Filter	
us-employees	Select User	select	addr_country = 'US'	x
uk-employees	Select User	select	addr_country = 'UK'	x
de-employees	Select User	select	addr_country = 'DE'	x

Atlas:

Proporciona una gestión de metadatos de Big Data utilizada para catalogar, descubrir y gestionar los metadatos en un ecosistema de big data. Su principal objetivo es proporcionar una vista centralizada y unificada de los metadatos de los activos de datos en todo el ecosistema:

Apache Atlas

Q SEARCH CLASSIFICATION GLOSSARY

Basic ☒ Advanced

Search By Type
hive_column

Search By Classification
PII

Search By Term
Search Term

Search By Text
Search by text

Clear Search

Favorite Searches
Save Save As
PII columns

Results for: (Type: hive_column) AND (Classification: PII)
If you do not find the entity in search result below then you can create new entity

Showing 8 records From 1 - 25

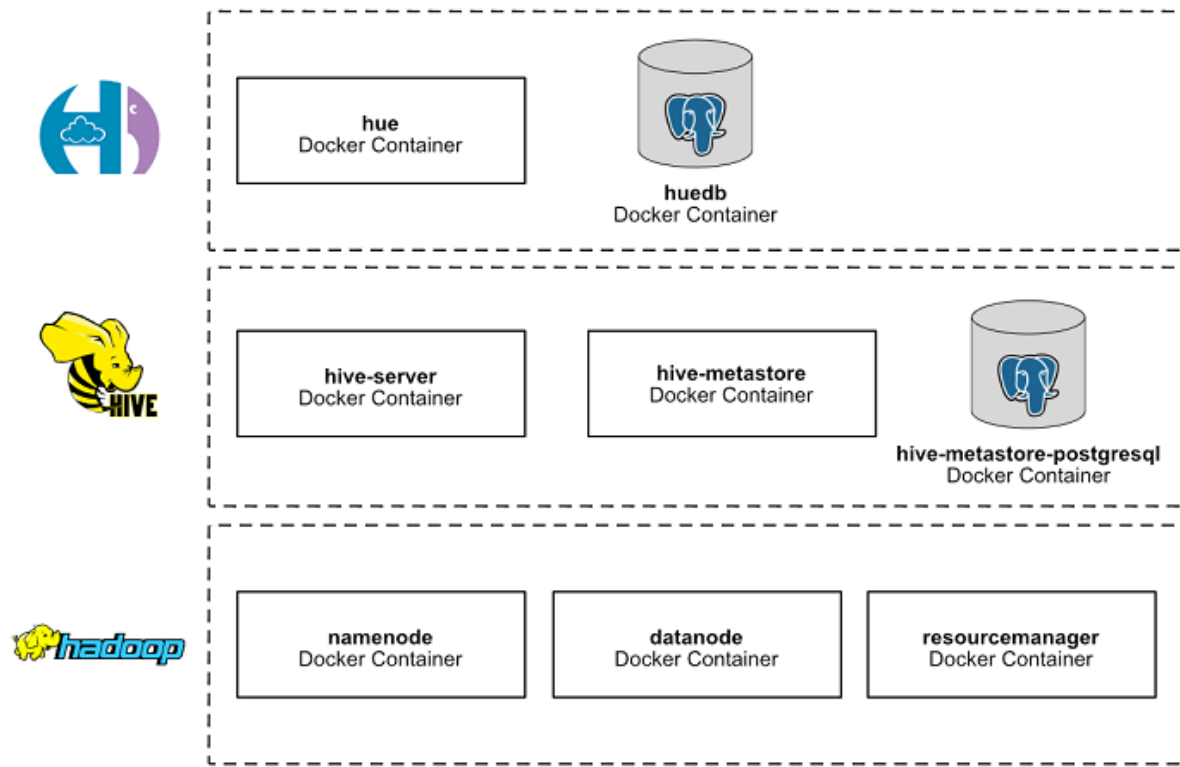
☐ Exclude sub-types ☐ Exclude sub-classifications ☐ Show historical entities Columns

Name	Owner	Type	Type	Classifications	Table	QualifiedName
providename	hive	hive_column	string	VENDOR_PII	prov_view	claim_prov_view.providename@cl1
emailaddress	hive	hive_column	string	PII	ww_customers	hortoniabank.ww_customers.emailaddress@cl1
ccnumber	hive	hive_column	string	PII	ww_customers	hortoniabank.ww_customers.ccnumber@cl1
nationalid	hive	hive_column	string	PII	ww_customers	hortoniabank.ww_customers.nationalid@cl1
nationalid	hive	hive_column	string	PII	us_customers	hortoniabank.us_customers.nationalid@cl1
ssn	hive	hive_column	string	FINANCE...	tax_2015	finance.tax_2015.ssn@cl1
providename	hive	hive_column	string	VENDOR...	provider_summary	claim_provider_summary.providename@cl1
ssn	hive	hive_column	string	FINANCE...	tax_2010	finance.tax_2010.ssn@cl1

Page Limit: 25

Laboratorio:

Se trabajará con el siguiente entorno de Docker Compose:



Instrucciones para su configuración:

- `git clone https://github.com/tech4242/docker-hadoop-hive-parquet.git`
- `cd docker-hadoop-hive-parquet/`
- `sudo docker-compose up`

Analizar Datos con Hive

La interfaz de login de Hue muestra el logo 'HUE' con el eslogan 'Query. Explore. Repeat.' Debajo del logo hay dos campos de entrada: el primero contiene el texto 'instructor' y el segundo contiene caracteres ocultos por puntos. En la parte inferior hay un botón azul que dice 'Iniciar sesión'.

Ingresa a Hue (ver imágenes)
`http://<ip>:8888/hue`

En la sección de archivos, cargar los archivos de la carpeta data y replicar la misma estructura de directorios en HDFS

En la sección de mis documentos, cargar el archivo clase-03.json y luego seleccionar el editor Hive.

Enlace sugerido para lectura:

Towardsdatascience : Big-data-with-hadoop-hive-parquet-hue-and-docker

Consulta

Search data and saved documents...

Jobs

default

Tablas

Filtrar...

access_log

javascript: void(0)

Explorador de archivos

Buscar nombre de archivo

Acciones

Eliminar permanentemente

Cargar

Nuevo

Inicio / user / instructor / data / bikeshare / stations

	Nombre	Tamaño	Usuario	Grupo	Permisos	Fecha
<input type="checkbox"/>	📁		instructor	instructor	drwxr-xr-x	June 21, 2020 03:31 PM
<input type="checkbox"/>	📁		instructor	instructor	drwxr-xr-x	June 21, 2020 03:31 PM

Mostrar 45 de 0 elementos

Página 1 de 1

Consulta

Search data and saved documents...

Jobs

default

Tablas

Filtrar...

access_log

Cargar a /user/instructor/data/bikeshare/stations

Seleccionar archivos

o arrástrelos y suéltelos aquí

austin_bikeshare_stations.csv 5.7kB

Explorador de archivos

Buscar nombre de archivo

Acciones

Eliminar permanentemente

Cargar

Nuevo

Inicio / user / instructor / data / bikeshare / stations

	Nombre	Tamaño	Usuario	Grupo	Permisos	Fecha
<input type="checkbox"/>	📁		instructor	instructor	drwxr-xr-x	June 21, 2020 03:31 PM
<input type="checkbox"/>	📁		instructor	instructor	drwxr-xr-x	June 21, 2020 03:51 PM

Mostrar 45 de 0 elementos

Página 1 de 1

Mis documentos

📄

👤

⋮

Papelera

Nombre	Descripción	Última mod
Clase-02		29/06/202

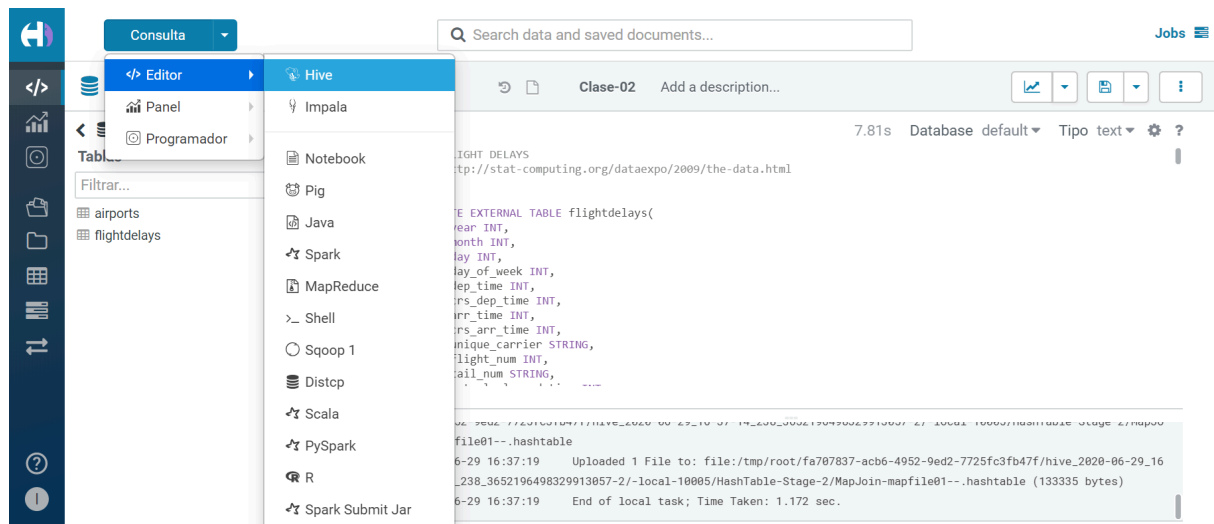
Copiar

Mover a papelera

Cambiar nombre de la carpeta

Exportar

Importar



Fuente: <https://towardsdatascience.com/making-big-moves-in-big-data-with-hadoop-hive-parquet-hue-and-docker-320a52ca175>