HENRY

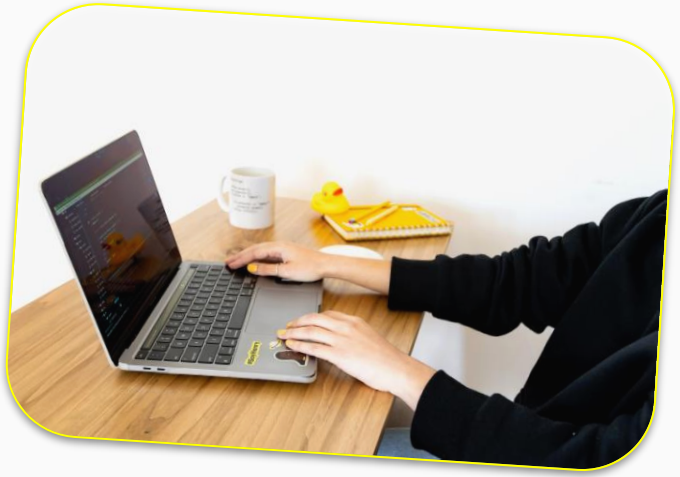
Hive

OBJETIVOS DE CLASE

-Utilizar Hive para crear un Datawarehouse sobre Hadoop

-Comprender cuáles son los Formatos de Almacenamiento utilizados

-Entender el concepto de la Governanza del Dato (Data Governance)



AGENDA

- Hive
- HiveQL
- Tipos de tablas
- Tipos de datos
- Formatos de Almacenamiento
- Particiones
- Data Governance

HIVE

¿Qué es HIVE y para qué sirve?

Hive se basa en la plataforma de software de Apache Hadoop y proporciona una interfaz de consulta y análisis similar a SQL.

- 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;

Metadata

```
TABLE Clientes(
   customer_id int,
   ....
)
```

/apps/hive/warehouse/clientes

HiveQL (Hive Query Language)

HiveQL

- Hive utiliza un subconjunto de comandos **SQL**.
- Data Definition Language
- Data Manipulation Language
- Las operaciones de **UPDATE** y **DELETE** no están habilitadas por defecto.

Tipos de tabla

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 en el parámetro **hive.metastore.warehouse.dir** y típicamente es **/user/hive/warehouse/**	Debemos agregar la palabra clave **EXTERNAL** y especificar el path de **HDFS** en la sección **LOCATION**
En caso de realizar una operación de tipo **DROP TABLE**, Hive eliminaría la metadata de la tabla y los datos	En caso de realizar una operación de tipo DROP **TABLE**, Hive eliminaría la metadata de la tabla pero no los datos

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:

- ARRAY<data_type>
- MAP<col_name : data_type, ...>
- STRUCT<primitive_type, data_type>

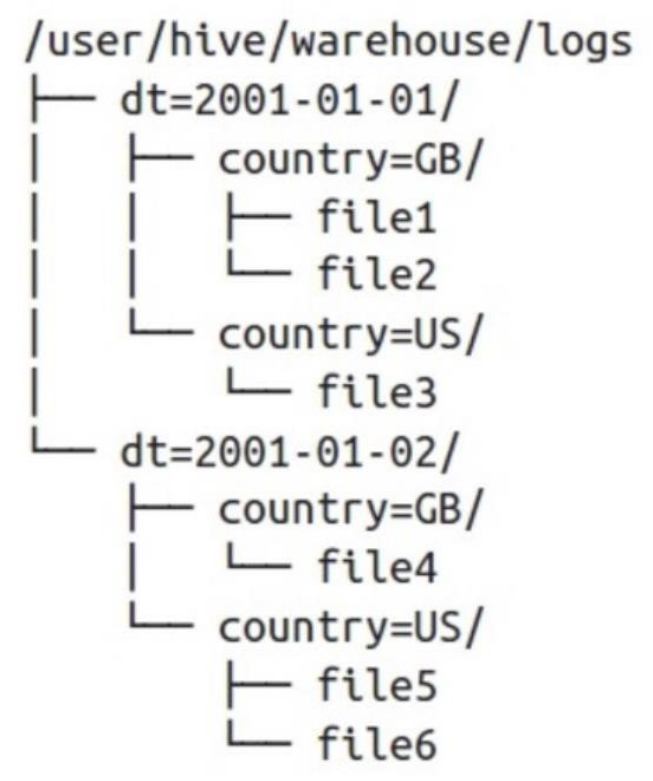
Formatos de Almacenamiento

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



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.

Hive SerDes

Acrónimo de Serializer/Deserializer. Permite interpretar diferentes formatos. SerDes disponibles en Hive:

SerDes disponibles en Hive:

- Avro (Hive 0.9.1 and later)
- ORC (Hive 0.11 and later)
- RegEx
- Thrift
- Parquet (Hive 0.13 and later)
- CSV (Hive 0.14 and later)
- JsonSerDe (Hive 0.12 and later in hcatalog-core)

```
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)

¿Qué es?

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
- Entorno de prueba gratuito

Data Governance

¿Qué es?

Propone considerar a los **datos como activos** de una empresa y su gestión debe estar alineada con los objetivos estratégicos y está cobrando importancia en las organizaciones.

Tiene que ver con darle al **ciclo de vida del dato**, una persona o grupo de personas, que sean responsables por conocer su recorrido completo, **desde las implicancias de cómo, dónde, por qué y por quién es generado, hasta de qué forma ese dato aporta información valiosa a la hora de tomar decisiones y evaluar nuestra posición frente a objetivos planteados.**

> Gestionar el dato no es sólo procesar y almacenar. Se trata de gestionar la seguridad, el cumplimiento de la normativa y la segregación del acceso, por no hablar de la lucha diaria por resolver los problemas causados por la **mala calidad de los datos.**


Ownership
Knowledge
Accessibility
Technology
People
Data
Governance
Process
Quality
Security

RESUMEN DE LA CLASE

- ✓ Hive
- ✓ HiveQL
- ✓ Tipos de tablas
- ✓ HUE
- ✓ Data Governance

¿PREGUNTAS?

¡Muchas gracias!