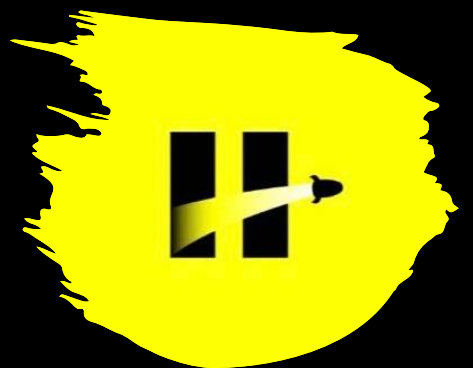
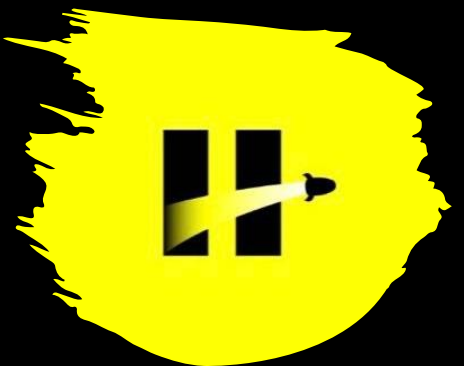
M4
BIG DATA

Clase 3



Objetivos de la clase

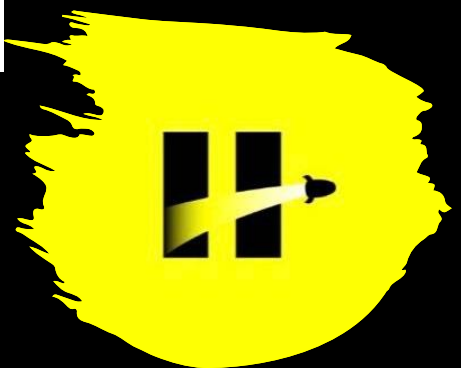
- Implementar una solución de Hive mediante el Uso de Docker
- Conocer Hive
- Tipos de Tablas en Hive
- Tipos de Datos
- Formatos de Almacenamiento
- Particiones
- Hue
- Data Governance



Frameworks Hadoop

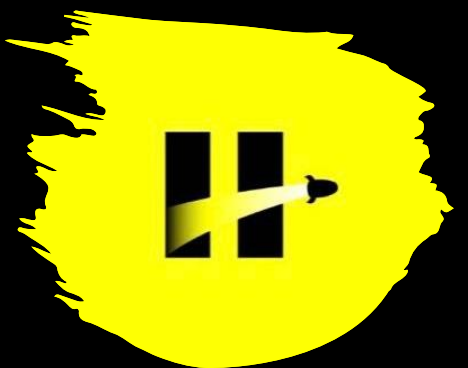
Apache Hive	Consultas SQL sobre Hadoop
Apache Sqoop	Transferencia de datos entre bases relacionales y Hadoop
Apache Spark	Procesamiento en memoria de ETL's, Streaming, Machine Learning y Grafos
Apache Kafka	Sistema de colas de mensajería que utiliza el patrón productor/consumidor
Apache HBase	Base de datos NoSQL de tipo columnar que se ejecuta sobre HDFS
Apache Ranger	Administración de políticas de seguridad sobre componentes de Hadoop
Apache Atlas	Herramienta que provee funcionalidades de Data Governance sobre Hadoop
Apache Nifi	Orquestación de flujos de datos desde y hacia Hadoop

Un Framework es una estructura base utilizada como punto de partida para elaborar un proyecto con objetivos específicos.



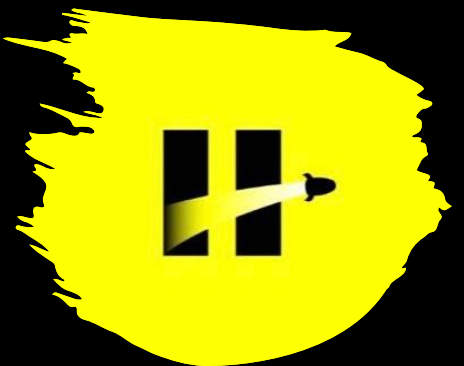
Apache Hive

Hive es una tecnología distribuida diseñada y construida sobre **Hadoop**. Que nos permite hacer consultas y analizar grandes cantidades de datos almacenados en **HDFS**. Tiene un lenguaje de consulta llamado **HiveQL** o HQL que internamente transforma las consultas SQL en trabajos MapReduce que ejecutan en Hadoop. El lenguaje de consulta HQL es un dialecto de SQL, que no sigue el estándar ANSI SQL, sin embargo es muy similar.



Hive

Hive o Apache Hive es una **herramienta de Data Warehousing y ETL (extraer, transformar, cargar)** construida para funcionar sobre Hadoop. Se encuentra, más concretamente, sobre el componente de **YARN (Yet Another Resource Negociator) y HDFS (Hadoop Data File System).**



Hive

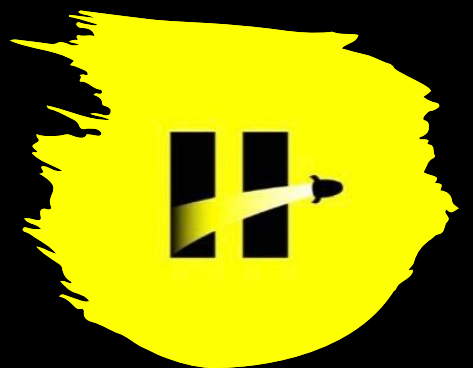
Además de saber qué es Hive, debes conocer que implementar esta herramienta te permite facilitar tres tareas principales dentro del procesamiento de los datos:

•**Análisis tipo SQL de *datasets* muy grandes:** esta es la tarea principal, puesto que es la que **permite que se lleve a cabo el procesamiento de los macrodatos**, es decir, de una gran volumen de datos.

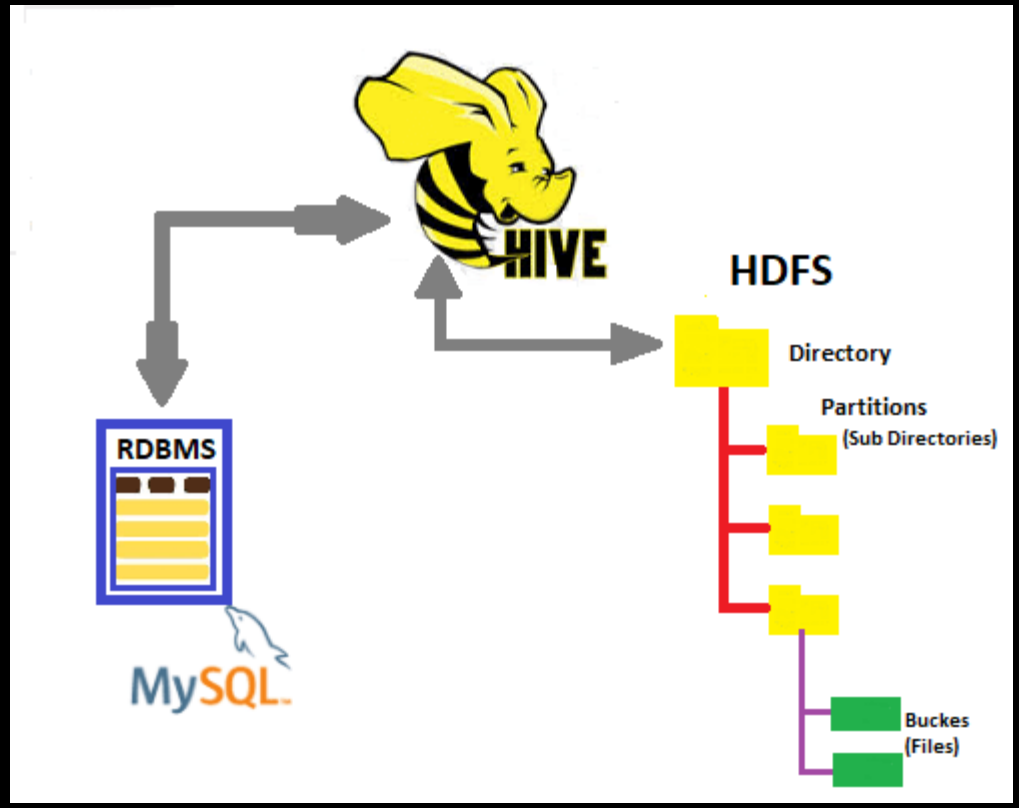
•***Queries Ad-Hoc*:** Apache Hive te permite generar *queries SQL Ad-Hoc* (para este propósito), es decir, **hacer consultas que solo pueden comprenderse a partir de la misma consultas.**

Por otra parte, qué es **Hive es comparable a cualquier arquitectura de sistemas de SQL**. De hecho, podrás contar con que hay un ***layer* de servidor**, en la que se administra el acceso a tablas, al disco, se ceden los permisos a los usuarios, el acceso a los esquemas, etc.

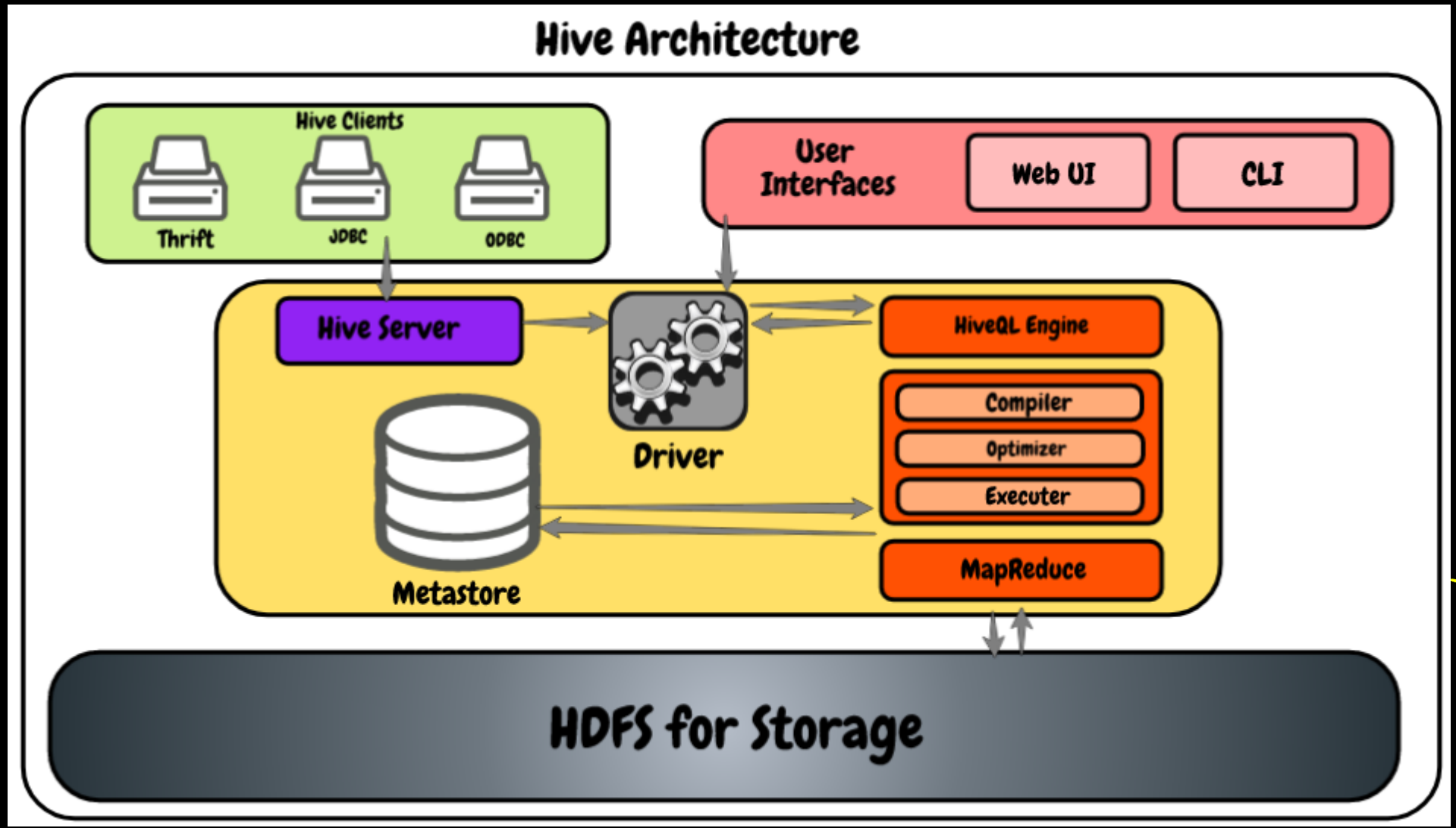
En suma, **en Apache Hive existe otro *layer* que es el cliente**, en la que permite conectar al servidor y ejecutar *queries*, entre otras **acciones más complejas de configuración de *queries* y de almacenamiento.**



Como Funciona Hive



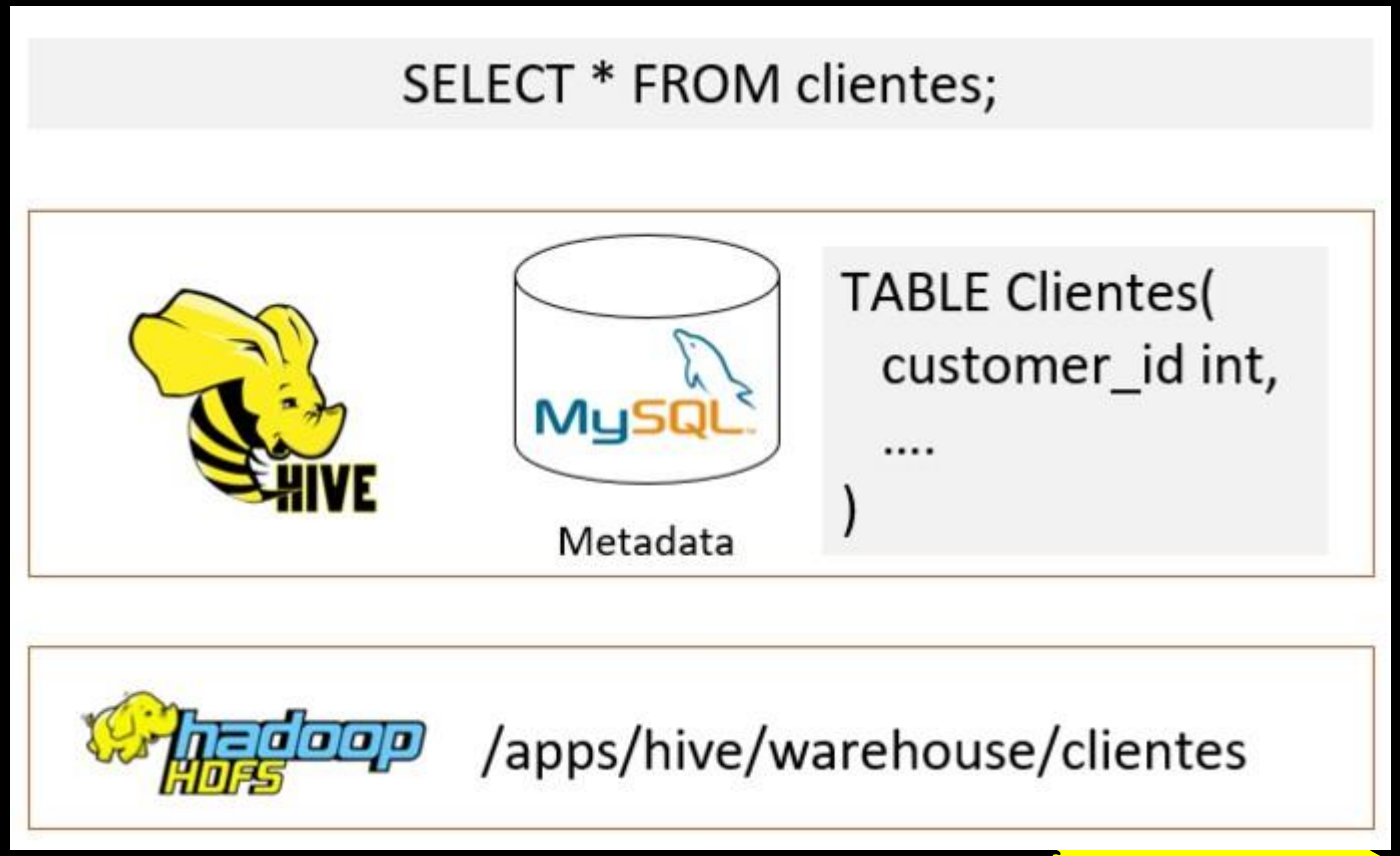
HADOOP - HDFS


Hive Architecture
Hive Clients
Thrift
JDBC
ODBC
User Interfaces
Web UI
CLI
Hive Server
Driver
HiveQL Engine
Compiler
Optimizer
Executer
MapReduce
Metastore
HDFS for Storage

Hive

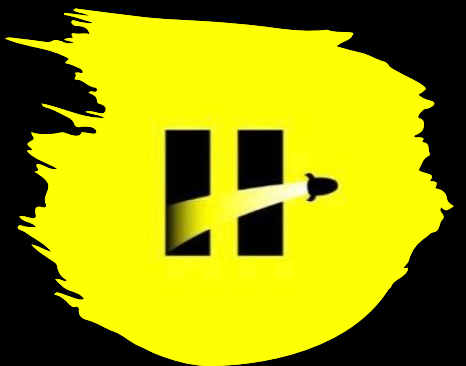
HiveQL (Hive Query Language)

- 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
- 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 en el parámetro hive.metastore.warehouse.dir y tipicamente 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

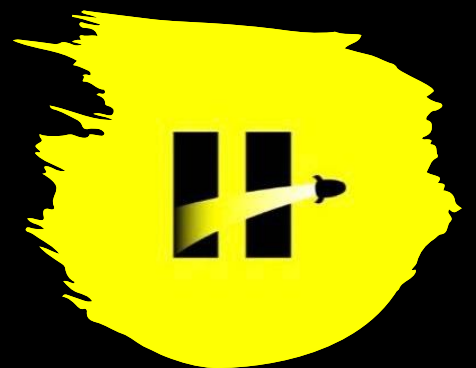


Tipos de Datos

•**Arrays**: Contienen una lista de elementos del mismo tipo. Estos elementos son accedidos por un índice. Por ejemplo, si el campo Mascota contiene a lista de elementos [‘perro’, ’gato’, ‘loro’], el elemento ‘gato’ puede ser accedido mediante Mascota[1].
•**Maps**: Contienen pares clave-valor. Cada elemento es accedido mediante su clave. Por ejemplo, un map lista_passwords conteniendo “Mauricio” como clave y “passDeMauricio” como valor, el password del usuario es accedido mediante lista_passwords[‘Mauricio’].
•**Structs**: Contienen elementos de diferente tipos. Cada elementos puede ser accedido mediante la notación punto (dot notation) . Por ejemplo, en una estructura auto, el color del auto puede ser recuperado mediante auto.color.

Hive, además de los tipos de datos comunes a todos los motores de bases de datos relacionales:

Tipo	TINYINT	1 byte
SMALLINT	2 byte	
INT	4 byte	
BIGINT	8 byte	
FLOAT	4 byte	
DOUBLE	8 byte	
BOOLEAN	TRUE/FALSE	
STRING	tamaño máximo 2GB. TIMESTAMP almacenar fecha y hora DATE almacenar solo fecha	

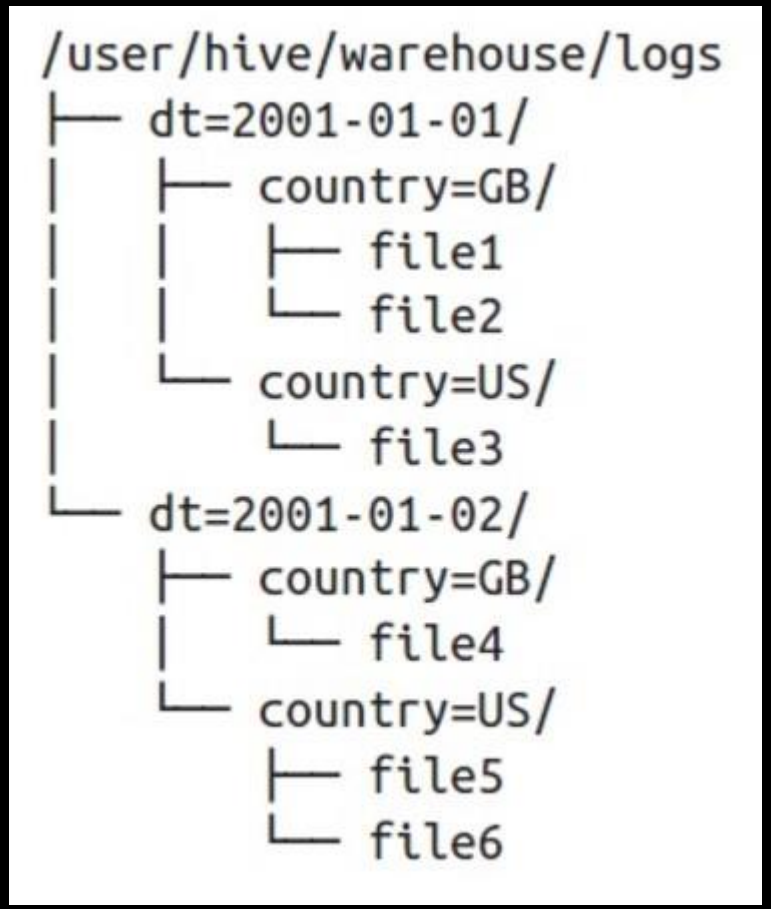


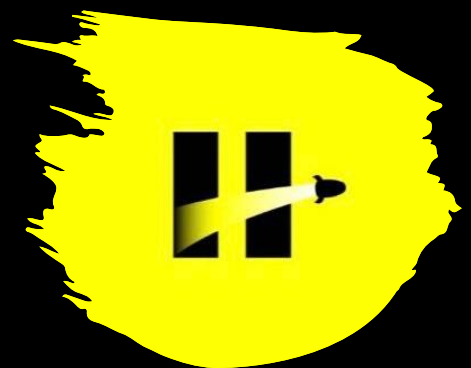
Formatos de Almacenamiento

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 mas eficientes.



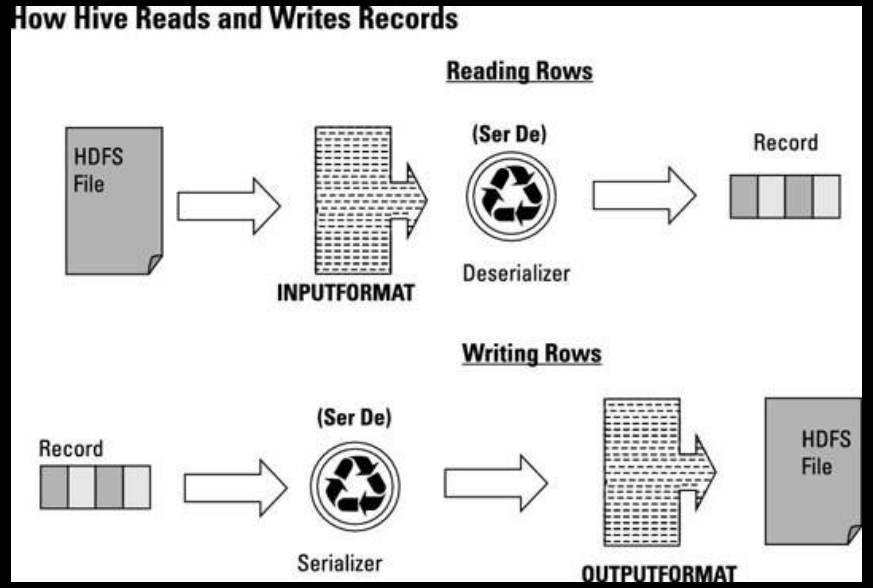


Hive SerDes

El interfaz SerDe permite indicarle a Hive como debe procesar un registro. SerDe es una combinación de Serializer y Deserializer.

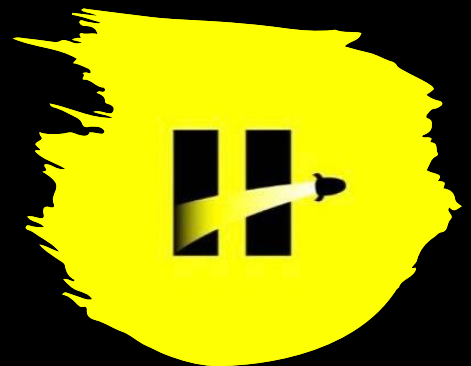
· **Deserializer** toma una representación string o binaria y lo convierte a un objeto Java que Hive puede manipular.

· **Serializer:** toma un objeto Java y lo convierte en algo que Hive puede escribir a HDFS.



SerDes disponibles en Hive:

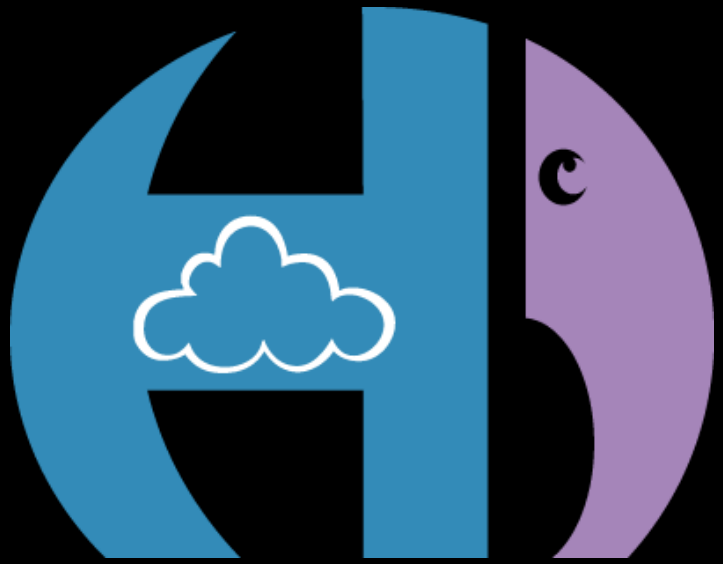
- Avro (Hive 0.9.1 and later)
- ORC (Hive 0.11 and later)
- RegEx
- Thrift
- Parquet (Hive 0.13 and later)
- CSV (Hive 0.14 and later)
- JsonSerDe (Hive 0.12 and later in hcatalog-core)

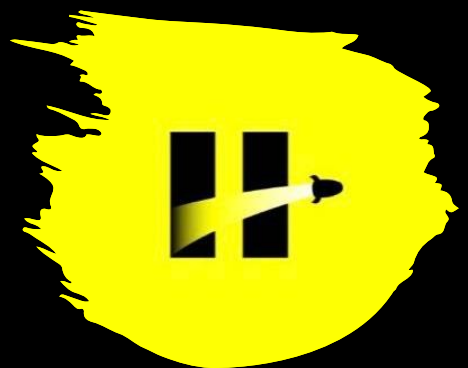


HUE

Hue (Hadoop User Experience)
Es una interfaz web que permite ejecutar consultas SQL hacia diferentes motores de bases de datos, principalmente relacionados a Big Data.

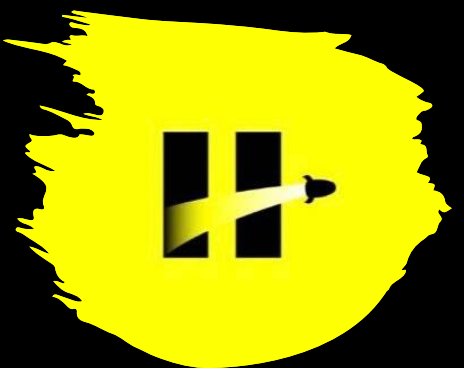
- Bases de datos soportadas (https://docs.gethue.com/administrator/configuration/connectors/)
- Entorno de prueba gratuito (https://demo.gethue.com/hue/accounts/login?next=/)





Que Es Hue?

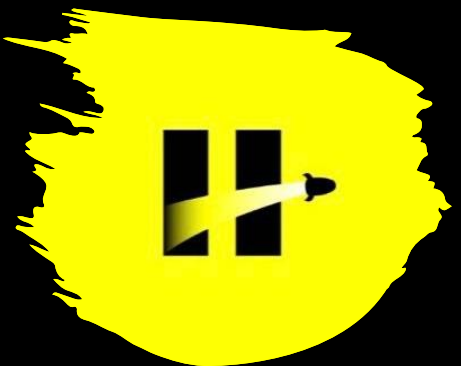
Para **simplificar el proceso de creación, mantenimiento y ejecución de muchos tipos de trabajos de Hadoop**, Hue (Hadoop User Experience) ofrece una **GUI web** para los usuarios de Hadoop. Básicamente, se compone de varias aplicaciones que interactúan con los componentes de Hadoop, y también tiene un SDK abierto para permitir la creación de nuevas aplicaciones.
En las primeras versionse de Hadoop no hubo ninguna interfaz de usuario diseñada, pero muchas de las operaciones en el ecosistema de Hadoop se operan a través de una interfaz de línea de comandos solamente. Entonces, una interfaz de usuario web que realiza algunas de las actividades comunes con el ecosistema Hadoop o los marcos basados en Hadoop es lo que llamamos Hue**. Básicamente, un marco de Hadoop de código abierto llamado Cloudera lanzó y desarrolló Hue.**



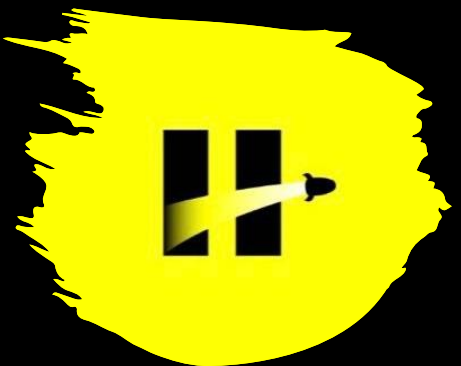
Data Governance

Que Es?

Es un concepto que 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.
Relacionado en gran medida en velar por la **calidad del mismo desde el momento de su generación**, ya que 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.



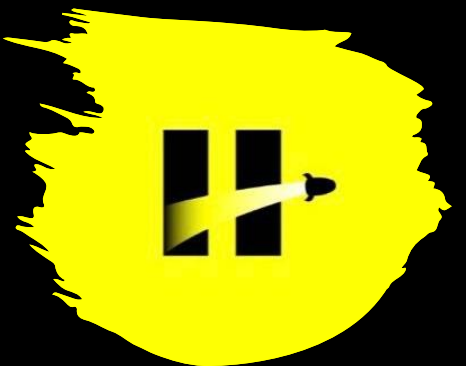
La Gobernanza de datos es la definición de las normas y el control sobre la gestión de datos, en términos de planificación, ejecución y seguimiento. Detrás de las operaciones está la estrategia. La gobernanza de datos es la capacidad de gestionar los datos como un verdadero activo empresarial. En otras palabras, al igual que la gestión de cualquier otro activo, por ejemplo, un depósito, una maquinaria, un servicio innovador o, en general, cualquier otro elemento que tenga valor para la empresa, los datos también necesitan unas reglas básicas para que puedan producir valor económico.



Que Se busca?

La gobernanza de los datos significa utilizar la influencia de la organización para garantizar que los datos se gestionen correctamente. ¿Significa esto crear gastos generales para las actividades de gestión? No necesariamente. Ciertamente, se trata de garantizar que haya claridad:

- **sobre el significado de cada dato;**
- **en las responsabilidades de la empresa y no en las de la informática;**
- **sobre los criterios para definir una figura de calidad;**
- **en garantizar que todo esto se conozca en toda la organización.**



Beneficios

•**Aumento de los ingresos.** La gobernanza de los datos aporta soluciones empresariales sólidas destinadas a aumentar la cuota de mercado, como sofisticados algoritmos de fijación de precios o metodologías para personalizar la experiencia del usuario.

•**La confianza.** Un director general y su línea de mando pueden confiar en los datos que utilizan para tomar decisiones, aumentando así la capacidad de respuesta de la empresa.

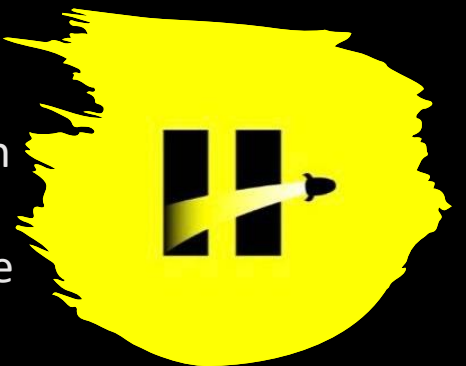
•**Mitigación de riesgos.** Incluso hoy en día, la mayoría de los programas de gobernanza de datos están impulsados por las necesidades de seguridad, privacidad y cumplimiento de la normativa. Gestionar correctamente los datos significa ser capaz de identificar, controlar y anticipar los riesgos.

•**Evitar el despilfarro y las actividades de escaso valor añadido**, como la comprobación de la bondad de los datos y la corrección de errores.

•**Monetización.** Crear productos de datos para venderlos en el mercado a otras organizaciones (por ejemplo, los datos que las compañías telefónicas facilitan para seguir el flujo de personas en una zona geográfica determinada).

•Difusión de conocimientos. Hacer explícito el conocimiento reduce los problemas de diferentes interpretaciones de los mismos datos.

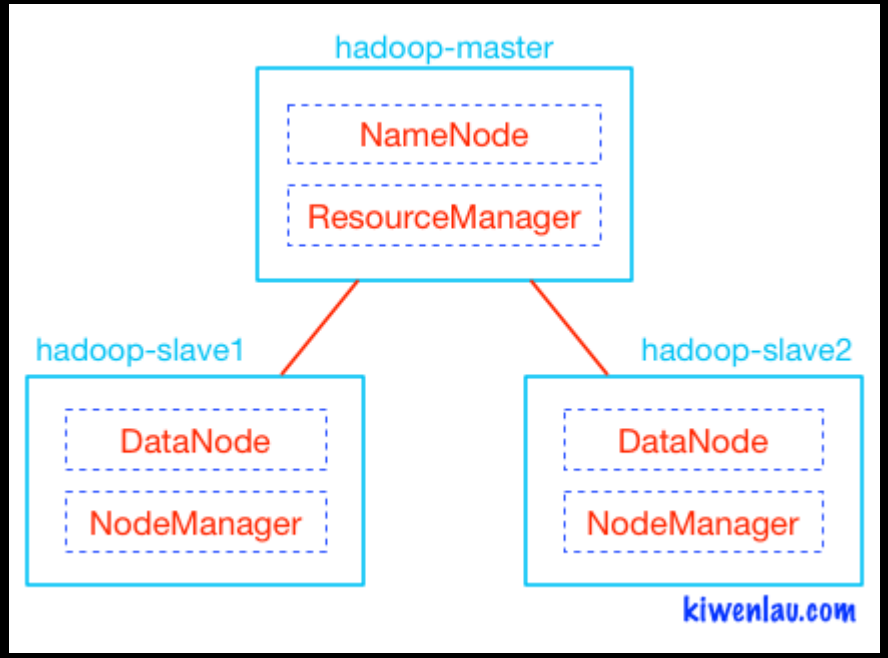
•**Apoyo a los programas empresariales.** La gobernanza de los datos interviene indirectamente en apoyo de los programas estratégicos que están en marcha en la organización, garantizando una gestión adecuada de los datos (por ejemplo, programas de gestión de datos maestros, cambio de ERP, etc.).

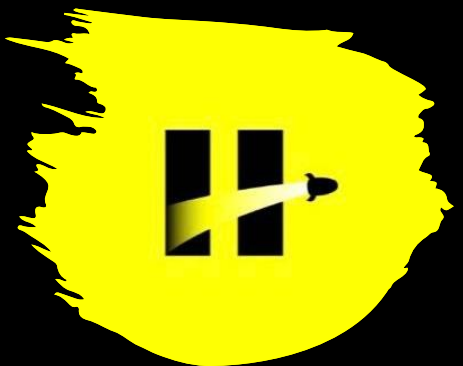


Actores Data Governance

Y que relación tiene esto con Docker?

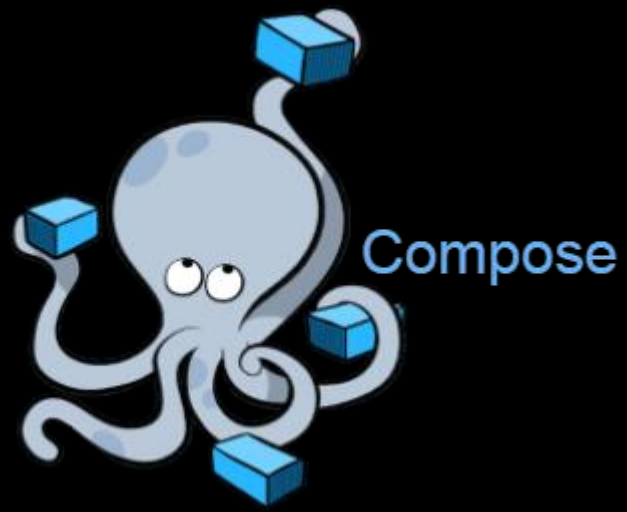
Mediante el uso de Docker podemos crear un entorno de Hadoop con Hue el cual utilizaremos para la practica



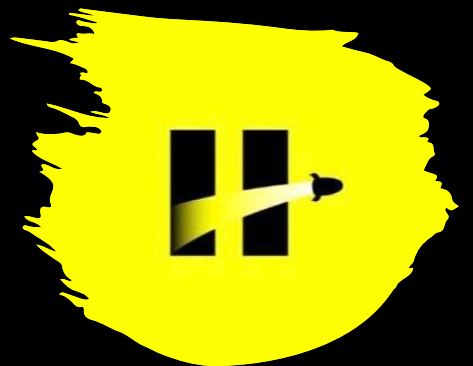


Docker Compose

Docker Compose **es una herramienta para definir y ejecutar aplicaciones de Docker de varios contenedores**

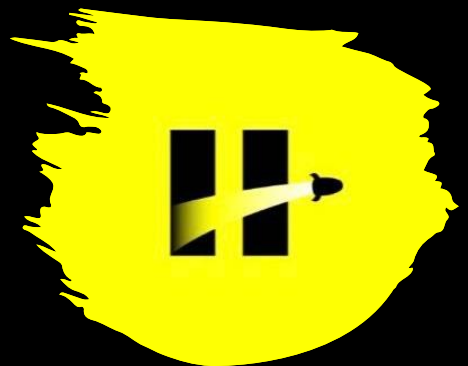


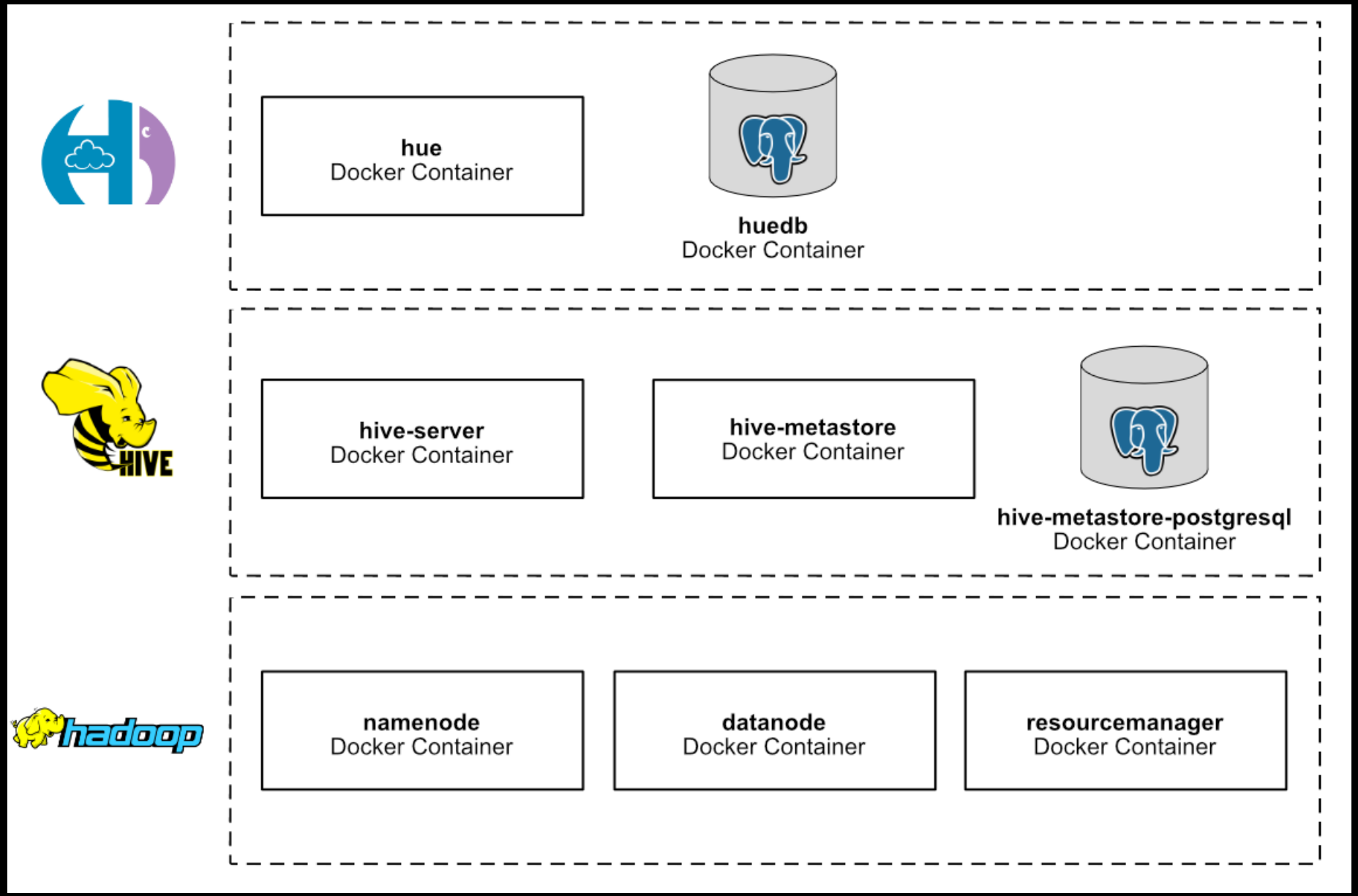
El archivo YAML define todos los servicios que se van a implementar. Estos servicios se basan en un archivo DockerFile o en una imagen de contenedor existente.




docker
Compose

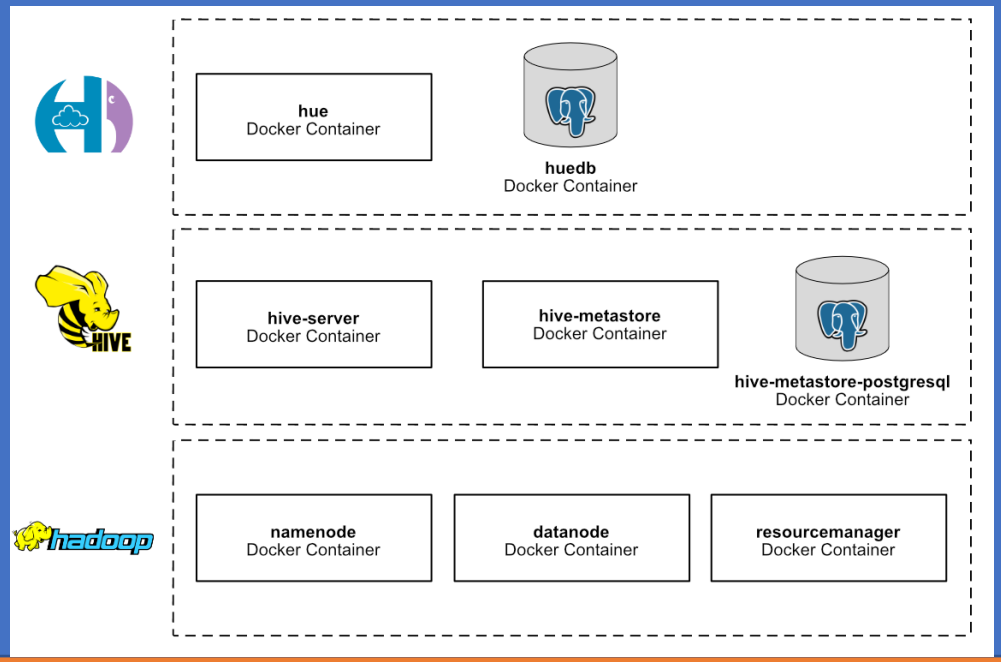

docker




hue
Docker Container
huedb
Docker Container
hive-server
Docker Container
hive-metastore
Docker Container
hive-metastore-postgresql
Docker Container
namenode
Docker Container
datanode
Docker Container
resourcemanager
Docker Container

PC VIRTUAL (LINUX)

DOCKER COMPOSE (HADOOP)



Vamos A GitHub