UAB
Universitat Autònoma de Barcelona

Antonio Espinosa

Apache Hive

Apache Hive

- Introducción
- Arquitectura y modelo de datos
- Uso y ejemplos

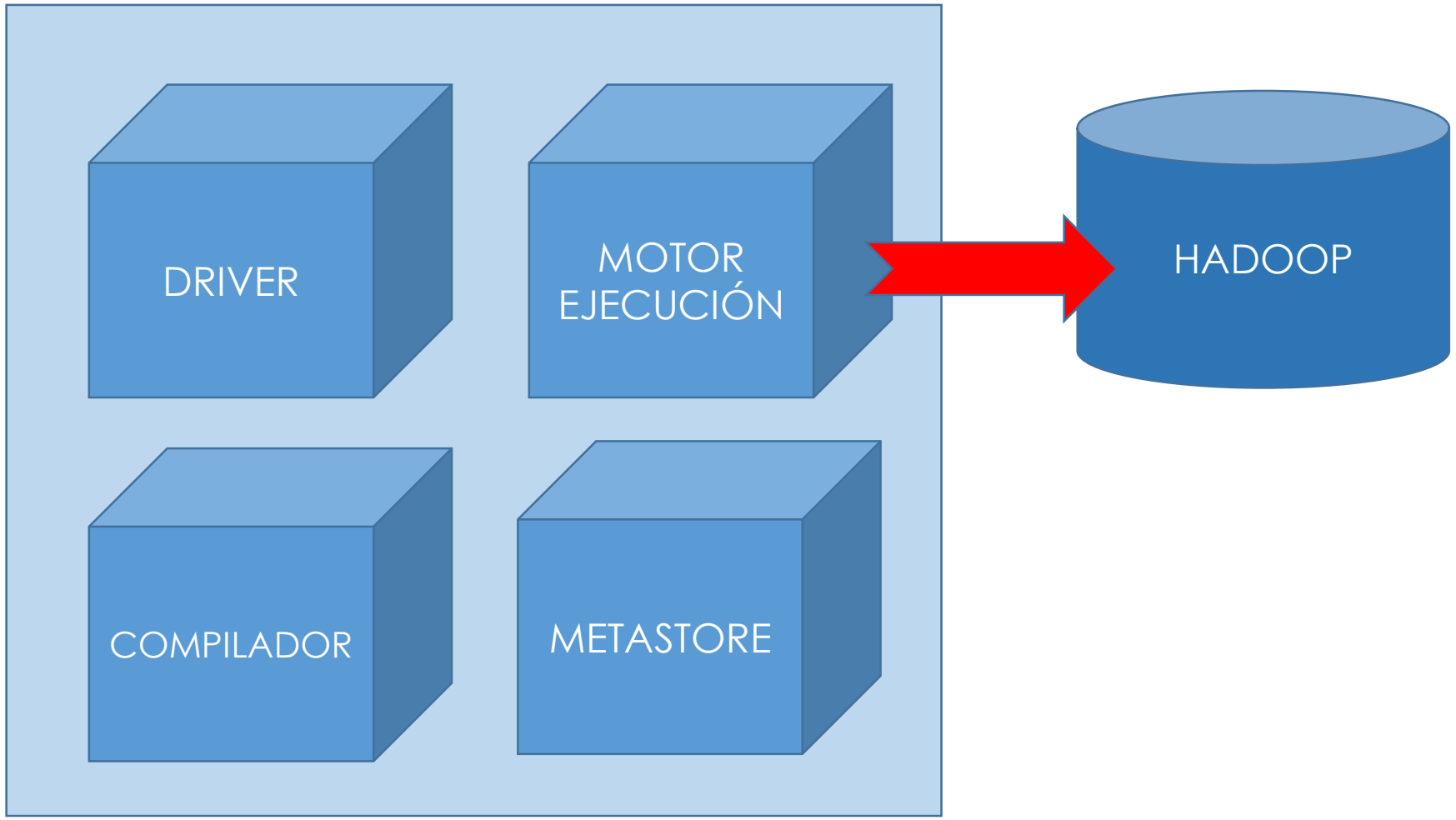
Apache Hive

- Apache Hive es un data warehouse de código abierto
- Sistema que facilita la lectura, escritura y la gestión de conjuntos de datos
- Encima de un sistema de ficheros distribuido HDFS
- Usando consultas SQL
- Mediante el uso de una herramienta de comandos o un driver JDBC

Para qué se usa Hive

- Repositorios de datos “batch”
- Visión resumida de los datos
- Análisis de los datos
- Consultas de datos
- Usando HiveQL que se traduce a trabajos Hadoop MapReduce

Arquitectura Apache Hive



Modelo de datos de Apache Hive

- **Tablas**: similares a las BD relacionales
- **Particiones**: cada tabla puede tener una o más claves de particionado para evitar procesar la tabla entera
- **Buckets**: en cada partición se pueden dividir los datos usando valores de una columna

Aplicaciones

- Informes resumen: agregación diaria de clicks en página web
- Minería de datos: evaluar actividad usuarios en función de sus operaciones
- Análisis ad-hoc: cuantos grupos de administradores según pais/provincia

Ejemplos de uso: importar datos

- Crear base de datos

```
create database clientes;
```

- Crear tabla de facturas

```
create table facturas(id INT, cliente
INT, producto STRING, coste DOUBLE)
row format delimited fields
terminated by ‘,’ stored as textfile;
```

Ejemplos de uso: importar datos

- Insertar datos en la tabla desde HDFS

```
LOAD DATA INPATH 'facturas.txt'
OVERWRITE INTO TABLE facturas;
```

Consultar datos Hive

- Contar número de registros

```
SELECT COUNT(*) from facturas;
```

Definición de particiones en Hive

```
CREATE TABLE paginas(url STRING, dominio STRING,
fecha DATETIME)
PARTITIONED BY (pais STRING)
STORED AS TEXTFILE;
```

- Cada partición corresponde a un valor en particular de país
- Se almacena en un directorio separado en HDFS

Uso de particiones en Hive

```
SELECT pagina.*
FROM paginas
WHERE fecha >='2017-12-24' AND pais = 'UK' AND
paginas.dominio like '%ejemplo.com'
```

Conclusiones

- Solución código abierto para data Warehousing
- Datos ya existentes en ecosistema Hadoop
- Lenguaje muy similar a SQL
- Análisis batch de los datos, no tiempo real
- Soporte para particionar datos

Apache Hive

UAB
Universitat Autònoma
de Barcelona