Hadoop and Spark for Data Scientists

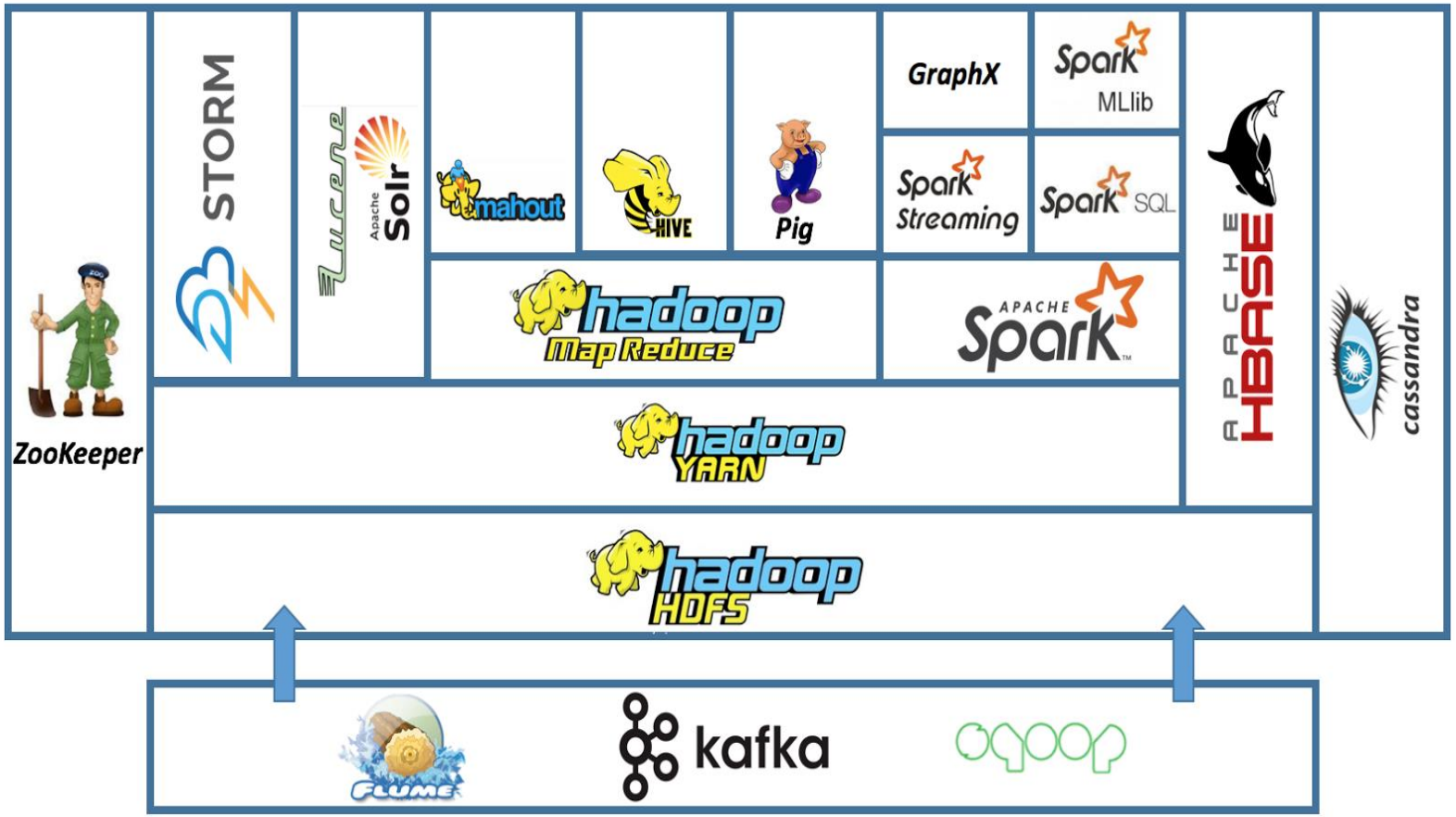
Lecture 2 : Apache Hive

(Data Warehousing at Scale)

Hassan Ahmadkhani

dayche.com . گروه دایکه

Hadoop Ecosystem



Hadoop High-Level Platform for Warehousing, Analysis and ETL

❑Apache Hive

- SQL abstraction layer for running MapReduce and others
- Hive data warehouse software enables reading, writing, and managing large datasets in distributed storage
- Apache Hive is a distributed data warehouse system built on top of Hadoop and is used for analyzing structured and semi-structured data

- Provides a mechanism to perform queries written in HQL (Hive Query Language)
- Users know SQL well
- According to a **Facebook** article, the data scaled from a 15 TB data set in 2007 to a 2 PB data in 2009 and 4 PB per day in 2020

- They(Facebook) needed a **scalable** and **economical** solution
- It is as an efficient ETL / E-LT (Extract, Transform, Load) tool
- Hive is good for :
  - Data Warehousing and ETL/ELT
  - Ad-hoc Analysis

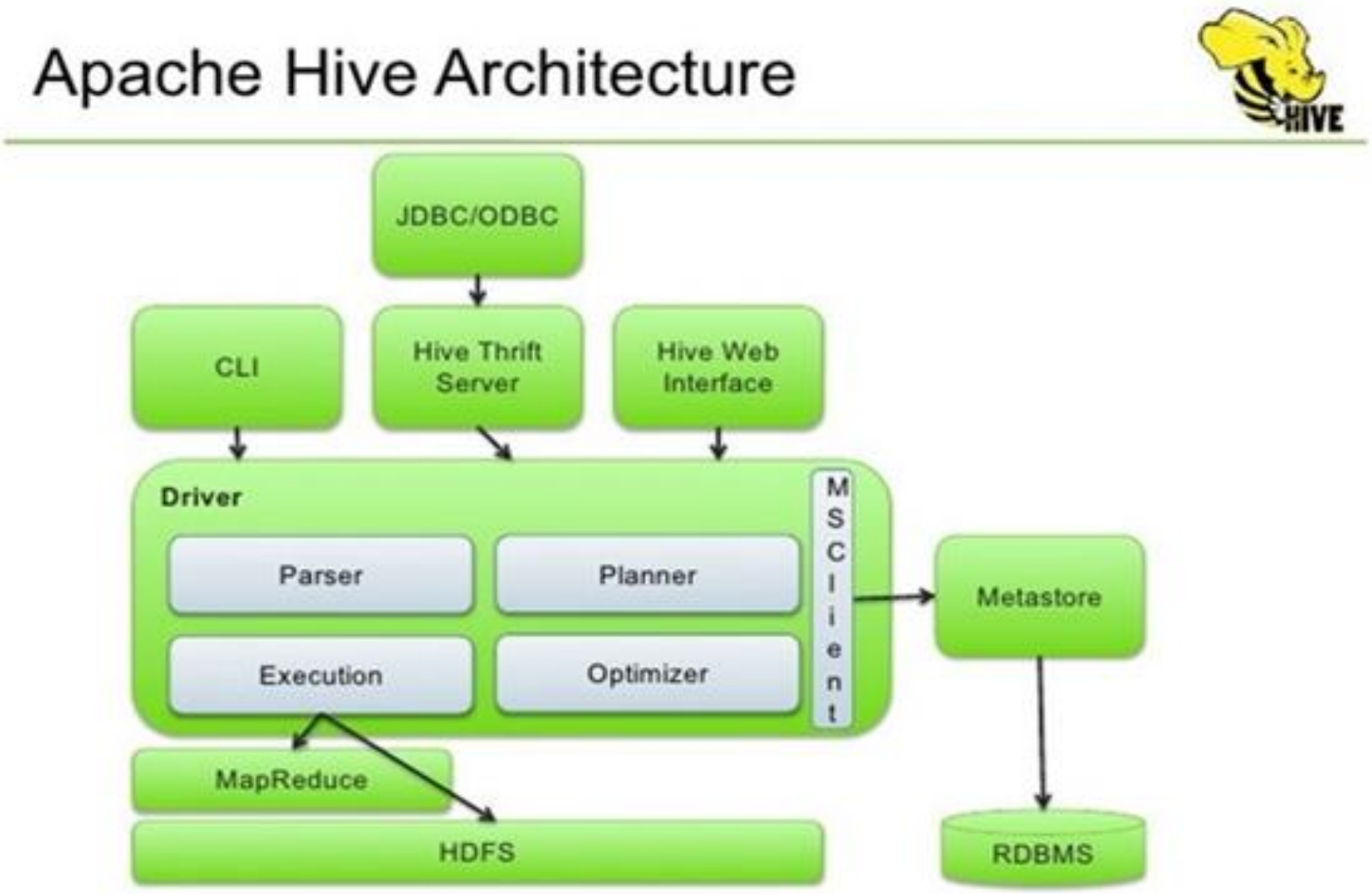
Use Cases for Hive

- Hive on MapReduce or Spark is best-suited for batch data preparation or ETL
- Large ETL sorts with joins to prepare data for Hadoop.
- Most data served to BI users in Impala / Presto is prepared by ETL developers using Hive
- You run data transfer or conversion jobs that take many hours.
- With Hive, if a problem occurs partway through such a job, it recovers and continues
- You receive or provide data in diverse formats,
- Hive **SerDes** and variety of **UDFs** make it convenient to ingest and convert the data
- What HIVE Is Not:
  - Not designed for **OLTP**
  - Does not offer **Real-time queries**

Apache Hive for Warehousing and ETL

- Hive is Good For:
  - Scalable SQL processing over data in Hadoop
  - Scales to 100PB+

Hive	RDBMS
SQL Interface.	SQL Interface.
Focus on analytics.	May focus on online or analytics.
No transactions.	Transactions usually supported.
Partition adds, no random INSERTs. In-Place updates not natively supported (but are possible).	Random INSERT and UPDATE supported.
Distributed processing via map/reduce.	Distributed processing varies by vendor (if available).
Scales to hundreds of nodes.	Seldom scale beyond 20 nodes.
Built for commodity hardware.	Often built on proprietary hardware (especially when scaling out).
Low cost per petabyte.	What's a petabyte?



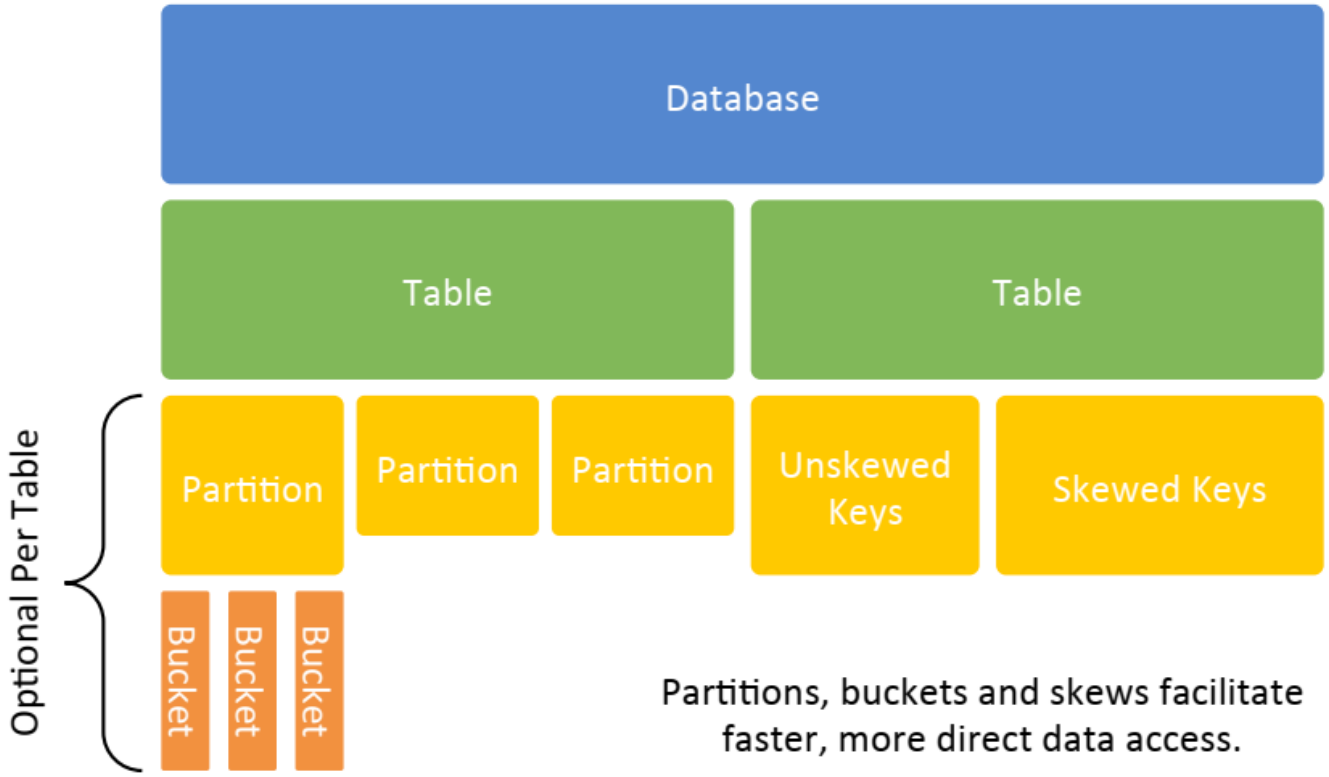
Apache Hive for Warehousing and ETL

❑ Apache Hive Data Types / Statements:

SQL Datatypes	SQL Semantics
INT	SELECT, LOAD, INSERT from query
TINYINT/SMALLINT/BIGINT	Expressions in WHERE and HAVING
BOOLEAN	GROUP BY, ORDER BY, SORT BY
FLOAT	CLUSTER BY, DISTRIBUTE BY
DOUBLE	Sub-queries in FROM clause
STRING	GROUP BY, ORDER BY
BINARY	ROLLUP and CUBE
TIMESTAMP	UNION
ARRAY, MAP, STRUCT, UNION	LEFT, RIGHT and FULL INNER/OUTER JOIN
DECIMAL	CROSS JOIN, LEFT SEMI JOIN
CHAR	Windowing functions (OVER, RANK, etc.)
VARCHAR	Sub-queries for IN/NOT IN, HAVING
DATE	EXISTS / NOT EXISTS
	INTERSECT, EXCEPT

Apache Hive for Warehousing and ETL

- ❑ Apache Hive Data Abstraction:



Partitions, buckets and skews facilitate faster, more direct data access.