**WHAT**

Apache Hive is a Data Warehousing package built on top of Hadoop and is used for data analysis. Hive is targeted towards users who are comfortable with SQL. It is similar to SQL and called HiveQL, used for managing and querying structured data. Apache Hive is used to abstract complexity of Hadoop. This language also allows traditional map/reduce programmers to plug in their custom mappers and reducers. The popular feature of Hive is that there is no need to learn Java.

Insert pic

**WHEN**

Hive, an open source peta-byte scale date warehousing framework based on Hadoop, was developed by the Data Infrastructure Team at Facebook. Hive is also one of the technologies that are being used to address the requirements at Facebook. Hive is  very popular with all the users internally at Facebook and is being used to run thousands of jobs on the cluster with hundreds of users, for a wide variety of applications. Hive-Hadoop cluster at Facebook stores more than 2PB of raw data and regularly loads 15 TB of data on a daily basis.

Let’s look at some of its features that makes it popular and user friendly:

* Allows programmers to plug in custom Mappers and Reducers.
* Has Data Warehouse infrastructure.
* Provides tools to enable easy data ETL.
* Defines SQL-like query language called QL.
* Shared Data Structures: Using HCatalog, a table and storage management layer for Hadoop, Hive metadata is exposed to other data processing tools, including Pig and MapReduce, as well as through a REST API. This allows users to easily read and write data without worrying about where the data is stored, what format it is, or redefining the structure for each tool.
* Faster Batch Processing Hive-on-Spark features the next generation of batch processing for Hive. With queries executed through Apache Spark™, a powerful data processing tool, users will see dramatic performance improvements compared to MapReduce.

**Apache Hive Use Case -**With its familiar interface, Hive is the tool-of-choice for a variety of batch processing workloads, including:

* Data preparation
* ETL
* Data mining
* Ad optimization

**Example:**

**Facebook:** Before implementing Hive, Facebook faced a lot of challenges as the size of data being generated increased or rather exploded, making it really difficult to handle them. The traditional RDBMS couldn’t handle the pressure and as a result Facebook was looking out for better options. To solve this impending issue, Facebook initially tried using Hadoop MapReduce, but with difficulty in programming and mandatory knowledge in SQL, made it an impractical solution. Hive allowed them to overcome the challenges they were facing.

## Where to Use Hive?

Insert pic

## Hive Architecture:

## Insert pic

Hive consists of the following major components:

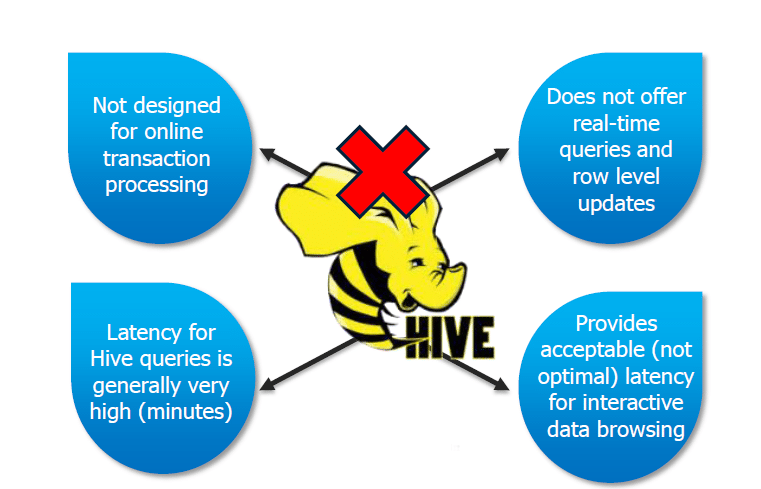
* Metastore – To store the metadata.
* JDBC/ODBC – Query Compiler and Execution Engine to convert SQL queries to a sequence of MapReduce.
* SerDe and ObjectInspectors – For data formats and types.
* UDF/UDAF – For User Defined Functions.
* Clients – Similar to MySQL  command line and a web UI.

## Components of Hive:

**Metastore:**

The Metastore stores the information about the tables, partitions, the columns within the tables. There are 3 ways of storing in Metastore: Embedded Metastore, Local Metastore and Remote Metastore. Mostly, Remote Metastore will be used in production mode.

## Limitations of Hive:



Hive has the following limitations and cannot be used  under such circumstances:

* Not designed for online transaction processing.
* Provides acceptable latency for interactive data browsing.
* Does not offer real-time queries and row level updates.
* Latency for Hive queries is generally very high.

**PIG VS HIVE**

Pig and Hive are the two key components of the Hadoop ecosystem. What does pig hadoop or hive hadoop solve? Pig hadoop and Hive hadoop have a similar goal- they are tools that ease the complexity of writing complex java MapReduce programs. However, when to use Pig Latin and when to use HiveQL is the question most of the have developers have. Apache HIVE and Apache PIG components of the Hadoop ecosystem are briefed. If we take a look at diagrammatic representation of the Hadoop ecosystem, HIVE and PIG components cover the same verticals and this certainly raises the question, which one is better? It’s Pig vs Hive (Yahoo vs Facebook).