Hue-Hive-Impala-Pig

Hue is a Web-UI to let users easily access to

Cloudera tools such as Hive, Hbase tables,

HDFS files, Jobs, Users, ...

Hue is introduced by Cloudera.

Default address to access:

– http://quickstart.cloudera:8888/hue

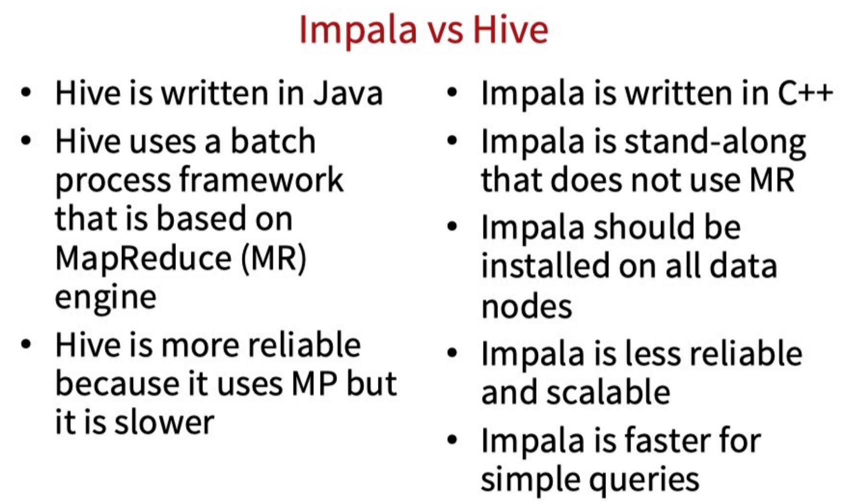
Hive

```
#Simple Exercise:

#– Display tables:
• Show tables;
#– Create table
• CREATE TABLE wordcount (word STRING, freq INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' STORED AS TEXTFILE;
#– Description of wordcount table
• DESCRIBE wordcount;
Load file from HDFS to Hive • Before loading a file into Hive:
#– Make sure that Hive has write access to the folder.
» Hdfs dfs –chmod 777 <folder_name>
• LOAD DATA INPATH "word_output" INTO TABLE wordcount;
# Display the input data:
• SELECT ∗ FROM wordcount;
# Find freq over 2
• SELECT ∗ FROM wordcount WHERE freq > 2 SORT BY freq ASC;
• SELECT freq, COUNT(1) AS f2 FROM wordcount GROUP BY freq SORT BY f2 DESC;
```

Impala

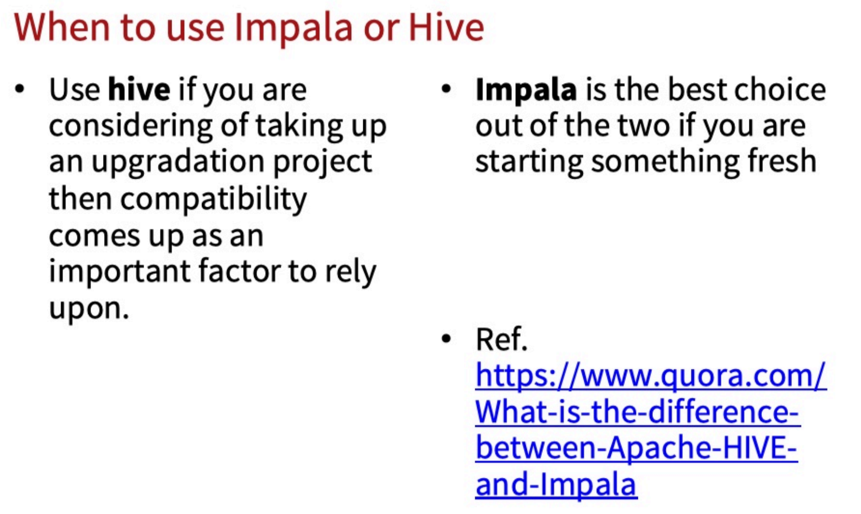
Cloudera Impala is the massively parallel processing (MPP) SQL query engine that runs natively in Apache Hadoop.



Impala vs Hive

- Hive is written in Java
- Hive uses a batch process framework that is based on MapReduce (MR) engine
- Hive is more reliable because it uses MP but it is slower

- Impala is written in C++
- Impala is stand-along that does not use MR
- Impala should be installed on all data nodes
- Impala is less reliable and scalable
- Impala is faster for simple queries



When to use Impala or Hive

- Use **hive** if you are considering of taking up an upgradation project then compatibility comes up as an important factor to rely upon.

- **Impala** is the best choice out of the two if you are starting something fresh
- Ref. https://www.quora.com/What-is-the-difference-between-Apache-HIVE-and-Impala

Pig

• Apache Pig is an abstraction over MapReduce.

• Apache Pig is a framework for analyzing large unstructured and semi-structured data on top of Hadoop.

• To write data analysis programs, Pig provides a high-level language known as Pig Latin.

• Pig Engine, then translates and converts the Pig Latin scripts into MapReduce tasks.

Features of Pig

• Rich set of operators:

– It provides many operators to perform operations like join, sort, filer, etc.

• Ease of programming:

– Pig Latin is similar to SQL and it is easy to write a Pig script if you are good at SQL.

• UDF’s:

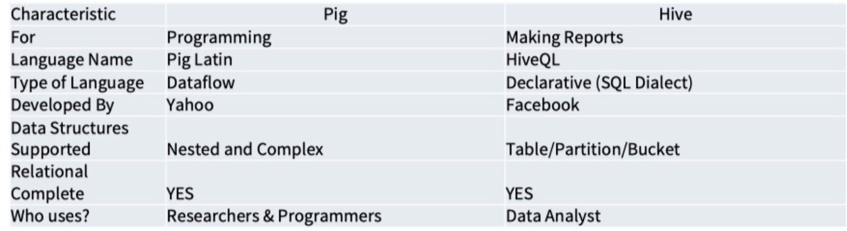
– Pig provides the facility to create UDF(User Defined Function) in other programming languages such as Java and call them in Pig Scripts.

• Handles all kinds of data:

– Apache Pig analyzes all kinds of data, both structured and unstructured.

– It can store the results in HDFS.

```
#Use pig to write a mapreduce
Lines=LOAD ‘input/hadoop.log’ AS (line: chararray);
Words = FOREACH Lines GENERATE FLATTEN(TOKENIZE(line)) AS word;
Groups = GROUP Words BY word;
Counts = FOREACH Groups GENERATE group, COUNT(Words);
Results = ORDER Words BY Counts DESC;
Top5 = LIMIT Results 5; STORE Top5 INTO /output/top5words;
```



Characteristic	Pig	Hive
For	Programming	Making Reports
Language Name	Pig Latin	HiveQL
Type of Language	Dataflow	Declarative (SQL Dialect)
Developed By	Yahoo	Facebook
Data Structures Supported	Nested and Complex	Table/Partition/Bucket
Relational Complete	YES	YES
Who uses?	Researchers & Programmers	Data Analyst

```
Pig # enter the environment
wordcount = LOAD 'wordcount' USING org.apache.hive.hcatalog.pig.HCatLoader();
#watch the table
Dump wordcount
SELECT freq, COUNT(1) AS f2 FROM wordcount GROUP BY freq SORT BY f2 DESC;
grpd = GROUP wordcount BY freq
cntd = FOREACH grpd GENERATE group, COUNT(wordcount) AS cnt;
fltrd = FILTER cntd BY cnt > 1;
#Watch results
Dumpt fltrd
#Store results
STORE fltrd INTO 'filtered_wc';
```