Impala - Overview

What is Impala?

- **MPP** (Massive Parallel Processing) SQL query engine.
- Fast, interactive SQL queries directly into HDFS, HBase or S3.
- Impala can read almost all the file formats such as Parquet, Avro, RCFile used by Hadoop.
- Distributed architecture based on daemon processes.

Impala vs Hive

- Impala uses the same metadata, SQL syntax (Hive SQL), ODBC driver, and user interface (Hue Beeswax) as Apache Hive.
- Each daemon is responsible for all the aspects of query execution that run on the same machine.
- By avoiding MapReduce Impala is faster than Apache Hive.
- However, Hive is best suited for long running batch jobs, e.g. ETL jobs.

Advantages of Impala

- Query data in HDFS at lightning-fast speed with traditional SQL knowledge.
- No data movement required: data processing is done where the data resides.

Impala vs Relational Databases

Feature	Impala	Relational DB
Language	SQL-like (HiveQL)	Uses SQL
Update/delete individual records	No	Yes
Support for transactions	No	Yes
Indices	No	Yes
Amount of data	Petabytes	Teradata

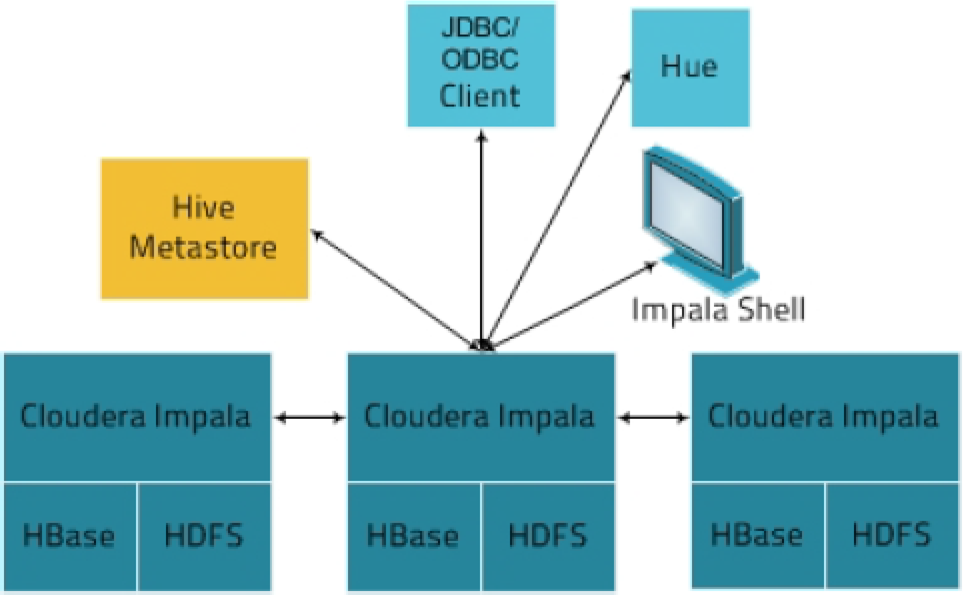
Impala vs HBase vs Hive

HBase	Hive	Impala
Wide-column store database	Data warehouse	Manage, analyze data
Data model is wide-column store	Relational data model	Relational data model
Schema-free	Schema-based	Schema-based
NOSQL	NOSQL	NOSQL
Open source	Open source	Open source

Drawbacks of Impala

- Impala can only read text files, not custom binary files.
- Whenever new records / files are added to the data directory in HDFS, the table needs to be refreshed.
- No support for triggers/transactions/indexing.

How does Impala fit in Cloudera?



How it all works together? (cont.)

- **Clients:** Hue, BI tools, R, Python, shell.
- **Metastore:** as you play around with Impala SQL, Hive Metastore is automatically updated with information about data available.
- **Impala:** Each process, running on the nodes, coordinates and executes queries. Nodes act as workers, executing query fragments in parallel.
- **Storage:** HBase and HDFS data to be queried.

Concepts and Architecture

Impala Daemon

- Represented physically by the `impalad` process.
- Reads and writes to data files.
- Accepts queries transmitted from the impala-shell command, Hue, JDBC, or ODBC.
- Parallelizes the queries and distributes work across the cluster.
- Transmits intermediate query results back to the central coordinator.
- In CDH 5.12 / Impala 2.9 and higher, it is possible to control which hosts act as query coordinators and which act as query executors.

Impala StateStore

- Checks the health of all Impala daemons in a cluster, and communicates its findings with them.
- Physically represented by a daemon process `statestored`.
- Only needed on one host in the cluster.
- If an Impala daemon goes offline due to hardware failure, network error, software issue or other reason, the StateStore informs all other daemons to avoid making requests to it.

Impala Catalog Service

- Relays the metadata changes from Impala SQL statements to all the Impala daemons in a cluster.
- The corresponding process is `catalogd`
- Usually running together with `statestored` on the same host.