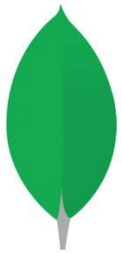
Big data SQL Engines

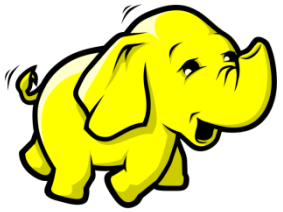
Michael Enudi

Journey through the world of databases and data engineering



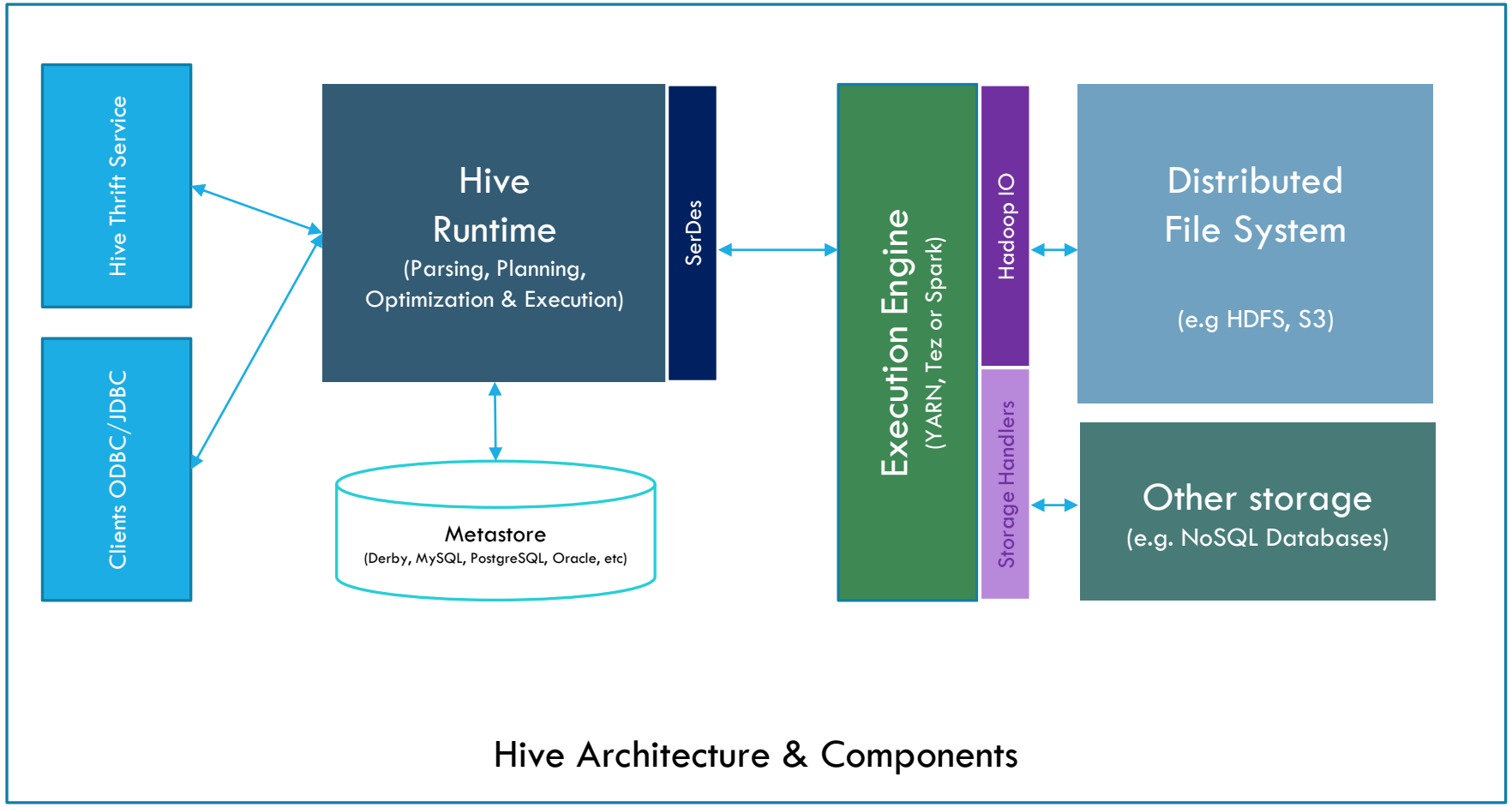




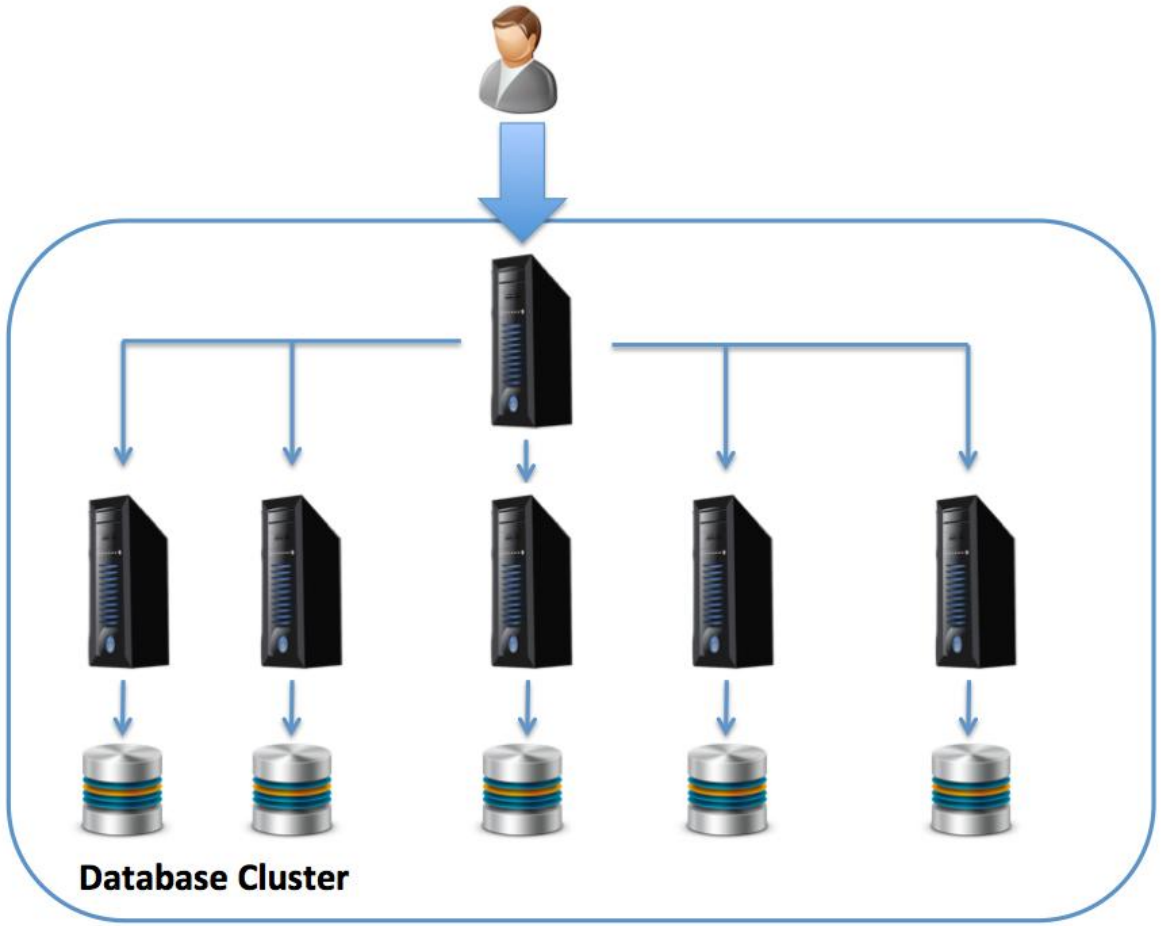

The Apache Hive™ data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL.

Function	MySQL	HiveQL
Selecting a database	`USE database;`	`USE database;`
Listing databases	`SHOW DATABASES;`	`SHOW DATABASES;`
Listing tables in a database	`SHOW TABLES;`	`SHOW TABLES;`
Describing the format of a table	`DESCRIBE table;`	`DESCRIBE (FORMATTED|EXTENDED) table;`
Creating a database	`CREATE DATABASE db_name;`	`CREATE DATABASE db_name;`
Dropping a database	`DROP DATABASE db_name;`	`DROP DATABASE db_name (CASCADE);`



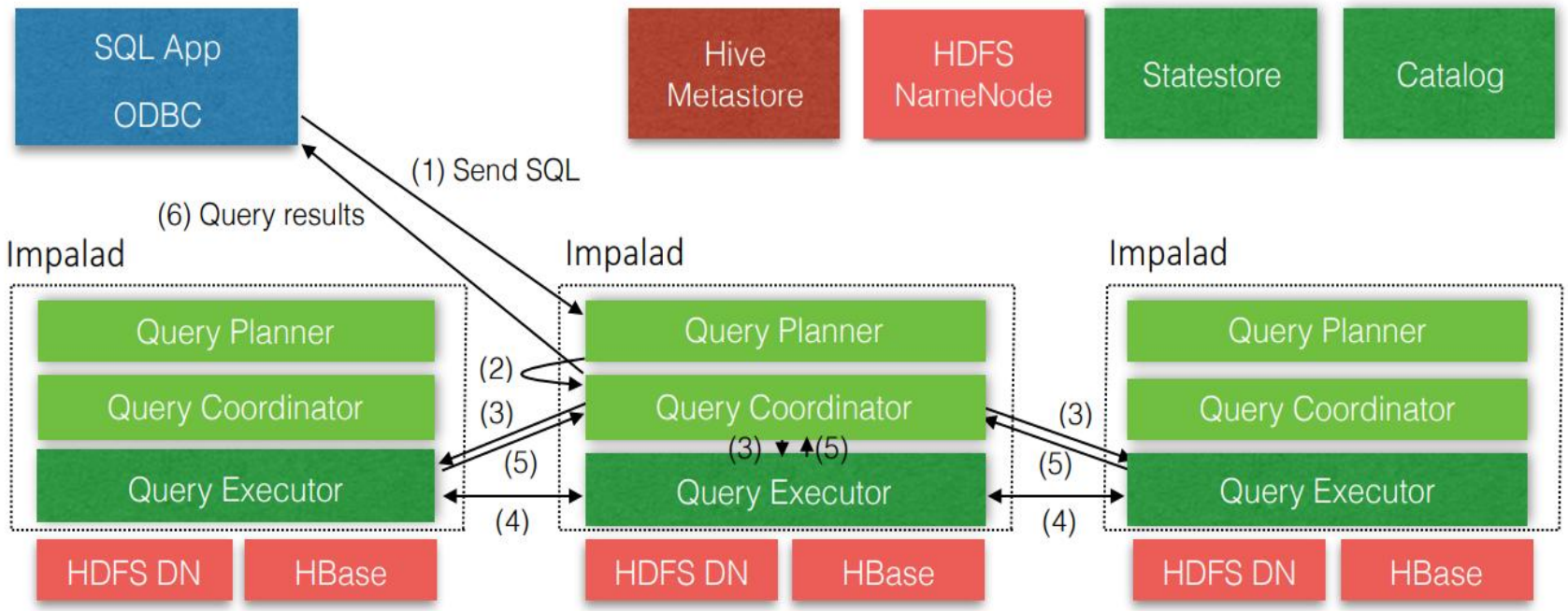
Hive Architecture & Components

MPP System Architecture



Apache Impala is an open source massively parallel processing (MPP) SQL query engine for data stored in Apache Hadoop

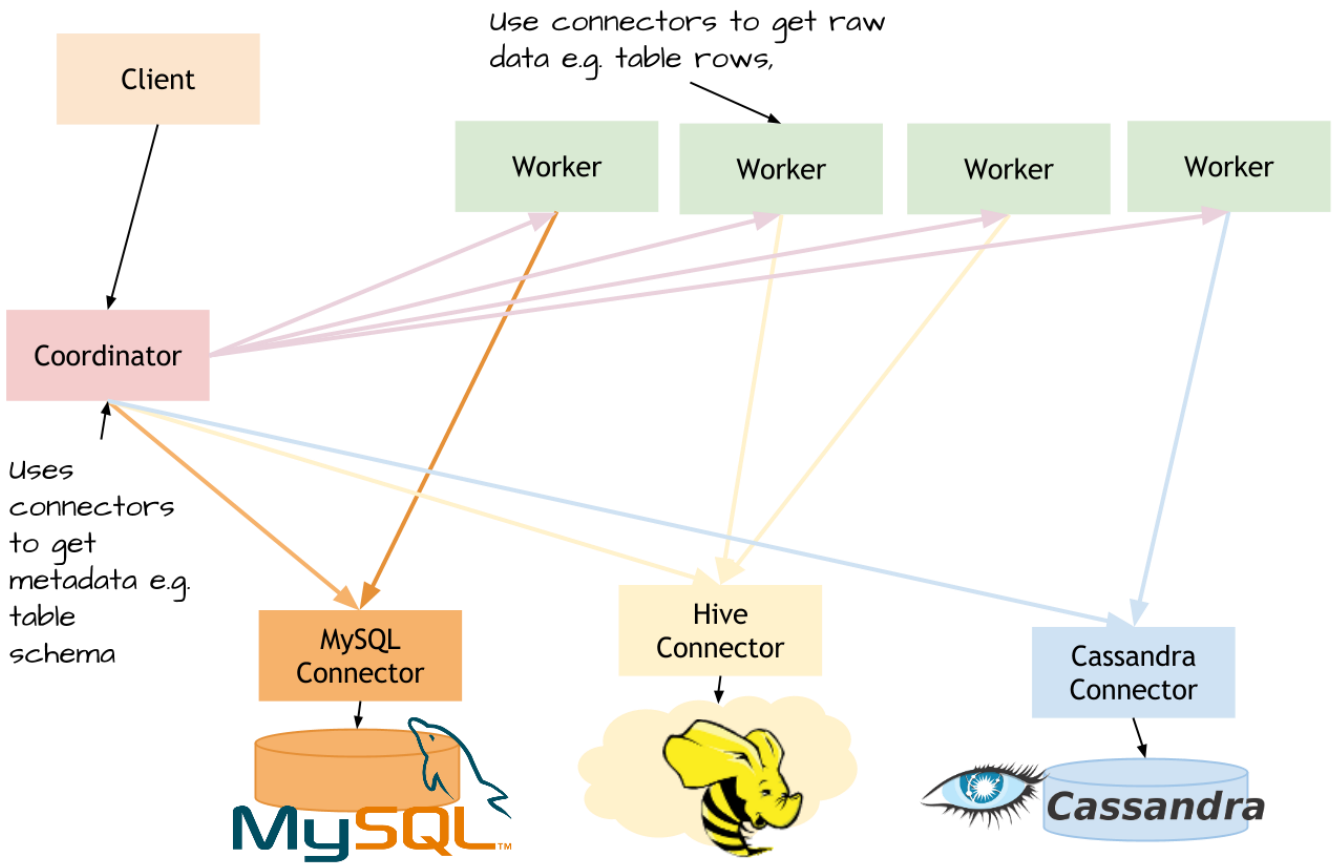
- Written in C++
- Support HDFS and HBase storage
- Uses Hive metastore
- Support ODBC/JDBC client connectivity
- Support simple and nested data structures



Impala Architecture & Components

Presto is a high-performance SQL query engine that was built for high analytical workload kind of operation and write and read data from a variety of sources including NoSQL databases

- Written in java
- Started at Facebook. Currently open-sourced
- Uses connector architecture to integrate with many storage platform include Relational and NoSQL systems.
- Support client's connectivity with ODBC/JDBC drivers
- Supports simple and nested data structures.



PrestoDB Connector Architecture

BIG DATA SQL-ENGINES WRAP UP