Big Data: **Hadoop EcoSystem**

**HDFS** Data Storage

* Scaleable and Fault tolerant due to Replication
* Namenode: Metadata
* Datanodes: Payload

Data processing Frameworks and Query Engines

* **Spark** (IMC) multiple capabilities MR batch processing, real time streaming, interactive queries
* **HiveQL** converts and executes SQL to Java based MR commands (long running batch oriented queries)
* **Impala** uses MPP for high speed interactive SQL query execution on top of Hive infra (metastore)

**Impala**

* Running as a service from Fusion Insights controller on specific servers
* All nodes run with impalad daemons process with one catalogD daemon as controller
* All metadata (table structure and partition details) is saved in cache of each Impala node

To add to other confusions, Spark and Hadoop often work together with Spark processing data that sits in HDFS, Hadoop’s file system. But they are distinct and separate entities, each with their own pros and cons and specific business-use cases.