Hive Vs Impala

Omid Vahdaty, Big Data ninja

Differences of Hive VS. Impala

	Hive	Impala
Author	Apache	Cloudera/Apache
design	Map reduce jobs	MPP database
Use cases	Hive which transforms SQL queries into MapReduce or Apache Spark jobs under the covers, **is great for long-running ETL jobs (for which fault tolerance is highly desirable**; for such jobs, you don't want to have to re-do a long-running query that failed after several hours)	Impala is a MPP analytic database on top of Hadoop and is largely written in C++ for speed, pushes data processing down to local DataNodes, avoiding network bottlenecks. **enables low-latency/interactive queries, especially under multi-user load**. This makes Impala very popular with data analysts who need and expect an **interactive "BI" experience**

Differences of Hive VS. Impala

	Hive	Impala
Read/write	parallel	Read in parallel, write on 1 virtual disk - may change.
Resource management	Yarn	128GB per node.Yarn supported.
SQL syntax	HiveSQL	?
Performance	Disk	In memory, All heavy calculations like group by, conversions would be memory based.
Querying	May start in a delay (batch jobs)	No delay
Query fault tolerance	Will restart on failure.	Start over in failure.

Differences of Hive VS. Impala

	Hive	Impala
Complex data types	yes	no
Anti pattern	Interactive / ad hoc.	?