Impala vs Hive: Difference between Sql on Hadoop components

06 Nov 2015

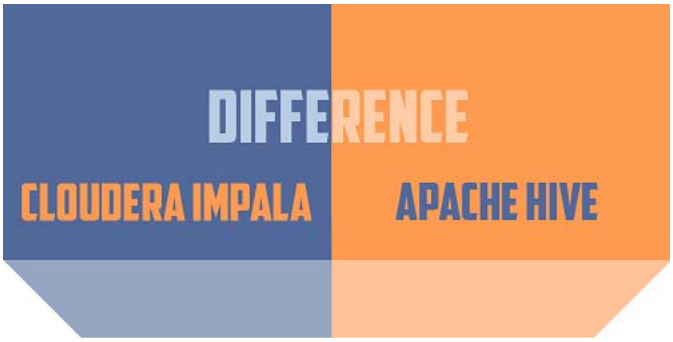
Latest Update made on January 10,2016.

Hadoop has continued to grow and develop ever since it was introduced in the market 10 years ago. Every new release and abstraction on Hadoop is used to improve one or the other drawback in data processing, storage and analysis. Apache Hive was introduced by Facebook to manage and process the large datasets in the distributed storage in Hadoop. Apache Hive is an abstraction on Hadoop MapReduce and has its own SQL like language HiveQL. Cloudera Impala was developed to resolve the limitations posed by low interaction of Hadoop Sql. Cloudera Impala provides low latency high performance SQL like queries to process and analyze data with only one condition that the data be stored on Hadoop clusters.

Data explosion in the past decade has not disappointed big data enthusiasts one bit. It has thrown up a number of challenges and created new industries which require continuous improvements and innovations in the way we leverage technology.

Big Data keeps getting bigger. It continues to pressurize existing data querying, processing and analytic platforms to improve their capabilities without compromising on the quality and speed. A number of comparisons have been drawn and they often present contrasting results. Cloudera Impala and Apache Hive are being discussed as two fierce competitors vying for acceptance in database querying space. While Hadoop has clearly emerged as the favorite data warehousing tool, the Cloudera Impala vs Hive debate refuses to settle down.

Hive vs Impala -Infographic



Upcoming Live Online Hadoop Tr

16 Aug Sun to Thurs (3 weeks) 6:30 PM - 8:30 PM PST LE

19 Aug Sat and Sun (4 weeks) 7:00 AM - 11:00 AM PST LE

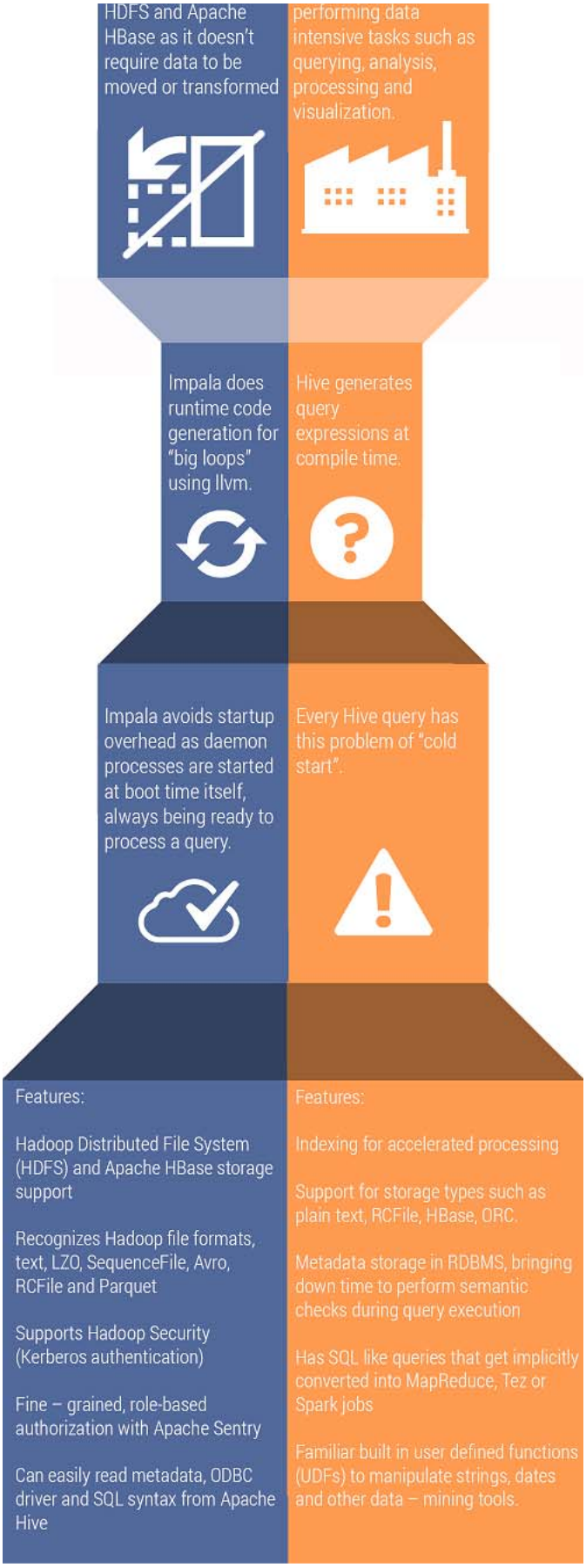
02 Sep Sat and Sun (4 weeks) 7:00 AM - 11:00 AM PST LE

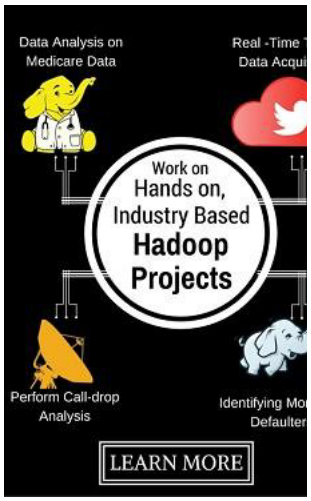
23 Sep Sat and Sun (4 weeks) 7:00 AM - 11:00 AM PST LE

Microsoft Professional Hadoop Certificati

DeZyre

Build Projects, Learn Skills, Get Hired REQUEST INFO





Relevant Courses

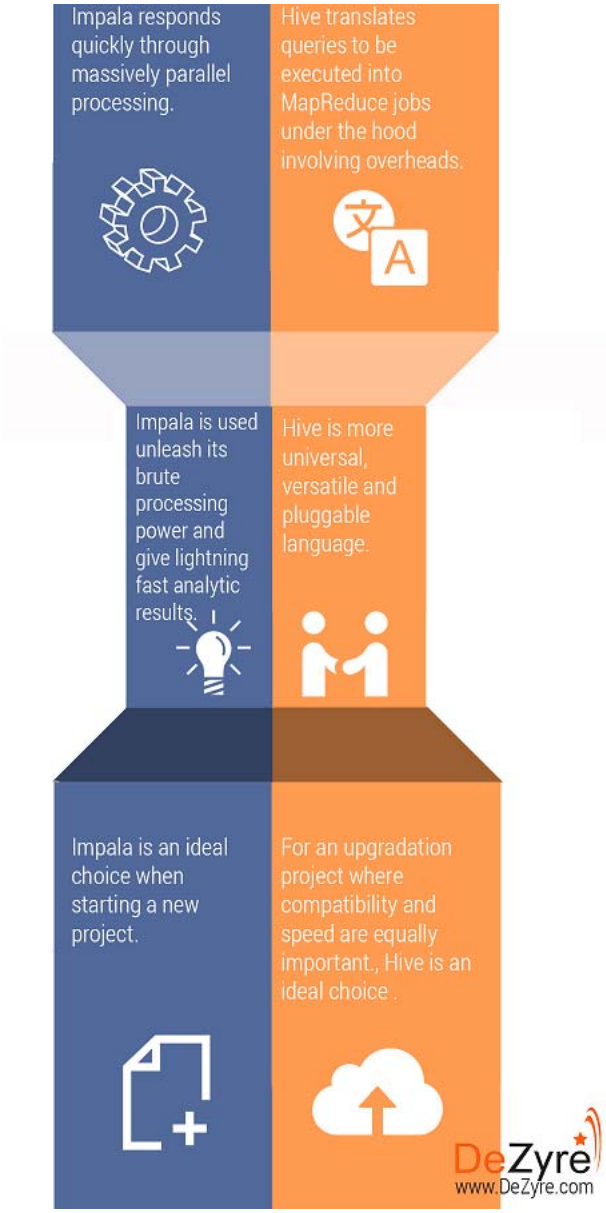
- Hadoop Online Training
- Apache Spark Training
- Data Science in Python Training
- Data Science in R Language Trainir
- Salesforce Certification Training
- NoSQL Database Training
- Hadoop Admin Training

You might also like

- Top 100 Hadoop Interview Questic Answers 2017
- Pig Interview Questions and Answe
- Hive Interview Questions and Ansv

DeZyre

Build Projects, Learn Skills, Get Hired REQUEST INFO



We try to dive deeper into the capabilities of Impala , Hive to see if there is a clear winner or are these two champions in their own rights on different turfs. We begin by prodding each of these individually before getting into a head to head comparison.

What is Impala?

Step aside, the SQL engines claiming to do parallel processing! Impala's open source Massively Parallel Processing (MPP) SQL engine is here, armed with all the power to push you aside. The only condition it needs is data be stored in a cluster of computers running Apache Hadoop, which, given Hadoop's dominance in data warehousing, isn't uncommon. Cloudera Impala was announced on the world stage in October 2012 and after a successful beta run, was made available to the general public in May 2013.

- Real-Time Hadoop Interview Quest and Answers
- Hadoop Admin Interview Question Answers
- Basic Hadoop Interview Questions Answers
- Apache Spark Interview Questions Answers
- Data Analyst Interview Questions a Answers
- 100 Data Science Interview Questic Answers (General)
- 100 Data Science in R Interview Qu and Answers
- 100 Data Science in Python Intervie Questions and Answers
- Recap of Data Science News for Jur
- Recap of Apache Spark News for Ju 2017
- Recap of Hadoop News for June 20
- Top Machine Learning Interview Q and Answers for 2017
- Hadoop Cluster Overview: What it how to setup one?
- Spark SQL for Relational Big Data Processing
- Getting to Know Hadoop 3.0 -Featu Enhancements
- Recap of Data Science News for Ma
- Recap of Apache Spark News for M
- Recap of Hadoop News for May 20

Blog Categories

- Big Data
- CRM
- Data Science
- Mobile App Development
- NoSQL Database

DeZyre

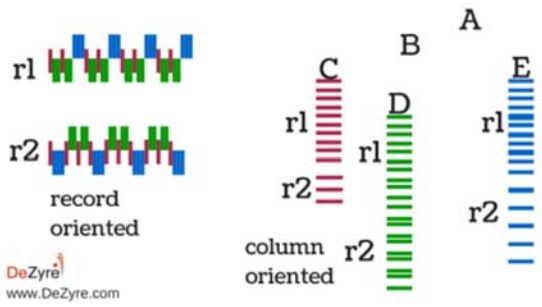
Build Projects, Learn Skills, Get Hired REQUEST INFO

Cloudera Impala is an excellent choice for programmers for running queries on HDFS and Apache HBase as it doesn't require data to be moved or transformed prior to processing. Cloudera Impala easily integrates with Hadoop ecosystem, as its file and data formats, metadata, security and resource management frameworks are same as those used by MapReduce, Apache Hive, Apache Pig and other Hadoop software. It is architected specifically to assimilate the strengths of Hadoop and the familiarity of SQL support and multi user performance of traditional database. Its unified resource management across frameworks has made it the de facto standard for open source interactive business intelligence tasks.

Cloudera Impala has the following two technologies that give other processing languages a run for their money:

Columnar Storage

Data is stored in columnar fashion which achieves high compression ratio and efficient scanning.



Tree Architecture

This is fundamental to attaining a massively parallel distributed multi – level serving tree for pushing down a query to the tree and then aggregating the results from the leaves.

Tutorials

DeZyre

Build Projects, Learn Skills, Get Hired REQUEST INFO

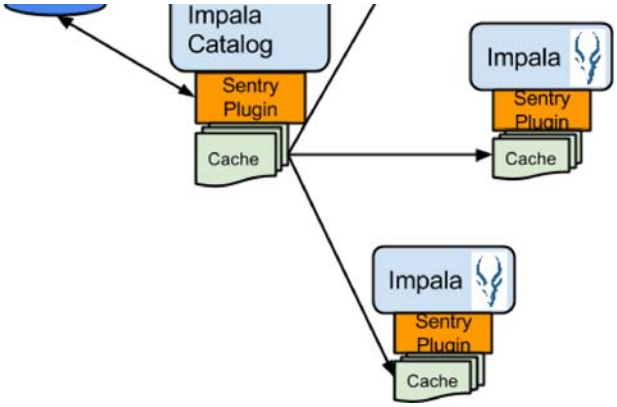


Image Credit:cwiki.apache.org

Impala massively improves on the performance parameters as it eliminates the need to migrate huge data sets to dedicated processing systems or convert data formats prior to analysis. Salient features of Impala include:

- Hadoop Distributed File System (HDFS) and Apache HBase storage support
- Recognizes Hadoop file formats, text, LZO, SequenceFile, Avro, RCFile and Parquet
- Supports Hadoop Security (Kerberos authentication)
- Fine – grained, role-based authorization with Apache Sentry
- Can easily read metadata, ODBC driver and SQL syntax from Apache Hive

Impala's rise within a short span of little over 2 years can be gauged from the fact that Amazon Web Services and MapR have both added support for it.

Apache Hive

Initially developed by Facebook, Apache Hive is a data warehouse infrastructure build over Hadoop platform for performing data intensive tasks such as querying, analysis, processing and visualization. Apache Hive is versatile in its usage as it supports analysis of huge datasets stored in Hadoop's HDFS and other compatible file systems such as Amazon S3. To keep the traditional database query designers interested, it provides an SQL – like language (HiveQL) with schema on read and transparently converts queries to MapReduce, Apache Tez and Spark jobs. Other features of Hive include:

- Indexing for accelerated processing
- Support for different storage types such as plain text, RCFile, HBase, ORC and others
- Metadata storage in RDBMS, bringing down time to perform semantic checks during query execution
- Has SQL like queries that get implicitly converted into MapReduce, Tez or Spark jobs

DeZyre

Build Projects, Learn Skills, Get Hired REQUEST INFO

leverage your familiarity with SQL (without writing MapReduce jobs separately) then Apache Hive is definitely the way to go. HiveQL queries anyway get converted into a corresponding MapReduce job which executes on the cluster and gives you the final output. Hive (and its underlying SQL like language HiveQL) does have its limitations though and if you have a really fine grained, complex processing requirements at hand you would definitely want to take a look at MapReduce.

For the complete list of big data companies and their salaries- CLICK HERE

Difference between Hive and Impala - Impala vs Hive

Impala has been shown to have performance lead over Hive by benchmarks of both Cloudera (Impala's vendor) and AMPLab. Benchmarks have been observed to be notorious about biasing due to minor software tricks and hardware settings. However, it is worthwhile to take a deeper look at this constantly observed difference. The following reasons come to the fore as possible causes:

1. Cloudera Impala being a native query language, avoids startup overhead which is commonly seen in MapReduce/Tez based jobs (MapReduce programs take time before all nodes are running at full capacity). In Hive, every query has this problem of "cold start" whereas Impala daemon processes are started at boot time itself, always being ready to process a query.
2. Hadoop reuses JVM instances to reduce startup overhead partially but introduces another problem when large haps are in use. Cloudera benchmark have 384 GB memory which is a big challenge for the garbage collector of the reused JVM instances.
3. MapReduce materializes all intermediate results, which enables better scalability and fault tolerance (while slowing down data processing). Impala streams intermediate results between executors (trading off scalability).
4. Hive generates query expressions at compile time whereas Impala does runtime code generation for "big loops".
5. Apache Hive might not be ideal for interactive computing whereas Impala is meant for interactive computing.
6. Hive is batch based Hadoop MapReduce whereas Impala is more like MPP database.
7. Hive supports complex types but Impala does not.
8. Apache Hive is fault tolerant whereas Impala does not support fault tolerance. When a hive query is run and if the DataNode goes down while the query is being executed, the output of the query will be produced as Hive is fault tolerant. However, that is not the case with Impala. If a query execution fails in Impala it has to be started all over again.

- Hadoop MapReduce Tutorial
- Big Data Hadoop Tutorial for Begir Hadoop Installation

Online Courses

- Hadoop Training
- Spark Certification Training
- Data Science in Python
- Data Science inR
- Data Science Training

DeZyre

Build Projects, Learn Skills, Get Hired REQUEST INFO

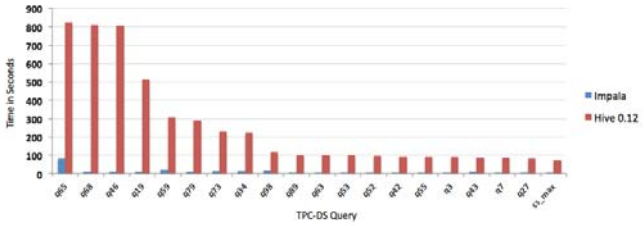


Image Credit : csdn.net

The above graph demonstrates that Cloudera Impala is 6 to 69 times faster than Apache Hive.To conclude, Impala does have a number of performance related advantages over Hive but it also depends upon the kind of task at hand. That being said, Jamie Thomson has found some really interesting results through dumb querying published on sqlblog.com, especially in terms of execution time. For all its performance related advantages Impala does have few serious issues to consider. Being written in C/C++, it will not understand every format, especially those written in java. If you are starting something fresh then Cloudera Impala would be the way to go but when you have to take up an upgradation project where compatibility becomes as important a factor as (or may be more important than) speed, Apache Hive would nudge ahead.

In practical terms, Apache Hive and Cloudera Impala need not necessarily be competitors. As both- Hive Hadoop, Impala have a MapReduce foundation for executing queries, there can be scenarios where you are able to use them together and get the best of both worlds – compatibility and performance. Hive is the more universal, versatile and pluggable language. Once data integration and storage has been done, Cloudera Impala can be called upon to unleash its brute processing power and give lightning fast analytic results.

Learn Hadoop to crunch your organizations big data.

Related Posts

How much Java is required to learn Hadoop?

Top 100 Hadoop Interview Questions and Answers 2016

Difference between Hive and Pig - The Two Key components of Hadoop Ecosystem

Make a career change from Mainframe to Hadoop - Learn Why

PREVIOUS NEXT

DeZyre

Build Projects, Learn Skills, Get Hired REQUEST INFO

Follow

0 Comments DeZyre Login

Recommend Share Sort by Newest

Start the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

Be the first to comment.

Big Data and Hadoop Training Courses in Popular Cities

- Hadoop Training in Texas
- Hadoop Training in California
- Hadoop Training in Dallas
- Hadoop Training in Chicago
- Hadoop Training in Charlotte
- Hadoop Training in Dubai
- Hadoop Training in Edison
- Hadoop Training in Fremont
- Hadoop Training in San Jose
- Hadoop Training in Washington
- Hadoop Training in New Jersey
- Hadoop Training in New York
- Hadoop Training in Atlanta
- Hadoop Training in Canada
- Hadoop Training in Abu Dhabi
- Hadoop Training in Detroit
- Hadoop Trainging in Germany
- Hadoop Training in Houston
- Hadoop Training in Virginia

Courses

Live Courses

About DeZyre

DeZyre

Build Projects, Learn Skills, Get Hired REQUEST INFO

One-on-One Training

Self-Paced Courses

Free Courses

Connect with us

COMODO SECURE