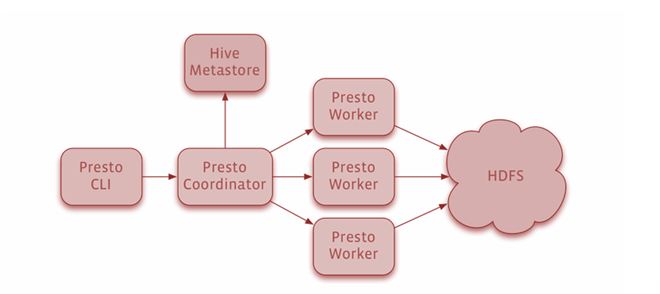
Presto vs Spark

* **What is Presto?**

**Presto is a distributed and open-source SQL query-engine that is used to run interactive analytical queries.** It can handle the query of any size ranging from gigabyte to petabytes. **Presto**was designed by Facebook people. It was designed to speed up the commercial data warehouse query processing. It can scale-up the organizational size matching with Facebook.

**Presto** runs on a cluster of machines. **Presto** setup includes multiple workers and coordinator. The **Presto** queries are submitted to the coordinator by its clients. **Presto** coordinator then analyzes the query and creates its execution plan. Later the processing is being distributed among the workers.



While working with petabytes or terabytes of data the user will have to use lots of tools to interact with HDFS and Hadoop. **Presto** can help the user to query the database through MapReduce job pipelines like Hive and Pig. **Presto** can help the user to operate over different kind of data sources like Cassandra and many other traditional data sources.

***Features of Presto***

* Can help in querying data from its resident location like that can be Hive, Cassandra, proprietary data stores or relational databases.
* Can combine the data of single query from multiple data sources
* The response time of Presto is quite faster and through an expensive commercial solution they can resolve the queries quickly
* It uses vectorized columnar processing
* Presto has pipelined execution
* Its architecture is simple and extensive

Everyday Facebook uses **Presto** to run petabytes of data in a single day. This may include several internal data stores. It also supports pluggable connectors that provide data for queries.

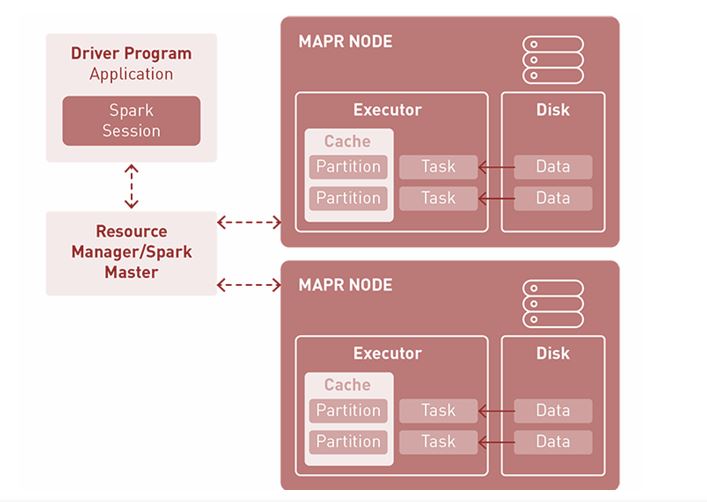
**Presto** supports the following connectors:

* TPC-H
* Cassandra
* Hadoop/Hive

As far as **Presto** applications are concerned then it supports lots of industrial application like Facebook, Teradata and Airbnb.**Presto** supports standard **ANSI SQL** that is quite easier for data analysts and developers. **Presto** is developed and written in Java but does not have Java code related issues like of Memory allocation and garbage collection. **Presto** has a Hadoop friendly connector architecture.

* **What is Spark?**

**Apache Spark is one of the most popular QL engines. It is a general-purpose data processing engine.** There are lots of additional libraries on the top of core **spark** data processing like graph computation, machine learning and stream processing. These libraries can be used together in an application. **Spark** supports the following languages like Spark, Java and R application development.



**Spark** applications run several independent processes that are coordinated by the **SparkSession** object in the driver program. A **Spark** application runs as independent processes that are coordinated by **Spark Session** objects in the driver program. Cluster or resource manager also assigns that task to workers. A task applies its units of work to the dataset, as a result, a new dataset partition is created. Final results are either stored and saved on the disk or sent back to the driver application.

**Spark** can handle petabytes of data and process it in a distributed manner across thousands of clusters that are distributed among several physical and virtual clusters.

**Spark** is being used for a variety of applications like

* Stream Processing
* Machine Learning
* Interactive Analytics
* Data Integration

**Spark** is being chosen by a number of users due to its beneficial features like speed, simplicity and support. **Spark’s** capabilities can be accessed through a rich set of APIs that are designed to specifically interact quickly and easily with data. **Apache Spark** community is large and supportive you can get the answer to your queries quickly and in a faster manner.