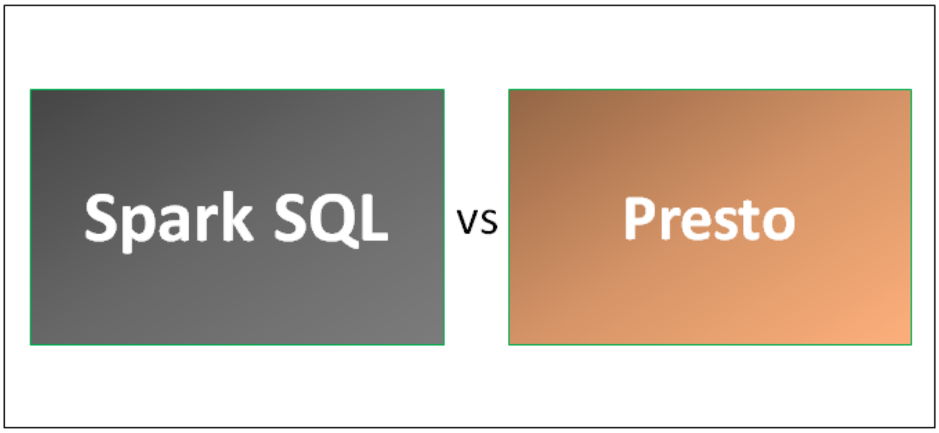
Spark SQL vs Presto | Top 7 Most Useful Distinction You Need to Know



Differences Between to Spark SQL vs Presto

Presto in simple terms is 'SQL Query Engine', initially developed for Apache Hadoop (https://www.educba.com/hadoop-vs-apache-spark/). It's an open source distributed SQL query engine designed for running interactive analytic queries against data sets of all sizes.

Spark SQL (https://www.educba.com/apache-hive-vs-apache-spark-sql/) is a distributed in-memory computation engine with a SQL layer on top of structured and semi-structured data sets. Since its in-memory processing, the processing will be fast in Spark SQL.

Hadoop, Data Science, Statistics & others

Head to Head Comparison Between Spark SQL and Presto (Infographics)

Below are the Top 7 comparison between Spark SQL and Presto:

Presto vs Spark SQL

1#. Eco-Systems / Platforms

Presto

Hadoop, Big Data Processing etc.

Spark SQL

Spark Framework, Big Data Processing etc.

2#. Purpose

Presto

Presto is designed for running SQL queries over Big Data (Huge workloads).

It was designed by Facebook to process their huge workloads

Spark SQL

Spark SQL is one of the components of Apache Spark Core.

Spark Core is the fundamental execution engine for spark platform

3#. Set Up

Presto

- Presto is a distributed SQL query engine for processing pet bytes of data and it runs on a cluster like set up with set of machines.
- A full Presto cluster setup includes a coordinator (Manager Node) and multiple workers. User submits the queries from a client which is the Presto CLI to the coordinator. The coordinator parses, analyzes, and plans the query execution and then it will distribute the query processing to the workers.

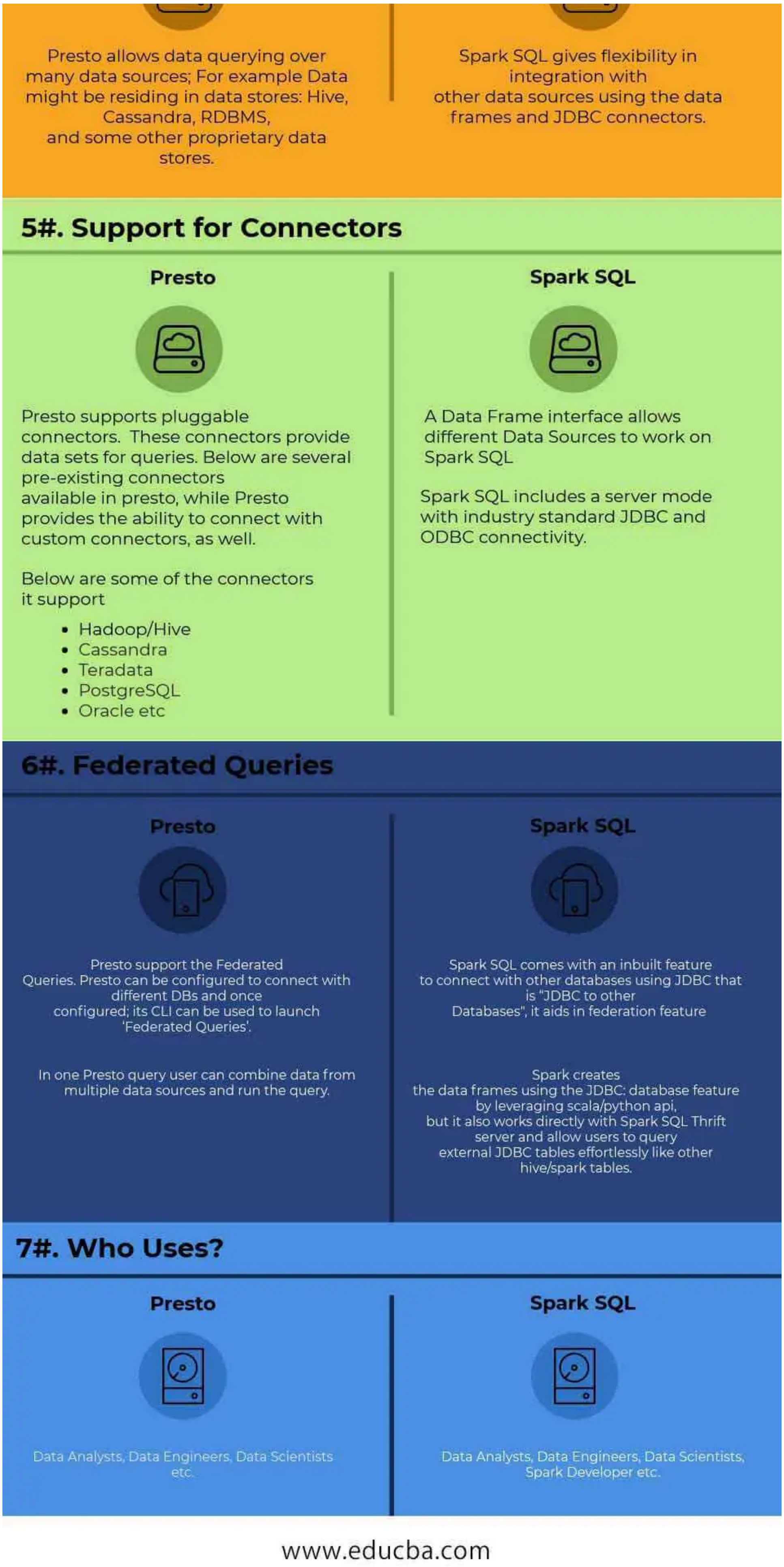
Spark SQL

- Spark SQL setup will be out of the box, if you install and configure Apache Spark Cluster
- Apache Spark is Hadoop's sub-project
- Apaches Spark is a cluster based Big Data processing technology, designed for fast computation.

4#. Capabilities/Features

Presto

Spark SQL



Presto allows data querying over many data sources; For example Data might be residing in data stores: Hive, Cassandra, RDBMS, and some other proprietary data stores.

Spark SQL gives flexibility in integration with other data sources using the data frames and JDBC connectors.

5#. Support for Connectors

Presto

Presto supports pluggable connectors. These connectors provide data sets for queries. Below are several pre-existing connectors available in presto, while Presto provides the ability to connect with custom connectors, as well.

Below are some of the connectors it support

- Hadoop/Hive
- Cassandra
- Teradata
- PostgreSQL
- Oracle etc

Spark SQL

A Data Frame interface allows different Data Sources to work on Spark SQL

Spark SQL includes a server mode with industry standard JDBC and ODBC connectivity.

6#. Federated Queries

Presto

Presto support the Federated Queries. Presto can be configured to connect with different DBs and once configured; its CLI can be used to launch 'Federated Queries'.

In one Presto query user can combine data from multiple data sources and run the query.

Spark SQL

Spark SQL comes with an inbuilt feature to connect with other databases using JDBC that is "JDBC to other Databases", it aids in federation feature

Spark creates the data frames using the JDBC: database feature by leveraging scala/python api, but it also works directly with Spark SQL Thrift server and allow users to query external JDBC tables effortlessly like other hive/spark tables.

7#. Who Uses?

Presto

Data Analysts, Data Engineers, Data Scientists etc.

Spark SQL

Data Analysts, Data Engineers, Data Scientists, Spark Developer etc.

www.educba.com

Key Differences Between Spark SQL and Presto

Below is the list, about the key difference between Presto and Spark SQL:

- Apache Spark (https://www.educba.com/apache-spark-vs-apache-flink/) introduces a programming module for processing structured data called Spark SQL. Spark SQL includes an encoding abstraction called Data Frame which can act as distributed SQL query engine.
- The motive behind the beginning of Presto was to enable interactive analytics and approaches to the speed of commercial data warehouses (https://www.educba.com/10-popular-data-warehouse-tools/) with the power to scale size of organizations matching Facebook.
- Presto was designed as an alternative to tools that query HDFS (https://www.educba.com/hdfs-vs-hbase/) data using MapReduce (https://www.educba.com/mapreduce-interview-questions/) jobs such as Hive (https://www.educba.com/hive-vs-hue/) or Pig, but Presto is not limited to HDFS.
- Spark SQL follows in-memory processing, that increases the processing speed. Spark is designed to process a wide range of workloads such as batch queries, iterative algorithms (https://www.educba.com/software-development/courses/java-course/), interactive queries, streaming etc.
- Presto is capable of executing the federative queries. **Below is the example of Presto Federated Queries**

Let us assume any RDBMS with table sample1

And HIVE with table sample2,

‘Testdb’ is the database in both hive and MYSQL. Using Presto we can evaluate data using in a single query once their connectors are configured correctly as shown below-

```
presto> <Function (select/Group by ..etc)> hive.Testdb.sample2
```

Function (select/Group by ..etc)>mysql.Testdb.sample1

- Spark SQL architecture consists of Spark SQL, Schema RDD, and Data Frame

  - Schema RDD: Spark Core contains special data structure called RDD. Spark SQL works on schemas, tables, and records. Therefore, a user can use the Schema RDD as a temporary table. So that user can call this Schema RDD as Data Frame (https://www.educba.com/data-analyst-interview-questions/)

- Data Frame supports different data formats (CSV, elasticsearch (https://www.educba.com/hadoop-vs-elasticsearch/), Cassandra (https://www.educba.com/hbase-vs-cassandra/) etc) and storage systems (HDFS, HIVE tables, MySQL, etc), It can be integrated with all Big Data (https://www.educba.com/big-data-vs-data-science/) tools/frameworks via Spark-Core and provides API for languages such as Python (https://www.educba.com/python-and-django-for-web-development/), Java, Scala, and R Programming.
- Companies using Presto: Facebook (https://www.educba.com/facebook-vs-twitter/), Netflix, Airbnd, Dropbox (https://www.educba.com/what-is-dropbox/) etc.
- Apache Spark Use Cases can be found in Industries like Finance, Retail, Healthcare, and Travel etc. Many e-commerce websites like eBay, Alibaba (https://www.educba.com/ecommerce-shopping-websites/), Pinterest are using Spark SQL to analyze hundreds of petabytes of data on its e-commerce platform.

Comparisons Table Spark SQL and Presto

Below is the topmost comparison between SQL and Presto.

Basis of comparison between SQL vs Presto	**Presto**	**Spark SQL**
Eco-Systems / Platforms	Hadoop, Big Data Processing etc	Spark Framework, Big Data Processing etc

Purpose	Presto is designed for running SQL queries over Big Data (Huge workloads). It was designed by Facebook to process their huge workloads..	Spark SQL is one of the components of Apache Spark Core. Spark Core is the fundamental execution engine for spark platform
Set up		
Capabilities/Features	Presto allows data querying over many data sources; For example, Data might be residing in data stores: Hive, Cassandra, RDBMS, and some other proprietary data stores.	Spark SQL gives flexibility in integration with other data sources using the data frames and JDBC connectors.
Support for Connectors	Presto supports pluggable connectors. These connectors provide data sets for queries. Below are several pre-existing connectors available in presto, while Presto provides the ability to connect with custom connectors, as well. Below are some of the connectors it support • Teradata (https://www.educba.com/teradata-career/)	A Data Frame interface allows different Data Sources to work on Spark SQL. Spark SQL includes a server mode with industry-standard JDBC and ODBC connectivity.

Federated Queries	Presto supports the Federated Queries. Presto can be configured to connect with different DBs and once configured; its CLI can be used to launch 'Federated Queries'. In one Presto query user can combine data from multiple data sources and run the query.	Spark SQL comes with an inbuilt feature to connect with other databases using JDBC that is "JDBC to other Databases", it aids in federation feature. Spark creates the data frames using the JDBC: database feature by leveraging scala/python API, but it also works directly with Spark SQL Thrift server and allows users to query external JDBC tables effortlessly like other hive/spark tables.
Who Uses?	Data Analysts, Data Engineers, Data Scientists etc	Data Analysts, Data Engineers, Data Scientists, Spark Developer etc

Conclusions

Spark SQL and Presto, both are SQL distributed engines available in the market.

Presto is very helpful when it comes to BI-type queries, and Spark SQL leads performance-wise in large analytics queries. When comparing with respect to configuration, Presto set up easy than Spark SQL. Both Spark SQL and Presto are standing equally in a market and solving a different kind of business problems.

Recommended Articles

This has been a guide to Spark SQL vs Presto. Here we have discussed Spark SQL vs Presto head to head comparison, key differences, along with infographics and comparison table. You may also look at the following articles to learn more –