Spark SQL - Quick Guide

Previous Page Next Page

Spark - Introduction

Industries are using Hadoop extensively to analyze their data sets. The reason is that Hadoop framework is based on a simple programming model (MapReduce) and it enables a computing solution that is scalable, flexible, fault-tolerant and cost effective. Here, the main concern is to maintain speed in processing large datasets in terms of waiting time between queries and waiting time to run the program.

Spark was introduced by Apache Software Foundation for speeding up the Hadoop computational computing software process.

As against a common belief, **Spark is not a modified version of Hadoop** and is not, really, dependent on Hadoop because it has its own cluster management. Hadoop is just one of the ways to implement Spark.

Spark uses Hadoop in two ways – one is **storage** and second is **processing**. Since Spark has its own cluster management computation, it uses Hadoop for storage purpose only.

Apache Spark

Apache Spark is a lightning-fast cluster computing technology, designed for fast computation. It is based on Hadoop MapReduce and it extends the MapReduce model to efficiently use it for more types of computations, which includes

interactive queries and stream processing. The main feature of Spark is its **in-memory cluster computing** that increases the processing speed of an application.

Spark is designed to cover a wide range of workloads such as batch applications, iterative algorithms, interactive queries and streaming. Apart from supporting all these workload in a respective system, it reduces the management burden of maintaining separate tools.

Evolution of Apache Spark

Spark is one of Hadoop's sub project developed in 2009 in UC Berkeley's AMPLab by Matei Zaharia. It was Open Sourced in 2010 under a BSD license. It was donated to Apache software foundation in 2013, and now Apache Spark has become a top level Apache project from Feb-2014.

Features of Apache Spark

Apache Spark has following features.

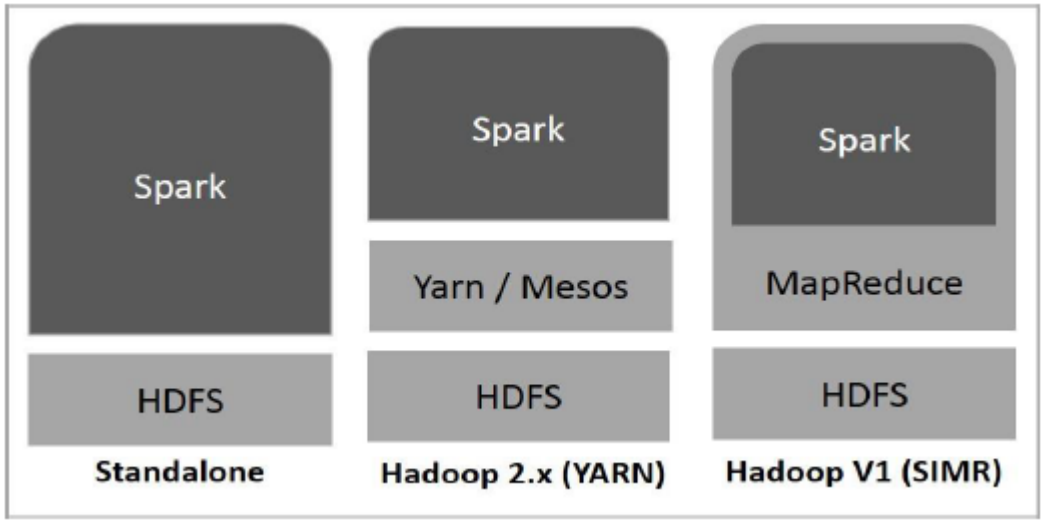
Speed – Spark helps to run an application in Hadoop cluster, up to 100 times faster in memory, and 10 times faster when running on disk. This is possible by reducing number of read/write operations to disk. It stores the intermediate processing data in memory.

Supports multiple languages – Spark provides built-in APIs in Java, Scala, or Python. Therefore, you can write applications in different languages. Spark comes up with 80 high-level operators for interactive querying.

Advanced Analytics – Spark not only supports 'Map' and 'reduce'. It also supports SQL queries, Streaming data, Machine learning (ML), and Graph algorithms.

Spark Built on Hadoop

The following diagram shows three ways of how Spark can be built with Hadoop components.



There are three ways of Spark deployment as explained below.

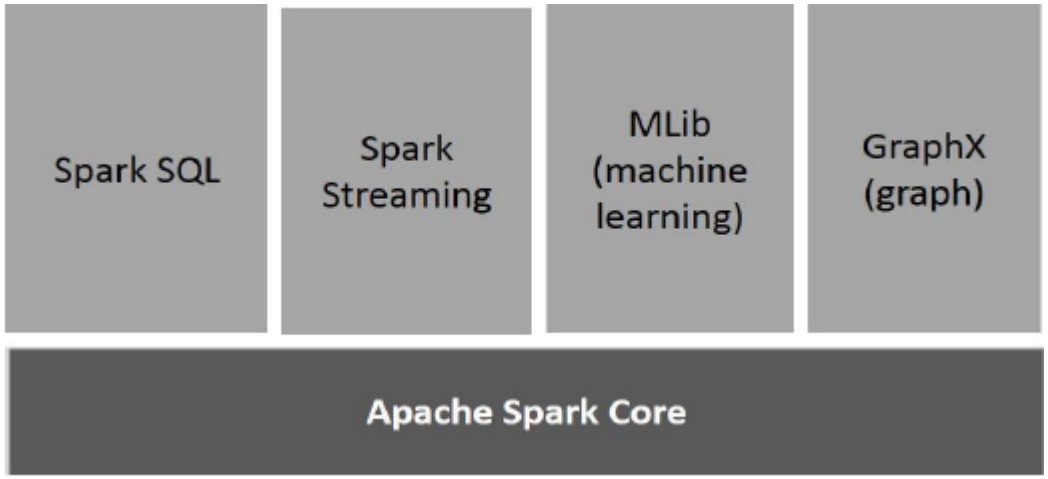
Standalone – Spark Standalone deployment means Spark occupies the place on top of HDFS(Hadoop Distributed File System) and space is allocated for HDFS, explicitly. Here, Spark and MapReduce will run side by side to cover all spark jobs on cluster.

Hadoop Yarn – Hadoop Yarn deployment means, simply, spark runs on Yarn without any pre-installation or root access required. It helps to integrate Spark into Hadoop ecosystem or Hadoop stack. It allows other components to run on top of stack.

Spark in MapReduce (SIMR) – Spark in MapReduce is used to launch spark job in addition to standalone deployment. With SIMR, user can start Spark and uses its shell without any administrative access.

Components of Spark

The following illustration depicts the different components of Spark.



Apache Spark Core

Spark Core is the underlying general execution engine for spark platform that all other functionality is built upon. It provides In-Memory computing and referencing datasets in external storage systems.

Spark SQL

Spark SQL is a component on top of Spark Core that introduces a new data abstraction called SchemaRDD, which provides support for structured and semi-structured data.

Spark Streaming

Spark Streaming leverages Spark Core's fast scheduling capability to perform streaming analytics. It ingests data in mini-batches and performs RDD (Resilient Distributed Datasets) transformations on those mini-batches of data.

MLlib (Machine Learning Library)

MLlib is a distributed machine learning framework above Spark because of the distributed memory-based Spark architecture. It is, according to benchmarks, done by the MLlib developers against the Alternating Least Squares (ALS) implementations. Spark MLlib is nine times as fast as the Hadoop disk-based version of **Apache Mahout** (before Mahout gained a Spark interface).

GraphX

GraphX is a distributed graph-processing framework on top of Spark. It provides an API for expressing graph computation that can model the user-defined graphs by using Pregel abstraction API. It also provides an optimized runtime for this

abstraction.

Spark – RDD

Resilient Distributed Datasets

Resilient Distributed Datasets (RDD) is a fundamental data structure of Spark. It is an immutable distributed collection of objects. Each dataset in RDD is divided into logical partitions, which may be computed on different nodes of the cluster. RDDs can contain any type of Python, Java, or Scala objects, including user-defined classes.

Formally, an RDD is a read-only, partitioned collection of records. RDDs can be created through deterministic operations on either data on stable storage or other RDDs. RDD is a fault-tolerant collection of elements that can be operated on in parallel.

There are two ways to create RDDs – **parallelizing** an existing collection in your driver program, or **referencing a dataset** in an external storage system, such as a shared file system, HDFS, HBase, or any data source offering a Hadoop Input Format.

Spark makes use of the concept of RDD to achieve faster and efficient MapReduce operations. Let us first discuss how MapReduce operations take place and why they are not so efficient.

Data Sharing is Slow in MapReduce

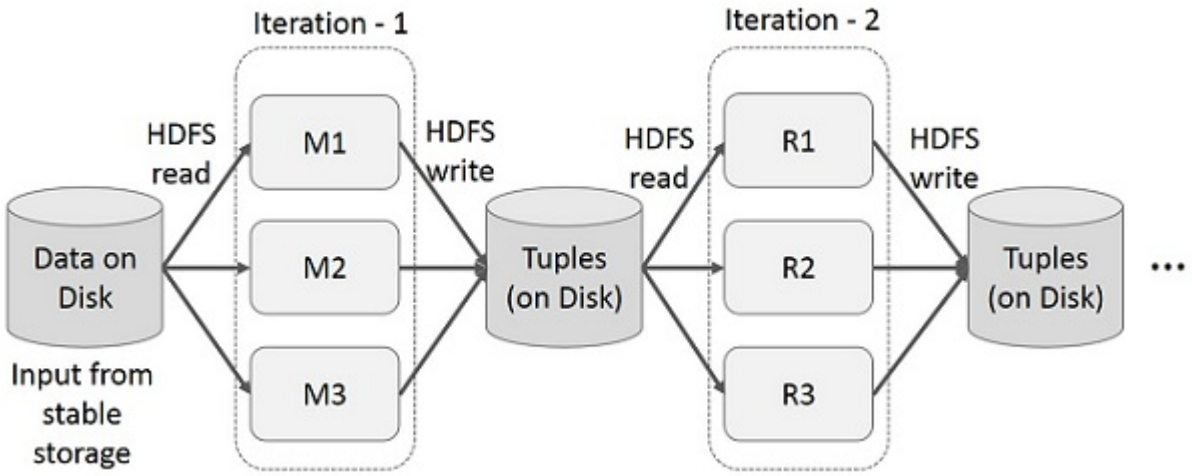
MapReduce is widely adopted for processing and generating large datasets with a parallel, distributed algorithm on a cluster. It allows users to write parallel computations, using a set of high-level operators, without having to worry about work distribution and fault tolerance.

Unfortunately, in most current frameworks, the only way to reuse data between computations (Ex: between two MapReduce jobs) is to write it to an external stable storage system (Ex: HDFS). Although this framework provides numerous abstractions for accessing a cluster's computational resources, users still want more.

Both **Iterative** and **Interactive** applications require faster data sharing across parallel jobs. Data sharing is slow in MapReduce due to **replication**, **serialization**, and **disk IO**. Regarding storage system, most of the Hadoop applications, they spend more than 90% of the time doing HDFS read-write operations.

Iterative Operations on MapReduce

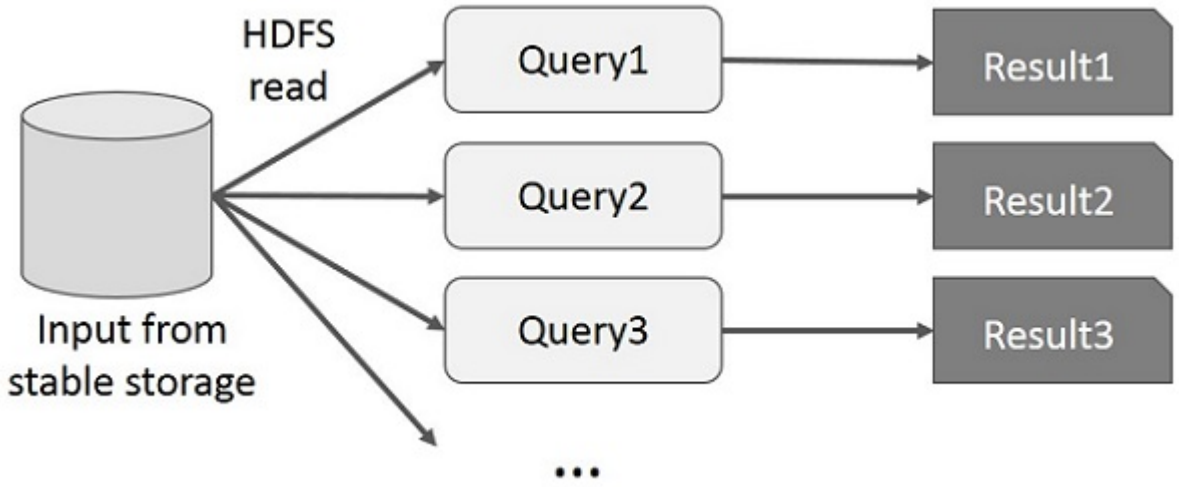
Reuse intermediate results across multiple computations in multi-stage applications. The following illustration explains how the current framework works, while doing the iterative operations on MapReduce. This incurs substantial overheads due to data replication, disk I/O, and serialization, which makes the system slow.



Interactive Operations on MapReduce

User runs ad-hoc queries on the same subset of data. Each query will do the disk I/O on the stable storage, which can dominates application execution time.

The following illustration explains how the current framework works while doing the interactive queries on MapReduce.



Data Sharing using Spark RDD

Data sharing is slow in MapReduce due to **replication**, **serialization**, and **disk IO**. Most of the Hadoop applications, they spend more than 90% of the time doing HDFS read-write operations.

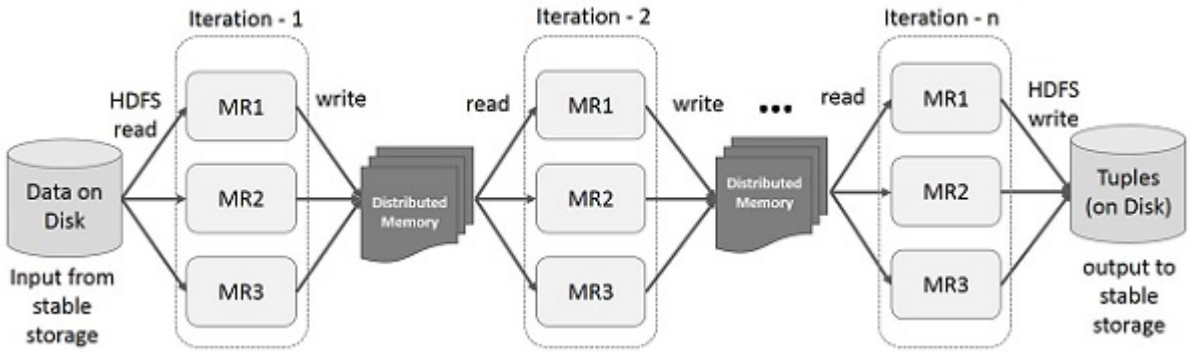
Recognizing this problem, researchers developed a specialized framework called Apache Spark. The key idea of spark is **R**esilient **D**istributed **D**atasets (RDD); it supports in-memory processing computation. This means, it stores the state of memory as an object across the jobs and the object is sharable between those jobs. Data sharing in memory is 10 to 100 times faster than network and Disk.

Let us now try to find out how iterative and interactive operations take place in Spark RDD.

Iterative Operations on Spark RDD

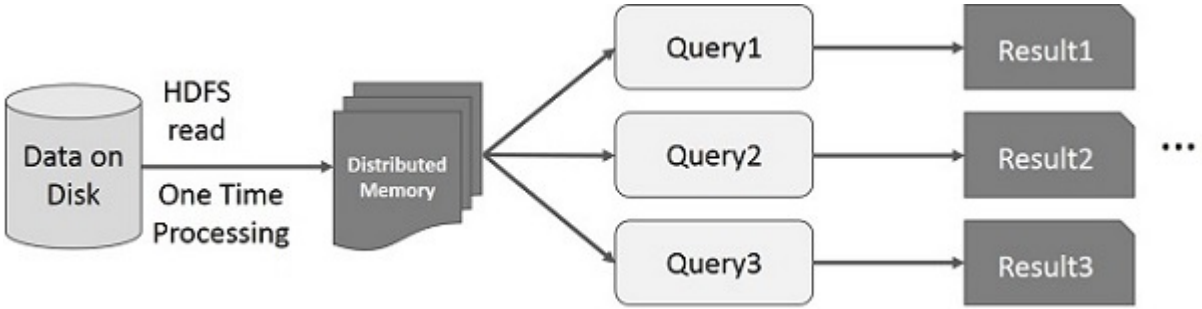
The illustration given below shows the iterative operations on Spark RDD. It will store intermediate results in a distributed memory instead of Stable storage (Disk) and make the system faster.

Note – If the Distributed memory (RAM) in sufficient to store intermediate results (State of the JOB), then it will store those results on the disk



Interactive Operations on Spark RDD

This illustration shows interactive operations on Spark RDD. If different queries are run on the same set of data repeatedly, this particular data can be kept in memory for better execution times.



By default, each transformed RDD may be recomputed each time you run an action on it. However, you may also **persist** an RDD in memory, in which case Spark will keep the elements around on the cluster for much faster access, the next time you query it. There is also support for persisting RDDs on disk, or replicated across multiple nodes.

Spark - Installation

Spark is Hadoop's sub-project. Therefore, it is better to install Spark into a Linux based system. The following steps show how to install Apache Spark.

Step1: Verifying Java Installation

Java installation is one of the mandatory things in installing Spark. Try the following command to verify the JAVA version.

```
$java -version
```

If Java is already, installed on your system, you get to see the following response –

```
java version "1.7.0_71"
Java(TM) SE Runtime Environment (build 1.7.0_71-b13)
Java HotSpot(TM) Client VM (build 25.0-b02, mixed mode)
```

In case you do not have Java installed on your system, then Install Java before proceeding to next step.

Step2: Verifying Scala Installation

You should Scala language to implement Spark. So let us verify Scala installation using following command.

```
$scala -version
```

If Scala is already installed on your system, you get to see the following response –

```
Scala code runner version 2.11.6 -- Copyright 2002-2013, LAMP/EPFL
```

In case you don’t have Scala installed on your system, then proceed to next step for Scala installation.

Step3: Downloading Scala

Download the latest version of Scala by visit the following link Download Scala . For this tutorial, we are using scala-2.11.6 version. After downloading, you will find the Scala tar file in the download folder.

Step4: Installing Scala

Follow the below given steps for installing Scala.

Extract the Scala tar file

Type the following command for extracting the Scala tar file.

```
$ tar xvf scala-2.11.6.tgz
```

Move Scala software files

Use the following commands for moving the Scala software files, to respective directory **(/usr/local/scala)**.

```
$ su –
Password:
# cd /home/Hadoop/Downloads/
# mv scala-2.11.6 /usr/local/scala
# exit
```

Set PATH for Scala

Use the following command for setting PATH for Scala.

```
$ export PATH = $PATH:/usr/local/scala/bin
```

Verifying Scala Installation

After installation, it is better to verify it. Use the following command for verifying Scala installation.

```
$scala -version
```

If Scala is already installed on your system, you get to see the following response –

```
Scala code runner version 2.11.6 -- Copyright 2002-2013, LAMP/EPFL
```

Step5: Downloading Apache Spark

Download the latest version of Spark by visiting the following link Download Spark . For this tutorial, we are using **spark-1.3.1-bin-hadoop2.6** version. After downloading it, you will find the Spark tar file in the download folder.

Step6: Installing Spark

Follow the steps given below for installing Spark.

Extracting Spark tar

The following command for extracting the spark tar file.

```
$ tar xvf spark-1.3.1-bin-hadoop2.6.tgz
```

Moving Spark software files

The following commands for moving the Spark software files to respective directory **(/usr/local/spark)**.

```
$ su –
Password:
# cd /home/Hadoop/Downloads/
# mv spark-1.3.1-bin-hadoop2.6 /usr/local/spark
# exit
```

Setting up the environment for Spark

Add the following line to **~/.bashrc** file. It means adding the location, where the spark software file are located to the PATH variable.

```
export PATH = $PATH:/usr/local/spark/bin
```

Use the following command for sourcing the ~/.bashrc file.

```
$ source ~/.bashrc
```

Step7: Verifying the Spark Installation

Write the following command for opening Spark shell.

```
$spark-shell
```

If spark is installed successfully then you will find the following output.

```
Spark assembly has been built with Hive, including Datanucleus jars on classpath
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
15/06/04 15:25:22 INFO SecurityManager: Changing view acls to: hadoop
15/06/04 15:25:22 INFO SecurityManager: Changing modify acls to: hadoop
disabled; ui acls disabled; users with view permissions: Set(hadoop); users with modify permissions: Set(hadoop)
15/06/04 15:25:22 INFO HttpServer: Starting HTTP Server
15/06/04 15:25:23 INFO Utils: Successfully started service 'HTTP class server' on port 43292.
Welcome to
```

```
   ____              __
  / __/__  ___ _____/ /__
 _\ \/ _ \/ _ `/ __/  '_/
/___/ .__/\_,_/_/ /_/\_\ version 1.4.0
   /_/

Using Scala version 2.10.4 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_71)
Type in expressions to have them evaluated.
Spark context available as sc
scala>
```

Spark SQL - Introduction

Spark introduces a programming module for structured data processing called Spark SQL. It provides a programming abstraction called DataFrame and can act as distributed SQL query engine.

Features of Spark SQL

The following are the features of Spark SQL –

Integrated – Seamlessly mix SQL queries with Spark programs. Spark SQL lets you query structured data as a distributed dataset (RDD) in Spark, with integrated APIs in Python, Scala and Java. This tight integration makes it easy to run SQL queries alongside complex analytic algorithms.

Unified Data Access – Load and query data from a variety of sources. Schema-RDDs provide a single interface for efficiently working with structured data, including Apache Hive tables, parquet files and JSON files.

Hive Compatibility – Run unmodified Hive queries on existing warehouses. Spark SQL reuses the Hive frontend and MetaStore, giving you full compatibility with existing Hive data, queries, and UDFs. Simply install it alongside Hive.

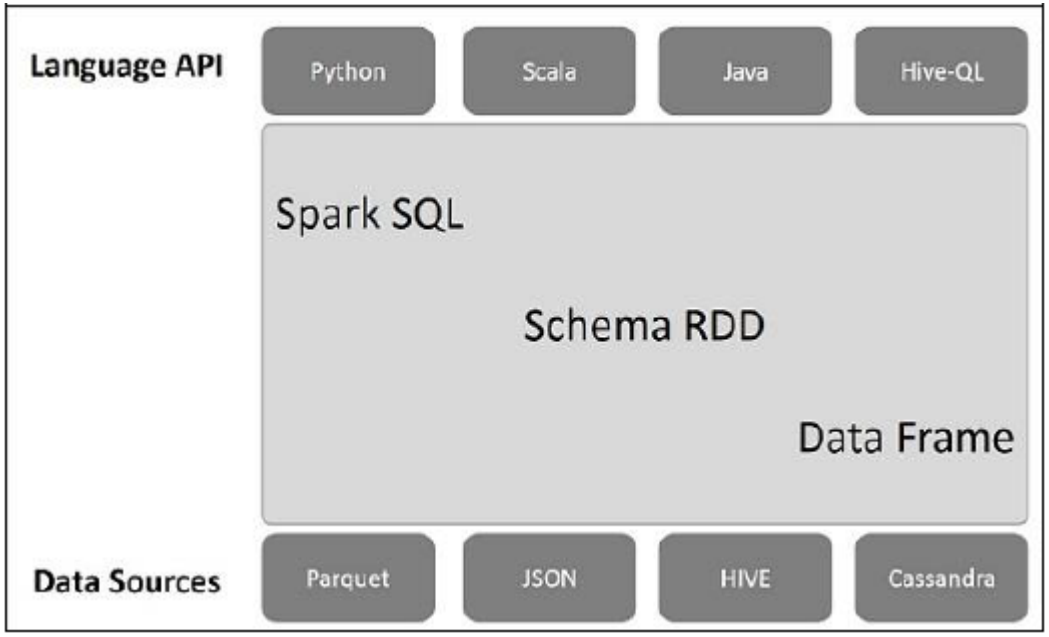
Standard Connectivity – Connect through JDBC or ODBC. Spark SQL includes a server mode with industry standard JDBC and ODBC connectivity.

Scalability – Use the same engine for both interactive and long queries. Spark SQL takes advantage of the RDD model to support mid-query fault tolerance, letting it scale to large jobs too. Do not worry about using a different

engine for historical data.

Spark SQL Architecture

The following illustration explains the architecture of Spark SQL –



This architecture contains three layers namely, Language API, Schema RDD, and Data Sources.

Language API – Spark is compatible with different languages and Spark SQL. It is also, supported by these languages- API (python, scala, java, HiveQL).

Schema RDD – Spark Core is designed with special data structure called RDD. Generally, Spark SQL works on schemas, tables, and records. Therefore, we can use the Schema RDD as temporary table. We can call this Schema RDD as Data Frame.

Data Sources – Usually the Data source for spark-core is a text file, Avro file, etc. However, the Data Sources for Spark SQL is different. Those are Parquet file, JSON document, HIVE tables, and Cassandra database.

We will discuss more about these in the subsequent chapters.

Spark SQL - DataFrames

A DataFrame is a distributed collection of data, which is organized into named columns. Conceptually, it is equivalent to relational tables with good optimization techniques.

A DataFrame can be constructed from an array of different sources such as Hive tables, Structured Data files, external databases, or existing RDDs. This API was designed for modern Big Data and data science applications taking inspiration from **DataFrame in R Programming** and **Pandas in Python**.

Features of DataFrame

Here is a set of few characteristic features of DataFrame –

- Ability to process the data in the size of Kilobytes to Petabytes on a single node cluster to large cluster.
- Supports different data formats (Avro, csv, elastic search, and Cassandra) and storage systems (HDFS, HIVE tables, mysql, etc).
- State of art optimization and code generation through the Spark SQL Catalyst optimizer (tree transformation framework).
- Can be easily integrated with all Big Data tools and frameworks via Spark-Core.
- Provides API for Python, Java, Scala, and R Programming.

SQLContext

SQLContext is a class and is used for initializing the functionalities of Spark SQL. SparkContext class object (sc) is required for initializing SQLContext class object.

The following command is used for initializing the SparkContext through spark-shell.

```
$ spark-shell
```

By default, the SparkContext object is initialized with the name **sc** when the spark-shell starts.

Use the following command to create SQLContext.

```
scala> val sqlcontext = new org.apache.spark.sql.SQLContext(sc)
```

Example

Let us consider an example of employee records in a JSON file named **employee.json**. Use the following commands to create a DataFrame (df) and read a JSON document named **employee.json** with the following content.

employee.json – Place this file in the directory where the current **scala>** pointer is located.

```
{
   {"id" : "1201", "name" : "satish", "age" : "25"}
   {"id" : "1202", "name" : "krishna", "age" : "28"}
   {"id" : "1203", "name" : "amith", "age" : "39"}
   {"id" : "1204", "name" : "javed", "age" : "23"}
   {"id" : "1205", "name" : "prudvi", "age" : "23"}
}
```

DataFrame Operations

DataFrame provides a domain-specific language for structured data manipulation. Here, we include some basic examples of structured data processing using DataFrames.

Follow the steps given below to perform DataFrame operations –

Read the JSON Document

First, we have to read the JSON document. Based on this, generate a DataFrame named (dfs).

Use the following command to read the JSON document named **employee.json**. The data is shown as a table with the fields – id, name, and age.

```
scala> val dfs = sqlContext.read.json("employee.json")
```

Output – The field names are taken automatically from **employee.json**.

```
dfs: org.apache.spark.sql.DataFrame = [age: string, id: string, name: string]
```

Show the Data

If you want to see the data in the DataFrame, then use the following command.

```
scala> dfs.show()
```

Output – You can see the employee data in a tabular format.

```
<console>:22, took 0.052610 s
+----+------+--------+
|age | id   |  name  |
+----+------+--------+
| 25 | 1201 | satish |
| 28 | 1202 | krishna|
| 39 | 1203 | amith  |
| 23 | 1204 | javed  |
| 23 | 1205 | prudvi |
+----+------+--------+
```

Use printSchema Method

If you want to see the Structure (Schema) of the DataFrame, then use the following command.

```
scala> dfs.printSchema()
```

Output

```
root
   |-- age: string (nullable = true)
   |-- id: string (nullable = true)
   |-- name: string (nullable = true)
```

Use Select Method

Use the following command to fetch **name**-column among three columns from the DataFrame.

```
scala> dfs.select("name").show()
```

Output – You can see the values of the **name** column.

```
<console>:22, took 0.044023 s
+--------+
|  name  |
```

```
+--------+
| satish |
| krishna|
| amith  |
| javed  |
| prudvi |
+--------+
```

Use Age Filter

Use the following command for finding the employees whose age is greater than 23 (age > 23).

```
scala> dfs.filter(dfs("age") > 23).show()
```

Output

```
<console>:22, took 0.078670 s
+----+------+--------+
|age | id   | name   |
+----+------+--------+
| 25 | 1201 | satish |
| 28 | 1202 | krishna|
| 39 | 1203 | amith  |
+----+------+--------+
```

Use groupBy Method

Use the following command for counting the number of employees who are of the same age.

```
scala> dfs.groupBy("age").count().show()
```

Output – two employees are having age 23.

```
<console>:22, took 5.196091 s
+----+-----+
|age |count|
+----+-----+
```

```
| 23 |  2  |
| 25 |  1  |
| 28 |  1  |
| 39 |  1  |
+----+-----+
```

Running SQL Queries Programmatically

An SQLContext enables applications to run SQL queries programmatically while running SQL functions and returns the result as a DataFrame.

Generally, in the background, SparkSQL supports two different methods for converting existing RDDs into DataFrames –

Sr. No	Methods & Description
1	**Inferring the Schema using Reflection** This method uses reflection to generate the schema of an RDD that contains specific types of objects.
2	**Programmatically Specifying the Schema** The second method for creating DataFrame is through programmatic interface that allows you to construct a schema and then apply it to an existing RDD.

Spark SQL - Data Sources

A DataFrame interface allows different DataSources to work on Spark SQL. It is a temporary table and can be operated as a normal RDD. Registering a DataFrame as a table allows you to run SQL queries over its data.

In this chapter, we will describe the general methods for loading and saving data using different Spark DataSources. Thereafter, we will discuss in detail the specific options that are available for the built-in data sources.

There are different types of data sources available in SparkSQL, some of which are listed below –

Sr. No	Data Sources
1	**JSON Datasets**

	Spark SQL can automatically capture the schema of a JSON dataset and load it as a DataFrame.
2	**Hive Tables** Hive comes bundled with the Spark library as HiveContext, which inherits from SQLContext.
3	**Parquet Files** Parquet is a columnar format, supported by many data processing systems.

Previous Page Next Page