



SparkSQL

tutorialspoint
SIMPLY EASY LEARNING

www.tutorialspoint.com



<https://www.facebook.com/tutorialspointindia>



<https://twitter.com/tutorialspoint>

About the Tutorial

Apache Spark is a lightning-fast cluster computing designed for fast computation. It was built on top of Hadoop MapReduce and it extends the MapReduce model to efficiently use more types of computations which includes Interactive Queries and Stream Processing.

This is a brief tutorial that explains the basics of Spark SQL programming.

Audience

This tutorial has been prepared for professionals aspiring to learn the basics of Big Data Analytics using Spark Framework and become a Spark Developer. In addition, it would be useful for Analytics Professionals and ETL developers as well.

Prerequisite

Before you start proceeding with this tutorial, we assume that you have prior exposure to Scala programming, database concepts, and any of the Linux operating system flavors.

Copyright & Disclaimer

© Copyright 2015 by Tutorials Point (I) Pvt. Ltd.

All the content and graphics published in this e-book are the property of Tutorials Point (I) Pvt. Ltd. The user of this e-book is prohibited to reuse, retain, copy, distribute or republish any contents or a part of contents of this e-book in any manner without written consent of the publisher.

We strive to update the contents of our website and tutorials as timely and as precisely as possible, however, the contents may contain inaccuracies or errors. Tutorials Point (I) Pvt. Ltd. provides no guarantee regarding the accuracy, timeliness or completeness of our website or its contents including this tutorial. If you discover any errors on our website or in this tutorial, please notify us at contact@tutorialspoint.com

Table of Contents

About the Tutorial	i
Audience.....	i
Prerequisite	i
Copyright & Disclaimer	i
Table of Contents.....	ii
1. SPARK SQL – INTRODUCTION	1
Apache Spark.....	1
Evolution of Apache Spark	1
Features of Apache Spark	1
Spark Built on Hadoop	2
Components of Spark	3
2. SPARK SQL – RDD	4
Resilient Distributed Datasets.....	4
Data Sharing is Slow in MapReduce	4
Iterative Operations on MapReduce	4
Interactive Operations on MapReduce	5
Data Sharing using Spark RDD	6
Iterative Operations on Spark RDD	6
Interactive Operations on Spark RDD	6
3. SPARK SQL – INSTALLATION	8
Step 1: Verifying Java Installation	8
Step 2: Verifying Scala installation	8
Step 3: Downloading Scala	8
Step 4: Installing Scala	9
Step 5: Downloading Apache Spark	9

Step 6: Installing Spark	10
Step 7: Verifying the Spark Installation	10
4. SPARK SQL – FEATURES AND ARCHITECTURE	12
Features of Spark SQL	12
Spark SQL Architecture	13
5. SPARK SQL – DATAFRAMES	14
Features of DataFrame	14
SQLContext	14
DataFrame Operations	15
Running SQL Queries Programmatically	17
Inferring the Schema using Reflection	18
Programmatically Specifying the Schema	21
6. SPARK SQL – DATA SOURCES	25
JSON Datasets	25
DataFrame Operations	26
Hive Tables	27
Parquet Files	29

1. SPARK SQL – 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

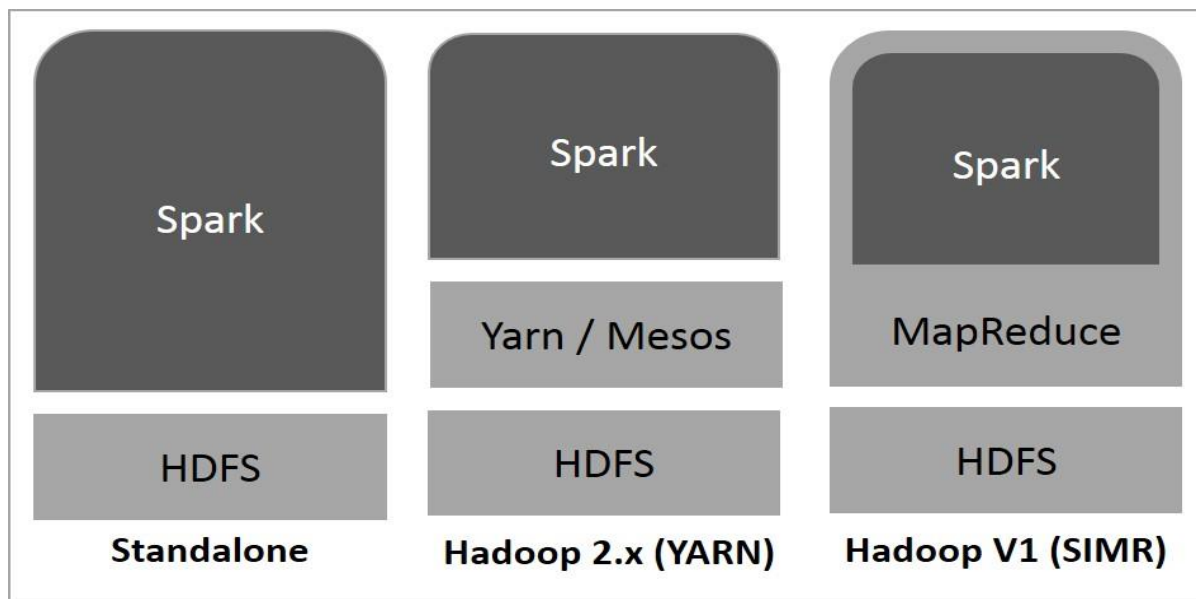
Apache Spark has following features.

- **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.

End of ebook preview

If you liked what you saw...

Buy it from our store @ <https://store.tutorialspoint.com>