Introduction to Big Data Computation Using Apache Spark

What will you learn?

Discover Apache Spark - the most advanced and fastest Big Data Computation Technology through hands-on experience and real life use cases. We will also be covering the basic concepts of Big Data and an overview of Hadop ecosystem.

Duration: 1 day

Course Objective: **Course Objective**

Have you ever wondered how airlines determine the air fare – even while you are trying to book tickets? Imagine the large amount of data that is processed within seconds to get you routes, prices and availability! That's the power of Big Data analytics and one of the most popularly used technology used is Apache Spark. Join us for the course and learn how to build Apache Spark applications to process large amount of data in real-time at a very high speed and efficiency.

Topics to be covered:

1. Introduction to the Big Data world
2. Big Data key concepts
3. Overview of Hadoop ecosystem
4. Different phases of Big Data analytics
5. Basic concepts of Apache Spark
6. Hands-on experience with Spark-Shell with a variety of datasets
7. Apache Spark streaming: hands-on real time analytics
8. Data analysis using Spark-SQL
9. Apache Spark specialized libraries for machine learning and graph analytics
10. Common Apache Spark use cases with examples

Pre-requisites:

All the hands-on exercises will be done using Cloudera VM. Participants must have Cloudera Quick-start VM set up in their machine to enjoy the workshop. Please note that min 4GB RAM should be available for Virtual Machine.

Cloudera Quick Start VM is available on following link.

http://www.cloudera.com/downloads/quickstart_vms/5-8.html

Please select the platform VMware or Virtual Box as per machine. Refer to following page before downloading Cloudera VM and VMware Player or Virtual Box:

http://www.cloudera.com/documentation/enterprise/5-3-x/topics/cloudera_quickstart_vm.html

Cloudera VM Zip file is around 4GB in size. All the prerequisites will be available half an hour before the workshop but it is highly recommended to have the set up ready before coming to workshop.

For Set up for Windows Machine, Please refer to following tutorial

https://www.youtube.com/watch?v=oNQ8f2My5Hs

For Set up for Ubuntu or MAC, Please refer to following tutorial.

https://www.youtube.com/watch?v=BeCtjd86YXo&t=53s

For common issues like BIOS setting, Please refer the page below.

https://community.cloudera.com/t5/Hadoop-101-Training-Quickstart/How-to-setup-Cloudera-Quickstart-Virtual-Machine/ta-p/35056

To understand Spark SQL and Dataframes better, it is advised to have installed MySQL on the machine.