As part of this topic, we will see details required to get started with Apache Spark – Core API

* A brief introduction to Spark
* Accessing the Documentation
* Connecting to the Environment
* Initializing Spark Jobs

## Agenda

* Objectives
* Problem Statement
* Introduction to Spark
* Initializing the job
* Create RDD using data from HDFS
* Read data from different file formats
* Standard Transformations
* Saving RDD back to HDFS
* Save data in different file formats
* Solution

## Objectives

* Convert a set of data values in a given format stored in HDFS into  
  new data values or a new data format and write them into HDFS.
  + Load RDD data from HDFS for use in Spark applications
  + Write the results from an RDD back into HDFS using Spark
  + Read and write files in a variety of file formats
  + Perform standard extract, transform, load (ETL) processes on data

## Problem Statement

* Use retail\_db data set
* Problem Statement
  + Get daily revenue by product considering completed and closed orders.
  + Data need to be sorted in ascending order by date and then descending  
    order by revenue computed for each product for each day.
* Data for orders and order\_items is available in HDFS  
  /public/retail\_db/orders and /public/retail\_db/order\_items
* Data for products is available locally under /data/retail\_db/products
* Final output need to be stored under
  + HDFS location – avro format  
    /user/YOUR\_USER\_ID/daily\_revenue\_avro\_python
  + HDFS location – text format  
    /user/YOUR\_USER\_ID/daily\_revenue\_txt\_python
  + Local location /home/YOUR\_USER\_ID/daily\_revenue\_python
  + Solution need to be stored under  
    /home/YOUR\_USER\_ID/daily\_revenue\_python.txt