Module 4: Deep Dive into Apache Spark Framework

Case Study I

edureka!

Case Study: Financial Regulation
Domain: Banking

All financial institutions have been introduced to a new financial regulation that will consider the impact of possible future events when calculating their risk exposure. For this change, all financial trading services will have to source data from far more data sources including both trade and risk data, which means large amounts of data that have not historically been required for accounting.

The data is of trading index listings for each trading day.

Tasks:

To adhere to the regulation a leading trading service provider has decided to create a new repository (Big Data) and a multi-facet platform that can cater to these models. As part of the R&D team, you are required to execute a POC exercise to maintain a master data model.

1. Test the spark environment by executing the spark's HdfsTest.scala example.
2. Try to implement the same example in pyspark and perform spark-submit.
3. Analyze the behavior of spark application on Spark web UI
4. Edit the application and add custom logs. Once executed check the Spark logs.
5. Transfer the sample dataset from RDBMS to HDFS
6. Validate the loaded data by comparing the statistics of data both in source and HDFS
7. Create a new directory EQ in HDFS and transfer the data where series is EQ
8. Set total trades which are less than 500 to 0 and transfer only updated rows.