8. Capstone Project

i. Develop a data science project showcasing data processing, analysis, and machine learning using Spark

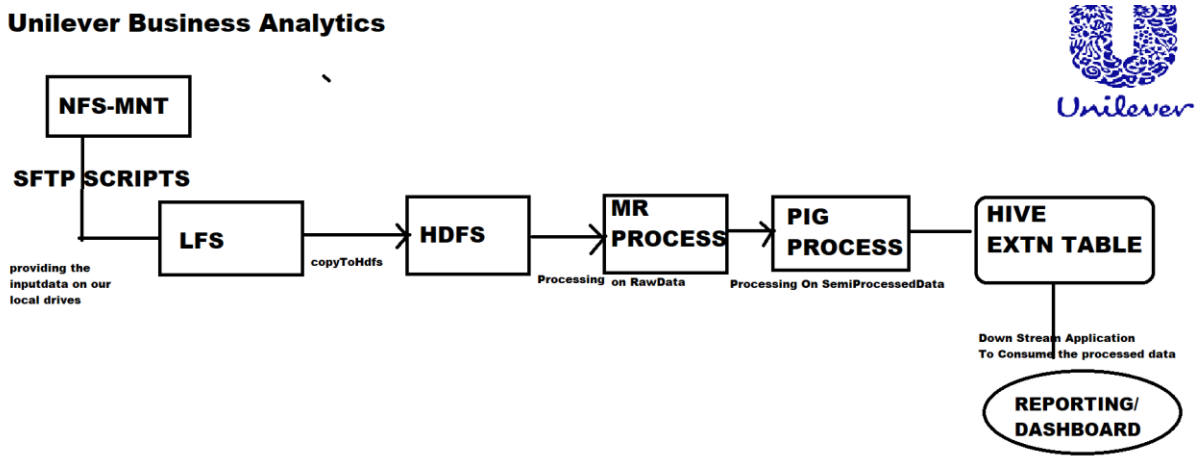
Unilever - Business Analytics

Project Name: Unilever – Business Analytics

Client: Unilever, USA (www.unilever.com)

Environment: hadoop, hdfs, map reduce, pig,hive,sql server,spark,scala,cdh

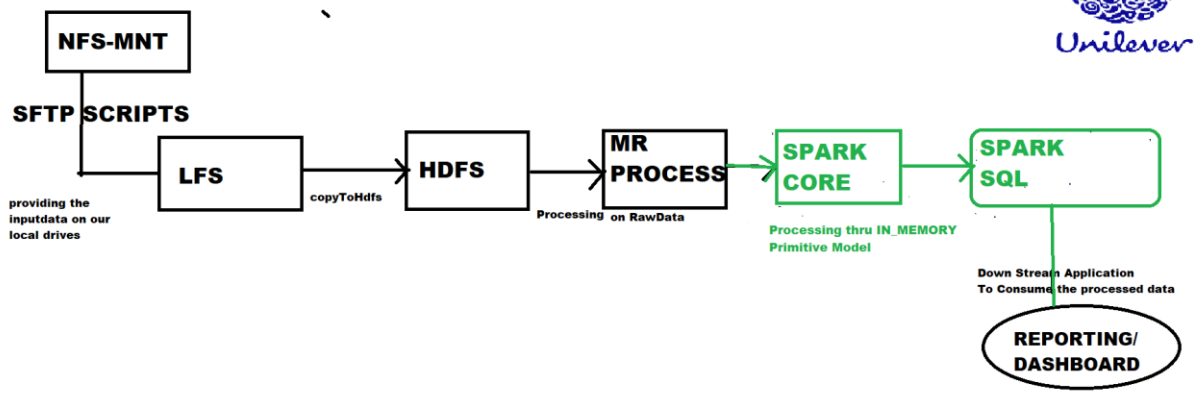
Description: Unilever is getting the source data from different source systems. As part of the same business each customer would be offered with different types of products based on their needs. Customer might have retail type of products, baby products, home electronics etc. To maintain this much of huge volumes of, different verities of data in traditional databases is a very tedious process. To meet the scaling needs of data of Unilever, re-plat forming of current data warehouse system to hadoop solution in a cost effective solution.



IN NEWER SPRINTS:

In Case of Map Reduce processing working with huge volume of data is a performance bottle neck with respect to multiple disk reads/writes...To overcome the same performance related challenges in place of PIG & HIVE, we are going to replace with SPARK processing where we will get performance optimisation with respect to IN_MEMORY Cluster Primitive model

Unilever Business Analytics



Responsibilities:

- Expertise in designing and deployment of Hadoop cluster and different Big Data analytic tools including Pig, Hive, HBase, Sqoop, flume, Apache Spark, with Cloudera Distribution.
- Involved in loading and transforming large sets of structured, semi-structured and unstructured data and analyzed them by running Hive queries and Pig scripts.
- Used Sqoop to dump data from relational database into HDFS.
- Real time streaming of data using Spark with Kafka.
- Responsible for developing data pipeline using flume, Sqoop and pig to extract the data from weblogs and Store in HDFS.
- Experienced with batch processing of data sources using Apache Spark.
- Experienced in implementing Spark RDD transformations, actions to implement business analysis.
- Implemented Spark using Scala and Spark SQL for faster testing and processing of data.
- Implemented partitioning, dynamic partitions and buckets in HIVE.
- Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Scala.
- Created Hive tables and involved in data loading and writing Hive UDFs.