Big Data Integration and Analysis

- Lab Mentor: Mohamed Nadjib Mami

Team Members:

- Gaurav Kumar
- Hectór Ugarte
- Miguel Mármol
- Tina Boroukhian

Requirement Introduction

- We have a scenario where we have big sizes of data and we are expected to clean the data and provide desired result after analyzing data. Expected results include analytic queries like aggregation eg. Finding max,min,count from a given set of data.

Functional Requirements

R01. Get raw sample data in CSV files.

R02. Move data to HDFS.

R03. Clean data (like date and number formatting as per standards).

R04. Reduce columns to keep only needed columns.

R05. Analyze data using analytical queries like aggregation eg. Finding max, min, and count.

R06. Get results from HDFS to local system.

Non-Functional Requirements

- Scalability on multiple machines (in our case we are targeting for 4 machines).
- High Performance - short response time even with high volumes of data.
- Fault tolerance.
- Work in distributed environment.