Case Study : Mobile App Store Domain: Telecom

A telecom software provider is building an application to monitor different telecom components in the production environment. For monitoring purpose, the application relies on log files by parsing the log files and looking for potential warning or exceptions in the logs and reporting them.

The Dataset contains the log files from different components used in the overall telecom application.

Tasks:

The volume of data is quite large. As part of the R&D team, you are building a solution on spark to load and parse the multiple log files and then arranging the error and warning by the timestamp.

1. **Load data into Spark DataFrame**
2. **Find out how many 404 HTTP codes are in access logs.**
3. **Find out which URLs are broken.**
4. **Verify there are no null columns in the original dataset.**
5. **Replace null values with constants such as 0**
6. **Parse timestamp to readable date.**
7. **Describe which HTTP status values appear in data and how many.**
8. **How many unique hosts are there in the entire log and their average request**
9. **Create a spark-submit application for the same and print the findings in the log**

Solution:

Jupyter Notebook with all explanation in comments and output has been attached with this submission.

Important Learnings from this case study:

1. Loading of the log file in DataFrame not as a CSV but as a text file
2. Parsing each line of the DataFrame by generating the REGEX pattern to extract individual fields and then using regex_extractor of Pyspark to extract the field from each text line of the DataFrame in order to form the tabular/relational type DataFrame to perform SQL queries over it.(very important learning)
3. Creating UDF functions
4. Performing interesting , real life queries over the Data.