Approach - JOB-A-THON June 2021

Created	@July 4, 2021
Created by	 Soundar Murugan
Tags	Work

Introduction

The data consists of two datasets namely the userTable and the visitorLogsData.

The objective was to create an input dataframe that will be used to train a machine learning model.

Hence, this was an ETL hackathon requiring the candidates to use the best coding practices to come up with an optimized and modular code to perform the required function.

My approach

I decided to go ahead with Apache Spark since such frameworks are widely in use in the current industry.

With the help of Spark, I could easily work on the data with multiple executors computing the results faster, and also the support of Big Data and SQL for other applications.

Pandas is a well-known alternative but I prefer Spark due to parallelization.

The steps I took are as follows:

1. Examine the userTable dataset
2. Convert the data types as required (Signup Date in this case) and save it as table1
3. Examine the visitorLogsData dataset

4. Convert both the DateTime formats (Timestamp and Unix Timestamp) to Timestamp format and save it as table2
5. Calculate User_Vintage as the difference between the current date and the user signup date
6. Filter the last 21 days from table2
7. Clean the table by formatting the records to follow a universal format
8. Use SQL queries to obtain the required tables
9. Join all these tables to form the final dataframe
10. Impute null values accordingly
11. Sort the columns and rows of this dataframe to fit the required format
12. Save the dataframe as solution.csv in the present working directory

---

Result

I obtained a result of 0.916 with my approach.