**Info 7390**

**Case Study 1**

**Problem 2**

**Sicheng Zhang**

**Mengqi Zhao**

Part2 report

**1. Introduction**

The purpose of this project is analyzing the EDGAR Log File Data Set. By a given year, program should get data for the first day of the month and programmatically generate the url. According to the guideline, we suppose to handle missing data, compute summary metrics, check anomalies and log operations.

**2. Basic flow**

**3. Analyzing Strategy**

Before we do the analyzing, we need our data prepared. So we replace and refill missing data and handle the amount and location to the log file at the beginning of our analyze methods And then replace NaN by .sample() or according to the ratio of exist data to pass value to missing data to keep lower the influence of the missing data. Because of the differences between the data, we choose different strategy for each data as followed.

1. IP:

IP address is a numerical label assigned to each device that connect to computer network. It’s format is xxx.xxx.xxx.\*\*\*. So we check each IP address with this format to see if there are illegal accessing.

1. Date:

Data in the same loop of month should have same date. So for this column, we just check if it match.

1. Code:

Code represent status code for the request. If we have code like 304, 404, 500, it means request meet error. Then we can treat this data as an abnormal data and handle it to log file

1. Time

Upload outputs

Combine summary and data

Analyze each data month by month

Download data by generated URL

4) Time:

We separate one day to 6 period to find highest access volume period in a day. By this, we can clearly have an impression of users’ preference. We can even know what is our main user group, which benefit further optimization.

5) CIK:

We can easily find the most interested company by using method “value\_counts” to calculate the frequency of each CIK.

6) File size:

One file should not be too larger or too small. Therefore, if the size is not in the given range that we set, an abnormal warning will be handle to log file.

7) Zone & Doc & Browser

We simply using Counter and df.most\_common(1) to extract the most frequent value

8) noagent & norefer & idx &crawler

These kinds of data only have 2 values at most: 0 and 1. So, when we loop the data, if we find data is not (1 & 0), an abnormal warning will be passed to log file. Then the most frequent value extract for summary.

After analyzation, we store the summary of every column in an empty data frame. While finish one month analyzation, we merge the summary. When we finish all 12 moth, we concatenate 12 csv and summary to 1 and pick browser and crawler as the example of drawing ratio graphic.

**4. Upload output**

We wrap our summary and output and upload it to AWS S3 by using package boto3