Advances in Data Sciences and Architecture

Assignment 1 – Part 2 : Missing Data Values

**Summary Report:**

We are basically accessing the website and scraping the data that we need using Beautiful Soup library. We first create a main list which has all the year links from the website.

When we take the input from the user, we basically travcerse through that list to get the particular year link.

**We then create two folders:**

First folder (the name of the folder is basically the year entered by the user) has first day files of every month in csv format (extracted files)

Second folder has (the name of the folder is the year entered by the user plus the timestamp) has all the CSVs which were created afrter the dataframes wwre cleaned and converted into CSVs.

**Now we are basically handling 2 anomalies:**

1. Missing values (No content)

* Checking number of values in each column and verifying all the columns have same number of values.
* Incase if the count does not match, it means that there are missing values.
* These missing values are then replaced with 0.

1. Does the column has sufficient valid values for analysis

* The key sheet has valid value for most of the columns.
* So we are iterating over each column and finding the invalid values.
* The invalid value indexes are then stored in a list and then we iterate through each of the indexes to replace the invalid value with 0.
* The reason that we are replacing it with zero is the number of non-zero values in a column corresponds to number of valid values.

**Summary Metrics:**

(As an example, we have considered the analysis for the year 2016 for our summary metrics result)

1. TOP 5 IPs on each day

The percentage of requests from the top IP

From our analysis, 2016-01-01 has

|  |  |
| --- | --- |
| ip | Number of requests |
| 54.175.49.ggf | 804147 |
| 68.180.229.abc | 589830 |
| 60.242.198.hbf | 533480 |
| 173.236.43.bbb | 419923 |
| 208.82.235.djf | 419131 |

percentage of requests from top 1 IP :54.175.49.ggf : 9.778663209252136

2. Request Response Code Stats:

The percentage of requests with 200 OK

The percentage of requests with 404 Not Found

|  |  |
| --- | --- |
| 200 | 7433434 |
| 301 | 427912 |
| 503 | 84968 |
| 304 | 80299 |
| 404 | 62430 |
| 403 | 37733 |
| 429 | 37388 |

The total number of 200 : 7433434

Percentage : 90.3927361218831%

The total number of 404 : 62430

Percentage : 0.7591670977490568%

3.Document Size Summary

Total data size accessed on the date : 2016-01-01

852713676021.0 MB

successful data size accessed on the date : 2016-01-01

822930860069.0 MB

percentage success96.50728998613288%

Average size of the dat being accessed : 103692 MB

4. Busy Hour

The summary of number of requests at all hours

[1: 394492, 2: 503588, 3: 344974, 4: 363013, 5: 310947, 6: 323605, 7: 350739, 8: 348616, 9: 358506, 10: 338545, 11: 417421, 12: 434628, 13: 381645, 14: 391679, 15: 375512, 16: 296722, 17: 286992, 18: 274203, 19: 284200, 20: 288066, 21: 284601, 22: 284256, 23: 293441, 24: 293095]

The busiest hour : 2 and Number of requests : 503588

percentage of requests : 6.1237776777391%

Average number of requests per hour : 342645

Average percentage : 4%

5. Extention Summary

TOP 10 extentions

|  |  |
| --- | --- |
| extention | cik |
| -index.htm | 2880953 |
| .txt | 2241964 |
| .hdr.sgml | 342373 |
| -index.html | 321964 |
| edgar.xml | 152197 |
| doc4.xml | 150526 |
| primary\_doc.xml | 125548 |
| -xbrl.zip | 98833 |
| xslFormDX01/primary\_doc.xml | 58526 |

Number of requests with the extention asked frequesntly : -index.htm

Percent :

35.0332328649918

Data Cleaning :

1.Reading the data in to Dataframes

2.Finding all Null values and Replacing them with 0

3.Column Validation and anamoly detection

4.Data Correction on each Column

5.Final Dataframe after the clean

6.Writing all the cleaned dataframe as csv

Logging:

1. Created a logging pipeline

2. Anomalies are logged as error

3. Log levels configuration

Docker Image:

