|  |
| --- |
|  |
| Problem 1: Data wrangling Edgar data from text files |
|  |
| Praising Data |

**Northeastern University**

October 6, 2017

Authored by: Madhumathi Prakash and Sonali Chaudhari

Problem 1: Data wrangling Edgar data from text files

Praising Data

# The Data Set

In this problem we are using the EDGAR (Electronic Data Gathering, Analysis, and Retrieval) system. It contains company’s performs automated collection, validation, indexing, acceptance, and forwarding of submissions.

# The Problem

We are accessing data from Edgar site. Given a CIK number and the accession number, we are creating a url to get the data but replacing the X, Y, and Z in the following url: [http://www.sec.gov/Archives/edgar/data/XXX/ZZZ/YYY- index.html](http://www.sec.gov/Archives/edgar/data/XXX/ZZZ/YYY-%20index.html) . From this url we are locating the 10Q file html link and then extracting all the tables. Lastly, we are loading all the tables into a CSV file. When looking for the tables in the 10Q files, we should also deal with logging all activities. For instance, we should deal with error if there is an invalid CIK or accession number or if the amazon keys aren’t valid.

We are using Docker for this problem and building a docker image which will automate the task. Put all zip file on Amazon S3.

# Our Approach

1. We first took a specific use case of the IBM url and worked with it: ([http://www.sec.gov/Archives/edgar/data/51143/000005114313000007/0000051143-13-000007- index.html](http://www.sec.gov/Archives/edgar/data/51143/000005114313000007/0000051143-13-000007-%20index.html)
2. To make the url general: [http://www.sec.gov/Archives/edgar/data/CIK/ZZZ/access number- index.html](http://www.sec.gov/Archives/edgar/data/CIK/ZZZ/access%20number-%20index.html)
   1. We added CIK number and accession number to a text file.
   2. Then we read each line in the text file and assigned it to the CIK and access number accordingly.
   3. To deal with “ZZZ” we realized that it was the access number without the “-“ so we removed it.
   4. We then replaced the

# Steps to Run the code