- *Problem Statement Formation:*

The SEC's EDGAR log file data set is a collection of web server log files that allow researchers to study the demand for SEC filings. This multiple terabyte data set provides researchers with a direct measure of demand for financial reports.

- *Context*

SEC's EDGAR log file data set records internet search traffic for EDGAR filings and compares three methods used in the literature for summarizing this data on a filing-day basis to facilitate analysis. The accounting and finance literature has used the EDGAR logs to study the demand for financial reporting information (e.g., Drake, Roulstone, and Thornock, 2015; Dechow, Lawrence, and Ryans, 2016; Loughran and McDonald, 2016; Bozanic, Hoopes, Thornock, and Williams, 2016; Ryans, 2017). While it is possible for investors to access SEC filings from other sources, such as a firm's investor relations website, it appears that EDGAR captures a significant fraction of financial disclosure demand. Monga and Chasan (2015) quote General Electric CFO Jeffrey Bornstein, who noted that GE's 2013 annual report was downloaded from their investor relations website just 800 times. For the same annual report, the EDGAR logs record 21,987 (4,325) downloads in the year (two months) following its filing. Some firms, such as Google (Alphabet, Inc), forward investors directly to the EDGAR web site to obtain their SEC filings, and as a result, EDGAR may capture an even greater fraction of filing views for some firms. Although the volume of EDGAR downloads for GE's 10-K appeared significant, it represents a relatively small number in comparison to GE's two million individual shareholders (Monga and Chasan, 2015). A caveat to EDGAR download data is that investors have other avenues for obtaining these filings in addition to investor relations websites. For example, EDGARonline, Bloomberg, and FactSet provide access to these reports, and we cannot currently observe download statistics from these sources. The SEC provides access to the EDGAR log files on their web site, with a 1 delay of approximately one year.2 The data is challenging to manipulate into a usable form, as it currently consists of 4,839 daily files containing a record of each EDGAR download. Each daily log file is large: the log file for a single day's downloads on March 31, 2016, contains approximately 14 million records in a 1.6-gigabyte file. Clearly, working with more than a few days' data in-memory is not feasible, so by definition researchers are presented with a big data challenge when analyzing the log files. However, when aggregated to the level of one record per filing-day, the dataset is reduced to a total of approx. 374 million observations from 2003 to 2017, and it becomes accessible to researchers with typical desktop hardware and statistical analysis software. Prior literature studies aspects of demand for financial disclosures in different settings and using a variety of proxies, such as index membership (e.g., Shleifer, 1986; Chen, Noronha, and Singal, 2004), shareholder composition (e.g., Lehavy and Sloan, 2008; Bushee, Core, Guay, and Hamm, 2010), and analyst following (e.g., Irvine, 2003). Demand for financial information is also associated with the concept of investor attention, as investor attention is likely to lead to consumption of firm disclosures. Studies of investor attention consider news media, trading volume, and extreme returns (e.g., Gervais, Kaniel, and Mingelgrin, 2001; Barber and Odean, 2008; Engelberg and Parsons, 2011). DellaVigna and Pollet (2009) study a setting where investor demand is lower due to inattention to Friday earnings announcements. Recently, more direct proxies of investor demand for information include studies of Google search volume for stock ticker symbols (e.g., Da, Engelberg, and Gao, 2011; Drake, Roulstone, and Thornock, 2012), and Yahoo Finance page views (Lawrence, Ryans, and Sun, 2016, 2017). 2https://www.sec.gov/data/edgar-log-file-data-set 2 EDGAR download logs allow for the study of demand for financial disclosures filed with the SEC. This paper compares the EDGAR download count methods used in Ryans (2017, hereinafter Ryans), Loughran and McDonald (2016, hereinafter LM), and Drake et al. (2015, hereinafter DRT). DRT described factors associated with demand for EDGAR filings and found that higher demand is associated with the speed at which earnings news is incorporated into returns. Dechow et al. (2016) show that SEC comment letters are downloaded at low rates compared to the 10-K reports on which they comment and that there is a stronger negative price response when SEC comment letters are disclosed if those comment letters are more heavily downloaded. Lee, Ma, and Wang (2015) create peer groups of firms by identifying those that are searched on EDGAR in close temporal proximity by individual users. LM noted that financial disclosures are in general little used as a source of information. Masden (2017) found that pre-earnings announcement returns are affected by attention to customers' downloads of EDGAR filings. Bozanic et al. (2016) use demand from IRS users to infer attention from tax authorities. Researchers intending to study investor attention need to screen out any programmatic, or "robot", downloads so that the remaining observed page views correspond to human readers.3 The EDGAR log files record the network (IP) address of each user downloading a document. The general robot-screening procedure is to calculate statistics about a user's download patterns over a day, then apply one or more tests to classify the user as a robot or a human. DRT developed the first method used in the literature to filter the EDGAR log files, though LM report that DRT's measure may provide misleading inferences due to mis-specified robot

● *Data source(s)*

Date Range : 2003 to 2017

Commented [1]: @yogitadroy@gmail.com - How about incremental data?

Commented [2R1]: @akhil.toronto.ca@gmail.com Added a section about incremental/new data

https://www.sec.gov/dera/data/edgar-log-file-data-set.html

Incremental Data:

The SEC provides access to the EDGAR log files on their web site, with a 1 delay of approximately one year. The website list log data till 2017(last update is Jan 25, 2018). I have emailed 'StructuredData@sec.gov.' to inquire about the Edgar log data after 2017.

● *Sneak Peak into the Data*

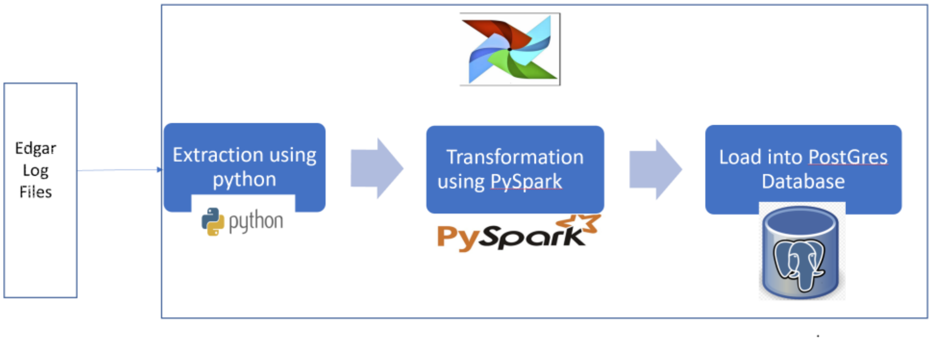
EDGAR Log Data The raw EDGAR log files contain a record for each filing download request.

Since a filing is associated with a unique firm by its CIK number, this is equivalent to firm-filing-day aggregation. The detailed log file record elements are described at https://www.sec.gov/files/EDGAR_variables_FINAL.pdf

Each record in the aggregated file contains some of the following fields:

1. date. YYYY-MM-DD format, Eastern Standard Time zone. 2. cik. The SEC's Central Index Key value associated with a filer (firm identifier). 3. accession. The SEC Accession Number, a document identifier (filing identifier)

● *Proposed architecture for the solution and rationale behind it*



Data will be extracted from daily Edgar Log Files from 2003-2017 using Python libraries, transformed into usable format using Python library of Apache Spark i.e. Pyspark and loaded into Postgres Database. Airflow will be used to automate this ETL data pipeline for historical and future data

The ETL solution will be scaled to Azure Cloud

●*Choice of technology for the solution and rationale*

Python:

Python is known for being the swiss army knife of programming languages. It's especially useful in data science, backend systems, and server-side scripting. That's because Python has strong typing, simple syntax, and an abundance of third-party libraries to use. Pandas, SciPy, Tensorflow, SQLAlchemy, and NumPy are some of the most widely used libraries in production across different industries.

Most importantly, Python decreases development time, which means fewer expenses for companies. For a data engineer, most code execution is database-bound, not CPU-bound. Because of this, it makes sense to capitalize on Python's simplicity, even at the cost of slower performance when compared to compiled languages such as C# and Java.

Spark/Pyspark:

Apache Spark is a unified analytics engine for large-scale data processing. It provides high-level APIs in Java, Scala, Python and R, and an optimized engine that supports general execution graphs. It also supports a rich set of higher-level tools including Spark SQL for SQL and structured data processing, MLlib for machine learning, GraphX for graph processing, and Structured Streaming for incremental computation and stream processing.

Postgres:

A multifunctional DBMS capable of processing complex queries and supporting massive databases

Airflow:

Airflow (https://airflow.apache.org/) is a configuration-as-code OSS solution for workflow automation. It is purely Python-based and there is no XML, YAML, etc. An Airflow workflow is defined as a DAG (Directed Acyclic Graph)coded in Python as a sequence of Tasks. It was originally developed at Airbnb in 2014; a top-level Apache Software Foundation project as of January 2019. It offers developers a way to programmatically author, schedule for execution, and monitor highly configurable complex workflows.

Azure Cloud:

TO DO