China Company Data Mining: Research, Scrapping, Processing, and Analysis for Strategic Insights

Subhashree Mangaraj — Abdulsamet Şahin — Omar El Qarchaoui

March 2, 2023

1 Introduction

Team China is pleased to submit our data competition report for the Global Data Competition by INNOSCRIPTA GMBH. We are excited to present our approach to business. Our objective for this competition is to develop a fast data scraping process and cleaning process, which will help us to present the data in a simplified form that is as understandable as possible.

To achieve this objective, we started with a description of our data sources and the process involved in using automated tools to extract continuously large volumes of data from these sources. We then provide an overview of the data preparation steps taken to clean and transform the data for analysis. We would also explain all the features in our data and their significance. We present our data in the form of an interactive reporting dashboard. We then conclude with a summary of our results and insights, along with our recommendations for further research.

We believe that our report demonstrates the strength of data scraping for company data and the importance of using data cleaning and analysis. We are honoured to have the opportunity to participate in this competition and are confident that our work will make a valuable contribution to the field of data.

2 Data Sourcing

2.1 Objective of the data source search

The objective is to find a single or multiple data sources which can provide us with authentic company information and maximum hits. The type of information we are aiming for is given below:

- Name (incl: Legal form)
- Registry information
- Address (Street, City, Postcode, State/Province, Country)
- Number of Employee
- Website
- E-mail id
- CEO information
  - Name (First and Last Names)
  - Email
  - Phone number
- Yearly Balance Sheet
- Yearly Revenue
- Shareholder structure

- C-Level contact persons
- History information
- Other Information

2.2 Data Source Selection

The first step has been to look for websites, blogs and research articles which are prospects for providing information on companies in China. In the process we came across the following websites:

- www.cec.org.cn
- https://companylist.org/China/Keywords/Registered_Companies/
- https://www.listofcompaniesin.com/china/registered/
- https://www.gongsi.com.cn/

Out of the above list, the website Gongsi.com.cn gave us the best results with the complete coverage of data and the maximum number of hits. Hence, we decided to stick to this particular website for scrapping and the data.

2.3 Website:Gongsi.com.cn

Gongsi.com.cn is a Chinese website that provides information about companies in China. The website offers a comprehensive database of Chinese companies, including their business scope, financial information, ownership structure, and contact information. Gongsi.com.cn provides a platform for companies to showcase their products and services, and for individuals and businesses to search and access information about these companies. Users of Gongsi.com.cn can search for companies using various filters, such as industry, location, and company name. The website also offers additional features, such as company news, industry analysis, and company rankings.

Gongsi.com.cn is a valuable resource for businesses and individuals who are interested in the Chinese market. It provides a wealth of information about companies in China and can be used for market research, competitive analysis, and business development. With its extensive database and user-friendly interface, Gongsi.com.cn is a trusted source of information for anyone looking to do business in China.

2.3.1 Company information:

Gongsi.com.cn provides detailed information about companies in China, including their legal name, registration number, business scope, registered capital, ownership structure, and contact information. This information can be useful for businesses and individuals who are looking to establish partnerships or conduct due diligence on potential partners or competitors.

2.3.2 Financial information:

Gongsi.com.cn also provides financial information about companies, including their revenue, profits, and assets. This information can be helpful for investors and analysts who are interested in evaluating the financial performance of companies in China.

2.3.3 Industry information:

Industry information: Gongsi.com.cn offers industry information and analysis, including industry trends, market size, and key players. This information can be useful for businesses that are looking to enter or expand in specific industries in China.

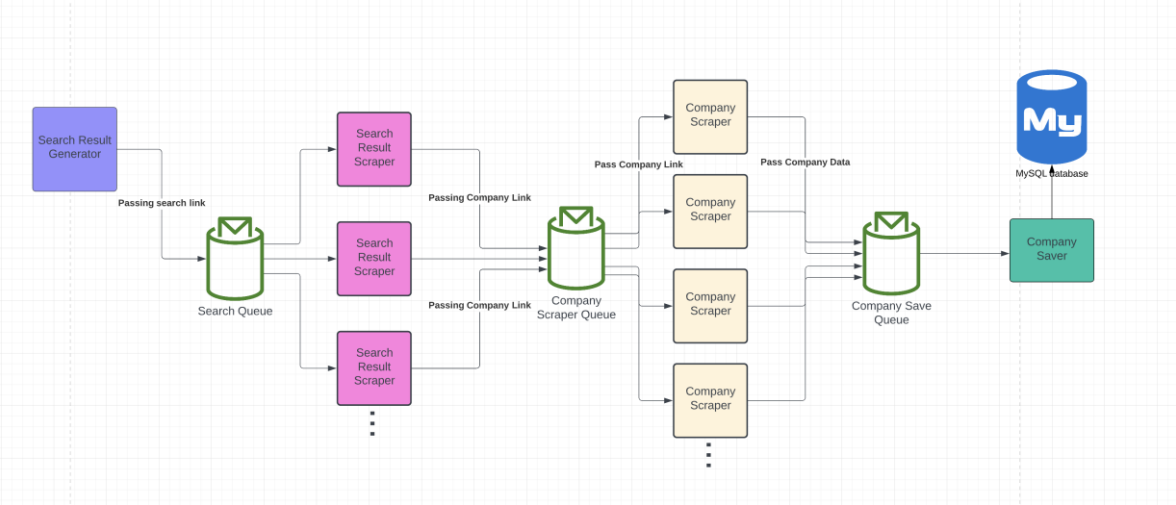

Figure 1: Workflow of scrapping to DB

2.3.4 News and updates:

Gongsi.com.cn provides the latest news and updates on companies and industries in China. This can be helpful for businesses and individuals who want to stay informed about the latest developments and opportunities in the Chinese market.

2.3.5 Rankings and ratings:

Gongsi.com.cn also offer rankings and ratings of companies in China, based on factors such as revenue, profits, and market share. This information can be useful for businesses and investors who are looking to identify leading companies in specific industries.

Gongsi.com.cn is a comprehensive resource for information about companies in China. Its extensive database and user-friendly interface make it a valuable tool for businesses and individuals who are interested in doing business in China.

3 Data Collection

3.1 Scrapping to DB Workflow

The complete workflow of the scrapping of data till the database is given in Fig1.

This scraper utilizes RabbitMQ for communication between the master and workers, enabling it to be highly scalable. The system is comprised of four different types of workers:

- Search result page generator
- Search result page scraper
- Company scraper
- Company saver worker

Search result page generator

There is a pagination limit in place for this system. Non-registered users have a limit of 100 pages, while registered users have a limit of 200 pages. Given that there are more than 200 pages of search results, filters must be applied to generate search results with fewer pages.

The search page generator is responsible for generating search pages and writing the links to the "search_page" queue, which is then scraped by the "Search page scraper" worker.

Search result page scraper

This scraper scrapes through all pages of the search result and writes company links onto "company_link" queue so that "Company scraper" can scrape them. We are not able to scrape more than

200 pages for every search result. This is the only scraper which needs authentication. Therefore there is another worker to generate valid cookies.

Note: To prevent long queue messages and maintain the performance of the scraper, it is designed to stop when the number of unscraped company links in the queue exceeds a certain threshold. Specifically, the scraper will pause when there are more than 30,000 unscraped links in the queue and resume when the number decreases to a safe level. This ensures that the scraper can continue to operate efficiently without overloading the RabbitMQ message queue.

Company scraper

The scraper is designed to extract the following types of company data:

- Basic company information, including the company's name, address, phone number, email, website, and other relevant details.
- Detailed company information, including the company's industry, type, size, and other relevant details.
- Branch locations and information about each branch.
- Key people associated with the company, such as executives and board members.
- information about investors who have provided funding to the company.
- Information about shareholders who own stock in the company.
- Annual reports filed by the company which contain financial information and employee count.

Company saver worker

To prevent simultaneous queries, the scraper utilizes queues to store data in the database. If a company already exists in the database, it will be updated.

3.2 Implementation of Scrapper

Running the scraper requires both RabbitMQ and MySQL. We can use existing installations or run them using Docker and the docker-compose.yml file provided in this repository. It is necessary to have **Python3** in order to run the scraper, as it has been written using this programming language. We can run the scraper using the instructions below:

1. Clone the repository
2. To run the Docker Compose file, navigate to the root folder of the project in your terminal and use the command **docker-compose up -d**. This will start the RabbitMQ and MySQL services in detached mode.
3. To configure the scraper, copy the config.example.py file to config.py using the command **cp config.example.py config.py**. Then, edit the contents of the config.py file to reflect your MySQL and RabbitMQ configurations
4. Install python dependencies using **pip install -r requirements.txt**
5. If you are not using docker, simply **import storage/scheme.sql** file into database.
6. To run the master script, use the command **python master.py** in your terminal. This will initiate the scraping process and coordinate communication between the various workers.
7. To run the worker script, use the command **python worker.py** in your terminal. This will start the worker process and enable it to receive and process tasks from the RabbitMQ message queue.

You can run as many workers as you need to scale the scraper. Once you have initiated the master script, you can run workers from any server, and they will be able to communicate with the master process through the RabbitMQ message queue. This enables you to easily scale the scraper to handle larger workloads.

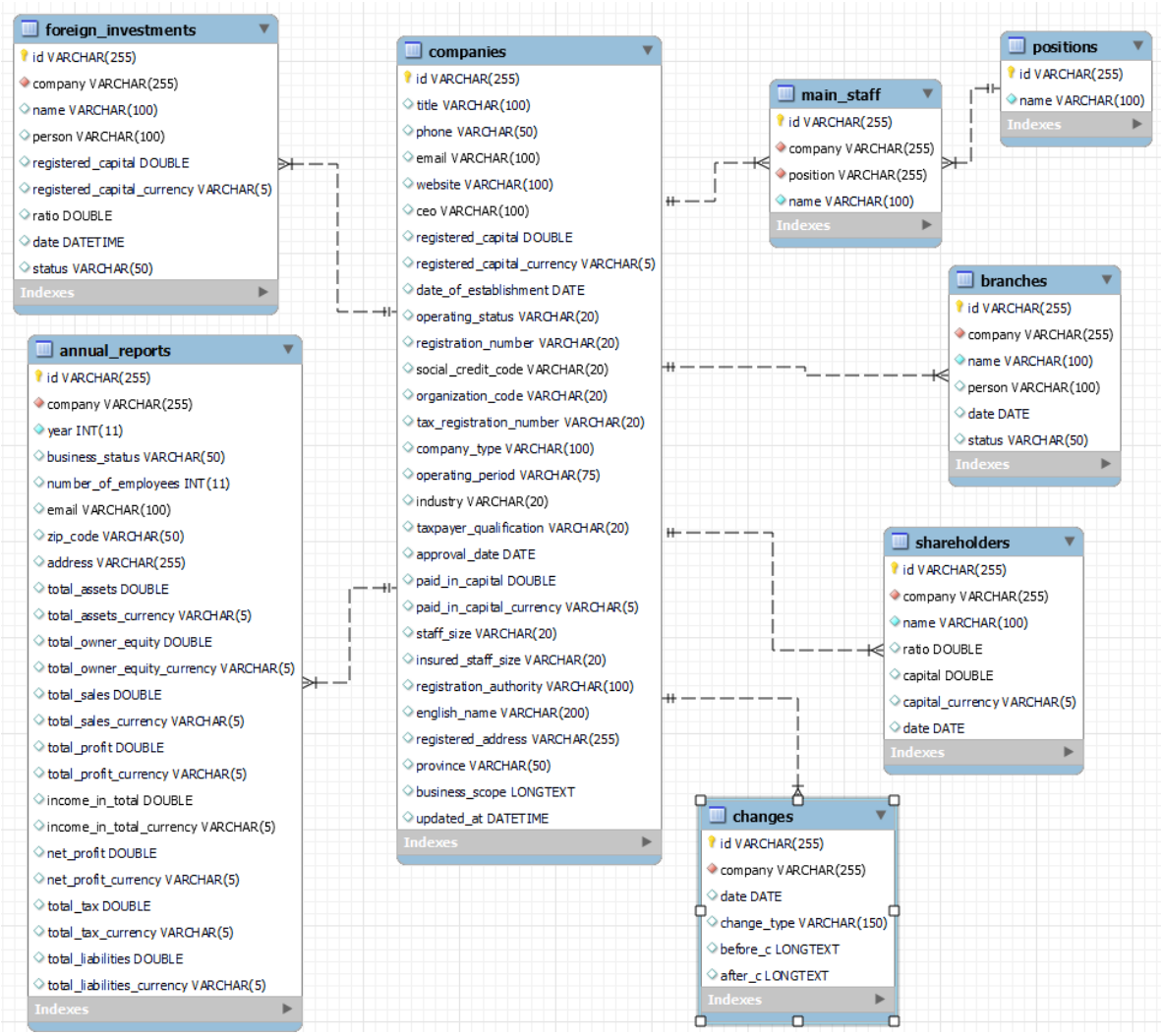

Figure 2: Entity Relationship Diagram

4 Database

MySQL is used as the database to store the data. The entity relationship diagram for the database schema is given n fig2.

There are seven tables in total.

The metadata regarding the tables and their columns are given in the tables.

Table: **companies**		
Column	Datatype	Remark
id	varchar(255)	PK(uinque id of the company)
title	varchar(100)	Name of the company
phone	varchar(50)	Contact number of the company
email	varchar(100)	Email id of the company
website	varchar(100)	Company website
ceo	varchar(100)	Name of the CEO of the company
registered_capital	double	Maximum amount of share capital that a company is authorized to raise
date_of_establishment	date	Date of Establishment
operating_status	varchar(20)	Operating state of the company
registration_numbe	varchar(20)	Company's registration number

organization_code	varchar(20)	used for grouping accounts under a specific revenue or administrative center
tax_registration_number	varchar(20)	Tax registration number
company_type	varchar(100)	Category of the company
operating_period	varchar(75)	Operational period of the company
industry	varchar(20)	The industry that the company belong to
taxpayer_qualification	varchar(20)	
approval_date	date	
paid_in_capital	double	
paid_in_capital_currency	varchar(5)	currency indicator for paid_in_capital
staff_size	varchar(20)	Number of employees in range
insured_staff_size	varchar(20)	
registration_authority	varchar(100)	State agency authorised to issue investment certificates in accordance with the Law
english_name	varchar(200)	English name of the company
registered_address	varchar(150)	Address of the company
province	varchar(50)	Province in which the headquarters are situated
business_scope	longtext	Description of the business of the company
updated_at	datetime	Timestamp when the information was last updated

Table: **annual_reports**		
Column	Datatype	Remark
id	varchar(255)	PK (generated)
company	varchar(255)	Foreign key to company id in companies table
year	int(11)	year of the balance sheet information
business_status	varchar(50)	Status in which the business was in the particular year
number_of_employees	int(11)	number of employees in the particular year
email	varchar(100)	email of the comapny
total_assets	double	Total assets of the company in Yuan
total_assets_currency	varchar(5)	Currency Indicator for total_assets
total_owner_equity	double	Total equity of the owner in the particular year
total_sales	double	Total sales in the particular year in Yuan
total_sales_currency	varchar(5)	Currency Indicator for total_sales
total_profit	double	Total profit in the particular year in Yuan
total_profit_currency	varchar(5)	Currency Indicator for total_profit
income_in_total	double	Totlal income in the particular year in Yuan
income_in_total_currency	varchar(5)	Currency Indicator for income_in_total
net_profit	double	Net profit of the company in the particular year in yuan
net_profit_currency	varchar(5)	Currency Indicator for net_profit
total_tax	double	Total tax for paid by the company in the particular year in Yuan
total_tax_currency	varchar(5)	Currency Indicator for total_tax
total_liabilities	double	Liabilities of the company in the particular year in Yuan
total_liabilities_currency	varchar(5)	Currency Indicator for total_liabilities

Table: **branches**		
Column	Datatype	Remark
id	varchar(255)	Primary Key

company	varchar(255)	Company id as Foreign key from companies table
name	varchar(100)	Name of the company branch
person	varchar(100)	Branch Manager's name
date	date	Date of establishment of the branch
status	varchar(50)	operating status of the branch

Table: **changes**		
Column	Datatype	Remark
id	varchar(255)	Primary Key
company	varchar(255)	Foreign key company id from companies table
date	date	Date when the change occurred
change_type	varchar(150)	The type of change that happened in the company
before_c	longtext	The status before the change
after_c	longtext	The status after the change or the new additions

Table: **foreign_investments**		
Column	Datatype	Remark
id	varchar(255)	Primary Key
company	varchar(255)	Foreign key company id from companies table
name	varchar(100)	Name of the entity invested in
person	varchar(100)	Person in charge of the entity in which the company has invested
registered_capital	double	Total capital invested on the entity
registered_capital_currency	varchar(5)	Currency Indicator for registered_capital
ratio	double	Ratio/Percentage of investment
date	datetime	Date of investment
status	varchar(50)	Status of investment

Table: **main_staff**		
Column	Datatype	Remark
id	varchar(255)	Primary Key
company	varchar(255)	Foreign key company id from companies table
name	varchar(100)	Name of the C-Level Employee
position	varchar(255)	id of the position to map to positions table

Table: **positions**		
Column	Datatype	Remark
id	varchar(255) PK	id of the position
name	varchar(100)	Designation of the C-Level employee

Table: **shareholders**		
Column	Datatype	Remark
id	varchar(255)	Primary Key
company	varchar(255)	Foreign key company id from companies table
name	varchar(100)	Name of the shareholder in the company
ratio	double	Percentage of shares held by the shareholder
capital	double	Capital of the shareholder

date	date	Date

4.1 Preprocessing of Data

The data that has been scraped is fairly clean. The pre-processing steps that are taken to refine the data are given below:

- The data has undergone additional cleaning to replace any strings that said 'the enterprise does not wish to publish a certain data' with "None" to indicate a lack of data.
- For values that should be numeric, non-numeric characters such as have been removed for example in the employee numbers field.
- Any currency value that was represented with 'ten thousand' or 'millions' as a suffix to the numeric value, has been removed and the values are converted to the unit currency metric.
- A problem was encountered that some currency fields were represented in Chinese Yuan, some in USD and some in EUR. After discussing this with the mentor, we added an additional currency indicator field for each of the currency fields.
- Main staff data has also been cleaned up, with multiple positions being separated into a new table to hold positions and the main staff positions being stored as a many-to-many relationship.

5 Real-time Power BI Dashboard for Data Tracking

The goal of introducing a BI Dashboard is to gather comprehensive information on the companies in the database, including their general information, revenue insights, foreign investments, and number of branches, in one place. Dashboards are an ideal tool to analyze data and the Power BI dashboard can provide a clearer and more accessible view of the Chinese company data. This real-time report allows users to view all relevant information on a single report, which is not easily achievable in the mySQL environment without writing queries each time. The dashboard is connected to the mySQL database and updates in real-time with any changes made by the real-time scrapper. Fig 3 shows how the dashboard looks. The report gives the option to select a company to view from the drop-down list. Once selected, it displays all the relevant information of the selected company that is saved in all different tables. The report can be refreshed or auto-refreshed when published to get the updated data from the database. Currently, the BI dashboard has not been published which means it is only accessible for local use. But there is a possibility to publish it with a subscription and make it available globally.

6 Results and Findings

The scraper runs continuously and updates the database in real-time. As of 1st of March 2023, the total number of companies scraped are **903685**. Some of the results are given in the table.

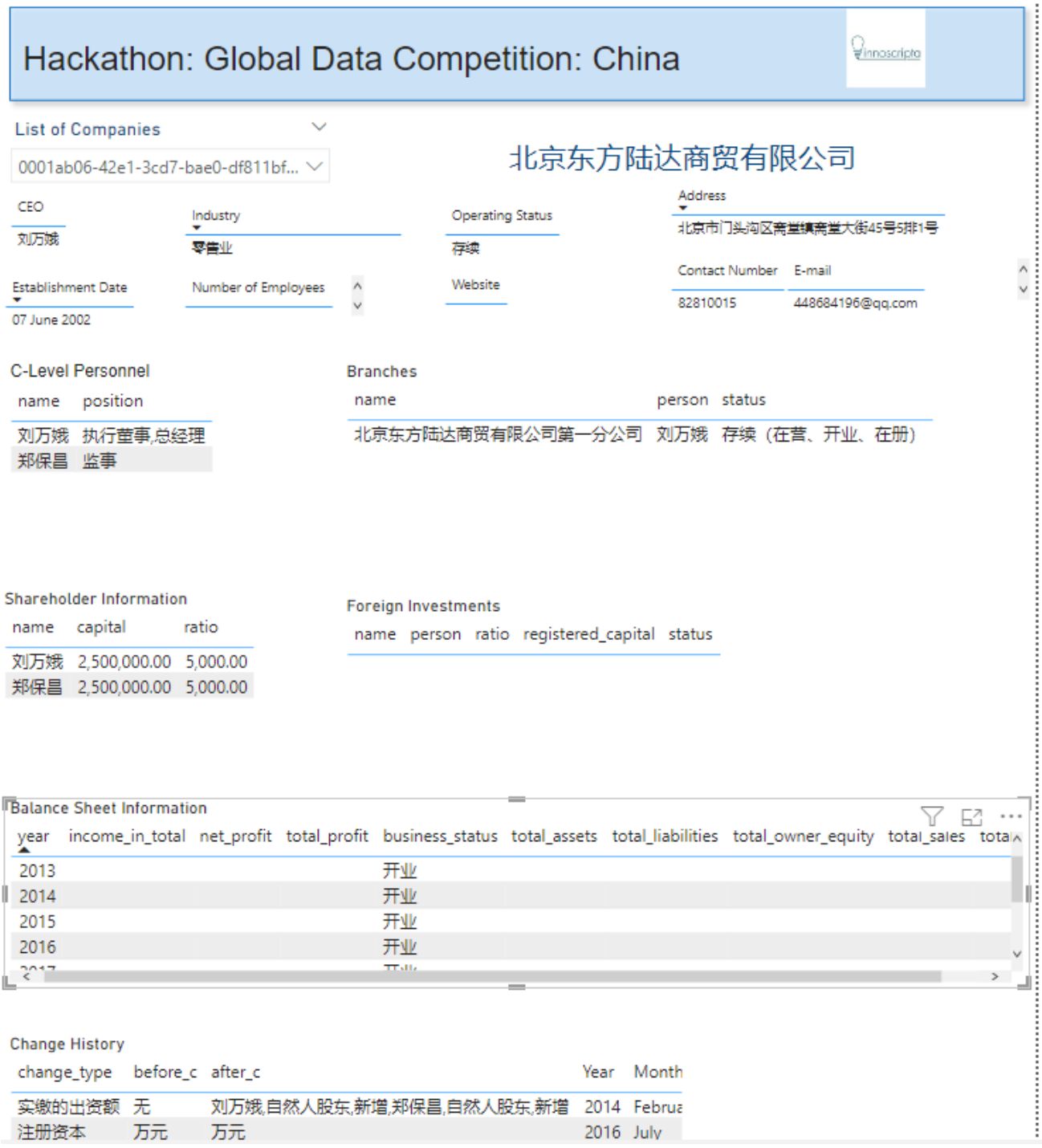

Figure 3: Power BI Dashboard

Query	Result
Total number of companies scraped	903685
Number of companies with contact information	450833
Companies with website information available	42461
Companies with CEO information available	902021
Companies with addresses available	901570
Companies with information on the number of employees	147376
Companies with industry information available	905870
Companies with registered address; atleast one contact info and ceo information	451457
Companies that have information on C-Level personnel	523753
Companies with information regarding company branches	46049
Companies with change history available	290241
Companies with shareholder information	404433
Companies with information on their foreign investments	34063
Companies with information on the balance sheet and revenue and profit	373353

Based on the above results, we can make several inferences about the dataset. The dataset includes a large number of companies, with over 903,000 companies scraped in total. A fraction of companies has updated their website information in the website. But when it comes to contact information, there are still are a large number of companies with either phone numbers or email ids, updated. The number of companies with CEO information available is higher than the number of companies with other types of information available, suggesting that CEO information may have been a focus of the website. A significant number of companies (over 404,000) have information on shareholders, suggesting that this type of information may also have been a focus of the website in maintaining data. The number of companies with information on their foreign investments is relatively low compared to other types of information, which may suggest that the companies in the dataset have a primarily domestic focus. The number of companies with information on C-Level personnel is significant. The subset of companies that have a registered address, at least one form of contact information, and CEO information may be particularly useful for certain types of analysis or research. Overall, these inferences suggest that the dataset contains a large amount of potentially valuable information and the gaps or missing information for certain companies have all the possibility to be updated in the database as the scraper will be running continuously.

6.1 Geographical Analysis

There are 30 different provinces and every company in the dataset belongs to one of them. In some cases where the address is not updated for a particular company, the province is set NULL.

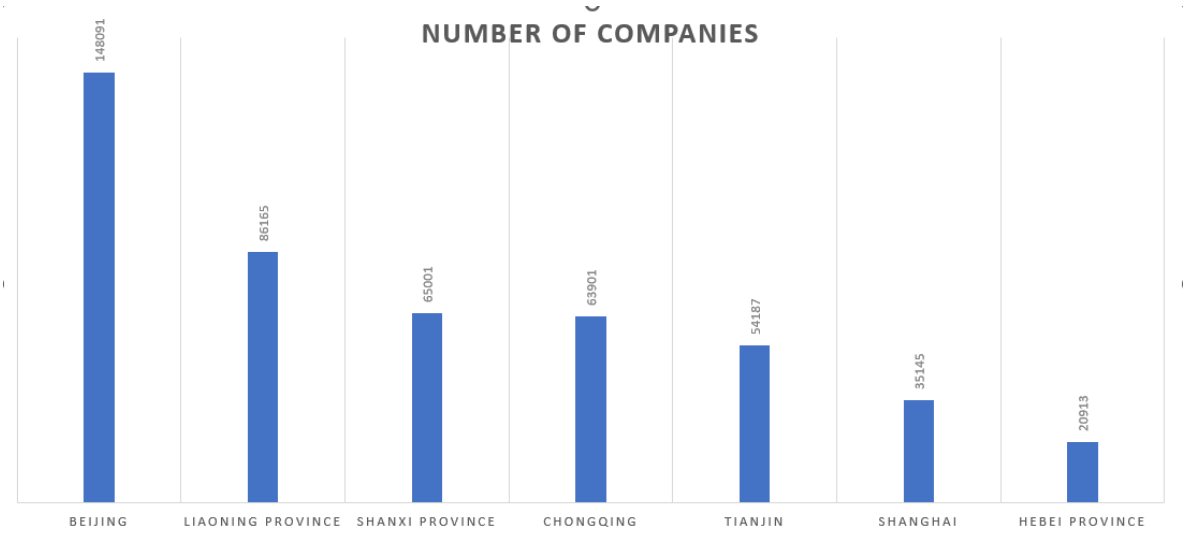

Figure 4: Geographical distribution of companies

Province	Number of Companies
Beijing	148091
Liaoning Province	86165
Shanxi Province	65001
Chongqing	63901
Tianjin	54187
Shanghai	35145
Hebei Province	20913
Shandong Province	100
Zhejiang Province	92
Inner Mongolia Autonomous Region	86
Yunnan Province	86
Jiangxi Province	85
Guizhou Province	79
Hainan Province	67
Shaanxi Province	65
Fujian Province	57
Gansu Province	54
Anhui Province	52
Sichuan Province	50
Guangdong Province	39
Jiangsu Province	38
Hubei Province	34
Heilongjiang Province	32
Hunan Province	29
Henan Province	25
Jilin Province	22
Ningxia Hui Autonomous Region	18
Qinghai Province	18
Xizang Autonomous Region	13
Guangxi Zhuang Autonomous Region	11

The above tables give the complete distribution of the number of companies in each of the provinces. Excluding the companies for which the address has not been updated, we can see that the majority of

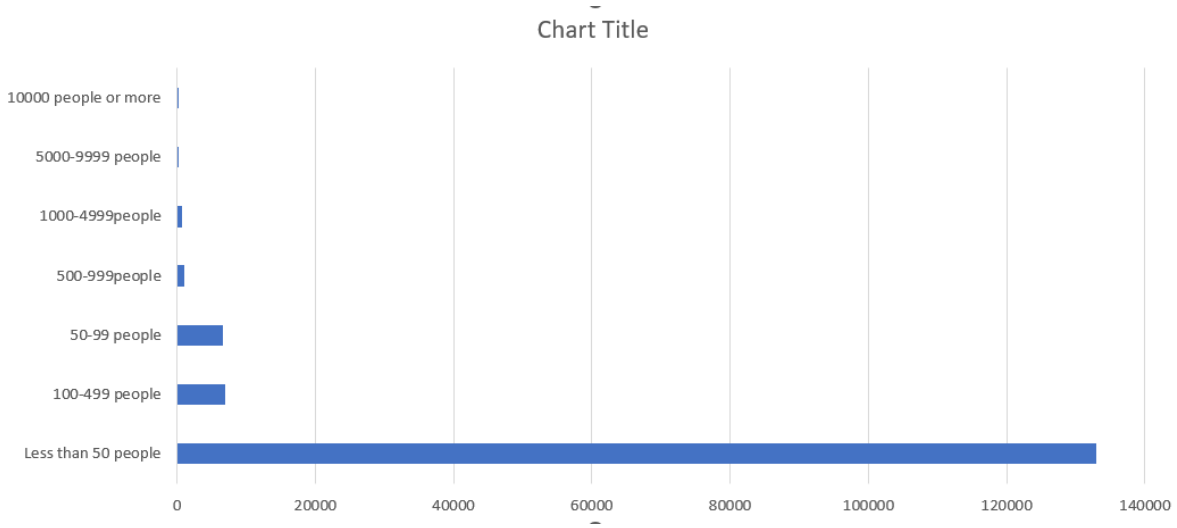

Figure 5: Number of employees to companies distribution

companies are concentrated in 6 provinces. Fig4 shows us the distribution of a number of companies over these 6 provinces. The translation of province names for this analysis is done by DeepL Translator.

6.2 Number of Employees Analysis

On the website, the staff size is given as an absolute number and also as a range. Excluding the companies for which the staff size has not been updated on the website, the distribution of companies over the staff size category is given by Fig5. The majority have less than 50 employees (132946 companies). This suggests that the dataset may include a large number of small or medium-sized businesses. Relatively few companies in the dataset have a large number of employees, with only 74 companies having 10,000 employees or more. This suggests that the dataset may be skewed towards smaller companies or companies in certain industries that do not typically have large workforces. The number of companies with staff sizes in the range of 100-999 employees is relatively low (around 6,800 companies in total), which suggests that there may be fewer mid-sized companies in the dataset compared to small or large companies.

Staff Size category	Number of Companies
Less than 50 people	132946
100-499 people	6882
50-99 people	6644
500-999people	987
1000-4999people	738
5000-9999 people	88
10000 people or more	74

6.3 Industry Analysis

The industry field refers to the category of the business to which the companies belong. The data has 94 categories of industries. We present here the top 20 industries and the number of companies that belong to them in the table and using a chart 6. The largest number of companies falls within the stevedoring and transport agency industry (33,685 companies). Other industries with a large number of companies include general equipment manufacturing, civil engineering construction, and construction and installation. Some industries with fewer companies represented in the dataset include textile manufacturing, building construction, and food manufacturing. It is possible that the composition of the dataset reflects the broader economy of the region or regions covered by the data collection effort, with certain industries being more prevalent or prominent than others. Overall, this information may be useful for conducting analyses or research on trends within specific industries or comparing

Figure 6: Industry to companies distribution

performance and characteristics across industries.

Industry	Number of Companies
Stevedoring and transport agency industry	33685
General equipment manufacturing	30256
Civil engineering construction industry	29284
Construction and installation	28219
Other service industries	28042
Metal products industry	27329
Accommodation	26752
Agriculture	25594
Road transport industry	23995
Textile, clothing and apparel industry	23543
Public facilities management industry	22155
Special equipment manufacturing	22074
Motor vehicles, electronic products and daily use products repair industry	20958
Entertainment industry	20593
Non-metallic mineral products industry	20359
Education, industry, sports and entertainment goods manufacturing	20035
Textile industry	18681
Building construction	18535
Food manufacturing	17894

6.4 Operating Status Analysis

Most companies in the database give us information about their operating status. Whether the company is running, suspended or closed. Removing Null, the data has 12 statuses. We here analyze how many companies are currently present in which status.

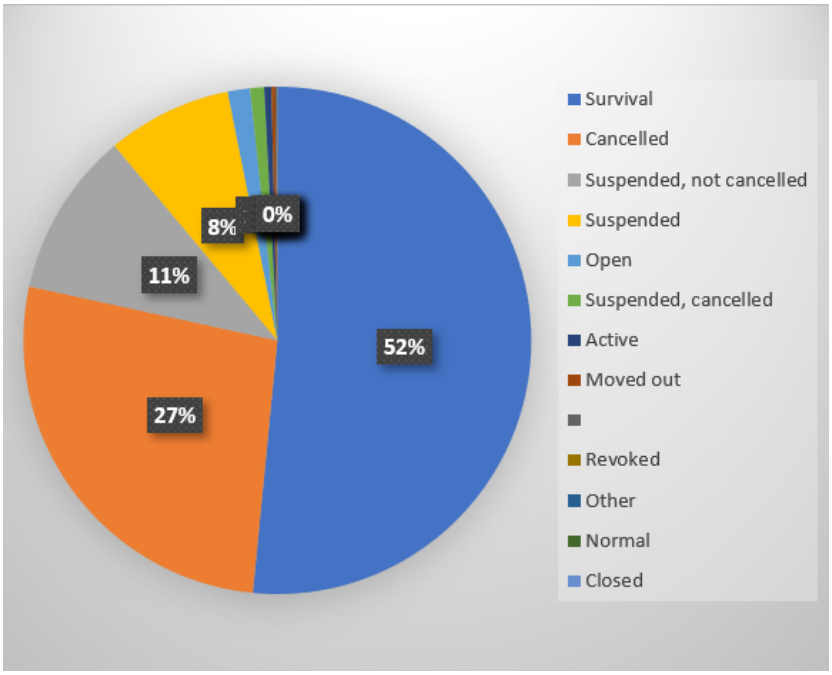

Figure 7: Distribution of companies over the operating statutes

Operating Status	Number of Companies
Survival	509097
Cancelled	265438
Suspended, not cancelled	103689
Suspended	78432
Open	14023
Suspended, cancelled	8761
Active	4554
Moved out	2406
Status not updated	1089
Revoked	134
Other	46
Normal	14
Closed	3

6.5 Annual Reports Analysis

In this section, we would analyze the top companies with respect to their, total income, assets, gross profit, net profit, tax paid by the companies and the liabilities they have. We are comparing the companies based on their annual reports for the year 2021.

Company	Total Income in CNY
China Galaxy Securities Co.	18877386900
China People's Property and Casualty Insurance Company Limited, Tianjin Branch	5529981300
China Construction Technology Group Co.	3832444598
Beijing Construction Engineering Civil Engineering Co.	3787891581
ConocoPhillips China Ltd.	2740350000
Lazard Optoelectronics Co.	1678112785
Shanghai Ming Hua Engineering Construction Co.	1494259640
Huadian (Beijing) Cogeneration Co.	1478857500
Shanghai Jingmeng Silicon Materials Co.	1109659300
China News Development Co.	998883000

Company	Total Assets in CNY
China Galaxy Securities Co.	18877386900
China People's Property and Casualty Insurance Company Limited, Tianjin Branch	5529981300
China Construction Technology Group Co.	3832444598
Beijing Construction Engineering Civil Engineering Co.	3787891581
ConocoPhillips China Ltd.	2740350000
Lazard Optoelectronics Co.	1678112785
Shanghai Ming Hua Engineering Construction Co.	1494259640
Huadian (Beijing) Cogeneration Co.	1478857500
Shanghai Jingmeng Silicon Materials Co.	1109659300
China News Development Co.	998883000

Company	Net Profit in CNY
China Galaxy Securities Co.	11723435700
Beijing Weifan Mountain Resort	633100100
Tianjin Dagang Oilfield Siyi Glass Products Factory	554807100
Lazard Optoelectronics Co.	379243602.4
Jinbin Intercity Railway Co.	349800000
Tianjin Shellwood Software Co.	240029400
Shanghai Jingmeng Silicon Material Co.	177392900
Tianjin Yichengtong Pipe Network Engineering Co.	131400000
National Bank (China) Co.	100179925.6
Matsushita Electric (China) Finance Co.	95437800

Company	Tax paid in CNY
Beijing Xin Yan Tong Hui Cleaning Service Co.	1529678000
Tianjin Dagang Oilfield Siyi Glass Products Factory	327651500
ConocoPhillips China Co.	216960000
Guoxin Securities Company Limited Beijing Branch	161031900
Shanghai Ming Hua Engineering Construction Co.	127410467.9
China Construction Technology Group Co.	114752665.3
Huadian (Beijing) Cogeneration Co.	112693700
Beijing Huadu Brewery and Food Co.	88045200
China Construction Bank Corporation Shanghai Songjiang Subbranch	68280000
China Union Qian Yuan Real Estate Fund Management Co.	40399693

Company	Liabilities in CNY
China Galaxy Securities Co.	3.77138E+11
Tianjin Dagang Oilfield Siyi Glass Products Factory	20747324200
National Bank (China) Co.	17544365205
Matsushita Electric (China) Finance Co.	11653891700
Jinbin Intercity Railway Co.	6718190000
Tianjin Jinran Thermal Power Co.	6639610000
Lazard Optoelectronics Co.	3577937134
Beijing Jiangyong Civil Engineering Co.	3265371005
Shanghai Ming Hua Engineering & Construction Co.	2049652643
Beijing Tianqiao Investment Development Company	1678198527

From the above tables, we can infer the financial performance of various companies in terms of their income, assets, net profit, total tax paid, and liabilities. For example, China Galaxy Securities Co. appears to be the top-performing company in terms of income, assets, and net profit. On the other hand, Tianjin Dagang Oilfield Siyi Glass Products Factory seems to have relatively high liabilities compared to its assets. Additionally, we can see that some companies, such as Lazard Optoelectronics Co., appear in multiple tables, indicating they have performed well across multiple financial metrics. Therefore, the tables provide a snapshot of the financial health and performance of the listed companies. However, it's important to note that these tables may not be comprehensive and may not provide a complete picture of a company's financial health.

6.6 Foreign Investment Analysis

Here we analyze which companies have the highest foreign investments.

Company	Foreign investments in CNY
State Grid Corporation	5.72733E+11
China National Petroleum Gas and Electric Group Co.	5.04427E+11
Export-Import Bank of China	4.55692E+11
China National Railway Group Co.	4.00028E+11
Beijing Guoyi Hospital Co.	3.5135E+11
China Development Financial Corporation	3.05203E+11
Beijing Xiongcai Education Technology Group Co.	2.82901E+11
China Development Bank	2.24419E+11
GCL Capital Management Co.	2.1426E+11
China Metallurgical Science and Industry Group Co.	2.06253E+11

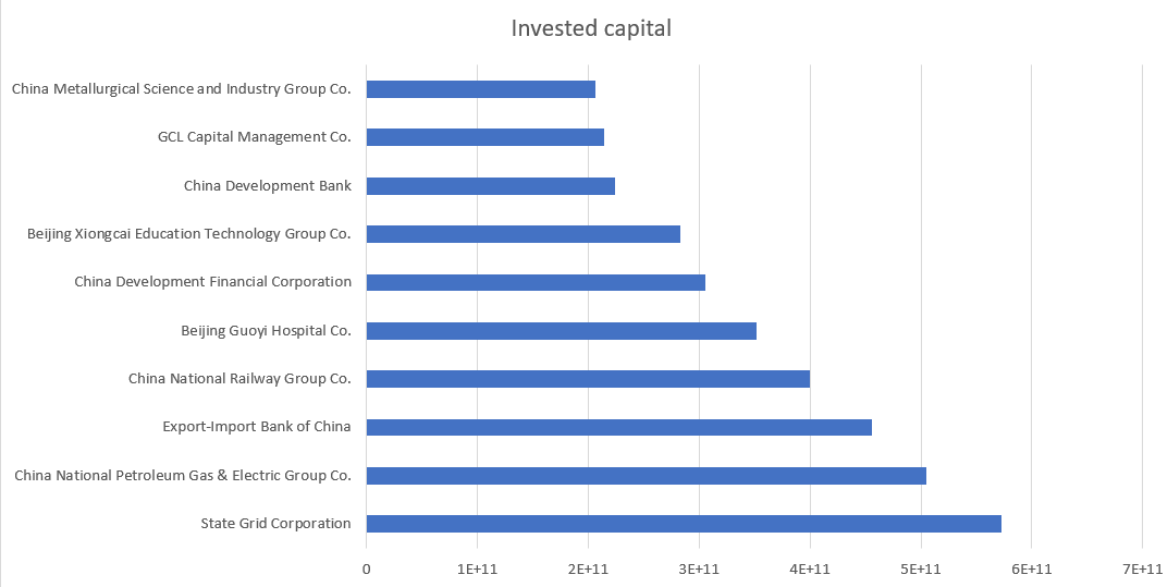

Figure 8: Companies with highest foreign investments

7 Future Work Recommendations

The major limitation that we faced in the course is that the data is in the Chinese language. Thus, one of the probable future works could be building a robust translator and integrating it with the scrapper. Using a pre-trained language model developed by Facebook AI to translate the names of the companies into English could be a possible option, as the data is not currently in the English language. The model, based on the BERT architecture, could be distilled to reduce its size and improve its efficiency. It could then be trained on a large corpus of text data using a masked language modelling objective to predict the missing words in a sentence. The model is then expected to achieve high accuracy in translating various languages, including low-resource languages. Additionally, its efficiency is expected to be improved compared to the original BERT model. This could make the "facebook/nllb-200-distilled-600M" model a more practical option for real-time translation applications, reducing the time and cost of translation projects. This will help the data get translated during pre-processing stage before it is stored in the database. Another future recommendation could be the use of stream processing(Kafka/Redpanda) and real-time databases (Rockset/Pinot) for storing and processing data. As we see the amount of data we scrap from the website is quite large and this future is likely to increase, it would a good option to switch to such big data managing technologies. Moreover, Real-time databases and stream processing systems provide low latency data access and processing, high

scalability, and flexibility for changing requirements. They are based on an event-driven architecture, which enables real-time processing of data and complex processing on data streams. These systems are ideal for applications that require real-time data processing and analysis as we are doing in this case. A more advanced analysis could be carried out with the data and can be integrated into the BI dashboard for real-time tracking of the data.

8 Conclusion

To conclude and summarize everything, in this project, we invested a good amount of time in finding the right data source and researching over scrapping the maximum amount of data despite having restrictions on the number of records being accessed per page. We found workarounds based on filters that to a great extent solved the access problem along with the guidance of mentors providing us with registration credentials. We have then built a robust scrapper which runs continuously to scrap data, preprocess it and store it in the relevant tables in the database. We have analysed the data to provide business insights from the data. This included general quality checks, geographical analysis, distribution of companies based on the number of employees, industries, operating status and last but not least, finding out the top 10 companies based on their financial aspects. We have also built a real-time local BI Dashboard to look into all the connected information of a company in a single place and track the updates. Last but not least we have provided recommendations for possible future work where we have invested time in researching a pre-trained language model that can be used as a translator in the step of pre-processing for the data to be available in English. We hope that we managed to build a solid foundation for complex analysis that could also provide a vast scope in integrating AI into businesses.