INNOSCRIPTA GMBH
Global Data Competition

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

Presented by :

- Abdulsamet Şahin
- Subhashree Mangaraj
- Omar El Qarchaoui

Img source: Google Images

Agenda

1. Introduction
2. Data Sourcing
3. Data Collection
4. Database
5. Results
6. Analysis
7. Real-time Power BI Dashboard
8. Future Work
9. Conclusion

Img source: Google Images

Data Sourcing

Objective of Data source search:

- Find a single or multiple data sources which can provide us with authentic company information and maximum hits.

Data Source Selection:

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

Img source: Google Images

Gongsi.com.cn

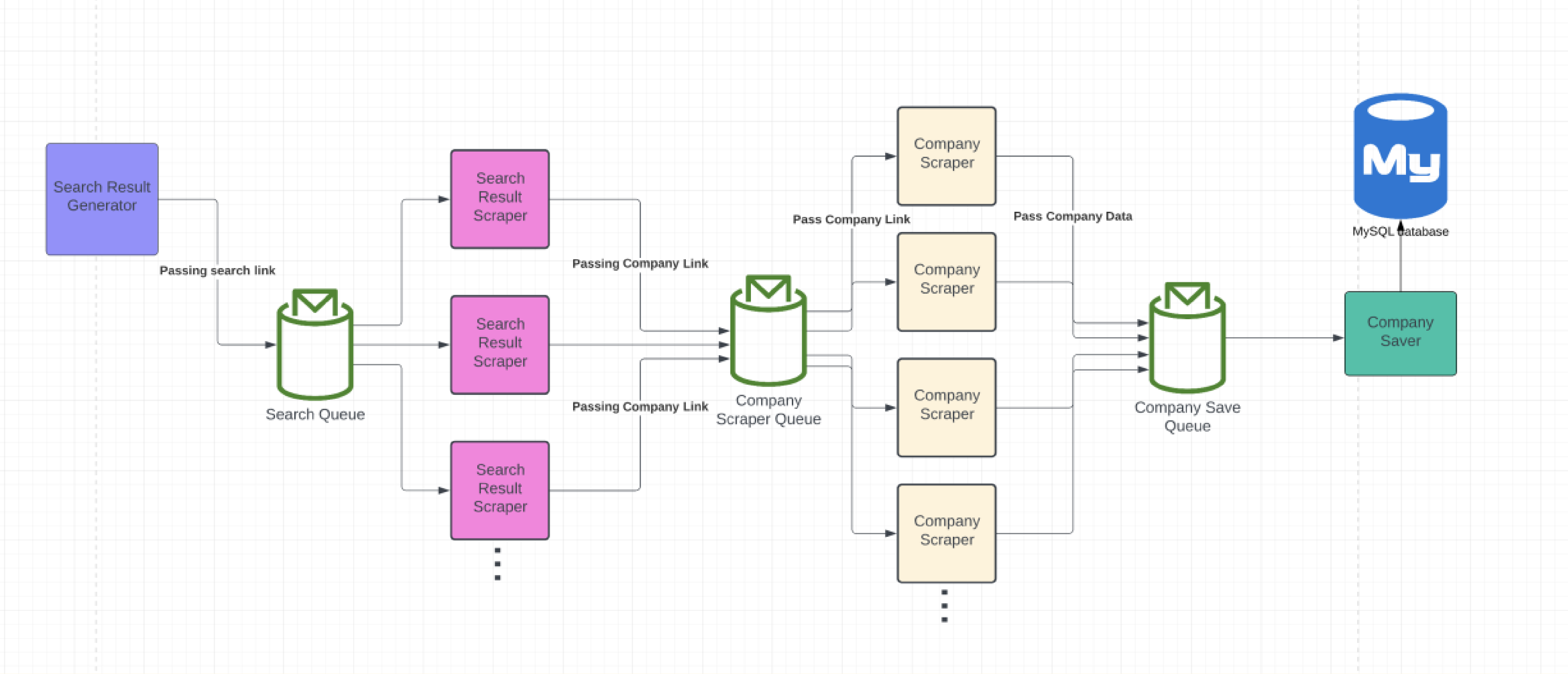
Features:

- Based on public data and available to the public.
- Contains a database of over 200 million enterprises.
- No rate limit in place.

Data Coverage:

- Basic company information
- Detailed company information
- Main staff
- Branches
- Shareholders
- Change history
- Investments
- Annual Reports

Workflow of Scraper

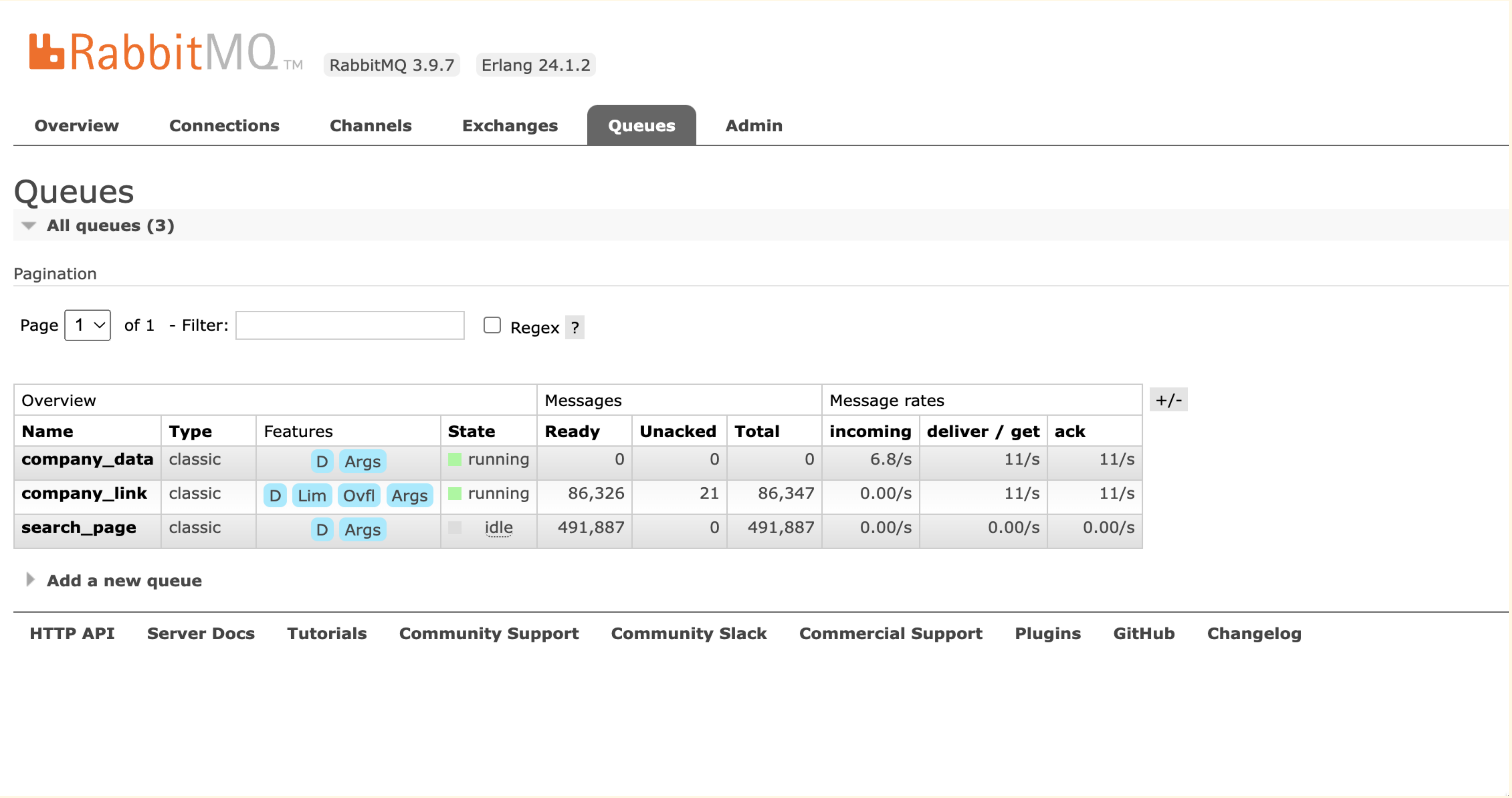


Technical Specifications

Features:

- Created using Python3
- Utilizes RabbitMQ to hold tasks and facilitate communication between workers.
- Thanks to RabbitMQ, the scraper can be easily scaled.
- Utilizes MySQL to store scraped data.
- Designed to run continuously.
- Multiple workers can scrape 1 million companies in a single day.

RabbitMQ



Queues

▾ All queues (3)

Pagination

Page 1 of 1 - Filter: □ Regex ?

Overview				Messages			Message rates		
Name	**Type**	Features	**State**	**Ready**	**Unacked**	**Total**	**incoming**	**deliver / get**	**ack**
company_data	classic	D Args	🟩 running	0	0	0	6.8/s	11/s	11/s
company_link	classic	D Lim Ovfl Args	🟩 running	86,326	21	86,347	0.00/s	11/s	11/s
search_page	classic	D Args	⬜ idle	491,887	0	491,887	0.00/s	0.00/s	0.00/s

Search Result Generator

Features:

- Starting point of the scraper.
- https://www.gongsi.com.cn/search/bs1in52eyg5eyl6
- Since there is a pagination limit, filters are used to reduce the number of pages.
- It uses combinations of province, category, and establishment year filters to generate search result.
- It will regenerate when then queue is empty.

Search Result Scraper

Features:

- The scraper collects company links by scraping the search result pages one by one, which are then scraped by the "Company Scraper" worker.
- To prevent overloading the RabbitMQ, this scraper will stop working when there are 100,000 companies unscraped.

Company Scraper

Features:

- The scraper will retrieve tasks from the previous worker (Search Page Scraper).
- It will scrape all the required data and pass it to the "Company Saver" worker.
- The scraper can be easily scaled by running as many workers as necessary.
- It is possible to run the scraper from other servers.

Company Saver

Features:

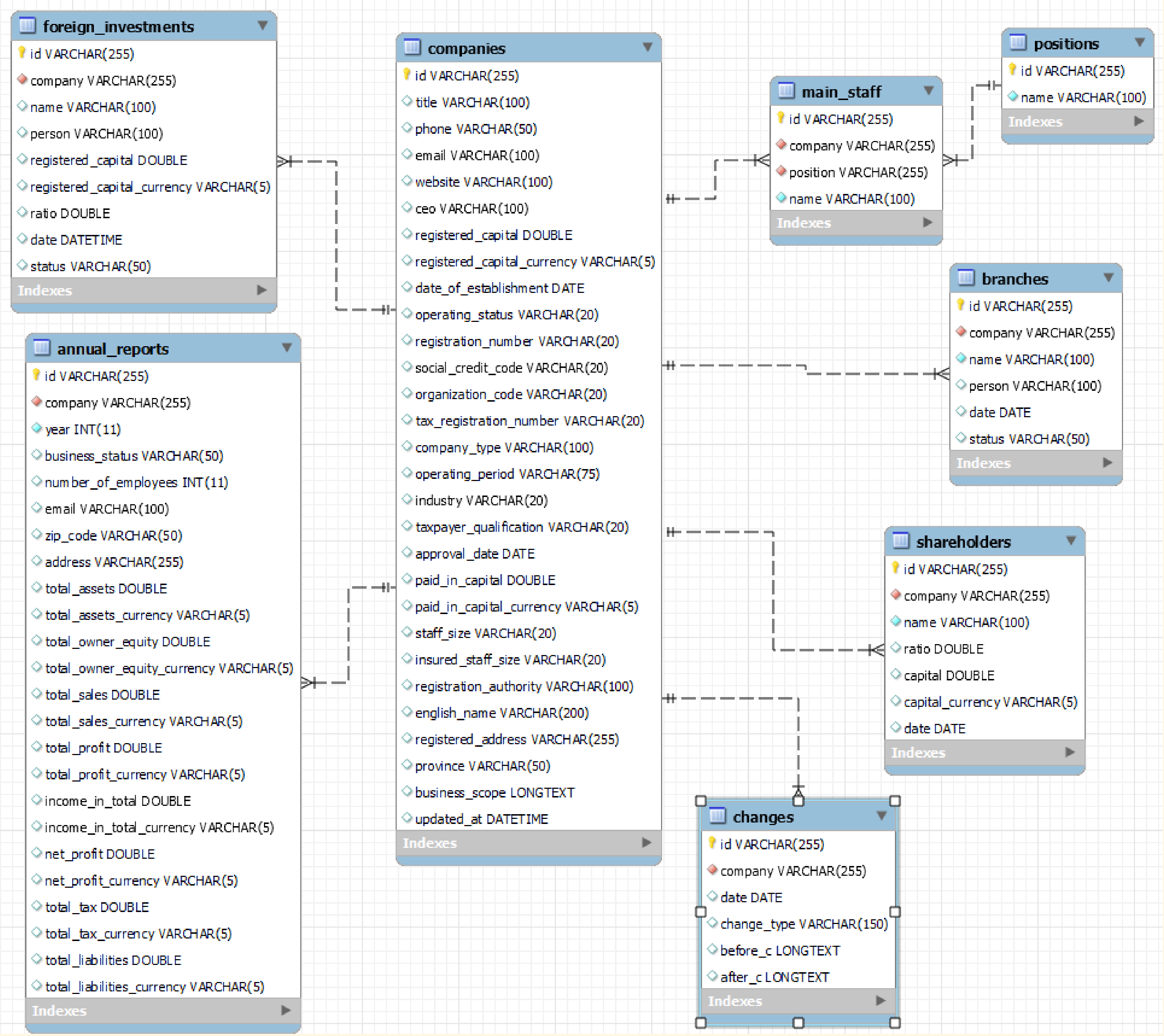
- This worker retrieves company data from the "Company Scraper" worker.
- It performs various cleaning operations on the data.
- The data is then stored in the database.
- Updates the company if already exists.
- If the company already exists, it is updated with the new data.

Data Pre-processing

Features:

- Replacing strings indicating a lack of data with "None,"
- Removing non-numeric characters from numeric values,
- Converting currency values to a unit currency metric,
- Adding currency indicator fields to address currency representation inconsistencies
- Cleaning up main staff data by separating multiple positions into a new table and storing main staff positions as a many-to-many relationship.
- The data is described as fairly clean after undergoing these pre-processing steps.

Database MySQL



foreign_investments
id VARCHAR(255)
company VARCHAR(255)
name VARCHAR(100)
person VARCHAR(100)
registered_capital DOUBLE
registered_capital_currency VARCHAR(5)
ratio DOUBLE
date DATETIME
status VARCHAR(50)
Indexes

annual_reports
id VARCHAR(255)
company VARCHAR(255)
year INT(11)
business_status VARCHAR(50)
number_of_employees INT(11)
email VARCHAR(100)
zip_code VARCHAR(50)
address VARCHAR(255)
total_assets DOUBLE
total_assets_currency VARCHAR(5)
total_owner_equity DOUBLE
total_owner_equity_currency VARCHAR(5)
total_sales DOUBLE
total_sales_currency VARCHAR(5)
total_profit DOUBLE
total_profit_currency VARCHAR(5)
income_in_total DOUBLE
income_in_total_currency VARCHAR(5)
net_profit DOUBLE
net_profit_currency VARCHAR(5)
total_tax DOUBLE
total_tax_currency VARCHAR(5)
total_liabilities DOUBLE
total_liabilities_currency VARCHAR(5)
Indexes

companies
id VARCHAR(255)
title VARCHAR(100)
phone VARCHAR(50)
email VARCHAR(100)
website VARCHAR(100)
ceo VARCHAR(100)
registered_capital DOUBLE
registered_capital_currency VARCHAR(5)
date_of_establishment DATE
operating_status VARCHAR(20)
registration_number VARCHAR(20)
social_credit_code VARCHAR(20)
organization_code VARCHAR(20)
tax_registration_number VARCHAR(20)
company_type VARCHAR(100)
operating_period VARCHAR(75)
industry VARCHAR(20)
taxpayer_qualification VARCHAR(20)
approval_date DATE
paid_in_capital DOUBLE
paid_in_capital_currency VARCHAR(5)
staff_size VARCHAR(20)
insured_staff_size VARCHAR(20)
registration_authority VARCHAR(100)
english_name VARCHAR(200)
registered_address VARCHAR(255)
province VARCHAR(50)
business_scope LONGTEXT
updated_at DATETIME
Indexes

main_staff
id VARCHAR(255)
company VARCHAR(255)
position VARCHAR(255)
name VARCHAR(100)
Indexes

positions
id VARCHAR(255)
name VARCHAR(100)
Indexes

branches
id VARCHAR(255)
company VARCHAR(255)
name VARCHAR(100)
person VARCHAR(100)
date DATE
status VARCHAR(50)
Indexes

shareholders
id VARCHAR(255)
company VARCHAR(255)
name VARCHAR(100)
ratio DOUBLE
capital DOUBLE
capital_currency VARCHAR(5)
date DATE
Indexes

changes
id VARCHAR(255)
company VARCHAR(255)
date DATE
change_type VARCHAR(150)
before_c LONGTEXT
after_c LONGTEXT
Indexes

Tables

Tables	Description
companies	General info table for companies. ‘id’ column is the primary key, columns for company details such as contact information, financial data, and registration details, and a datetime column to track updates
annual_report	contains financial information related to different companies for different years
branches	contains information related to different branches of a company
changes	information related to changes that occurred in different companies.
Foreign_investmen	contains information about the foreign investments made by a company.
Main_Staff	information about the staff of a company
positions	information about the positions of C-Level employees in a company.
shareholders	contains information about the shareholders of a company.

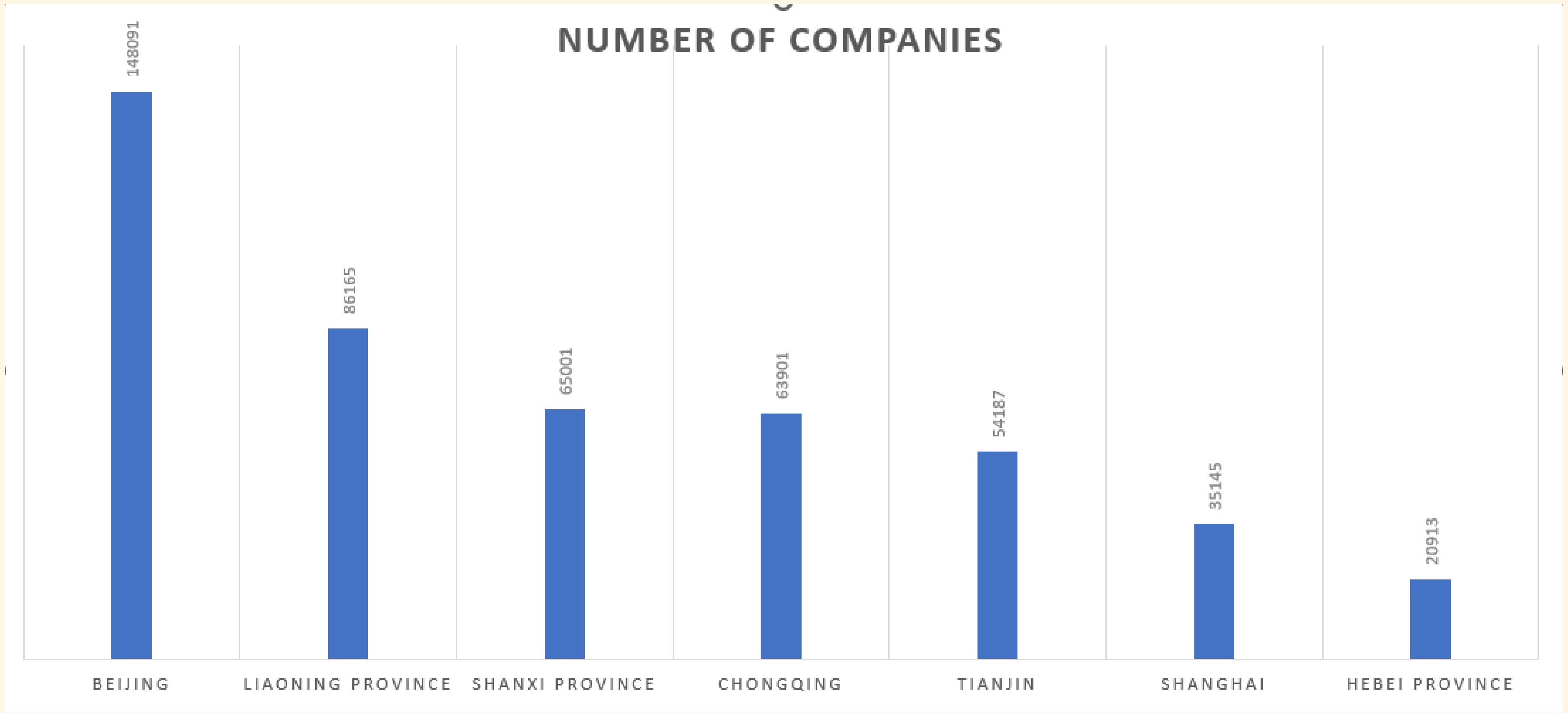
Results and Findings : 903685 total number of companies

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

Analysis

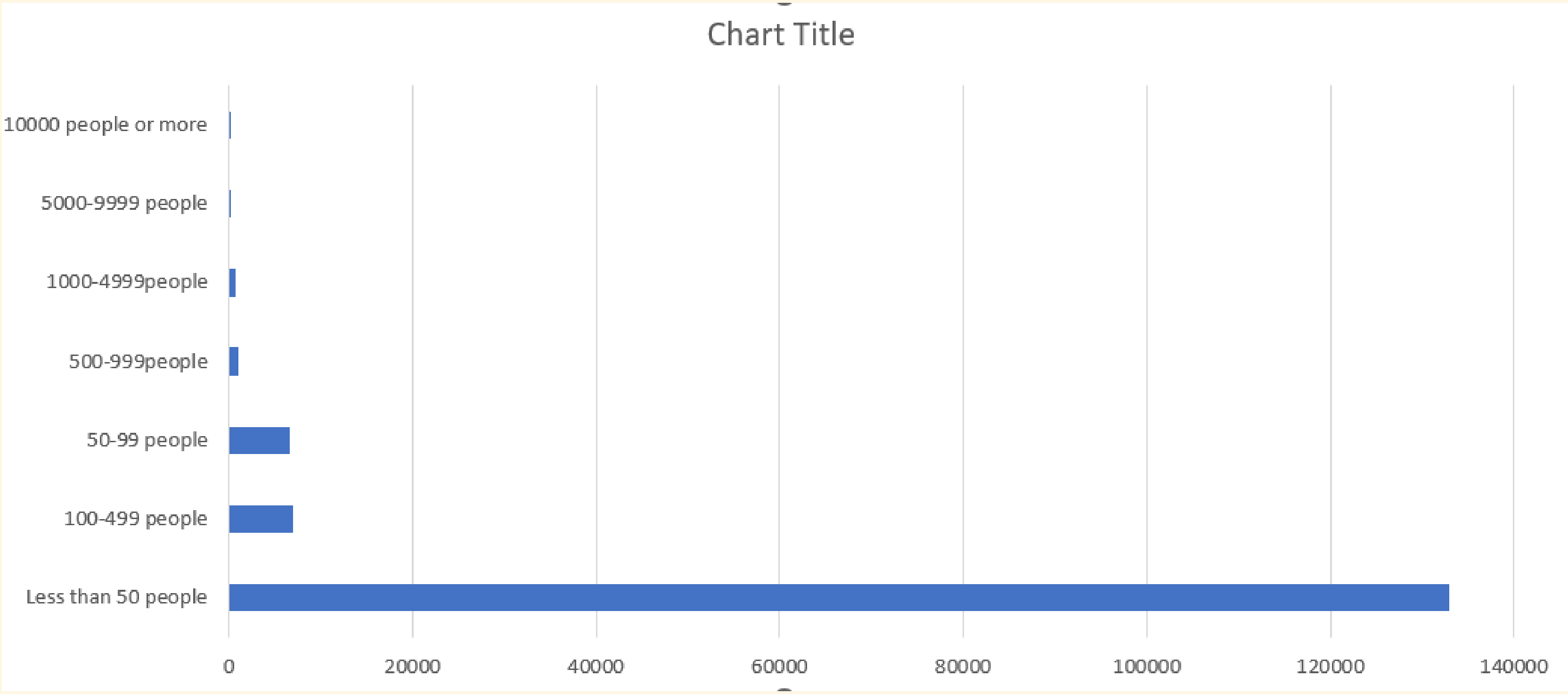
Geographical Analysis:

- There are 30 provinces
- The top 7 province where maximum of the companies reside are given below



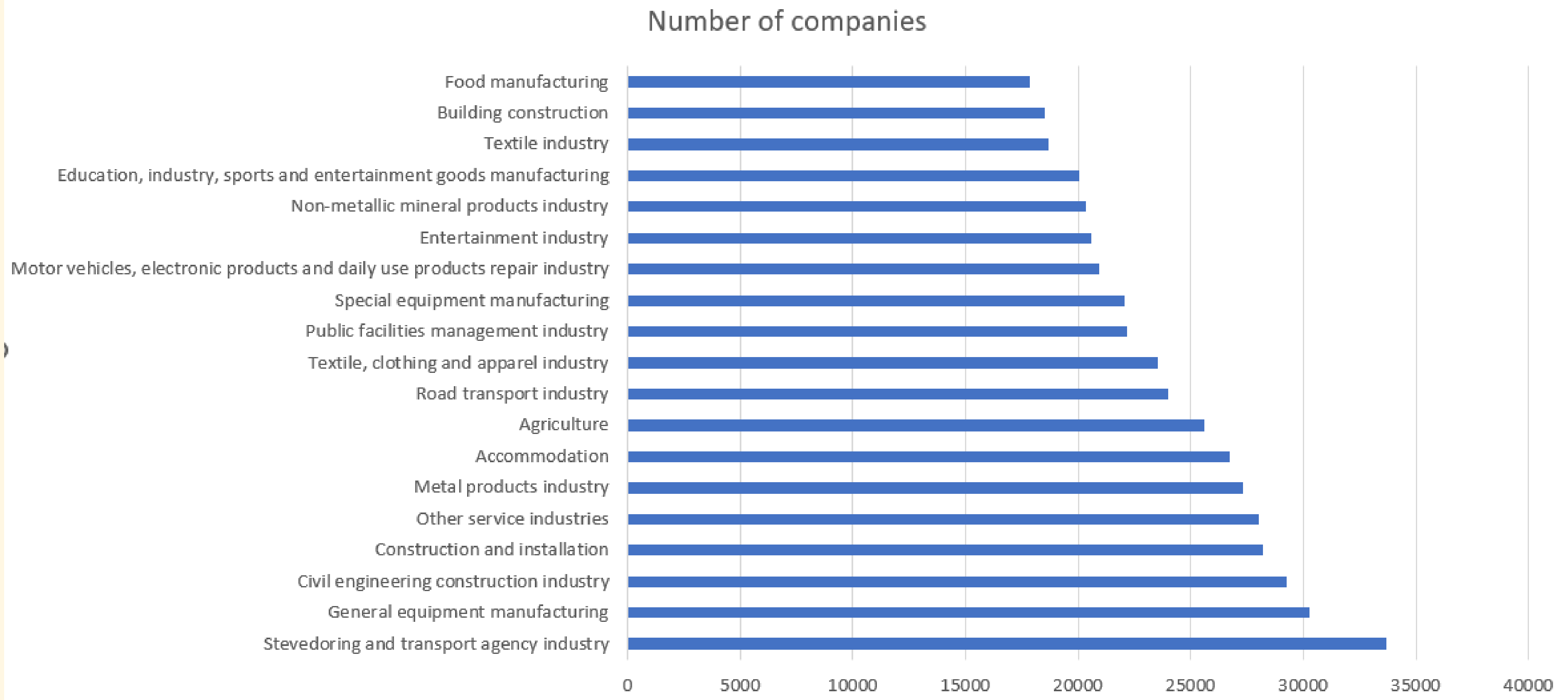
Number of Employees Analysis:

- There are 7 categories in which the companies are divided as per staff size
- Majority of the companies have staff size less than 5
- Only 74 companies having 10,000 employees or more
- Fewer mid sized companies as number of companies with staff sizes in the range of 100-999 employees is relatively low



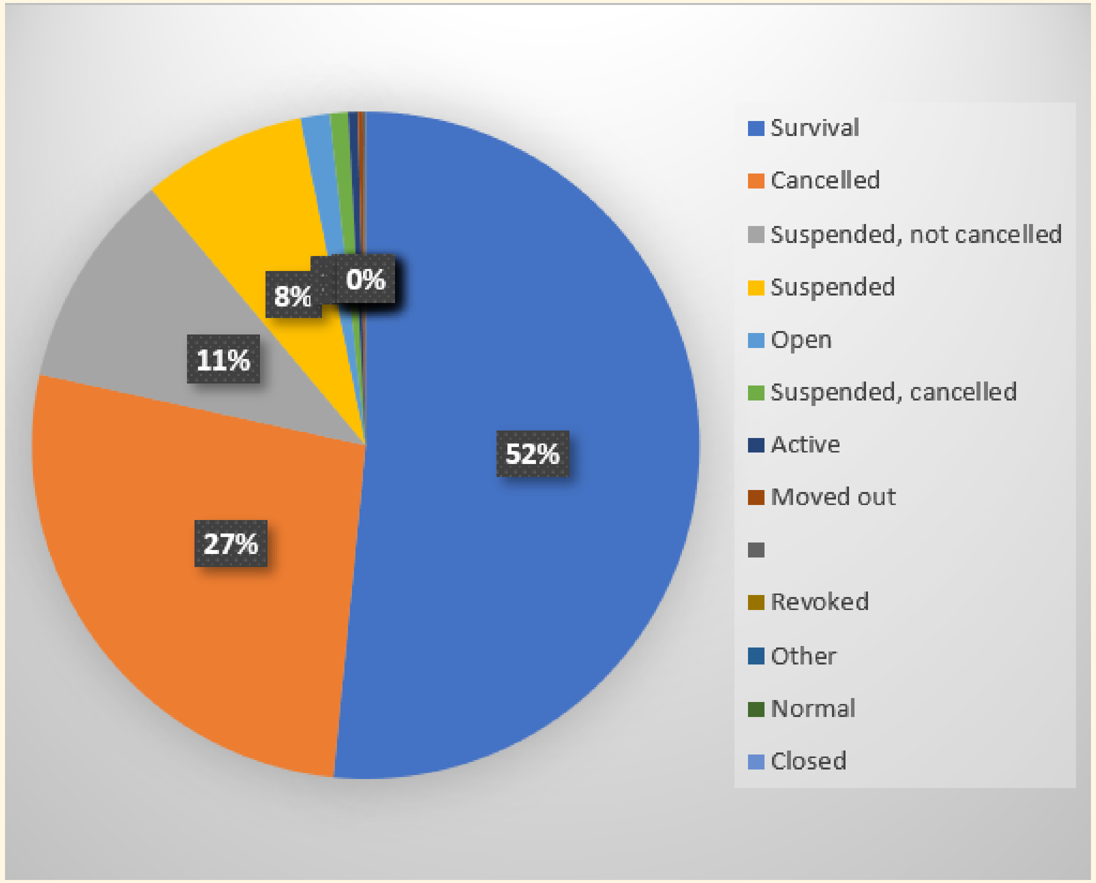
Industry Analysis:

- There are 94 categories of industries
- We present here the top 20 industries and number of companies in that category
- The largest number of companies falls within the stevedoring and transport agency industry



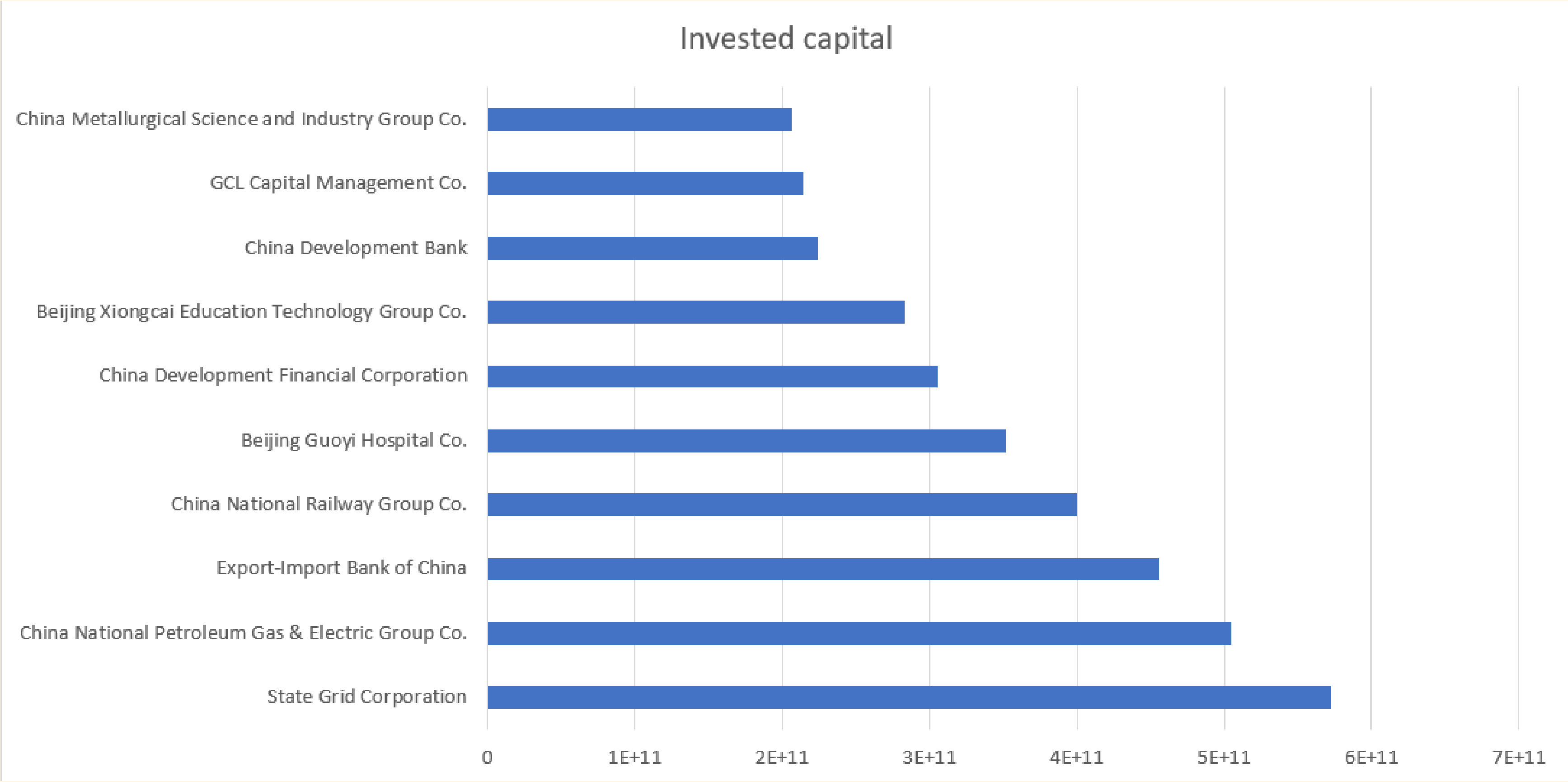
Operating Status Analysis:

- Removing Null, the data has 12 statuses
- More than 50 percent of companies in the database are on ‘Survival’ status
- Followed by companies in ‘cancelled’ status

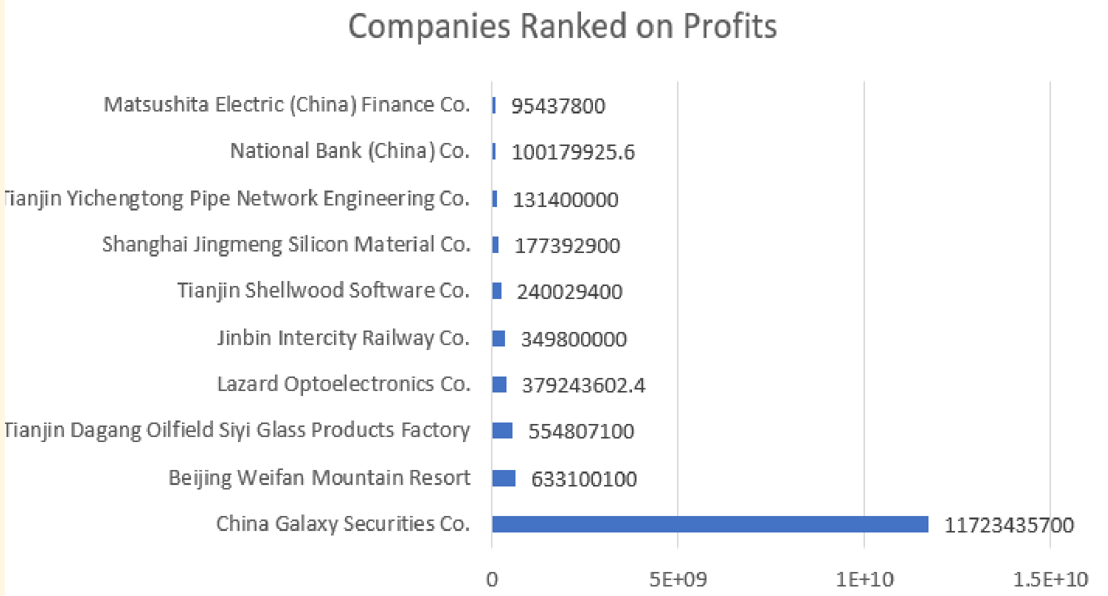


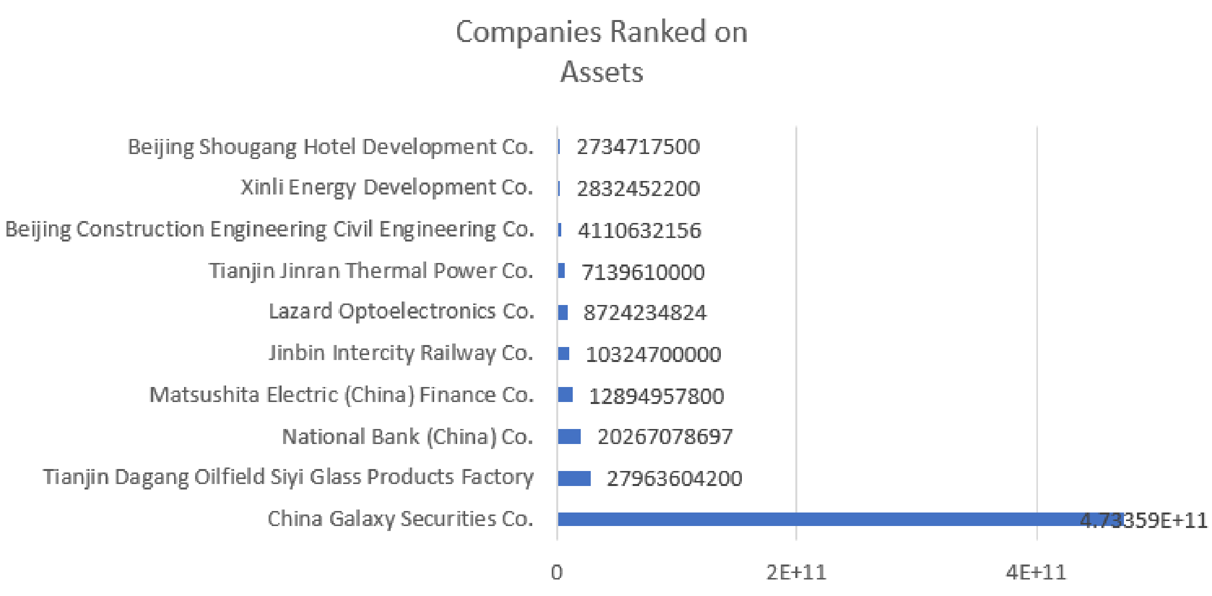
Foreign Investment Analysis:

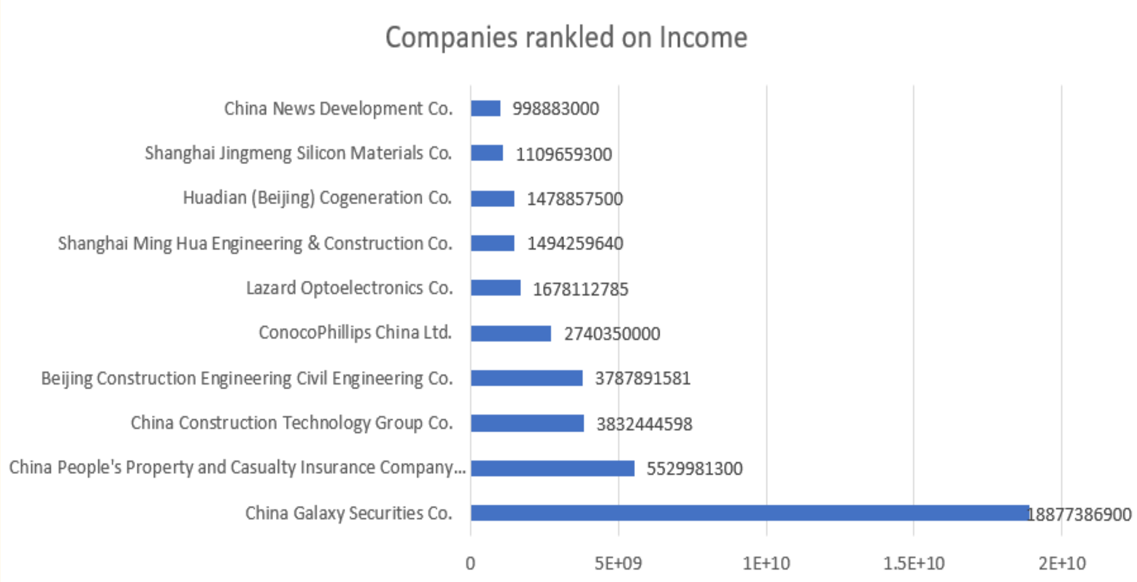
- Top 10 companies with highest foreign investments

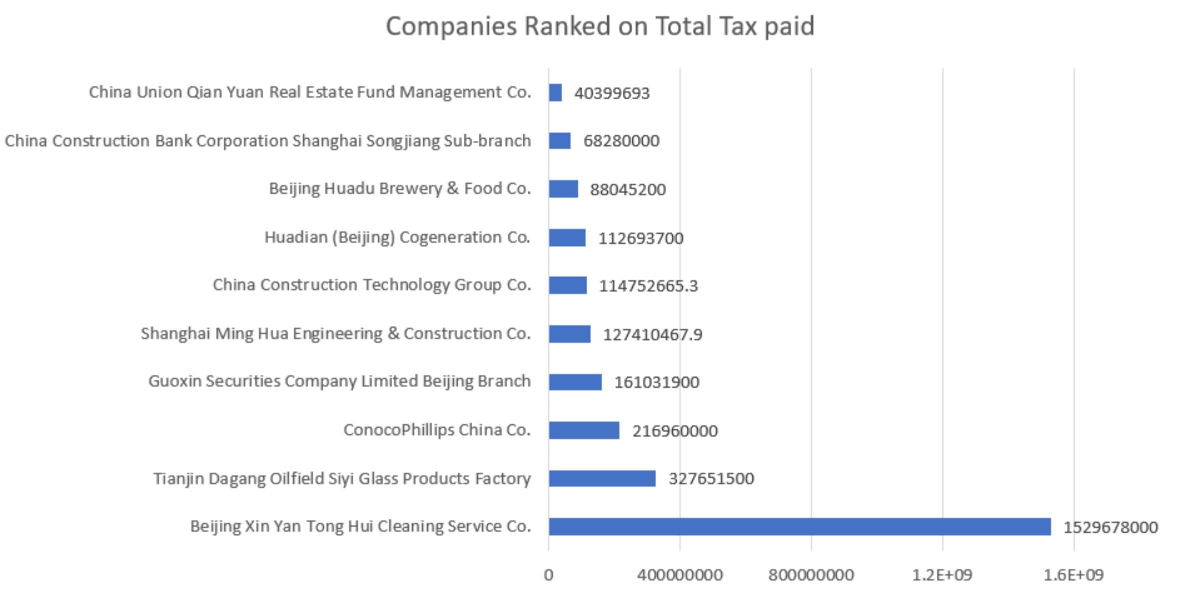


Balance Sheet and Revenue Analysis:



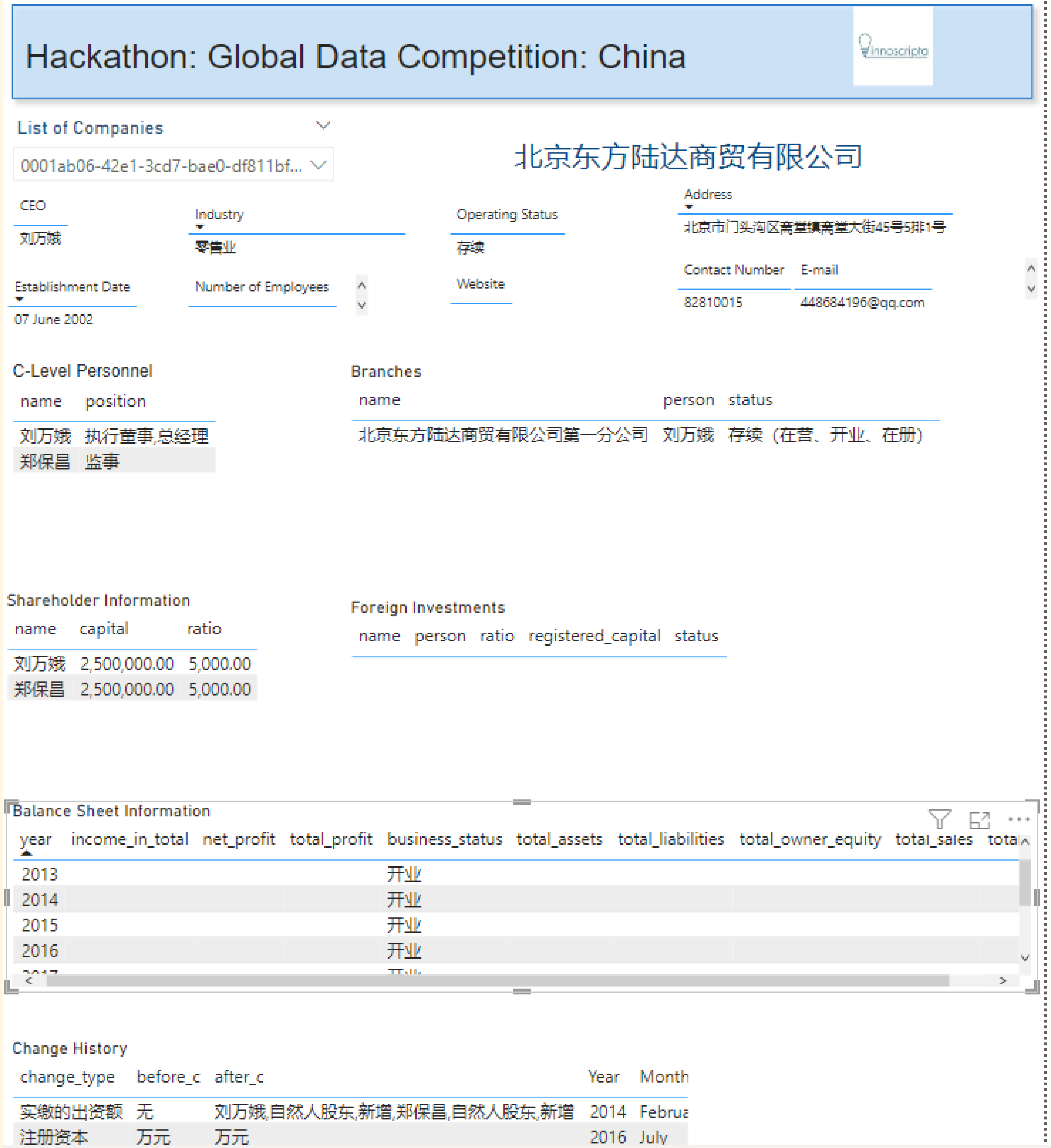






Real Time Power BI Dashboard:

- Objective is to analyse and track the complete details of each company in a single report.
- Connected to database for real time update of data
- Information displayed based on selected company id



Hackathon: Global Data Competition: China

List of Companies

0001ab06-42e1-3cd7-bae0-df811bf...

北京东方陆达商贸有限公司

CEO: 刘万娥

Industry: 零售业

Operating Status: 存续

Address: 北京市门头沟区斋堂镇斋堂大街45号5排1号

Establishment Date: 07 June 2002

Number of Employees:

Website:

Contact Number: 82810015

E-mail: 448684196@qq.com

C-Level Personnel

name	position
刘万娥	执行董事,总经理
郑保昌	监事

Branches

name	person	status
北京东方陆达商贸有限公司第一分公司	刘万娥	存续（在营、开业、在册）

Shareholder Information

name	capital	ratio
刘万娥	2,500,000.00	5,000.00
郑保昌	2,500,000.00	5,000.00

Foreign Investments

name	person	ratio	registered_capital	status

Balance Sheet Information

year	income_in_total	net_profit	total_profit	business_status	total_assets	total_liabilities	total_owner_equity	total_sales	tota
2013				开业					
2014				开业					
2015				开业					
2016				开业					
2017				开业					

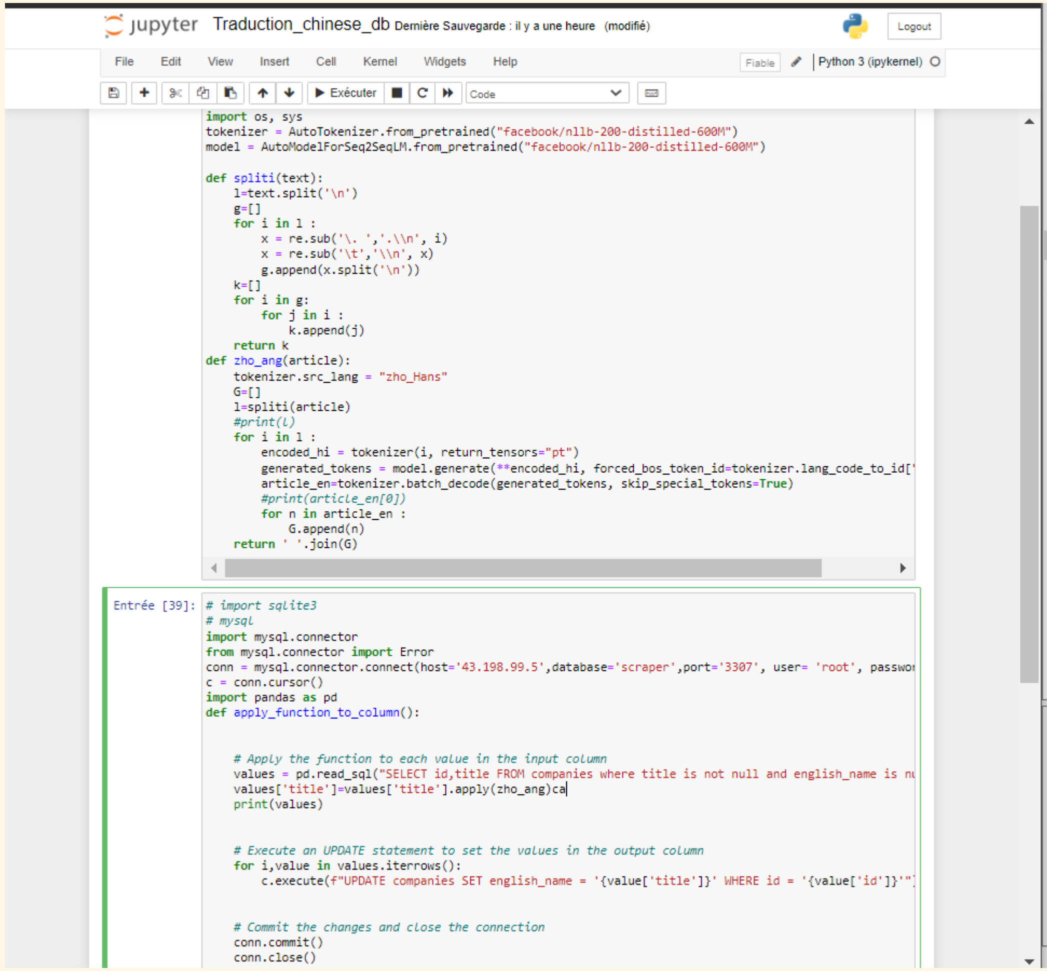
Change History

change_type	before_c	after_c	Year	Month
实缴的出资额	无	刘万娥,自然人股东,新增,郑保昌,自然人股东,新增	2014	Februa
注册资本	万元	万元	2016	July

Future Work

Pre-trained language model for tranlation

- The "facebook/nllb-200-distilled-600M" model is a highly accurate and efficient pre-trained language model that can be customized for specific languages or domains to improve its translation quality.



```python
import os, sys
tokenizer = AutoTokenizer.from_pretrained("facebook/nllb-200-distilled-600M")
model = AutoModelForSeq2SeqLM.from_pretrained("facebook/nllb-200-distilled-600M")

def spliti(text):
    l=text.split('\n')
    g=[]
    for i in l :
        x = re.sub('\. ','.\\n', i)
        x = re.sub('\t','\\n', x)
        g.append(x.split('\n'))
    k=[]
    for i in g:
        for j in i :
            k.append(j)
    return k
def zho_ang(article):
    tokenizer.src_lang = "zho_Hans"
    G=[]
    l=spliti(article)
    #print(l)
    for i in l :
        encoded_hi = tokenizer(i, return_tensors="pt")
        generated_tokens = model.generate(**encoded_hi, forced_bos_token_id=tokenizer.lang_code_to_id['
        article_en=tokenizer.batch_decode(generated_tokens, skip_special_tokens=True)
        #print(article_en[0])
        for n in article_en :
            G.append(n)
    return ' '.join(G)
```

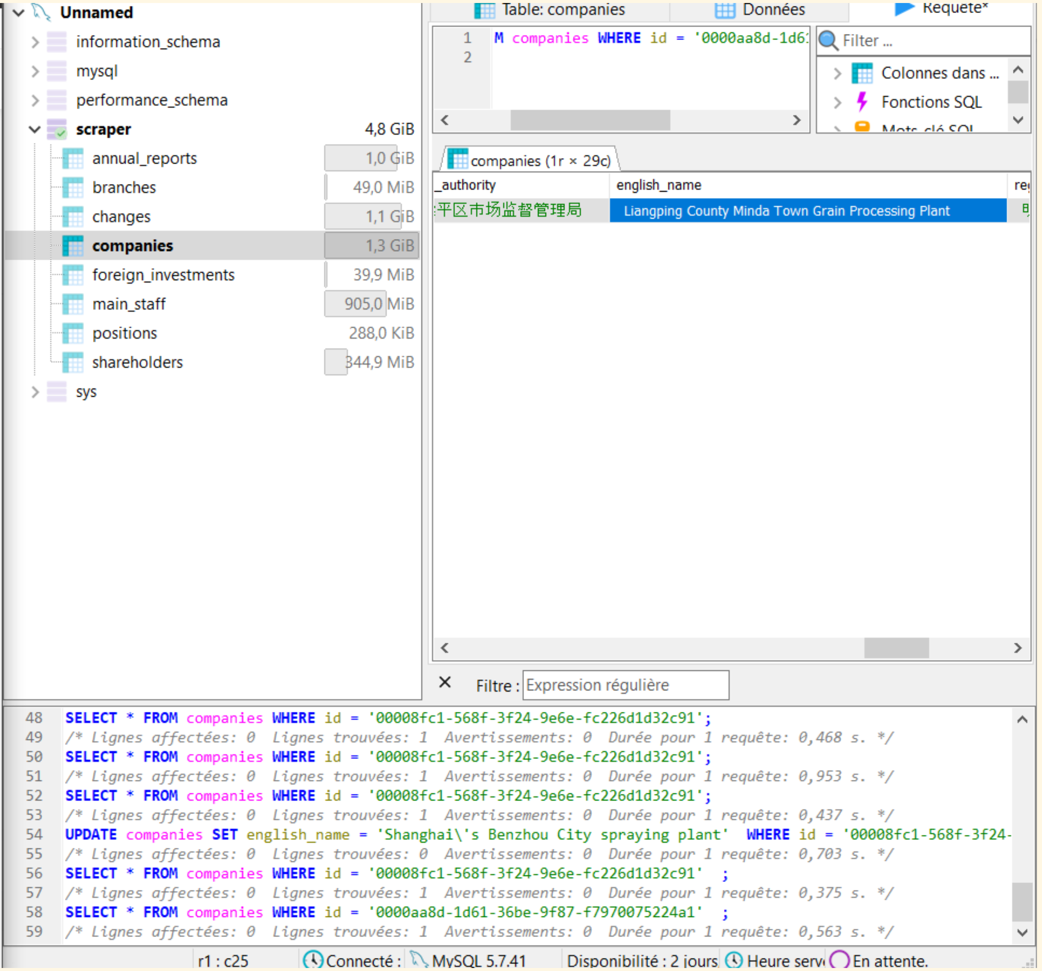
Entrée [39]:

```python
# import sqlite3
# mysql
import mysql.connector
from mysql.connector import Error
conn = mysql.connector.connect(host='43.198.99.5',database='scraper',port='3307', user= 'root', passwor
c = conn.cursor()
import pandas as pd
def apply_function_to_column():


    # Apply the function to each value in the input column
    values = pd.read_sql("SELECT id,title FROM companies where title is not null and english_name is nu
    values['title']=values['title'].apply(zho_ang)ca
    print(values)


    # Execute an UPDATE statement to set the values in the output column
    for i,value in values.iterrows():
        c.execute(f"UPDATE companies SET english_name = '{value['title']}' WHERE id = '{value['id']}'"


    # Commit the changes and close the connection
    conn.commit()
    conn.close()
```



Unnamed	
information_schema	
mysql	
performance_schema	
scraper	4,8 GiB
annual_reports	1,0 GiB
branches	49,0 MiB
changes	1,1 GiB
companies	1,3 GiB
foreign_investments	39,9 MiB
main_staff	905,0 MiB
positions	288,0 KiB
shareholders	344,9 MiB
sys	

```sql
M companies WHERE id = '0000aa8d-1d6
```

companies (1r × 29c)

_authority	english_name	re
:平区市场监督管理局	Liangping County Minda Town Grain Processing Plant	[illegible]

```
SELECT * FROM companies WHERE id = '00008fc1-568f-3f24-9e6e-fc226d1d32c91';
/* Lignes affectées: 0  Lignes trouvées: 1  Avertissements: 0  Durée pour 1 requête: 0,468 s. */
SELECT * FROM companies WHERE id = '00008fc1-568f-3f24-9e6e-fc226d1d32c91';
/* Lignes affectées: 0  Lignes trouvées: 1  Avertissements: 0  Durée pour 1 requête: 0,953 s. */
SELECT * FROM companies WHERE id = '00008fc1-568f-3f24-9e6e-fc226d1d32c91';
/* Lignes affectées: 0  Lignes trouvées: 1  Avertissements: 0  Durée pour 1 requête: 0,437 s. */
UPDATE companies SET english_name = 'Shanghai\'s Benzhou City spraying plant'  WHERE id = '00008fc1-568f-3f24-
/* Lignes affectées: 0  Lignes trouvées: 0  Avertissements: 0  Durée pour 1 requête: 0,703 s. */
SELECT * FROM companies WHERE id = '00008fc1-568f-3f24-9e6e-fc226d1d32c91'  ;
/* Lignes affectées: 0  Lignes trouvées: 1  Avertissements: 0  Durée pour 1 requête: 0,375 s. */
SELECT * FROM companies WHERE id = '0000aa8d-1d61-36be-9f87-f7970075224a1'  ;
/* Lignes affectées: 0  Lignes trouvées: 1  Avertissements: 0  Durée pour 1 requête: 0,563 s. */
```

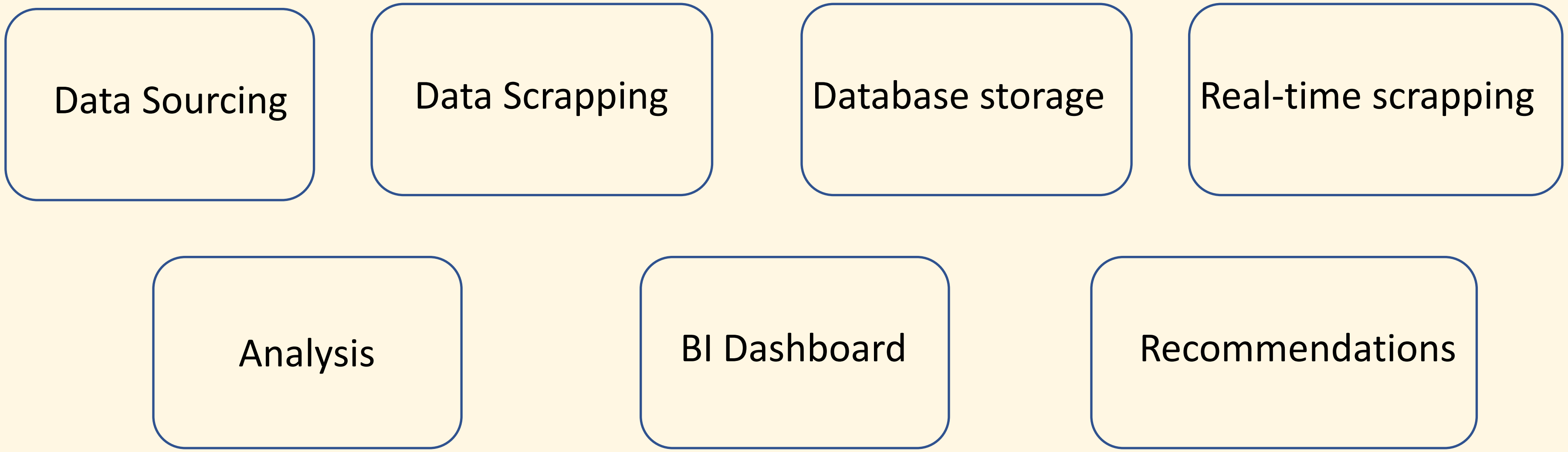
r1 : c25 | Connecté : | MySQL 5.7.41 | Disponibilité : 2 jours | Heure serv | En attente.

Future Work

Stream processing(Kafka/Redpanda) and real-time databases (Rockset/Pinot)

- Low latency data access
- High processing,
- High scalability, and
- Flexibility for changing requirements
- Ideal for applications that require real-time data processing

Conclusion



Data Sourcing

Data Scrapping

Database storage

Real-time scrapping

Analysis

BI Dashboard

Recommendations

Thank you