JACOBS UNIVERSITY

MRD006-340002 – Advanced Project 2

Project Report

Data Researcher in Data Hunting

Topic Description:

Data Researcher in Data Hunting Internship: **Munich Re (Group)**

Written By:
Vijaya Nawale
(30002931)

Supervised by:

Dr. Tobias Hirzinger
&
Prof. Dr. Adalbert F.X. Wilhelm

February 2, 2021

Abstract

This paper describes a report on **Internship at Munich Re (Group), Munich**.

Work Profile: Data Research Intern in a Data Hunting Team (RID1.6).
(Internship of 6 months - August 1, 2020 to January 31, 2021)
Under the supervision of **Dr. Tobias Hirzinger.**

With this profile in Munich Re, I gained knowledge of fundamentals of Reinsurance concept, and received an opportunity to work on Data hunting, Geospatial Data Visualization and Data Enrichment with Python. This work profile truly sparked my interest and built curiosity towards how interestingly the world's Leading Reinsurance company **Munich Re** works on Data Engineering technologies.

At the same time, I am a master student of Data Engineering in Jacobs University Bremen, Germany. The internship at Munich Re is a part of my studies for **Advanced Project II**. I am writing this report on real time industrial experience gained during Internship under the guidance of **Prof. Dr. Adalbert Wilhelm**. The paper will briefly introduce you about Munich Re, and the details on the tasks performed in the Data Hunting Department of **Munich Re (Group)**, Munich.

TABLE OF CONTENTS

1. Introduction

Munich Re (Group) - A world-leading provider of reinsurance, primary insurance and insurance-related risk solutions.

“With gross premiums written of €33.8 bn from reinsurance alone, Munich Re is one of the world's leading reinsurers and operates in life, health and property-casualty industry.”

As a reinsurer, in the form of exceptional strategic alliances, Munich Re writes commercial undertakings in direct cooperation with prevailing insurers. Munich Re provides a broad variety of tailored products and personalized insurance solutions and services to customers handling industrial and large-scale projects.

With over 4,000 corporate customers in 160+ countries, Munich Re does business.

As a global reinsurer worldwide, Munich Re is equipped with in-depth awareness of the international and local market, immense information inventories and experience in analytics. Munich Re integrates these skills and puts them at your fingertips exactly according to your needs and strategic priorities, in line with their emphasis on long-lasting relationships and value-added for each customer.

The Integrated Analytics group uses scientific records, data engineering, and cutting-edge tools to provide transformative possibilities for risk assessment and management of life insurance plans. In each modelling and deployment, Munich Re’s fast execution and steady innovation set them apart in the industry [1].

Munich Re - Big Data Transformation

- Dedicated Data Intake Team and Data Strategies.
- Top-down system of data alongside topics of creativity and core market domains.
- Systematic cataloging of all data sources.
- Only relevant data is ingested and cleaned by Dedicated community to develop pipelines and products for Data.
- The Data Lake strategy of Munich Re and its network have a powerful infrastructure, expert knowledge and leading-edge tools to assist with ideas and projects from Analytics.

It's always been my great pleasure to be a part of such a large organisation and working with the data hunting team at Munich Re. The aim of this report is described in the following sections about tasks performed with the profile Data Researcher at Munich Re.

2. Data Hunting

In technology, the rising importance of Data to businesses is a recent trend. Data Transformation is the latest Digital Transformation.

The Advanced Data Hunting team in Munich Re (Group):

- Identifies the right new source of Data and ideation of cases for conceivable use. Identification of the proprietor and marketing of a commercial enterprise.
- The Data search, pre-processing and ingestion of data. Cooperation with large data providers, start-ups and large web platforms.
- Search for relevant data to guide normal and new opportunities for growth.

Munich Re developed a Data Lake to be used by scientists for its actuarial data, it observed that new tasks were needed to help everyone get the high quality out of it. Not all data originate from inside the organization. How did the department build a team for Data Hunting? Searching for potential data within and outside of the organization to maximize business outcomes. The best job title is Data Hunter. Munich Re allows practitioners to consider the right algorithm for a particular space of problem. It is half of the problem that getting the best out of the data, and these are the tasks performed by Data hunters.

As a major international reinsurer, Munich Re is equipped with in-depth awareness of the world and nearby markets, broad inventories of data and experience in analytics. Data hunting department deals with variety of data like Building data, Cyber data, Company data etc. [1].

During my internship I worked mostly with building data team.

Building Data:

In all its complexities, the venture for insurers on building projects is also to truly identify and evaluate the risks. Indeed, each and every structure is unique from an engineering perspective. After all, the many common characteristics, such as the state, the substances and the technical specifications, can all vary greatly from structure to structure. This poses challenges for each person involved in a project, including the insurers, over and over. The insurance sector bears the burden of obtaining old infrastructure, especially in the segment of the property insurance plan. Storm, fire and water accidents occur more often and with greater consequences; buildings are collapsing and operations of commercial businesses are being disrupted. Civil liability eventualities may also occur.

With Munich Re’s new level management of hurricane claims damage predictions can be made 4 days advance to the landfall and 4 days after the event to detect damage. This requires the vicinity and characteristics of any building in the portfolio, making this genuinely a solution for handling hurricane claims in the next period. [1][8].

2.1. Requirements of data

Why reinsurance company needs data?

For the administration of their loss recoveries and reinsurance book, reinsurance companies have constantly processed large amount of data as properly as for actuarial modeling, trend and general modeling. Historically, research has been carried out by add-on software, spreadsheets or even on paper. This technique of replication and/or guidance has led to inaccuracies, discrepancies and inconsistencies on a daily basis.

Hierarchy of needs:

- To collect all the necessary data and to be able to bring it together in an acceptable processing environment.
- To make it convenient to analyze and process, this data needs to be modeled in a uniform manner.

Visually the hierarchy looks like:

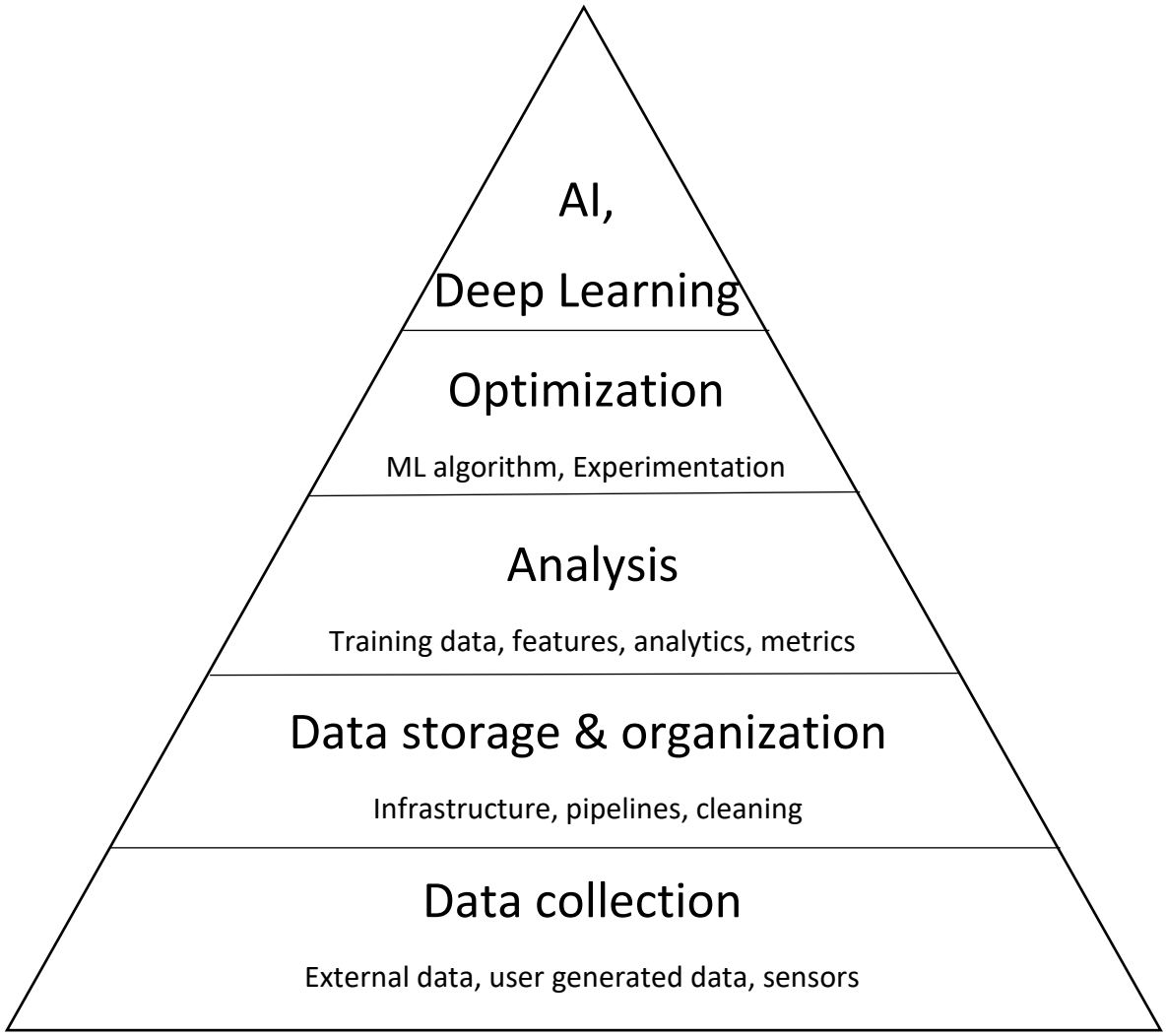

Figure 1: Hierarchy of needs of the data

2.2. Data Gathering

There are several methods to collect large datasets as per requirement. We could buy data from data provider groups, or use a data collection tools to obtain data from website. Techniques for data gathering could be structured as follows:

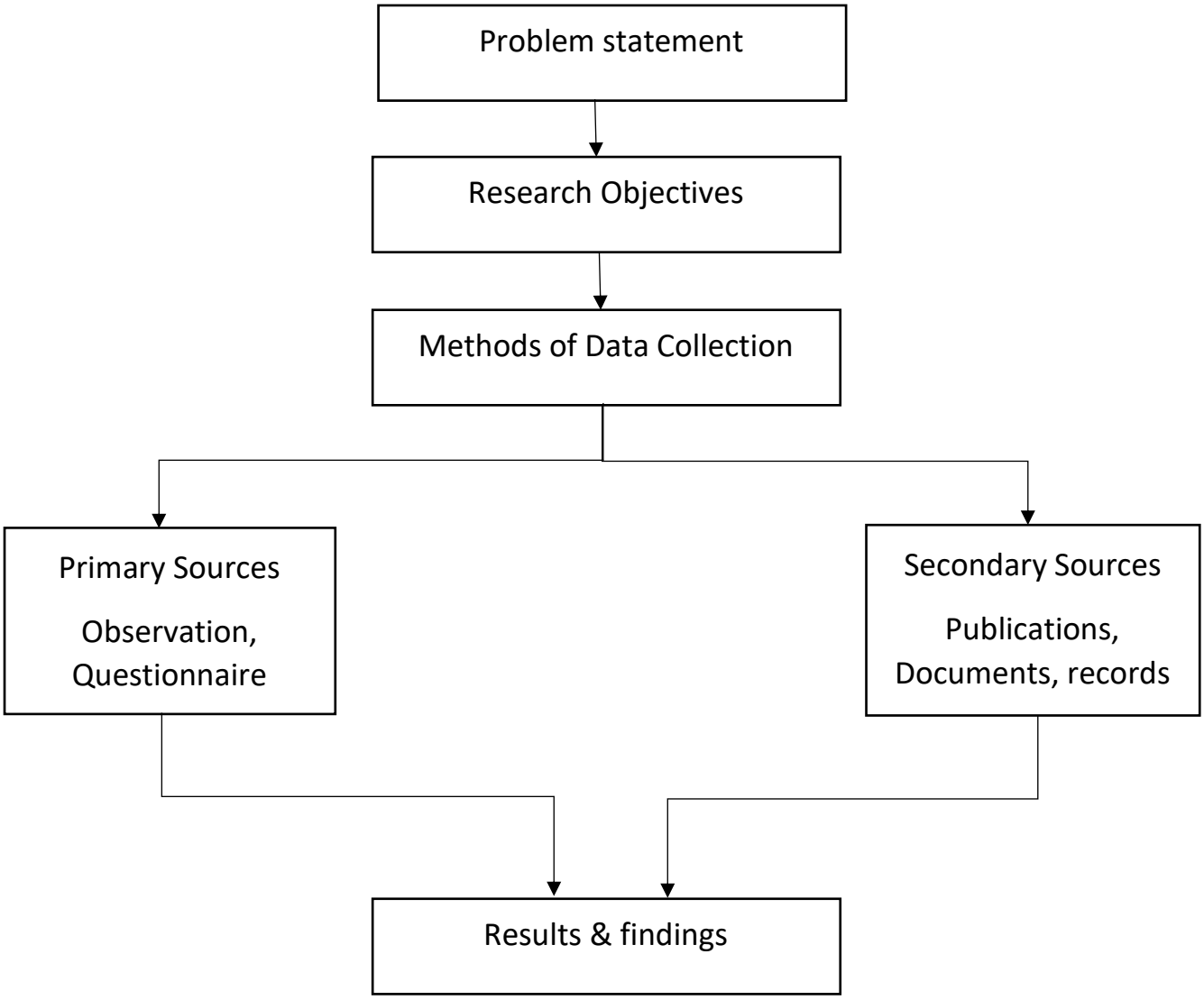

Figure 2: Data collection techniques

Following are some methods we used for data collection:

2.2.1. Databases

If the necessary data sets are available in the databases of our organization, then we can use SQL queries to obtain the data we want from them without difficulty.

2.2.2. Sites

Many sites are out there available where, data sets are stored in splendid formats to be downloaded for practice/competitions by individuals. Some government and private paid providers of the large data sets are also reachable from internet search.

2.2.3. API or Web-scraping

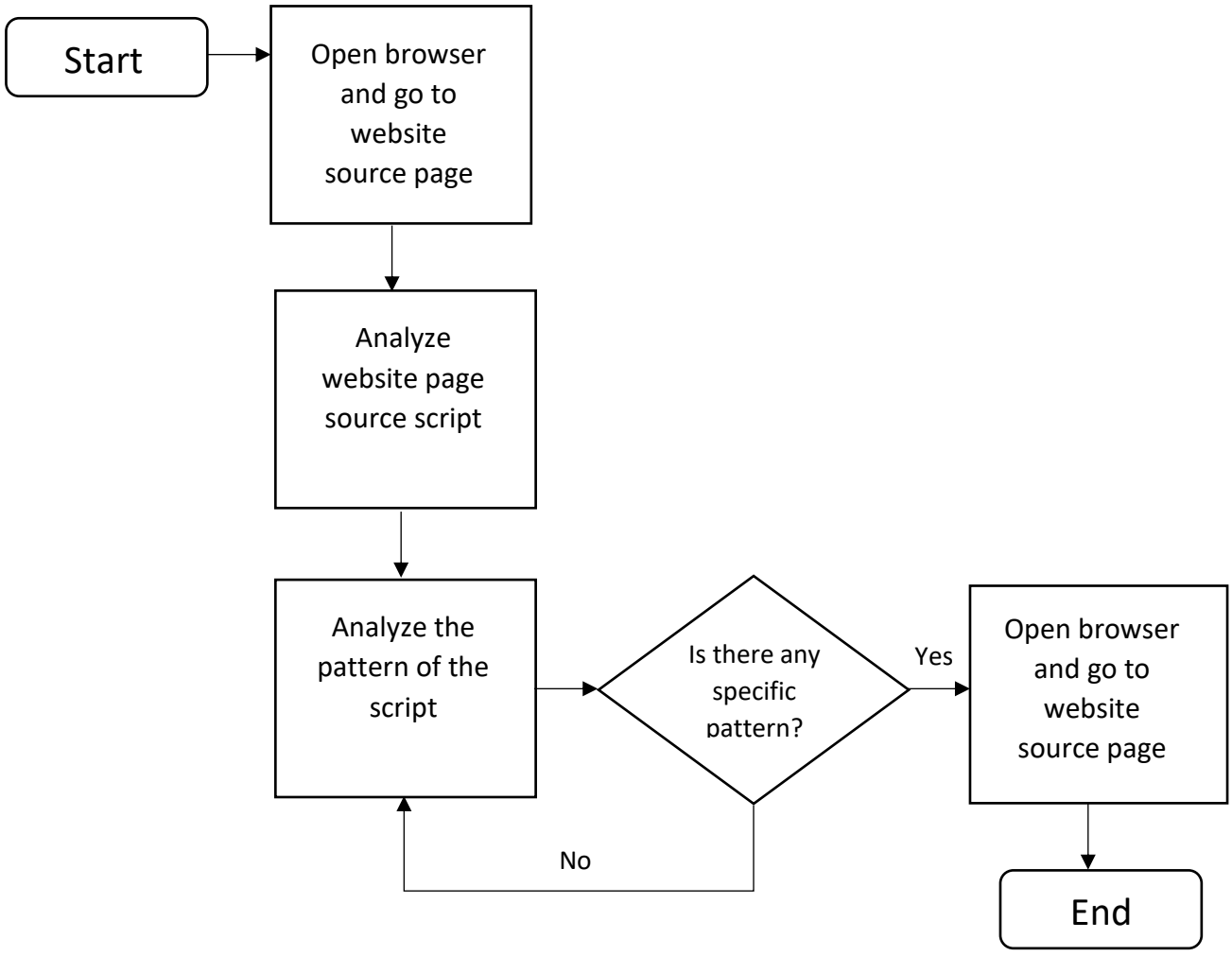

Figure 3: Web scraping algorithm

If availability of the required data is there on websites, then we can collect and store information in our local storage/databases using the website API (if available) or Web Scraping strategies. Statistics obtained from the Internet are often saved in a JSON file, and further processing is required to convert JSON to commonly used format ".csv" format [2].

Web scraping is the way to get a massive amount of public data from websites. It automates data collection and transforms scraped data to formats such as HTML, txt, Excel, CSV and JSON [2].

We used Python libraries for web scraping which primarily consists following parts:

- **Parse a - HTML website:**
  Installing the appropriate Python libraries-
  Beautifulsoup - helps to scrape web pages to extract data and provide Python expressions to change the parse tree, iterate and scan.

  ```
  pip install BeautifulSoup4
  ```

- **Extract the requisite data/information:**
  Accessing the website's HTML content. the soup object holds all the data in structure of nested tree. We could extract this data easily.

- **Store the data:**
  Finally, in required file format like CSV, we'd save all our results [2].

3. Data Enrichment & Evaluation

Munich Re's pilot project is a principle that has tremendous loss reduction potential. The IoT is growing. Sensors are being used in more and more places to convey calculated information to IoT systems while taking into account IT security. The records are analyzed in real-time on these channels, primarily to pick out what action is needed without delay and ensure timely response [1].

Requirement of data enrichment?
We wanted to enrich building database with the large data (millions of records) from data lake for pilot project. There is already existing company data enrichment tool but, with the tool it was a time-consuming task to enrich large data. To automate this procedure as data enrichment from Data Lake we decided to develop python script which will improve performance and quality of enriched data [1].

The purpose of data enrichment is to borrow company data like parent company details about the company from Data Lake to make our building database more robust for pilot project of Munich Re, and increase its usefulness in business approach.

Data enrichment determines a process that involves cleaning and incorporating existing data sets, as well as integrating it with external data from various sources. Anything that refines and strengthens the quality of data by filling in gaps and fixing wrong data can be called as data enrichment [3].

- Data cleaning refers to the elimination of corrupt, outdated and inaccurate entries from data sets.
- Appending data requires updating missing or outdated archives with accurate, up-to-date data.

Steps followed to enrich data (Company data) from Data Lake:

1. **Defining our need:**

   Extraneous data may bring distractions and uncertainty, so enriching current data with additional data that are not necessary for our purposes is counterproductive. Making sure to clear on what we need and what we are going to use it for before enriching data

2. **Evaluate our datasets:**

   Before embarking the enrichment process, we checked the accuracy, consistency and completeness of our current data. Searched for the gaps in our files so that we don't spend time digging through outdated, unnecessary data.

3. **Restrictions/conditions to enrich data:**

   Segmentation must be constantly served by enrichment, and vice versa. The more our data units can be avoided and "tightened" to concentrate on unique target segments, the more response data we receive on our data enrichment activities.

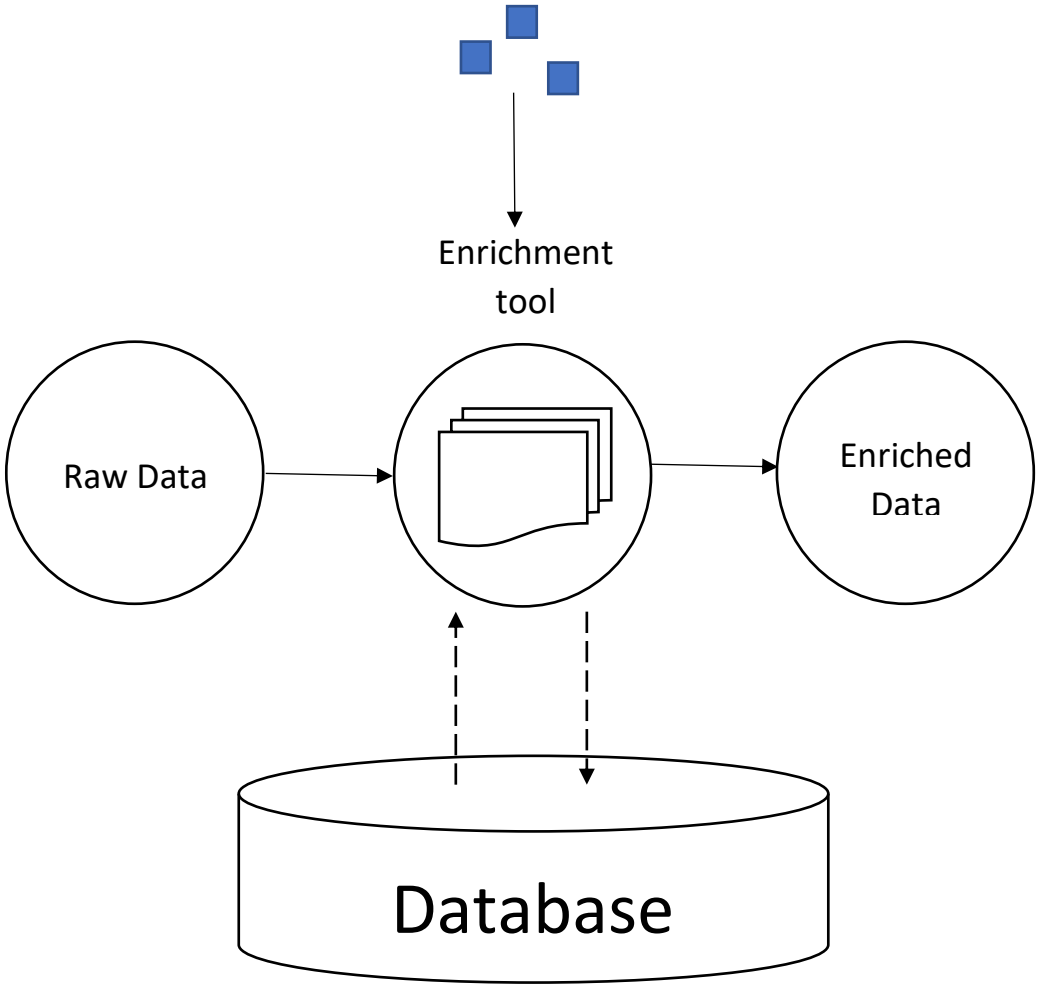

Figure 4: Data Enrichment workflow

Data enrichment is the way of enhancing current data through the supplementation of missing or incomplete data [3].

Our scenario to develop data enrichment algorithm:

- Customer data, for example, can be represented and saved with distinct formatting and degree of completeness. It is now not rare to see that one system stores a specific company's name and address, while for some reason the other system holds company's parent company details.

- The logic behind the idea is therefore, enriching data can be achieved by combining two data sources. In addition, if the one source data appear to have some embedded intelligence, then we can enrich a record by using the information from the data themselves.

- For better understanding of the concept lets go, dive in:
  We have data of company records that we want to enrich. The data is portrayed as follows:

ID	Name	State	Parent company name	Address
1234567	LLOYD BERNIECE	NC	-	-

Table 1: Initial dummy data from building database

- The preliminary data consists only of the name, ID and state. It would probably be tempting to look at other sources of data to see if we can find additional attributes to complete our golden data. But assess first whether or not there is any embedded intelligence in our current data.

- Suppose we now have another source of data supply that grants data as observed:

Name	Address	State	Parent company name
Lloyd B.	108 D	NC	Hapag-Lloyd

Table 2: Dummy data from other source data from Data Lake

- All looks precise until one statement is put forward, whether this 'Lloyd B.' With our 'LLOYD is the same or not?

  Due to the fact that data is so scattered in various structures, each and every structure can be owned by excellent departments. There is also the risk of typos and other human errors that make it worse to be counted.

This is where Fuzzy-matching API plays the role, this API is developed by Data Engineering team of Munich Re. Basically, the API compares all these data sets to determine whether they represent the same entity. First, all attributes of records are predominantly evaluated on the basis of their similarity and then given a final rating to decide if the data is the same. The higher the ranking, the greater the certainty that the records represent the same person.

- On completion of our scenario, after requesting for data from the Fuzzy-matching API, we are now confident that these 2 names are the same company. And now,

ID	Name	State	Parent company name	Address
1234567	LLOYD BERNIECE	NC	Hapag-Lloyd	108 D

Table 3: Enriched dummy data stored in new table in building database.

And here, we finished with the data enrichment!

Note: Extra effective information is stored in the enriched data and, as a result, it is riskier to be violated. The data security should always be maintained at the very top.

3.1. API Requests with Python

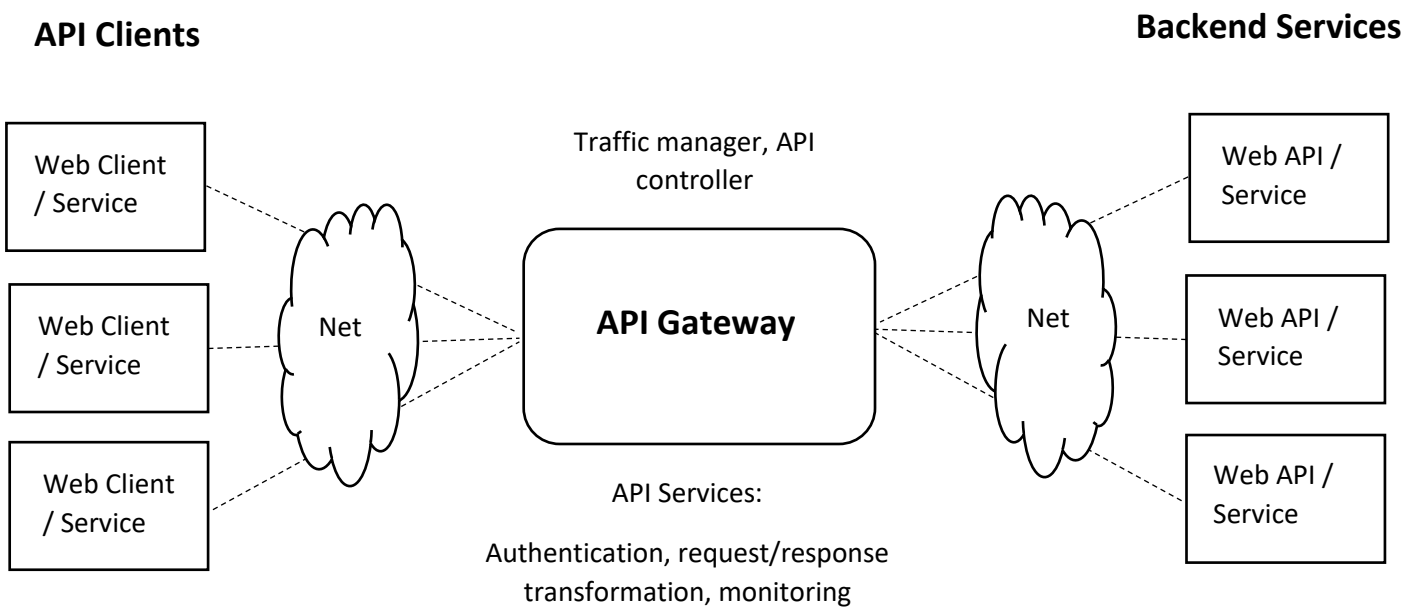

Figure 5: API Gateway services

Steps followed while implementing data enrichment with Python:

- Installation of Python libraries – psycopg2, datetime, csv, requests, json.
- Database connection, query data from building database to get matching from company Data
- Request a token for authorization for Fuzzy Matching API server - needs to be refreshed every 1 hour.
- API configuration to enrich data from data lake
- Make batches of 100 records to request – iterating over batches
- API call – request and response
- Format response data in dictionary or list
- Create new table in building database to store enriched data back.
- Insert enriched data into newly created table in building database

3.2. Data Evaluation

Quality and performance check:

Problems occurred while enriching data from Fuzzy-matching API:

- API server Performance
- Quality of the data
- Entity match rate

Performance improvement techniques:

- Sending data records in batches of 100 for single request which eventually reduces API traffic and gives response faster
- Difference between sequential and parallel requesting to API

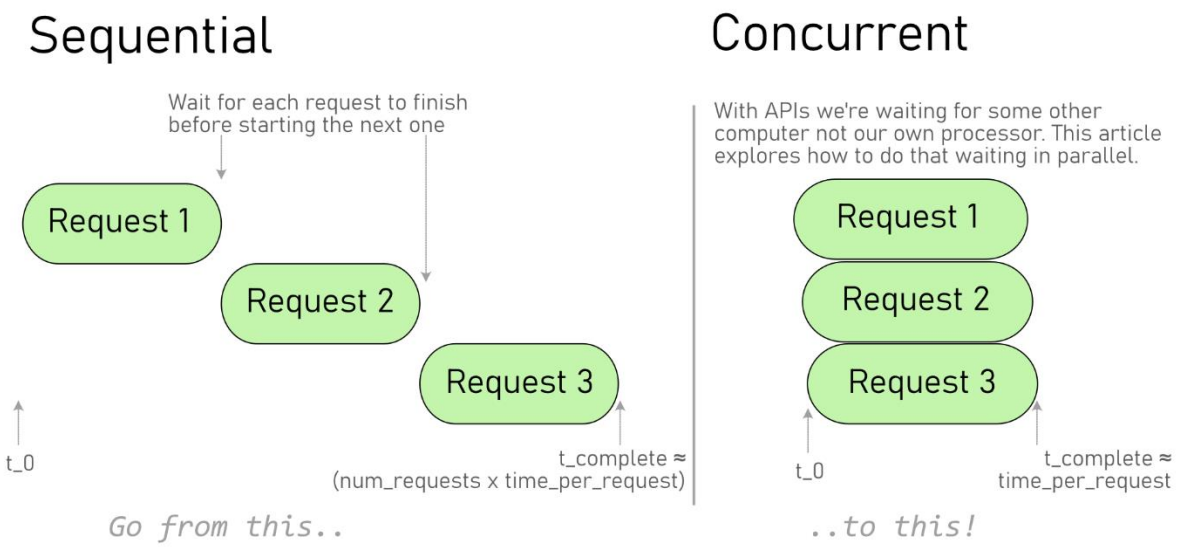

Figure 6: Sequential and parallel API requests [6].

Initially to get idea of the quality of data, performance and match rate from the API response, we performed manual evaluation on small set of data in excel sheets.

Data evaluation additionally encompass the following tasks:

- Comparing enriched data with raw data - to conclude on matching rate of the API server
- Identifying tremendous data gaps (if any) - data gaps gives degree of uncertainty; however, project team may decide if the degree of uncertainty associated with the data gap is right or whether extra data gathering is required.
- Performing statistical evaluations
- Developing visual representations of data as some patterns can be identified easily by just visualization.
- Grouping records - as in our case there is a group of 3 ranks in enriched data which differs based on some other fields.
- Summarizing results in pivot

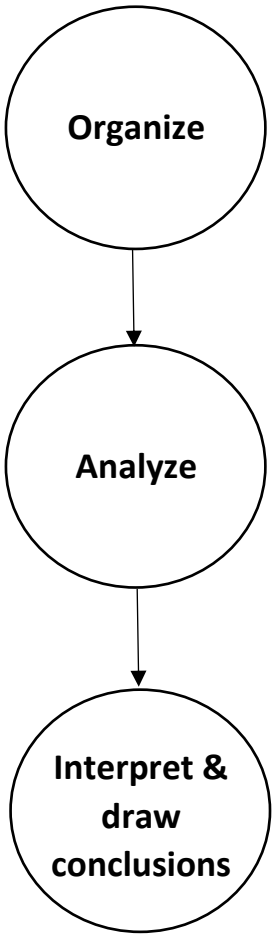

Figure 7: Strategy followed for data evaluation

4. Data Visualization

4.1. GIS (Geographic Information System)

Munich Re is involved in variety of activities for Geo Risks Research. Innovation in combination with particular specialist knowledge, make certain progress and development in the markets.

For several years, satellite data has been used by **Munich Re** to determine damage. New satellite and assessment techniques provide timely, loss estimates of high-quality. Being able to model and depict earthquake events in a three-dimensional manner opens up new possibilities to gain a deeper understanding of the intense forces in major earthquakes [1].

Munich Re performs analytics on satellite imagery and visualize the results of the analytics for better and quick understanding of risks from natural disasters. Following image shows how Munich Re's new digital technologies measure wildfire risks from the regions of North America to automate the recording of hurricane damage [7].

Figure 8: Visualization of Munich Re’s evaluation for wildfire risks [7].

Remote Industries Munich Re takes hurricane claims management to new level. Help you improve damage and risk assessment, speed up response time, and deliver high-quality viable customer experience by using remote sensing AI, machine learning and aerial imagery [8].

4.2. ArcGIS Pro

ArcGIS Pro is an application that is ribbon-based, which provided accessibility to many instructions from the ribbon at the top of the window; more specialized and advanced functionalities can be opened as necessary on panes (dockable windows).

Geoprocessing tools perform essential role for spatial analysis, they also provide other applications. In order to construct an output dataset, most geoprocessing devices operate on an input dataset. Some software adjust the characteristics or geometry of an input dataset. A few devices have other results, such as producing layer choices or creating messages or reports [4].

Learned ArcGIS activities which includes spatial analysis:

- comparing places
- determining how places are related
- finding best locations
- detecting patterns

In certain cases, natural threats such as floods, windstorms and earthquakes end up contributing to major claims. Where do such events appear, and how to calculate the probability that they will occur? Insurers support these tasks with the help of geoprocessing tools.

For rental insurance of regions from US we analyzed US residential data. To give instant results to client with visualization we performed visualization of 5-digit postcode areas from US [1].

Analysis is performed according to insurer’s values. This shows 5-digit postcode areas from US states with different variations such as:

- Total of average tax assessed value
- Average tax assessed value improvements
- Average gross area in Sqf
- Total of average tax market value
- Average tax market value improvements

Following screenshot shows example of the visualization performed for above task.

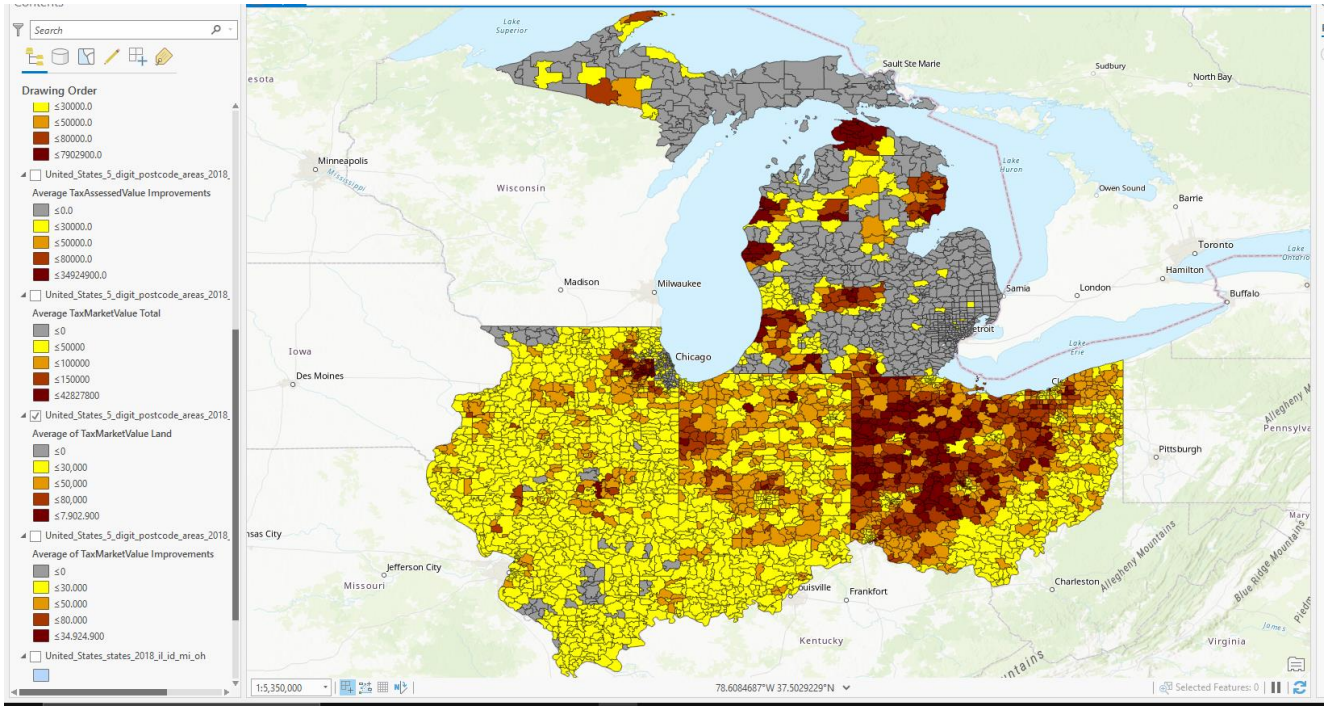

Figure 9: AcrGIS Pro visualization for 5-digit postcode areas from US

5. Conclusions and Future Scope

Working in a Data Hunting team is kind of making my foundation firm on Data such as how data is hunted and pre-processed. It's a great exposure to develop curiosity and excitement towards leading technologies of Data Engineering.

Technologies learned during internship:

- Azure DevOps, Databricks, ArcGIS Pro, Postgres, Python expertise.
- Data hunting – data hunters look for external sources of data that can be combined with internal information to generate new insights and optimize processes.
- Types of data worked on - Geospatial data, Building data, Company data, Farm data, Machinery Data.
- Web Scraping - the manner of downloading data from websites and extracting valuable data from that data.
- Data Enrichment – API request with Python
- Data Evaluation techniques
- Data Visualization – ArcGIS Pro

Future scope

Based on knowledge gained during internship and with the cooperation of team members in data hunting, I received opportunity to work with Data Engineering team of Munich Re (Group), Munich for master's Thesis on following topic:

Thesis topic: A generic learning-to-rank model for matching entities with optional entity attributes.

Motivation: Our data science group is looking to leverage the experience of fuzzy matching of geographical and legal-entities, in order to empower a framework to learn similarities between generic entity types, with variable number of attributes, and offer matching-as-a-service to internal business use-cases.

6. References

[1] https://www.munichre.com/en.html

[2] https://pypi.org/project/beautifulsoup4/

[3] https://towardsdatascience.com/understanding-data-enrichment-with-simple-example-12ae97bb8f04

[4] https://pro.arcgis.com/en/pro-app/latest/get-started/get-started.htm#:~:text=ArcGIS%20Pro%20is%20the%20latest,Online%20or%20ArcGIS%20Enterprise%20portal.

[5] https://stackoverflow.com/

[6] https://medium.com/@dmort.ca/part-5-api-request-timing-comparison-sequential-multiprocessing-threading-and-async-c4f699552ab3

[7] https://news.europawire.eu/press-releases-tagged-with/munich-re-remote-industries/

[8] https://www.munichre.com/content/dam/munichre/mram/content-pieces/pdfs/RemoteIndustries_FS_09102020.pdf/_jcr_content/renditions/original./RemoteIndustries_FS_09102020.pdf