REFLECTION

The project from task one which primarily involves "Reading and extracting given data from various formats, wrangling the data to solve inconsistencies present, bringing the data together into a singular format and finally mapping the unified records to a relational database using Pony ORM Entities", has provided me with the opportunity to improve my data parsing, wrangling and database skills as I had to learn and utilize new programming concepts.

I started out the assignment by obtaining the needed data which came in four different data formats (.CSV, JSON, XML and .TXT) with each data format carrying separate information. Each of these data formats have their own uniqueness and peculiarities:

- The .CSV (Comma-Separated Values) file is a delimited text file that uses a comma to separate values with each line of the file representing a data record (row) and each record consisting of one or more fields (Columns) that are also separated by commas. In summary, the CSV file typically stores tabular data (numbers and text) in plain text (Wikipedia, 2021a). The CSV data given basically contains vehicle information of each customer with 1000 rows/records (each record corresponds to one customer) and 8 columns.

- The .XML (Extensible Markup Language) file is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. The design goals of XML emphasize simplicity, generality, and usability across the Internet. Although the design of XML focuses on documents, the language is widely used for the representation of arbitrary data structures such as those used in web services (Wikipedia, 2021b). The XML data provided for the project gives job information of about 1000 customers.

- The. JSON (JavaScript Object Notation) file is an open standard file format and data interchange format that uses human-readable text to store and transmit data objects consisting of attribute–value pairs and arrays (or other serializable values) (Wikipedia, 2022a). The provided json data for the project gives credit card information of the customers.

- The TXT file is a plain text document that contains a mix of messages sent to/from the company, Internal messages between colleagues, E-mail message sent to a customer and Confidential internal memo.

Due to the differences and peculiarities that exists between each of the given data-sets as discussed above, parsing the data-sets into a unified record therefore becomes much more challenging, hence the very next step taken in the course of the assignment after obtaining and loading the data into a Python notebook was to convert/extract all the data into same data format to ease the unification process. All the respective data formats except the TXT file were converted/extracted into a Pandas Data frame while the information presented in the text file was manually read and extracted. The pandas Data Frame is a two-dimensional size-mutable tabular data structure with labeled axes (rows and columns) and also a feature of the pandas' library, a software tool for data manipulation and analysis (Python | Pandas DataFrame 2019, Wikipedia, 2020).

The conversion of all data formats to a singular form (Pandas Data-Frame) thus allows for easy data wrangling, cleaning, and unification hence greatly easing the challenge earlier presented in parsing four different datasets in different data formats. The combination of the four data-frames into a unified data frame was done using the pandas merge function, another feature of the pandas' tool which merge Data Frames or named Series objects with a database-style join (pandas.DataFrame.merge).

The Data-frames were merged on a newly created column FULL NAME (which contains the combination of the first name and last name) serving as the primary key. The next and final process in the Assignment after creating the unified record entails creating a relational database using pony ORM and mapping the unified record into the created database.

Another particular challenge aside those discussed above was from the text data. Unlike the other data formats, the text file cannot be parsed to a data frame format, information present in the text file can only be extracted manually by reading, thus posing a huge challenge to the possibility of automating the entire project work-flow.

RELATIONAL DATABASE

The last aspect of Task 1 entails utilizing Pony ORM (a Python library that aids the convenient use and interaction with objects that are stored as rows in a relational database) to store the combination of customer data into MySQL, a relational database management system.

The use of Pony ORM as a choice of object-relational mapper offers several advantages such as an exceptional convenient syntax for writing queries, automatic query optimization, an elegant solution for the N+1 problem and an online database schema editor (What is Pony ORM, 2017). While the use of MySQL as a choice of relational database management system also offers several advantages such as Round-the-clock Uptime, Comprehensive Transactional Support, Complete Workflow Control, Free installation, Simple syntax, mild complexity, and Cloud compatibility, MySQL however might not have been the best choice of Database management system for a small and medium-sized business looking for expansion due to the limitation of Scalability. It also has limited compliance with SQL standard and is only partially open-sourced (Comparing Database Management Systems, 2021). Other alternatives to MySQL as a choice of Database Management System are:

- SQLite, a C-language library (with bindings to other programming languages e.g., python) embedded into end programs in contrast to many other database management systems which are client–server database engine (Wikipedia, 2022b). It is the most used database engine in the world (Wikipedia, 2022b).

- PostgreSQL: PostgreSQL is a powerful, open-source object-relational database system with more than 30 years of active development that has earned it a strong reputation for reliability, feature robustness, and performance (PostgreSQL, 2021). It features Transactions, Atomicity, Consistency, Isolation, Durability (A.C.I.D.) properties, automatically updatable views, materialized views, triggers, foreign keys, and stored procedures (Wikipedia, 2021c). PostgreSQL runs on all major operating systems, including Linux, UNIX, Mac OS and Windows.

- Oracle: Oracle is a relational database management system mainly used for running online transaction processing (OLTP), data warehousing, or both. It is the first and one of the most popular relational database management systems. Oracle Database is cross-platform i.e. It can run on various hardware across operating systems including and also ACID (Atomicity, Consistency, Isolation, Durability) compliant.

Other alternative ORMs to Pony also exists, one of which is SQLAlchemy; a Python SQL toolkit and Object Relational Mapper that gives application developers the full power and flexibility of SQL. SQLAlchemy's philosophy is based on the idea that relational databases behave less like object collections as size increases and object collections like-wise behave less like tables with more abstractions. SQLAlchemy aims to solve this problem by adopting the data mapper pattern rather than the active record pattern used by a number of other object-relational mappers thereby allowing the library to takes on the job of automating redundant tasks while the developer remains in control of how the database is organized and how SQL is constructed (SQLAlchemy, 2021, Wikipedia, 2022c). Peewee, a simple and small ORM with few concepts is also another alternative. Peewee is easy to learn and intuitive to use, it supports SQLite, MySQL, PostgreSQL and cockroach db.

The Python program also has the ability to connect with Databases directly without the use of any ORM as intermediary. The general approach to doing this for all DBMS involves first creating a connection object with the database in the DBMS of choice and thereafter interacting with the databases by creating a cursor object and using the Cursor. Execute ("SQL query goes into this bracket ") command.

To make recommendations based on the above explanations, I'd say the best DBMS alternative for the company being a SMB is PostgreSQL because it is fully open sourced and hence free. It is also good for scalability; therefore, the business will be able to use it for a long time before outgrowing it. It is also recommended that the Python/SQL syntax be used for interacting with the DBMS directly rather than using an intermediary ORM because it could be difficult to debug ORM's and ORM's has higher learning curves.

BIG DATA ISSUES

Big data refers to extremely large volume of data (combination of structured, semi structured and unstructured) whose size ranges in exabytes and more, yet growing exponentially with time. Such data cannot be stored or processed using the traditional methods. According to Doug Laney, Big data is marked by three main characteristics viz: high volume, high velocity and high variety.

As the company expands, scaling up is done by establishing new branches in various countries and regions of the world. The volume of data the company accrues will no doubt increase rapidly and with time level up to that of big data as defined above. When this occurs, there are three main big data challenges the company will then have to address technologically viz: The Challenge of Big Data Transport/Storage, The Challenge of Big Data Processing and Management Issues.

One of the main characteristics of big data according to the 3V's of big data by Doug Laney is *High Volume* which could range up to exabytes. Storage hence becomes a challenge as current disk technology limits are about 4 terabytes per disk which means 1 exabyte would require 250,000 disks, an enormous amount of disk impossible to be attached to one computer system based on current technology (Kaisler et. Al 2013). Transport also becomes another challenge as it would take a huge period of time to successfully transfer such volume of data from the collection or storage point to the processing point. To solve this problem, it is advised that the company subscribe to the service of more suitable database management systems like Oracle and employ more big data tools e.g., Hadoop.

Aside the challenge of storage and transport of the big data accrued, another possible challenge the company will experience is the challenge of processing the accrued data. Assume an exabyte of data is to be processed and the data is chunked into blocks of 8 words, so 1 exabyte = 1K petabytes. If a processor expends 100 instructions on one block at 5 gigahertz, the time required for end-to-end processing would be 20 nanoseconds. To process 1K petabytes would require a total end-to-end processing time of roughly 635 years. Thus, effective processing of exabytes of data will require extensive parallel processing and new analytics algorithms in order to provide timely and actionable information (Kaisler et. al 2013). A popular solution proposed to this problem is "bringing the code to the data", unlike the traditional method of "bringing the data to the code" (Bajaj et. al, 2014).

The last challenge of management might perhaps be the most difficult one to solve. Data in the company will come from various sources (financial reports, e-mails, customer logs) and from various regions/countries as the company expands. Combining all this data to a unified record may be a challenging task.

Legally, there are also challenges the company will encounter as they accrue more data from customers and even internally. These challenges include: Data ownership and security, Consumer privacy, Third-party contracts, Regulatory compliance, Underlying contracts, Legal discovery (Henriquez, 2021). However, the most pressing amidst these challenges for the company in the scenario is that of Data Security. Data Security simply refers to the process of protecting data from unauthorized access and data corruption. An expanding company as painted in the scenario above is saddled with the responsibility of protecting the data of customers entrusted to them. Failure in this will lead to customers/employee data falling into the hands of a third party where it can be used for fraudulent activities such as phishing scams and identity thefts. One good solution to this problem is to employ the use of cloud security services.

REFERENCES

1. Wikipedia. (2021). *Comma-separated values*. [online] Available at: https://en.wikipedia.org/w/index.php?title=Comma-separated_values&oldid=1060298461 [Accessed 10 Jan. 2022].

2. Wikipedia. (2021). *XML*. [online] Available at: https://en.wikipedia.org/w/index.php?title=XML&oldid=1062340378 [Accessed 10 Jan. 2022].

3. Wikipedia. (2022). *JSON*. [online] Available at: https://en.wikipedia.org/w/index.php?title=JSON&oldid=1063500160 [Accessed 10 Jan. 2022].

4. GeeksforGeeks. (2019), Python | Pandas DataFrame, https://www.geeksforgeeks.org/python-pandas-dataframe/

5. Wikipedia. (2020). *pandas (software)*. [online] Available at: https://en.wikipedia.org/wiki/Pandas_(software). [Accessed 10 Jan. 2022].

6. pandas.DataFrame.merge, https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.merge.html

7. What is Pony ORM? (2017) https://ponyorm.readthedocs.io/en/latest/

8. Comparing Database Management Systems: MySQL, PostgreSQL, MSSQL Server, MongoDB, Elasticsearch, and others, (2021) https://www.altexsoft.com/blog/business/comparing-database-management-systems-mysql-postgresql-mssql-server-mongodb-elasticsearch-and-others/

9. Wikipedia. (2022). *SQLite*. [online] Available at: https://en.wikipedia.org/w/index.php?title=SQLite&oldid=1063639993 [Accessed 10 Jan. 2022].

10. PostgreSQL: The World's Most Advanced Open-Source Relational Database, 2021 https://www.postgresql.org/

11. Wikipedia. (2021). *PostgreSQL*. [online] Available at: https://en.wikipedia.org/w/index.php?title=PostgreSQL&oldid=1061904134 [Accessed 10 Jan. 2022].

12. SQLAlchemy, 2021, https://www.sqlalchemy.org/

13. Wikipedia. (2022). *SQLAlchemy*. [online] Available at: https://en.wikipedia.org/w/index.php?title=SQLAlchemy&oldid=1063436001 [Accessed 10 Jan. 2022].

14. Kaisler, Stephen & Armour, Frank & Espinosa, J. & Money, William. (2013). Big Data: Issues and Challenges Moving Forward. Proceedings of the Annual Hawaii International Conference on System Sciences. 995-1004. 10.1109/HICSS.2013.645.

15. R. H. Bajaj and P. Ramteke, "Big data–the new era of data," International Journal of Computer Science and Information Technologies, vol. 5, pp. 1875-1885, 2014.

16. Jaime Henriquez, Big data: Six critical areas of legal risk. https://www.techrepublic.com/article/big-data-six-critical-areas-of-legal-risk/