Reflection Paper on Project Experience   By: Brian and Amani

Throughout this project, our team worked on several technical aspects, from data selection and cleaning to setting up an ETL pipeline, performing data analysis, and managing cloud storage. The project presented numerous challenges that required both technical expertise and effective teamwork. As we reflect on these experiences, we recognize the valuable lessons learned and the skills gained, as well as the areas for further development.

The data selection process was more challenging than anticipated, requiring thorough exploration across platforms like Kaggle and Google Dataset Search to find movie datasets focusing on revenue. Many datasets we encountered were either incomplete, poorly documented, or contained fake and unusable entries, such as fabricated movies with implausible box office revenues. Additionally, several datasets were formatted incorrectly or couldn't be opened, further narrowing our options. We prioritized finding reliable and well-structured data, ensuring it could provide insights into revenue-related factors. This process required patience and careful evaluation to avoid datasets that would compromise the integrity of our analysis. Ultimately, we selected the TMDB dataset for its emphasis on real-world data and comprehensive information, which we believed would allow us to uncover meaningful insights while maintaining analytical rigor.

One of the primary challenges faced during the project was the selection and cleaning of data. We opted to use MySQL instead of MongoDB because of structure purposes, as MySQL's relational format better suited the project's needs. The dataset, particularly the TMDB (The Movie Database) dataset, was large and filled with null values, which made it difficult to parse through and clean. The presence of over 3 million missing values in the dataset created significant roadblocks, as it took considerable time to identify and address them. Additionally, the data was not consistently formatted, with some columns containing irrelevant or out-of-context entries, which added another layer of complexity. We employed various techniques to manage these issues, such as imputing missing values, removing unusable data points, and applying more advanced methods like predictive imputation and grouping for similar entries. However, this was still a labor-intensive process that required constant iteration. The sheer size of the dataset further compounded these challenges, requiring us to implement more efficient methods for handling and processing the data, such as parallel processing and batch loading. This experience underscored the importance of thorough data cleaning and the need for strategies to deal with large, messy datasets effectively.

In addition to the challenges of data cleaning, the ETL process presented its own difficulties. One of the major obstacles was ensuring reproducibility within the pipeline. Due to the complexity of the datasets and the unique structure of each one, we found ourselves hardcoding various aspects of the ETL setup, which made it challenging to replicate the process in different environments or for different team members. This lack of flexibility highlighted the need for a more scalable and modular approach in the future. We also encountered issues with data schema mismatches between the source data and the target MySQL database, which required extra work to ensure that data was properly mapped and transformed. Ensuring that the ETL process is easily reproducible and maintainable would save time and reduce errors, making the project more efficient in the long run.

Another challenge we faced was integrating cloud storage into our workflow. Uploading data to the cloud proved more difficult than anticipated, as we lacked a clear understanding of how the Software Development Kit (SDK) integrated with our local environment. Setting up the connection between our local systems and the cloud was confusing, and we struggled to grasp how to properly configure the SDK from the command prompt. Moreover, managing large datasets in the cloud introduced additional challenges related to bandwidth, storage limits, and data retrieval speeds, which further slowed down our workflow. While we eventually managed to establish a connection, the process was not as straightforward as we had hoped, and we realized that we needed a deeper understanding of cloud storage solutions and their integration processes. This experience emphasized the importance of thoroughly understanding cloud tools and SDKs before embarking on a project that requires their use.

Despite these challenges, several lessons were learned throughout the course of the project. On a technical level, we learned the importance of clean and well-structured data. The data cleaning process was more time-consuming than expected, and we quickly realized that without a solid foundation in data preparation, the analysis and subsequent steps would be unreliable. We also gained experience in optimizing ETL pipelines, focusing on how to balance the speed of data extraction with data accuracy and consistency. Additionally, the need for a reproducible and scalable ETL pipeline became evident as we faced difficulties in adapting the process for different datasets and team members. From a teamwork perspective, the project also taught us the value of regular communication and coordination. Given the complexity of the tasks involved, it was easy for team members to become siloed in their work. By holding regular check-ins and sharing progress updates, we were able to ensure that everyone was aligned on the project's goals and the steps needed to achieve them. This was an important takeaway for future projects, as collaboration and knowledge-sharing within the team were crucial to overcoming obstacles.

Looking ahead, there are several areas for improvement. One key takeaway is the need for a more modular approach to the ETL setup. By breaking down the pipeline into smaller, reusable components, we could make the process more flexible and less reliant on hardcoded values. Additionally, gaining more hands-on experience with cloud storage solutions and SDKs would help to eliminate the confusion we encountered during this project. It would also be beneficial to dedicate more time to understanding the specific tools we use, ensuring that we can integrate them seamlessly into our workflow. Another area for improvement is in data visualization. While we were able to analyze the data, presenting the results in a more visually compelling and accessible way could significantly enhance the impact of our findings, particularly when sharing the results with stakeholders or non-technical audiences.

Throughout this project, we gained several valuable skills. We learned how to clean and preprocess large datasets, which is an essential skill in any data-driven project. Working with the ETL pipeline also improved our coding and automation skills, particularly in Python. Moreover, we gained hands-on experience in cloud storage and learned about the complexities of integrating cloud platforms into our workflow. These technical skills, along with our increased understanding of data wrangling and cloud storage, will serve us well in future projects. However, there is always room for growth. Developing a deeper understanding of cloud SDKs, mastering more advanced data visualization techniques, and further refining our ETL workflows are all areas where we can continue to improve.

In conclusion, this project provided both challenges and valuable learning experiences. We faced difficulties in data cleaning, setting up the ETL pipeline, and integrating cloud storage, but these challenges offered important lessons in data management, technical problem-solving, and teamwork. As we reflect on our experiences, we recognize the skills we have gained and the areas where we can continue to develop. Moving forward, we will be better equipped to handle complex datasets, automate workflows, and work more effectively with cloud-based tools, setting us up for success in future projects.