Industrial Work

Report

On

Weekly Project Work

Developed By: -	Guided By:-
Krupal Patel (20162121007)	Prof. Aniket Patel (Internal)
	Mr. Jignesh Patel (External)

Submitted to
Faculty of Engineering and Technology
Institute of Computer Technology
Ganpat University

Ganpat University
॥ विद्यया समाजोत्कर्षः ॥

Institute of Computer Technology

National Assessment and Accreditation Council
NAAC
A
Grade
Accredited University

Year – 2024

ACKNOWLEDGEMENT

IBM/Industry Internship is a golden opportunity for learning and self-development. I consider myself very lucky and honored to have so many wonderful people lead me through in completion of this project. First and foremost, I would like to thank Dr. Rohit Patel, Principal, ICT, and Prof. Dharmesh Darji , Head, ICT who gave us an opportunity to undertake this project. My grateful thanks to Prof. Aniket Patel & Mr. Jignesh Patel (Internal & External Guides) for their guidance in project work Predicting Application Rating of Google Play Store, who despite being extraordinarily busy with academics, took time out to hear, guide and keep us on the correct path. We do not know where would have been without his/her help. CSE department monitored our progress and arranged all facilities to make life easier. We choose this moment to acknowledge their contribution gratefully.

Krupal Patel (20162121007)

TABLE OF CONTENT

Introduction to Job Role

The role of a Cloud Data Engineer is pivotal in architecting, building, and maintaining data solutions on cloud platforms such as AWS, Azure, or Google Cloud. Responsible for designing scalable data architectures, developing efficient ETL processes, and optimizing performance, they collaborate closely with cross-functional teams to ensure data integrity, security, and compliance. By staying updated with emerging technologies and leveraging advanced cloud services, Cloud Data Engineers empower organizations to harness the full potential of their data assets, driving insights, innovation, and competitive advantage in today's data-driven landscape.

Week I

Learning relevant tools and technology

This week, I'll dive into mastering the essential tools and technologies crucial for our cloud data engineering journey. Starting with AWS, I'll explore its key services and functionalities, gaining a solid understanding of cloud infrastructure management. Additionally, I'll delve into Terraform, learning how to provision and manage cloud resources efficiently through infrastructure as code. To streamline our development workflow, I'll also familiarize myself with GitHub Actions, enabling seamless automation and integration within our projects. This week sets the foundation for our cloud data engineering endeavors, equipping me with the skills needed to tackle upcoming challenges effectively.

Week II

Started on new project ACTEN3 as data engineer.

As a data engineer, I've embarked on a new project, ACTEN3, diving into its intricacies. This week, I'm immersing myself in understanding the project flow, conducting background research, and familiarizing myself with its components. I've begun by setting up the project locally, ensuring smooth functionality, and rigorously testing all its functions. Moreover, I'm actively connecting with new team members, fostering collaboration and knowledge exchange. This phase marks the initial steps toward effectively contributing to ACTEN3, laying the groundwork for successful project implementation and integration.

Week III

Getting familiar with the MW

This week, I'm immersing myself in understanding the middleware (MW) component of our project. I'm delving deep into the codebase, dissecting its functionality, and gaining insights into its operations. Additionally, I'm focusing on learning about pre-processing files, understanding how data is transformed and prepared for further analysis or processing. By familiarizing myself with MW and its intricacies, I'm equipping myself with the knowledge and skills necessary to effectively contribute to our project's development and optimization.

Week IV

Learning new software Gathr.

This week, I'm dedicating time to learning a new software called Gathr. I'm diving into its documentation, absorbing its features and functionalities. Alongside, I'm actively testing the software to understand its potential applications and capabilities. In this process, I'm also identifying any limitations of the software and devising strategies to overcome them by leveraging existing functions or alternative approaches. By gaining proficiency in Gathr and addressing its limitations, I aim to enhance our toolkit and streamline our data engineering processes effectively.

Week V

Working on Gathr

This week, my primary focus is on Gathr, where I'm dedicated to optimizing its functionalities and addressing any outstanding tasks. Simultaneously, I'm tasked with developing a series of ETL (Extract, Transform, Load) processes tailored to meet specific business requirements. These include ETLs for managing Business Data, Business Relation Data, Driver Data, Truck Data, Trailer Data, Stop Data, Location Data, and Assignment Data. Following development, I meticulously test each ETL to ensure accuracy, reliability, and alignment with business objectives. This week's efforts aim to streamline data management processes and enhance the effectiveness of our data workflows within Gathr.

Week VI

Resolving bugs and daily ticket.

This week is dedicated to maintaining project stability by resolving bugs and addressing daily tickets promptly. Additionally, I'm implementing an automated process to trigger ETLs upon file uploads through the user interface, enhancing efficiency and reducing manual intervention. I'm also adapting ETLs to accommodate changes in business requirements, ensuring alignment with evolving needs. Furthermore, I'm creating comprehensive workflows for all ETL processes, optimizing data management procedures. Lastly, I'm integrating TigerGraph with the ETLs to leverage its capabilities for enhanced data analysis and insights. This week's focus is on enhancing productivity, adaptability, and connectivity within our data ecosystem.

Week VII

Migration from gathr to python driven ETL

This week, we initiated the development of a configuration-driven ETL (Extract, Transform, Load) process focused on the "stop.csv" file. The primary steps included creating a dedicated configuration file to streamline and customize the ETL operations for "stop.csv". Following this, extensive testing was conducted to ensure the ETL process operates as intended, accurately handling data extraction, transformation, and loading phases. A significant enhancement was the integration of a Pydantic model, which was implemented to validate the data integrity and structure, ensuring that the information processed adheres to predefined schemas and standards. This addition not only improved the reliability of the data handling process but also enhanced the overall robustness of our ETL solution.

Next Steps: Further refine the ETL process by incorporating advanced data validation techniques and explore optimization opportunities to enhance performance and scalability.

Week VIII

Migration from gathr to python driven ETL

This week marked a significant expansion in our ETL (Extract, Transform, Load) processes, extending our operational scope to include a variety of crucial data files. The focus was on developing and implementing configuration-driven ETL processes for the following datasets:

- Business Data: Established a foundational structure for handling core business metrics and information.
- Load Data: Developed processes to manage data related to cargo and freight loads.
- Load_BusinessRelation Data: Created a specialized ETL to handle the nuanced relationships between loads and business entities.
- Driver Data: Focused on aggregating and transforming data concerning our fleet drivers.
- Truck Data: Developed a configuration to process data related to the trucks in our fleet, focusing on operational metrics.
- Trailer Data: Implemented an ETL process for managing data associated with our trailers, crucial for logistics and planning.
- Location Data: Created a process to handle geospatial and location-based data, essential for route planning and optimization.
- Assignment Data: Developed a system to manage and transform data related to assignments, including driver, truck, and load assignments.

Each of these data files now has a corresponding configuration file designed to tailor the ETL process to the specific needs and structures of the data. This approach ensures flexibility, scalability, and precision in handling diverse datasets.

Testing Phase: Following the development of these ETL processes, extensive testing was carried out. This phase aimed to validate the efficacy, accuracy, and reliability of our ETL operations, ensuring that data transformation and loading meet our stringent standards for data integrity and quality.

Week IX

Resolving bugs and daily tickets.

This week, our team focused on further refining the ETL (Extract, Transform, Load) processes, with a special emphasis on integrating geocoding capabilities and improving our handling of data within the blob storage system. The enhancements are designed to elevate our data processing framework's accuracy and efficiency, particularly in dealing with spatial data and storage organization. Below are the key accomplishments:

Geocoding and Reverse Geocoding Integration: Implemented geocoding functionalities to convert addresses into geographical coordinates, enhancing our datasets with precise location data. Similarly, reverse geocoding was integrated, enabling the conversion of geographical coordinates back into readable addresses. These capabilities are crucial for improving the richness and usability of our location data, facilitating better analysis and decision-making.

Blob Storage Management: A systematic approach was adopted for managing files within the blob storage, focusing on:

- **Success_Records:** Created a dedicated storage container for records processed successfully, ensuring easy access and analysis of successful data transactions.

- **Error_Records:** Established a separate container for error records, streamlining the process of identifying and troubleshooting data processing issues.
- **Archive:** Implemented an archiving strategy for historical data preservation, allowing for efficient storage management and data retrieval when necessary.
- **Dropzone**: Set up a specific area (dropzone) for incoming files, serving as a staging area before files undergo the ETL process. This helps in organizing and prioritizing files for processing.

Environment-Dependent Configurations: All changes made to the ETL processes, including geocoding, reverse geocoding, and storage management, have been made environment-dependent. This approach allows for greater flexibility and scalability, ensuring that our ETL framework can adapt to different operating environments and configurations seamlessly. It facilitates customization and optimization based on specific operational needs and scenarios.

Week X

Creating a preprocessor.

This week, we developed a sophisticated parsing script capable of extracting multiple datasets from a single input file, incorporating critical business insights directly into the data processing workflow. The script intelligently segments data according to predefined business logic, enhancing our analytical capabilities. We've implemented version control for ongoing enhancements and automated the script's execution through a cron job, ensuring timely data processing. Future efforts will focus on advancing parsing techniques and integrating AI for improved data categorization. This development marks a significant step forward in automating and optimizing our data analysis and reporting processes.

Week XI

Resolving bugs and daily ticket.

This week, we initiated a significant upgrade to our parsing script, aimed at enhancing its efficiency and integrating advanced data processing capabilities. Concurrently, we embarked on the migration of our data storage to a new blob system, streamlining our data management and accessibility. A pivotal development was the creation of environment-specific filesystems, optimizing performance across different operational contexts. Additionally, we designed and implemented APIs to facilitate seamless retrieval of files from the blob storage, ensuring smoother data flows and improved access for analysis and reporting. These developments represent a comprehensive effort to modernize our data infrastructure and improve our analytical and operational agility.

Week XII

Creating a feature to get logs

This week, we enhanced our data processing framework by incorporating a logging feature into both our ETL (Extract, Transform, Load) and preprocessing stages. This new functionality captures detailed operational logs, significantly improving our ability to monitor, debug, and optimize these critical processes. To complement this advancement, we developed a dedicated API designed to efficiently fetch logs from both the ETL and preprocessor. This API streamlines the access to log data, enabling rapid analysis and troubleshooting. These updates mark a strategic improvement in our data management capabilities, offering deeper insights into our workflows and fostering a more robust data processing environment.

Week XIII

Resolving bugs and daily ticket.

This week's focus was on enhancing the parser script to efficiently generate home time (ht) loads, alongside addressing key issues for improved accuracy and functionality. Notably, we resolved a critical bug affecting driver name accuracy and corrected improper handling of <NA> field values, ensuring data integrity. Additionally, modifications were made to accurately generate ht loads and adjust status reporting anomalies. These strategic updates have significantly bolstered the script's reliability and efficiency, laying a solid foundation for processing large data volumes and maintaining high data quality standards, essential for our operational success and analytical precision.

Week XIV

Stabilizing ETL

I focused on enhancing the stability of our ETL (Extract, Transform, Load) process by diligently addressing potential loopholes. This involved thorough testing by running the ETL multiple times to identify and rectify any issues. By ensuring seamless operation and eliminating loopholes, we aim to enhance the reliability and efficiency of our data pipeline, ultimately optimizing our data management system.

Week XV

Optimization of ETL

I concentrated on optimizing the execution time of our ETL process. This involved refining Tiger Graph calls and strategically adjusting delays to their optimal values. By fine-tuning these parameters, we aim to minimize the total time taken for the ETL process to run, thereby improving overall efficiency, and enhancing the responsiveness of our data pipeline.

Week XVI

Documentation and KT

I've dedicated efforts to documentation and knowledge transfer for our project's new processes. This included meticulously crafting comprehensive documentation detailing the intricacies of the newly created processes. Additionally, I conducted Knowledge Transfer sessions with our new team member to ensure they have a clear understanding of the project's objectives, methodologies, and intricacies. These efforts aim to facilitate smooth onboarding and empower team members with the necessary insights to contribute effectively to the project.

Conclusion

Over the past weeks, I've embarked on an enriching journey as a Cloud Data Engineer, diving deep into various tools, technologies, and project tasks. From mastering essential cloud platforms like AWS to exploring infrastructure as code with Terraform, each week has been filled with learning and growth. Delving into new projects, understanding middleware components, and troubleshooting bugs have honed my problem-solving skills and adaptability. Creating and testing ETL processes, integrating with new software like Gathr, and connecting with TigerGraph have expanded my expertise in data management and analysis. As I conclude this period, I'm equipped with a diverse skill set and a deeper understanding of cloud data engineering, ready to tackle upcoming challenges and drive innovation in our projects.