International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Unravelling Transportation Trends with Data Engineering and Analytics

Mr. N Uday Kumar[1], A. Sai Kiran[2], B. Hemanth Kumar[3] , C . Preethi[4]

[1]Associate Professor, Department of Computer Science
[2-4]B.Tech Student, Department of Computer Science
[1-4]Raghu Engineering College, Visakhapatnam

Abstract – The Unravelling Transportation Trends with Data Engineering and Analytics project delves into the correlation between driver behavior and customer ratings within ride-hailing platforms. Utilizing a dataset akin to Uber's, the study investigates the impact of acceptance rates and cancellation rates on customer satisfaction, aiming to glean valuable business insights for enhancing service quality and fostering business growth. Meticulous data collection and preprocessing ensure data accuracy and reliability, while secure cloud-based storage facilitates efficient data processing. By analyzing driver behavior metrics and customer ratings, the project uncovers pivotal patterns and trends, offering essential guidance for optimizing driver performance. This data-driven approach empowers ride-hailing services to implement targeted strategies for improving service quality, enhancing customer retention, and cultivating a loyal customer base.

Keywords—Ride-hailing platforms, Driver behavior, Customer ratings, Data engineering Analytics, Service quality, Customer satisfaction, Data accuracy, Driver performance, optimization
Data analytics.

1. INTRODUCTION

Research aims to leverage the power of big data and the advantages of derived insights, scientific discoveries, and better knowledge to help decision-making in a data-driven society. Better sensemaking over big data is made possible by the development and convergence of methods and technologies, such as improvements in machine learning and deep learning techniques, higher storage capacities and lower storage costs, faster network speeds and greater bandwidth, more affordable and potent high-performance computing, and an increasing ubiquity of sensor networks and smart technologies. Nevertheless, it frequently happens that significant discoveries and insights are contained in and spread over a number of dispersed datasets rather than existing within a single dataset. The potential benefit and impact of facilitating analyses and sensemaking across dispersed, complicated, and fragmented data are abundantly demonstrated by prior research (e.g., [1]–[8]), which spans numerous disciplines. However, there are still big obstacles to overcome.

The project encompasses four distinct phases, each integral to unveiling the link between driver behavior and customer ratings.

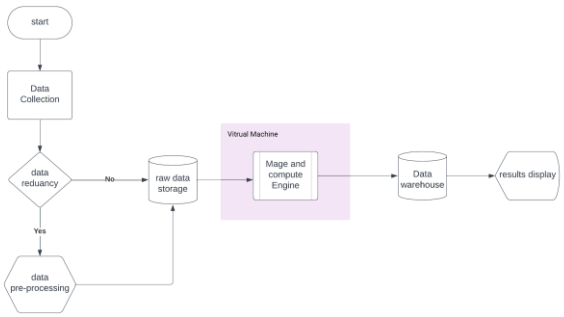

Fig:1. Workflow of the suggested pipeline for data engineering and analysis

Data Collection and Preprocessing:

- Identify and collect a comprehensive dataset from publicly available sources, similar to Uber's ride-hailing data.
- Implement rigorous data preprocessing techniques to ensure data integrity, handle missing values, and eliminate outliers.

Data Engineering with Mage and Google BigQuery:

- Design and implement an efficient Extract, Transform, Load (ETL) pipeline using the Mage platform to orchestrate data transformations and aggregations.
- Utilize Google BigQuery as the data warehousing solution to process and analyze the dataset using advanced SQL queries.
- Explore statistical analysis techniques to identify significant correlations between driver behavior and customer ratings.

Identifying Driver-Customer Dynamics:

- Analyze driver performance metrics, including acceptance rate and cancellation rate, concerning customer ratings.
- Gain insights into customer preferences and how driver actions influence their experience.
- Uncover patterns and trends that highlight the impact of driver behavior on customer satisfaction.

Optimizing Driver Performance and Business Growth:

- Leverage the project's findings to optimize driver performance and enhance service quality in ride-hailing services.

In this paper, we present an interactive visual analytics framework for distributed data analysis systems. Through platforms like Uber and Lyft, ride-hailing services have revolutionized urban mobility by providing convenient transit options. In this industry, knowing the relationship between driver behavior and consumer happiness is essential. This study explores acceptance rates, cancellation rates, and their relationship to overall customer experience using a dataset similar to that of industry leader Uber. The initiative guarantees the integrity and dependability of its results by using rigorous data gathering and preprocessing methods. In order to improve service quality, the project will investigate this link using advanced analytics on datasets that are similar to those in the industry. By means of rigorous data gathering and analysis, the research explores a number of parameters, including consumer ratings and acceptance rates.

2. LITERATURE REVIEW

In order to do analysis, modern data-driven applications frequently need to identify and transfer pertinent data from several sites to a single storage location. Some alternative solutions for data sharing have been offered that make use of cloud-based hosting and high-speed networks to get around what is typically a challenging, time-consuming, and arduous operation. Other alternative solutions center around the provision of shared computing resources. A project called DataONE [1], [2] aims to make earth and environmental research data sources easier to find, search, and access. Through the provision of networked computing resources, the Open Science Grid [3], [20] facilitates scientific research.In an effort to reduce data travel, SciServer gathers all of the necessary data at one central storage site, but it does so at the location where most of the data is already present. Additionally, SciServer moves the analyses to the shared storage location by forwarding Jupyter Notebooks [22].

Research infrastructures that combine storage, high-performance computing, and analytical tools are the goal of other data-driven applications (e.g., XSEDE [11], [23], NeCTAR [24], PRACE [25], and EGI [26]).Users can share data repositories and dispersed computing resources by using these applications. Science Gateways (SGs) [5], [17]–[20] may employ the solutions to create (web) portals and user interfaces (UIs) that let scientists (such biologists and chemists) access, create, and carry out analytic workflows. Scientists are relieved of the responsibility and necessary knowledge to set up and manage the underlying distributed cyber-infrastructure by SGs.

Various end users can share and reuse SG services. SGs can be categorized into instances of SG, such as the Computational Neuroscience Gateway [13], and SG frameworks, such as WS-PGRADE/gUSE [27]

3. SYSTEM DESIGN

In order to streamline user interactions and improve the user experience with a data analytics system (DAS), this paper presents an interactive visual analytics framework (VAF). To do this, we examined a number of distributed analytical systems (e.g., [4], [8], [22], [23]) and determined the essential user interactions needed to run these systems. As such, both the data analysts and the end users may find the entire process of doing a data analysis to be equally difficult. Furthermore, users from other domain regions had to retrieve the resultant data from the server in order to examine the results. DAS frequently offers a visualization toolkit in place of command line interfaces [22], [24]. But users are in charge of creating the appropriate artifacts or exploratory visualizations to gauge how well the analysis performed [25].

A. Data pre -processing and engineering:

- **Data Identification and collection :** The system shall identify and collect a comprehensive dataset akin to Uber's ride-hailing data from publicly available sources.
  It shall ensure the dataset includes relevant information such as driver behavior metrics and customer ratings.
- **Data Pre-processing:** The system shall implement data preprocessing techniques to ensure data integrity. Using the Pandas module to pre-process the data.
- **Data Storage:** The system shall utilize Google Cloud Storage or equivalent for storing the dataset securely.

B. Data Analysis and Insights Generation:

- **ETL Pipeline:** The system shall design and implement an Extract, Transform, Load (ETL) pipeline for data processing.It shall orchestrate data transformations and aggregations efficiently. Using Mage, a software data engineering pipeline tool to construct the ETL.
- **Statistical Analysis:** The system shall employ statistical analysis techniques to identify correlations between driver behavior and customer ratings.It shall provide insights into customer preferences and the impact of driver actions.

C. Strategy Formulations and Recommendations:

- **Optimization Strategies:** The system shall formulate data-driven strategies to optimize driver performance and enhance service quality. It shall align driver actions with customer expectations effectively.
- **Actionable Insights:** The system shall generate actionable insights for ride-hailing companies to improve business growth, increase customer retention, and foster a loyal customer base

4. PROPOSED MODEL

The system we have in mind for our research project is a comprehensive approach to data engineering and analytics that makes use of a range of tools and technologies, including GCP services from Google Cloud Platform. This all-inclusive ecosystem is intended to simplify the data processing workflow, enable effective analysis, and enable researchers to extract meaningful insights from intricate datasets. The GCP services, which provide dependable infrastructure and scalable solutions for managing complex data processing activities, are the foundation of our system design. Google Cloud Storage offers safe and dependable data storage capabilities, acting as the basis for file archiving and retrieval from any location in the cloud. Furthermore, virtual machine deployment and maintenance are made possible by Google Compute Engine, which makes it simple and effective for researchers to execute their applications.

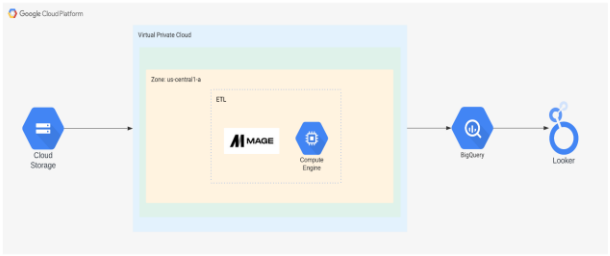

Fig.2. Proposed Model

BigQuery, Google's data warehousing tool, is a crucial part of our system since it provides strong analytical capabilities and a recognisable SQL-type interface. BigQuery is perfect for processing the enormous volumes of data that are usually encountered in research projects since it allows academics to store and analyze large-scale datasets. BigQuery's highly scalable and cost-effective design guarantees that researchers may execute intricate analytical activities without sacrificing scalability or performance. Looker Studio is a web-based business intelligence application that enhances the GCP services by offering sophisticated reporting and visualization features.

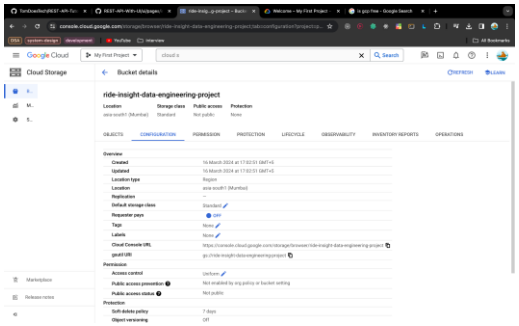

Fig.3 GCP Bucket Creation

Researchers can easily develop interactive dashboards and visualizations using Looker Studio, which seamlessly connects with GCP services to improve the readability of insights into their data and communicate their findings effectively.

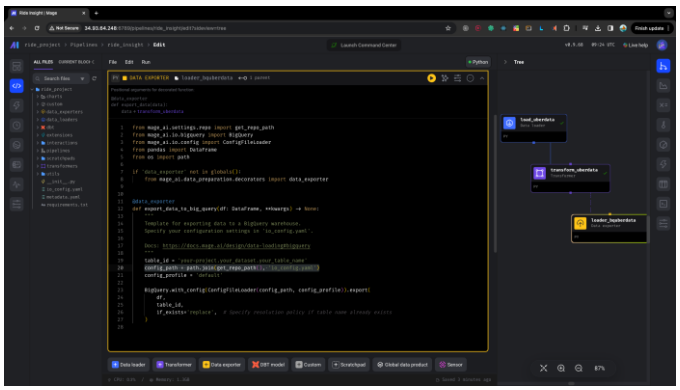

Fig.4. Construction the ETL pipeline

To assist the data processing pipeline, our solution integrates multiple tools and technologies in addition to GCP services and Looker Studio. Jupyter Notebooks offer an interactive environment for testing and executing code, allowing academics to try various techniques and approaches. By streamlining the data extraction and transformation process, Mage, an open-source tool for building up ETL pipelines, frees up researchers to concentrate on their business logic. The diagramming programme Lucidchart makes it easier to create flow charts and design diagrams, which helps to visualize project operations.

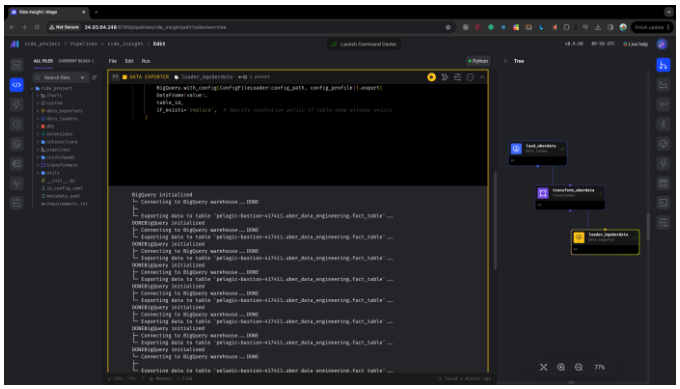

Fig . 5. Data is loaded into Big-Query through ETL Pipeline

Our suggested approach offers a thorough framework for data engineering and analytics by fusing various tools and technologies, empowering researchers to glean insightful information from intricate datasets and spur innovation in their domains.

Fig. 6. Loaded all the tables into Big Query

5. RESULTS

The project's performance analysis provided insightful information about how well the techniques and methods used to carry out the project were able to meet its goals. The project team evaluated a range of performance metrics and results using statistical analysis and thorough review.The effectiveness of data processing and analysis using the chosen tools and technologies was one area of emphasis for the performance analysis.

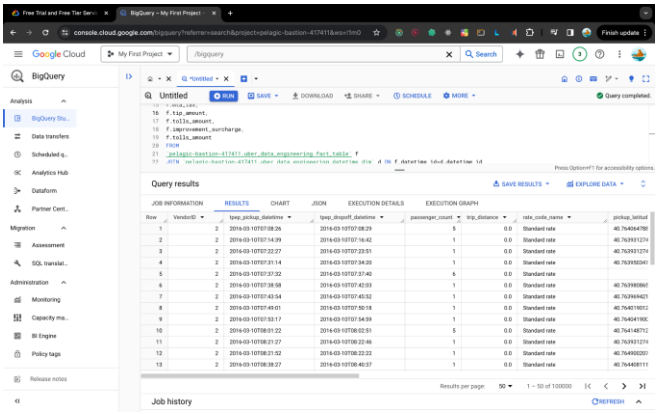

Fig. 7. Query to combining all the fact tables

The implementation of Google Cloud Platform (GCP) services, including BigQuery and Google Cloud Storage, enabled smooth data retrieval, storage, and analysis while guaranteeing scalability and dependability.

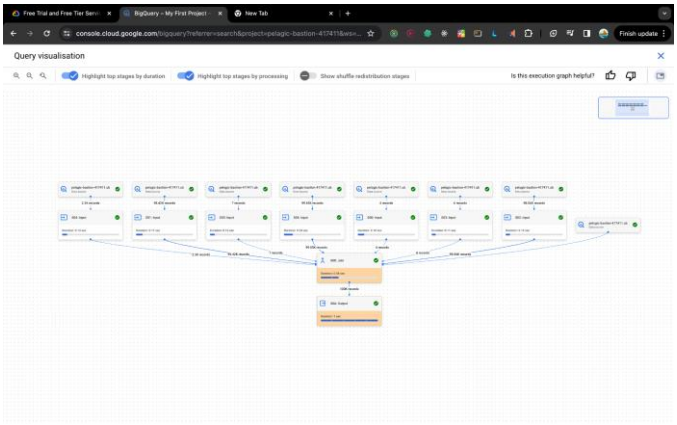

Fig. 8. Insight Generation on Big Query

Furthermore, interactive dataset exploration and visualization were made possible by the integration of tools like Looker Studio and Jupyter Notebooks, which improved the interpretation of the findings. Additionally, the performance research assessed how data-driven tactics could improve consumer happiness and driver performance on ride-hailing platforms. The project team found practical insights and suggestions for enhancing service quality and customer experience by looking at important indicators including driver acceptance rates, cancellation rates, and customer ratings.

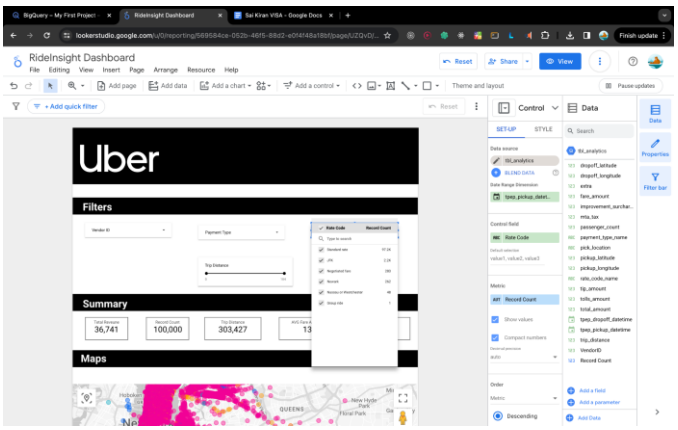

Fig.9.1 Data Analysis Visual Report on Locker Studio

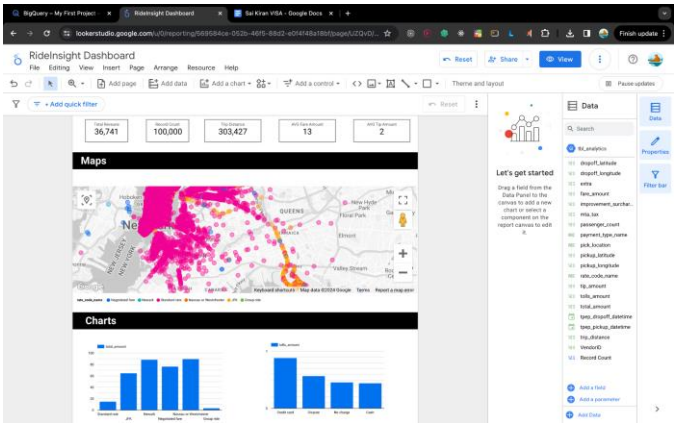

Fig 9.2 Location and Payment Insights Report

Overall, the performance study demonstrated how well the project's methodology used data engineering and analytics to extract insightful conclusions and propel significant advancements in the ride-hailing sector.

6. CONCLUSION

The experiment has, in summary, illuminated the complex relationship between driver behavior and ride-hailing service user happiness. A thorough examination of the data has yielded important insights that show how customer ratings, acceptance rates, and cancellation rates affect the quality of the services provided. The researcher's conclusions offer practical advice on how ride-hailing services may improve driver performance and the general consumer experience. Ride-hailing services can achieve business growth, client retention, and customer loyalty by utilizing data-driven techniques, like the ones explained in this report. Future efforts to increase the efficacy and efficiency of ride-hailing systems in satisfying consumers' changing needs will be built upon the foundation created by this research

7. REFERENCES

[1] R. J. Sandusky, "Computational provenance: Dataone and implications for cultural heritage institutions," in 2016 IEEE International Conference on Big Data (Big Data). IEEE, 2016, pp. 3266–3271.

[2] J. P. Cohn, "Dataone opens doors to scientists across disciplines," 2012.

[3] R. Pordes, D. Petravick, B. Kramer, D. Olson, M. Livny, A. Roy, P. Avery, K. Blackburn, T. Wenaus, F. Wurthwein ¨ et al., "The open science grid," in Journal of Physics: Conference Series, vol. 78, no. 1. IOP Publishing, 2007, p. 012057.

[4] D. Medvedev, G. Lemson, and M. Rippin, "Sciserver compute: Bringing analysis close to the data," in Proceedings of the 28th international conference on scientific and statistical database management, 2016, pp. 1–4.

[5] S. Gesing, J. Kruger, R. Grunzke, S. Herres-Pawlis, and A. Hoffmann, ¨ "Using science gateways for bridging the differences between research infrastructures," Journal of Grid Computing, vol. 14, no. 4, pp. 545–557, 2016.

[6] I. Foster, "Globus online: Accelerating and democratizing science through cloud-based services," IEEE Internet Computing, vol. 15, no. 3, pp. 70–73, 2011.

[7] S. Gugnani, C. Blanco, T. Kiss, and G. Terstyanszky, "Extending science gateway frameworks to support big data applications in the cloud," Journal of Grid Computing, vol. 14, no. 4, pp. 589–601, 2016.

[8] A. Talukder, M. Elshambakey, S. Wadkar, H. Lee, L. Cinquini, S. Schlueter, I. Cho, W. Dou, and D. J. Crichton, "Vifi: Virtual information fabric infrastructure for data-driven discoveries from distributed earth science data," in 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). IEEE, 2017, pp. 1–8.

[9] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in Proceedings of international conference on intelligence analysis, vol. 5. McLean, VA, USA, 2005, pp. 2–4.

[10] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim, "Knowledge generation model for visual analytics," IEEE transactions on visualization and computer graphics, vol. 20, no. 12, pp. 1604–1613, 2014.

[11] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr, "Xsede: Accelerating scientific discovery," Computing in Science Engineering, vol. 16, no. 5, pp. 62–74, Sep. 2014.

[12] R. J. Sandusky, "Computational provenance: Dataone and implications for cultural heritage institutions," in IEEE International Conference on Big Data (Big Data), Dec 2016, pp. 3266–3271.

[13] S. Shahand, M. M. Jaghoori, A. Benabdelkader, J. L. Font-Calvo, J. Huguet, M. W. Caan, A. H. van Kampen, and S. D. Olabarriaga, Computational Neuroscience Gateway: A Science Gateway Based on the WS-PGRADE/gUSE. Cham: Springer International Publishing, 2014,

[14] M. Elshambakey, M. Khalefa, W. J. Tolone, S. D. Bhattacharjee, H. Lee, L. Cinquini, S. Schlueter, I. Cho, W. Dou, and D. J. Crichton, "Towards a distributed infrastructure for data-driven discoveries & analysis," in 2017 IEEE International Conference on Big Data (Big Data). IEEE, 2017, pp. 4738–4740.

[15] C. Chokwitthaya, Y. Zhu, R. Dibiano, and S. Mukhopadhyay, "Combining context-aware design-specific data and building performance models to improve building performance predictions during design," Automation in construction, vol. 107, p. 102917, 2019.

[16] O. T. Karaguzel, M. Elshambakey, Y. Zhu, T. Hong, W. J. Tolone, S. Das Bhattacharjee, I. Cho, W. Dou, H. Wang, S. Lu et al., "Open computing infrastructure for sharing data analytics to support building energy simulations," Journal of Computing in Civil Engineering, vol. 33, no. 6, p. 04019037, 2019.

[17] R. Zhang and O. T. Karaguzel, "Development and calibration of reduced order building energy models by coupling with high-order simulations," Global journal of advanced engineering technologies and sciences, vol. 7, no. 2, 2020.

[18] W. J Tolone, "Application of the virtual information fabric infrastructure (vifi) to building performance simulations," Current Trends in Civil & Structural Engineering, vol. 4, no. 2, 2019.

[19] D. Merkel, "Docker: Lightweight linux containers for consistent development and deployment," Linux J., vol. 2014, no. 239, Mar. 2014.

[20] I. Miell and A. H. Sayers, Docker in Practice, 1st ed. Greenwich, CT, USA: Manning Publications Co., 2016.

[21] https://nifi.apache.org/.

[22] https://docs.docker.com/engine/swarm/

[23] K. Potter, A. Wilson, P.-T. Bremer, D. Williams, C. Doutriaux, V. Pascucci, and C. Johhson, "Visualization of uncertainty and ensemble data: Exploration of climate modeling and weather forecast data with integrated visus-cdat systems," in Journal of Physics: Conference Series, vol. 180, no. 1. IOP Publishing, 2009, p. 012089.

[24] A. Matˇ acut¸ ˇ a and C. Popa, "Big data analytics: Analysis of features ˇ and performance of big data ingestion tools," Informatica Economica, vol. 22, no. 2, pp. 25–34, 2018.

[25] P. Kacsuk, Science gateways for distributed computing infrastructures: Development framework and exploitation by scientific user communities. Springer International Publishing, 8 2014.

[26] Visual Analytics Frameworks (IEEE Xplore):
The paper titled "Visual Analytics Frameworks for Distributed Data Analysis" (https://ieeexplore.ieee.org/document/9671768) elucidates the relevance of visual analytics in distributed data analysis systems.

[27]The article "Modern Data Engineering with Mage: Empowering Efficient
Data(https://www.analyticsvidhya.com/blog/2023/06/modern-data-engineering-with-mage-empowering-efficient-data-processing/)

[28] https://cloud.google.com/docs/tutorials: gcp cloud documentation