Data Science and Engineering (DSE) Record, Volume 4, issue 1.

Development of Batch Data Pipeline System for Flight Delay Prediction

Suchada Manowon [1] and Pruet Boonma [2]

[1] Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand
[2] Department of Computer Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand
suchada_manowon@cmu.ac.th

Abstract. Flight delays persist as a challenge, which impacting airline and airport productivity, passenger experience, and financial resources. Nowadays, air transportation data predominantly rely on administrative records from various institutions. This study aims to designing and implementing an effective data pipeline system with the capacity to capture high-frequency data from diverse sources through batch processing. This comprehensive pipeline encompasses the entire of end-to-end data pipeline stages; including data sourcing, ingestion, processing, storage, and analysis.

The proposed pipeline system extracts data from various datasets, including flight data, airport information, airline details, airplane specifications, and routes. It employs a variety of methods such as web scraping, APIs, and database loading for data ingestion. It efficiently consolidates flight information, transforming and cleaning data and then loading it into a designated destination database. Additionally, this study establishes an automated batch processing platform using Apache Airflow. This platform is characterized by a comprehensive evaluation across three essential aspects; 1. System metrics, including memory and disk usage, 2. Job metrics extracted from Airflow metrics, which are utilized to monitor processes, ensuring smooth execution, 3. Data quality metrics that assess six dimensions – accuracy, validation, completeness, consistency, uniqueness, and timeliness – to ensure the usability of the defined data.

Leveraging the flight dataset for data analysis and data visualization, this approach involves the comparison of various base regression models for flight delay prediction. Additionally, flight data dashboards offer data insights. The implications of this multifaceted approach extend to enhancing air transportation statistics, predictive modeling capabilities, and facilitating data-driven decision-making processes.

Keywords: ETL, Data Monitoring, Data Analysis, Data Visualization.

1 Introduction

Data and information have always been vital in aviation. Their significance has amplified in the 21st century due to the development of new aircraft and the ongoing advancements in technology. Data plays a huge role in every aspect of aviation from

passenger comfort to engine efficiency, such as economy, business, government, cargo, people mobility, travel and tourism [1]. As the air traffic have a main role in economy of agencies and airports in the transportation industry, it is necessary for them to increase quality of their services. One of the important challenges of airports and airline agencies is flight delay. Flight delay is a longstanding problem with the transportation industry, which massively affects the productivity of airlines and airports around the world. In addition, delay in flight makes passengers concerned and this matter causes extra expenses for the agency and the airport itself [2].

Currently, the Air Transportation Statistics rely on administrative data from various institutions. it is necessary to have a new data source that can serve as an alternative reference for air transportation activities which is faster and more granular. Activity flight data is extremely high volume and pipelines present new design challenges. Therefore, this research proposes to design and implement the effective data pipeline, gathering data from multiple sources and formats, providing insights into flight delay prediction and offering a flight data dashboard for data-driven decision-making within the aviation domain.

The one of the important challenges of pipeline system is the analysis of flight delays, which is performed through batch processing. This approach necessitates waiting for a certain volume of raw data to accumulate before initiating the pipeline system. Typically, this means data is between an hour to a few days old before it is made available for analysis [3]. In this research, batch process jobs are scheduled to run every 24 hours, ensuring a recurring execution. During each run, data is extracted from the source, subjected to various operations, and subsequently published to the data sink. This process continues until all data has been processed.

The handling of a batch data stream requires the distributed operational end-to-end data pipeline. This system should possess the capability to efficiently capture high-frequency data from various sources on a schedule. These stages of batch processing involve in figure 1 overview data pipeline architecture including stages; data sourcing, data ingestion, data processing, data storage, and data analysis.

The following are the main objectives of this independent study:

1. To implement data pipeline for analysis of flight delay prediction and data visualization of flight data.
2. To monitor and evaluate performance of data pipeline.

2 Literature Review

2.1 Data pipeline architecture

Juan Carlos Farah [4] conducted a study on a architecture designed the challenges of managing Learning Analytics (LA) data, acquisition, visualization, anonymization, and sharing and enabled users to access their datasets on demand. Although currently optimized for teacher-mediated contexts, the architecture can be adapted to other

domains like digital healthcare and e-government, supporting open data initiatives across various fields. Overall, this architecture represents a valuable contribution to enhancing data management and promoting open data practices in learning analytics.

In the context of processing data streams pertinent to COVID-19 policy discussions in the UK, Gaythorpe [5] presents their approach to processing data streams used in COVID-19 policy discussions in the UK. The pipeline showcased robustness and flexibility, yet the authors emphasize that human validation, understanding the data's context and provenance, remains irreplaceable, especially when dealing with evolving data and errors. Human validation can be systematically integrated as data pipeline, complementing the increasing emphasis on code and analysis reproducibility. Finally, more validation with reproducibility ensures a more reliable approach to data processing and analysis in pandemic scenarios.

S.Sajida [6] aims to explore the research efforts and opportunities in backstage of data warehousing. The survey begins with an examination of open source and commercial ETL tools, such as SIRIUS, ARKTOS, PYGMATEL, Talend Open Studio, and Microsoft ETL solution (SSIS). Subsequently, the focus shifts to ETL modeling and design, covering works that utilize various formalisms like UML and web technologies.

2.2 Data quality assessment

In the data quality assessment within transportation systems, Shawn Turner [7] study in quality of traffic data in transportation systems. This focused on various aspects of traffic data quality and organizing stakeholder’s workshops in Columbus, Ohio, and Salt Lake City, Utah, to gather feedback and insights for the action plan. The recommended action plan provides guidelines for defining and assessing traffic data quality. Six key data quality measures are highlighted: accuracy, completeness, validity, timeliness, coverage, and accessibility, along with examples and definitions. However, further research is needed to establish clear calculations and applications of these data quality measures in different transportation domains. With the action plan, stakeholders can make informed decisions and implement effective strategies to enhance the overall quality of traffic data for various transportation functions.

Jeusfeld [8] studied in approach the data warehouse's quality through a semantically rich model of quality management. The model enables stakeholders to define abstract quality goals, which are then translated into executable analysis queries on quality measurements stored in the data warehouse's meta database. The implementation of this approach is currently underway using the ConceptBase meta database system. This methodology provides a structured and efficient means of evaluating and maintaining data warehouse quality, facilitating better decision-making processes for users.

Aparna Nayak [9] focus on determined by its suitability for operational, decision-making, and planning purposes. Despite the abundance of linked data available on the web. To address these quality issues, quality assessment, and data refinement. However, existing frameworks often focus on individual aspects and are largely tailored to DBpedia-related challenges. The study provides insights into the current state of data

quality evaluation and proposes a solution based on ontology for developing an end-to-end system capable of analyzing the root causes of quality violations.

2.3 Flight delay prediction models

The prediction of flight delays has garnered significant attention within the aviation research. In 2018, Bhuvan Bhatia [10] has studied in two parts. The first part involves utilizing flight data, weather information, and demand data to predict flight departure delays using techniques such as Logistic Regression, Random Forest, and Support Vector Machine (SVM). The results indicate that the Random Forest method outperforms the SVM model in terms of predictive performance. The second part of the study explores the potential for predicting flight delay patterns solely based on the volume of concurrently published tweets, along with their sentiment and objectivity.

Choi and team [11] developed binary classification to predict scheduled flight delays, with a specific focus on addressing the challenges posed by data imbalancing during the training process. To achieve this goal, they employed techniques including Decision Trees, AdaBoost, and K-Nearest Neighbors (KNN) for predicting individual flight delays.

In 2020, Yogita Borse [12] focused on scale back monetory loss and for the higher and smooth operation and develop a system to predict the delay in flights. They have use methodology like classification or regression ways are often accustomed determine the delay which includes Feed forward network, Neural Network, Random Forrest, decision tress, Naïve Bayes Classification Tree, Regression Tree and etc.

Jingyi Qu [13] proposes a flight delay regression prediction model named Att-Conv-LSTM. This model harnesses spatio-temporal neural networks and an attention mechanism module, while also incorporating meteorological information to enhance the accuracy of flight delay predictions. The model's efficacy is show in data from four airports in Beijing, Tianjin, and Hebei. The study's principal findings reveal that concurrently extracting time series and spatial features from the data leads to superior accuracy when contrasted with the utilization of a singular temporal neural network.

3 Methodology

Methodology is divided into 3 components, following the data pipeline architecture shown in Figure 1. These components contain, the data pipeline implementation, system monitoring, and data consumption, combining into the end-to-end data pipeline system. All these elements are orchestrated and managed through Apache Airflow, [14] a platform that allows for programmatically, scheduling, and monitoring workflows operating within its Docker container environment. The runtime version used is 4.16.2 (95914), and the environment is allocated 8GB of RAM, 8GB of SWAP memory, and 2 processors.

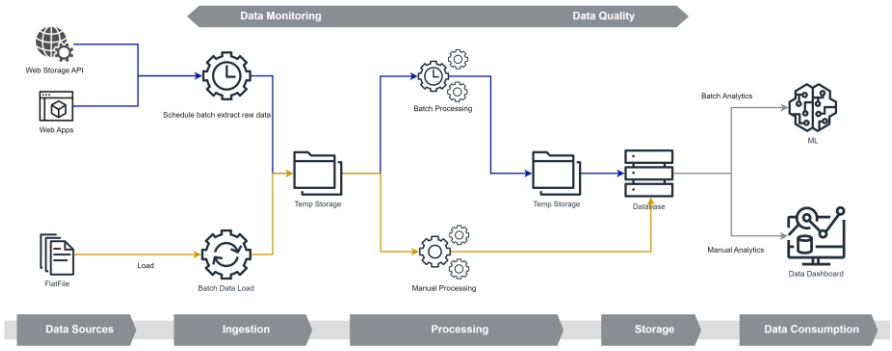

Figure 1: Overview data pipeline architechture of this study.

3.1 Data extraction, transformation, and loading (ETL)

In the extract-transform-load (ETL) process, data is extracted from diverse sources, transformed for analysis, and loaded into data warehouses or lakes for further processing, following the guidelines of Kimball and Rose [15][16]. The extraction phase gathering relevant flight data from various sources, in different methods of exrtraction base on their data format such as direct loading, API calls and web scraping. Then the transformation stage converts these data into format for target system's requirements and analysis, relying on techniques such as cleaning, aggregating, and joining. Subsequently, in the loading phase with strategies like Incremental Load, data is inserted or updated into the target system [17] in each ETL run, aligned with the system's schema. This step involves mapping, integrating, validating, and writing data to appropriate target tables.

3.2 Evaluation indicators

The evaluation of the data pipeline is divided into two domains: data pipeline performance and data quality assessment. Performance and monitoring indicators were measured during the ETL process to assess the pipeline's efficiency. System and job metrics were employed to gauge performance, while data quality was evaluated for integrated data.

3.2.1 System Monitoring and Error Handling

Following the guidance from "Data Pipeline with Apache Airflow" [18], the system is continually monitored using key signals: latency, traffic, errors, and saturation. Task failures trigger email alerts for admin.

1) Latency measures response times, including webserver responses and task scheduler durations.
2) Traffic metrics monitor task workload, expressed as averages per duration.
3) Error metrics track zombie tasks, non-HTTP 200 responses, and timeouts.
4) Saturation metrics assess system resource utilization, aiding in understanding capacity limits.

The implementation of task monitoring workflow utilized Prometheus's server to store metrics from Airflow. StatsD client and Prometheus StatsD exporter convert and store metrics for visualization through Grafana.

3.2.2 Data quality

The "ydata-profiling" Python library [19] is utilized to extract descriptive statistics and insights from data profiling reports, including structure, completeness, uniqueness, correlation, and distribution. This informs effective data quality checks for perform data quality assemennts that incorporate Apache Airflow and a Python framework to evaluate integrated data using the six primary dimensions for data quality assessment [20][21]: Completeness, Uniqueness, Timeliness, Validity, Accuracy, and Consistency.

1) Completeness reveals varying levels of missing data or incompleteness across different schemas, with measures all columns, rows and cells.
2) Uniqueness assessment confirms not have duplicate rows, ensuring distinct and reliable data.
3) Timeliness is maintained through monthly and daily updates of airport, airline, flight schedule, and flight status data by a batch processing scheduler.
4) Validity is ensured via SQLColumnCheckOperator checks for accurate column names, data types, and null values.
5) Accuracy assessment involved comparing data extracted from flight, airport, and airline records against the validated information stored in the official database. Due to the large volume of data, a comprehensive manual comparison was impractical. This subset data from randomly chosen rows, allowing for effective re-identification and manual comparison to ensure accuracy.
6) Consistency verification enhance data reliability in format and pattern in all columns and rows.

3.3 Visualization and Analytic

3.3.1 Data visualization

Utilized Power BI for flight data dashboard, Following "**Better Data Visualization**" [22] guidelines, emphasizing clear presentation of essential information, reducing clutter, integrating graphics and text, employing small multiples, and strategically using color.

3.3.2 Flight delay prediction model

Following CRISP-DM methodology [23] for implement flight delay prediction involves phases such as business understanding, data understanding, data preparation, modeling, evaluation, and the deployment phase is not utilized in this study's approach.

4 Results

4.1 Data sources

To develop an effective flight delay prediction model, essential columns are required, which necessitates the gathering of data from multiple sources. Each data source demands a different extraction method due to its unique characteristics. The table 1 provides infomation into the data sources and their corresponding datasets, including the volume of data, number of columns, format, and extraction methods.

Table 1: Data sources information with their corresponding datasets, the volume of data, number of columns, format, and extraction method

Source name	Dataset	Number of rows	Number of columns	Format	Extraction
Bureau of Transportation Statistics (BTS)	flight	73623704	28	csv	download
Datahub	airport	57421	12	json	api
OpenFlight	airport	7698	14	json	api
	airline	6161	8	json	api
	route	67662	9	json	api
FlightRadar24	airport	5094	7	Json	api
	airline	1964	3	json	api
	flight	-	-	json	Web scraping

4.2 Data extraction, transformation, and loading

The 4 destination relational tables (table 2) are the culmination of the ETL process, providing structured and refined data that can be utilized for various analytical and decision-making tasks, such as a flight data dashboard and developing flight delay prediction models. For all destination database schemas are shown in **Appendix 1: Database schema.**

Table 2: Destination database information

	Table name	Description	Number of rows	Number of columns
1	airline	Airline information contains all airline detail contains index link to airport	6158 (21/5/2023)	12
2	airport	Airport information contains all airport detail including index, location and exist_flightradar24-additional column to check airport available in website for scraping.	57421 (21/5/2023)	23
3	Flight schedule	Flight schedule information contains schedule time departure and arrival with flight detail, gathering from web scraping between 2022-09-18 to now	200,000 (21/5/2023)	22
4	Flight status	Flight status information contains schedule time and their status with flight detail, gathering from web scraping between 2022-09-18 to now	83,201,329 (21/5/2023)	29

4.3 Monitoring and Data Quality

4.3.1 Monitoring job performance

All job's metrics which available from Apache Airflow document website [24], will be calculated and visualized via the Grafana client in three sub-dashboards: scheduler, execution pool, and DAG stats, source code template for Grafana dashboards provide by Github repository @rozhok's account [25]. These sub-dashboards cover job monitoring in four categories as described in book guidelines from "Data Pipeline with Apache Airflow": Latency, Traffic, Errors, and Saturation. The system will track all processes and send email alerts in case of task failures as design.

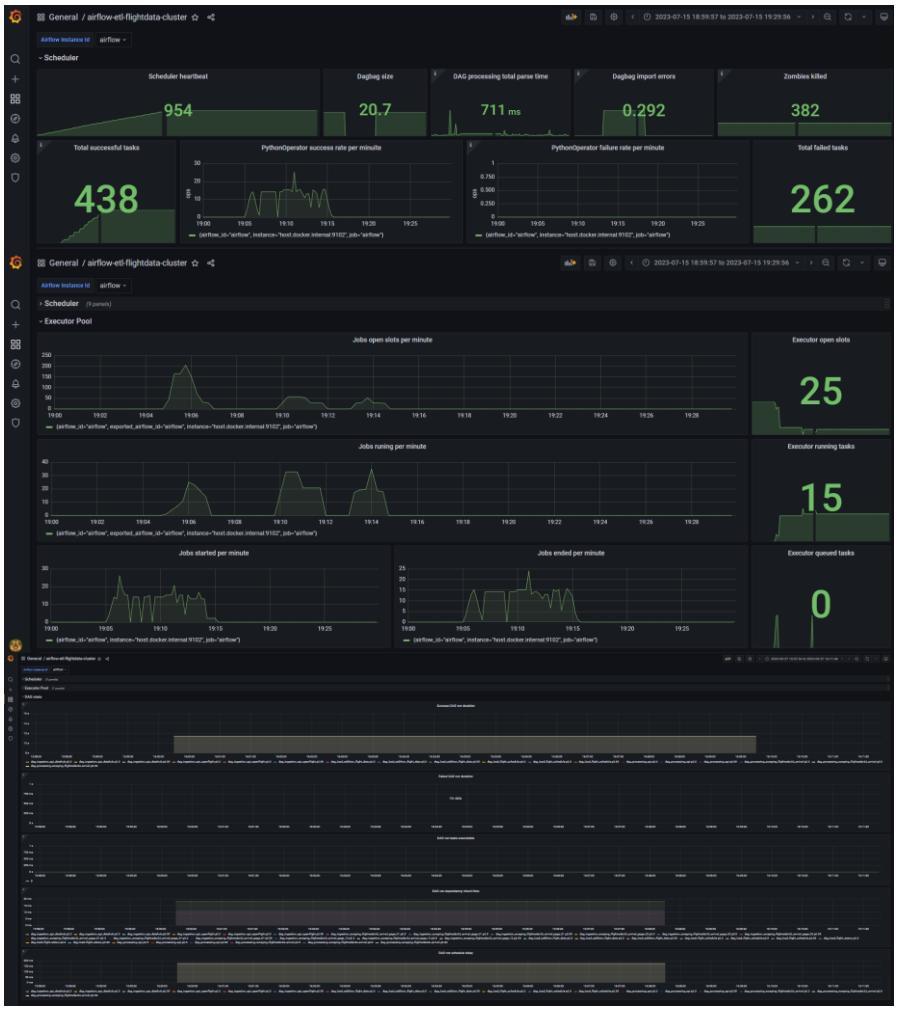

Figure 2: The job monitoring dashboard in the latency approach, which was implemented using Grafana on July 15, 2023

4.3.2 Monitoring system performance

The study object to monitor system performance during the execution of processes with a focus on resource usage and disk usage, implement by Docker Desktop program to collect system logs while various services were running. System monitoring is observation of various services in containers in real-time and dynamic change in table 3.

Table 3: The dynamic changes in system resource usage when Docker is running and its influence on disk usage

Name	CPU (%)	Memory Usage/Limit	MEM (%)	Disk Read/Write	Network I/O
airflow-init	0.00	0	0.00	0	0
airflow-web-server	1.46	220.7MB/7.69GB	2.80	0B/0B	26MB/8.67MB
airflow-scheduler	569.54	5.03GB/7.69GB	65.43	0B/0B	629MB/95.8MB
postgres	0.00	26.7MB/7.69GB	0.34	0B/0B	1.98Kb/0B
grafana	0.35	27.45MB/7.69GB	0.35	0B/0B	565KB/173Kb
Prometheus	0.00	73.41MB/7.69GB	0.93	0B/0B	3.81MB/276KB
statsd-exporter	2.23	48.21MB/7.69GB	0.61	0B/0B	1.34MB/3.64MB

4.3.3 Data quality

1) Completeness

In airline table has all 12 columns, 7 of 12 is not missing value, two of columns have more 95% completeness, one has more 85% completeness and two have more high missing value more than 50%, and overall have 8.3% missing cells.

In airport table has all 23 columns, 11 of 23 have no missing value. 3 columns have more than 85% completeness, 3 columns have more than 50% completeness and six have higher 50% missing rate. and overall have 4.4% missing cells.

In flight schedule table has all 22 columns, 12 of 23 have no missing value. One is no data, 5 columns have more than 95% completeness, 4 have higher 50% missing rate. and overall have 18.3% missing cells.

In flight status has all 29 columns, 17 of 29 no missing value. 1 column are all no data, 1 column is less than 25% completeness and 10 columns have more than 98 percent completeness. And overall have 6.9% missing cells.

2) Uniqueness

The results indicate that all datasets in the airline, airport, flight schedule, and flight status categories do not have duplicate rows. This suggests that the data is well-maintained and free from any redundant entries, ensuring accuracy and reliability for future analytical purposes.

3) Timeliness

To ensure the freshness and timeliness of the airport and airline data, a monthly update is scheduled, while the flight schedule and flight status tables are updated daily using a web scraping technique. By regularly fetching the latest information from reliable sources and incorporating batch data extraction is scheduled, the database remains up-to-date with accurate flight details.

4) Validity

In order to ensure data quality in validity, a series of checks are performed before inserting data into the "flightschedule" and "flightstatus" tables. This is accomplished through the implementation of the SQLColumnCheckOperator in the load Directed Acyclic Graphs (DAGs) from Airflow. The checks include verifying the existence of required columns and ensuring that they match the specified names. Additionally, null checks are performed, with null values equating to zero in certain columns. The date and time are validated in correct format and correspond to their respective data types. These data quality checks guarantee that only accurate and valid data is loaded into the destination database, minimizing errors and inconsistencies.

5) Accuracy

Data accuracy will be implemented through manual observation to verify data from the source and destination. This randomly check will ensure that the data between processes in the ETL (Extract, Transform, Load) process matches the valid data in the database, maintaining the correct meaning in rows and columns.

6) Consistency

All information stored in the "airline" table followed a consistent pattern, as did the data in the "airport" table. Flight information stored in each column exhibited coherence. Date and time information were consistently stored in the same format across all extracted tables. Delay times were recorded in minutes as float values. Additionally, year, month, day, and day of the week were stored in integer format, while boolean types were used for "cancelled," "diverted," "dep_del15," and "arr_del15."

4.4 Flight Delay Analysis

4.4.1 Flight delay dashboard

The Flight Delay Dashboard, developed using Power BI Desktop and sourced from the 'flightdata' database, provides insights into flight delays and answers various aspects of the aviation business. It offers valuable information related to flight delay patterns over time, flight status ratios and categorization, manual data filtering for in-depth analysis, financial impact of delay analysis, airport and airline delay analysis, crew and resource allocation, and customer experience improvement.

By analyzing and visualizing data, the dashboard assists airlines and airports in making informed decisions, optimizing operations, enhancing customer satisfaction, and strategically planning for improved performance. For example, Figure 3 illustrates the airline delay analysis within the dashboard, offering insights into flight volume, status, and average delay times for departures and arrivals of competitive airlines among the top flight traffics.

The visualizations provide insights for comparing delay data across airlines, which can offer valuable information regarding the operational performance of different players in the industry. This information can guide competitive strategies and benchmarking efforts.

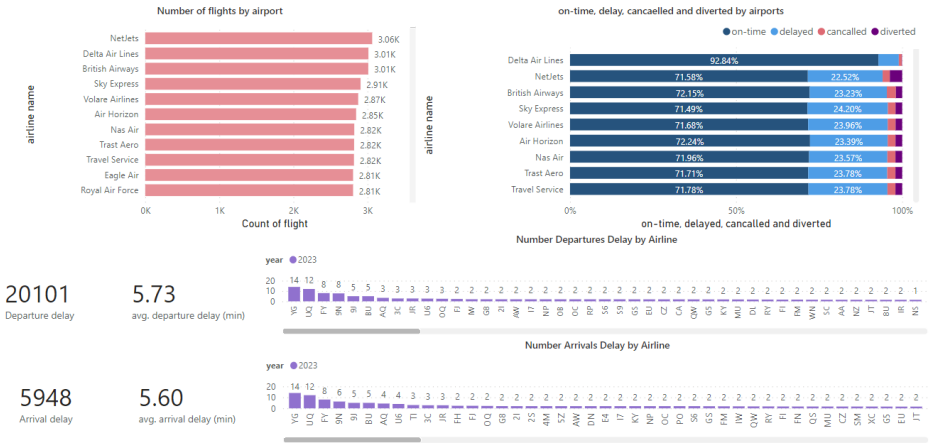

Figure 3: Airline delay information.

4.4.2 Flight delay prediction model

The flight delay prediction model guideline, based on Fabien Daniel's "Predicting flight delays" project, using "2015 Flight Delays and Cancellations" dataset with available source code on Kaggle [26]. The process followed by CRISP-DM methodology, encompassing stages from business understanding to model evaluation, can be automated through an Airflow DAG named "dag_analysis_flight_delay_prediction_model." This script-driven approach fetches data from “flightstatus” table, conducts exploratory analysis, performs feature engineering, trains regression models using sklearn, evaluates their performance, and selects the best model for deployment, as metric results shown in table 4. Scheduled execution via Airflow ensures model currency and accuracy by updating it every three months.

Based on the evaluation results presented in table 4 ($R^{2,}$ MSE, MAE and RMSE metrics), the Random Forest Regressor emerges as the best-performing model, exhibiting the lowest values for MSE (128.20), MAE (8.00), RMSE (11.32), and the highest R^2 (0.916) among all the tested models, which were trained and evaluated in July 2023.

Table 4: The regression models' evaluation results are from July 2023

Model	Testing set regression metrics			
	R^2	MSE	MAE	RMSE
Linear Regression	0.898	154.96	8.99	12.45
Boosted Linear	0.895	159.78	9.30	12.64
Bagged Linear	0.898	154.96	8.99	12.45
Lasso	0.897	157.13	9.04	12.54
Boosted Lasso	0.895	159.44	9.25	12.63
Bagged Lasso	0.897	157.12	9.04	12.53
Ridge	0.898	154.96	8.99	12.45
Boosted Ridge	0.895	159.18	9.27	12.62
Bagged Ridge	0.898	154.96	8.99	12.45
Random forest Regressor	**0.916**	**128.20**	**8.00**	**11.32**
Decision Tree Regressor	0.827	263.55	11.62	16.23

5 Conclusion and Discussion

5.1 Conclusion

This project has successfully implemented an end-to-end data pipeline system, orchestrated by Apache Airflow, for scheduled batch processing of tasks using Python scripts. The pipeline encompasses data extraction from reliable sources through an ETL process, followed by transformation and loading stages based on the flight delay prediction model guidelines. Monitoring approaches ensure data quality and performance. The analysis is performed through flight data visualization and flight delay prediction models, with the Random Forest Regressor emerging as the best-performing model.

5.2 Discussion

The project's focus on flight delay prediction introduces a challenge of limited domain knowledge in the aviation industry, an area in which I lack expertise. Having a shallow understanding of the aviation domain could potentially impact informed decisions during data processing and analysis, potentially affecting the accuracy and relevance of the prediction models.

This project has been developed with complexity and involves numerous stages, which has occasionally resulted in sections being less detailed or bypassed. While this approach may have been necessary to handle the complexity, it could potentially affect the overall understanding of the data pipeline and its processes. Future studies could consider offering a more comprehensive overview of each stage to improve clarity.

To dealing with multiple stages and processes introduces various technical challenges. Incompatibility issues with software versions and resource limitations, such as

low-performance computers and limited data storage, pose constraints on the efficiency of the data pipeline. Additionally, web scraping processes require high RAM and large storage for data collection and backup storage for further compound the technical challenges.

In section evaluation of data pipeline system, I have provided the real-time dynamic performance monitoring approach for the data pipeline, but it may not fully capture the pipeline's overall performance. Future studies could involve comparing different data pipeline design systems, considering factors such as time, resource utilization, and RAM performance across various environments to determine the most efficient and suitable system.

The section on data quality checks aims to ensure reliable data for analysis by implementing a data quality assessment across six dimensions. This assessment ensures that the dataset is suitable for further analysis. In comparison to missive BTS dataset, which relies on a single dataset from the US state, my project offers a more extensive and global perspective. By incorporating data from various years and locations worldwide, it reduces bias and provides a more diverse dataset for flight delay prediction, leading to widely applicable findings in the aviation domain. However, there are some issues related to data accuracy. Manual observation introduces the potential for human error in assessing accuracy, and the lack of a process to delete outdated data may impact data timeliness. Future improvements could involve developing automated methods for data quality assessment and addressing concerns related to data freshness.

The consumption stage of the end-to-end data pipeline involves the Flight Data Dashboard. This flight data visualization approach offers insights for data-driven decision-making. To enhance its utility, a deeper understanding of the business domain is necessary. Further exploration of dashboard design and interactive visualization techniques could improve the user experience and facilitate more in-depth data exploration. Obtaining feedback from business stakeholders to clarify requirements is also advisable. Additionally, adding more steps to the deployment process could enhance data accessibility and its applicability to real-world situations.

Flight Delay Prediction Models, the study compares various regression models for flight delay prediction without default hyper-parameter tuning during model training and result revealed that the Random Forest Regressor emerged as the best-performing model. When comparing with Mrs. Yogita Borse's work, her results using Decision Tree Logistic achieved an R2 score of 0.93, Regression achieved 0.92, and Neural Network achieved 0.91. In contrast, the Random Forest model in my project achieved an R2 score of 0.916, which is close and still demonstrates strong predictive accuracy.

In Jingyi Qu's study, they propose a flight delay prediction method based on Att-Conv-LSTM, and their results indicate that the prediction error of the Conv-LSTM model is reduced by 11.41 percent compared to the single LSTM, and the prediction error of the Att-Conv-LSTM model is reduced by 10.83 percent compared to the Conv-LSTM. In my project, the Random Forest Regressor achieved an RMSE value of 11.32256, which suggests strong predictive accuracy. Therefore, adding hyper-parameter tuning steps could lead to improved prediction accuracy and model efficiency.

References

1. Satria, P., & Setia, P. (2021). Development of Automated Flight Data Collection System for Air Transportation Statistics. Journal Title, 1863(1), 012020. doi:10.1088/1742-6596/1863/1/012020.
2. Rodrigo, B. A., Martin, D., & Augusto, V. (2012). The impact of flight delays on passenger demand and societal welfare. Journal Title, doi:10.1016/j.tre.2011.10.009.
3. Levy, E. (2021). "Streaming Data and Data Lake Architecture: The Ultimate Guide". Upsolver [accessed 2022 May 11] from https://www.upsolver.com/resources/whitepapers/streaming-data-and-data-lake-architecture-the-ultimate-guide.
4. Farah, J.C., Machado, J.S., Cunha, P.T.D., Ingram, S., & Gillet, D. (2021). An End-to-End Data Pipeline for Managing Learning Analytics. In Proceedings of the 2021 19th International Conference on Information Technology Based Higher Education and Training (ITHET), Sydney, Australia, 04-06 November 2021. IEEE. DOI: 10.1109/ITHET50392.2021.9759783.
5. Gaythorpe, K.A.M., Fitzjohn, R.G., Hinsley, W., Imai, N., Knock, E.S., Perez Guzman, P.N., Djaafara, B., Fraser, K., Baguelin, M., & Ferguson, N.M. (2023). Data pipelines in a public health emergency: The human in the machine. Epidemics, 43, 100676.
6. Sajida, S., M.C.A, M.Tech, M.Phil (2015). A Study of Extract–Transform–Load (ETL) Processes. International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, NCACI-2015 Conference Proceedings.
7. Turner, S. (2007). Defining and measuring traffic data quality: White paper on recommended. Transportation Research Record: Journal of the Transportation Research Board, pp. 62-69.
8. Jeusfeld, M., Quix, C., & Jarke, M. (1998). Design and analysis of quality information for data warehouses. In Proceedings of the 17th International Confer-ence on Conceptual Modeling.
9. Nayak A., Božić B., Longo L. (2022) Linked Data Quality Assessment: A Survey. In: Xu C., Xia Y., Zhang Y., Zhang LJ. (eds) Web Services – ICWS 2021. ICWS 2021. Lecture Notes in Computer Science, vol 12994. Springer, Cham. DOI: 10.1007/978-3-030-96140-4_5
10. Bhatia, B. (2018). Flight delay prediction: A project. Master's thesis, California State University, Sacramento. Department of Computer Science.
11. Choi, S., Kim, Y.J., Briceno, S., & Mavris, D. (2016). Prediction of weather-induced airline delays based on machine learning algorithms. In Proceedings of the 35th Digital Avionics Systems Conference (DASC).
12. Borse, Y., Jain, D., Sharma, S., Vora, V., & Zaveri, A. (2020). Flight Delay Predic-tion System. International Journal of Engineering Research & Technology (IJERT), 9(03), 123-129. Department of Information Technology, K.J Somaiya College of Engineering, Mumbai, India.
13. Qu, J., Xiao, M., Yang, L., & Xie, W. (2023). Flight Delay Regression Prediction Model Based on Att-Conv-LSTM. Entropy, 25(5), 770.

14. Krists Kreics. (2019). Quality of Analytics Management of Data Pipelines for Retail Forecasting. Thesis submitted for examination for the degree of Master of Science in Technology. School of Science, Espoo, 29.07.2019.
15. Kimball, R., & Ross, M. (2010). The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence. John Wiley & Sons.
16. [Online]. Available: https://research.aimultiple.com/scraping-techniques/
17. Seenivasan, D. (2023). Improving the Performance of the ETL Jobs. International Journal of Computer Trends and Technology, 71(3), 27-33. ISSN: 2231 – 2803. pp. 31
18. HarensLak, B., & de Ruiter, J. (2021). Data Pipelines with Apache Airflow. pp. 312-321
19. [Online]. Available: https://ydata-profiling.ydata.ai/docs/master/index.html
20. Askham, N. (October 2013). The six primary dimensions for data quality as-sessment: Defining data quality dimensions.
21. Thota, S., 2017. Big Data Quality. Springer International Publishing, Cham. pp. 1–5
22. Schwabish, J.A. (2021). Better Data Visualization: A Guide for Scholars, Researchers, and Wonks (p. 44).
23. Chapman, P., Clinton, J. & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. Published 2000. Computer Science. Corpus ID: 59777418.
24. [Online]. Available: https://airflow.apache.org/docs/apache-airflow/stable/administration-and-deployment/logging-monitoring/metrics.html
25. [Online]. Available: https://github.com/databand-ai/airflow-dashboards
26. [Online]. Available: https://www.kaggle.com/code/fabiendaniel/predicting-flight-delays-tutorial/notebook

Appendix

Appendix 1: Database schema