1 Introduction

In the contemporary landscape of data-driven decision-making, organizations worldwide are grappling with the ever-increasing volume, variety, and velocity of data. Data pipelines have emerged as indispensable components in data engineering, facilitating the seamless flow of information and empowering businesses to extract valuable insights from vast datasets. Amidst the complexities of data pipeline management, data queues play a pivotal role in ensuring the efficient, reliable, and scalable movement of data from producers to consumers.

As data pipeline architectures become more sophisticated, the strategic utilization of data queues gains paramount importance in optimizing data flow and preserving data integrity. However, the dynamic nature of data demands, coupled with the diversity of data queue technologies available, poses significant challenges for organizations seeking to identify the most effective data queue solutions tailored to their unique requirements.

2 Objectives

The aims of this project fall primarily on developing a data queuing model to analyse and optimize data pipelines.

3 Background

This brief background introduction touches upon the significance of data pipelines, introduces the role of queuing theory, and outlines the specific focus of the master's thesis. It provides readers with a concise overview of the subject matter, setting the tone for the subsequent sections of the thesis.

With the steady increase of datafication, that has been estimated to double very 1,5 years [OLBO15], increases the need for structures that are capable of swiftly and efficiently, pull, save and gather data for analysis and derive insights from such newly sourced data.

Therefore, multiple tools were developed to deal with these new requirements such as data pipelines. Data pipelines provide the foundation for a range of data projects which can range from exploratory data analyses, data visualizations and machine learning tasks. [IBM] However, with the total volume of data predicted grow rates, ensuring the continuous and error-free is a significant challenge.

As data pipelines performance depends on multiple variable factors, queueing theory has emerged as the mathematical framework to analyse and address these new challenges. Queueing theory, is the mathematical study of queues, for systems with a steady inflow of entries (customers) and a limited number of servers where the analyst wants to know if the current system is capable of serving all the inflow demands. The aim is to calculate the multiple queue performance metrics. [Tho12]

In this research

4 Model Key assumptions

For the analysis of our case study, I have selected a M/D/1 queuing model, whose assumptions best fit the our case study. Starting from the M"in the first slot stands for memoryless" [HB14], in this model interval of arrival of new entries into the queue (customers) follow a Poisson distribution , the D"in the second slot stands for deterministic, where the server processes each costumer taking a fixed deterministic amount of time, idling when there are no queuing customers [ANT21], the "1"in the third slot represents the number of servers, which is this case is a single server and the forth slot is empty to represent that this queue has an infinite capacity and follows a First-Come-First-Served (FCFS) policy [HB14].

5 Methodology Overview

5.1 Research Design

Describe the overall approach you took to address your research question. Is it an experimental study, a case study, a survey, or another type of research design?

My case study start by designing a prototype of a data pipeline that simulates the streaming of critical heart health metrics data that are then processed and stored in the hospital data centre to be further analysed.

The data processed in this case study is real data collected from the hospital of TODO CITATION.

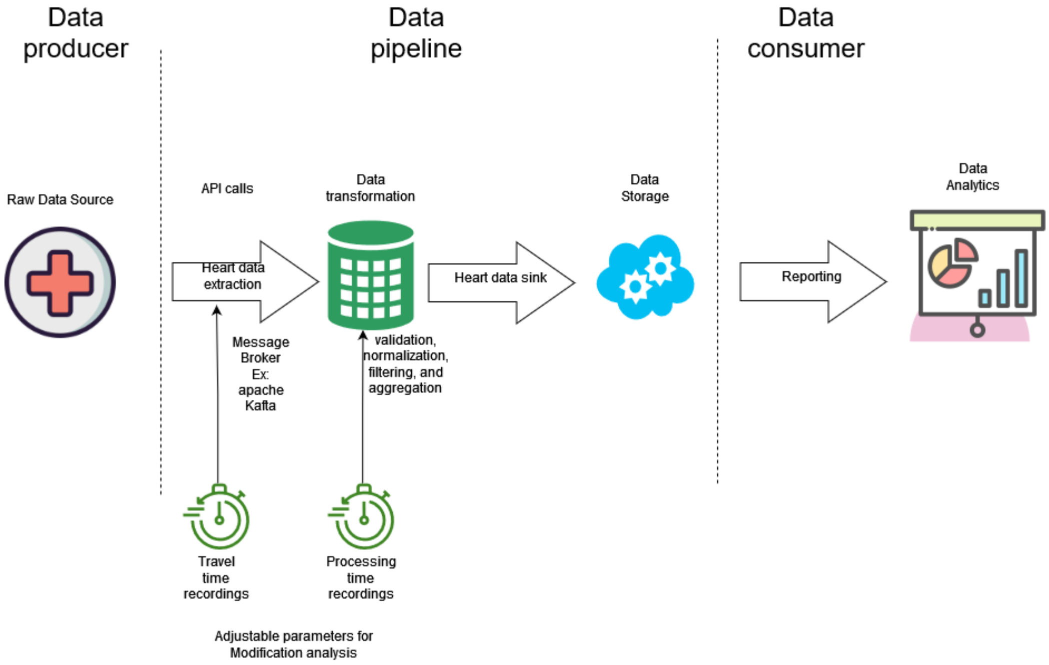
The data processing is divided into 3 steps, the data producer, which is responsible for feeding the heart data to the data pipeline, emulating a patient whose data is being recorded for a diagnosis and sending it to the second step, the data broker. The data broker is responsible for receiving the data from multiple patients and emulate central control system that will receive all requests and will forward each request to the correct service handler, sending the heart data to the third and final step. On this final step the data of the patience, as well as the data from the metrics collected during the life cycle of this customer will be stored for future analysis.

Through the processing of this data I am able to record the necessary performance metrics to analyse this data pipeline from a data queueing perspective.

5.2 Data Collection

Detail how you collected the data for your study. Specify the sources, methods, and instruments you used, whether it’s surveys, interviews, observations, experiments, archival research, etc.

In the case study, data collection happens at the entry and exit of each of our 3 steps.



Figur 1: Enter Caption

5.3 Data Analysis

Describe the techniques you used to analyze the collected data. This could involve qualitative analysis (thematic analysis, content analysis) or quantitative analysis (statistical tests, modeling).

5.4 Tools and Software

6 Underlying Theory

6.1 Literature Review Recap

6.2 Theoretical Framework

6.3 Research Question

7 Thesis Structure

The structure of this thesis is as follows: Chapter 2 presents an extensive review of literature on data pipelines, data queue technologies, and related studies in the UK context. Chapter 3 outlines the research methodology employed in this study. Chapter 4 provides a detailed analysis of data queue usage and its impact on data pipeline efficiency based on real-world case studies. Chapter 5 presents optimization strategies and best practices for data queue utilization. Finally, Chapter 6 concludes the thesis by summarizing the key findings, discussing their implications, and proposing avenues for future research.

Through this research, we endeavor to contribute to the advancement of data engineering practices in the UK and empower organizations with actionable insights to optimize their data pipelines through strategic data queue utilization.

8 Analyzing Data Pipeline Efficiency through Data Queue Utilization

In this chapter, we delve into the practical implementation of data queue usage within data pipelines. The implementation phase focuses on deploying data queues as integral components of data pipelines and configuring them to optimize data flow, ensure reliability, and enhance overall data pipeline performance.

8.1 Data Pipeline Architecture

Our data pipeline architecture is designed to handle the complex data processing needs of modern organizations. It encompasses data ingestion, data transformation, data enrichment, and data delivery stages. Data queues are strategically integrated at key points in the pipeline to facilitate smooth data movement and improve data processing throughput.

8.2 Selecting the Data Queue Technology

The selection of the most suitable data queue technology is crucial for the success of our analysis. After an extensive evaluation of data queue technologies available in the UK, we opted for Apache Kafka, a distributed messaging system renowned for its scalability and low-latency capabilities. Kafka's ability to handle high-throughput data streams and support real-time data processing aligns perfectly with the demands of our data pipeline analysis.

8.3 Data Queue Configuration for Analysis

Configuring the data queues to serve our analysis goals is essential. We employed Kafka's partitioning feature to parallelize data processing and achieve load balancing during performance tests. Additionally, we enabled detailed logging and monitoring to capture valuable data queue metrics, including message throughput, latency, and consumer lag.

8.4 Integration of Data Queueing with Data Pipeline Analysis

Data queues play a central role in our data pipeline analysis framework. We instrumented the pipeline components to publish and consume data through Kafka topics, enabling us to capture fine-grained data flow details. Furthermore, we used Kafka Connect to capture the broker data from the data pipeline.

8.5 Maximum Likelihood Estimation

This prototype has one data source that has generated messages for T time for the M/D/1 data queue that will enter be processed inside the queue.

These messages are defined by the random variable X.

The variable X follows a sequence of independent and identically distributed(i.i.d) random variables that can be defined as,

$$X = (X_1, X_2, X_3, ..., X_n)$$

,

From previous work of TODO CITE THIS: https://doi.org/10.1080/03610926.2014.950750 page 5827-5828

we can easily derive that the sum of the i.i.d random events can be defined from a Poisson distribution where we can define the probability of X as,

$$P(X = x) = \frac{e^{-\rho}\rho^x}{x!}, \quad x \in \mathbb{N}_0$$

TODO

$$\max \mathscr{L}(\rho; x \sim) = -n\rho + \sum_{i=1}^{n} x_i \cdot \log \rho - \log \prod_{i=1}^{n} x_i!$$

8.6 Real-Time Data Analysis with Data Queues

Incorporating real-time data analysis into the pipeline analysis is a critical aspect of our implementation. We explored Kafka Streams, a powerful library for real-time data processing, to conduct continuous analysis on data flowing through the data queues. This facilitated instant insights and feedback on the data pipeline's performance.

8.7 Case Study: Enhancing Data Pipeline Efficiency with Data Queue Analysis

In this section, we present a detailed case study of a UK-based organization that utilized data queueing to analyze and optimize their data pipeline. We highlight the challenges faced, the data queue configuration employed, and the resulting improvements in data pipeline efficiency.

9 Limitations

Throughout the course of this research on data queue utilization in data pipeline analysis, several limitations have been identified that may impact the scope and accuracy of the findings:

9.1 Limited Computational Power

As a student conducting this research with access to university resources, the computational power available for conducting extensive simulations and performance evaluations may be limited. Large-scale simulations with high data volumes and complex data pipeline configurations may not be feasible due to computational constraints, potentially leading to a restricted exploration of certain scenarios.

9.2 Limited Knowledge or Research on Closed Form Expressions

The probability of n customers in a data queue is a fundamental metric in data queueing theory. However, due to limited prior research and theoretical knowledge in this specific context, it may not be possible to derive closed form expressions for the probability of n customers in our data pipeline scenario. This limitation may require reliance on numerical methods or simulation-based approaches for estimating queue performance metrics.

9.3 Consideration of Traffic Intensity Parameter

Analyzing data queue utilization effectively requires considering the traffic intensity parameter, which combines both the arrival rate and the service rate. Estimating the accurate traffic intensity for our data pipeline becomes challenging, as it may vary based on unpredictable data fluctuations and processing workloads. Consequently, the analysis might involve certain assumptions or approximations that could impact the accuracy of the results.

9.4 Inaccuracy in Recording Processing and Service Time

Recording the processing time and service time of data queues, especially in a real-world environment, may be subject to inaccuracies or measurement errors. The precision of timestamps and the synchronization of monitoring tools might not be perfect, leading to potential inaccuracies in data collection and performance measurements.

9.5 Limited Generalizability of Results

The data pipeline architecture and data queue utilization explored in this research are based on specific configurations and use cases. Therefore, the findings and optimization strategies may not be universally applicable to all data pipeline scenarios. The generalizability of the results may be limited to similar data pipeline architectures with comparable traffic characteristics.

9.6 Complexity of Real-World Environments

Real-world data pipeline environments can be highly complex, involving interactions with various external systems, data sources, and business logic. This research might not encompass all aspects of these complexities, potentially affecting the completeness of the analysis and the ability to capture every relevant factor that impacts data queue utilization.

Referenser

[ANT21] Azam Asanjarani, Yoni Nazarathy, and Peter Taylor. A survey of parameter and state estimation in queues. *Queueing Systems*, 97(1–2):40, 2021.

[HB14] Mor Harchol-Balter. *Performance modeling and design of computer systems: Queueing theory in action*, page 236. Cambridge University Press, 2014.

[IBM] IBM. What is a data pipeline.

[OLBO15] Peter O'Donovan, Kevin Leahy, Ken Bruton, and Dominic TJ O'Sullivan. An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities. *Journal of big data*, 2(1):1–26, 2015.

[Tho12] Nick T. Thomopoulos. *Fundamentals of queuing systems: Statistical methods for analyzing queuing models*. Springer, 2012.