**Customer Churn Prediction System with Data Integration**

**Martin Gitau Kimani**

**100994**

**An Information System Project 2 proposal submitted to the School of**

**Computing and Engineering Sciences in partial fulfilment of the requirements**

**For the award of the bachelor’s degree in Business Information Technology of**

**Strathmore University**

**School of Computing and Engineering Sciences**

**Strathmore University**

**Nairobi, Kenya.**

**December 2023**

# Declaration and approval

I hereby declare that this work has not been previously submitted and approved by this or any

Other University. To the best of my knowledge and belief, the report contains no material

Previously published or written by another person except where due reference is made in the

Research proposal itself.

Student’s signature

Sign: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Supervisor’s signature

Sign: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Abstract

In the contemporary information technology landscape, data has evolved into a valuable asset, often likened to the significance of oil. Recognizing data's pivotal role in organizational success, effective utilization has become imperative for staying competitive and creating value. However, the raw nature of data requires transformation into meaningful insights, a task typically undertaken by professionals such as Data Analysts, BI Analysts, and Data Scientists. This underscores the essential need for robust data integration within businesses to extract actionable insights.

Customer churn, the phenomenon where consumers disengage from online products or services, poses a significant challenge for businesses, resulting in reduced revenue and sluggish growth. While this phenomenon may lack a straightforward explanation, leveraging data can provide crucial insights into understanding and predicting customer churn. This emphasizes the central focus of ensuring proper data integration in business operations, utilizing data to not only comprehend the issue but also to anticipate and mitigate the occurrence of customer churn. Creating a data eco system that will allow for the support a customer churn predictive system is paramount.

Table of content

[Declaration and approval ii](#_Toc153623812)

[Abstract iii](#_Toc153623813)

[Table of content iv](#_Toc153623814)

[List of Figures vii](#_Toc153623815)

[Chapter 1: Introduction 1](#_Toc153623816)

[1.1 Background 1](#_Toc153623817)

[1.2 Problem Statement 2](#_Toc153623818)

[1.3 Objectives 3](#_Toc153623819)

[1.3.1 General Objective 3](#_Toc153623820)

[1.3.2 Specific objectives 3](#_Toc153623821)

[1.4 Research Questions 3](#_Toc153623822)

[1.5 Justification 3](#_Toc153623823)

[1.6 Scope and limitations 4](#_Toc153623824)

[Chapter 2: Literature Review 5](#_Toc153623825)

[2.1 Introduction 5](#_Toc153623826)

[2.2 Problem of Data integration 5](#_Toc153623827)

[2.2.1 State of data integration 6](#_Toc153623828)

[2.2.2 Data integration methods and techniques 7](#_Toc153623829)

[2.2.3 Data Warehousing 8](#_Toc153623830)

[2.2.4 Causes of customer churn 8](#_Toc153623831)

[2.2.5 Customer purchase behavior 8](#_Toc153623832)

[2.2.6 Predictive analysis 9](#_Toc153623833)

[2.2.7 Develop Customer Churn Predictive Analysis and implementing data pipelines 10](#_Toc153623834)

[2.3 Related works 12](#_Toc153623835)

[2.3.1 Machine Learning for Environmental Data Integration in Conservation Initiatives 12](#_Toc153623836)

[2.3.2 DataOps 13](#_Toc153623837)

[2.3.3 Customer churn prediction in telecom using machine learning in big data platform 13](#_Toc153623838)

[2.4 Gaps Customer churn with data integration 15](#_Toc153623839)

[2.5 Conceptual Frame work 16](#_Toc153623840)

[2.5.1 Generation 16](#_Toc153623841)

[2.5.2 Ingestion 17](#_Toc153623842)

[2.5.3 Transformation 17](#_Toc153623843)

[2.5.4 Storage 18](#_Toc153623844)

[2.5.5 Serving for Machine Learning / Data Analytics: 19](#_Toc153623845)

[2.5.6 Underlying Factors 20](#_Toc153623846)

[Chapter 3: Methodology 22](#_Toc153623847)

[3.1 Introduction 22](#_Toc153623848)

[3.2 Development Approach 22](#_Toc153623849)

[3.3 Justification of the Methodology 22](#_Toc153623850)

[3.4 System Analysis 22](#_Toc153623851)

[3.4.1 Context Diagram 23](#_Toc153623852)

[3.4.2 DFD Level 0, 1 23](#_Toc153623853)

[3.4.3 ERD (Entity-Relationship Diagram) 24](#_Toc153623854)

[3.4.4 Database Schema 24](#_Toc153623855)

[3.5 System Design 24](#_Toc153623856)

[3.6 System Deliverable/Milestones 25](#_Toc153623857)

[Deliverables 25](#_Toc153623858)

[Milestones 26](#_Toc153623859)

[3.6.1 System proposal 27](#_Toc153623860)

[References 28](#_Toc153623861)

[Appendix 30](#_Toc153623862)

List of Figures

[Figure 1‑1 Data Engineering life cycle 2](#_Toc153623108)

[Figure 2‑1 Components of ETL process 7](#_Toc153623109)

[Figure 2‑2 Conceptual Frame work 16](#_Toc153623110)

[Figure 0‑1 Gant Chart 32](#_Toc153623111)

# Introduction

## Background

Data integration is a big part of businesses as data has become a key factor in business growth. It may seem that every business should jump onto the Data analytics and Data science band wagon in order to tap into insights they produce, but this can be detrimental and may yield insufficient results (Joe Reis, 2022). This is because data integration is important for businesses and depending on the level of business or organization its data maturity stage may differ. Data maturity stages may vary. These stages are; starting with data, Scaling with data and Leading with data from start-ups to big companies respectively (Joe Reis, 2022). So this means it is very important for data engineers to know at what stage of maturity the organization is for proper data integration. For this scenario we will be looking at an ecommerce problem of customer churn.

Data engineering can be defined as implementation and maintenance of systems that allow for raw data intake and production of consistent information that is of high quality that supports downstream use cases such as analysis and machine learning, also alongside this it would involve data management, DataOps and software engineering (Joe Reis, 2022). All this is essential for the integration of data in any company. Nowadays it is seen that data analyst and data scientist are spending most of their time around 70% and 80% of their time cleaning and organizing data. This is due to failure of implementing proper data infrastructure from immature data engineering and data science practices (Ralph Kimball, 2019).

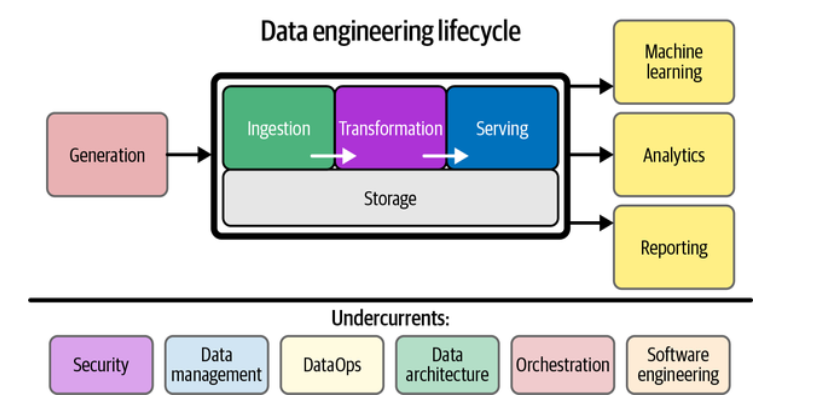
This can be approached by implementing and focusing on the data engineering life cycle that deals with the data engineering lifecycle. Let’s for example a cosmetics company that has thousands of users and are active on social media. The company draw data from the various social media platforms, store them and serves the data for downstream use cases then provide a visualization with tools like power BI to provide a live feed of their product and how it is performing based on what users are saying about it on social media. This project aims to underscore the pivotal role of data integration within companies, with a specific focus on addressing the customer churn challenge. Our approach involves implementing a robust data pipeline designed to seamlessly gather data from diverse sources, transform it into a structured format, and conduct thorough exploratory and analytical processes. Leveraging machine learning techniques, the project aims to predict customer churn, highlighting how strategic data integration plays a crucial role in deriving actionable insights for informed decision-making.

Figure 1‑1 Data Engineering life cycle

## Problem Statement

Customer churn, a prevalent challenge in business, refers to the phenomenon where customers discontinue their association with a product or service (Nonis, 2022). In the context of our exploration, it poses a significant threat to organizational revenue and growth. Identifying and mitigating customer churn is crucial for sustaining a healthy and thriving business. In tackling the issue of customer churn, data integration emerges as a pivotal framework. By seamlessly consolidating and harmonizing data from various sources, data integration provides a holistic view of customer interactions, behaviors, and preferences. This unified data landscape becomes the foundation for implementing advanced analytics and machine learning algorithms.

With a robust data integration framework in place, predictive analytics can be employed to anticipate customer churn. By analyzing historical customer data and identifying patterns or indicators that precede churn, organizations can proactively take measures to retain valuable customers (Momin, 2020). Efficient data integration enables real-time monitoring and analysis of customer interactions. Streaming data from diverse sources, such as user engagement on platforms or social media sentiments, allows organizations to stay abreast of evolving customer sentiments.

This dynamic insight, integrated through data pipelines, empowers swift decision-making in response to emerging churn risks. Data integration also addresses the challenge of inaccurate insights caused by disparate data sources. By unifying data through a well-orchestrated integration process, organizations can ensure the accuracy and reliability of the insights derived, further enhancing the effectiveness of customer churn prediction models. In essence, data integration serves as the linchpin in proactively addressing customer churn. By leveraging a unified data landscape, organizations can not only comprehend the nuances of customer behavior but also employ advanced analytics to predict and preemptively respond to potential churn, thereby fostering customer retention and business sustainability.

## Objectives

### General Objective

Creating data pipelines involves efficiently collecting and preprocessing diverse data sources for the subsequent creation and execution of customer churn prediction models

### Specific objectives

1. To investigate data integration methods, techniques and investigating challenges in its implementation
2. To analyze customer purchase behaviors and data in online services
3. To design and develop a predictive model for customer churn using a functional streaming data pipeline
4. To test the functionality and effectiveness of the customer churn predictive model by evaluating its performance within the streaming data pipeline.

## Research Questions

1. How do various factors contribute to customer churn within the industry
2. In which way do various factors influence customer purchasing decisions
3. How can we address challenges to effective data integration
4. How are various data science techniques and algorithms used in predictive analysis
5. How can we implement current technologies to implement data pipelines

## Justification

In the absence of an automated system that is not capable of predicting customer churn and facilitating seamless data integration, business will be left vulnerable to significant risks. Extensive research by (Argelaguet, 2021) has shown that companies without robust predictive analytics systems often experience a notable surge in customer churn rates. Also the lack of efficient data integrating solution can lead to data silos that hinder critical decision making processes. This greatly undermines the organization’s competitiveness

## Scope and limitations

This project will only cover the data pipelines for cases of medium sized companies will not include big data and IOT devices. It’s possible to integrate big data and data from IOT devices in future but for case it is focusing on small to medium ecommerce business.

# Literature Review

## Introduction

In this chapter we will dive into some of the existing body of knowledge in the field of data integration and data engineering. The aim of the chapter is to provide a more comprehensive understanding of the current state of data engineering and data integration along with the challenges it presents. We will also look at the current solutions that have been put in place or proposed. Customer churn will be another problem we will look at and try to understand it much better as it will be our downstream use-case that will allow see how the whole process of data engineering fits in and how effective it can be in developing predictive system for customer churn.

## Problem of Data integration

The process of data integration, which involves extracting data from various sources and transforming it into a standardized and analyzable format, can be fraught with challenges. In many cases, there are no established best practices for efficient data integration, leading to a need for more effective processes (Hira, 2023). Industries and research institutions often find that their data storage infrastructure is insufficient to handle the volume and variety of data they generate. This necessitates advancements in workflows, including data storage, data management, data maintenance, data integration, and data interoperability (Joe Reis, 2022).

Among these, data integration and data interoperability are of particular importance for organizations seeking to enhance their workflows. However, these challenges can become particularly complex for companies or organizations planning to implement big data architectures. This is because the data they work with is often heterogeneous in nature (J. M., 2021). The complexity of data integration is further exacerbated by the need to integrate data from various sources, each with its own unique characteristics and formats. This can lead to issues with data consistency, quality, and reliability, which can in turn impact the effectiveness of data-driven decision-making processes (Joe Reis, 2022).

Despite these challenges, advancements in data integration technologies and practices continue to evolve, providing organizations with more effective tools and methods for managing and leveraging their data (Joe Reis, 2022).

### State of data integration

Currently, many companies may not fully grasp the vital importance of data integration in their data analytics and data science endeavors. This oversight becomes apparent when organizations hastily adopt popular data analysis and data science practices, believing they will yield immediate benefits. In such cases, a disconcerting pattern emerges where data scientists and analysts within these companies, which have not integrated robust data integration practices, find themselves disproportionately occupied with the arduous task of data cleaning and reformatting before they can even commence meaningful analytical computations (Murphy, 2022).

This prevailing scenario underscores a profound issue: the failure to appreciate data integration's foundational role in the data ecosystem. It is not merely a preparatory step; it is the linchpin upon which the edifice of data-driven insights and informed decision-making rests. Without adequate data integration, data remains a disparate and unwieldy assortment, a far cry from the harmonious, structured dataset necessary for sound analytics. Consequently, organizations inadvertently squander substantial resources, both in terms of time and human capital, on the avoidable "data wrangling" stage.

The consequences of neglecting data integration are manifold. Not only does it impede the pace of data-driven projects, but it also compromises the quality of insights and, by extension, the accuracy of decision-making. The value that data scientists and analysts bring to the table is undercut, as they find themselves bogged down by data-related drudgery rather than focusing on the core analytics and innovation.

Therefore, it is imperative for organizations to recognize that data integration is not an optional component but a fundamental prerequisite for meaningful data utilization. By embracing robust data integration practices, businesses can streamline their data workflows, unlock the full potential of their data assets, and empower their data experts to engage in value-adding activities. In doing so, they position themselves to harness the true benefits of data analytics and data science.

As of the moment most companies are not aware how crucial data integration is. This can be seen in instances where companies jump on the band wagon of the most popular data analysis and data science practices that have been seen to have very beneficial for some organization. What will happen is that the data scientist and analysts that are employed in these organization which have not incorporated any kind of data integration practices will spend most of their time trying to clean the data and format it to proper structure before performing any analytical calculation on them

### Data integration methods and techniques

Understanding effective data integration techniques is crucial for the development of efficient data pipelines. There are two primary approaches to achieve this

#### ETL

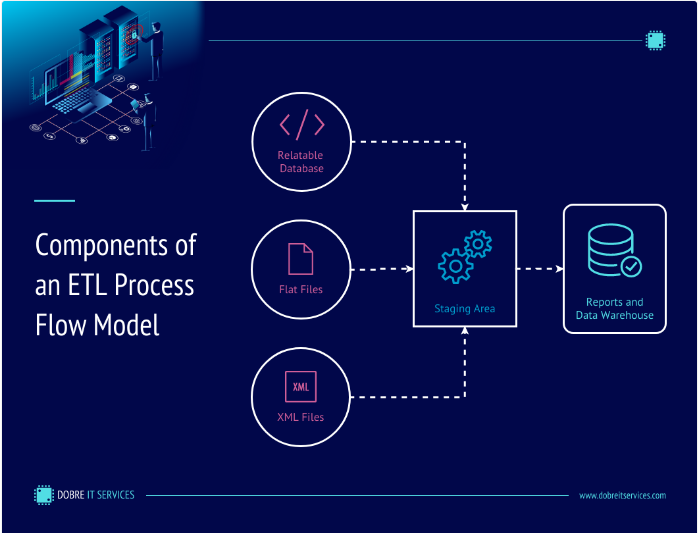


Figure 2‑1 Components of ETL process

First we have ETL Processes, which stands for Extract, Transform, Load, are fundamental processes in data integration. This involves the extraction of raw data from source systems, transforming it into a usable format, and loading it into the target system. Tools like Talend and Informatica are widely employed for ETL processes. These tools streamline the data transformation journey, making data ready for analysis and other downstream operations.

### Data Warehousing

Secondly we have Data warehousing where it focuses on centralizing data from various sources into a data warehouse. This centralized repository is designed for analysis and reporting purposes. Solutions like Amazon Redshift and Snowflake are popular choices for data warehousing. These platforms offer scalability and high performance, making them well-suited for handling and analyzing large volumes of data (Bhatia, 2019). These two approaches, ETL processes and data warehousing, are critical components of efficient data integration strategies. They enable organizations to consolidate, transform, and analyze data from diverse sources, empowering data-driven decision-making and insights (Bhatia, 2019)

### Causes of customer churn

Understanding the factors contributing to customer churn is a critical endeavor in enhancing customer retention. Various methodologies have been employed to delve deeper into this issue. One such approach involves conducting customer surveys, where companies and organizations gather valuable feedback from their clientele. These surveys serve as a means to identify pain points and unearth the underlying reasons that drive customers away. By carefully analyzing the responses, recurring themes and issues can be uncovered, offering actionable insights for improvement (Çelik, 2019).

Another method, cohort analysis, plays a pivotal role in comprehending customer behavior. This technique involves segregating customers into distinct groups based on specific characteristics, such as when they signed up or how they use a product (Fedushko, 2022). By examining these cohorts, organizations can identify patterns associated with churn. This data-driven approach provides a means to understand why certain customer groups are more susceptible to churn, enabling proactive measures to be taken to address the issue. These methods collectively empower organizations to gain insights into customer churn, fostering the development of effective strategies for customer retention and overall business enhancement.

### Customer purchase behavior

This endeavor involves a detailed analysis of customer purchasing behavior in online platforms, scrutinizing various elements like user interactions, buying patterns, and conversion rates. There are established methods to tackle this challenge effectively. One approach leverages tools such as Google Analytics and Mix panel, which provide valuable insights into user behavior on websites and mobile applications. These tools enable organizations to gain a comprehensive understanding of how users interact with their platforms, identify trends in purchasing behavior, and assess the efficacy of conversion strategies (McGuirk, 2023).

Another solution in the arsenal is A/B testing, a method to evaluate the impact of changes on user behavior. This is achieved by comparing the performance of two or more versions of webpages or apps. Organizations can systematically analyze which design or content elements lead to improved conversion rates, allowing for data-driven decisions in enhancing user experiences and achieving higher conversion rates (Yu, 2020). These strategies collectively empower businesses to optimize their online platforms and bolster customer engagement.

Within the existing strategies for analyzing customer behavior on online platforms, a critical gap lies in the absence of a dedicated mechanism for predicting and addressing customer churn. The challenge of foreseeing potential customer departures before they occur is a crucial aspect that hasn't been adequately addressed in the current framework. Our project seeks to bridge this gap by introducing a customer churn prediction system integrated with ETL (Extract, Transform, Load) pipelines. These pipelines will facilitate the seamless collection, transformation, and loading of data, enabling organizations to harness valuable insights. By leveraging historical data and employing predictive analytics, our system aims to proactively identify patterns and indicators that precede customer churn.

The integration of ETL pipelines ensures that the predictive models are fueled with accurate and timely data, enhancing their efficacy. This comprehensive solution empowers organizations to not only anticipate customer churn but also implement targeted strategies to retain valuable customers. In essence, our project endeavors to enhance the existing framework by providing organizations with a proactive and integrated approach to customer churn mitigation.

### Predictive analysis

In this phase, the project harnesses advanced machine learning algorithms, including decision trees, logistic regression, and random forests, to delve deep into the data and make accurate predictions, particularly focused on customer churn. These algorithms excel in identifying intricate patterns and relationships within the data, providing a foundation for highly precise predictions (Fedushko, 2022). For instance, decision trees dissect data into branches of conditions, logistic regression evaluates the probability of a specific event like customer churn, and random forests aggregate the power of multiple decision trees to enhance accuracy. This ensemble of algorithms collectively forms a robust framework for making informed predictions regarding customer churn (Fedushko, 2022).

Additionally, the project employs statistical modeling techniques such as survival analysis and regression models to gain comprehensive insights into the factors influencing the likelihood of customer churn (Fedushko, 2022). Survival analysis, specifically adept at assessing the time to an event, is instrumental in understanding when and why customers are likely to churn. On the other hand, regression models establish relationships between various factors and the propensity for customer churn, further enhancing the project's ability to provide insightful predictions (Fedushko, 2022).

In the context of the project's focus on customer churn prediction, these sophisticated algorithms and statistical modeling techniques contribute to a nuanced understanding of the variables influencing customer behavior. By leveraging these tools, the project aims to provide organizations with actionable insights that go beyond simple predictions, enabling them to implement targeted strategies for customer retention effectively.

### Develop Customer Churn Predictive Analysis and implementing data pipelines

In the endeavor to implement comprehensive data pipelines and cater to downstream use cases, organizations often depend on data integration platforms and tools as linchpins of their data strategy. These platforms play a multifaceted role in the data lifecycle and data engineering lifecycle, particularly in streamlining the Extract, Transform, Load (ETL) processes, and ensuring that data is not only efficiently processed but also readily accessible for analysis (Joe Reis, 2022).

Data integration platforms offer a robust infrastructure that simplifies the intricacies of ETL. They provide the means to extract data from a myriad of sources, ranging from databases to cloud services, and subsequently transform this data into a standardized and analyzable format (Joe Reis, 2022). The transformation phase is vital, as it makes the data usable for a variety of applications, including analytics and machine learning. Moreover, these platforms ensure the seamless loading of processed data into the target systems, which can encompass data warehouses, databases, or other analytics tools. This strategic approach sets the stage for informed decision-making, by rendering data insights accessible to business analysts, data scientists, and other stakeholders (Ralph Kimball, 2019).

In the broader context, the seamless integration of data platforms not only streamlines ETL processes but also establishes a foundation for advanced analytics, including machine learning. As organizations extract, transform, and load data through these platforms, the standardized and enriched datasets become instrumental for conducting sophisticated machine learning analyses (Ralph Kimball, 2019). Specifically, when it comes to predicting customer churn, this integrated approach ensures that the data required for machine learning models is prepared, consistent, and readily available.

The transformation phase within the data integration platforms plays a crucial role in shaping the data into a format suitable for training predictive models. This enriched data, free from inconsistencies, serves as the fuel for machine learning algorithms, allowing organizations to derive meaningful insights into customer behavior and make informed predictions about potential churn. Therefore, the utilization of data integration platforms is not just about optimizing ETL processes; it becomes a strategic enabler for organizations to seamlessly transition from data processing to advanced analytics, ultimately enhancing their capability to predict and address customer churn effectively.

## Related works

We will now look at some of the technologies and practices that have been put in place to tackle the need for data integration

### Machine Learning for Environmental Data Integration in Conservation Initiatives

Machine Learning for Environmental Data Integration in Conservation Initiatives, delves into the Santos Project's approach to addressing challenges in integrating environmental data for conservation and monitoring. The project proposed the use of machine learning techniques, specifically Random Forest (RF) for univariate data and a hybrid strategy combining Self-Organizing Map (SOM) and RF for multivariate data (Fonseca, 2023). They address challenges such as non-linearity, covariation, interactive effects, missing or noisy data, model optimization, accurate predictions, and imbalanced observations.

It emphasizes the critical need for conservation efforts in the face of environmental challenges. The project highlight the complexities of integrating diverse environmental data and outline the analytical challenges, which are contextualized within the Santos Project aimed at understanding the spatio-temporal dynamics of benthic, pelagic, and physical systems (Fonseca, 2023). The machine learning techniques highlighted in the project, RF for univariate data and a hybrid Self Organizing Map SOM - Random Forest (RF) strategy for multivariate data, are justified based on their ability to handle non-linearity, covariation, and interactive effects (Fonseca, 2023). The project emphasizes the advantages of RF, such as its ability to deal with non-linear data and variable importance analysis. For multivariate data, the combination of SOM and RF is presented as an effective approach.

Project touches the importance of addressing missing values, highlighting the superiority of the bagging imputation technique in handling missing or noisy data. The resilience of the machine learning techniques against noisy data is noted, and the identification of noisy data based on model outputs is highlighted. In the context of imbalanced data sets, the project explores the correlation between the overall statistics of the RF model and those of individual classes. This joint interpretation is deemed valuable for understanding model limitations and discussing environmental mechanisms shaping observed patterns.

They showed two analytical workflows that enable the exploration and enhancement of model accuracy while facilitating the investigation of cause-and-effect relationships in the data. These workflows are positioned as foundational for implementing long-term learning algorithms, crucial for monitoring initiatives. This is one of the cases we see how data integration is vital when it comes to analytical processing of data and machine learning

### DataOps

This is a new innovative approach to data integration that emphasizes on collaboration, automation and continuous delivery in the context of data-related processes (Munappy, 2020). It is a set of principles that aim to improve the collaborating and communication between data engineers, data scientists and other data professionals while also enhancing the efficiently, quality and agility of data related processes. It draws its inspiration from DevOps which its focus is improving software development and IT operations.

Some of the key elements of DataOps include close collaboration between various teams involved in data such as data engineers and data scientists to foster better communication and alignment of goals. Automation is another aspect that involves the automation of data related tasks such as data ingesting transformation and deployment to minimize manual effort and reduce chances of errors (Munappy, 2020). The other is continuous delivers in promoting the concept of continuous delivery of data, which means making data available for analytics, reporting and decision making in real-time or near real-time. Lastly is monitoring of feedback that is done for data pipelines, workflows and data quality and providing timely feedback for continuous improvement (Munappy, 2020).

### Customer churn prediction in telecom using machine learning in big data platform

The project addressed the significant issue of customer churn, particularly in the tele-communications industry, where retaining customers is crucial for revenue stability. The aim was to develop an efficient churn prediction model using machine learning techniques on a big data platform, focusing on feature engineering and selection. The research utilized a vast dataset from SyriaTel Telecom Company, spanning nine months and comprising structured, semi-structured, and unstructured data formats, totaling about 70 Terabytes on the Hadoop Distributed File System (HDFS). The project leveraged the Apache Spark environment to experiment with four machine learning algorithms: Decision Tree, Random Forest, Gradient Boosted Machine Tree (GBM), and Extreme Gradient Boosting (XGBOOST). The XGBOOST algorithm yielded the best results for churn prediction classification.

One notable contribution of the project was the incorporation of Social Network Analysis (SNA) features extracted from customer social networks into the prediction model. The SNA features enhanced the model's performance significantly, improving the Area Under Curve (AUC) standard measure from 84% to 93.3%. The dataset's challenges included its massive size, varied data formats, and rapid influx, necessitating a suitable big data platform for handling and processing. The project showcased the advantages of big data platforms in collecting, storing, and processing diverse data efficiently. The study also emphasized the importance of addressing class imbalance in the dataset, with three scenarios explored: oversampling, under sampling, and without re-balancing. The evaluation utilized AUC due to its suitability for unbalanced datasets. The research highlighted the superiority of big data systems in handling complex computations, particularly in the context of Social Network Analysis measures on large-scale networks.

The project's relevance was underscored by the telecom industry's focus on customer retention strategies, acknowledging the higher costs associated with acquiring new customers compared to retaining existing ones. The project demonstrated that big data technologies facilitated feature engineering, a critical aspect of predictive model development, allowing for the extraction of richer features. The model's evaluation and impact testing were conducted using a new dataset, and the successful deployment of the model to production further validated its effectiveness. The project provided an overview of existing studies in churn prediction, emphasizing the project's unique contribution in the Syrian telecommunications context. The research bridged gaps by addressing feature engineering from raw data, handling class imbalance, and utilizing big data technologies for enhanced predictive modeling. The project's methodology and findings contribute valuable insights to the field of churn prediction in the telecommunications sector.

## Gaps Customer churn with data integration

In our case of customer churn prediction systems with integrated data, there are notable opportunities for improvement. Currently, existing models predominantly concentrate on batch data processing, introducing a delay in obtaining insights and taking proactive measures. Recognizing this gap, our objective is to introduce real-time data integration capabilities. By facilitating the seamless integration of data sources in real-time, even on a small scale, we aim to enhance the system's agility and responsiveness to evolving customer behaviors. This shift toward real-time processing represents a crucial advancement, ensuring that businesses can promptly identify and address factors contributing to customer churn.

Furthermore, scalability and efficiency in data integration processes are essential for the seamless operation of predictive models. We recognize the need to explore solutions that not only enhance the scalability of the system but also improve its overall efficiency. This involves optimizing data integration workflows to handle growing datasets without compromising on performance. A scalable and efficient data integration process is fundamental to ensuring that the churn prediction system can accommodate increasing volumes of data while maintaining a high level of accuracy and responsiveness.

Ethical considerations are paramount, particularly in the realm of customer churn prediction where fairness, bias, and data privacy play crucial roles. As part of our commitment to ethical practices, we aim to address these concerns comprehensively. In churn prediction, biases may inadvertently impact certain customer segments, and the potential for privacy breaches is a serious consideration. Our approach involves incorporating ethical safeguards into the system, mitigating biases, ensuring fairness, and prioritizing data privacy. By proactively addressing these ethical considerations, we aim to build a customer churn prediction system that not only delivers accurate insights but also upholds the highest standards of fairness and privacy in its operations.

## Conceptual Frame work

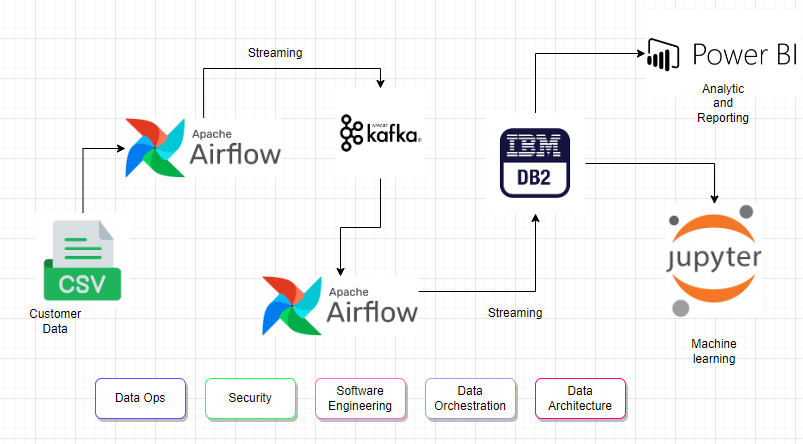


Figure 2‑2 Conceptual Frame work

Above we have the conceptual frame work of how the project is going to look like. Let us see what each step represent and what will happen.

### Generation

Within the initial stage of our project, the generation of data unfolds as a pivotal foundation for our entire processing pipeline. Our primary data source resides in a CSV file located on the local computer, representing the genesis of information pertinent to customer interactions and behaviors. This file adheres to a structured format, characteristic of CSV conventions, requiring a nuanced understanding for effective handling throughout the pipeline. As we direct our focus toward customer churn prediction, the data's velocity adopts a more measured pace, aligning with a batch processing approach suited for monthly reports and periodic updates from the CSV file. Although the volume of data isn't extensive, careful handling is essential to extract meaningful insights. Even at this early stage, data quality assumes paramount importance, necessitating the implementation of validation checks and cleansing procedures to ensure accuracy.

Metadata, encompassing timestamps and source information, becomes a vital asset for tracking the data's provenance. Our data collection technique involves direct access to the CSV file on the local computer, ensuring efficient and accurate acquisition. Temporary storage considerations involve utilizing the computer's local storage or a designated data lake. Effective data governance practices, encompassing ownership, responsibility clarification, and defined retention policies, are integral even at this nascent stage, establishing the groundwork for a robust processing pipeline tailored for our customer churn prediction objectives.

### Ingestion

In the dynamic process of data ingestion, the ingestion stage serves as the cornerstone of our data ecosystem. Here, data flows in from diverse sources, each with its unique characteristics, formats, and structures. To orchestrate this diverse influx, connectors and specialized data ingestion tools are deployed, ensuring a harmonious integration of real-time, batch, and scheduled data. The intricacies of data quality are meticulously addressed at this juncture, employing robust validation processes and data profiling tools to cleanse and refine the incoming data. This meticulous cleansing addresses errors, inconsistencies, and duplicates, fostering a foundation of high-quality and reliable data.

As data arrives in varied formats, the focus then shifts to data transformation and normalization processes. The aim is to standardize the incoming data, aligning data types, and ensuring a consistent format. This not only enhances the accessibility of the data but also paves the way for more accurate and meaningful insights during subsequent processing stages. The ultimate objective is to centralize this refined data into a unified repository, laying the groundwork for a streamlined and efficient data processing pipeline.

The role of Apache Airflow becomes pivotal in this context, serving as the orchestrator for loading data into Kafka. This strategic integration ensures a seamless flow of data within our ecosystem, setting the stage for further processing. The centralized repositories, such as data lakes and warehouses, become the nexus where integrated data converges, ready for comprehensive analysis, and facilitating a cohesive data strategy for our project.

### Transformation

The transformation stage is almost similar with data cleaning and processing. The raw data may contain errors or inconsistencies. This is where we apply a series of cleaning techniques to rectify the issues for example applying statistical methods and imputation of missing values and data inconsistencies may also be resolved. This stage will ensure data is accurate and lack irregularities setting a good foundation for subsequent analysis (Bhatia, 2019).

Feature Engineering is one of the critical components for the transformation stage and particularly for machine learning application. It will involve the creations selection, or modification of variables in the data set. The aim here is to extract meaningful insights from the data by crafting relevant attributes. This may involve creating interaction terms, encoding categorical variables, scaling features and dimensionality reduction. Effectiveness of this will enhance performance of machine learning models (Joe Reis, 2022).

For scenarios requiring data summarization or aggregation for reporting and analysis, the "Transformation" stage is the hub for such operations. Aggregation involves grouping and summarizing data to glean higher-level insights, employing operations like averaging, summing, or identifying maximum and minimum values within specific time periods or categories. This yields aggregated data offering a more concise and comprehensive view of trends and patterns within the dataset.

In our project context, the role of Kafka in the transformation stage is instrumental. It serves as a dynamic platform facilitating data transformation, ensuring formatted and structured data aligns with the requirements of analytical tools, models, and reporting. Kafka's integration streamlines the transformation process, making the data interpretable, compatible with chosen tools, and ready for subsequent stages of modeling and analysis.

### Storage

In the intricate landscape of our data-driven project, the storage phase emerges as a critical chapter, orchestrated seamlessly by Apache Airflow. This phase marks the transition of data from our Kafka setup to the robust embrace of IBM Db2, forming the nexus where the intricacies of customer churn predictions are securely held and meticulously organized. Apache Airflow takes center stage as the orchestrator extraordinaire, ensuring the fluid and secure transfer of data from Kafka to its designated haven, IBM Db2. Beyond its technical prowess, Apache Airflow introduces an element of reliability and consistency crucial for sustaining the integrity of the stored data. This reliability is paramount, acting as the bedrock for subsequent analyses that hinge on the accuracy and completeness of the stored information.

Enter IBM Db2, our data sanctuary and analytical playground. Serving as the central repository, IBM Db2 not only safeguards our data with robust security measures but also ensures its accessibility for the intricate dance of customer churn predictions and analytical model extractions. The strategic choice of IBM Db2 aligns harmoniously with scalability requirements, empowering it to gracefully manage expanding data volumes while upholding peak performance standards. In summary, the storage phase stands as the linchpin of our project's success story. The collaborative synergy between Apache Airflow and IBM Db2 crafts a narrative of dependability, scalability, and meticulous organization, setting the stage for insightful analyses and predictions in the realm of customer churn.

### Serving for Machine Learning / Data Analytics:

In the "Machine Learning / Data Analytics" phase, we embark on constructing and training machine learning models tailored for customer churn prediction within our project. Utilizing powerful tools such as Jupyter Notebooks, we delve into sophisticated feature selection and engineering techniques. This involves not only selecting pertinent variables but also crafting insightful metrics like customer lifetime value and churn history to elevate the predictive prowess of our models (Nonis, 2022).

Our validation and evaluation processes are robust, employing industry-standard techniques such as cross-validation. We carefully scrutinize metrics like accuracy, precision, and recall to ensure the reliability and effectiveness of our models. Power BI and Jupyter Notebooks take center stage during this phase, aiding in the visualization and interpretation of model performance. These tools enable us to present results in an accessible manner, ensuring our models align with predetermined criteria. Upon successful validation, our models seamlessly transition into production, facilitated by APIs and integrated with Power BI. This integration extends the reach of our insights, providing real-time and actionable information directly within familiar CRM systems, empowering end-users, particularly support agents (Raj, 2020).

Complementing our machine learning endeavors, we apply data analytics techniques such as data mining and statistical analysis. Power BI continues to play a pivotal role by transforming complex data into intuitive reports and dashboards. These visualizations serve as a compass for end-users, aiding them in deciphering insights, making informed decisions, and responding effectively to customer churn scenarios. The synergy between machine learning and data analytics, coupled with the user-friendly interfaces of Power BI and Jupyter Notebooks, forms the backbone of our comprehensive approach to address the challenges of customer churn prediction

### Underlying Factors

#### Security

In the realm of e-commerce, security is paramount. Ensuring the security of customer data is not only a legal requirement but also a trust-building measure. This involves measures like encrypting customer data during storage and transmission, implementing access controls to restrict data access to authorized personnel only, and monitoring for any unusual or unauthorized activities that may indicate a security breach. Compliance with data protection regulations, such as the GDPR, is essential to maintain customer trust and avoid legal issues (Ralph Kimball, 2019).

#### Data Management

E-commerce platforms generate a vast amount of data daily. Effective data management is essential for organizing and maintaining this data. It includes structuring data storage to handle the large volume of transactional data, implementing indexing mechanisms for quick data retrieval, version control to track changes and ensure data integrity, and data cataloging to ensure that data is well-organized and easily accessible. Proper data management helps in efficiently handling and using the wealth of customer data generated (Joe Reis, 2022).

#### DataOps

In the e-commerce context, DataOps can significantly improve the efficiency of the data pipeline. It combines data engineering and operations to streamline data processes. Collaboration among data engineers, data scientists, and operations teams is vital to ensure data flows seamlessly. Automation of routine data tasks, monitoring for data quality, and rapid deployment of updated models are essential aspects of DataOps. For an e-commerce platform, DataOps can lead to timely insights into customer behavior, allowing for swift responses to potential churn risks (Joe Reis, 2022).

#### Data Architecture

Data architecture plays a critical role in an e-commerce platform's ability to efficiently store and manage data. Decisions about databases, data models, and the overall structure of the data ecosystem are key. In e-commerce, data architecture needs to consider the scalability and performance requirements of handling vast quantities of transactional data. Ensuring that the architecture supports both real-time and historical data is crucial for effective customer churn prediction (Joe Reis, 2022).

#### Orchestration

Effective orchestration is vital for coordinating data processes and tasks in the pipeline. In e-commerce, it's about managing the flow of data from different sources, including online transactions, customer interactions, and other data streams. Workflow automation ensures that data is processed efficiently and in the right sequence. For example, orchestrating the flow of data from website clicks to purchase history can provide a comprehensive view of customer behavior (Joe Reis, 2022).

#### Software Engineering

Software engineering principles are central to the development and maintenance of the software components of our data pipeline. This includes building APIs for data access, creating data transformation scripts to preprocess and clean data, and developing machine learning models for customer churn prediction. In e-commerce, efficient software engineering ensures that the churn prediction models are up-to-date, scalable, and responsive to real-time customer interactions (Joe Reis, 2022).

# Methodology

## Introduction

In this chapter we will look at the methodology that is going to be selected for our project and in this case is the building a data pipeline for our customer churn prediction in the e-commerce sector. We will look at overview of the development approach that we will select and deep dive in to the different phases of the project. The chapter will entail and introduction to the selected development approach, a justification for its selection and in depth system analysis and a thorough system design.

## Development Approach

For our project, we have carefully considered that the selection of a development approach that is well suited for the complexities in addressing data integration for a customer churn prediction in the ecommerce industry. And after evaluation we have chosen to go with the structured systems analysis design methodology. SSAD is the systematic approach that will provide well-defined framework for analyzing, designing and implementing complex systems also offering a structured and organized way to tackle intricate projects making it valid choice for our data engineering project

## Justification of the Methodology

The choice for the methodology chosen which the SSAD is because of its effectiveness in handling complex projects especially in this case where we are looking to integrate data into a customer churn prediction in the e-commerce sector. SSAD provides clear and organized approach to system development, which is crucial for a project with multiple components and data integration requirements. It emphasizes the importance of thorough analysis, logical design and efficient implementation ensuring that the project aligns with the specific needs and objectives of the e-commerce business

## System Analysis

In this phase, we will utilize various SSAD tools and techniques to comprehensively understand the requirements of our project and build our system in accordance. The focuses will be on defining the technologies, assessing team workloads, identifying limitations, establishing timeframes, and managing the budget. This methodology will provide a structured and systemic approach to system analysis, which will be invaluable in ensuring the success of our data engineering project

### Context Diagram

A context diagram is a simplified visual representation used in data analytics to illustrate interactions between a data pipeline and its external environment. In the context of a data analytics project for an e-commerce company, the system (data pipeline) is depicted along with external entities like the online store database, customer data, and a management dashboard. Processes, such as data cleaning and sales analysis, are represented within the system, connected by arrows indicating the flow of data.

The significance of a context diagram lies in its ability to provide a clear and concise overview of the system's interactions. It helps define the project's scope, facilitates communication among team members and stakeholders, and serves as a foundational tool for understanding the flow of data within our system.

### DFD Level 0, 1

In the context of our project, the Level 0 Data Flow Diagram (DFD) serves as a strategic starting point, offering a bird's-eye view of the entire system. Here, a single process encapsulates the entirety of our data analytics pipeline, connecting with external entities such as databases, users, and potentially a reporting system. The Level 0 DFD lays out the fundamental processes, key data stores, and the essential data flows that illustrate how information is exchanged between the system and its external elements.

As you progress to the Level 1 Data Flow Diagram (DFD), the diagram provides a more detailed breakdown of the major processes identified at Level 0. Each primary process, such as data cleaning or sales analysis, is further dissected into specific sub processes. This decomposition creates a more granular perspective, highlighting the intricacies of the internal workings of our data integration system. Additionally, Level 1 incorporates data stores, representing where data is stored during various stages, and depicts the data flow paths, outlining precisely how data moves between sub processes and stores. This level of detail is instrumental for a comprehensive understanding of the system's internal processes and the flow of data within them.

### ERD (Entity-Relationship Diagram)

In our data engineering project, the Entity-Relationship Diagram (ERD) serves as a valuable tool for modeling the relationships between entities within our data pipeline. Entities, representing distinct elements in our system like datasets or tables, are visualized along with their attributes, offering a comprehensive view of the data's characteristics. For instance, in an e-commerce analytics scenario, entities might include "Customer" and "Product," each with attributes defining customer details and product information.

The ERD's primary function is to illustrate how these entities are interrelated. Relationships, depicted by lines connecting entities, showcase how data flows and interacts within the system. This visual representation aids in understanding the data's structure and organization, providing a foundation for designing an efficient and cohesive data analytics pipeline.

### Database Schema

In our data engineering project, the Database Schema acts as a crucial blueprint, detailing the structure and organization of our database. It outlines tables such as "Customers," "Products," and "Sales," each with specific fields like customer ID, product ID, and sales data. The schema not only defines the data types and attributes within these tables but also establishes relationships between them, providing a clear roadmap for how data is stored and interrelated.

The significance of the Database Schema lies in its ability to offer structural clarity and define relationships between different elements of data. For instance, the schema specifies how customer information in the "Customers" table is linked to sales data in the "Sales" table through unique identifiers. This clear representation facilitates efficient data retrieval and storage, forming a foundational element in the design and maintenance of a robust data engineering pipeline.

## System Design

In the system design phase of our data engineering project aimed at creating an E-commerce Customer Churn Prediction system, we will leverage various Structured Systems Analysis and Design (SSAD) tools and techniques to ensure a comprehensive understanding of project requirements. The focus will be on defining technologies, assessing team workloads, identifying limitations, establishing timeframes, and managing the budget. This systematic approach aims to lay a solid foundation for the success of our project.

The Context Diagram will serve as a crucial guide in visualizing the interactions between our data pipeline and external elements. Representing the system (data pipeline) alongside entities like the online store database, customer data, and a management dashboard, this diagram will help define the project's scope and facilitate effective communication among team members and stakeholders.

The Data Flow Diagrams (DFDs) at Levels 0 and 1 will provide a structured overview of our data analytics pipeline. The Level 0 DFD offers a bird's-eye view, encapsulating the entire system and illustrating fundamental processes, data stores, and essential data flows. As we move to the Level 1 DFD, a more detailed breakdown of major processes, such as data cleaning and sales analysis, will be presented. This granular perspective, including sub processes, data stores, and data flow paths, is crucial for a comprehensive understanding of the internal workings of our customer churn prediction system.

The Entity-Relationship Diagram (ERD) will play a pivotal role in modeling relationships between entities within our data pipeline. Entities like "Customer" and "Product," each with specific attributes, will be visualized, demonstrating the interrelations crucial for predicting customer churn. This visual representation aids in understanding the data's structure and organization, providing the foundational structure for designing an efficient and cohesive E-commerce Customer Churn Prediction system.

The Database Schema not only defines data types and attributes within tables but also specifies how different elements of data are connected. For instance, it outlines how customer information in the "Customers" table is linked to sales data in the "Sales" table through unique identifiers. The Database Schema's significance lies in providing structural clarity and facilitating efficient data retrieval and storage. This schema forms a foundational element in the design and maintenance of a robust data engineering pipeline, ensuring the seamless functioning of our E-commerce Customer Churn Prediction system.

## System Deliverable/Milestones

### Deliverables

The feasibility study report will systematically evaluate technical feasibility by assessing the technology stack's compatibility with the project requirements. Operational feasibility will delve into how well the proposed data pipeline aligns with existing business processes. Economic feasibility will include a detailed cost-benefit analysis, considering both initial development costs and long-term operational expenses.

Requirements Specification Document will not only list functional and non-functional requirements but also detail the rationale behind each requirement. It will provide a traceability matrix to link requirements to specific project objectives, ensuring transparency and alignment with the overarching goals of predicting customer churn in the e-commerce sector.

Building on the context diagram, the system architecture diagrams will include an expanded DFD (Data Flow Diagram) hierarchy, illustrating the flow of data through different system components. Additionally, the ERD (Entity-Relationship Diagram) will model the relationships between various entities in the e-commerce data environment, guiding the database schema design.

The implementation codebase for workflow orchestration and machine learning algorithm will be organized following best practices and coding standards. Detailed comments and documentation within the code will facilitate ease of understanding for future developers or maintainers. Version control systems like Git will be used to track changes, ensuring code reliability and collaboration among team members.

The testing documentation will include unit tests for individual components, integration tests to ensure seamless interaction between different modules, and system tests to validate the overall functionality. Performance testing will be conducted to assess the system's response under various loads, ensuring its scalability.

User manuals will cater to various stakeholders, providing different levels of detail for end-users, system administrators, and developers. They will include step-by-step guides, screenshots, and troubleshooting sections to enhance usability and facilitate a smooth user experience.

### Milestones

The completion of the feasibility study will involve presenting findings to stakeholders, engaging in discussions to address any concerns, and obtaining formal approval to proceed. It marks a critical point where the project's potential challenges and risks are well-understood.

The sign-off on requirements analysis will involve stakeholders acknowledging that the specified requirements accurately reflect their needs. This milestone is crucial to ensure that subsequent development phases are aligned with the business objectives and expectations.

These phases will be organized using an agile framework, with sprints and iterative development. Milestones may include the completion of specific user stories, successful integration of data sources, and the achievement of predefined testing criteria. Regular sprint reviews will keep stakeholders informed and engaged.

Project completion involves deploying the data pipeline to the production environment, monitoring its initial performance, and transitioning to the maintenance and support phase. A comprehensive handover to the operations team will ensure a smooth transition and ongoing stability.

### System proposal

The proposed Data-Driven Customer Churn Prediction System is a comprehensive solution designed to tackle the persistent challenge of customer churn faced by businesses. Leveraging sophisticated data integration techniques, the system seamlessly incorporates diverse data sources, offering a holistic view of customer interactions. By employing advanced machine learning algorithms, the system analyzes historical customer data to predict potential churn indicators, identifying patterns and correlations. Real-time monitoring ensures continuous tracking of customer behavior, enabling immediate responses to evolving patterns and potential churn triggers. The system presents its findings through an intuitive dashboard, providing stakeholders with actionable insights for informed decision-making. Expected outcomes include proactive churn mitigation, improved decision-making through data-driven insights, and enhanced customer retention rates. The implementation plan encompasses phases such as system integration, data onboarding, algorithm implementation, real-time monitoring setup, and dashboard development, ensuring a seamless and effective deployment of the proposed system into existing business operations.

# System Analysis and Design

## Introduction

This chapter delves into the analysis and design of the Data-Driven Customer Churn Prediction System, a comprehensive solution designed to tackle the persistent challenge of customer churn faced by businesses. The system leverages sophisticated data integration techniques and advanced machine learning algorithms to predict potential churn indicators, identifying patterns and correlations.

We will explore the system requirements, both functional and non-functional, that underpin the system's design and operation. This will include the data pipelines, the data exploration and machine learning methods that we will use.

We will also examine the various system analysis and design diagrams that have been drawn up to illustrate the system's structure and workflow. These include Context Diagram, Data Flow Diagrams, Activity Diagrams, ERD Diagrams, Logical Database Schema, and Class Diagrams.

The chapter will also discuss the implementation plan, which encompasses phases such as system integration, data onboarding, algorithm implementation, real-time monitoring setup, and dashboard development. This plan ensures a seamless and effective deployment of the proposed system into existing business operations.

By the end of this chapter, the reader will have a comprehensive understanding of the system's design, its requirements, and the steps taken to ensure its successful implementation and operation.

## System Requirements/Requirements Gathering

The Data-Driven Customer Churn Prediction System is designed with a set of specific functional and non-functional requirements in mind. These requirements are crucial to ensure the system's effectiveness in predicting customer churn and providing actionable insights for businesses.

### Functional Requirements

1. Data Integration: The system is designed to seamlessly incorporate diverse data sources, offering a holistic view of customer interactions. This is achieved through sophisticated data integration techniques in our case would be for example extracting data from various files such as csv.
2. Machine Learning Algorithm: The system employs advanced machine learning algorithms to analyze historical customer data. This allows the system to predict potential churn indicators by identifying patterns and correlations. Some of the algorithms we will see will be for example Random Forest and Decision Trees.
3. Real-Time Monitoring: The system is equipped with real-time monitoring capabilities to continuously track data flow within pipelines. This high data through put and high data quality and integrity.
4. Dashboard Development: The system presents its findings through an intuitive dashboard, providing stakeholders with actionable insights for informed decision-making.

### Non-Functional Requirements

1. System Integration: The system is designed to be integrated into existing business operations seamlessly. This is achieved through a detailed implementation plan that includes system integration, data onboarding, algorithm implementation, real-time monitoring setup, and dashboard development.
2. System Security: The system is designed with robust security measures to protect the data and the integrity of the system. This includes the use of secure data transfer protocols and stringent access controls.
3. Scalability: The system is designed to handle varying loads, ensuring its scalability. This is achieved through performance testing, which assesses the system's response under various loads.
4. Usability: The system is designed with an intuitive user interface and detailed user manuals. This ensures a smooth user experience and facilitates ease of understanding for various stakeholders.

## System Analysis Diagrams

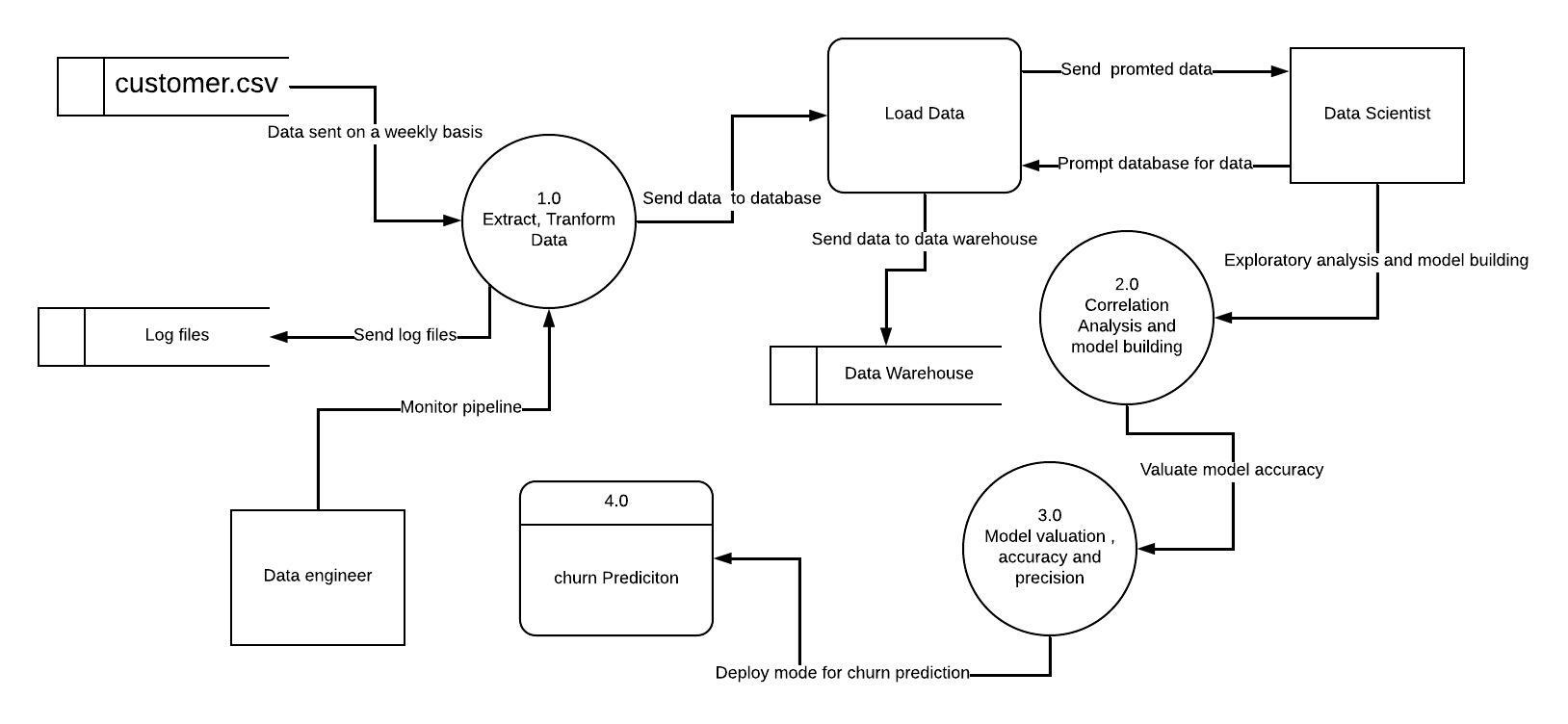
System analysis diagrams provide a visual representation of the system's components and their interactions. They are crucial in understanding the system's functionality and identifying potential areas of improvement. In the context of the Data-Driven Customer Churn Prediction System, several diagrams were used to analyze the system.

### Context Diagrams

The context diagram provides a high-level overview of the system, illustrating the system's interactions with external entities. It shows how the system exchanges data with its environment, including users, data sources, and other systems.

### Data Flow Diagrams (Level 0, Level 1)

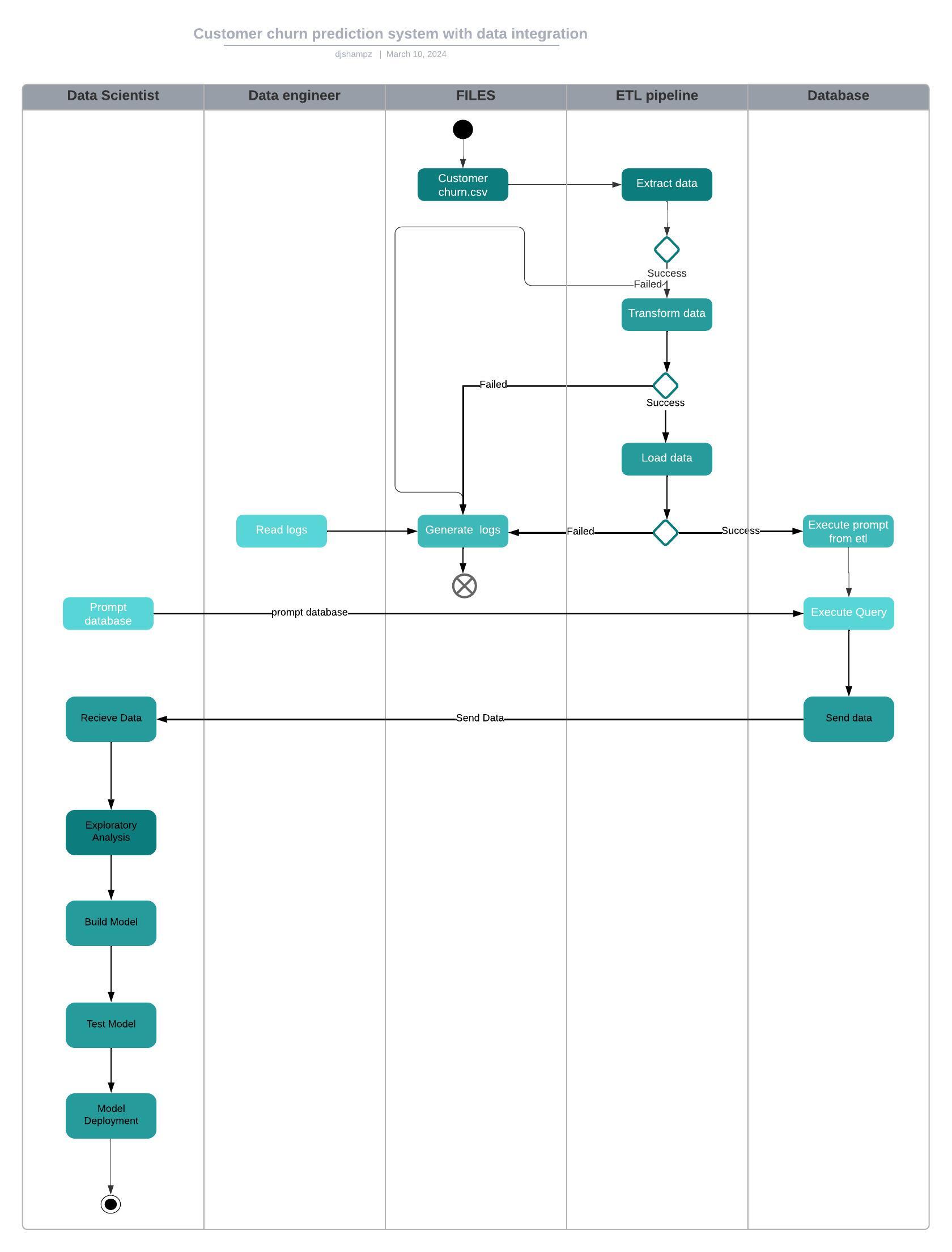
Data Flow Diagrams (DFDs) provide a detailed view of the system's data flows. The Level 0 DFD, also known as the context diagram, shows the system as a whole and how it interacts with external entities. The Level 1 DFD breaks down the system into its major high-level processes, illustrating how data flows between them.



### Activity Diagrams

Activity diagrams depict the system's dynamic behavior, showing the sequence of activities that occur in a particular use case. They are particularly useful in visualizing the flow of control in the system.

These diagrams were instrumental in understanding the system's data flows, interactions, and control flows. They provided valuable insights into the system's functionality and helped identify potential areas of improvement.

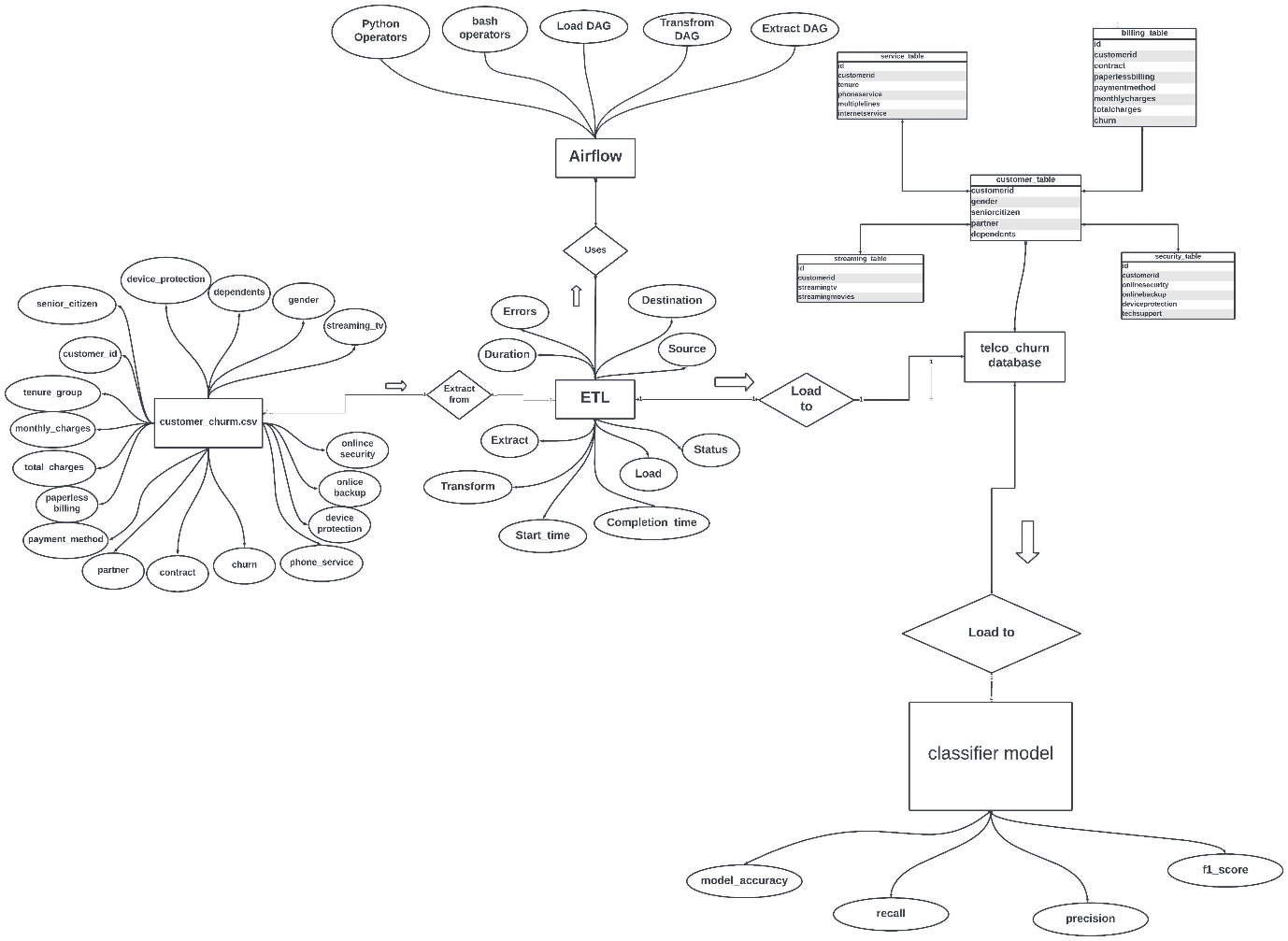


## System Design Diagrams

System design diagrams are instrumental in visualizing the structure of the system and how its components interact with each other. They provide a clear picture of the system's architecture, making it easier to understand and modify. In the context of the Data-Driven Customer Churn Prediction System, two types of diagrams were primarily used: Entity-Relationship Diagrams (ERD) and Logical Database Schema.

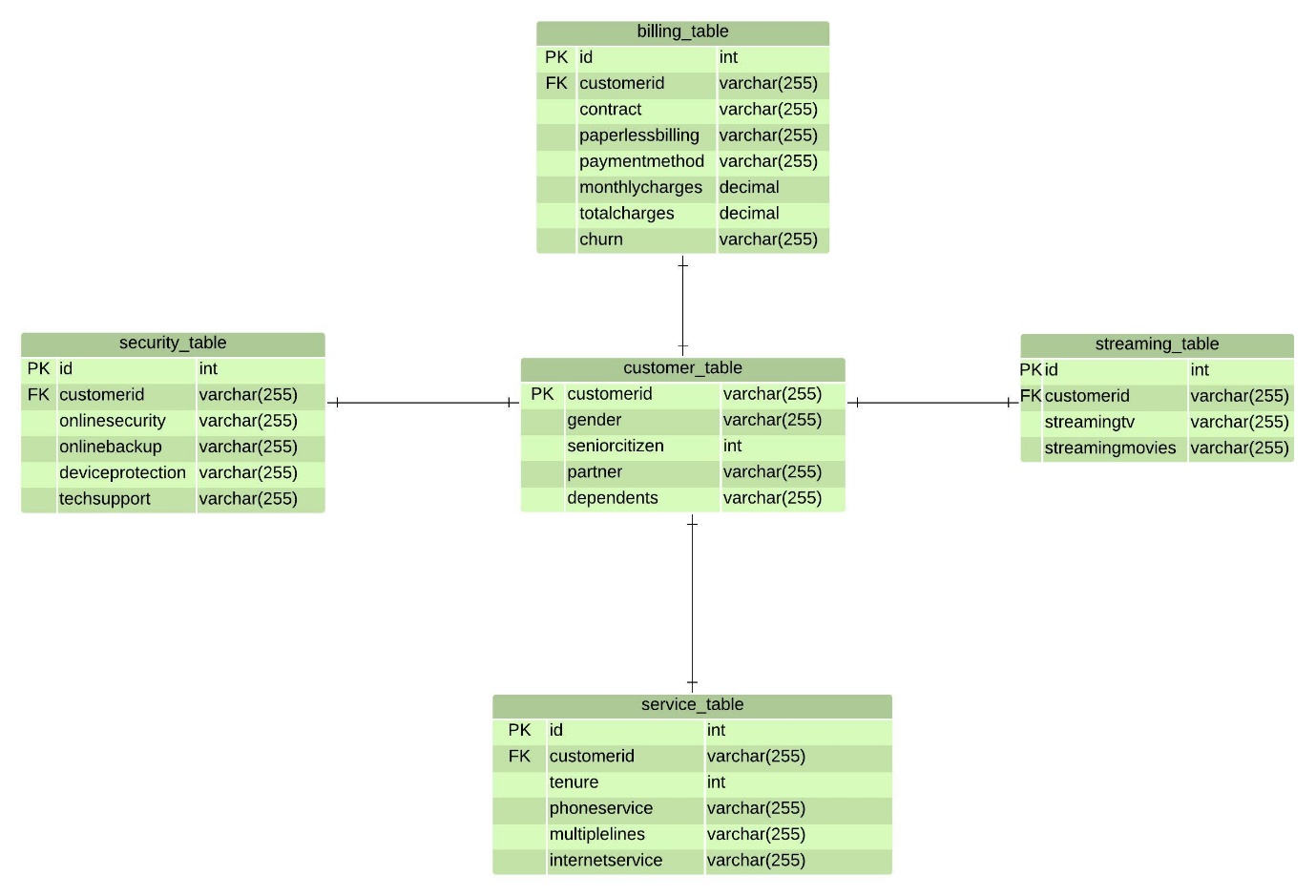
### Entity-Relationship Diagrams (ERD)

ERDs are used to model the relationships between various entities in the system. They provide a visual representation of the system's data and how it is organized, making it easier to understand the system's structure and the relationships between its components. In the context of the Data-Driven Customer Churn Prediction System, the ERD illustrates the relationships between various entities in the e-commerce data environment, guiding the database schema design.



### Logical Database Schema

The Logical Database Schema shows how the actual database will be formed by logically setting out how the tables and relationships will be formed. It provides a detailed view of the system's data structure, including tables, fields, and the relationships between them. In the context of the Data-Driven Customer Churn Prediction System, the Logical Database Schema illustrates how customer data is organized and how it can be accessed and manipulated to predict customer churn.



These diagrams were instrumental in designing the system's data structure and understanding how data flows within the system. They provided valuable insights into the system's design and helped guide the implementation of the Data-Driven Customer Churn Prediction System.

# System Implementation and testing

## Introduction

This chapter delves into the practical aspects of the Data-Driven Customer Churn Prediction System, focusing on its implementation and testing. The implementation environment, both hardware and software, is described in detail, providing an understanding of the infrastructure that supports the system.

For machine learning projects, a comprehensive description of the dataset used is provided, including its retrieval, analysis, and division into training, testing, and validation data. The chapter also outlines the testing strategies employed to ensure the system's robustness and reliability. It discusses the various testing paradigms used, such as white box testing, black box testing, and unit testing, and presents the results of these tests.

The aim of this chapter is to provide a comprehensive understanding of how the system was brought to life from the design and analysis stages, and how its functionality and performance were validated. This will provide insights into the practical challenges encountered and how they were addressed, thereby offering a complete picture of the system's development lifecycle.

## Description of the Implementation Environment

### Hardware Specifications

The development of the Data-Driven Customer Churn Prediction System was carried out on a laptop. The laptop's specifications are as follows:

* Processor: Core i5 2.1GHz
* RAM: 12 GB
* Storage: 500GB
* Operating System: Windows 10

The laptop provided a reliable and efficient environment for the development and testing of the system. Its processing power and memory capacity were sufficient to handle the computational demands of the machine learning algorithms and data processing tasks. The operating system supported all the necessary software tools and provided a user-friendly interface for development.

### Software Specifications

The software tools used for the development of the system are as follows:

* Apache Airflow: Used for orchestrating the data pipeline.
* Jupyter Notebook: Used for the development and testing of the machine learning algorithms.
* MySQL: Used as the primary database management system for the project.
* Python: The primary programming language used for the development of the system.

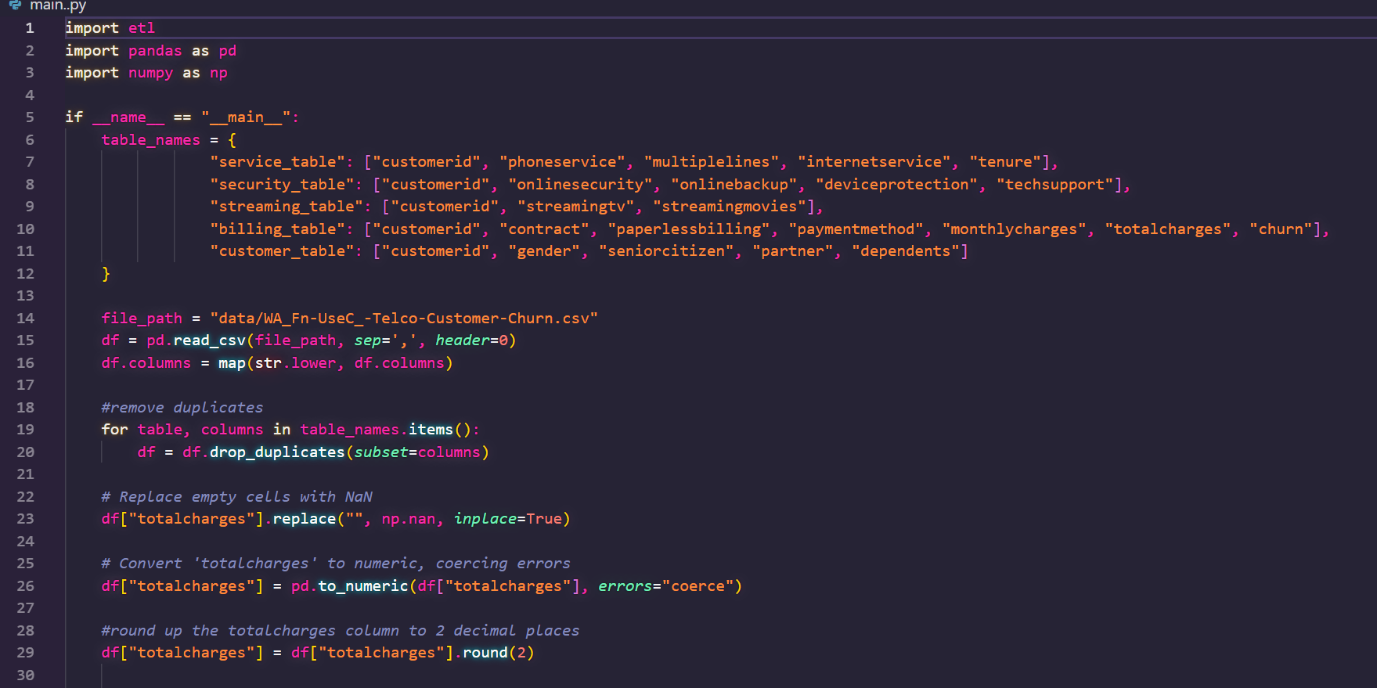
These software tools were chosen for their robustness, efficiency, and compatibility with the project requirements. They provided a comprehensive environment for the implementation of the system, from data processing and machine learning to database management.

## Description of the Dataset

The dataset used for the development of the Data-Driven Customer Churn Prediction System is a CSV file containing customer data. The data was retrieved from the file WA\_Fn-UseC\_-Telco-Customer-Churn.csv and includes various customer attributes such as gender, senior citizen status, partner status, dependents, tenure, phone service, multiple lines, internet service, online security, online backup, device protection, tech support, streaming TV, streaming movies, contract, paperless billing, payment method, monthly charges, total charges, and churn status.

The dataset was analyzed using Python, with the help of pandas and numpy libraries. The data was divided into training, testing, and validation sets to ensure a robust machine learning model. The features used for the model were extracted from the dataset, and one-hot encoding was applied to convert categorical data into a format that could be provided to the machine learning algorithms.

The dataset was processed using an ETL (Extract, Transform, Load) pipeline, as described in the etl.py script. The script includes functions for creating a database connection, extracting column names, creating column types, creating tables, inserting data into tables, and adding primary and foreign keys. The script also includes data cleaning steps such as removing duplicates, replacing empty cells with NaN, and converting the 'totalcharges' column to numeric. The dataset provided a comprehensive view of customer interactions, which was crucial for the prediction of customer churn. The features in the dataset allowed the machine learning model to identify patterns and correlations that could indicate potential churn. The use of this dataset was integral to the successful implementation of the Data-Driven Customer Churn Prediction System.



The above screen shots just shows some of the operation that will be conducted on the data this will happen within our data pipeline that will extract the data from a csv file named WA\_Fn-UseC\_-Telco-Customer-Churn.csv. The above shows a dictionary with lists of table columns for values and keys are the table names, that was used for create our table structure that will be used as a basis of creating our database schema

All this process will be executed in one of the dags in our airflow orchestration. There will be three stages of our data pipeline which will be

1. Extraction – this will basically be the extraction of data from the source destination and there won’t be much going on here as for our case the file will be available locally there could be cases where data would be from APIs and cloud databases which our system is well equipped to do so
2. Transformation of the data – this will entail the transformation of data from the csv files into data frames for further processing. There will be conversion of our columns from capital or uppercase to lowercase for consistency purposes when dealing with the column names. Also, some of the columns need to be converted into their proper datatypes
3. Loading – This stage will load data into our MySQL server into their respective tables also creating various keys and references / relationships that we will need when dealing with the database and queries

## Description of Testing

In this section we will look at the various ways of testing we used for our pipeline in general and our machine learning model. This will entail the various test case scenarios of how the pipe line should behave when put under certain situations and stress. For the model it would involve using various testing methods on how accuracy, f1-score and other metrics that we will discuss

### Data pipeline testing and model training

For the testing paradigm we will be using unit testing as we have many different parts of the parts of our system that will interact with each other within the system and it is crucial that all the parts of the project are working as individual on a stand alone basis. We also employed white box texting for our model specifically as it is crucial in that it will allow us to have predictable outcome for our various accuracy and scores.

## Testing results and verdict

# CONCLUSIONS, RECOMMENTATIONS AND FUTURE WORKS

## Conclusions

## Recommendations

## Future works

# References

Argelaguet, R. C. (2021). *Computational principles and challenges in single-cell data integration.* Nature biotechnology.

Bhatia, P. (2019). Data mining and data warehousing. *Principles and practical techniques*.

Çelik, O. &. (2019). Comparing to techniques used in customer churn analysis. *Journal of Multidisciplinary Developments*, 30-38.

de Lima Lemos, R. A. (2022). Propension to customer churn in a financial institution: A machine learning approach. *Neural Computing and Applications*, 11751-11768.

Fedushko, S. &. (2022). E-commerce customers behavior research using cohort analysis: A case study of COVID-19. *Journal of Open Innovation: Technology, Market, and Complexity*, 12.

Fonseca, G. &. (2023). Overcoming the challenges of data integration in ecosystem studies with machine learning workflows: an example from the Santos project. *Ocean and Coastal Research*, 71.

Hira, S. &. (2023). Automated heuristic based context dependent ETL process to generate multi‐dimensional model for tabular data. *Concurrency and Computation: Practice and Experience*, 35.

J. M., M. J. (2021). Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences*, 118.

Joe Reis, M. H. (2022). *Fundamentals of Data Engineering.* Sebastopol, CA: O'Reilly Media, Inc.

Kosvyra, A. F. (2021). Towards data integration for AI in cancer research. *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2054-2057.

McGuirk, M. (2023). Performing web analytics with Google Analytics 4: a platform review. *Journal of Marketing Analytics*, 1-15.

Momin, S. B. (2020). Prediction of customer churn using machine learning. *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing: BDCC 2018*, 203-212.

Munappy, A. R. (2020). From ad-hoc data analytics to dataops. *In Proceedings of the International Conference on Software and System Processes*, 165-174.

Murphy, G. (2022). The 5 Tenets of Data Integration. *The 5 Tenets of Data Integration*, 1-3.

Nonis, P. R. (2022). Customer retention and addressing customer churn through Predictive Analytics in telecom industry. *Doctoral dissertation*.

Oluwatoyin, A. M. (2022). Customer Churn Prediction in Banking Industry Using Power Bi. *In Proceedings of Third International Conference on Computing, Communications, and Cyber-Security:*, 767-774.

Raj, A. B. (2020). Modelling data pipelines. *2020 46th Euromicro conference on software engineering and advanced applications*, 13-20.

Ralph Kimball, M. R. (2019). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling.* John Wiley and Sons, Inc.

Yu, M. L. (2020). A new framework for online testing of heterogeneous treatment effect. *Proceedings of the AAAI Conference on Artificial Intelligence*, 10310-10317.

.

Appendix

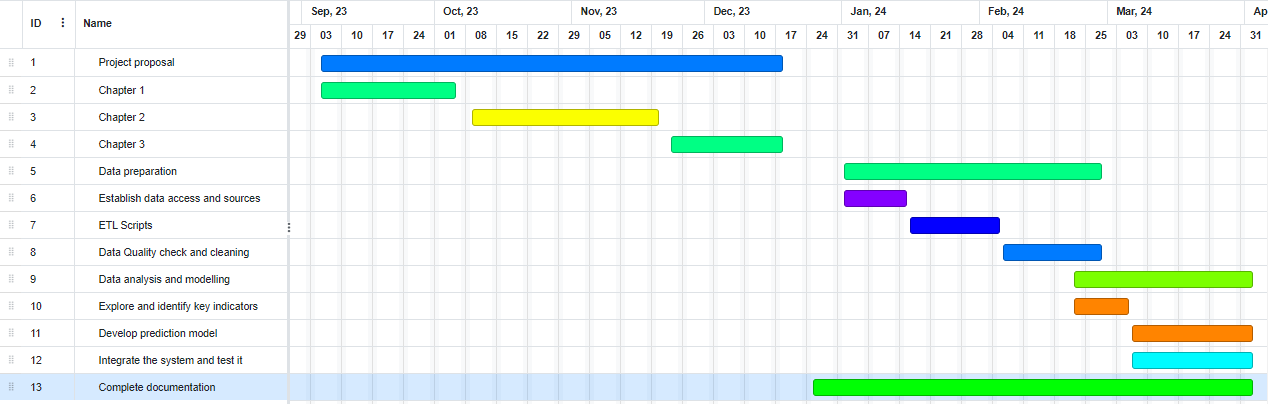


Figure 0‑1 Gant Chart