

**Customer Churn Prediction for Ecommerce and data Integration with data engineering**

**Kimani Martin Gitau**

**100994**

**Submitted in partial fulfilment of the requirements of the Bachelor of Business in Information Technology at Strathmore University**

**School of Computing and Engineering Sciences**

**Strathmore University**

**Nairobi, Kenya**

**October 2020**

Declaration and approval

I declare that this work has not been previously submitted and approved for the award of a

Degree by this or any other University. To the best of my knowledge and belief, the research

Proposal contains no material previously published or written by another person except where

Due reference is made in the research proposal itself.

Student Name: Martin Gitau

Admission Number: 100994

Student Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Supervisor Name: Benjamin Mundama

Supervisor Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Abstract

Data has become a very valuable asset in the information technology world and other spaces. Data has become an asset for many organization even being compared to the value of oil. This means that data needs to be utilized for any organization that needs to stay ahead of the game, and also create values for the organization. Data by itself is not that meaningful and it tends to be messy. There needs to be a way to take in this data and change it to meaningful insights, this work is normally done by people like Data analysts, BI analysts and Data Scientists. So this means that there needs to be a form a data integration in a business in order for the organization to gain insights

Customer churn is a phenomena that happens when consumers stop interacting with an online product or service. This can be a problem for businesses and organization as it leads to less revenue and slow business growth. This phenomena can’t be explained but with enough information form data we can gain insight into the problem. Which brings us to the main focus of the problem which is making sure data integration in a business is done properly and using the data to gain insights and try to foresee when the problem will occur.

Table of content

[Declaration and approval ii](#_Toc147059986)

[Abstract iii](#_Toc147059987)

[Table of content iv](#_Toc147059988)

[Chapter 1 Introduction 1](#_Toc147059989)

[1.1 Background 1](#_Toc147059990)

[1.2 Problem Statement 2](#_Toc147059991)

[1.3 Objectives 3](#_Toc147059992)

[1.3.1 General Objective 3](#_Toc147059993)

[1.3.2 Specific objectives 3](#_Toc147059994)

[1.4 Research Questions 3](#_Toc147059995)

[1.5 Justification 3](#_Toc147059996)

[1.6 Scope and limitations 4](#_Toc147059997)

[Figure 1‑1Data Engineering life cycle 7](#_Toc147055027)

# 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. 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 up to big companies respectively (Fundamentals of Data Engineering, 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.

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 (Fundamentals of Data Engineering, 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.

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. The purpose of this project is to show the importance of data integration in companies and we will use the customer churn problem for our downstream use case.

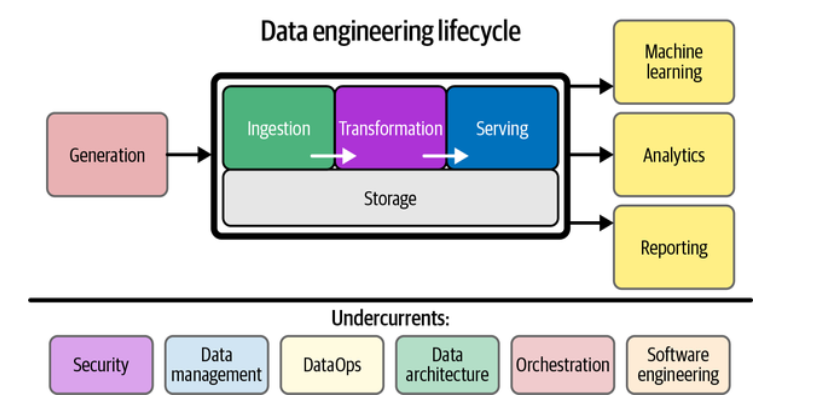


Figure ‑ Data Engineering life cycle

## Problem Statement

In light of the above context the problem at hand is twofold. Is that first, many organization struggle cause of poor data integration processes and lead to a lot of time spent on data cleaning and organization. The current state of data integration varies from company to company so it is crucial for data engineers and data analysts and scientists be aware of the maturity level of the organization.

Most data integration implementation that have been done such as batch processing and streaming. Streaming is when data is collected and processed continuously as it is being updated offering up to date data for real time analysis and insights. Tools used for this integration can be Apache Kafka, Apache storm etc. For batch processing is where data is collected in large amounts and then later processed and tools used for this can be Apache Hive, Apache Spark and Apache Hadoop.

All these process fall under workflows called ETL (Extract, Transform, Load). All the above technologies have been made to support such workflows but not integrated together since company systems vary from each other.

## Objectives

### General Objective

1. To implement data pipelines to collect and preprocess data for our downstream use case of Customer churn prediction

### Specific objectives

1. To investigate causes of customer churn
2. To analyze customer purchase behaviors and interaction with online services
3. To study of various data integration methods and techniques
4. To investigate various data science method used for predictive analysis
5. To study on challenges of current data integration behaviors adopted
6. To investigate tools used in data integration
7. To develop data pipelines for data preprocessing
8. To implement a down streams use case for our data pipelines

## Research Questions

1. What is customer churn
2. What causes customer churn
3. Who does customer churn effect
4. What is data integration
5. Why is data integration important
6. Where is the problem of lack proper data integration arising
7. Who is affected by lack of proper data integration
8. How much of a problem is lack of proper data integration

## Justification

The system we want to implement will not only show and predict customer prediction but also the valuable nature of data integration and how it can streamline business processes that require data and also make life more easier for the data analysts and data scientists.

## 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 ecommerce business.