**A Synopsis**
**On**

# “Cloud-Based Retail Data Lake & Analytics Platform”

Submitted in partial fulfilment of the requirements for the award of the Degree of

**Bachelor of Computer Application of**
**Poornima University, Jaipur**

**Session: 2023-2026**

**Submitted By:**
**Aakib Khan [PUFCEBASX13766]**
**Aditya Singh [PUFCEBASX15907]**
**Aditya Singh [PUFCEBASX14211]**

**III Year, AI&DS**

**Submitted to:** Department of Computer Applications

**Faculty of Computer Science & Engineering,**
Poornima University Ramchandrapura, Sitapura Ext., Jaipur, Rajasthan

# Cloud-Based Retail Data Lake & Analytics Platform

## Abstract

- The rapid growth of digital retail and e-commerce has resulted in the generation of massive volumes of transactional data from customers, products, orders, and payments.
- Traditional data management systems face limitations in handling such large and heterogeneous datasets efficiently.
- This project aims to design and implement a Cloud-Based Retail Data Lake & Analytics Platform that enables scalable storage, efficient data processing, and analytical reporting.
- The proposed system ingests raw retail data, processes it through structured layers, and stores analytics-ready data in MS SQL Server.
- The project demonstrates industry-level data engineering practices and provides meaningful business insights through structured queries and analysis.

## Introduction

- Retail organizations generate data from multiple sources such as online transactions, customer profiles, product catalogs, and payment systems.
- Managing and analyzing this data efficiently is crucial for business decision-making.
- With the advent of cloud computing and big data technologies, data lakes have become a preferred solution for handling large-scale data.
- This project focuses on building a cloud-based data lake system that stores raw and processed retail data and enables analytics using MS SQL Server.
- The system follows a layered architecture to ensure scalability, data quality, and efficient reporting.
- The project is developed using modern data engineering concepts and tools to simulate real-world industry scenarios.

## Problem Statement

- Retail data is generated in large volumes and diverse formats, making it difficult to store, process, and analyze efficiently using traditional systems.

- The absence of a centralized and scalable data platform leads to data inconsistency, poor performance, and delayed insights.
- This project aims to solve these challenges by designing a cloud-based data lake and analytics system.

## Objectives & Scope

### Objectives:

- To design a cloud-based retail data lake architecture
- To ingest and store raw retail transaction data
- To clean, transform, and organize data for analytics
- To implement structured analytical reporting using MS SQL Server

### Scope:

- The project focuses on backend data engineering and analytics. It does not include frontend dashboards but provides structured datasets for reporting and analysis.

## Literature Review

- Various studies highlight the importance of data lakes in managing big data for analytics.
- Existing retail analytics systems use centralized data warehouses and cloud platforms to process large datasets.
- Research indicates that layered architectures improve data quality and scalability.
- This project builds upon these concepts by implementing a simplified yet industry-relevant data lake model using SQL-based analytics.

## Feasibility Study

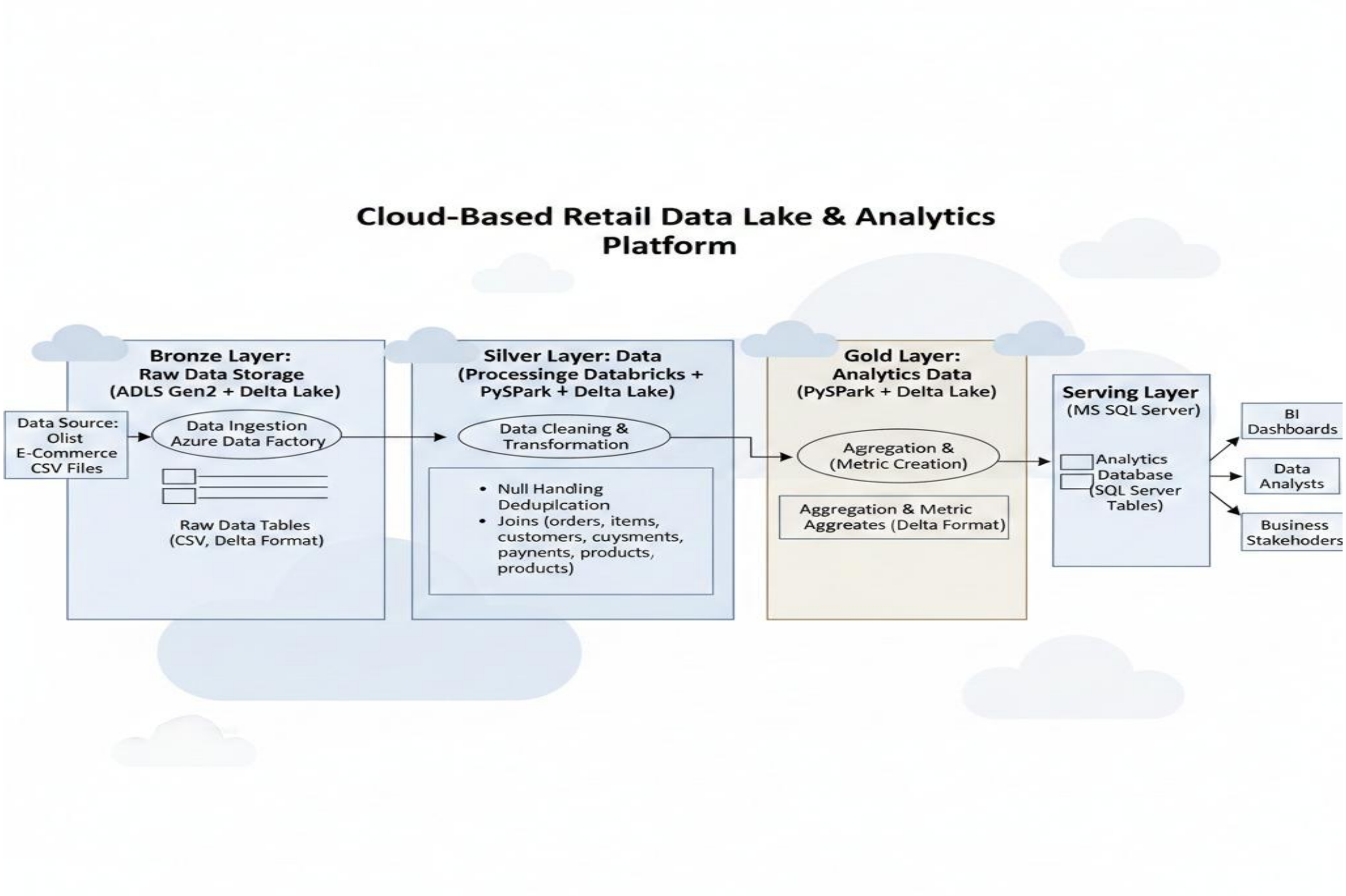
- **Technical Feasibility:** The project uses MS SQL Server and open-source datasets, which are easily accessible and suitable for large data processing.

- **Economic Feasibility:** The system is easy to use and maintain, with structured data layers and clear workflows.
- **Operational Feasibility:** The project uses freely available tools and datasets, making it cost-effective.
- **Need & Significance:** The project is significant as it solves the analytical problems in real-world scenarios.

## Modules in Proposed Project

- **Data Ingestion Module – Imports raw retail data into the system**
- **Data Storage Module – Stores raw and processed data**
- **Data Transformation Module – Cleans and structures data**
- **Analytics Module – Performs SQL-based analysis and reporting**
- **Reporting Module – Generates insights using queries**

## Data Flow Diagram (DFD)

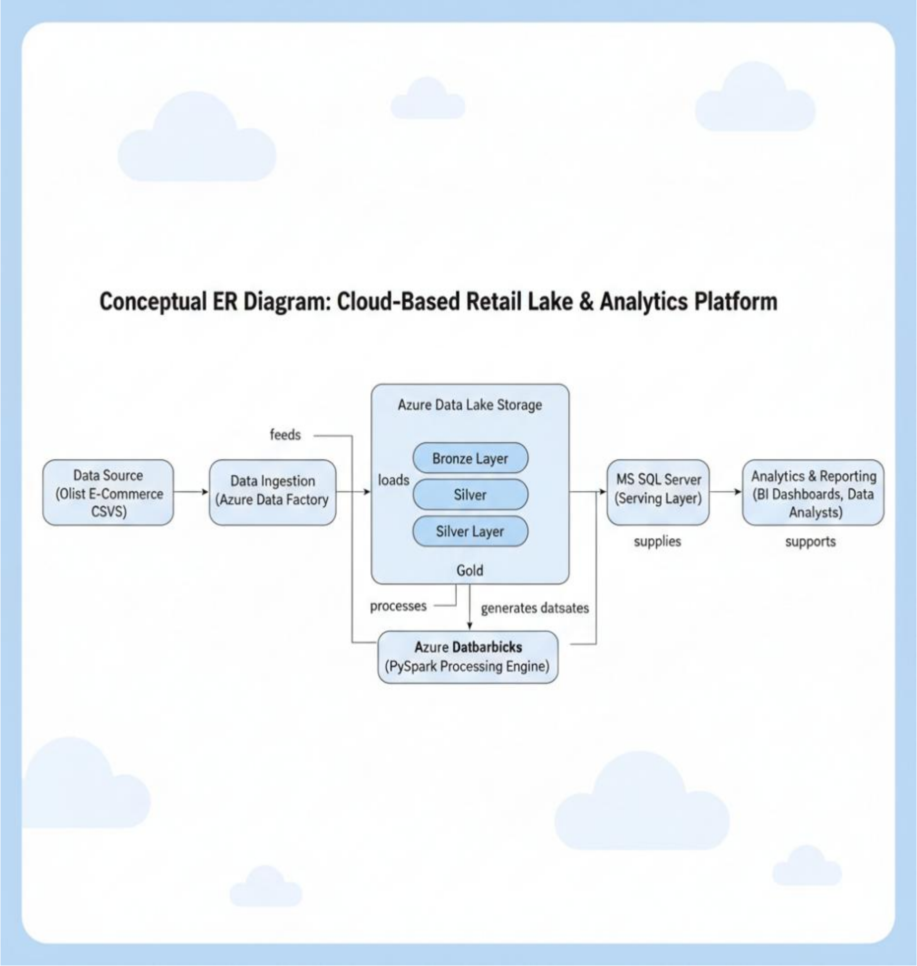

Cloud-Based Retail Data Lake & Analytics Platform
Bronze Layer: Raw Data Storage (ADLS Gen2 + Delta Lake)
Data Source: Olist E-Commerce CSV Files
Data Ingestion Azure Data Factory
Raw Data Tables (CSV, Delta Format)
Silver Layer: Data (Processinge Databricks + PySPark + Delta Lake)
Data Cleaning & Transformation
• Null Handling Dedupilcation
• Joins (orders, items, customers, cuysments, paynents, products, products)
Gold Layer: Analytics Data (PySPark + Delta Lake)
Agregation & (Metric Creation)
Aggregation & Metric Aggreates (Delta Format)
Serving Layer (MS SQL Server)
Analytics Database (SQL Server Tables)
BI Dashboards
Data Analysts
Business Stakehoders

- The DFD illustrates the **end-to-end flow of retail data** from source systems to analytics consumption using **Azure-based cloud services**.
- **Olist Brazilian E-Commerce Dataset** acts as the external data source, providing raw CSV files such as orders, customers, payments, and order items.
- Raw data is ingested into **Azure Data Lake Storage (Bronze Layer)**, where data is stored in its original form without transformations.
- The **Bronze layer** ensures data durability and acts as the single source of truth for future processing.
- Data from the Bronze layer is processed using **Azure Databricks with PySpark** in the **Silver Layer**, where data cleaning, null handling, deduplication, schema validation, and joins are performed.
- The **Silver layer** contains structured and quality-checked datasets suitable for analytical transformations.
- Further aggregations and business logic are applied in the **Gold Layer**, creating curated and analytics-ready datasets using **Delta Lake**.
- The **Gold layer** stores business metrics such as sales trends, order values, freight analysis, and customer behavior insights.
- Curated Gold-layer data is loaded into **MS SQL Server**, which acts as the serving layer for reporting and downstream analytics.
- **Business users, data analysts, and BI tools** consume data from MS SQL Server to generate dashboards and insights.
- The DFD clearly demonstrates **Medallion Architecture (Bronze–Silver–Gold)** and shows a **left-to-right data flow**, ensuring scalability, data quality, and separation of concerns.

**Entity-Relationship (ER) Diagram**

- The ER diagram represents a **high-level conceptual view** of the system components involved in the retail analytics platform.
- The **Data Source** entity represents the Olist Brazilian E-Commerce Dataset, which provides raw retail data to the system.
- The **Data Ingestion Process** entity initiates the flow of data by collecting raw files from the data source and loading them into the cloud environment.
- The **Data Lake (Azure Data Lake Storage)** entity acts as the central storage system and stores data across different layers.
- The **Bronze Layer** stores raw, unprocessed retail data exactly as received from the source.

- The **Silver Layer** represents processed and cleaned data where transformations such as null handling, deduplication, joins, and schema validation are performed.
- The **Gold Layer** contains curated and aggregated datasets that are optimized for analytics and reporting.
- The **Data Processing Engine (Azure Databricks with PySpark)** entity interacts with all Medallion layers to perform transformations and analytics logic.
- The **Serving Layer (MS SQL Server)** receives curated data from the Gold layer and makes it available for querying.
- The **Analytics & Reporting** entity represents BI tools and analysts who consume the processed data to generate business insights.
- Relationships between entities indicate **data flow, dependency, and interaction**, rather than primary–foreign key constraints.


Conceptual ER Diagram: Cloud-Based Retail Lake & Analytics Platform
Data Source (Olist E-Commerce CSVS)
Data Ingestion (Azure Data Factory
feeds
loads
Azure Data Lake Storage
Bronze Layer
Silver
Silver Layer
Gold
processes
generates datsates
Azure Datbarbicks (PySpark Processing Engine)
MS SQL Server (Serving Layer)
supplies
Analytics & Reporting (BI Dashboards, Data Analysts)
supports

## Technical Details

- ○ **Software Requirements**
  - Programming Language: Python, SQL
  - Big Data & Processing Frameworks: PySpark, Apache Spark
  - Cloud Services**:** Azure Data Lake Storage (ADLS), Azure Databricks
  - Data Storage & Management**:** Delta Lake, MS SQL Server
  - IDE / Tools**:** Azure Portal, Databricks Workspace

- ○ Hardware Requirements
  - Processor: Intel i5 or above
  - RAM: Minimum 8 GB
  - Storage: Minimum 20 GB free disk
  - Internet Connection**:** Required for cloud access

- ○ **Platform**
  - Development Platform: Windows
  - Deployment Platform: Cloud-based Application (Microsoft Azure)

## Methodology

**Definition**: The project follows a structured data engineering methodology to design and implement a cloud-based retail data lake and analytics platform.

**Purpose:** To ensure systematic data ingestion, processing, transformation, and analytics using scalable cloud technologies.

**Content:**

- **Approach:** Medallion Architecture (Bronze, Silver, Gold) with iterative development
- **Techniques:** Data ingestion, data cleansing, deduplication, joins, aggregations, window functions, and analytics processing
- **Tools:** Python, PySpark, Azure Databricks, Azure Data Lake Storage, Delta Lake, MS SQL Server
- **Steps:** Requirement analysis, data ingestion, data processing, data transformation, analytics creation, testing, and deployment

## Proposed Work Flow

| Phase | Description | Duration (Weeks) | Start Week | End Week |
|---|---|---|---|---|
| Phase 1 | Requirement Analysis | 1 | 1 | 1 |
| Phase 2 | Literature Review (Cloud Data Lakes & Big Data Analytics) | 1 | 2 | 2 |
| Phase 3 | Feasibility Study | 1 | 3 | 3 |
| Phase 4 | System & Architecture Design (Medallion Architecture) | 2 | 4 | 5 |
| Phase 5 | Development (Data Ingestion, PySpark Processing, Bronze–Silver–Gold) | 4 | 6 | 9 |
| Phase 6 | Testing & Data Validation | 2 | 10 | 11 |
| Phase 7 | Documentation & Deployment | 1 | 12 | 12 |

## Conclusion / Expected Outcome and Benefits to Society

- The proposed system provides a scalable and efficient solution for managing and analyzing large volumes of retail data using cloud-based technologies. It enables organizations to transform raw data into meaningful insights, supporting data-driven decision-making. The platform improves data quality, transparency, and accessibility, benefiting businesses, analysts, and stakeholders.
- By leveraging modern data engineering practices such as Medallion Architecture, big data processing, and cloud analytics, the project demonstrates the practical application of data engineering and cloud computing technologies. This system can support the retail and e-commerce sector by improving sales analysis, customer behavior understanding, operational efficiency, and strategic planning, thereby contributing to digital transformation and economic growth.

## Team Members

**Student-1 Name:** Aakib Khan **Registration No:** 13766
**Role:** Code and Project Developer

**Student-2 Name:** Aditya Singh **Registration No:** 15907
**Role:** Documentation & Development

**Student-3 Name:** Aditya Singh **Registration No:** 14211
**Role:** Deployment

## Guide Details

**Guide Name:** Ms. Ankita Kumari
**Designation:** Assistant Professor

## Bibliography

- Microsoft Azure. *Azure Data Lake Storage Documentation.* https://learn.microsoft.com/en-us/azure/storage/data-lake-storage/
- Databricks. *Delta Lake Documentation.* https://docs.databricks.com/delta/index.html
- Zaharia, M., Chowdhury, M., Das, T., et al. (2016). *Apache Spark: A Unified Engine for Big Data Processing.* Communications of the ACM, 59(11), 56–65.
- Olist E-Commerce Dataset. *Olist Public Dataset on Kaggle.* https://www.kaggle.com/olistbr/brazilian-ecommerce
- Ghodsi, A., et al. (2017). *The Medallion Architecture: Delta Lake Design Patterns for Data Lakes.* Databricks Technical Blog.
- Microsoft SQL Server. *SQL Server Documentation.* https://learn.microsoft.com/en-us/sql/sql-server/