REQUIREMENT DOCUMENT

DATA LAKE ANALYTICS

1. INTRODUCTION

1.1 PURPOSE

The purpose of this project is to design and implement a Data Lake Analytics solution using Azure services. The solution aims to centralize raw data from multiple sources, transform it into curated datasets, and enable advanced analytics, reporting, and machine learning. This will empower the organization with faster insights, improved decision-making, and scalable data management.

1.2 BACKGROUND

In the retail/e-commerce industry, vast amounts of structured and unstructured data are generated from:

- Customer interactions
- Sales transactions
- Product catalogs
- Social media engagement
- Supply chain operations

Traditional systems struggle with storing and processing this variety and volume of data. By leveraging Azure Data Lake Storage, Azure Data Factory, and Azure Databricks, the organization can build a modern cloud-based data platform that ensures:

- Cost-effective storage
- Scalable data processing
- Advanced analytics and machine learning capabilities.

1.3 SCOPE

1.3.1 In Scope

- Building a centralized Data Lake in Azure.
- Using Azure Data Factory for data ingestion and orchestration.
- Creating curated datasets for reporting and analytics.
- Integrating with Power BI for dashboarding and visualization.
- Implementing basic security and role-based access control for data governance.

1.3.2 Out of Scope

- On-premises data warehouse modernization.
- Real-time streaming ingestion (only batch ingestion considered initially).
- Advanced ML model deployment (only data preparation included).
- Third-party tool integration outside Azure ecosystem.

2. BUSINESS REQUIREMENTS

2.1 BUSINESS OBJECTIVES

1. Centralize sales data into one reliable platform (Bronze, Silver, Gold).
2. Improve data quality with cleaning and standardization.
3. Provide analytics-ready data for decision-making and dashboards.

2.2 BUSINESS OUTCOMES

- Single source of truth for retail sales analysis.
- Faster insights with automated daily/weekly dashboards.
- Better understanding of top products and high-value customers.

2.3 SUCCESS CRITERIA

- Pipelines achieve ≥ 99% success rate.
- ≥ 95% of data is valid and clean in Silver zone.
- Power BI dashboards match source data and are used by key stakeholders.

3. REQUIREMENTS BREAKDOWN

3.1 FUNCTIONAL REQUIREMENTS

Defines what the system should do – the core features and functionalities (e.g., data ingestion, transformation, reporting).

ID	Requirement	Description
FR1	Data Ingestion	Ingest data from multiple structured (SQL, CSV, APIs) and unstructured (JSON, logs) sources.
FR2	ETL/ELT Pipelines	Automate data workflows using Azure Data Factory.

FR3	Data Storage	Store ingested data in Azure Data Lake Storage (ADLS) with raw, curated, and analytics zones.
FR4	Data Transformation	Process, clean, and enrich data using Azure Databricks notebooks.
FR5	Curated Datasets	Provide ready-to-use datasets for analytics and BI tools.
FR6	Metadata Management	Maintain schema, lineage, and catalog for datasets.
FR7	User Access Control	Enable role-based access (Admin, Analyst, Engineer, Business User).

3.2 NON-FUNCTIONAL REQUIREMENTS

Non-Functional Requirement: Defines how the system should perform – quality attributes like scalability, security, performance, and reliability.

ID	Requirement	Description
NFR1	Scalability	System should handle increasing data volumes without performance loss.
NFR2	Reliability	Ensure fault tolerance and recovery for data pipelines.
NFR3	Performance	Optimize Spark jobs for faster data processing.
NFR4	Security	Implement RBAC, encryption (at rest & transit), and compliance policies.
NFR5	Cost Optimization	Use serverless/on-demand compute where possible.
NFR6	Usability	Provide seamless integration with Power BI for end users.

3.3 REQUIREMENTS PRIORITIZATION

The requirements are prioritized into High, Medium, and Low, based on their impact on project success.

Priority Level	Requirements
High Priority	Data ingestion, Data storage, Data transformation, Curated datasets for BI

Medium Priority	Metadata management, Security & Role-Based Access Control (RBAC)
Low Priority	Advanced ML preparation, Integration with third-party tools

4. DELIVERABLES

1. Data Lake Zones – Azure Data Lake Storage structured into Bronze (raw), Silver (cleaned), and Gold (analytics) layers.
2. ADF Pipelines – Automated pipelines for data ingestion, cleansing, transformation, and aggregation.
3. Analytics Layer – Synapse SQL views and Power BI dashboards showing KPIs (sales trends, top products, top customers).

5. CONCLUSION

The Data Lake Analytics project successfully implemented a Medallion Architecture (Bronze, Silver, Gold) using Azure Data Factory (ADF), Azure Data Lake Storage (ADLS), and Power BI.

- Raw retail data was ingested into Bronze, cleansed and standardized in Silver, and aggregated into Gold for analytics.
- Automated ADF pipelines ensured reliable ingestion, transformation, and governance of data.
- Power BI dashboards delivered clear business insights, including sales trends, top products, customer analysis, and revenue growth patterns.
- The solution achieved its goals of centralizing data, improving quality, enabling decision-making, and reducing manual effort.

This project demonstrates a scalable, cost-efficient, and business-ready data platform that can be extended in the future with real-time ingestion, advanced analytics, and AI/ML use cases.