Technical Product Requirement Document: Databricks

I. Introduction

Purpose

Databricks is a comprehensive cloud platform that streamlines data analytics and machine learning by enabling seamless integration between data engineering and data science workflows. It leverages Apache Spark to efficiently process large volumes of data, providing users with a unified workspace for collaborative data analysis, ETL processes, and real-time machine learning model development. The platform's ability to scale resources dynamically ensures that organizations can handle growing data needs while improving operational efficiency and decision-making.

Scope

- **In-Scope:**
  - Implementation of Databricks as the central data processing and analytics platform.
  - Development of ETL pipelines to process data from multiple sources.
  - Collaboration capabilities through shared notebooks for team-based data analysis.
  - Integration with machine learning libraries to support data science workflows.

- **Out-of-Scope:**
  - Migration of legacy systems and data into Databricks.
  - Development of custom machine learning models (focus is on the platform, not specific models).
  - Support for non-cloud environments (Databricks is deployed on the cloud).

II. Project Overview

Background

- **Product Name:** Databricks Unified Analytics Platform
- **Product Description:** Databricks is a cloud-based platform for big data processing, machine learning, and analytics. It offers a collaborative environment for data engineers, data scientists, and business analysts to work on large-scale data projects using Apache Spark.

Objectives

- Enhance operational efficiency through automation.
- Improve data accuracy and reporting.
- Provide a user-friendly interface for end-users.
- Ensure compliance with industry regulations.

III. Functional Requirements

Functional Requirements:

- **Data Ingestion:** The platform should support batch and real-time data ingestion from various sources, including cloud storage (AWS S3, Azure Blob Storage) and databases.
- **Collaborative Workspaces:** Databricks must allow multiple users to collaborate using shared notebooks, supporting Python, SQL, Scala, and R languages.
- **Data Processing:** It must support Apache Spark for distributed computing to process large datasets.

- **Machine Learning Integration:** Databricks must integrate with popular machine learning libraries like TensorFlow, scikit-learn, and Databricks' MLlib for model development.
- **ETL Automation:** The platform must enable automated Extract, Transform, and Load (ETL) pipelines to process and analyze data.

Non-Functional Requirements:

- **Scalability:** Databricks must scale up and down to handle data volumes ranging from terabytes to petabytes.
- **Performance:** The platform should provide low-latency processing for real-time analytics and handle large-scale data processing efficiently.
- **Security:** Databricks must offer robust security features, including Single Sign-On (SSO), role-based access control, and encryption of data at rest and in transit.
- **Compliance:** The platform must comply with relevant data governance and regulatory requirements such as GDPR and HIPAA.

Stakeholders

- **Data Engineers:** Responsible for setting up and maintaining data pipelines.
- **Data Scientists:** Use Databricks for developing machine learning models and performing data exploration.
- **Business Analysts:** Analyze data and generate reports to support business decisions.
- **IT Security Team:** Ensures the platform complies with security and regulatory standards.
- **Executives:** Use insights from the platform to make strategic business decisions.IV. Technical Specifications

Constraints

- **Cloud Provider Limitations:** The platform is dependent on the capabilities and limitations of the cloud provider chosen (e.g., AWS, Azure).
- **Budget Constraints:** Costs related to cloud resources may limit the scope of the project (e.g., scaling for very large datasets).
- **User Training:** Team members will need training on using Databricks effectively.

System Architecture

- **Cloud Infrastructure:** Databricks will be hosted on AWS/Azure. It will integrate with cloud storage services (e.g., S3, Azure Data Lake) for data ingestion and storage.
- **Data Flow:** Data will flow from cloud storage to Databricks for processing, and the processed data will be stored in cloud databases or sent to other analytics tools for reporting.
- **Integration with Other Tools:** Databricks will connect with BI tools like Power BI or Tableau for visualizing processed data.

Success Metrics

- **Data Processing Time:** Reduced time to process large datasets by 40% compared to previous tools.
- **Collaboration Efficiency:** Increased collaboration among teams through shared notebooks, with a 30% improvement in project completion time.
- **Scalability:** The ability to scale data processing operations to handle 5x the current data volume without performance degradation.

Brittany Jones
Febuary 15, 2024