Alternative Cloud-Native Data Platform Architecture on Microsoft Azure

This guide details the implementation of a modern, cloud-native data platform architecture on Microsoft Azure, focusing on services such as **Azure Synapse Analytics**, **Snowflake**, **Azure Databricks**, and other Azure-native components for robust data processing, storage, transformation, and analytics capabilities.

Table of Contents

---

1. Core Architecture Overview

This architecture leverages Azure's managed, serverless, and scalable services to:

- Enable **real-time data processing** and **advanced analytics**.
- Minimize operational overhead and infrastructure management through cloud-native solutions.
- Utilize **Snowflake** as the data warehouse and **Azure Databricks** for big data processing and machine learning workflows, providing flexibility for both structured and unstructured data.

---

2. Recommended Architecture Components

2.1 OLTP Database (Transactional Layer)

- **Azure SQL Database**: Use Azure SQL Database as the main OLTP database, offering scalability and high availability.

2.2 NoSQL Database (Catalog Data)

- **Azure Cosmos DB**: Store catalog and product metadata in Cosmos DB, which supports low-latency, high-throughput operations for distributed data.

2.3 Staging Layer (Real-Time Stream Processing)

- **Azure Event Hubs**: Capture and process real-time streaming data with Azure Event Hubs.

2.4 Data Warehouse

- **Snowflake on Azure**: Use Snowflake for scalable, high-performance data warehousing, optimized for analytics and easy integration with Azure services.

2.5 Data Lake Storage

- **Azure Data Lake Storage Gen2 (ADLS Gen2)**: Use ADLS Gen2 as the data lake to store raw and processed data, accessible by both Snowflake and Databricks for a unified data repository.

2.6 Big Data and Machine Learning Platform

- **Azure Databricks**: Employ Databricks for scalable data processing, analytics, and machine learning. Utilize Databricks' Delta Lake to provide lakehouse architecture features.

2.7 Data Orchestration

- **Azure Data Factory**: Use Azure Data Factory for orchestrating ETL processes and data workflows.

2.8 Business Intelligence

- **Power BI**: Utilize Power BI for data visualization, directly connecting it to Snowflake and Databricks for real-time analytics.

---

3. Detailed Step-by-Step Implementation

Module 1: Transactional Database Design and Real-Time Ingestion

1. **Set Up Azure SQL Database for OLTP**:
    - Create an **Azure SQL Database** instance to manage transactional data.
    - Configure firewall settings to allow access from necessary Azure resources and your local IP.
2. **Enable Change Data Capture (CDC)**:
    - Enable CDC on the necessary tables in Azure SQL Database.
3. **Stream Real-Time Changes with Azure Event Hubs**:
    - Configure **Azure Data Factory** or **SQL Data Sync** to stream real-time data changes from Azure SQL Database to **Azure Event Hubs**.

Module 2: NoSQL Catalog Database Setup

1. **Set Up Azure Cosmos DB for Metadata**:
    - Create **Azure Cosmos DB** and configure it to support high-volume catalog data.
    - Select the **API** based on data access requirements (e.g., SQL API for JSON data, Cassandra API for key-value data).
2. **Integrate Cosmos DB with Data Lake**:
    - Use **Azure Data Factory** to export data periodically from Cosmos DB to **ADLS Gen2** for historical analysis and integration with the data lake.

Module 3: Data Lake and Data Warehouse Architecture

1. **Create an ADLS Gen2 Storage Account**:
    - Set up an **Azure Data Lake Storage Gen2** account to serve as the data lake.
    - Organize the storage with appropriate folder structures (e.g., raw, curated, enriched).
2. **Set Up Snowflake on Azure**:
    - Provision a **Snowflake** account on Azure and connect it to ADLS Gen2 for seamless data access.
    - Use **Snowflake's External Tables** feature to directly query data stored in ADLS Gen2 without loading it into Snowflake.
3. **Establish Lakehouse Architecture with Delta Lake**:
    - Use **Azure Databricks Delta Lake** to enable ACID-compliant transactions and create a unified data layer for real-time and historical analytics.

Module 4: Data Transformation and Processing

1. **Data Ingestion and Transformation with Azure Databricks**:
    - Set up **Azure Databricks** workspaces for data engineering and machine learning.
    - Use **Databricks Delta Lake** to transform raw data into refined datasets.
2. **Process and Enrich Data in Snowflake**:
    - Leverage **Snowflake's SQL** capabilities for data transformation and analytics on structured data.
3. **Implement Data Cleansing and Transformation in Azure Data Factory**:
    - Configure **Azure Data Factory** pipelines for batch ETL processes, cleansing, and joining datasets before loading them into the data warehouse.

Module 5: Orchestration and Workflow Management

1. **Orchestrate ETL Pipelines with Azure Data Factory**:
    - Use **Azure Data Factory** for automating and scheduling ETL workflows.
    - Set up data pipelines that connect **Azure SQL Database**, **Cosmos DB**, **Event Hubs**, and **ADLS Gen2** for data movement and transformation.
2. **Use Databricks Notebooks for Real-Time Data Processing**:
    - Develop **Databricks Notebooks** to handle real-time data transformations and analytics on event-driven data from Event Hubs.

Module 6: Machine Learning Model Deployment

1. **Setup Azure Databricks MLflow for Model Management**:
    - Use **Databricks MLflow** within Azure Databricks for model experimentation, tracking, and management.
2. **Deploy Models to Production with Azure Machine Learning**:
    - Register models in **Azure Machine Learning** and deploy them as web services.
    - Configure these services to connect with Snowflake and Power BI for real-time model inference.

Module 7: Business Intelligence and Visualization

1. **Create Dashboards in Power BI**:

    - Connect **Power BI** to Snowflake and Databricks for interactive data visualizations.
    - Use **DirectQuery** mode in Power BI to enable near real-time reporting on Snowflake data.
2. **Integrate Power BI with Azure Synapse Analytics**:
    - Connect **Power BI** with Synapse for additional analytical power and near real-time insights.

---

4. Benefits of the Modernized Architecture

- **Scalability and Cost Efficiency**: Azure's cloud-native services scale dynamically with demand, and Snowflake's pay-as-you-go pricing reduces costs for sporadic workloads.
- **Real-Time Processing**: Azure Event Hubs and Databricks Delta Lake support real-time streaming and processing, offering up-to-date analytics and insights.
- **Operational Simplification**: Managed services such as Azure Data Factory and Databricks reduce the need for manual infrastructure management and maintenance.
- **Unified Lakehouse Architecture**: Integrating Snowflake with Databricks Delta Lake provides a seamless platform for structured and unstructured data.
- **Comprehensive ML and BI Capabilities**: Combining Databricks, Azure Machine Learning, and Power BI allows for advanced analytics and machine learning model deployment at scale.

This architecture provides a robust, flexible, and scalable platform for end-to-end data management, analytics, and machine learning on Microsoft Azure.