D607 TASK 1

DATA ANALYTICS AT SCALE- D609

5/25/2025

WESTERN GOVERNORS' UNIVERSITY
Zainab Abbas

Big Data Architecture Proposal: STEDI Human Balance Analytics

Executive Summary

The purpose of this proposal is to design and implement a scalable big data lakehouse architecture for STEDI's Human Balance Analytics project using Microsoft Azure services. Due to contractual obligations, the organization must utilize Azure platforms for all data analytics operations. As a data engineer for STEDI, my objective is to develop a secure and efficient infrastructure that supports the ingestion, transformation, curation, and querying of sensor data from the STEDI Step Trainer and its companion mobile app. Azure services such as Synapse Analytics, Data Factory, and Data Lake Storage Gen2 offer a robust and integrated platform for handling big data workflows, making them ideal for this initiative (Microsoft, n.d.-a).

The proposed architecture will facilitate ethical and compliant data curation by filtering customer data based on research consent. It will also enable robust analytics, improve operational insights, and position STEDI as an innovator in health-tech by leveraging Azure's scalable ecosystem. The solution is expected to deliver highly curated datasets suitable for machine learning, allowing for the detection of human balance patterns and anomalies with precision (Microsoft, n.d.-b).

Business Problem Recap

STEDI's core offering is a Step Trainer device and mobile app that collect motion data to help users improve balance. However, challenges such as flawed serial number assignment during device fulfillment and varied customer consent for research use present significant barriers. The data collected is semi-structured and high in volume, and without a scalable infrastructure, it

becomes difficult to perform real-time analytics or model training. Additionally, regulatory and ethical constraints require strict filtering based on user consent, which must be enforced through the data pipeline (Udacity, n.d.). Without these capabilities, STEDI risks producing unreliable insights and violating user trust.

Needs Assessment

Big data analytics provides the tools needed to overcome the challenges presented by STEDI's current data environment. Azure Synapse Analytics, Azure Data Lake Storage, and Azure Data Factory can be used to create automated workflows that extract, filter, and prepare data for analysis and model training. These services provide seamless integration and high throughput, with the ability to process semi-structured data at scale (Microsoft, n.d.-a). Synapse's Spark pools allow for large-scale data processing using both Python and SQL interfaces, ideal for the transformations required for STEDI's data (Microsoft, n.d.-c).

To manage STEDI's dataset, distributed data processing using Synapse Spark pools and schema-on-read capabilities via Data Lake Gen2 will be implemented. These features support dynamic data exploration without requiring pre-defined schemas, making them ideal for handling diverse formats like JSON from the mobile app and sensor devices. Azure Data Factory will orchestrate ETL processes to clean and align customer, accelerometer, and step trainer records, ensuring that only valid and consented data is retained for downstream use (Microsoft, n.d.-b; Udacity, n.d.).

Solution Design

To address the identified needs, a lakehouse architecture will be developed using Microsoft Azure services. Azure Data Lake Storage Gen2 will function as the central repository for storing

raw, trusted, and curated datasets (Microsoft, n.d.-b). Azure Synapse Analytics will serve as the data processing engine for both batch and interactive querying (Microsoft, n.d.-a). Azure Data Factory will handle orchestration of data pipelines, enabling automated and scalable ETL workflows. Azure SQL Database will serve as a structured output layer, while Power BI will support business intelligence and stakeholder reporting (Microsoft, n.d.-d).

The data model follows a zoned architecture. The Landing Zone ingests raw JSON data from mobile and device sources. Data Factory pipelines apply filters based on customer research consent and store processed data in the Trusted Zone. Curated datasets are created through transformations and joins performed in Synapse and are stored in a Curated Zone within Data Lake Storage or exported to Azure SQL Database for analytics. Power BI dashboards will draw from these curated sources to provide actionable insights (Microsoft, n.d.-a; Microsoft, n.d.-d).

Justification of Choices

Microsoft Azure is a required platform based on organizational agreements. Azure Synapse Analytics provides versatile processing capabilities, including support for Apache Spark and T-SQL, making it a suitable choice for handling large volumes of data with varying structures (Microsoft, n.d.-a). Data Lake Storage Gen2 supports hierarchical namespace and is optimized for analytical workloads, which aligns well with the requirements of a zoned data lakehouse architecture (Microsoft, n.d.-b).

Power BI is selected for visualization due to its native integration with the Azure ecosystem and its ability to create customizable dashboards (Microsoft, n.d.-d). Azure SQL Database is leveraged for storing cleaned and structured data that can be queried easily by analysts. These

tools meet both technical and business needs, offering scalability, governance, and clarity in presentation (Microsoft, n.d.-d).

Future Enhancements

Several enhancements can further improve this architecture. Azure Stream Analytics can be introduced for real-time ingestion and processing of IoT data, enabling live monitoring of user balance activity. Azure Purview can be deployed to manage data lineage and enforce governance policies. Azure Machine Learning can be used to build, train, and deploy predictive models directly on curated datasets, improving STEDI's ability to derive insights and offer personalized recommendations (Microsoft, n.d.-c).

Implementation Plan

Implementation will begin by uploading the JSON data into Azure Data Lake Storage Gen2, organized into folders representing Landing, Trusted, and Curated zones (Microsoft, n.d.-b). Azure Data Factory pipelines will be developed to clean and transform landing data into trusted formats, applying filters to exclude non-consenting records (Microsoft, n.d.-a). Synapse Notebooks will perform the necessary joins and aggregations to create curated datasets. These curated outputs will be stored in Azure SQL Database and visualized using Power BI dashboards (Microsoft, n.d.-d). Version control and reproducibility will be ensured by storing all code and configuration in a Git repository (Udacity, n.d.).

References

Microsoft. (n.d.-a). Azure Synapse Analytics Documentation. https://learn.microsoft.com/en-us/azure/synapse-analytics/

Microsoft. (n.d.-b). Azure Data Lake Storage Gen2 Documentation. https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction

Microsoft. (n.d.-c). Azure Machine Learning Documentation. https://learn.microsoft.com/en-us/azure/machine-learning/

Microsoft. (n.d.-d). Power BI Documentation. https://learn.microsoft.com/en-us/power-bi/

Udacity. (n.d.). STEDI Human Balance Analytics Project Resources. https://github.com/udacity/nd027-Data-Engineering-Data-Lakes-AWS-Exercises