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 like 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.

Section 2: 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). For instance, Synapse's Spark pools allow for large-scale data processing using both Python and SQL interfaces, ideal for the types of transformations required for STEDI's data.

To manage STEDI's dataset, I will implement distributed data processing using Synapse Spark pools, along with schema-on-read capabilities via Data Lake Gen2. These features support dynamic data exploration without requiring data to be pre-structured, making it easier to handle 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 only valid and consented data is retained for downstream use (Microsoft, n.d.-b).

Section 3: Solution Design

To address the identified needs, I propose a lakehouse architecture built entirely within the Azure ecosystem. The core components include Azure Data Lake Storage Gen2 for staging and managing files, Azure Synapse Analytics for Spark-based transformations and SQL-based analytics, and Azure Data Factory for orchestration and workflow management. These components are well-integrated and allow for scalable processing of large datasets, with the flexibility to accommodate future enhancements (Microsoft, n.d.-a).

The data model mirrors the original AWS logic but implemented in Azure. Data from the mobile app and Step Trainer devices is ingested into the Landing Zone. ETL pipelines created using Azure Data Factory then sanitize the data based on customer research consent and populate the Trusted Zone. This ensures compliance with privacy standards and prepares the data for further enrichment. Curated datasets are then generated through joins in Synapse and written to curated storage containers, where they can be accessed by business users or exported to machine learning platforms (Microsoft, n.d.-b).

Justification of Choices

Microsoft Azure is mandated by contractual agreements and is fully capable of supporting the required data architecture. Azure Synapse Analytics supports both SQL-based and Spark-based processing, which is necessary for building robust and flexible ETL pipelines. Azure Data Lake Storage Gen2 is optimized for big data workloads, offering hierarchical namespace and POSIX compliance, which improve manageability and performance (Microsoft, n.d.-a).

Power BI is chosen for data visualization due to its deep integration with the Azure data stack and its ability to create intuitive dashboards tailored to various stakeholders. These tools align with STEDI’s need to scale data analytics operations, ensure ethical use of data, and deliver actionable insights to healthcare partners and end users alike (Microsoft, n.d.-b).

Future Enhancements

Future improvements can include integrating Azure Stream Analytics for near real-time ingestion from IoT Step Trainers. This will enable immediate feedback loops and enhance real-time monitoring capabilities. Implementing Azure Purview can enhance data cataloging and lineage tracking, while Azure Machine Learning will facilitate automated model training and deployment workflows. These services support rapid iteration and continuous improvement, essential for STEDI’s mission to drive innovation in balance analytics (Microsoft, n.d.-a).

Implementation Plan

The implementation will begin with uploading the raw JSON files to Azure Data Lake Gen2, organized into Landing Zone containers. Azure Data Factory pipelines will be created to ingest and process the data into the Trusted Zone, applying filters based on user consent. Synapse Notebooks will be used to filter, join, and curate datasets, which will then be saved to curated containers or Azure SQL Database. Power BI dashboards will be built using these curated datasets to communicate insights to stakeholders. All artifacts including scripts, transformation logic, and SQL code will be stored in a Git repository for version control and reproducibility (Microsoft, n.d.-b; 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

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

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

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2016). Apache Spark: The Definitive Guide. O’Reilly Media.