# READHUB CLOUD ARCHITECTURE: DATA PLATFORM FOR INTELLIGENT BOOKSTORE ANALYTICS

Article

By: Himani

# INTRODUCTION

We developed ReadHub, a cloud-native intelligent bookstore platform built to transform raw customer and book data into actionable insights. Leveraging a **Lakehouse architecture** on **Azure**, we have created a scalable and secure platform that processes both real-time and historical data. This platform enables businesses to drive key decisions using the insights generated from their data.

# MISSION AND OBJECTIVES

•**Mission**: Our goal is to deliver a unified, scalable, and secure platform designed to empower data analytics and AI capabilities.
•**Objectives**:
•Integrate diverse data sources to offer a comprehensive view of customer behavior and sales trends.
•Transform and organize data using Lakehouse layers to refine insights for various stakeholders.
•Enable BI, ML, and application-driven analytics by providing high-quality datasets for decision-making.

# KEY DATA SOURCES

We incorporated multiple data sources to power the analytics engine, including:

- **Customer Profiles**: Detailed user information, reading history, and preferences.
- **Sales Transactions**: Comprehensive records of book purchases and interactions.
- **Web Logs**: Real-time user activity, navigation patterns, and search behaviors.
- **Book Metadata**: Key information about each book, including title, author, genre, and ratings.
- **User Reviews**: Customer feedback and ratings that contribute to understanding sentiment.

# DATA PROCESSING AND TRANSFORMATION

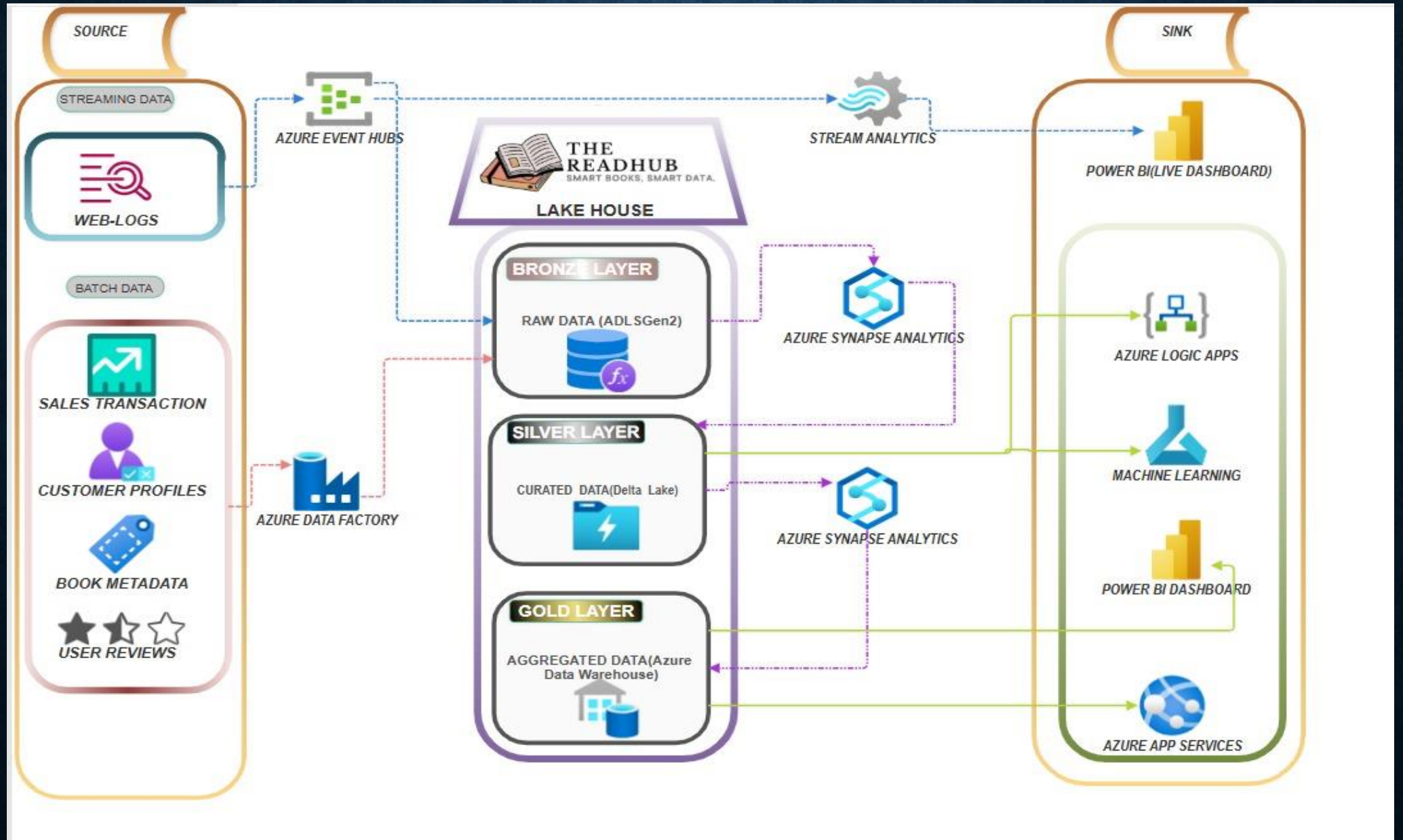
We designed the data processing pipeline to ensure that raw data is efficiently transformed into meaningful insights at every stage. Here's how we handled it:

**Bronze Layer: Ingest Data**

We used two methods for data ingestion:

- **Batch Ingestion**: For stable data sources like sales transactions, customer profiles, and reviews, we scheduled batch ingestion via **Azure Data Factory** to ensure regular updates.
- **Streaming Ingestion**: For real-time data such as web logs, we implemented **Event Hub** for continuous streaming, processed using **Stream Analytics**.
  All raw data was stored in **Azure Data Lake Gen2** under the Bronze Layer for auditing, recovery, and further processing.

# CLOUD ARCHITECTURE Model



# BRONZE LAYER: INGEST DATA

We used two methods for data ingestion:

- **Batch Ingestion**: For stable data sources like sales transactions, customer profiles, and reviews, we scheduled batch ingestion via **Azure Data Factory** to ensure regular updates.
- **Streaming Ingestion**: For real-time data such as web logs, we implemented **Event Hub** for continuous streaming, processed using **Stream Analytics**.
  All raw data was stored in **Azure Data Lake Gen2** under the Bronze Layer for auditing, recovery, and further processing.

# SILVER LAYER: CURATE DATA

Once data was ingested, we cleaned and structured it to make it usable for analysis. Key tasks included:

- Handling inconsistent schemas and formats.
- Removing duplicates and null values.
- Normalizing dates and IDs.
- Joining different datasets, such as combining sales data with customer profiles and reviews with book metadata. The curated data was stored in **Azure Data Lake Storage Gen2** and formatted using **Delta Lake**, ensuring high-quality and reliable data ready for downstream analytics.

# GOLD LAYER: AGGREGATE DATA

In this layer, we aggregated and summarized the curated data to produce business-ready datasets optimized for reporting, dashboarding, and machine learning. Key outputs included:

- **Sales Summary**: Revenue by genre, author, and sales trends.
- **Customer Segments**: Behavior-based user groups to help with targeted marketing.
- **Book Recommendations**: Personalized book suggestions based on previous customer interactions.
- **Review Sentiment Scores**: Insights into how customers feel about books based on review data.

# KEY BUSINESS OUTPUTS

The processed data supported several important use cases and generated business insights:

- **Sales Insights Dashboard**: We provided a live Power BI dashboard with trends in book sales, revenue by genre and author, and seasonal demand patterns.
- **Customer Segmentation**: By analyzing customer behavior, we grouped users based on demographics, reading habits, and purchase history. This enabled personalized marketing efforts using **Power BI** and **machine learning-driven segmentation**.
- **Book Recommendation Engine**: We built an Azure-based recommendation system that suggested personalized books to users based on their reading and review history. The recommendations were delivered through **Azure App Services API**.
- **Marketing and Campaign Analytics**: We generated detailed campaign performance reports that measured user engagement and promotional effectiveness. These insights were automated using **Azure Logic Apps**.
- **User Behavior Analytics**: We gained insights into user navigation, session durations, and bounce rates using **Power BI dashboards** and **Azure App Services APIs**.
- **Sentiment Analysis on Reviews**: We analyzed customer reviews for sentiment, providing positive, neutral, and negative feedback summaries and identifying emerging themes using **Power BI dashboards**.

# PIPELINE ORCHESTRATION AND AUTOMATION

We orchestrated the end-to-end data flow using **Azure Data Factory**. This included automating the process with a **master pipeline** that runs daily at 12:05 AM, initiating various sub-pipelines:

- Sales ingestion.
- Reviews ingestion.
- Metadata API.
- Web logs streaming. Each pipeline is designed to trigger sequentially, ensuring that each step only runs if the previous one succeeds.

# PIPELINE FAILURE HANDLING

In the event of a pipeline failure, we ensured that the system could automatically retry:

- **Retry Attempts**: The system is configured to retry up to 3 times.
- **Retry Interval**: Each retry occurs after an hour of delay.
  If all retries fail, the pipeline is marked as **failed** in **Azure Data Factory** or **Synapse Pipelines**. Our data engineering team then manually reviews the logs, diagnoses the root cause, and can re-run the pipeline on demand.

# CONCLUSION

Through the design of ReadHub's **Lakehouse architecture** on **Azure**, we successfully integrated batch and streaming data pipelines to process real-time and historical data. By utilizing **Synapse Analytics** and **Delta Lake**, we created reliable and actionable datasets that power dashboards, machine learning models, and APIs.

- The platform supports vital business use cases such as tracking **sales trends**, offering **personalized book recommendations**, and performing **sentiment analysis** on customer reviews. With scalability, automation, and machine learning readiness, ReadHub enables businesses to make data-driven decisions, positioning itself as a robust, future-proof solution.