ReadHub – Cloud-Native Intelligent Bookstore Analytics

A Modern Data Lakehouse Architecture on Azure

DINESHMURUGAN-LinkedIn

Executive Summary

Introduction

ReadHub is a cloud-native intelligent bookstore platform that enables real-time data-driven decision making. Its core objective is to unify diverse data sources—customer profiles, web logs, transactions, book metadata—and transform them into curated datasets that drive business intelligence, personalized recommendations, and predictive analytics.

Mission & Purpose

Mission: Deliver a unified, scalable, and secure data platform for analytics and AI.

Purpose: Provide real-time and historical insights to drive smarter business decisions.

Objectives

- Integrate structured and semi-structured data sources
- Organize data using Lakehouse architecture (Bronze, Silver, Gold)
- Deliver curated datasets for BI tools, ML models, and applications

Data Sources and Use Cases

Input Sources

- Customer Profiles: Reading history and preferences
- Sales Transactions: Purchase records
- Web Logs: User activity and behavior
- Book Metadata: Author, title, ratings
- User Reviews: Customer feedback

Key Use Cases

- Sales Insights Dashboard
- Customer Segmentation
- Book Recommendation Engine
- Campaign and Marketing Analytics
- Sentiment Analysis

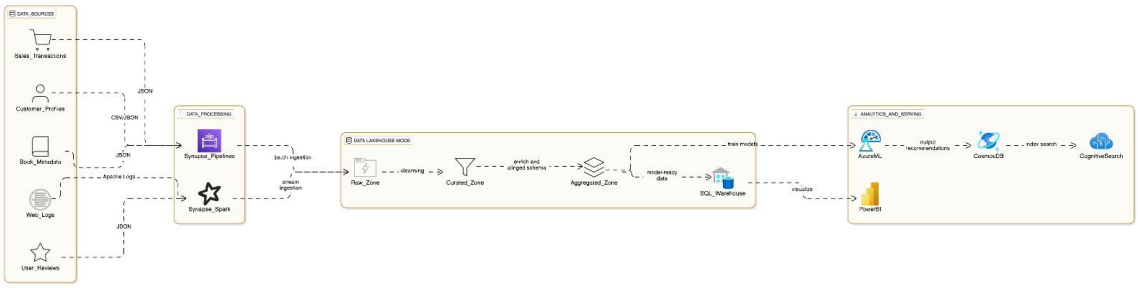
Lakehouse Architecture Phases

The architecture evolved in three phases:

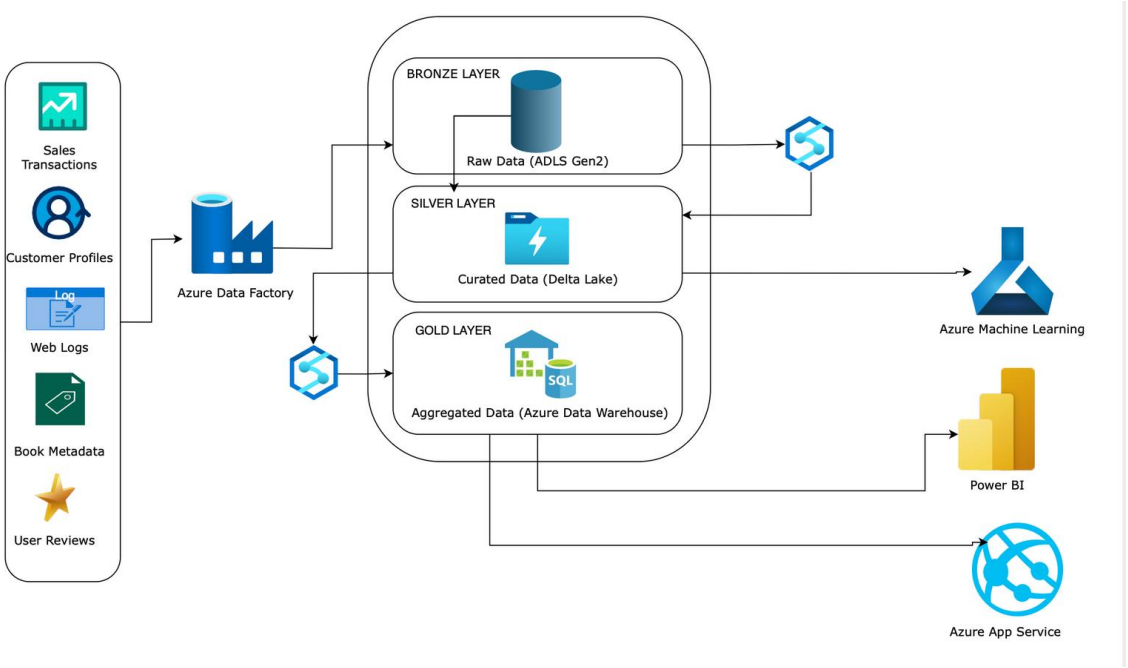
1. Phase 1: Initial Design – Basic ingestion and data flow blueprint
2. Phase 2: Refinement – Defined Lakehouse layers, schema validation
3. Phase 3: Final Design – Complete orchestration, failure handling, and automation

Each phase incrementally improved system reliability and analytical readiness.

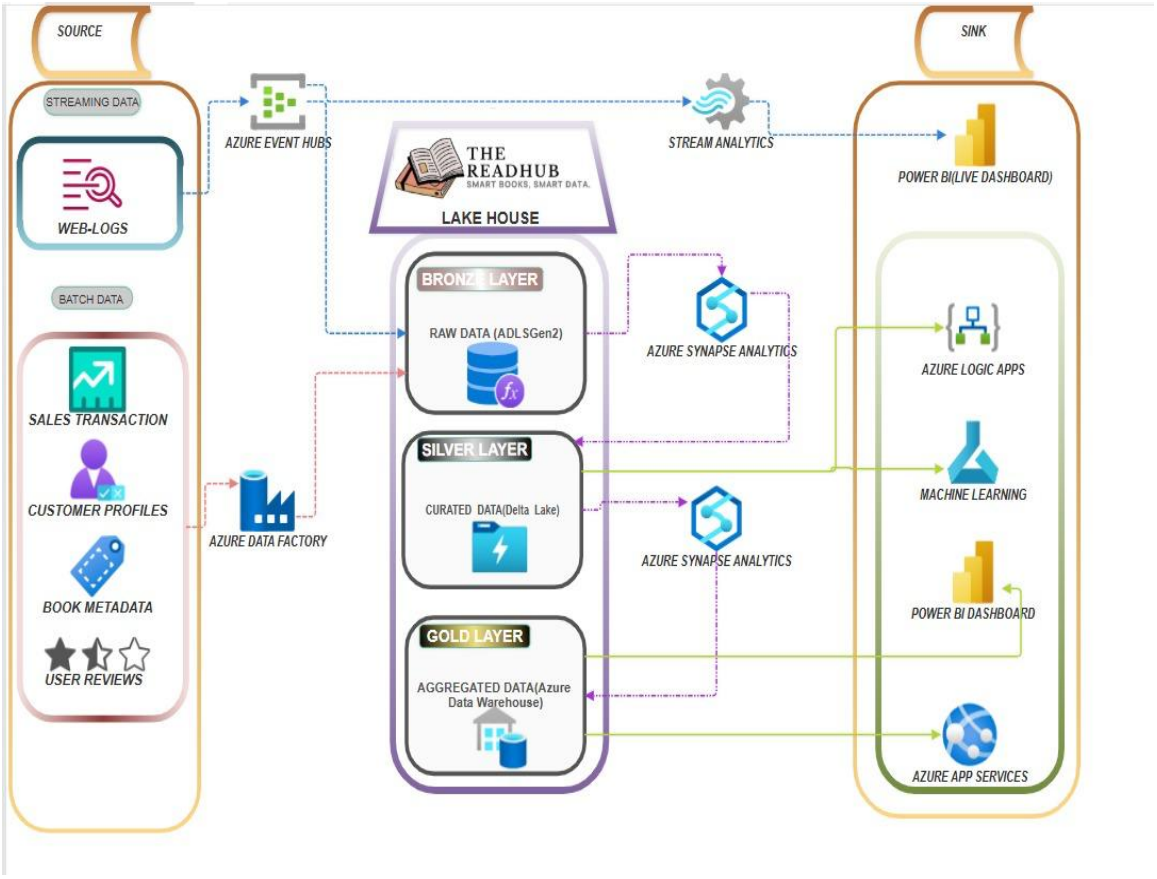
Phase -1:



Phase-2:



Phase-3(FINAL DESIGN):



Lakehouse Data Layers

Bronze Layer – Raw Ingestion

Handles batch (ADF) and streaming (Event Hubs) ingestion of data like sales, customer info, and logs. Stored in Azure Data Lake Gen2 under /bronze/.

Silver Layer – Curated Data

Performs cleaning, deduplication, and schema unification using Delta Lake format. Stored in /silver/ for analytics readiness.

Gold Layer – Aggregated Data

Provides summary tables for dashboards and ML models, hosted in Synapse SQL Pools.

Pipeline Orchestration

Automated pipeline flow managed by Azure Data Factory.
- Master pipeline runs daily at 12:05 AM
- Includes Sales Ingestion, Metadata API, Reviews and Web Logs
- Sequential execution with dependency chaining

This ensures reliable data delivery across all layers.

Pipeline Failure Handling

Failures can occur due to:
- Network issues
- Malformed files
- API rate limits
- Storage access errors

Retry Strategy:
- 3 retries
- 1-hour delay per retry

Persistent failures are logged for manual investigation and re-run.

Platform Deliverables

Dashboards (Power BI):
- Sales Trends
- Customer Segments
- Campaign Metrics
- Sentiment Analytics

APIs:
- Personalized Recommendations
- Behavior Analytics

Reports:
- Weekly Campaign Reports
- Engagement Metrics Summaries

Conclusion

The ReadHub platform leverages modern cloud-native architecture to unify batch and streaming pipelines. By utilizing Delta Lake, Synapse SQL, and Power BI, it delivers trusted, ML-ready datasets for business users. Its scalable and automated architecture is designed to support evolving analytics and AI use cases in the retail bookstore space.