Readhub Article
Written By: Amritpal Singh.

WHAT IS READHUB?

ReadHub is a cloud-native intelligent bookstore platform that transforms raw customer and book data into actionable insights using a modern Lakehouse architecture on Azure.

MISSION

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

OBJECTIVES

- Integrate diverse data sources
- Transform and organize data using Lakehouse layers
- Enable BI, apps, and ML through refined datasets

PURPOSE

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

DESIGN AND DISCOVERY PHASE

DATA SOURCES (INPUT)

- **Customer Profiles:** User information, reading history, and preferences
- **Sales Transactions:** Records of book purchases and interactions
- **Web Logs:** User activities, navigation patterns, and search behaviour
- **Book Metadata:** Title, author, genre, and ratings
- **User Reviews:** Customer feedback and ratings

DATA PROCESSING (TRANSFORMATION)

The ReadHub System acts as the central processing unit where raw data is transformed into actionable insights.

DATA UTILIZATION (OUTPUT)

Sales Insights Dashboard

- **Delivery Method:** Power BI (Live Dashboard), Scheduled Reports
- **What We Deliver:** Trends in book sales, revenue by genre/author, seasonal demand patterns

Customer Segmentation

- **Delivery Method:** Power BI dashboards, ML-driven segmentation reports
- **What We Deliver:** User grouping based on demographics, reading behavior, and history for personalization and marketing

Book Recommendation Engine

- **Delivery Method:** Azure App Services API, Integrated in Web/Mobile App
- **What We Deliver:** Personalized suggestions, related titles, curated lists

Marketing & Campaign Analytics

- **Delivery Method:** Power BI dashboards, Azure Logic Apps automation
- **What We Deliver:** Campaign performance, promotion effectiveness, engagement metrics

User Behavior Analytics

- **Delivery Method:** Power BI (Behavioral Dashboards), Azure App Services APIs
- **What We Deliver:** Insights on navigation, session duration, bounce rates, feature usage

Sentiment Analysis on Reviews

- **Delivery Method:** Power BI dashboards, Text Analysis Reports
- **What We Deliver:** Review sentiment summary (positive, neutral, negative), emerging themes, book-level scores

---

CLOUD ARCHITECTURE PHASES

PHASE 1 - INITIAL DESIGN

By Himani Makwana

PHASE 2 - REFINEMENT

By Himani Makwana

PHASE 3 - FINAL ARCHITECTURE

By Dinesh Murugan

PROCESS OF CREATING PIPELINE

BRONZE LAYER: INGEST DATA

- **Ingestion Types:** Batch and Streaming

- **Batch via Azure Data Factory:** Sales, Customer Profiles, Book Metadata, Reviews
- **Streaming via Event Hub:** Real-time Web Logs
- **Stored In:** Azure Data Lake Gen2 → /bronze/
- **Purpose:** Store raw, unprocessed data for auditing, recovery, and downstream processing

SILVER LAYER: CURATE DATA

- **Curation Steps:** Handle inconsistent schemas, remove duplicates/nulls, normalize dates/IDs, join datasets
- **Storage:** Azure Data Lake Storage Gen2
- **Format:** Delta Lake (ACID-compliant)
- **Purpose:** Clean, validate, and enhance data from Bronze Layer for analysis

GOLD LAYER: AGGREGATE DATA

- **Hosted In:** Azure Synapse SQL Pool
- **Optimized For:** Reporting, dashboarding, model training
- **Key Outputs:**
    - sales_summary
    - customer_segments
    - book_recommendations
    - review_sentiment_scores
- **Purpose:** Provide trusted, aggregated datasets for BI, ML, and APIs

---

PIPELINE ORCHESTRATION

- **Tool Used:** Azure Data Factory
- **Sub-pipelines:** Sales Ingestion, Reviews Ingestion, Metadata API, Web Logs Stream
- **Master Pipeline:** Runs daily at 12:05 AM
- **Execution Strategy:** Dependent, sequential execution (each step runs only if the previous one succeeds)

PIPELINE FAILURE HANDLING

By Amritpal Singh

A pipeline failure occurs when data cannot be processed, transferred, or loaded.

Causes:

- Network issues
- Incorrect formats (malformed JSON/CSV)
- Storage access errors
- API rate limits/timeouts

Retry Strategy:

- **Attempts:** Up to 3
- **Interval:** 1-hour between retries

If All Retries Fail:

- Pipeline is marked Failed in Azure Data Factory/Synapse
- Data engineers manually inspect logs
- Root cause diagnosed and pipeline re-run on demand

CONCLUSION

Conclusion (Personal Analysis by Amritpal Singh)

Through the ReadHub project, I analyzed how modern cloud-native architectures can transform raw bookstore data into meaningful business intelligence. By implementing a Lakehouse architecture on Azure, I gained hands-on experience in integrating both batch and streaming data sources and orchestrating them through automated pipelines.

One of my key takeaways was understanding the importance of layered data processing—starting from raw ingestion (Bronze), cleansing and structuring (Silver), and finally to aggregated insights (Gold). Tools like Azure Synapse, Delta Lake, and Data Factory played a crucial role in making the data trustworthy, scalable, and AI-ready.

I also focused on the pipeline failure handling mechanism, which deepened my understanding of error management in production-level systems. Overall, this project enhanced my skills in data engineering, pipeline orchestration, and designing cloud-based analytics systems that can power real-time decision-making in the retail domain.

CONTRIBUTIONS

- **Dinesh:** Phase 3 Final Architecture, Pipeline Design
- **Apple:** Mission & Objectives, Sources & Sink

- **Himani:** Phase 1 Initial Design, Phase 2 Refinement
- **Amritpal:** Pipeline Failure Handling

Thank you.