DataFlix
Data management for streaming

Meet the Producers

PRITESH SINGH

RAFFI MANNARELLI

VICTORIA LIU

SIMRAN KAUR

Project Overview

Data and Data tools

- We look at how data is **generated, stored and used.**
- What are the factors that decide **how** and **what** to data use?
- **User privacy** and other factors

Pipelines

- Our data stack and the reasons for our choices
- Overview of how the **data flows** within the organization

Analytics and beyond

- **Why** do we need to **collect and maintain** data
- What **business value** is derived from the metrics
- What **changes** can we propose based on our findings of the data

Next steps

- Other **use-cases or data strategies** used in a streaming service
- The future of analytics and data

Stakeholders

VIDEO PLATFORM TEAM →

Role: Develops and manages the core technology behind video streaming.

Data Use: Utilizes '*PlayEvents*' and '*BufferTime*' from the data to **optimize the video player stack**, ensuring smooth streaming with minimal buffering.

SUBSCRIPTION TEAM →

Role: Oversees **subscriber management** and service access control.

Data Use: Leverages '*Subscriptions*' data to manage user subscriptions effectively, tracking start and end dates, as well as monitoring subscription statuses for renewals or cancellations.

ENGAGEMENT TEAM

Role: Focuses on maintaining and increasing **viewer engagement.**

Data Use: Analyzes '*PauseEvents*', '*StartEvents*', and '*StopEvents*' to understand viewing patterns, helping to create targeted content recommendations and improving overall user engagement with the platform.

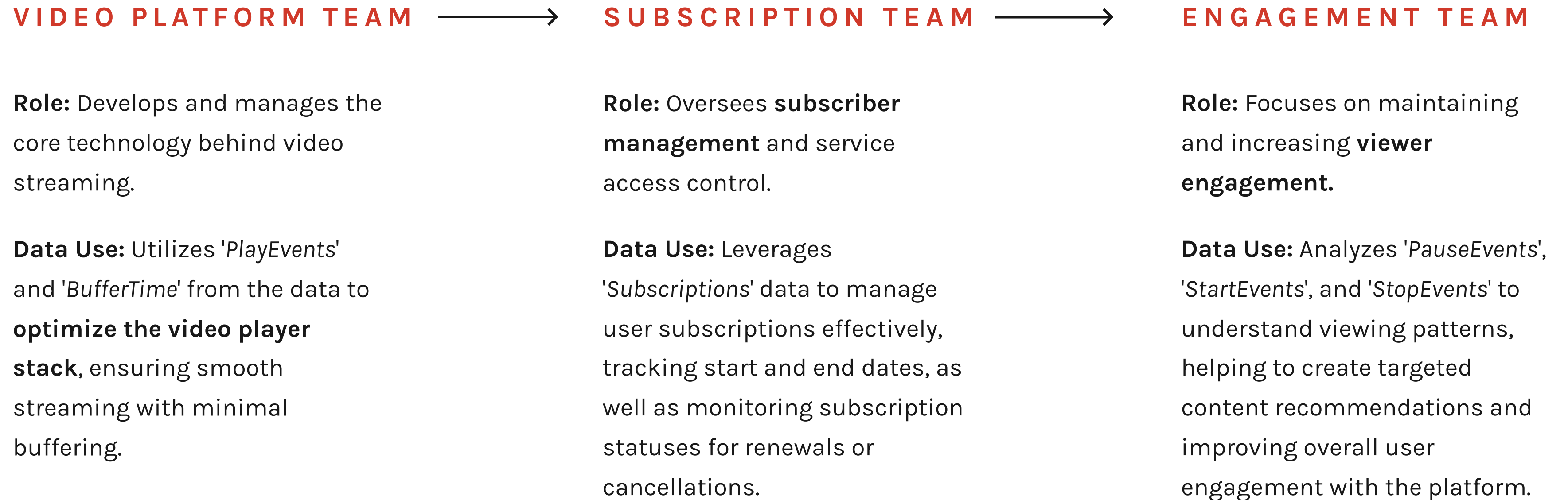Project Overview 01 Business Stakeholders 02 <u>Data Source 03</u> Data Strategy & Stack 04 NoSQL Pipeline 05 Results/Dashboarding 06 Conclusion & Next Steps 07

Data Source

DATA CREATION

The dataset we used was a fabricated representation structured to simulate actual user interaction analytics within a streaming service application such as Netflix or Hulu etc.

COMPOSITION

- **Video Playback Events**: Details of video play, pause, buffer and error instances and type including session IDs and timestamp.
- **User Subscriptions**: Subscription activity, covering the lifecycle from start to end date, with subscription types
- **Recommendations**: Data on content suggested to users, aligned with the algorithmic approach.
- **CMS**: Data related to the content, Langugage, rating, etc with ContentID as the key

AGENDA

Project Overview 01 Business Stakeholders 02 Data Source 03 Data Strategy & Stack 04 NoSQL Pipeline 05 Results/Dashboarding 06 Conclusion & Next Steps 07

Data Stack & Strategy

Azure for Data Storage

Microsoft Azure

Power BI for Dashboarding

Power BI

Azure Cosmos for NoSQL

Cost-Effective: Offers a range of options that fit various budgets without compromising on quality.xyz

High Performance: Provides rapid execution speeds and robust parallel computing capabilities.

Intuitive Interface: Power BI's user-friendly design makes it easy to create and share dashboards.

Integrated Analytics: Seamlessly integrates with Azure for consistent data analysis and reporting.

Flexible Data Model: Schema-less architecture allows for the storage and combination of various data types, ideal for handling streaming data.

Scalable Performance: Allows for real-time analysis, essential for managing large and growing datasets in streaming service environments.

Data Stack & Strategy
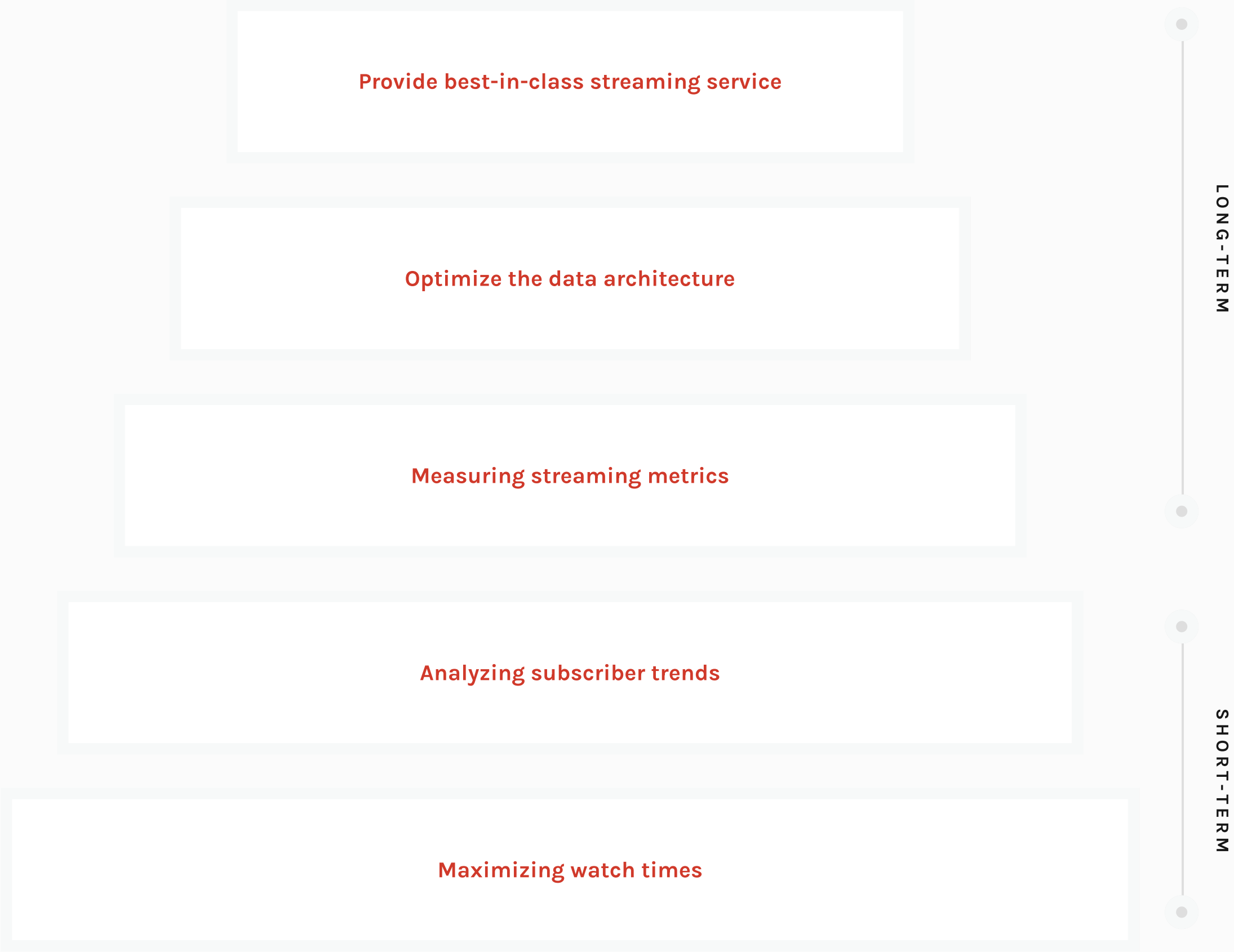
Main keywords

What is the motivation?

Insights at large a scale

Enhancing Product performance

User delight and retention

What are the benefits?

Excellent user experience

Predictive analytics

Targeted marketing

What are the risks?

Content piracy

Managing personal & payment data

Provide best-in-class streaming service

Optimize the data architecture

Measuring streaming metrics

Analyzing subscriber trends

Maximizing watch times

LONG-TERM

SHORT-TERM

Project Overview 01 Business
Stakeholders 02 Data Source 03 Data
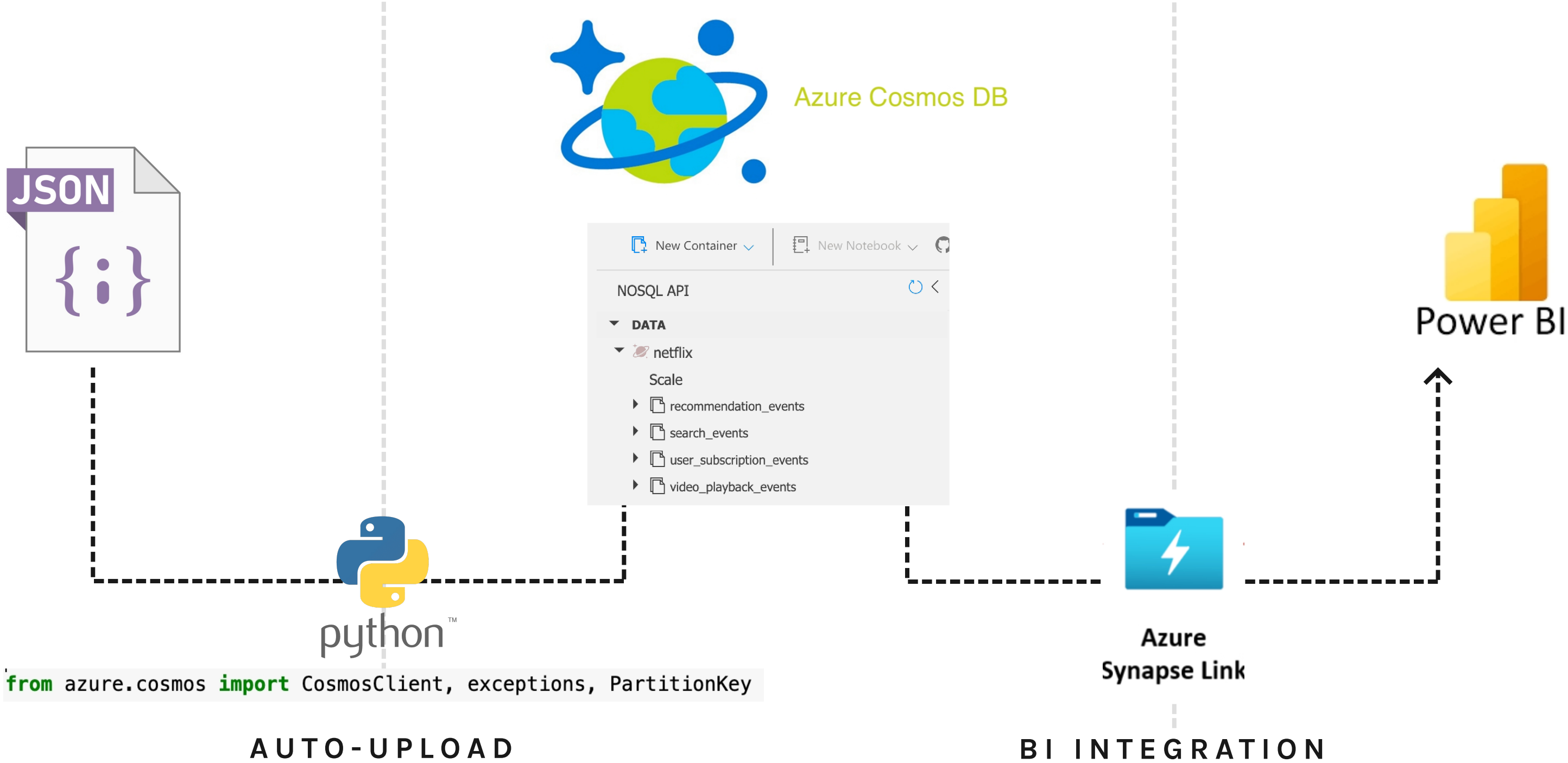Strategy & Stack 04 NoSQL Pipeline 05
Results/Dashboarding 06 Conclusion &
Next Steps 07

NoSQL Pipeline

DATA POPULATION   **NOSQL STORAGE LAYER**   **BUSINESS INTELLIGENCE**

AUTO-UPLOAD   **BI INTEGRATION**

Cosmos DB Highlights

SCALABILITY

Designed for horizontal scaling. Easy to handle more traffic and larger data sizes by distributing the load across multiple servers

MULTI-MODEL SUPPORT

Schema-less. Flexible data storage of multiple data types including document, key-value, graph, and column-family.

INDEXING & QUERYING

Automatically indexes all data without requiring explicit index management. Enable faster data retrieval through sql-like querying.

LOW LATENCY

It supports replicating data across multiple Azure regions, ensuring low-latency access to data regardless of where users are located.

COST-EFFECTIVENESS

Avoid continuous server operation, even when the workload is low or inconsistent. Pay only for the resources used.

AGILE DEVELOPMENT

The schema-less nature and the serverless databases enable faster development cycles and agility.

Project Overview 01 Business
Stakeholders 02 Data Source 03 Data
Strategy & Stack 04 NoSQL Pipeline 05
Results/Dashboarding 06 Conclusion &
Next Steps 07

PowerBI Dashboard
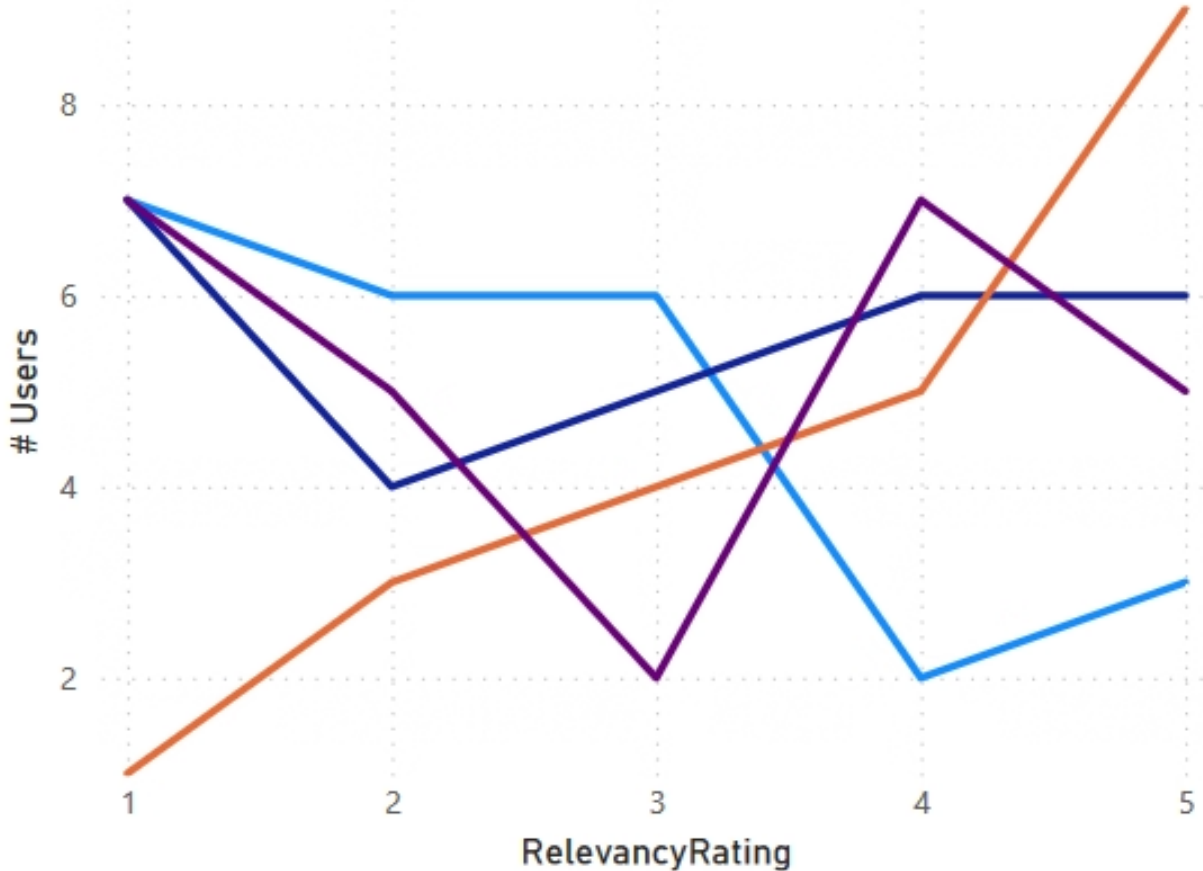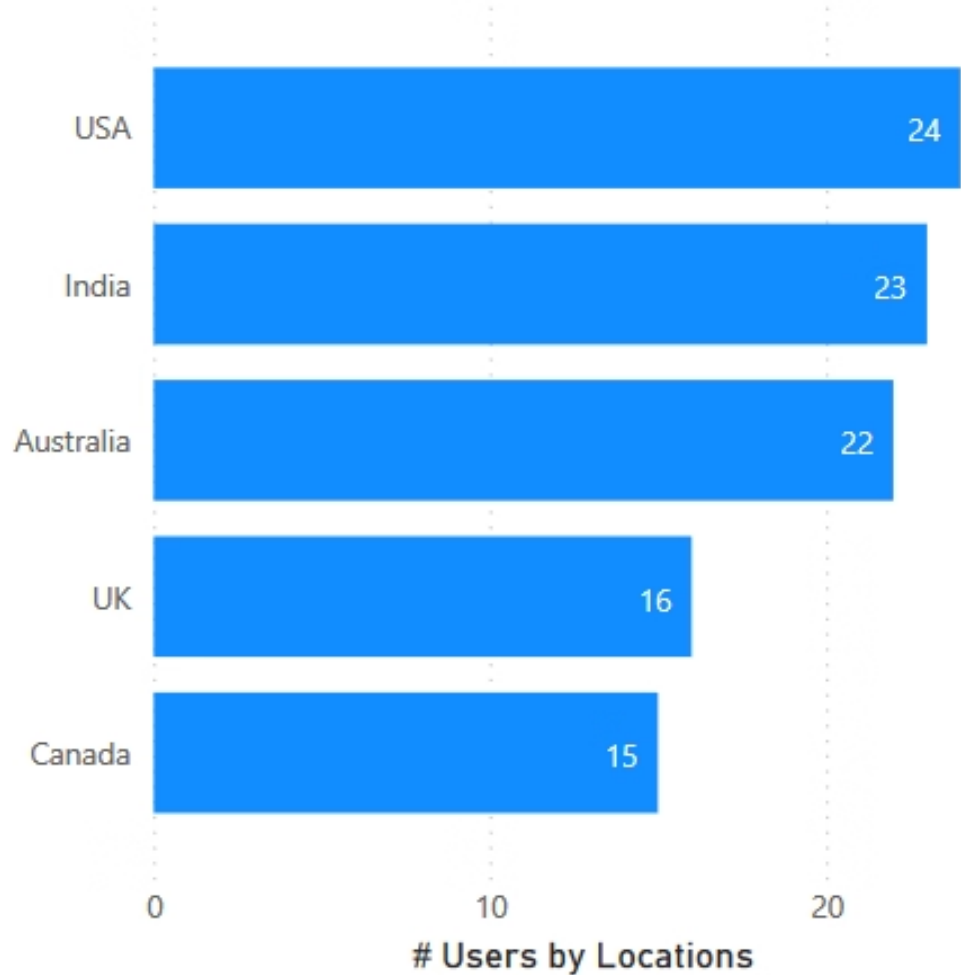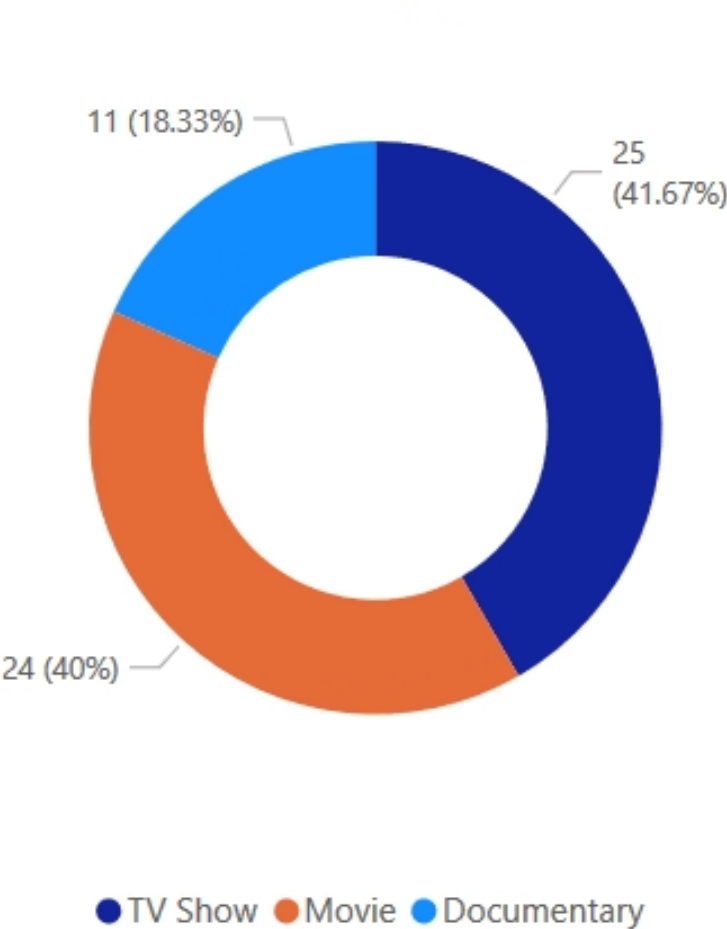
Error Analysis Dashboard

100
Events
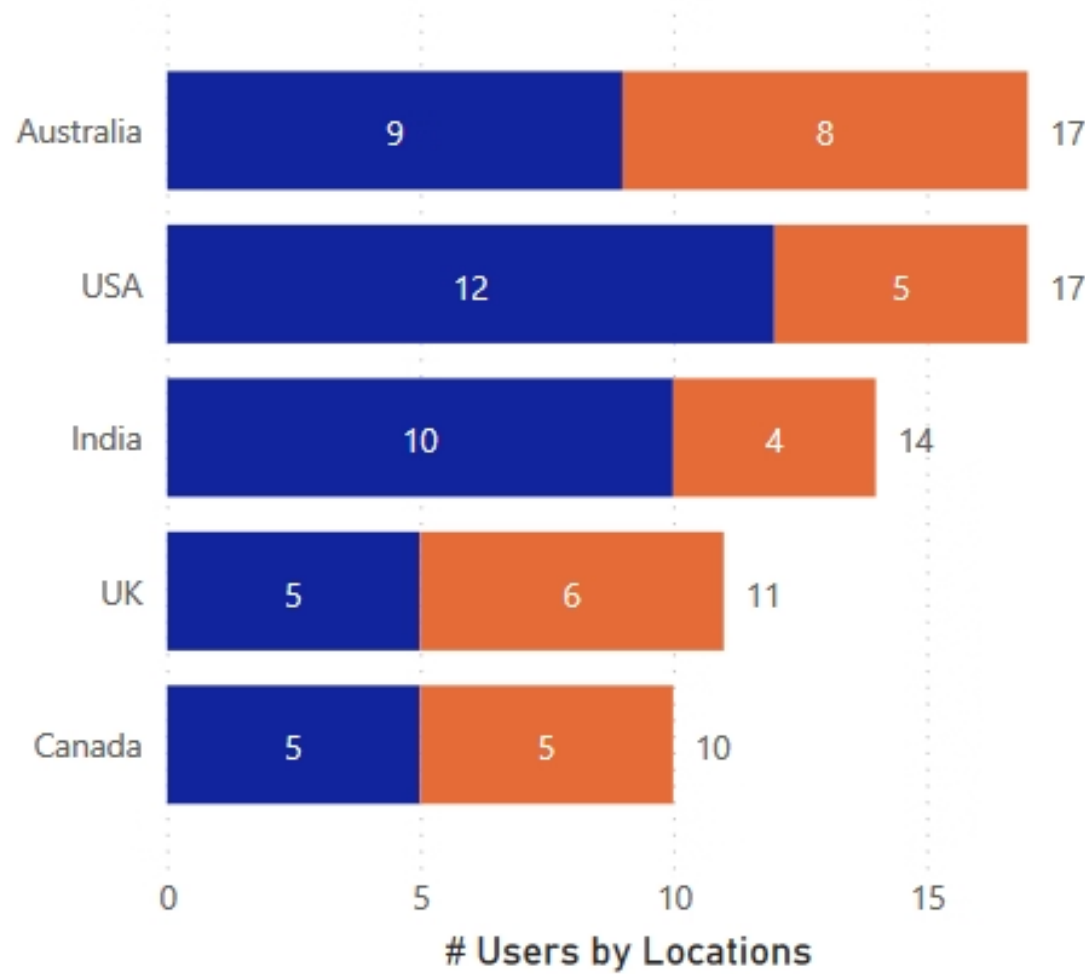
69
Errors

22
Buffers

Errors by Location

ErrorType ● Buffer Error ● Connection Error

Location	Buffer Error	Connection Error	Total
Australia	9	8	17
USA	12	5	17
India	10	4	14
UK	5	6	11
Canada	5	5	10

Users by Locations

Errors by Event Type

Play
Connection Error 12
Buffer Error 11
None 9

Buffer
Buffer Error 10
None 7
Conn...

Error
Buffer Error 14
None 8
Connection Error 4

Pause
Connection Error 7
None 7
Buffer Er...
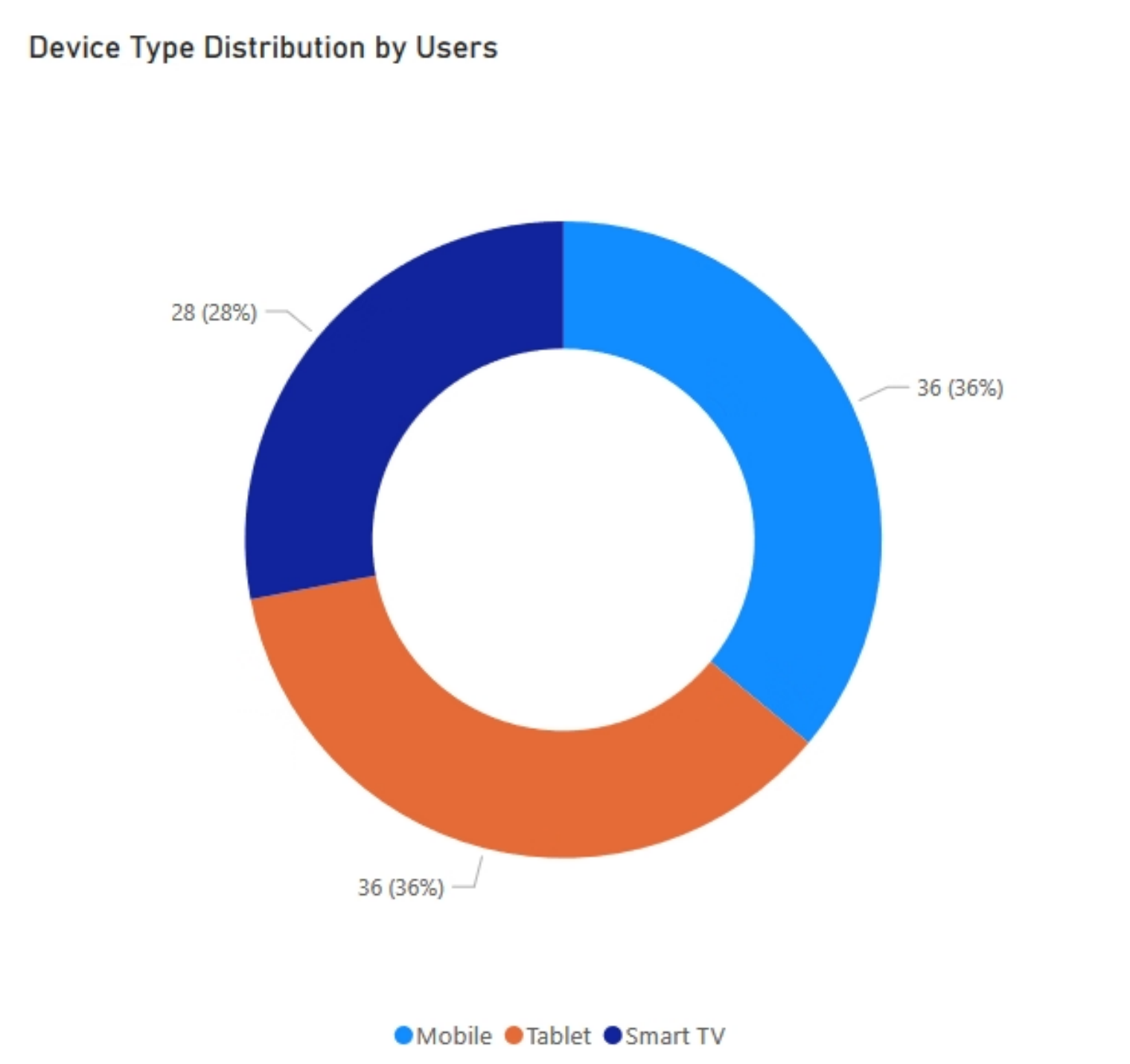
Revenue by SubscriptionType

Device Type Distribution by Users
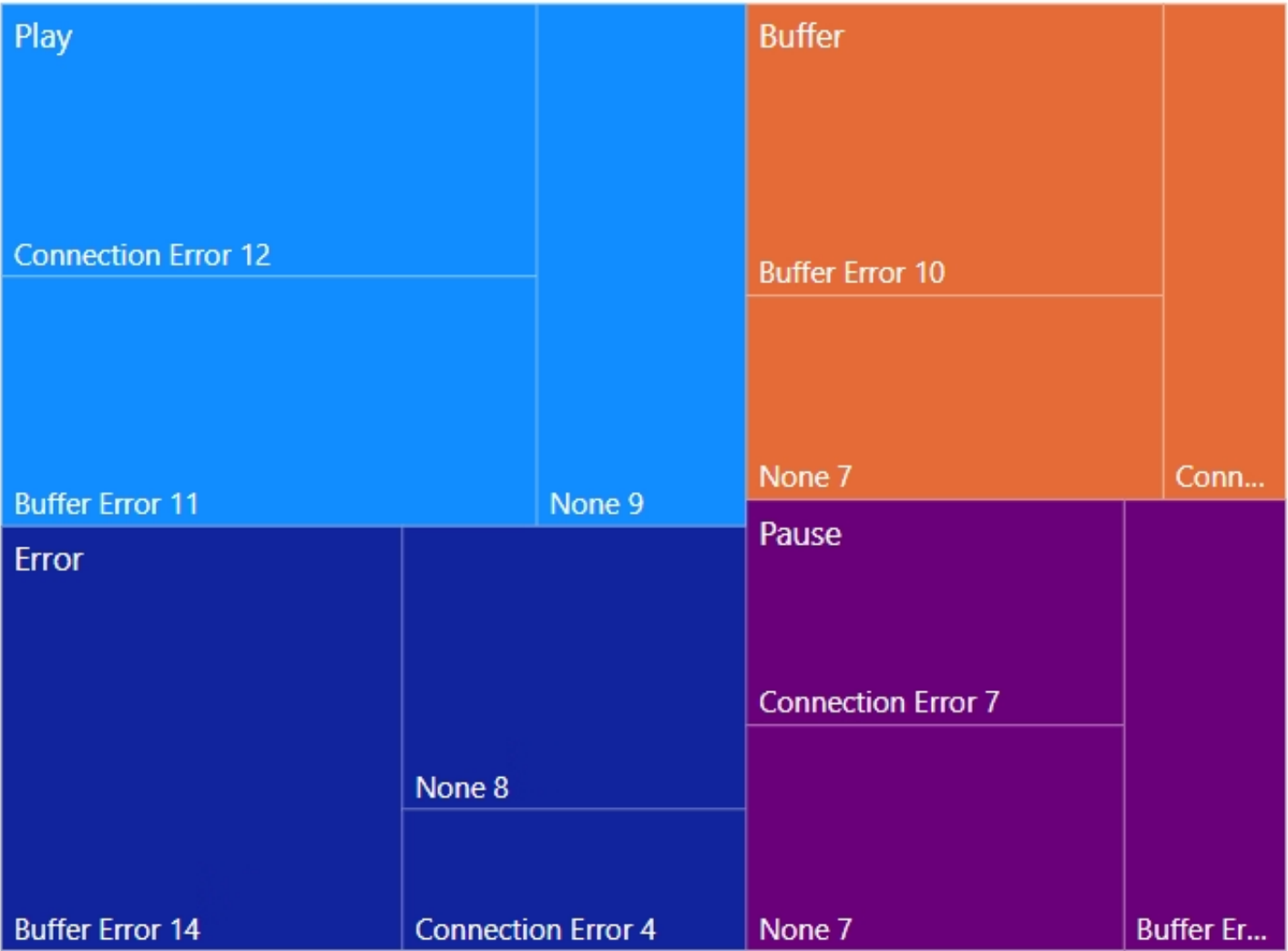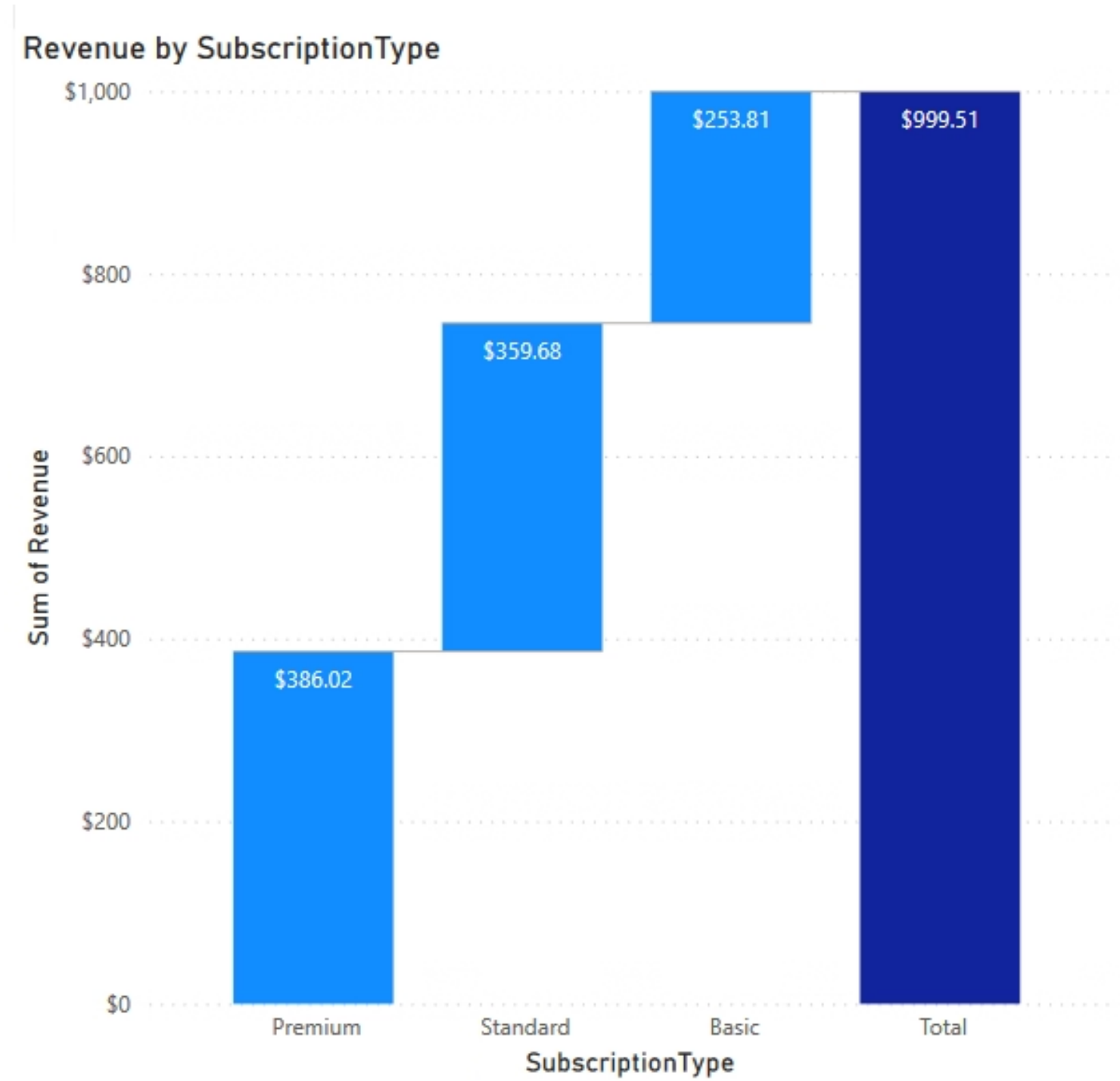
Most of the users have opted for either a Premium or a Standard membership

There is almost equl distribution of Device Type amongst users

Project Overview 01 Business Stakeholders 02 Data Source 03 Data Strategy & Stack 04 NoSQL Pipeline 05 Results/Dashboarding 06 <u>Conclusion & Steps 07</u>

Conclusion & Next Steps

Takeaway 1

- Implemented a Database Strategy and worked on full data pipeline from designing Architecture till the final dashboard

Takeaway 2

- Business needs define what Data to process
- Data then, gives business insights or identifies problems
- This virtuous cycle is at the heart of the BigData revolution

Next Steps

- With LLMs generating complex data such as natural language interactions, our flexible service infrastructure is equipped to adjust and interpret these advanced data streams.

NETFLIX VENTURES

Thank you!