PROJECT PROPOSAL

CLOUD ARCHITECTURE & DATA PIPELINE DESIGNS

HELLEN GAITAN
DATA CONSULTANT

Contents

Introduction....2
Objectives....3
Business Requirements....4
Overview of Airbnb's Business Model and Data Needs....5
Data Sources....6
Data Cloud Architecture....8
Components of the architecture....8
Ingestion layer....8
Transformation layer....9
Server layer....9
Consume Layer....10
Streaming Data Layer....10
How It Helps with Airbnb's New Features:....11
Data Pipeline....11
Design and implementation of the data pipeline: Parent-Child Pipeline....11
Orchestration and Scheduling of Data Pipeline Tasks:....12
Failures Responses:....13
Conclusion....14
References....15

Introduction

This project aims to transform Airbnb's hospitality operations by leveraging data analytics and cloud technologies to optimize resource allocation, drive increased profitability, and promote sustainable growth. By analyzing large datasets in a productive way, Airbnb will be able to implement data-driven strategies, to enhance its operational efficiency and customer experience.

This document is dedicated to outlining the most effective strategies and considerations for data processing, with a focus on efficiency and fostering a data-driven culture within the organization. The goal is to establish practices that ensure the ongoing maintenance and optimization of the data architecture and pipeline.

It is important to note that this project is centered around cloud architecture and the design of a data pipeline, drawing from the insights of data engineers Shara Pineda, Lovedeep Mehta, Ricardo Schmid, and Hellen Gaitan. However, the development of this document and the recommendations contained within are primarily driven by Hellen Gaitan.

Objectives

1. Utilize data analytics to identify patterns and trends in resource utilization, allowing Airbnb to allocate resources more effectively and reduce waste.

2. By optimizing resource allocation and streamlining operations, the project aims to increase Airbnb's profitability and financial performance.

3. Implement strategies that support long-term growth and sustainability, ensuring that Airbnb can continue to thrive in a competitive market.

4. Redesign existing processes to be more efficient and effective, reducing the time and effort required to complete tasks.

5. Create a user-friendly platform for host partners that simplifies the hosting experience and reduces friction in interactions with Airbnb.

6. Improve the speed and efficiency of interactions between host partners and Airbnb, leading to a more seamless experience for both parties.

7. Use data-driven insights to personalize customer interactions and improve engagement, leading to higher customer satisfaction and loyalty.

8. Enable Airbnb to make informed decisions based on data analysis, leading to better outcomes and improved performance across all aspects of the business.

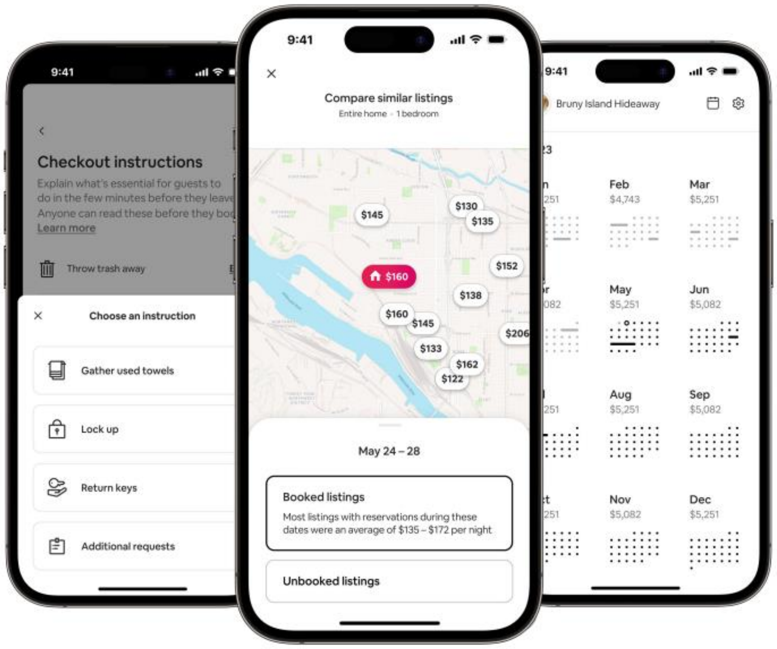
Business Requirements

Current Situation

Airbnb's current business model relies on providing a platform for individuals to rent out their properties to travelers. This model has been successful in creating a vast network of hosts and guests, leading to a significant amount of data being generated daily. For this reason, it's essential to respond to the need of enhance an automated system to make the flow simple, easy to update scalability and keep the real time information on place. This project is created to enforce Airbnb cloud architecture and Data pipeline to accomplish the requirements they need to perform these updates.

As part of the 50+ new features and upgrades, they're launching 25 improvements for Hosts, including new pricing tools to help Hosts set competitive prices, easily add weekly and monthly discounts, and compare their listing to similar ones in their area. They're also including a yearly view in calendar, the ability to easily enter checkout instructions, read receipts and new quick replies in messaging, and more.

airbnb 2023 Summer Release



Overview of Airbnb's Business Model and Data Needs

Airbnb's business model revolves around providing a seamless and enjoyable experience for both hosts and guests. This includes features such as easy booking, secure payments, and reliable customer support. To support these features, Airbnb requires robust data analytics capabilities to understand user behavior, optimize resource allocation, and enhance customer satisfaction.

Specific Requirements for the Data Cloud Architecture and Data Pipeline

1. The data cloud architecture should be scalable and flexible to accommodate Airbnb's growing data needs.

2. The data pipeline should be able to ingest, process, and analyze large volumes of data in real-time to provide timely insights. Specifically, this should be a must for bookings, reservations fee and cancellations, to have the calendar up to date.

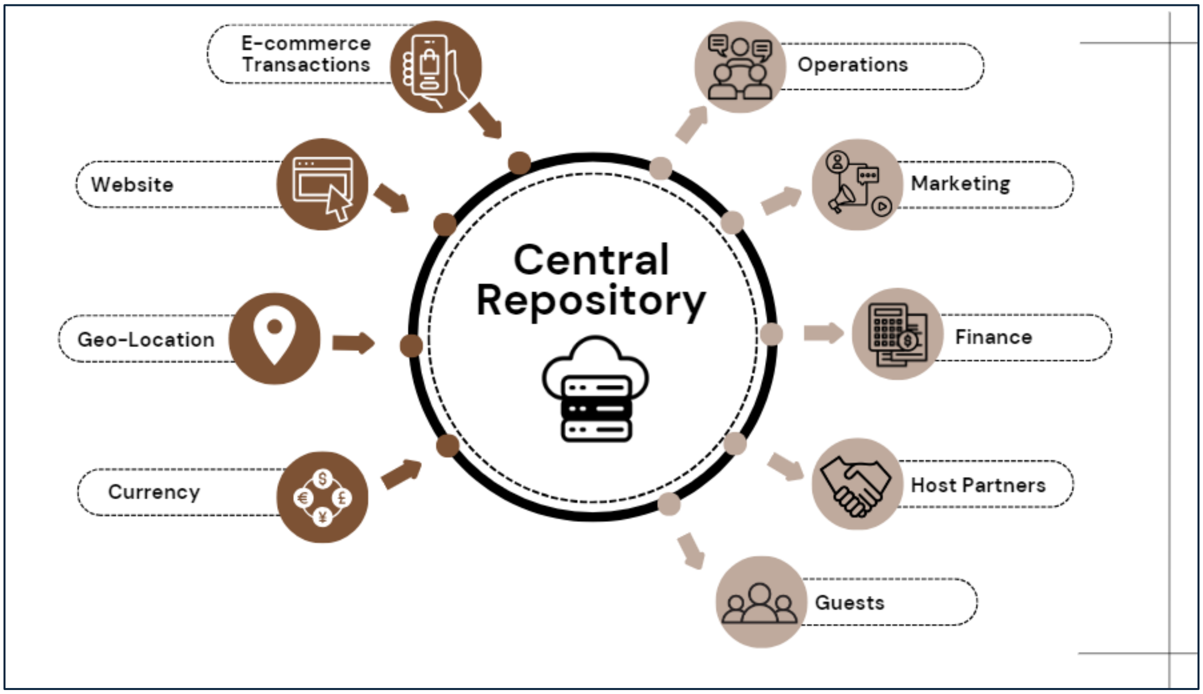
3. Data security and privacy should be prioritized to protect sensitive information.

4. The architecture should support seamless integration with existing systems and tools used by Airbnb. Azure platform is a great alternative to sustain this enhancement.

5. The data pipeline should be reliable and fault-tolerant to ensure uninterrupted operation.

6. This project provides challenges and possible responses the data pipeline and architecture could have on the performance of the platforms used in this project.

Data Sources



Sources overview:

- Ecommerce Transaction:

Includes all data related to bookings, reservations fee, transactions, cancellations, earnings, number of purchases, calendar up to date, and payments in overall.

- Website:

Encompasses all activities and interactions on the website excluding ecommerce transactions. This source includes information about click to contacts, form submissions, site sessions, traffic source, user interactions, behavior overview and customer patterns. We are going to use log files creating an automated process, extracting the relevant information and store it the data lake.

- Currency (API): Depend on website you are using (ex. .ca, .br)

For currency conversion, we are going to use a REST API like OPEN EXHANGE RATES. We are going to configure an HTTP call activity in Azure Data Factory to fetch the exchange rates and store it in the data lake.

- **Geolocation (API):** Website tracking, Filling out of details, Search for location/properties,

For geo-location, we are going to set Azure data factory linked with HTTP as the service type with the URL of Google Maps Geocoding API.

Consumers:

- **Operations:** Utilizes data for identifying trends and making strategic decisions based on the current situation.

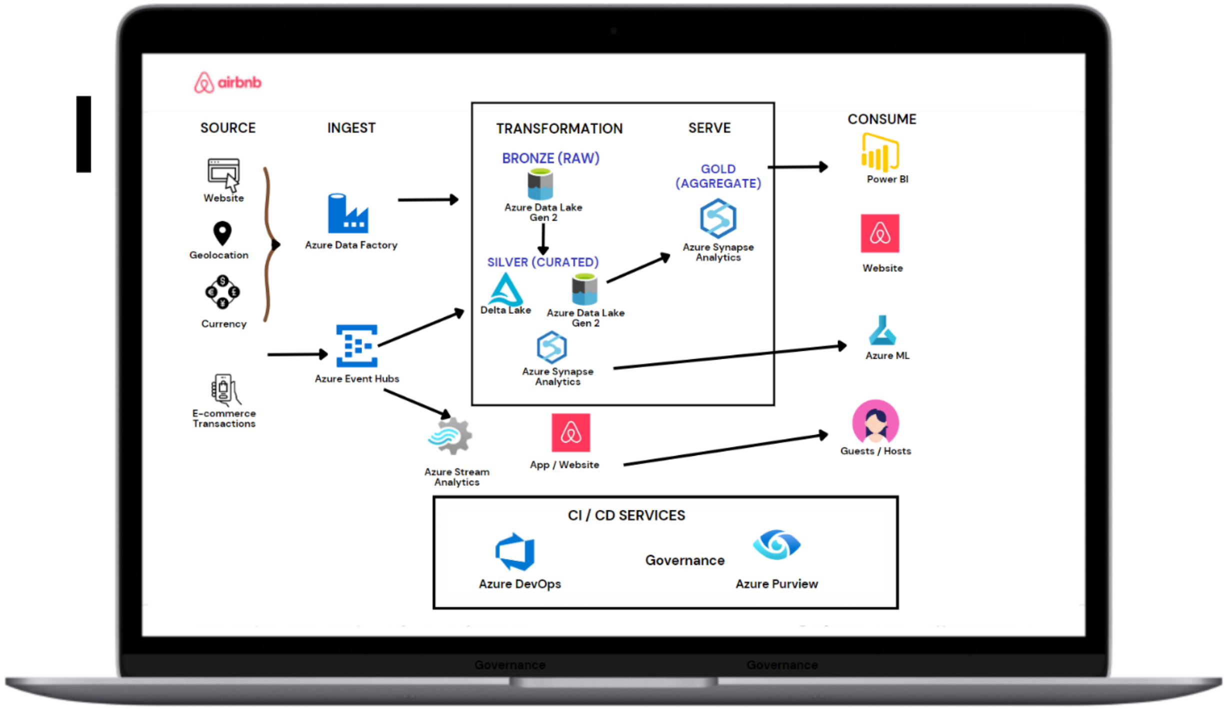
- **Marketing:** Leverages data for targeted marketing campaigns, adjusting prices, and identifying market trends.

- **Finance:** Uses data for strategic decision-making based on the current financial situation.

- **Host Partners:** Analyzes data for understanding trends, patterns, and managing bookings effectively. They are going to receive help Hosts to set competitive prices, easily add weekly and monthly discounts, and compare their listing to similar ones in their area. They're also including a yearly view in calendar, the ability to easily enter checkout instructions, read receipts and new quick replies in messaging, and more.

- **Guests:** Represents the end consumers who make bookings on the platform.

Data Cloud Architecture



Components of the architecture

Ingestion layer

1. **Azure Data factory:** ADF would serve as the primary service for batch ingestion, responsible for collecting and processing large volumes of data from various sources. For Airbnb, this would include data related to website tracking (user interactions, page views, etc.), geolocation (location data of properties, user locations, etc.), and currency (exchange rates, currency conversions, etc.).

   ADF would utilize connectors to extract data from different sources, such as web logs for website tracking, location services for geolocation data, and external APIs for currency data. The extracted data would then be transformed and loaded into a data lake or data warehouse for further processing and analysis.

2. **Azure Event Hubs**: plays a crucial role in this data architecture by acting as a highly scalable and real-time data streaming platform. In the ingestion layer, it's necessary to transmit the E-commerce Transactions, since this holds information needed in real-time, to keep updated the calendar. Event

Hubs can ingest and collect large volumes of event data from various sources, such as applications, devices, sensors, and logs. This data is collected in real-time, ensuring that no events are missed.

Transformation layer

In the data architecture described, there are three main layers: bronze, silver, and gold. Each layer serves a specific purpose in the data processing, with different technologies used to store and manage the data at each stage.

Bronze Layer:

Azure Data Lake: Azure Data Lake is used as the storage service for the bronze layer. It is a scalable and secure data lake storage solution that can handle large volumes of raw data. In this layer, raw data from various sources, such as website tracking, geolocation, and currency, is stored in its original format without any transformation.

Silver Layer:

Delta Lake: Delta Lake is used in the silver layer to leverage different versions of the data. Delta Lake provides ACID transactions, scalable metadata handling, and unified streaming and batch data processing. It allows for efficient data versioning and management, ensuring data integrity and consistency.

Azure Data Lake (Curated Data): Azure Data Lake is also used in the silver layer to store curated data. This curated data has undergone some level of transformation and cleansing to make it suitable for analysis. This layer acts as a staging area for data before it is further processed in the gold layer.

Azure Synapse Analytics: Azure Synapse Analytics is used for big data analysis in the silver layer. It provides a unified analytics service that enables the integration of big data and data warehousing. Synapse Analytics can handle large volumes of data and perform complex analytics, making it suitable for processing data in the silver layer.

Server layer

Gold Layer:

In the gold layer of the Airbnb architecture, Azure Synapse Analytics plays a crucial role in performing advanced analytics and deriving insights from the curated and processed data. Azure Synapse Analytics provides powerful analytics capabilities, including machine learning, data warehousing, and big data processing. In the gold layer, Synapse Analytics can be used to perform complex analytics tasks, such as predictive modeling, trend analysis, and anomaly detection, to derive valuable insights from the data.

Synapse Analytics integrates seamlessly with Azure Data Lake and Delta Lake, allowing for easy access to curated and processed data. This integration enables data scientists and analysts to work with the data in a familiar environment, using tools like Apache Spark and SQL Server to query and analyze the data.

Synapse Analytics integrates with other Azure services, such as Power BI for data visualization and Azure Machine Learning for building and deploying machine learning models. This integration enables a seamless end-to-end data analytics workflow within the Airbnb architecture.

Consume Layer

In the consumption layer of the Airbnb architecture, the focus is on enabling various consumers, including marketing, finance, operations, and end-users, to access and derive insights from the processed data.

1. **Marketing, Finance, and Operations (Power BI):** These teams use Power BI as a powerful tool for visualizing and analyzing data. Power BI connects to Azure Synapse Analytics to access the curated and processed data. Marketing teams can analyze trends, customer behavior, and campaign performance. Finance teams can track financial metrics and analyze revenue streams. Operations teams can monitor operational performance, customer satisfaction, and efficiency metrics.

2. **End-Users (Airbnb Website):** End-users access the data through the Airbnb website, where they can view listings, make bookings, and interact with the platform. The website uses data from the consumption layer to provide personalized recommendations, pricing information, and availability status to users.

3. **Data Engineers (Azure ML):** Data engineers use Azure Machine Learning (Azure ML) to develop and deploy machine learning models for various purposes, such as pricing optimization, fraud detection, and demand forecasting. Azure ML integrates with Azure Synapse Analytics and Azure Data Lake to access the required data for model training and inference.

Streaming Data Layer

Azure Stream Analytics: Streaming data from the website is processed in real-time using Azure Stream Analytics after ingestion process from Event hubs. Azure Stream Analytics will perform the streaming data before transferring the data to website to be seen by the guests and hosts in real-time.

A user-friendly interface for hosts and guests to interact with the platform. This includes features such as listing management, pricing tools, calendar views, messaging, and more.

How It Helps with Airbnb's New Features:

The enhanced Airbnb cloud architecture and data pipeline described would greatly benefit Airbnb's new features and upgrades in several ways:

- The new pricing tools for hosts can leverage the data pipeline to analyze market trends, competitor pricing, and demand patterns in real-time. This information can help hosts set competitive prices and easily add discounts, enhancing their listing's visibility and bookings.
- The data pipeline can provide historical booking data to enable a yearly view in the calendar. Hosts can easily visualize their availability and plan, improving their booking management efficiency.
- The data pipeline can facilitate the easy entry of checkout instructions and provide read receipts and quick replies in messaging. This improves communication between hosts and guests, leading to a better overall experience.
- The data pipeline can compare a host's listing to similar ones in their area, providing valuable insights for pricing and marketing strategies. Hosts can make data-driven decisions to optimize their listings and maximize bookings.
- The data pipeline ensures that all information, including pricing, availability, and messaging, is kept up-to-date in real-time. This enhances the user experience by providing accurate and timely information to hosts and guests.

Data Pipeline

Design and implementation of the data pipeline: Parent-Child Pipeline

The Parent-Child Pipeline design for the Airbnb data pipeline was likely chosen for its ability to handle complex data processing tasks in a scalable and efficient manner. This will enable to answer business questions along the process.

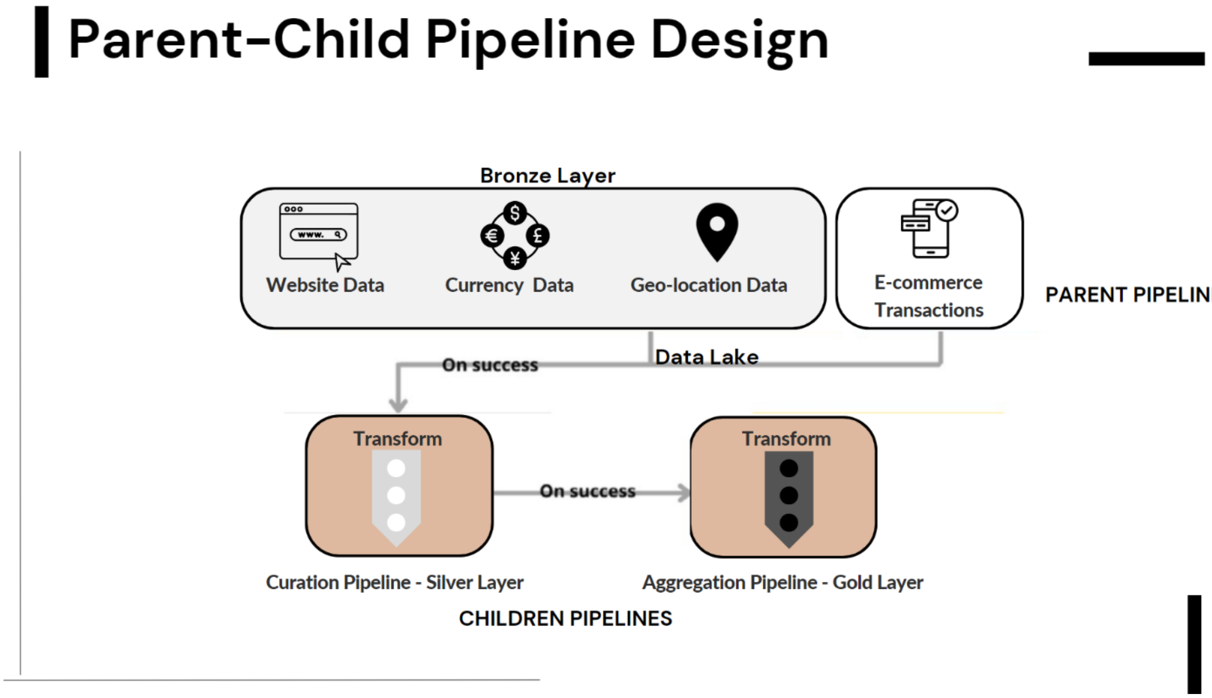
The Parent-Child Pipeline design allows for the creation of modular and reusable components. Each "parent" pipeline can serve as a template for different data processing tasks, with "child" pipelines being created as instances of these templates for specific use cases. This modularity and reusability make it easier to maintain and scale the data pipeline as Airbnb's data processing needs grow.

The design allows for parallel processing of data by breaking down the processing tasks into smaller, independent units (the "child" pipelines). This parallel processing capability can significantly improve the overall performance and efficiency of the data pipeline, especially when processing large volumes of data.

The Parent-Child Pipeline design facilitates dependency management between different stages of the data processing workflow. "Child" pipelines can be configured to depend on the output of their parent pipelines, ensuring that data is processed in the correct order and that downstream tasks are not started until the necessary data is available.

The design is inherently scalable, allowing Airbnb to easily scale up or down the number of "child" pipelines based on demand. This scalability is essential for handling fluctuations in data volume and processing requirements, ensuring that the data pipeline can meet Airbnb's needs as its business grows.

The modular nature of the design makes it easier to maintain and update the data pipeline over time. Changes to the processing logic or the addition of new features can be implemented by modifying the parent pipeline template or creating new child pipelines, without affecting the overall structure of the pipeline.



Parent-Child Pipeline Design

Orchestration and Scheduling of Data Pipeline Tasks:

For orchestration and scheduling of data pipeline tasks in Airbnb's architecture, Azure Data Factory will be used. This tool provides a graphical interface for designing workflows and scheduling tasks, making it easy to orchestrate complex data processing pipelines.

Azure Data Factory provides capabilities for orchestrating and scheduling data pipeline tasks in a cloud environment. It allows to create pipelines that can ingest, transform, and load data, with built-in support for scheduling and monitoring tasks.

Failures Responses:

Timeout:

A timeout limit of 24 hours is set for each data processing task. If a task exceeds this limit, it is considered failed, and an appropriate response is triggered.

Retry Interval:

Tasks that fail due to transient issues are retried up to 3 times, with a retry interval between attempts. The interval is designed to allow for temporary issues to resolve before retrying.

Alerts for Failures:

Alerts are configured to notify relevant stakeholders of failures in the data pipeline. Pop-up notifications are displayed on monitoring dashboards to alert operators and developers of ongoing issues. Additionally, email notifications are sent to designated recipients to ensure timely awareness and response to failures.

Conclusion

The Airbnb architecture and pipeline project aimed to enhance the data processing capabilities of the platform to support its successful business model of connecting hosts and guests. The project focused on implementing a scalable, efficient, and real-time data pipeline using Azure services. This new architecture enables the ingestion, processing, and analysis of large volumes of data generated daily by the platform. By optimizing resource allocation and streamlining operations, the project aims to increase Airbnb's profitability and financial performance. Redesigning existing processes to be more efficient and effective reduces the time and effort required to complete tasks. Utilizing data-driven insights to personalize customer interactions and improve engagement results in higher customer satisfaction and loyalty by hosts and guests as well. Basically, with the implementation of this data architecture and pipeline design. Airbnb is going to make informed decisions based on data analysis leads to better outcomes and improved performance across all aspects of the business.

References

1. Cloud architecture project – Group 4
2. Airbnb news - Sharing more about technology

https://news.airbnb.com/sharing-more-about-the-technology-that-powers-airbnb/

3. Medium.com - Upgrading Data Warehouse Infrastructure at Airbnb

https://medium.com/airbnb-engineering/upgrading-data-warehouse-infrastructure-at-airbnb-a4e18f09b6d5