Cloud Architecture Project

Shara Vanessa Pineda

Southern Alberta Institute of Technology

Data Science - DATA-040-003

Table of Contents

1. The Company

1.1. Overview

The company chosen for this project is Airbnb and was based on several factors. Firstly, Airbnb being a big, international company makes it a household name, familiar to millions of travelers worldwide. It has a global reach in the hospitality industry and its widespread recognition makes it a fitting candidate for the project. Secondly, Airbnb operates entirely online, making it a good example for studying cloud architecture's role in facilitating digital transactions. Lastly, Airbnb always had an innovative approach to redefining travel and hospitality through technology. By trying to understand and create a cloud architecture for the company, the project aims to collect insights that can showcase understanding of different data concepts.

1.2. Airbnb

Founded in 2007, Airbnb is an online marketplace whose goal is straightforward yet impactful: to provide travelers with one-of-a-kind accommodations and experiences offered by locals. At its core, Airbnb serves as a platform where travelers can explore a diverse collection of lodging options, from cozy apartments to lavish villas, tailored to suit various tastes and preferences.

Fun fact about the company, Airbnb was originally named "Airbed and breakfast" when the founders rented out an air mattress in their living room to help cover rent expenses. This ingenious idea has since grown into a global phenomenon, with Airbnb now boasting over 7.7 million listings worldwide (as of December 2023). In essence, Airbnb defies the norms of traditional hospitality, offering travelers a gateway to immersive experiences and authentic connections wherever they go.

1.3. Objectives

The objectives identified in this article are divided into three segments, each representing a key stakeholder within Airbnb.

a. **For Airbnb: Optimize resource allocation driving increased profitability and sustainable growth**. Given Airbnb's status as an online marketplace and its digital-native nature, anything being utilized in the cloud resources is crucial. The goal is to

utilize essential resources effectively while also implementing measures to save on non-essential expenditures.

b. **For Host Partners: Design efficient processes and user-friendly platform that facilitates a faster and more streamlined experience for host partners.** Host partners play a crucial role in the company as they provide the "product" that the company sells. Because these partners are typically consisting of non-technical individuals who list their properties on the platform, one of the goals is making the platform's design and data presentation intuitive and accessible to individuals without technical expertise.

c. **For Guests: Enhance customer engagement and satisfaction by allowing data-driven decisions.** Customers should be presented with data that is pertinent and useful to them, such as options within their price range or properties that include their preferred amenities. From this data, customers can make well-informed decisions that meet their needs and preferences.

2. Vision

2.1. Overview

In the context of this article, the Vision refers to a conceptual framework that serves as a visual representation of the company's vision, illustrating the flow of data from its sources to the central repository and ultimately to the users as illustrated in Figure 1.

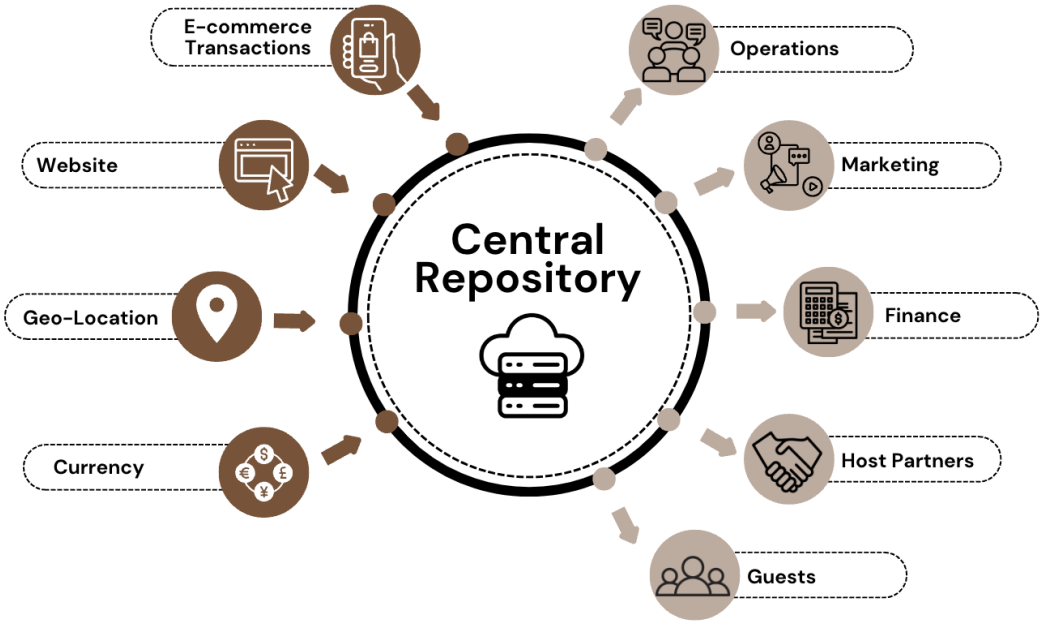

Figure 1: Airbnb Vision

The sources shown on the vision board represent the locations where data essential for analytics is stored. The data obtained from these sources are then stored within the central repository.

The central repository serves as a hub where data is stored and managed. This centralized approach ensures that decision-makers have access to timely and accurate information to inform their decision-making processes.

Lastly, the users shown on the vision board represent the decision-makers who utilize the processed data to make informed decisions.

2.2. Data Source

There are four sources identified for this project, namely:

a. **E-commerce Transactions**: This includes all aspects related to bookings including reservations, reservation fees, cancellations, and maintaining an up-to-date calendar.
b. **Website**: This includes data from both the website and the mobile application. It incorporates Website Tracking, which provides insights into click-to-contact actions, form submissions, site sessions, traffic sources, and user behavior. Additionally, it covers Host Partners and Users Information, offering details about users.
c. **Geolocation:** This includes data on the geographical locations from which host partners, guests, and other visitors access the website. It also aids in providing information about the proximity of properties to specific sites including user's current location. This data source will be collected by downloading from an HTTP source.
d. **Currency:** This enables the reflection of the correct currency used by customers, considering Airbnb's international presence and diverse user base. This data source will be collected by using a REST API.

2.3. Data Users

There are five users identified for this project, each representing the three main stakeholder groups previously mentioned namely Airbnb, Host Partners, and Guests.

a. **Operations:** This user group utilizes data to identify trends and patterns, guiding strategic decision-making and setting directions based on situational analysis.
b. **Marketing:** This user group uses data to analyze trends, execute targeted marketing campaigns, and make price adjustments as necessary to optimize performance.
c. **Finance:** This user group relies on data to analyze trends, generate forecasts, and make informed decisions regarding financial strategies and resource allocation.
d. **Host Partners:** This user group benefits from access to data that enables them to assess their sales performance, compare the performance of their properties over time, analyze their rates relative to competitors, and gather insights from guest feedback to enhance their offerings.
e. **Guests:** This user group utilizes all available data to make informed booking decisions that align with their needs and preferences, ensuring a personalized and satisfactory experience.

3. Cloud Architecture

3.1. Overview

Cloud architecture is the backbone of modern data-driven enterprises, providing a framework for managing and processing massive amounts of data with efficiency and scalability. It's like the blueprint of a building, ensuring every component works seamlessly together. Whether it's sourcing data from different repositories, processing it, or analyzing it, cloud architecture ensures that everything works flawlessly, empowering organizations to thrive in today's fast-paced digital landscape. Cloud architecture is comprised of a series of interconnected stages, starting from data source and ultimately to consumption as illustrated in Figure 2.

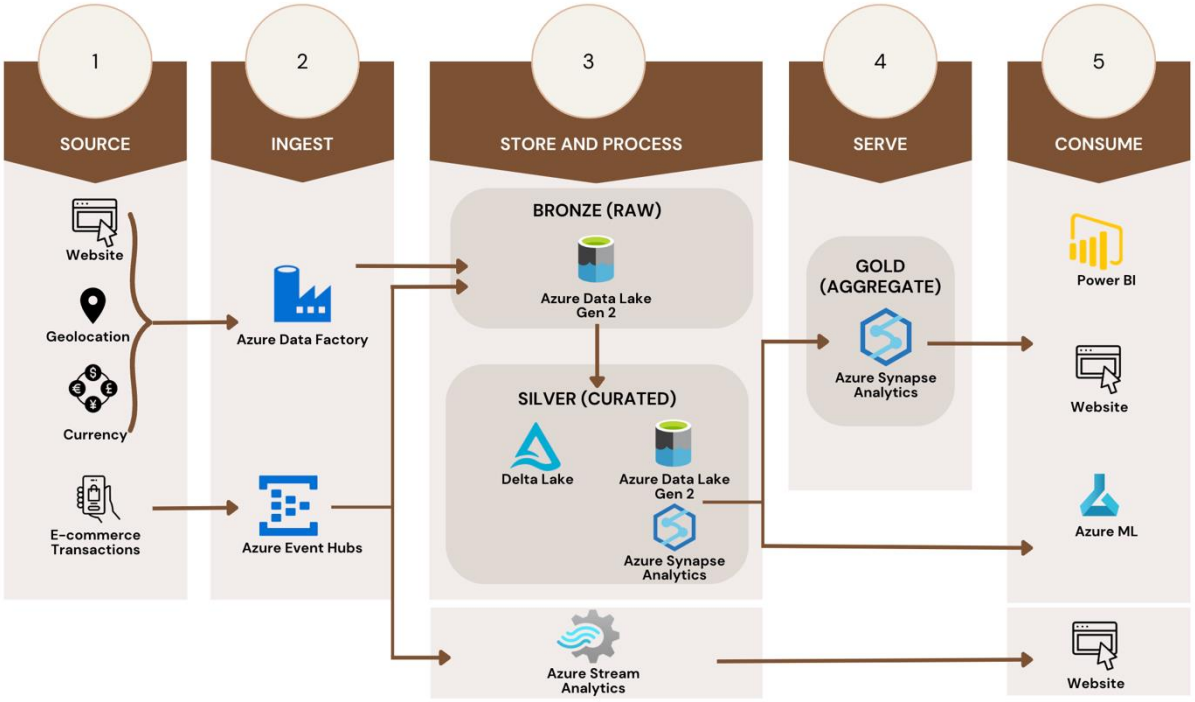

Figure 2: Airbnb Cloud Architecture

3.2. Data Source

The Sources, as previously mentioned, are the locations where data essential for analytics is stored. For consistency, we'll rely on the same sources throughout this project. For more detailed information about these sources, please refer to Article 2.2.

3.3. Ingest

In the Ingest phase, data is brought from the identified sources through two distinct methods: Batch Ingestion and Streaming Ingestion. By incorporating these two approaches, diverse range of data sources and processing requirements can be effectively handled, ensuring that the cloud architecture is prepared to deliver actionable insights in a timely manner while also optimizing resource utilization and minimizing waste when data updates are not needed.

a. **Batch Ingestion:** This will be used to ingest data from Website, Geolocation and Currency. Batch ingestion involves the process of collecting and loading data in predefined intervals or batches. This approach is ideal for scenarios where data updates do not require real-time processing. For this project, **Azure Data Factory** is utilized for batch ingestion. Each batch of data will be loaded every 8 hours, aligning

with the frequency of updates for currency and geolocation data, as well as non-real-time analysis for website-related data.

b. **Streaming Ingestion**: This will be used to ingest data from E-commerce Transactions. Streaming ingestion involves the continuous and real-time collection and processing of data as it becomes available. This approach enables near-instantaneous data updates and is suited for scenarios requiring real-time insights. For this project, **Azure Event Hubs** is utilized for streaming ingestion. Data will be loaded in real-time with a minimal lag of 2-3 seconds.

3.4. Streaming Data

In this cloud architecture, streaming data that doesn't require immediate processing or analysis follow a distinct pathway from other types of data. This real-time data flow is managed through a series of interconnected services, beginning with Azure Event Hub, then proceeding to **Azure Stream Analytics** before being reflected on the website.

3.5. Store and Process

The Store and Process phases, along with the Serve phase, collectively form the Lakehouse architecture. Within the Store and Process phases, data is organized into two layers: Bronze and Silver. By employing a structured approach through these layers, we can effectively manage and process data within the architecture, fostering data reliability, integrity, and usability throughout the data lifecycle.

a. **Bronze Layer (Raw Data):** The Bronze layer serves as the landing zone for raw, unprocessed data. Here, data exists in its most native form without any modifications or transformations. For this project, **Azure Data Lake Gen 2**, a storage solution offering vast storage capacity and compatibility with various data types, is utilized. Additionally, it implements an access control model, ensuring secure data management. Access to the Bronze layer is recommended to be restricted to seasoned data engineers, emphasizing the importance of data governance and integrity at this stage.

b. **Silver Layer (Curated):** The Silver layer represents the subsequent stage where data is cleaned, transformed, and curated to ensure usability and reliability. Data in this layer is typically structured, organized, and enriched with defined schemas and data types. Business rules are adhered to during data transformation to ensure compliance and alignment with organizational requirements. For the project, **Azure Data Lake Gen 2** stores data requiring minimal transformation, while **Azure Synapse**

Analytics facilitates data transformation and analysis, enabling valuable insights to be derived. Additionally, **Delta Lake** is leveraged for schema evolution, ensuring flexibility and scalability in managing evolving data requirements.

In the Lakehouse architecture, some data from the Silver layer can be sent directly to Azure ML (Machine Learning) for analysis. This streamlines the process, allowing organizations to derive insights more efficiently and make faster decisions based on machine learning models. Giving direct access to the Silver layer also enables users to have the option to analyze or aggregate data beyond the standard set of aggregations provided in the gold layer.

3.6. Serve

In the Serve Phase of this cloud architecture, data from the curated Silver layer undergoes further refinement and transformation in the Gold Layer.

c. **Gold Layer (Aggregate)**: The Gold layer is the last phase where data undergoes further curation, refinement, and transformation to meet the specific needs of users. Datasets and views are created to support reports and dashboards, providing valuable insights to aid in the decision-making process. This phase focuses on answering specific business questions tailored to the users' requirements. For this project, **Azure Synapse Analytics** is utilized for both further analysis and storage of structured data within the Gold layer. This powerful tool enables seamless integration of analytics and storage functionalities, allowing for efficient processing and management of data.

3.7. Consume

In the Consume Phase, users access curated and aggregated data from the Silver and Gold Layer for analysis and decision-making. Key tools in this phase include Power BI, the website interface, and Azure ML. **Power BI** enables users to create interactive reports and dashboards, providing quick insights. The **website interface** offers easy access to curated data and insights, serving as a central hub for users. **Azure ML (Machine Learning)** facilitates advanced analytics, empowering users to uncover patterns and make data-driven decisions confidently.

4. Data Pipeline

4.1. Overview

Data pipeline, in the simplest terms, is a set of instructions that make data flow from one location to another. It includes all the processes required to transform raw data into prepared data that is ready for consumption by users. Data pipelines also automate repetitive tasks, enhancing efficiency and consistency in data processing.

4.2. Data Pipeline Design

For this project, the pipelines were designed using a **Parent-Child Approach** as illustrated in Figure 3. In the parent-child approach, a master pipeline, known as the parent pipeline, orchestrates the execution of one or more subordinate pipelines, known as child pipelines. Each child pipeline performs specific tasks or processes, while the parent pipeline manages the overall workflow and ensures that dependencies between the child pipelines are met.

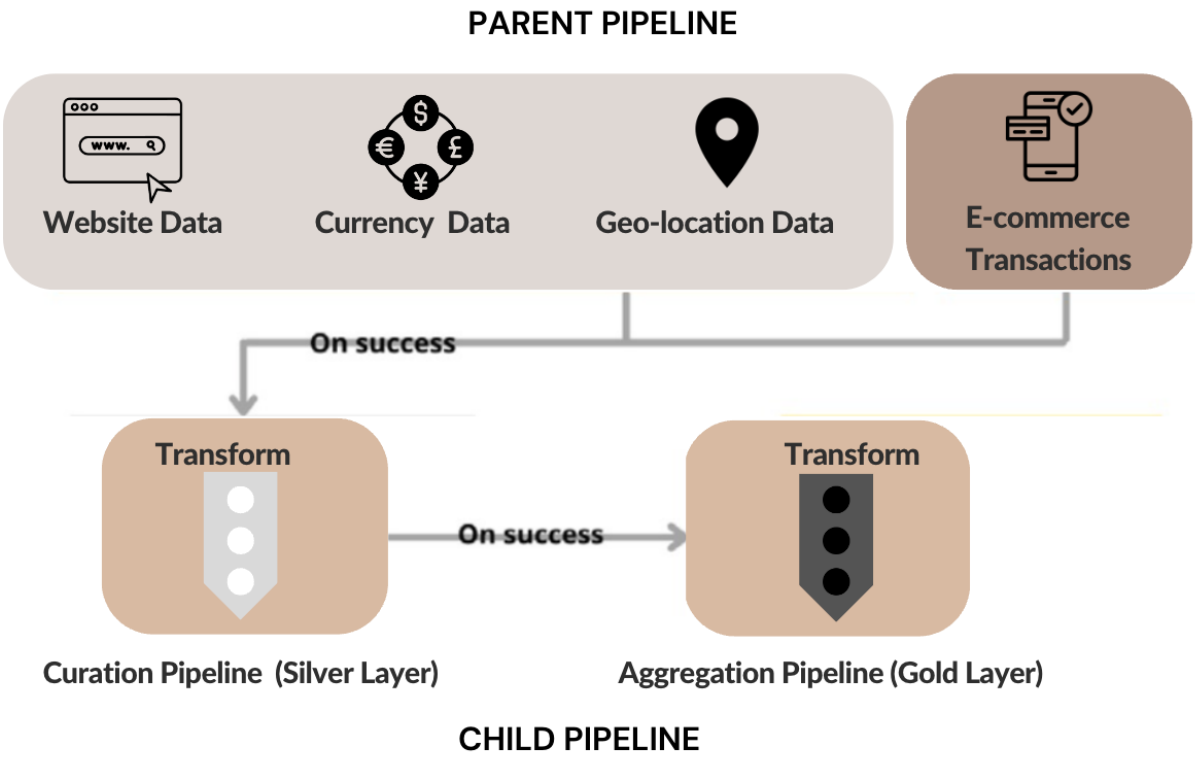

Figure 3: Parent-Child Pipeline

One of the key benefits of the parent-child approach is that each dataset is ingested only once, and subsequent child pipelines can reuse it. This reduces redundancy and saves on storage and compute costs by eliminating the need to duplicate data for each pipeline. Additionally, the parent-child approach provides better execution control over child pipelines by enforcing dependencies with the parent. Child pipelines can be triggered only after the parent pipeline has successfully completed its tasks, ensuring that data is processed in the correct sequence and minimizing errors in the data processing workflow.

4.3. Data Pipeline Deployment

In data pipeline deployment, there are typically three methods: Manual, Time-based (Scheduled), and Event-based. For this project, two methods will be utilized: Time-based and Event-based.

a. **Time-based (Scheduled):** In the Time-based approach, the data pipeline is triggered by a scheduler and configured to run at specified intervals. For this project, the Time-based method can be employed for financial systems, including daily runs for host partner payouts and monthly runs for internal accounting procedures.
b. **Event-based:** The Event-based approach initiates pipeline execution based on specific dependencies or conditions, such as the success or failure of previous events or the availability of new data. In this project, majority of pipelines utilized the Event-based method. This ensures that pipeline execution is contingent on the completion of preceding tasks, such as updating currency based on fetched exchange rates and integrating new host partner information following a successful registration process.

4.4. Monitoring Data Pipelines

For the data pipeline deployment, it is essential to anticipate and prepare for potential failures that may occur during operation. Monitoring all activities and implementing strategies to address these failures is crucial for ensuring the reliability and effectiveness of the pipelines. To increase durability, the following measures are suggested to be incorporated into the pipelines:

a. **Timeout:** Timeout defines the maximum duration that a specific activity should run before it is terminated, regardless of its status (successful or failed). This setting helps prevent activities from running indefinitely and consuming excessive

resources. For this project, a timeout of 24 hours is configured, after which the activity is automatically terminated.

b. **Retry Intervals:** The retry mechanism introduces a delay between each retry attempt, allowing for a more gradual and controlled approach to handling failures. Retry intervals can be configured at both the pipeline and activity levels, specifying the duration between retry attempts and the maximum number of retry attempts. For this project, a configuration of 3 retry attempts with a 5-minute interval for batch processing and a 10-second interval for streaming is implemented.
c. **Alerts:** Alerts are essential for notifying users of any failures that occur within the pipelines. This ensures that users are promptly informed of any critical activities that have failed and need attention. In this project, alerts will be provided through pop-up notifications and emails, enabling users to stay informed and take appropriate actions in response to failures.

5. Summary

The company chosen is Airbnb as its status as an online marketplace makes it an ideal candidate for examining the role of cloud architecture in facilitating digital transactions. Utilizing the Azure ecosystem, the designed cloud architecture seamlessly integrates various applications to ensure the smooth flow of data from its source to consumption. The implementation of a parent-child pipeline design also enhances operational efficiency, with a master pipeline orchestrating the execution of subordinate pipelines. With this cloud architecture and the strategies designed for the deployment of the pipelines, the project aims to achieve specific objectives tailored to Airbnb as a company, its host partners, and its guest customers. These objectives encompass optimizing resource allocation, enhancing user experiences, and enabling data-driven decision-making processes. Throughout the project, careful consideration has been given to anticipate and address potential challenges and failures, with strategies devised to mitigate risks and ensure operational reliability.

In conclusion, this project represents a comprehensive exploration of cloud architecture's transformative potential within online marketplaces like Airbnb. By embracing innovative technologies and strategic deployment strategies, the project seeks to exceed the objectives set forth.