ST 590-651 Project 2 Part 1: The Build-up of Hertz Big Data Platform

Group G: Raiden Han, Jingjing Li, Nataliya Peshekhhodko

June 27, 2022

1 Introduction

The Hertz Corporation, one of the largest worldwide vehicle rental companies sited in Estero, Florida, USA, operates in 160 countries all over the world. Traditionally, Hertz relies on customer satisfaction surveys to dictate its marketing strategy toward customers, which is time-consuming, inaccurate, site-limited, and inefficient. Understanding customers' needs, strengthening its relationship with customers, and predicting the markets call for new technology to facilitate the business development of Hertz.

Big data refers to data sets that are too large or complex to be dealt with by traditional data-processing application software. It can be described with five "V"s feature: Volume, Variety, Velocity, Veracity, Value which correspond to the quantity of generated and stored data, type and nature of the data, data generation speed, truthfulness or reliability of the data and the worth in information that can be achieved by the processing and analysis of large datasets respectively. Companies frequently adopt big data to improve operational efficiency, increase customer service quality, predict personalized markets, and so on, resulting in revenue and profits.

In this report, a big data platform recently built for Hertz, which aims to extract value and insights from a large amount of data and answer business questions, will be introduced. The following includes data transformation, big data pipeline, data input services, monitoring services, and orchestration tools.

2 Data Transformation

In order to treat big data, data transformation is usually the first step. Data transformation is converting data from one form or structure into another that can be used for data warehousing, integration, and computation. Data transformation can help businesses resolve compatibility issues and improve data consistency to meet the destination system's requirements. Many companies have enormous systems to produce data, while some company services are legacy which only generate batched data. Unlike many businesses, Hertz also consumes data from external resources like manufacturers. Considering the problem of decentralized data storage with many legacy systems which can not be directly applied to quick searches, analysis, computation, etc., Hertz recently undertook a massive digital transformation to evolve its technology landscape. The center of all of it is the data that needs to be centralized, scaled quickly, analyzed, stored with less cost, and secured. The goals Hertz was trying to achieve are

- providing a better view of customers (retail and corporate);
- improving customer journey from car reservation and all the way through;
- building a global solution that allows better asset management;
- reducing time to market;
- allowing businesses to have insights about new opportunities from data analysis.

3 AWS Pipeline

The platform for big data management of Hertz company is built on Amazon Web Services, Inc. (AWS), a subsidiary of Amazon that provides on-demand cloud computing platforms and Application Programming Interfaces (APIs) to customers. The schematic diagram of the AWS diagram is shown in Figure 1.

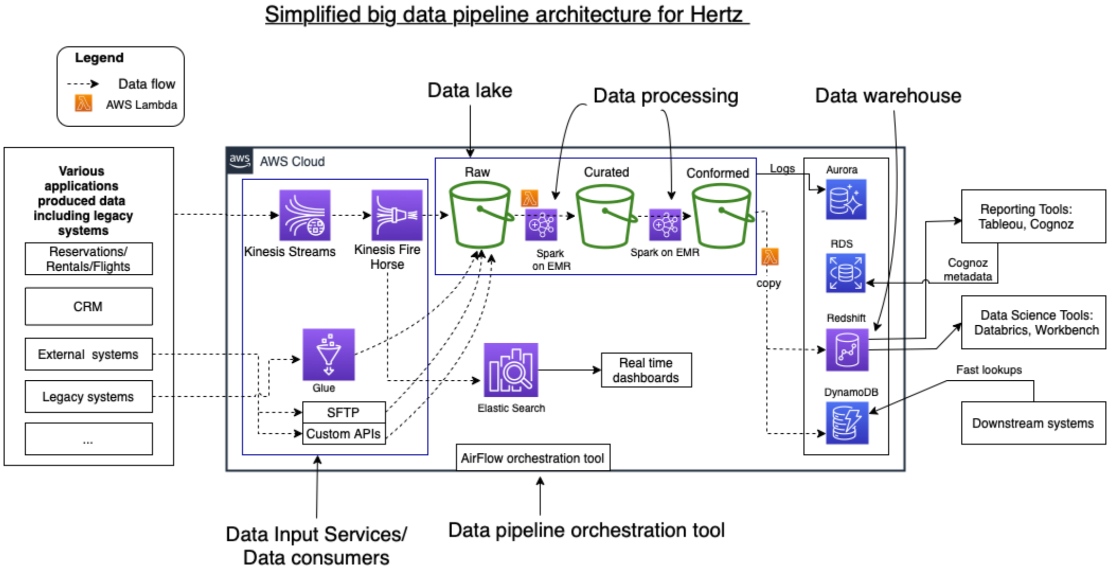

Figure 1: Hertz Big Data Pipeline

Data from many systems are entered into cloud solutions via several data input services/data consumers such as Kinesis services, Glue, Custom APIs, and Custom SFTPs (Secure File Transfer Protocol from external systems), which depend on the source of the data. AWS Kinesis Stream, which sends data to AWS Kinesis FireHorse, is used for the streamed data. Batched data from legacy systems, which are not able to stream data, is consumed by AWS Glue.

Data consumers write data to Data Lake, a set of Amazon's Simple Storage Service (S3) buckets with three consecutive levels: raw, curated, and conformed. The raw level denotes untouched data that data consumers directly write. The curated level contains processed data, and the processing is done by Spark on EMR. Spark jobs are triggered by computing serverless AWS lambda. Spark on EMR is a data processing tool with auto-scaling. Conformed level accommodates data processed by another set of jobs for Spark on EMR. Besides the S3 approach, data consumers also deliver data to Elastic Search, which is used for data discoveries such as more insights.

Once data arrives at the conformed level, it is copied by AWS lambda and conveyed to Data warehouse Redshift with Spectrum and DynamoDB. AuroraDB is used for different types of log stores, which are used for debugging purposes. Amazon Redshift with Spectrum is used as a data warehouse and source for upstream systems, which are reporting and data science tools. High-performance DynamoDB is used for quick look-ups for the processes when a decision is time-sensitive. Data from the warehouse and DynamoDB is ready to be consumed by downstream systems, reporting tools, and data science tools.

Another critical tool is workflow orchestration, and the one Hertz uses is Apache Airflow.

So as to retain data fidelity, data manipulations, transformations, flows, and results of any actions against data have to be tracked with AWS monitoring services, such as AWS Cloud Watch and AWS Cloud Trail. Hertz also used Cloud Watch alerts to bring immediate attention to some critical metrics or actions that require close attention.

In an effort to protect data, all data in transit and at rest need to be encrypted.

4 Hertz Data Platform Architecture

Detailed Hertz data platform architecture is provided in Figure 2. In this report, we will touch only on services directly involved in big data management: Data Input Services, Data Storage, Data Processing, Databases, Data Warehouse, Monitoring Services, and Orchestration Tool. Most AWS services involved in the big data pipeline are serverless (no need to manage infrastructure for these services, AWS will provide infrastructure under the hood). Hertz uses the Infrastructure as Code(IaC) approach for provisioning infrastructure (Infrastructure as Code or IaC is the process of provisioning and managing infrastructure defined through code instead of doing so with a manual process).

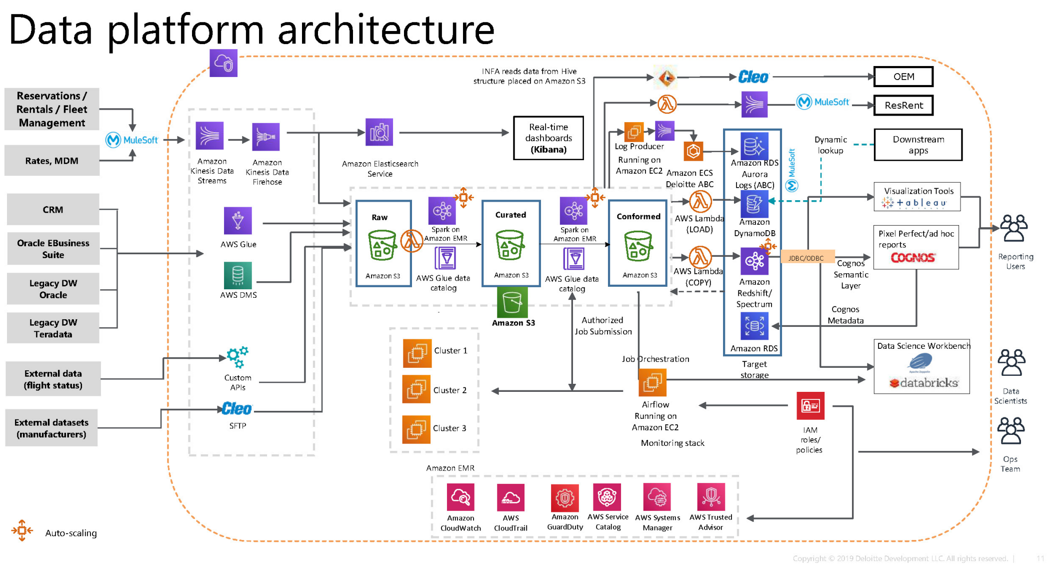

Figure 2: Hertz Data Platform Architecture

4.1 Data Input Services

Data input of Hertz data platform is done through the following three major approaches:

- Amazon Kinesis Data Stream is a massively scalable, fully managed, serverless, highly durable data ingestion and processing service optimized for streaming data. It allows configuring hundreds of thousands of data producers to put data into a Kinesis data stream continuously. Data will be available within milliseconds to Amazon Kinesis applications, and those applications will receive data records in the order they were generated.
- Amazon Kinesis Data Firehorse is a fully managed service for delivering real-time streaming data to destinations such as Amazon S3, Amazon Redshift, Amazon OpenSearch Service, Splunk, and any custom HTTP endpoint or HTTP endpoints.
- AWS Glue is a serverless data integration service that makes it easy to discover, prepare, and combine data for analytics, machine learning, and application development.

Hertz uses Kinesis Data Streams to read massive data, mostly event-based and IoT. Then, Kinesis Data FireHorse consumes data from Kinesis Data Streams and stores all raw data at the S3 row level. AWS Glue consumes data from legacy systems and writes data to the S3 row level. Hertz utilizes SFTP (Secure Transfer File Protocol) or custom APIs to communicate with external systems.

4.2 Data Storage

For the generation of data lake and storage of processed data, Hertz takes the way of S3, an object storage service offering industry-leading scalability, data availability, security, and performance. This means customers of all sizes and industries can use it to store and protect any amount of data for various use cases, such as data lakes, websites, cloud-native applications, backups, archives, machine learning, and analytics. Amazon S3 is designed for 99.999999999% (11 9's) of durability and stores data for millions of customers all around the world.

Hertz has three different layers of S3: raw, curated, and conformed, where untouched data from sources come to the S3 raw layer, and Firehorse or Glue writes to the S3 raw layer. Processed data moves from the S3 raw layer to S3 curated, and S3 conformed. Spark on EMR is used as a processing tool. S3 conformed level is the final destination of the data. Redshift/Spectrum, Aurora, and Dynamo DB read data from S3 Conformed.

4.3 Data Processing

In the Hertz data platform, Spark on Amazon EMR act as a data processing tool. Apache Spark is a multi-language engine executing data engineering, data science, and machine learning on single-node machines or clusters. Amazon EMR (Amazon Elastic MapReduce) is a managed cluster platform that simplifies running big data frameworks, such as Apache Hadoop and Apache Spark, on Amazon to process and analyze vast amounts of data. Using these frameworks and related open-source projects, you can process data for analytical purposes and business intelligence workloads. Amazon EMR also lets you transform and move large amounts of data into and out of other Amazon data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB. Lambdas are used to trigger Spark jobs on EMR. Lambda is also used to copy data from S3 conformed layer to Redshift data warehouse. AWS Lambda is a serverless, event-driven compute service that lets you run code for virtually any type of application or backend service without provisioning or managing servers.

4.4 Databases

Hertz data platform comprises three typical types of databases, Amazon Aurora, DynamoDB, and Amazon Relational Database Service (Amazon RDS). Amazon RDS is a collection of managed services that makes it simple to set up, operate, and scale databases in the cloud. Amazon Aurora is a fully managed relational database engine compatible with MySQL and PostgreSQL. Many companies prefer Aurora for log storage. Amazon DynamoDB is a proprietary, fully managed, serverless, key-value NoSQL database designed to run high-performance applications at any scale. DynamoDB offers built-in security, continuous backups, automated multi-Region replication, in-memory caching, and data export tools.DynamoDB is usually used for fast exploration.

4.5 Data Warehouse

Hertz data platform uses Amazon Redshift with Spectrum as the data warehouse. Amazon Redshift is a fully managed, petabyte-scale data warehouse service in the cloud. Spectrum is a tool within Amazon Redshift that allows fast, complex analysis of objects stored on the AWS cloud. With Redshift Spectrum, an analyst can perform SQL queries on data stored in Amazon S3 buckets. Data from conformed S3 bucket is copied to Redshift via Lambda.

4.6 Monitoring Services

To ensure the fidelity of data, Hertz data platform uses Amazon Cloud Watch and Amazon Cloud Trail to track all data activities and as well as throw alerts. Amazon Cloud Watch is a monitoring and observability service. CloudWatch is a tool to monitor applications, respond to system-wide performance changes, and optimize resource utilization, resulting in the messages in the form of logs, metrics, and events. Amzon Cloud Trail is a service that helps enable governance, compliance, and operational and risk auditing of AWS account. Actions taken by a user, role, or an AWS service are recorded as events in CloudTrail. Events include actions taken in the AWS Management Console, AWS Command Line Interface, and AWS SDKs and APIs.

4.7 Orchestration Tool

Apache Airflow is a top-rated workflow management platform, and its open-source and free nature has attracted a large number of users to it. Today, the Apache Airflow ecosystem is maintained by hundreds of members. Apache Airflow is commonly used to run ETL jobs, manage machine learning pipelines, and automate development tasks.

Today, using Amazon-managed Apache Airflow workflows, the user can write, schedule, and monitor workflows using airflow from within AWS without facing the everyday challenges of running the airflow environment. Managed workflows take care of the airflow environment setup, scaling, security, and handling upgrades and monitoring. This means the user can spend less time managing your airflow environments and more time using them to perform your data processing workflows in the cloud.

Airflow workflows are typically represented as a collection of all the tasks you want to run, organized in a way that reflects their relationships and dependencies. In technical terms, this collection of tasks is called a directed acyclic graph, or DAG. To define workflows in airflow, users typically write DAGs in Python.

5 Summary

In summary, utilizing AWS services allowed Hertz successfully complete big data transformation. Created big data architecture is secure, reliable, and cost-effective. As a result, the company could centralize data from different sources in one place, significantly improving data analytic and reporting processes and achieving all defined business goals.

References

[1] IBM Corporation. How big data is giving hertz a big advantage.

[2] AWS Events. Invent 2019: Accelerated analytics: Building the next-gen data platform for hertz, 2019.

[3] Afnan Rehan. What is data transformation and how it optimizes business processes, 2020.

[4] Amazon Web Services. Best practices for using apache spark on aws, 2016.

[5] Amazon Web Services. Free data lakes and analytics on aws, 2022.

[6] Wikipedia. Amazon web services, 2022.