Final Presentation - Cloud Native Pipeline

Team members: ***Lavender Juan, Fengze Han, Shradha Tiwari, Vineesha Devathi, Tejas Dhomne, Wangzhen Yu***

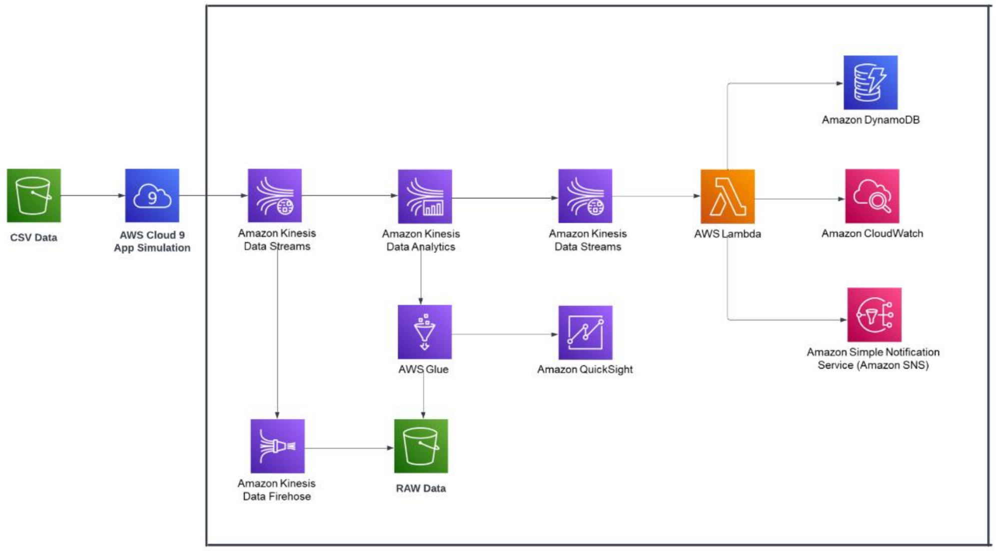
The name of our product is Cloud-Native Analytical Pipeline, and it is a real-time scalable product that enables a faster intelligent decision-making process for retail businesses by eliminating the latency gap between data engineering and business analysis through processing real-time data and extracting timely insights. Moreover, it delivers key insights to answer specific business problems based on brand popularity, purchase behaviors, and customer details. This product provides an affordable, easy-to-use service for better operation of successful deliveries/purchases and a delightful customer experience for the growth of retail businesses. The scalability of the service enables businesses to store millions of records in a database considering the change in volume of data based on seasonality or long weekends for timely delivery of the item/product. The time-sensitive data is processed immediately in real-time to eliminate delays in operations for a positive customer experience. Moreover, batch data processing can be done for the historical data to gain impactful business insights to optimize the item/product delivery process for the business. Our product helps to solve key business questions by simulating the logs of user purchases, product views, cart history, and the user's journey. Using insights on customer details, businesses can determine the number of unique visitors per day during a certain time to increase the purchase conversion rate, analyze customer behavior reports on when users add products to their carts but don't buy them to avail sponsored ads on an item, and determine the top categories per hour or weekday to promote discounts based on trends for higher sales. Finally, the brand popularity to know which brands need more marketing for higher reachability of the products converts to purchases to increase the overall revenue of the business.

Analytics is the procedure of gathering statistics from all the assets that influence numerous provider providers. Analysts can then make use of these statistics to infer modifications in client conduct and retail buying patterns. Retail analytics spans the complete client adventure, from discovery through acquisition, conversion, and finally retention and support. We will use an ecom retail dataset to simulate the logs of person purchases, product views, cart history, and the person's adventures on the web platform to create analytical pipelines in batch and Real-time. The batch processing will contain statistics ingestion, Lake House architecture, processing, visualization, the use of Amazon Kinesis, Glue, S3, and QuickSight to attract insights concerning the following: Unique site visitors according to day, During a positive time, the customers upload merchandise to their carts, but do not purchase them, Top classes according to hour or weekday (promotions/discounts) and to recognise advertising strategies for manufacturers.

Because we intend to work with e-commerce websites, we may make our services accessible to customers via Big Data channels. We can use the data in this way to quickly identify trends, spot patterns, and adjust our marketing tactics as necessary. Additionally, clients typically favor cloud-based systems and web apps over server-based ones. Using a cloud-based platform like Customer Relationship Management (CRM) or Salesforce will enable you to have an abstract interface. The POS data will then be collected by connecting this to Amazon Kinesis Data Streams. After doing analysis on the streaming data, Quicksight can offer insights. Since we'll be promoting our business using the cloud, it can be looked up as a Platform as a Service (PaaS).

Architecture Link -

https://lucid.app/lucidchart/947c2f5a-e4e6-427a-8674-385a29926a48/edit?viewport_loc=-120%2C-115%2C2711%2C1128%2C0_0&invitationId=inv_37748d2e-63f2-4b0f-b4c2-908022ea6299



Performance, security, manufacturing scalability, and data availability are all attributes of the object storage service **Amazon S3**. Users may store and retrieve any quantity of data with

Amazon S3 at any time, from any location. **Amazon Kinesis Data Streams** is a real-time data collecting and processing service from Amazon. Utilizing Kinesis Data Streams, applications may be created for data processing. Kinesis Data Firehose is a component of the Kinesis streaming data platform, which also includes Amazon Kinesis Data Analytics, Kinesis Data Streams, and Kinesis Video Streams. Kinesis Data Firehose users are not required to manage resources or create applications. If the data producers are set up to send data to Kinesis Data Firehose, the data will be automatically sent to the specified destination. Before delivery, data can also be transformed using Kinesis Data Firehose for analytics, machine learning, and application development. It is simple to find, prepare, and combine data using a server-less data integration service. It runs Spark/Python programs at a minimal cost without maintaining Infrastructure. Only while the job is in progress do you pay. Additionally, you must pay storage costs for the items in the Data Catalog. The **AWS Glue** Data Catalog may have tables added by a crawler. The vast majority of people that utilize AWS Glue do it in this manner. In a single run, a crawler can browse many data sources. The crawler then adds to or modifies one or more tables in your Data Catalog. Data doesn't need to be loaded while utilizing **AWS Athena**, an interactive query service for S3, because it stays in S3. It supports a number of data formats, including AVRO, CSV, JSON, and ORC, and is server-less. **Apache Flink** is a scalable platform for data analytics and a distributed processing engine. Huge data streams can be managed with Flink, which can also give your streaming application real-time analytical insights on the processed data. Any size of in-memory calculations may be performed with Flink in a variety of cluster setups. For distributed computations over data streams, Flink also offers data distribution, communication, and fault tolerance. Flink applications employ unbounded or bounded data sets to process streams of events. Unbounded streams lack a defined ending and are handled indefinitely. Bounded streams have a start and finish that are known and may be handled in batches. **Amazon Quick Sight** is a scalable, embeddable, serverless, machine learning-powered business intelligence (BI) tool created for the cloud. By only charging you when your customers use their dashboards or reports, it is the first BI solution to provide pay-per-session pricing, enabling cost-effective large-scale deployments. Just a few of the sources it can connect to include Redshift, S3, Dynamo, RDS, files in the JSON, text, CSV, and TSV formats, Jira, Salesforce, and an on-premises Oracle SQL server. Users may clean and normalize data with **AWS Glue DataBrew's** visual data preparation tool without writing any code. Compared to bespoke data preparation, DataBrew speeds up the process of preparing data for analytics and machine learning. There are several pre-built transformations available to automate data preparation activities including checking for anomalies, converting data to standard formats, and fixing inaccurate numbers. **Amazon DynamoDB** is a fully managed key-value NoSQL database service that provides quick, dependable performance and simple scaling. Developers are relieved of the management responsibilities associated with maintaining and growing a distributed database thanks to DynamoDB. You may build database tables using DynamoDB that can accommodate any number of queries and any quantity of data storage and retrieval. You may alter the throughput capacity of your tables without impacting their performance or availability. Amazon DynamoDB supports PartiQL, an open-source SQL-compatible query language that enables efficient data querying independent of where or how it is stored.

Strength: An efficient data pipeline that includes techniques for transferring data from one system to another. Whether the data is updated or not, it can be processed in real time instead of batch. Additionally, data pipelines include ingesting data using various methods, storing raw data, cleaning data, validating and transforming data into a queryable format, displaying KPIs, and managing the above processes. Includes important tasks.

Weakness: There are some limitations on acquiring data resources consisting of purchase records of the consumers.

Opportunities: As the retail cloud industry grows rapidly, this service will likely be in high demand, increasing the needs for our product. There are many beneficial aspects of our technology for our customers, such as Cloud Scaling and integration with third-party applications. Additionally, providing recommendations to new users based on the newly created networks.

Threats: Because cloud infrastructure is intended to be simple to use and facilitate quick data exchange, organizations find it challenging to guarantee that data is only accessible to authorized parties. The cloud security posture management techniques in our software, however, could be insufficient to safeguard their cloud-based infrastructure. A security breach can result in data leaks and the release of a lot of personal information about individuals, which would be extremely harmful to our business and our customers.

Speaking about the market, the United States is the region that we want to concentrate on as it has the biggest market share in the retail sector. The eCommerce retail industry is expected to see an increase in the need for big data applications and analytics, which will increase the emphasis on data-determined models with the goal of understanding customer preferences and meeting their demands. Global corporations like Kofana, Salesforce and SAP Sales Cloud, who are now dominating the market for cloud-based business intelligence services, are our major rivals. However, our product offers certain distinctive qualities like an openly integrated system on AWS that has an understandable technical process for our clients. Second, more stability and compatibility, given that AWS now holds the largest market share in the IaaS Cloud sector will elevate the expectation of our clients' experience. Third, our solution will have a user interface (UI) that is simple to use and engaging. The environment will be thoughtfully built to meet consumer expectations, and our product will do so without sacrificing the processing power provided by our competitors.

Our product is a membership and subscription based SaaS business, therefore, our products will be charged weekly, monthly or annual membership fees. For revenue from membership fees, Gross Sale Revenue, each period prices (weekly, monthly, and annual) x number of customers in corresponding period, is a good metric to measure. It represents the total income from sales. And Net Sale Revenue is the other great metric, using the Gross Sale Revenue minus discounts and returns.

Not to mention, as our product is a SaaS, we will charge our clients through the licensing and subscriptions of our software in order to fund our revenue goals. A pay-based on-usage revenue model is what we intend to develop. Pay as you go is seen favorably because our model is a SaaS business model. The estimation of the quantity of new clients we will gain over time is the first stage in creating our revenue model. We can divide them into 3 different subscription options once we know how many additional clients we will bring on board.

The main cost of our product is made up of five parts: AWS service fee, marketing expense, personnel cost, infrastructure cost and attorney fee. Firstly, we need to pay AWS a certain amount of money to use their services, including Amazon S3, Amazon Kinesis Data Streams, Amazon Kinesis Data Analytics, Amazon Kinesis Data Firehose, AWS Glue, Amazon QuickSight, AWS Lambda, Amazon DynamoDB, Amazon CloudWatch and Amazon SNS. Secondly, we need to pay a marketing team to help us advertise and find potential customers. Since it is a new product, we are in great need of promotion and advertising to expand the market and build relationships with potential customers. For the personnel cost, we need to hire relevant technical personnel to help us build our business model, build our web pages or computer programs, and design an easy-to-use and attractive user interface (UI). Meanwhile, we also need operators to monitor data in real time and deal with problems arising from data analysis. Infrastructure costs are the costs we incur from renting office space, purchasing office supplies and related electronics while developing and operating our products. Attorney fees are paid to attorneys who provide us with legal services, such as drafting contracts with clients. Estimate growth rate.

Monitoring a company's SaaS growth rate is crucial since it'll indicates how successful our products, it can be calculated using the formula for revenue growth rate i.e. (Second Year Revenue – First Year Revenue) / First Year Revenue * 100 = % Revenue Growth Rate. This number can be useful if our product has to make any long-term adjustments to its marketing tactics or service delivery. To monitor and evaluate growth over a five-year period, we can compute the growth rates by using either of the five strategies such as Monetization, Moving into new markets, Moving upmarket, Moving down market or Product expansion. Using these strategies, if our product is widely accepted we can estimate an increase of more than 100% growth rate for a year and steady for upcoming 2 years. Therefore, the overall estimated growth rate can reach up to at least 400% (Approx) during the 5th year after our product launch.

We will be performing product analysis and customer analysis to enhance the features and gain insights which lines with the user needs and boosts sales for the grocers. Analysis on the product will give us information about the consumer behavior and guides us in making further decisions. Analytics on the sales data and customer data is a vital source for this product. To increase the sales, we need to understand the customer shopping patterns, festival shopping patterns and most purchased items. Moreover, competitor analytics such as discounts and hyped products need to be identified. The next step would be targeting, which would incorporate behavioral factors (recent activity, category clusters), derived factors (transaction data, propensity models), and demographic factors (gender, geography, consumer/business).