Name : Lauren Abigail

NIM : 2602108426

Major : Data Science

Class : LC09

Course : Big Data Infrastructure and Technology

Big Data Analytic [LO 2 – 50 Points]

1. **Analyze and explain each part of the ELT process in Figure 1.**

   Answer:

   - **Extraction:** This is the first step in the ELT process, where data is collected from various sources. These sources could be databases, cloud storage, IoT devices, or external APIs. The key objective here is to gather raw data in its most granular form.
   - **Loading:** In this step, the extracted data is loaded into a staging area in a data warehouse or data lake. This staging area acts as a temporary storage location where data can be transformed and cleaned before being moved to its final destination.
   - **Transformation:** This is the crucial step where the raw data is transformed into a more useful format. This can include cleaning (removing duplicates or errors), normalizing (converting data into a consistent format), aggregating (summarizing data), and enriching (combining with additional data sources). Once the data is transformed, it is then loaded into the final data storage system where it can be analyzed and used for business insights.

2. **Choose one case or problem and use the ELT process to express and perform data operation for your case.**

   Answer:

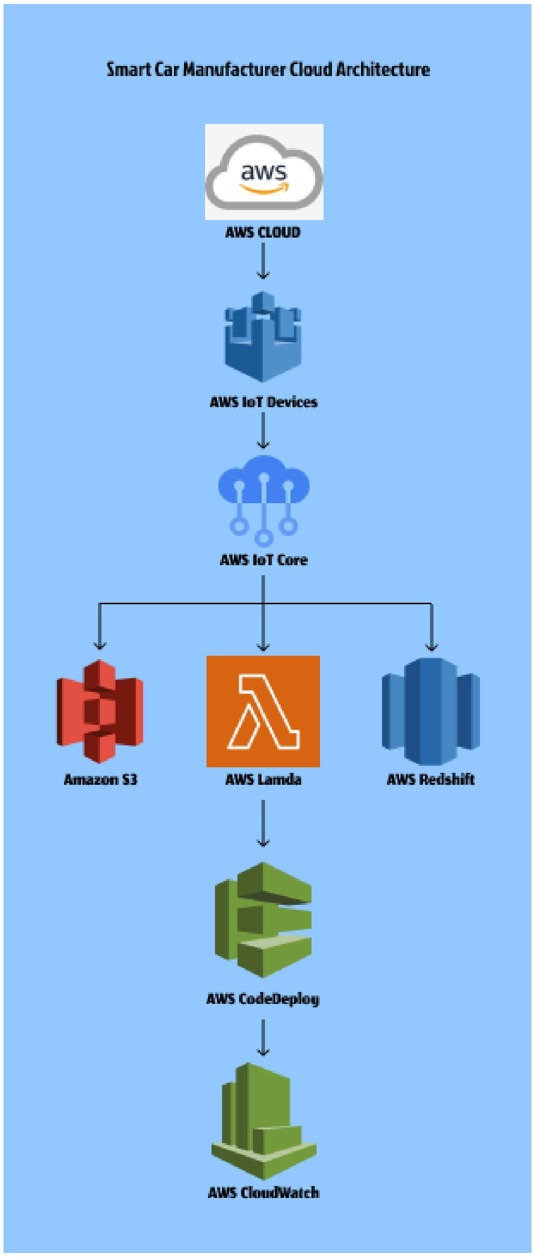
   - **Case:** Analyzing customer purchase behavior for a retail company.
   - **Extraction:** Collect data from multiple sources such as online sales databases, in-store sales records, customer feedback forms, and loyalty program data.

- **Loading:** Load the extracted data into a staging area within the company's data warehouse. Ensure that data from all sources is loaded in its raw form without any preprocessing.
- **Transformation:** Clean the data by removing duplicates and correcting errors. Normalize data formats (e.g., date formats, currency formats). Aggregate data to calculate total sales per customer, average purchase value, and frequency of purchases. Enrich the data by integrating customer demographic information from the loyalty program database.
- **Loading (Final):** Load the transformed data into the final data storage system, which could be a data warehouse optimized for analytical queries. This data can now be used to perform detailed analyses on customer purchase behavior, identify trends, and make data-driven decisions for marketing and inventory management.

Cloud Architecture [LO 3 – 50 Points]

3. **Design a simple cloud architecture for a Smart Car manufacturer that utilizes the cloud for data collection and improving the overall features of their products via OTA updates.**

   Answer:



- **Data Collection:** Utilize IoT devices installed in the smart cars to collect data such as vehicle performance, sensor readings, GPS data, and user preferences. These devices send data to the cloud in real-time.
- **Data Ingestion:** Use AWS IoT Core to securely connect and ingest data from the smart cars. AWS IoT Core can handle massive volumes of data from millions of devices.
- **Data Storage:** Store the ingested data in Amazon S3 for scalable and cost-effective storage. Amazon S3 provides high availability and durability for the collected data.
- **Data Processing:** Use AWS Lambda for serverless data processing. Lambda functions can transform and process the data as it arrives, preparing it for further analysis.
- **Data Analysis:** Store processed data in Amazon Redshift for

analytical queries. Redshift allows for complex SQL queries and data analysis, enabling insights into vehicle performance and user behavior.

- ○ **OTA Updates:** Use AWS CodeDeploy to automate the deployment of OTA updates to the smart cars. CodeDeploy ensures that updates are rolled out efficiently and reliably to all vehicles.
- ○ **Monitoring and Management:** Utilize AWS CloudWatch to monitor the entire architecture. CloudWatch provides insights into system performance and alerts for any issues.

4. **A company providing chatbot services using generative AI currently accepts input only in text format. They plan to add a feature allowing users to input images for analysis based on user requests. Please select the most appropriate AWS services for this case and describe their implementation with respect to the following aspects:** a. **Cost optimization** b. **Reliability**

   Answer:

   - ○ **AWS Services Selection:**
     - ■ **Amazon Rekognition:** Use Amazon Rekognition for image analysis. Rekognition provides powerful image and video analysis capabilities, including object detection, facial recognition, and text extraction from images.
     - ■ **Amazon S3:** Store user-uploaded images in Amazon S3. S3 provides cost-effective and scalable storage.
     - ■ **AWS Lambda:** Use AWS Lambda for serverless processing of the images. Lambda can be triggered by S3 events (e.g., when a new image is uploaded), and it can invoke Rekognition for analysis.
     - ■ **Amazon API Gateway:** Set up an API Gateway to handle user requests for image analysis. API Gateway integrates seamlessly with Lambda and provides a secure and scalable interface for the chatbot service.
   - ○ **Cost Optimization:**
     - ■ **Use Amazon S3 Infrequent Access storage class** for storing images that are not frequently accessed. This reduces storage costs.

  - **Leverage AWS Lambda’s pay-as-you-go pricing** to ensure that you only pay for the compute time you consume, eliminating the need for dedicated servers and reducing costs.
  - **Amazon Rekognition’s pay-per-use pricing** ensures that you only pay for the image analysis tasks performed, without any upfront costs.
- **Reliability:**
  - **Amazon S3’s high durability and availability** ensure that user-uploaded images are reliably stored and accessible when needed.
  - **AWS Lambda and API Gateway** provide highly reliable and scalable processing capabilities, automatically handling varying loads and ensuring that the image analysis service is always available.
  - **Amazon Rekognition’s robust image analysis capabilities** are backed by AWS’s infrastructure, ensuring consistent and reliable performance.

Reasons for Choosing These Options:

- **Cost Optimization:** The selected services offer cost-effective solutions that scale with usage, ensuring that the company only pays for what it uses while maintaining performance.
- **Reliability:** AWS’s managed services provide high availability and fault tolerance, ensuring that the chatbot’s new image analysis feature is reliable and performs well under varying loads.