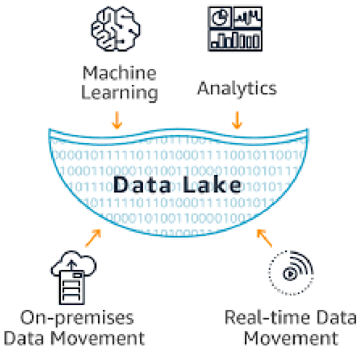
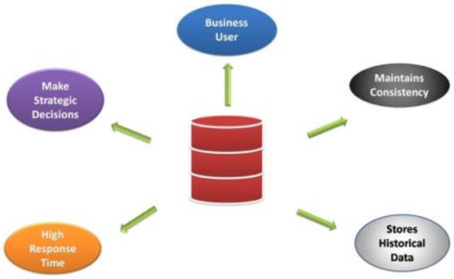
Assignment

1. **What is a Data Lake? Explain its benefits, how it differs from a data warehouse, and how it might benefit a client.**

- ♦ Data Lake is the centralized repository for hosting unprocessed raw data, consider as initial data landing zone for analytical insights. The format of the data is usually blobs, objects, or files.

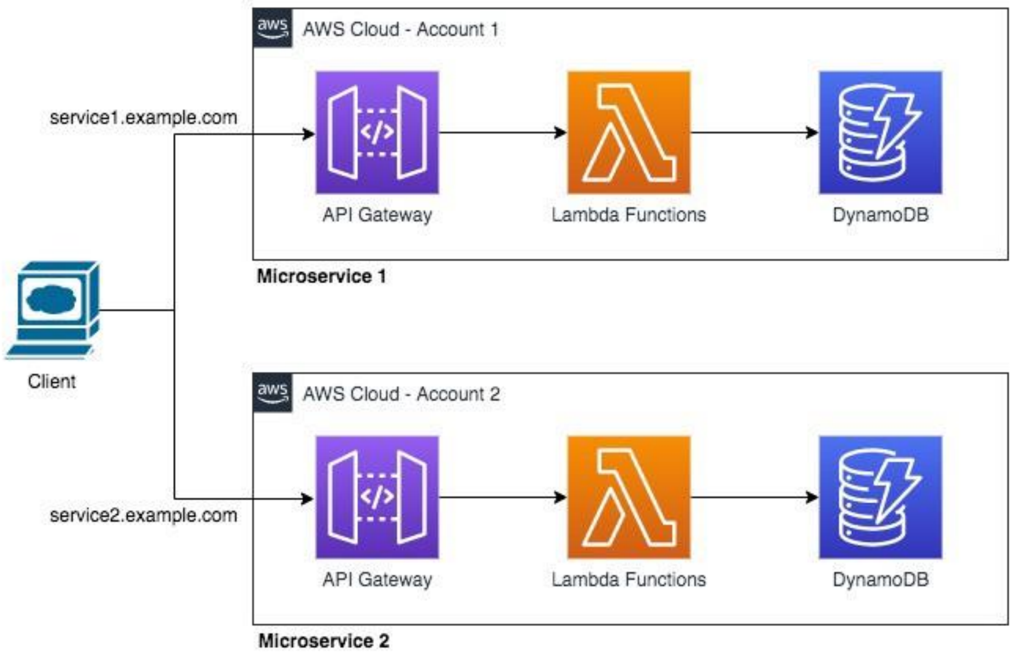
- ♦ Below is the difference between Data Lake and Data Warehouse.

Data Lake	**Data Warehouse**
 Machine Learning Analytics Data Lake On-premises Data Movement Real-time Data Movement	 Business User Make Strategic Decisions Maintains Consistency High Response Time Stores Historical Data
⇒ Raw format of the data (Unprocessed).	⇒ Structured format of the process data.
⇒ Data Lake use extract load and transfer (ELT) method.	⇒ Data Warehouse use extract transfer and load (ETL) method.
⇒ Ideal for in predictive and advance analytics.	⇒ Ideal for operational users.
⇒ Data is kept in raw format and is transformed only when it is ready to use.	⇒ The processed data faced problems of changes need to make in them.
⇒ Storing data in raw format is inexpensive.	⇒ Storing data is costlier and time consuming.
⇒ Data Lake follows schema on read format.	⇒ Data Warehouse follows schema on write format.

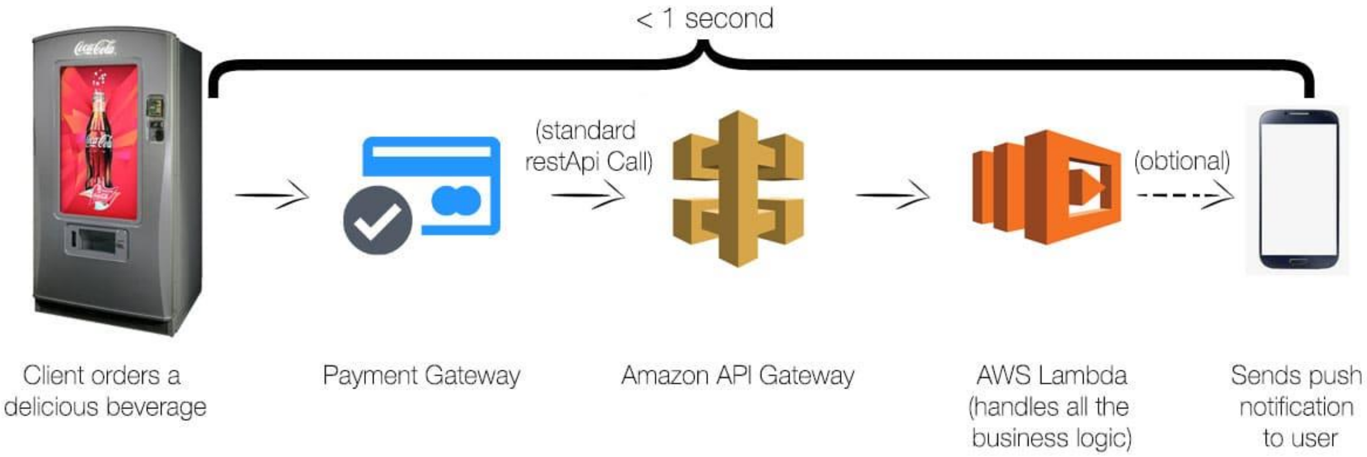
- ♦ Benefits of data lake:
  - o Always ensure data availability.
  - o Cheap scalability.
  - o Provides quality data for real time analytics.
  - o Handling data at speed.
  - o Stores data in any format.

Assignment

2. Explain serverless architecture. What are its pros and cons?



- Serverless architecture is the is an approach to software design that allows developers to build and run services without having to manage the underlying servers. It is easy for developers to write the code and cloud providers take care of the servers to run application, database, and storage systems. As shown in the above figure microservice model where client go to the website (Front end side) and on the back side AWS cloud environment handles the request and load the content for the user.
- The architecture is useful for developers to focus on writing the applications and can offload the responsibilities like managing the hardware, software, security updates and create backups in case of failure.
- The best example of serverless architecture is the vending machine of Coco Cola.


< 1 second
(standard restApi Call)
(obtional)
Client orders a delicious beverage
Payment Gateway
Amazon API Gateway
AWS Lambda (handles all the business logic)
Sends push notification to user

Assignment

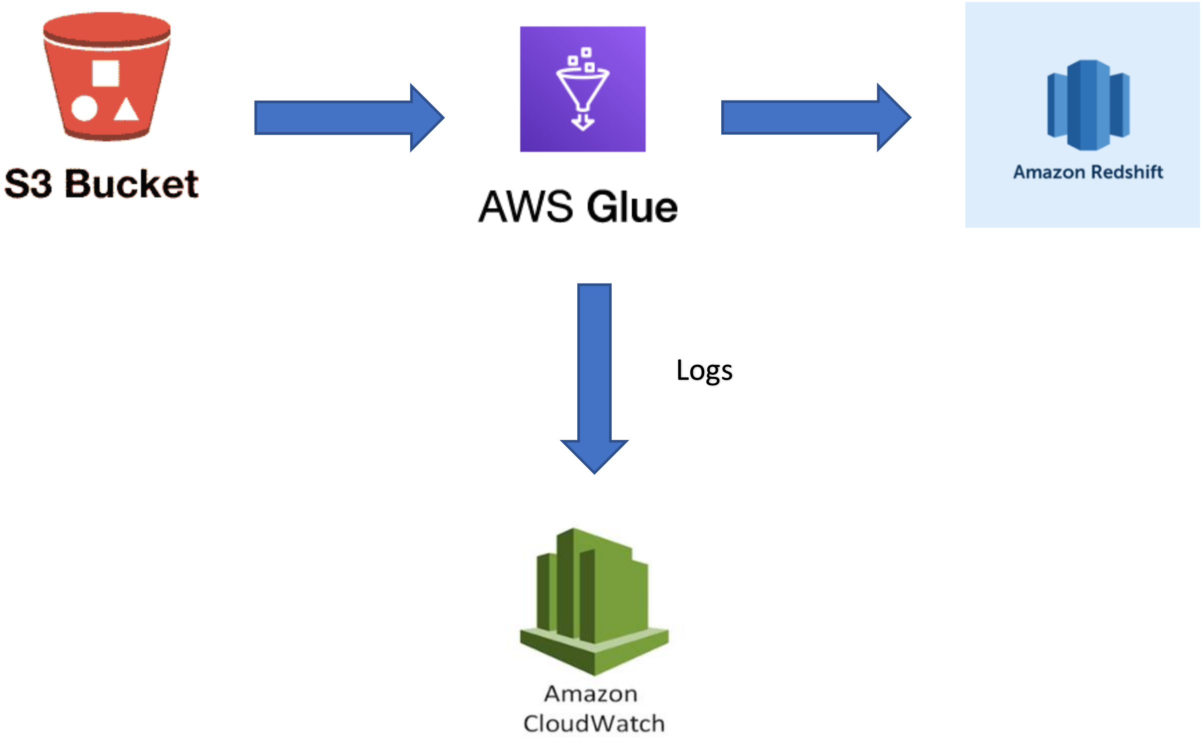
- Client buys Coco Cola product, the machine calls the payment gateway (offered by different vendors) to verify the purchase and it makes API call with Amazon API Gateway that triggers the AWS Lambda. The AWS lambda will handle all the logic behind the purchase and send the notification to user for payment.

- **Pros of Serverless Architecture:**
    - Cloud providers charge on per invocation basis, which prevents on paying for the unused servers or virtual machines.
    - Scalability (Autoscaling & Downscaling) can be performed based on web traffic.
    - Productivity of developers increase without managing the hardware's.

- **Cons of Serverless Architecture:**
    - In case of data center outage or other issues that impact the servers managing the serverless application customer must rely on cloud provider to fix the issue.
    - There are chances of the application data exposition if the architecture is not configured properly.
    - Performance impact is common cause (cold start) adding latency to the code execution when functions are invoked after period of inactivity.

3. Please provide a diagram of the ETL pipeline from Section 1 using serverless AWS services. Describe each component and its function within the pipeline.

- Below is the end-to-end flow of data from s3 bucket to redshift data warehouse.



Assignment

S3 Bucket: Initial landing zone for raw file to be stored in the s3 bucket. Files of any format can be stored in the bucket, over here the file format is CSV.

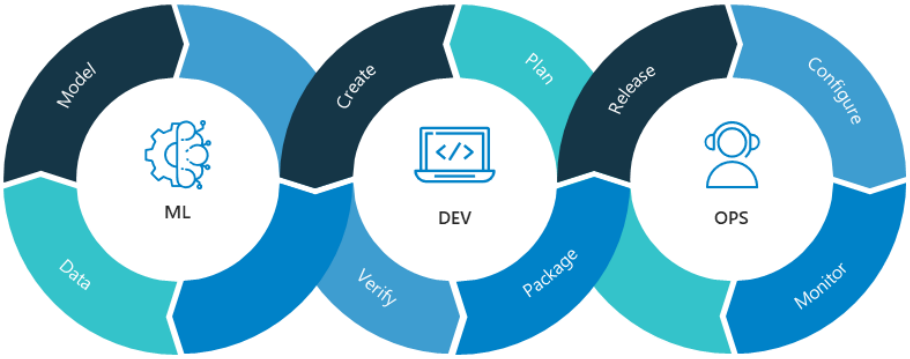
AWS Glue: It is an ETL tool use for the data integration that makes easy to transform the data for analytics, machine learning and software development purpose. Below are the steps performed inside the glue.

- Grant read/write access for raw data store in s3 bucket.
- Add crawler in glue studio to scan the raw data stored in the s3 bucket.
- Create IAM role (If not exist) and select the frequency to run the task as per need.
- Configure the crawler output to add the database. It will crawl the metadata from s3 bucket and load the data into the table inside the database.
- Run the crawler. Table will be created in the catalog.
- Establish the connection with the redshift by providing the username and password. Select the database, VPC, Subnet and security group. Test the connection by adding the IAM role.
- Implement crawler for redshift to add data into the table inside the schema defined inside the redshift database. Run crawler for testing purpose.
- Create job in the studio and apply the aggregation as per need.
- The data will be available inside the redshift after job get successfully executed.

AWS Redshift: The processed data get stored in the data warehouse for further operational use.

4. **Describe modern MLOps and how organizations should be approaching management from a tool and system perspective.**

- MLOPS is Machine Learning Operations, it is concept used to take machine learning models into the production and then monitor the performance for the same.
- MLOPS goals are hard to accomplish without a solid framework to follow. Automating model development and deployment means faster go-to-market times and lower operational costs. It's a useful framework for managers and developers to be more agile and strategic in their decisions.



Assignment

- Below are the approaches need to be considered while designing the framework.

  - Deployment and automation.
  - Reproducibility of models and predictions.
  - Diagnostics.
  - Governance and regulatory compliance.
  - Scalability.
  - Collaboration.
  - Business uses.
  - Monitoring and management.