Section 2: Consulting Soft Skills

➢ **Data Lake,its benefits, how it differs from a data warehouse, and how it might benefit a client.**

❑ Data Lakes allows us to store relational data such as operational databases and data from line of business applications, and non-relational data like mobile apps, IoT devices, and social media. They also gives us the ability to understand what data is in the lake through crawling, cataloging, and indexing of the data.

❑ **BENEFITS OF DATA LAKE :**

- Democratize Data;
- Get Better Quality Data;
- Data storage in native format;
- Scalability;
- Versatility;
- Schema Flexibility;
- Supports not only SQL but more languages;
- Advanced Analytics;

❑ **DATA LAKE VS DATA WAREHOUSE :**

	Data Lake	**Data Warehouse**
Data Structure	**Raw**	**Processed**
Purpose of Data	Not yet determined	Currently in use
Users	Data scientists	Business professionals
Accessibility	Highly accessible and quick to update	More complicated and costly to make changes

❑ **BENEFITS TO A CLIENT :**

A Data Lake provides the flexibility needed to store raw data and a common pool to combine multiple points and shape the data to provide useful insights that can be customized to meet the customers need and requirements.

➢ **Serverless architecture, its pros and cons.**

❑ A serverless architecture is a way to build and run applications and services without having to manage infrastructure. Your application still runs on servers, but all the server management is done by AWS.

❑ **PROS OF SERVERLESS ARCHITECTURE :**

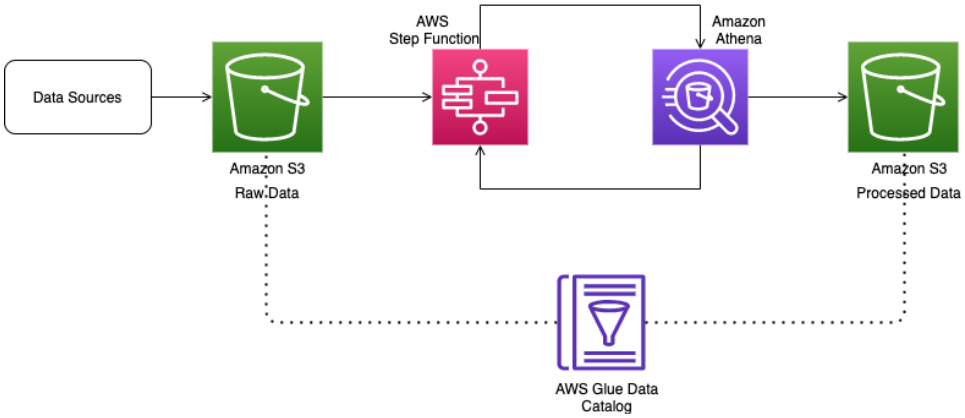
- No server management.
- Cost.
- Scalability.
- Security.

- Quicker time to market. Development environments are easier to set up and not having to manage servers leads to accelerated delivery and rapid deployment
- Reduced latency.
- Security.

❑ **CONS OF SERVERLESS ARCHITECTURE :**

- Vendor lock-in is a risk.
- Serverless architectures are not built for long-running processes.
- Debugging and testing is a challenge.
- Performance impact.

➢ **<u>Diagram of the ETL pipeline from Section 1 using serverless AWS services, description of each component, and its function within the pipeline.</u>**



- **DATA SOURCES :**

A data source is the location where data that is being used originates from.

- **AMAZON S3 RAW DATA :**

The raw data layer contains ingested *data* that has not been transformed and is in its original file format (for example, JSON or CSV).

- **AWS STEP FUNCTION:**

AWS Step Functions is a low-code, visual workflow service that developers use to build distributed applications, automate IT and business processes, and build data and machine learning pipelines using AWS services.

- **AMAZON ATHENA :**

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to manage, and you pay only for the queries that you run. Athena is easy to use

- **AMAZON S3 PROCESSED DATA :**

Processed data is known as information. This act of processing data generally involves the collection and manipulation of items of data, to create meaningful information, which then can be helpful to make certain decisions.

- **AWS GLUE DATA CATALOG :**

The AWS Glue Data Catalog is an index to the location, schema, and runtime metrics of your data.

➢ **Describtion of modern MLOps, and how organizations should be approaching management from a tool and system perspective.**

MLOps stands for **Machine Learning Operations**. MLOps is a core function of Machine Learning engineering, focused on streamlining the process of taking machine learning models to production, and then maintaining and monitoring them.

MLOps platforms provide value by making the business of machine learning more efficient. These platforms ultimately lead to more productive data scientists and more performant models, accelerating the revenue generation or cost savings targets of the models themselves.

The business and marketing people have to make use of ML in a broader and innovative approach to get desired functionality.