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

As the name suggest, Data Lake is a giant pool of data. It is a storage repository which holds large amount of data in its raw form. Since it stores the data in raw format, it can include both unstructured as well as structured data. It is beneficial for those clients that want the large volume of data storage without the need of processing it. In better words, companies that require large data storage without rapid transformation, data lake serves them best.

Data lake is different from data warehouse in terms of data processes and transformations. A traditional data warehouse is nothing but a transformed and processed data storage. Data warehouse holds the data in a structured format to facilitate advanced querying and reporting.

Some of the key differences between the two includes the following

* Data lake stores unstructured and structured data while data warehouse stores the historical data in a relational and structured manner.
* Data lake is often used by data scientist and data engineers as it a cost-effective data storage. Data warehouse is mainly used when there is a need for data driven business solutions. Basically, it is used by analysts to analyze the meaningful information affecting the business decision.
* The other important difference between the two could be in terms of size. Data lake can store petabytes of data including real time analytics and deep learning. Data warehouse only stores the data which is relevant for business analysis. Therefore, it mainly helps in aggregating and summarizing data with the help of queries.

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

The serverless architecture is a system where you can use services without having the manage the underlying infrastructure. In better words, it is a system where application is designed using Baas (Backend as a service), Saas (Software as a service) or Faas (Function as a service). Serverless architecture has basically done away the need for traditional always on server component.

Serverless architecture has made it possible for the developers to focus on code without worrying the need for maintaining and updating the containment architecture. In serverless architecture, all the responsibilities are undertaken by the third party (eg AWS, GCP, Microsoft Azure).

Now let’s study the pros and cons of serverless architecture:

* Effective cost: Cloud providers charge based on instance. You can turn off the service when not in use and wouldn’t be billed for that time duration.
* Increased Productivity: It has no doubt increased the productivity of the developers and engineers as they can focus only now on code.
* Lack of control: this is one of the main disadvantages as you lack the control over the software stack on which you run your code. In simple words, you are dependent on cloud provider. If any day the cloud services are failing, you cannot use until failure has been fixed.
* Data Security: Data security is very important as there can be chances of data leak in case of improper configuration. Since it is a shared server, data can be exposed to other server users.

**Please provide a diagram for an ETL pipeline (e.g., Section 2) using serverless AWS services. Describe each component and its function within the pipeline.**

AWS services use glue as an ETL tool and glue uses glue workflows to define and deliver the ETL pipeline and flowchart. The diagram can be easily made, all we need to do is provide the triggers, dependencies, and time duration of when the job is supposed to run. We can use Airflows, Matillion (a new ETL tool) to show the diagrams of workflow.

Diagram

Description automatically generated

The above diagram shows the functioning of ETL on AWS Server. The first step is to ingest the data within S3 buckets either by sync or through Http. Once the data is ingested within S3 bucket we can crawl the data to the staging area (landing skid). To crawl the data, we can use glue crawlers along with classifiers depending on the file format. Once we get the data in the landing skid, we can create glue jobs to relationalize, transform and parse the data within these 4 data layers (data layer, snapshot layer, analysis layer and reporting layer). Each layer has a certain specific task depending on the infrastructure. With the help of Boto3 package, we can use pyspark and sql spark to transform and load the data to the next layer. We can use Athena query engine for querying the data within these layers and can export the data for dashboarding and other reporting purposes.

In snapshot layer, we can partition the data, in analysis layer we can do the analysis and transformation and finally in the reporting layer, we can generate the data for query-based reporting.

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

Machine learning operations in layman terms refers to a process of deploying and maintaining the machine learning models in the production environment efficiently and effectively. These are practices used for collaborating and communicating so that it becomes easier to regulate and align the machine learning models with the business needs.

The modern MLOps are slowly progressing into machine learning life cycle management. It involves entire life cycle starting from information and data gathering to data cleaning, model training to model fitting and validating, and monitoring and retraining by tuning the hyper parameters.

The organizations should not approach MLOps as DevOps as the execution of both may vary despite being fundamentally similar.

MLOps is more experimental in nature as it requires tuning the parameters. Also, the testing would not only involve unit testing or code testing but would also include model validation and model training.

The deployment of machine learning model is more complex as it involves automating the task done manually by data scientists before deployment and model training. The team would include both data scientists as well as data engineers.

MLOps and DevOps being fundamentally similar are different in CI/CD. Monitoring machine learning models is complex as statistics in the production region should be monitored, in case of deviation from the expectation, roll backs should be done. Other than continuous integration and continuous deployment, continuous training should be done which involves automatic retraining of the models.