**DE-Fellowship – Week 10**

Task#3 Date: 17-May-2023

**AWS services that are related to DE.**

### **Data Ingestion Tools**

**Amazon Kinesis**

Amazon Kinesis offers several managed cloud-based services to collect and analyze streaming data in real time. Data engineers use Amazon Kinesis to build new streams, easily specify requirements, and begin streaming data. Additionally, Kinesis enables engineers to get data instantly and analyze it rather than waiting for a data-out report.

**AWS Snowball**

Snowball is the ultimate tool to deliver your enterprise data from on-premises databases to Amazon S3. To solve the problem of replicating data from on-site data sources to Cloud Storage, AWS uses a Snowball device to ship to the data source location and then connects it to the Local Network.

**AWS Storage Gateway**

Many companies have running on-site machines that are essential for daily tasks, but they may need regular data backup on Amazon S3. Fortunately, AWS Data Engineering features a Storage Gateway that allows organizations to transfer data from on-site data sources to Amazon S3 using the File Gateway configuration of Storage Gateway. It uses an NFS (Network File System) connection to share data to Amazon S3.

### **Data Storage Tools**

**Amazon S3**

Amazon Simple Storage Service or Amazon S3 is a data lake that can store any volume of data from any part of the internet. Since it is an incredibly scalable, quick, and affordable choice, Data engineers have the flexibility to duplicate their S3 storage across various Availability Zones with Amazon S3. Data engineers can effectively create web-based cloud solutions that expand automatically and have flexible setups owing to Amazon S3.

### **Data Integration Tools**

**AWS Glue**

AWS Glue is a fully managed ETL (extract, transform, and load) service for easily and affordably processing, improving, and migrating data between different data stores and data streams. Data engineers may interactively analyze and process the data using AWS Glue Interactive Sessions. Data engineers can visually develop, run, and monitor ETL workflows in AWS Glue Studio with a few clicks.

### **Data Warehouse Tools**

**Amazon Redshift**

Amazon Redshift is a petabyte-scale data warehouse cloud service that enables you to use your data to discover new insights about your clients and organization. Data engineers can gain insights from data with Redshift Serverless by easily importing and querying data in the data warehouse. Additionally, engineers can build schemas and tables, import data visually, and explore database objects using Query Editor v2.

### **Data Visualization Tool**

**Amazon QuickSight**

Amazon QuickSight is Amazon’s best tool that can easily help you create BI Dashboards within a few clicks. It can deliver Machine Learning insights. You can use Amazon QuickSight from a Web browser, Mobile device or embed QuickSight dashboard in websites, portals, or from various applications. AWS Data Engineering also focuses on integrating Amazon Redshift to many Businesses Intelligence and Business Analytics tools.

**Other Some Tools**

**AWS CloudWatch**

With the help of AWS CloudWatch, you can merge all your system, application, and AWS service logs into a single, highly scalable service. Data engineers can discover the logs for the services they run in CloudWatch, and keeping a debug log while developing is beneficial. Engineers can schedule the services they want to launch during a specific period using CloudWatch Events.

**Amazon IAM**

 AWS Identity and Access Management (IAM) is another popular AWS service that enables you to control access to AWS resources. The IAM service offers features for managing authorizations for actions against AWS services like Amazon SageMaker and Amazon S3 and identity management, including support for identity federation. IAM allows data engineers to design roles that adhere to the least privilege principle for each AWS service they use.

**AWS Lambda**

AWS Lambda is a serverless computing AWS service that executes your code in response to events and manages the underlying computing resources effortlessly. Lambda comes in handy when collecting the raw data is essential. Data engineers can develop a Lambda function to access an API endpoint, obtain the result, process the data, and save it to S3 or DynamoDB.

**Amazon EMR**

AWS Elastic Map Reduce (EMR) is one of the primary AWS Services for developing large-scale data processing that uses Big Data Technologies like Apache Hadoop, Apache Spark, Hive, etc. Data engineers can use EMR to launch a temporary cluster to run any Spark, Hive, or Flink task. It allows engineers to define dependencies, set up cluster setup, and find the underlying EC2 instances.

**Amazon DynamoDB**

Amazon DynamoDB supplies an alternative to relational database systems by using several data types, such as document, graph, key-value, memory, and search. Data engineers can use it to store semi-structured data with a unique key. To prevent race conditions, they can utilize DynamoDB to track the state of other services like Step Functions.

**Amazon Athena**

Amazon Athena is an interactive query tool for easily assessing data in Amazon S3 using SQL. Data engineers can start using Athena to extract some insights from the data once the metadata goes into the Data Catalog. When accessing GBs of data in Parquet format with strong partitions, engineers typically receive the results in a matter of seconds.

**Here is a rough sketch of a typical data pipeline flow in AWS:**

**Data Ingestion:** Raw data from various sources is ingested into the pipeline. This could be done using services like Kinesis Data Streams, AWS Data Pipeline, or directly from applications using SDKs.

**Data Storage:** The ingested data is stored in Amazon S3, acting as a central data lake. S3 supplies durable and scalable storage for the raw data.

**Data Preparation**: AWS Glue is used to transform and prepare data for downstream processing. Glue can perform data cleaning, normalization, schema evolution, and other transformations.

**Data Processing**: Processed data can be further analyzed or transformed using services like AWS Lambda or AWS EMR (Elastic MapReduce). Lambda functions or EMR clusters can be triggered based on events or schedules.

**Data Warehouse:** Processed data can be loaded into Amazon Redshift or other data warehousing solutions for efficient querying and analytics.

**Data Visualization and Analysis:** Analytical tools like Amazon QuickSight or third-party BI tools can be used to visualize and analyze the data stored in the data warehouse.