

BYTEWISE LIMITED

Data Engineering Track

Task: Week – 1 (Third Month)

Task No: 1

Task Date: 15-5-2023

Internee Name: Umer Farooq


Mentor Name: Ahtisham

❖ **Task Details:**

This task includes the following:

1. AWS Cloud Computing
2. AWS For Data Engineering
3. S3,
4. RDS,
5. GLUE,
6. Redshift,
7. IAM role,
8. Crawler,
9. Athena,
10. Quicksight,
11. Lambda,
12. EC2


❖ What is Cloud Computing:

 What is Cloud Computing?

cloud com·put·ing

noun

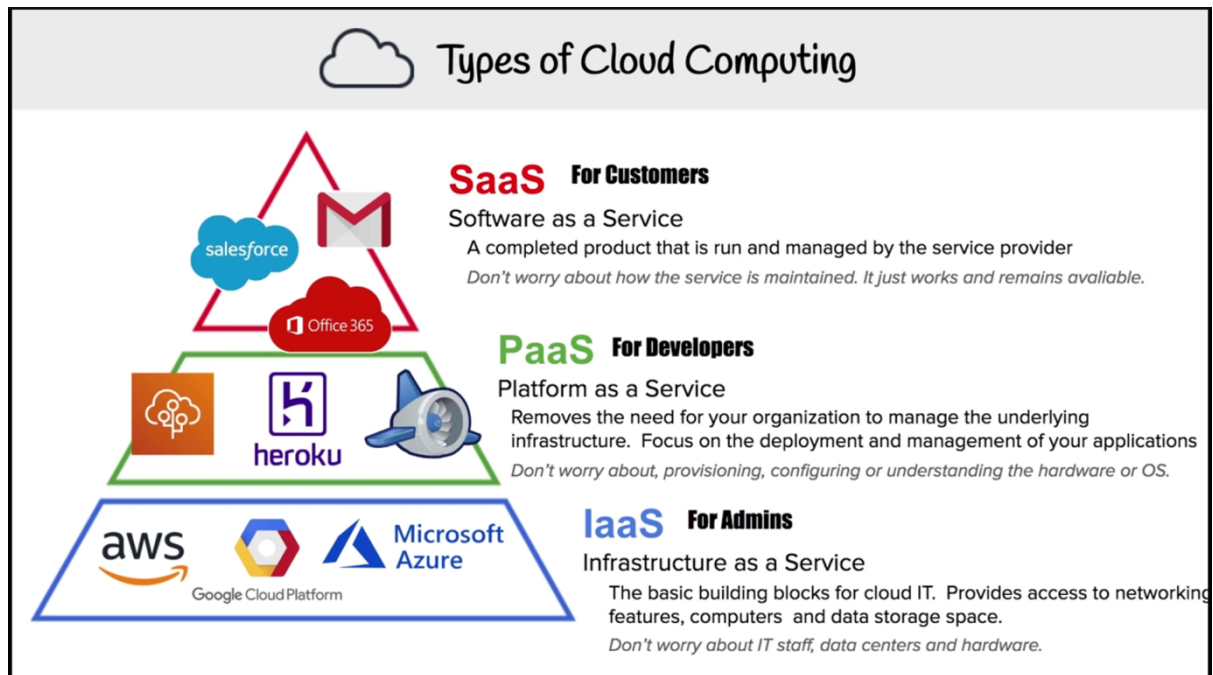
the practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer.



On-Premise	Cloud Providers
<ul style="list-style-type: none">• You own the servers• You hire the IT people• You pay or rent the real-estate• You take all the risk	<ul style="list-style-type: none">• Someone else owns the servers• Someone else hires the IT people• Someone else pays or rents the real-estate• You are responsible for your configuring cloud services and code, someone else takes care of the rest.

Cloud computing is the delivery of **hosted services over the Internet (the cloud)**. These services can include anything from data storage to computing power to software applications. Cloud computing is a broad term that encompasses a wide range of services.

Amazon Web Services (AWS) is a **cloud computing platform** that offers a broad set of global compute, storage, database, analytics, application, and deployment services that help organizations move faster, lower IT costs, and scale applications. AWS is the most comprehensive and broadly adopted cloud platform in the world, offering over 200 fully featured services from data centers globally.

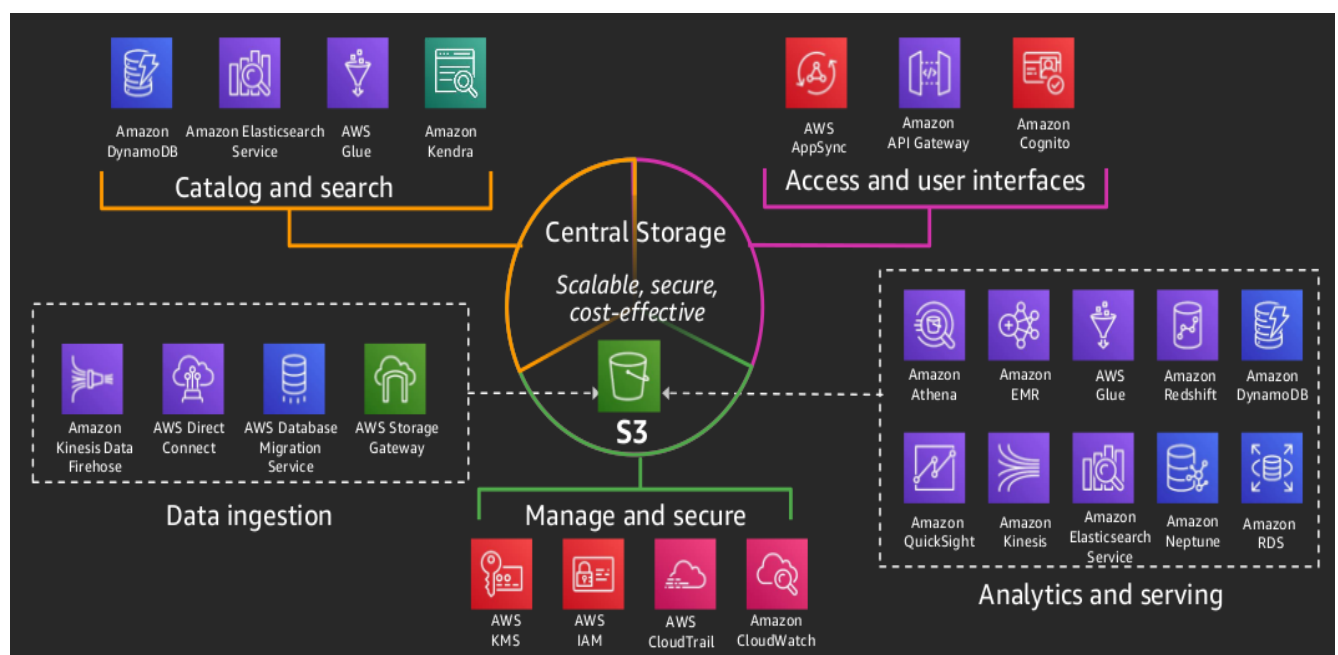


❖ How AWS Can be Used for Data Engineering:

AWS can be used for data engineering in a variety of ways. Some of the most common uses include:

- **Storing data:** AWS offers a variety of storage services, including Amazon Simple Storage Service (S3), Amazon Relational Database Service (RDS), and Amazon DynamoDB. S3 is a highly scalable and durable object storage service that can be used to store any type of data and any amount of data. RDS is a fully managed relational database service that offers a variety of database engines, including MySQL, PostgreSQL, and Oracle. DynamoDB is a fully managed NoSQL database service that offers high performance and scalability.

- **Processing data:** AWS offers a variety of compute services, including Amazon Elastic Compute Cloud (EC2), Amazon Elastic MapReduce (EMR), and AWS Lambda. EC2 is a web service that provides resizable compute capacity in the cloud. EMR is a managed Hadoop and Spark service that can be used to process large amounts of data. Lambda is a serverless compute service that can be used to run code without provisioning or managing servers.
- **Analyzing data:** AWS offers a variety of analytics services, including Amazon QuickSight, Amazon Athena, and Amazon Redshift. QuickSight is a cloud-powered business intelligence service that makes it easy to analyze data. Athena is a serverless, interactive query service that makes it easy to analyze data in S3. Redshift is a fully managed, petabyte-scale data warehouse that offers fast performance and scalability.



❖ AWS Services:

1). S3 Object Storage:

- Amazon Simple Storage Service (S3) is an object storage service that offers durability, scalability, availability, security, and performance. S3 can be used to store any type of data, including text, images, videos, audio, and binary files.
- Amazon S3 stores any type of data by breaking it down into objects. **An object is a uniquely identified collection of data that consists of a key and a value.** The key is a string that identifies the object, and the value is the data itself. Objects can be any size, up to 5TB, and they can be stored in any order.
- GCP offers Cloud Storage and Azure offers Blob Storage which provides the same functionality.

2). RDS:

- Amazon Relational Database Service (RDS) is a **fully-managed relational database service** that makes it easy to set up, operate, and scale a relational database in the cloud. RDS provides you with six familiar

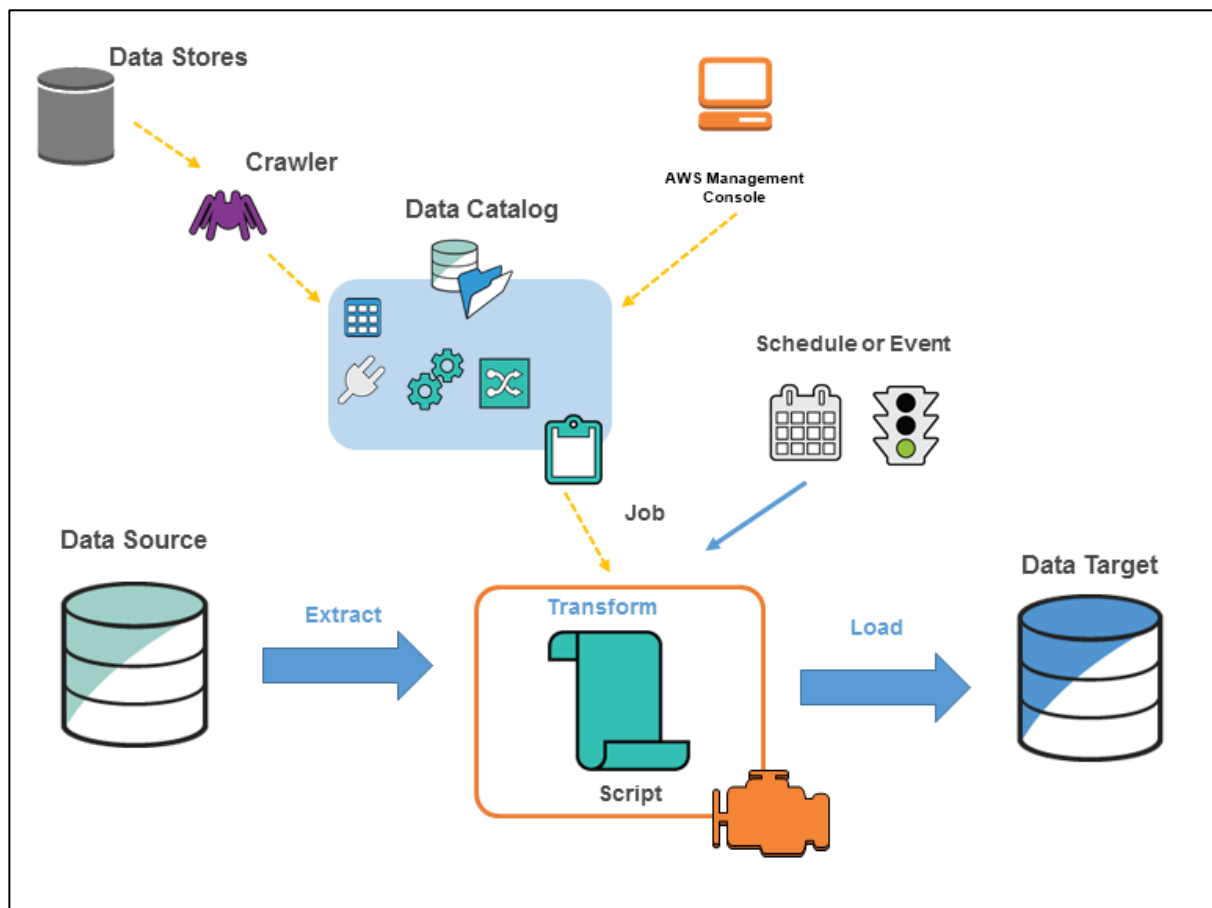
database engines to choose from, including Amazon Aurora, MySQL, MariaDB, PostgreSQL, Oracle, and SQL Server.

- **With RDS, you can focus on your applications and not on database administration.** RDS takes care of routine database maintenance tasks, such as backups, software patching, and hardware failure recovery. You can also scale your database up or down as needed, without having to worry about managing the underlying infrastructure.
- Similar services that GCP provides is Cloud SQL and Azure provides Azure SQL Database.

3). Glue:

- AWS Glue is a **fully-managed, serverless ETL** (extract, transform, and load) service that makes it easy to prepare data for analytics, machine learning, and application development. GLUE can be used to process data from a variety of sources, including S3, RDS, and DynamoDB.
- AWS Glue automates much of the effort required for data integration, including discovering data sources, determining data formats, and creating ETL jobs.

- **AWS Glue works by first crawling data sources to discover their schema and content.** Once the schema is known, AWS Glue can **create ETL jobs** that extract data from the sources, transform it into a desired format, and load it into a target data store. AWS Glue also provides a visual interface that makes it easy to manage ETL jobs and track their progress.
- Similar services that GCP provides is Google Cloud Data Fusion and Azure provides Azure Data Factory.



4). Redshift:

- Amazon Redshift is a **fully-managed, petabyte-scale data warehouse service in the cloud** that offers fast performance and scalability. It is **built on top** of technology from the **Massive Parallel Processing (MPP)** data warehouse company ParAccel, to handle large scale data sets and database migrations.
- Redshift works by storing data in columnar format, which is more efficient for analytical workloads. Redshift also uses a distributed architecture, which allows it to scale to handle large amounts of data.
- Similar services that GCP provides is Google BigQuery and Azure provides Azure SQL Data Warehouse.

5). IAM Role:

- An IAM role is an IAM entity that defines permissions for a user, group, or service. IAM roles can be used to grant permissions to access AWS resources.
- **IAM roles are needed for a variety of reasons**, including:

- To allow AWS services to access resources on your behalf. For example, Amazon S3 can assume an IAM role to access your Amazon S3 buckets.
- To allow users to access resources in different AWS accounts. For example, you can create an IAM role in your account and then grant a user in another account permission to assume the role.
- To allow applications to access resources without storing credentials in the application. For example, you can create an IAM role for an application and then configure the application to assume the role.

How to use IAM Role:

- To use an IAM role, you first need to create the role. You can create a role in the AWS Management Console, the AWS CLI, or the AWS SDKs. Once you have created the role, you need to attach a permissions policy to the role. The permissions policy defines the permissions that the service or user will have when they assume the role.
- Once you have created the role and attached a permissions policy, you can then assume the role. You can assume a role in the AWS Management Console, the AWS CLI, or the AWS SDKs.

6). Crawler:

- An AWS crawler is a tool that **helps you discover and catalog your data in AWS**. It can crawl data stored in Amazon S3, Amazon RDS, Amazon DynamoDB, and other data stores. Once the crawler has discovered your data, it will create a table in the AWS Glue Data Catalog that contains metadata about the data, such as the schema, location, and format.
- You can use the AWS Glue Data Catalog to manage your data and to create ETL jobs that process your data.
- **To use an AWS crawler**, you first need to create a crawler in the AWS Glue console. You can then specify the data stores that you want the crawler to crawl. The crawler will then start crawling the data stores and creating tables in the AWS Glue Data Catalog.
- Once the crawler has finished crawling your data, you can use the AWS Glue Data Catalog to manage your data and to create ETL jobs that process your data.

7). Athena:

- AWS Athena is an **interactive query service** provided by Amazon Web Services (AWS). It allows you to **analyze and query data stored in**

Amazon S3 (Simple Storage Service) using SQL statements, without the need to set up and manage traditional database infrastructure.

- Differences between Athena and Redshift is that Athena is a serverless, ad-hoc query service that directly queries data in S3, while Redshift is a fully-managed data warehousing service optimized for fast query performance on structured data. Athena is suitable for interactive querying of large datasets, while Redshift is more suitable for complex analytics, frequent data updates, and longer-running workloads.

8). QuickSight:

- AWS QuickSight is a business intelligence (BI) and data visualization service provided by Amazon Web Services (AWS). It enables users to create interactive dashboards, perform ad-hoc analysis, and generate insights from their data.
- It offers a user-friendly experience, facilitates collaboration, and leverages the scalability and integration capabilities of the AWS ecosystem.

9). Lambda:

- AWS Lambda is a serverless computing service provided by Amazon Web Services (AWS) that enables you to run your code in response to events or triggers, without the need for managing servers or infrastructure.
- It is a fully managed service, which means AWS takes care of the underlying infrastructure, including provisioning, scaling, and monitoring of the compute resources needed to execute your code.

10) . EC2:

- Amazon Elastic Compute Cloud (EC2) is a web service that provides resizable compute capacity in the cloud.
 - It is a powerful tool that can be used to run a variety of applications, including web applications, high-performance computing (HPC) applications, and machine learning (ML) applications.
-