**PERFORMING ETL AND VISUALIZATION THROUGH AWS**

**What is ETL?**

**Diagram, schematic

Description automatically generated**

ETL (Extract, Transform, Load), ETL is used in data warehousing.

If we talk about the whole process show in the above diagram. First step is to find a datamodel. For that you can take the data from data sources like Kaggle, google data set etc or if you want real time data we use API and web scraping to take the data from the website. So after having the datamodel we have to store that in a relational database. Then we need to extract the data and keep it in the storage (eg: aws s3) the we will create dimensional modeling then we write etl script to load the data into data warehouse and then from the datawarehouse we create visualizations to dashboard

**Performing everything on cloud**

Graphical user interface, diagram

Description automatically generated

We have different tools in amazon to do the job in the cloud.

We have aws rds it acts like a relational database in the cloud, then we have aws s3 where we extract the data from rds and keep the required data into aws s3. Then we have aws glue where we transform the data by using dimensional modeling. And then write some etl scripts and then load the data to aws redshift after that we can use ani BI tools to visualize the data in aws quicksight.

More details on AWS features

**What is AWS RDS?**

* Amazon RDS is a Relational Database Cloud Service
* Amazon RDS minimizes relational database management by automation
* Amazon RDS creates multiple instances for high availability and failovers
* Amazon RDS supports PostgreSQL, MySQL, Maria DB, Oracle, SQL Server, and Amazon Aurora

# **What is Amazon S3?**

Amazon Simple Storage Service (Amazon S3) is an object storage service that offers industry-leading scalability, data availability, security, and performance. Customers of all sizes and industries can use Amazon S3 to store and protect any amount of data for a range of use cases, such as data lakes, websites, mobile applications, backup and restore, archive, enterprise applications, IoT devices, and big data analytics. Amazon S3 provides management features so that you can optimize, organize, and configure access to your data to meet your specific business, organizational, and compliance requirements.

# **What is AWS Glue?**

AWS Glue is a serverless data integration service that makes it easy for analytics users to discover, prepare, move, and integrate data from multiple sources. You can use it for analytics, machine learning, and application development. It also includes additional productivity and data ops tooling for authoring, running jobs, and implementing business workflows.

With AWS Glue, you can discover and connect to more than 70 diverse data sources and manage your data in a centralized data catalog. You can visually create, run, and monitor extract, transform, and load (ETL) pipelines to load data into your data lakes. Also, you can immediately search and query cataloged data using Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum.

AWS Glue consolidates major data integration capabilities into a single service. These include data discovery, modern ETL, cleansing, transforming, and centralized cataloging. It's also serverless, which means there's no infrastructure to manage. With flexible support for all workloads like ETL, ELT, and streaming in one service, AWS Glue supports users across various workloads and types of users.

Also, AWS Glue makes it easy to integrate data across your architecture. It integrates with AWS analytics services and Amazon S3 data lakes. AWS Glue has integration interfaces and job-authoring tools that are easy to use for all users, from developers to business users, with tailored solutions for varied technical skill sets.

With the ability to scale on demand, AWS Glue helps you focus on high-value activities that maximize the value of your data. It scales for any data size, and supports all data types and schema variances. To increase agility and optimize costs, AWS Glue provides built-in high availability and pay-as-you-go billing.

# **What is Amazon Redshift?**

Amazon Redshift is a fully managed, petabyte-scale data warehouse service in the cloud. You can start with just a few hundred gigabytes of data and scale to a petabyte or more. This allows you to use your data to gain new insights for your business and customers.

The first step to create a data warehouse is to launch a set of nodes, called an Amazon Redshift cluster. After you provision your cluster, you can upload your data set and then perform data analysis queries. Regardless of the size of the data set, Amazon Redshift offers fast query performance using the same SQL-based tools and business intelligence applications that you use today.

# **What is Amazon QuickSight?**

Amazon QuickSight is a cloud-scale business intelligence (BI) service that you can use to deliver easy-to-understand insights to the people who you work with, wherever they are. Amazon QuickSight connects to your data in the cloud and combines data from many different sources. In a single data dashboard, QuickSight can include AWS data, third-party data, big data, spreadsheet data, SaaS data, B2B data, and more. As a fully managed cloud-based service, Amazon QuickSight provides enterprise-grade security, global availability, and built-in redundancy. It also provides the user-management tools that you need to scale from 10 users to 10,000, all with no infrastructure to deploy or manage.

QuickSight gives decision-makers the opportunity to explore and interpret information in an interactive visual environment. They have secure access to dashboards from any device on your network and from mobile devices.