## AWS DATA ANALYTICS VIRTUAL INTERNSHIP

Internship-I report submitted in partial fulfillment of requirements for the award of degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

By

**PASUPULETI REVANTH (21131A05K8)**

**PATAMSETTI SREE VIDYA (21131A05G8)**

**RANKIREDDY HEMA DURGA NANDINI (22135A0519)**

**KURMA DHANA LAKSHMI (22135A0514)**

****

Under the Esteemed Guidance of

Name of Course Coordinator Name of Course Mentor

Dr. CH. SITA KUMARI Dr. K. NARASIMHA RAJU

(Associate Professor) (Associate Professor)

Department of Computer Science and Engineering

**GAYATRI VIDYA PARISHAD COLLEGE OF ENGINEERING (AUTONOMOUS)**

(Affiliated to JNTU-K, Kakinada)

**VISAKHAPATNAM**

**2023 – 2024**

**Gayatri Vidya Parishad College of Engineering (Autonomous)**

##### Visakhapatnam

##### Gayatri Vidya Parishad College of Engineering

##### CERTIFICATE

This report on

##### AWS DATA ANALYTICS VIRTUAL INTERNSHIP

is a bonafide record of the Internship – I work submitted

By

**PASUPULETI REVANTH (21131A05K8)**

**PATAMSETTI SREE VIDYA (21131A05G8)**

**RANKIREDDY HEMA DURGA NANDINI (22135A0519)**

**KURMA DHANA LAKSHMI (22135A0516)**

In their V semester in partial fulfilment of the requirements for the Award of Degree of

**Bachelor of Technology in**

**Computer Science and Engineering**

During the academic year 2023-2024

##### Name of Course Coordinator Head of the Department

**Dr.CH. SITA KUMARI Dr. D. HARINI**

**( Associate Professor ) (Associate Professor)**

##### 

##### 

##### Name of Course Mentor

Dr. K. NARASIMHA RAJU

**(Associate Professor)**

**CERTIFICATES:**

** PASUPULETI REVANTH – 21131A05K8**

**PATAMSETTI SREE VIDYA – 21131A05G8**

****

** RANKIREDDY HEMA DURGA NANDINI - 22135A0519**

 **KURMA DHANA LAKSHMI - 22135A0516**

**ACKNOWLEDGEMENT**

We would like to express our deep sense of gratitude to our esteemed institute **Gayatri Vidya Parishad College of Engineering (Autonomous)**, which has provided us an opportunity to fulfill our cherished desire.

We thank our internship mentor **Dr. K. Narasimha Raju, Associate Professor** and our Course coordinator **Dr. Ch. Sita Kumari**, **Associate Professor**, **Department of Computer Science and Engineering,** for the kind suggestions and guidance for the successful completion of our internship.

We are highly indebted to **Dr. D.N.D. HARINI, Associate Professor and Head of the Department of Computer Science and Engineering, Gayatri Vidya Parishad College of Engineering (Autonomous),** for giving us an opportunity to do the internship in college.

We express our sincere thanks to our Principal **Dr. A.B. KOTESWARA RAO, Gayatri Vidya Parishad College of Engineering (Autonomous)** for his encouragement to us during this project, giving us a chance to explore and learn new technologies in the form of mini projects.

We are grateful for **EDUSKILLS,** **AICTE** and **AWS** for providing us this learning opportunity

Finally, we are indebted to the teaching and non-teaching staff of the Computer Science and Engineering for all their support in completion of our project.

**PASUPULETI REVANTH (21131A05K8)**

**PATAMSETTI SREE VIDYA (21131A05G8)**

**RANKIREDDY HEMA DURGA NANDINI (22135A0519)**

**KURMA DHANA LAKSHMI (22135A0514)**

### ABSTRACT

Data analytics is the science of analyzing raw data to make conclusions about that information. Data analytics relies on a variety of software tools ranging from spreadsheets, data visualization, and reporting tools, data mining programs, or open-source languages for the greatest data manipulation. Data Analysis mainly deals with data collection, data storage, data preprocessing and data visualization.

In the course-1 we have learnt about the **cloud computing**, Cloud computing is the on-demand delivery of compute power, database, storage, applications, and other IT resources via the internet with pay-as-you-go pricing. These resources run on server computers that are located in large data centers in different locations around the world. When you use a cloud service provider like **AWS**, that service provider owns the computers that you. This course deals with the following main concepts of compute services, storge services, management services, database services, compliance services, AWS cost management services.

As a part of our course-2 we learnt **data analytics** which deals with the raw data to draw solutions. In this course we have learnt about big data which is the main and foremost important tool for the data analysis so, this big data is very important for data analytics, and here comes the problem of storing the data for that we have learnt different tools used for data storage and how to analyze the big data and preprocess the data the following are the concepts that we learnt in this course. The main concepts includes storage using **AMAZON** includes amazon S3, amazon athena, amazon redshift, amazon glue, amazon sagemaker and amazon IOT analysis.

### INTRODUCTION

In 2006, Amazon Web Services (AWS) began offering IT infrastructure services to businesses as web services—now commonly known as cloud computing. One of the key benefits of cloud computing is the opportunity to replace upfront capital infrastructure expenses with low variable costs that scale with your business.

With the cloud, businesses no longer need to plan for and procure servers and other IT infrastructure weeks or months in advance. Instead, they can instantly spin up hundreds or thousands of servers in minutes and deliver results faster.

**Cloud computing** is providing developers and IT departments with the ability to focus on what matters most and avoid undifferentiated work like procurement, maintenance, and capacity planning. As cloud computing has grown in popularity, several different models and deployment strategies have emerged to help meet the specific needs of different users. Each type of cloud service, and deployment method, provides you with different levels of control, flexibility, and management.

**Cloud security** at AWS is the highest priority. As an AWS customer, you will benefit from a data center and network architecture which is built to meet the requirements of the most security- sensitive organizations.

AWS provides networking services to meet the needs of your applications. You can provision virtual servers, set up a firewall, configure Internet access, allocate and route IP addresses, and scale your infrastructure to meet increasing demand.

AWS provides the broadest selection of **analytics services** that fit all your data analytics needs and enables organizations of all sizes and industries to reinvent their business with data. From data movement, data storage, data lakes, big data analytics, log analytics, streaming analytics, business intelligence, and machine learning (ML) to anything in between, AWS offers purpose-built services that provide the best price-performance, scalability, and lowest cost.

Today, AWS provides a highly reliable, scalable, low-cost infrastructure platform in the cloud that powers hundreds of thousands of businesses in 190 countries around the world.

### INDEX

#### AWS CLOUD FOUNDATIONS

|  |  |  |
| --- | --- | --- |
| **S.NO** | **Topic Name** | **Page no** |
| MODULE 1 | **Introduction to cloud computing**   * 1. Introduction to cloud computing   2. Advantages of Cloud   3. Introduction to AWS   4. Moving to the AWS cloud | 13-14 |
| MODULE 2 | **Cloud economics and billing**   * 1. Fundamentals of pricing   2. Total cost of Ownership, case study   3. AWS Organization   4. AWS Billing & Cost Management   5. Technical Support Models | 15-16 |
| MODULE 3 | **AWS global infrastructure**   * 1. AWS Global Infrastructure   2. AWS Services & Services Categories | 17 |
| MODULE 4 | **AWS cloud security**   * 1. AWS Shared Responsibility Model   2. AWS IAM   3. Securing Accounts   4. Securing Data   5. Working to Ensure Compliance | 18-19 |
| MODULE 5 | **Networking and content delivery**   * 1. Networking Basics   2. Amazon VPC   3. VPC Networking   4. VPC Security   5. Route S3   6. CloudFront | 20-22 |
| MODULE 6 | **Compute**   * 1. Compute Services Overview   2. Amazon EC2   3. Amazon EC2 Cost Optimization | 23-24 |

|  |  |  |
| --- | --- | --- |
|  | * 1. Container Services   2. Introduction to AWS Lambda   **Activity:**  Introduction to AWS Elastic Beanstalk |  |
| MODULE 7 | **Storage**   * 1. AWS EBS   2. AWS S3   3. AWS EFS   4. AWS S3 Glacier | 25 |
| MODULE 8 | **Databases**   * 1. Amazon RDS   2. Amazon Dynamo DB   3. Amazon Redshift   4. Amazon Aurora | 26-27 |
| MODULE 9 | **Cloud Architecture**   * 1. AWS Well-Architected Framework Design Principles   2. Reliability & High Availability   3. AWS Trusted Advisor | 28 |
| MODULE 10 | **Auto scaling and monitoring**   * 1. Elastic Load Balancing   2. Amazon CloudWatch   3. Amazon EC2 Auto Scaling | 29 |

**AWS DATA ANALYTICS**

|  |  |  |
| --- | --- | --- |
| 11 | Big Data | 30 |
| 12 | Big data pipeline | 30-31 |
| 13 | Big data tools | 31-32 |
| 14 | Big data collection | 32 |
| 15 | Big data storage | 33 |
| 16 | Big data ingestion | 33-34 |
| 17 | Big data processing and analysis | 34 |
| 18 | Big data visualization | 35 |
| 19 | LABS 1-8 | 36-42 |
| 20 | CASE STUDY | 43-49 |
| 21 | CONCLUSION | 50 |
| 22 | REFERENCE LINKS | 51 |

# COURSE 1:

# AWS CLOUD FOUNDATIONS

### MODULE 1

#### CLOUD CONCEPTS OVERVIEW

##### Introduction to cloud computing:

Cloud computing is the on-demand delivery of compute power, database, storage, applications, and other IT resources via the internet with pay-as-you-go pricing. These resources run on server computers that are located in large data centre in different locations around the world. When you use a cloud service provider like AWS, that service provider owns the computers that you are using. These resources can be used together like building blocks to build solutions that help meet business goals and satisfy technology requirements.

The services provided by cloud computing are: 1.IAAS (infrastructure as a service)

2.PAAS (platform as a service) 3.SAAS (software as a service)

##### Advantages of cloud computing:

###### On-Demand Self-service

* + - Multi-tenancy

###### Offers Resilient Computing

* + - Fast and effective virtualization

###### Provide you low-cost software

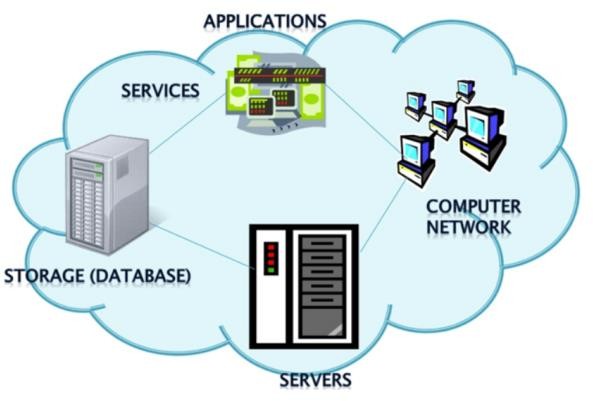


Fig. 1.1 Applications of Cloud Computing

##### Introduction to Amazon Web Services (AWS):

**WHAT IS AWS?**

AWS is designed to allow application providers, ISVs, and vendors to quickly and securely host your applications – whether an existing application or a new SaaS-based application. You can use the AWS Management Console or well-documented web services APIs to access AWS's application hosting platform.

**AWS ORGANIZATIONS:** AWS Organizations is a free account management service that enables you to consolidate multiple AWS accounts into an organization that you create and centrally manage. AWS Organizations include consolidated billing and account management capabilities that help you to better meet the budgetary, security, and compliance needs of your business.

##### BENEFITS:

* + - Policy-based account management
    - Application programming interfaces (APIs) that automate account management Consolidated billing

##### HOW TO ACCESS AWS ORGANIZATIONS?

* + - AWS Management Console
    - AWS Command Line Interface (AWS CLI) tools
    - Software development kits (SDKs)
    - HTTPS Query application programming interfaces (API)

#### Moving to the AWS Cloud – The AWS Cloud Adoption Framework (AWS CAF):

* + - The AWS Cloud Adoption Framework (AWS CAF) provides AWS experience and best practices to help you digitally transform and accelerate your business outcomes
    - The AWS CAF was created to help organizations develop efficient and effective plans for their cloud adoption journey.
    - These capabilities provide best practice guidance that helps you improve your cloud readiness.
    - The AWS CAF organizes guidance into six areas of focus, called perspectives.

1. Business
2. People
3. Governan

### MODULE 2

##### CLOUD ECONOMICS AND BILLING

* 1. **Fundamentals of pricing:**

There are three fundamental drivers of cost with AWS:

-compute

-storage

-outbound data transfer

* The best way to estimate costs is to examine the fundamental characteristics for each AWS Service.
* The service pricing strategy gives you the flexibility to choose the services that you need for each project and to pay only for what you use.

There are several free AWS services, including:

* Amazon VPC
* Elastic Beanstalk
* AWS CloudFormation
* IAM

##### Total cost of ownership:

AWS total cost of ownership (TCO) varies from organization to organization, as many have differing compliance and security requirements, customer expectations for service and availability, and DevOps goals.

Some of the costs that are associated with data center management shown in fig1.2

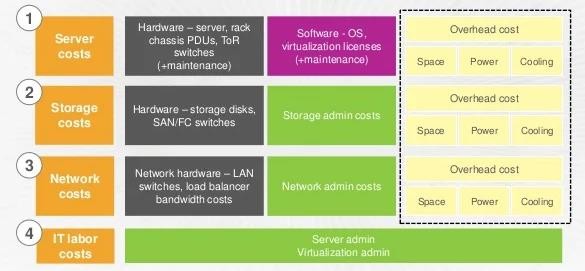


Fig 1.2 Total Cost of Ownership

##### AWS Organizations:

* + - AWS Organizations is a free account management service that enables you to consolidate multiple AWS accounts into an organization that you create and centrally manage.
    - AWS accounts are natural boundaries for permissions, security, costs, and workloads.
    - AWS Organizations provides many features for managing multi-account environments, including:

- Simplify account creation by programmatically creating new accounts using the AWS Command Line Interface (CLI), SDKs, or APIs.

-Group accounts into organizational units (OUs), or groups of accounts that serve a single application or service.

-Apply tag polices to classify or track resources in your organization and provide attribute-based access control for users or applications.

AWS Organizations is available in two feature sets:

* + - Consolidated billing.
    - All features.

##### AWS Billing and Cost Management:

The AWS Billing Console collects information from Cost and Usage Reports and presents it for the user to pay.

The AWS Billing Console lets you access all previous payments, access AWS credits, and perform consolidated bills.

AWS Cost Management is one of the five pillars of the AWS Well-Architected Framework. By using AWS inbuilt tools, people can use previous data to forecast and optimize future spending.

##### Technical Support:

* Provide unique combination of tools and expertise
  + AWS Support
  + AWS Support Plans
* Support is provided for:
  + Experimenting with AWS
  + Production use of AWS
  + Business-critical use of AWS
* Proactive guidance:
  + Technical Account Manager (TAM)
* Best practices:
  + AWS Trusted Advisor

### MODULE- 3

##### AWS GLOBAL INFRASTRUCTURE OVERVIEW

* 1. **AWS Global Infrastructure:**

The AWS Global Infrastructure consists of Regions and Availability Zones.

1.Your choice of a Region is typically based on compliance requirements or to reduce latency. 2.Each Availability Zone is physically separate from other Availability Zones and has redundant power, networking, and connectivity.

3.Edge locations, and regional edge caches improve performance by caching content closer to users.

The global infrastructure is shown in below figure 1.3

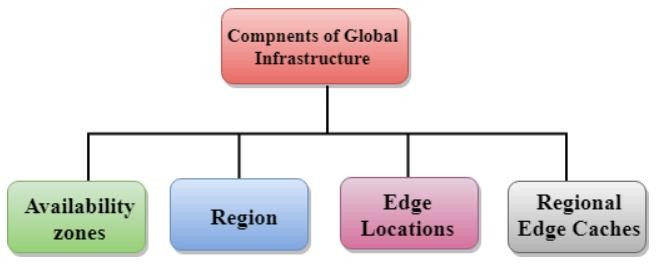


Fig 1.3 Components of Global Infrastructure

##### AWS services and service category overview:

AWS categories of services shown in below fig 1.4

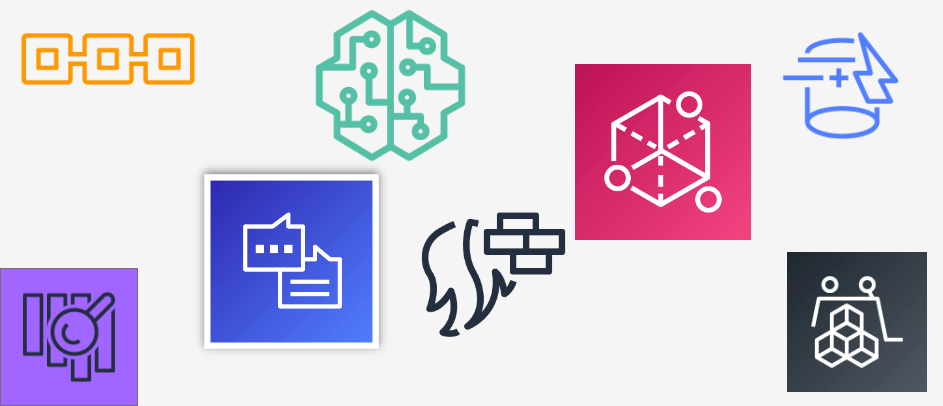


Fig 1.4 AWS Services

* + - Compute
    - Containers
    - Storage
    - Database

**AWS CLOUD SECURITY**

### MODULE-4

##### AWS shared responsibility model:

The AWS Shared Responsibility Model is a collection of security practices that is divided between the customer and AWS such that they can stress less and take equal part in the cloud security and compliance.

Some key takeaways from this section of the module include:

* + - AWS and the customer share security responsibilities–
      * AWS is responsible for security of the cloud
      * Customer is responsible for security in the cloud

##### AWS Identity and Access Management (IAM):

* + - IAM policies are constructed with JavaScript Object Notation (JSON) and define permissions.
      * IAM policies can be attached to any IAM entity.
      * Entities are IAM users, IAM groups, and IAM roles.
    - An IAM user provides a way for a person, application, or service to authenticate to AWS.

Fig 1.5 AWS identity and access management

##### Securing a new AWS account:

The following are the main steps to secure the AWS account

* Secure logins with multi-factor authentication (MFA).
* Delete account root user access keys.
* Use groups to assign permissions to IAM users.
* Configure a strong password policy.
* Delegate using roles instead of sharing credentials.
* Monitor account activity using AWS CloudTrail.

##### Securing accounts:

AWS offers many tools to help secure your account

* Safeguard your passwords and access keys
* Limit AWS account root user access to your resources
* Audit IAM users and their policies frequently
* Use AWS Git projects to scan for evidence of unauthorized use
* Monitor your account and its resources

##### Securing data on AWS:

AWS supports encryption of data. It provides two approaches:

* Client-side encryption
* Server-side encryption

In Client-side encryption, the data is encrypted outside of the AWS Cloud and then sent to storage. It is stored on an AWS storage in encrypted form.

In Server-side encryption, AWS takes care of the encryption of the data in its storage. The encryption process is transparent for the client, who writes or reads this data.

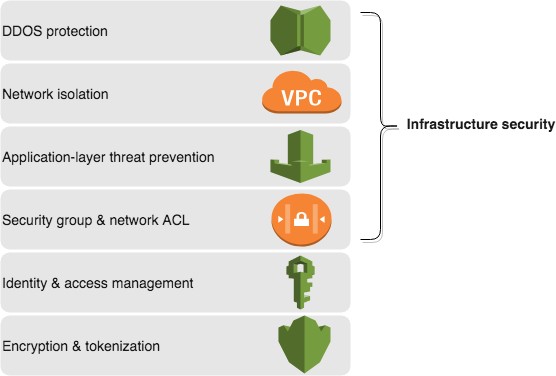


Fig 1.6 Security Infrastructure

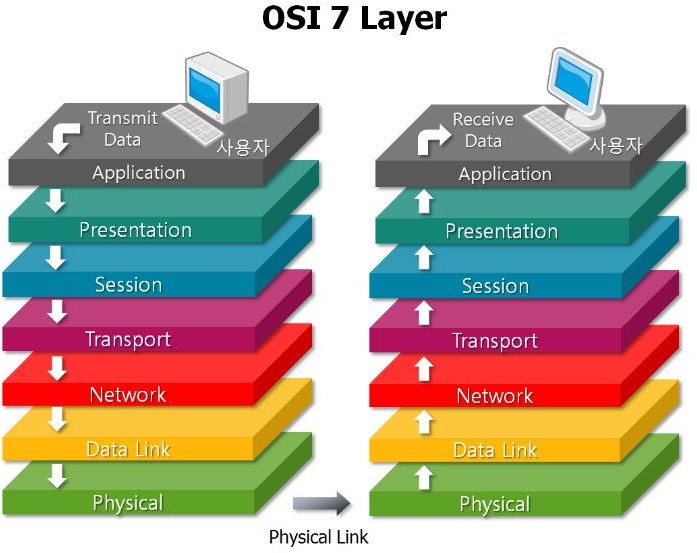
##### Working to ensure compliance:

* + - AWS Compliance empowers customers to understand the robust controls in place at AWS to maintain security and data protection in the AWS Cloud
    - AWS security compliance programs provide information about the policies, processes, and controls that are established and operated by AWS.
    - AWS Config is used to assess, audit, and evaluate the configurations of AWS resource

**MODULE -5**

##### NETWORKING AND CONTENT DELIVERY

* 1. **Networking basics:**
     + A computer network is two or more client machines that are connected together to share resources
     + Each client machine in a network has a unique Internet Protocol (IP) address that identifies it. An IP address is a numerical label in decimal format. Machines convert that decimal number to a binary format.
     + A 32-bit IP address is called an IPv4 address. IPv6 addresses, which are 128 bits, are also available. IPv6 addresses can accommodate more user devices.
     + A common method to describe networks is **Classless Inter-Domain Routing (CIDR).**
     + The **Open Systems Interconnection (OSI)** model is a conceptual model that is used to explain how data travels over a network.



##### Amazon VPC:

Fig 1.7 OSI layer

Amazon Virtual Private Cloud (Amazon VPC) enables you to launch AWS resources into a virtual network that you've defined

* A VPC is a logically isolated section of the AWS Cloud.
* A VPC belongs to one Region and requires a CIDR block.
* A VPC is subdivided into subnets.
* A subnet belongs to one Availability Zone and requires a CIDR block.
* Route tables control traffic for a subnet.

##### VPC networking:

* + - Amazon Virtual Private Cloud (Amazon VPC) gives you full control over your virtual networking environment, including resource placement, connectivity, and security.
    - There are several VPC networking options, which include:
      * Internet gateway: Connects your VPC to the internet
      * NAT gateway: Enables instances in a private subnet to connect to the internet
      * VPC endpoint: Connects your VPC to supported AWS services
      * VPC peering: Connects your VPC to other VPCs
      * VPC sharing: Allows multiple AWS accounts to create their application resources into shared
      * AWS Site-to-Site VPN: Connects your VPC to remote networks
      * AWS Direct Connect: Connects your VPC to a remote network by using a dedicated network connection
      * AWS Transit Gateway: A hub-and-spoke connection alternative to VPC peering
    - You can use the VPC Wizard to implement your design.

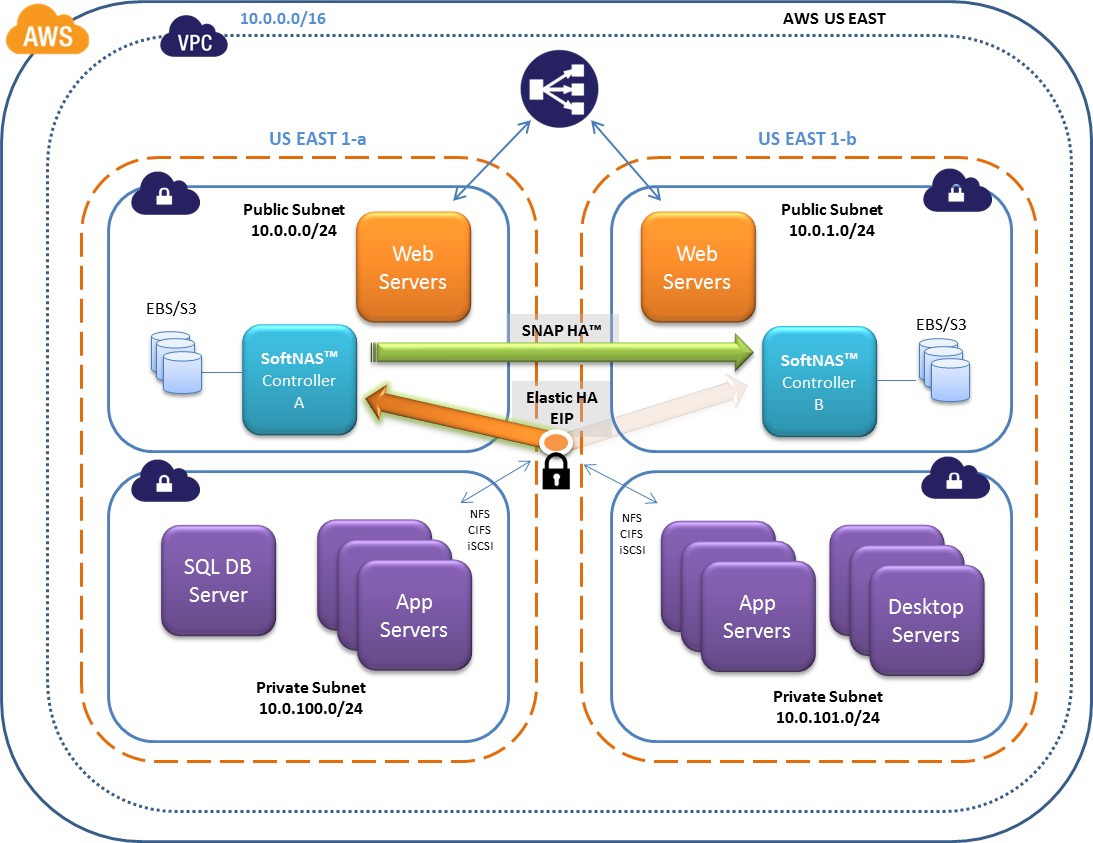


Fig 1.8 VPC

##### VPC security:

* Build security into your VPC architecture.
* Security groups and network ACLs are firewall options that you can use to secure your VPC.

##### Amazon Route S3:

* Amazon Route 53 is a highly available and scalable cloud DNS web service that translates domain names into numeric IP addresses.
* Amazon Route 53 supports several types of routing policies.
* Multi-Region deployment improves your application’s performance for a global audience

##### Amazon CloudFront:

* + - Amazon CloudFront is a fast content delivery network (CDN) service that securely delivers data, applications, and APIs to customers globally with low latency, high transfer speed
    - Amazon CloudFront offers many benefits, including:
      * Fast and global
      * Security at the edge
      * Highly programmable

**COMPUTE**

## MODULE -6

##### Compute services overview:

* + - Amazon Web Services (AWS) offers many compute services like Amazon EC2, Amazon Elastic Container Registry (Amazon ECR), Amazon Elastic Container Service (Amazon ECS), AWS Elastic Beanstalk, AWS Lamba, Amazon Elastic Kubernetes Services (Amazon EKS), Amazon Faregate

##### Amazon EC2:

* EC2 stands for Amazon Elastic Compute Cloud.
* Amazon EC2 is a web service that provides resizable compute capacity in the cloud.
* You can scale the compute capacity up and down as per the computing requirement changes.
* Amazon EC2 provides the developers with the tools to build resilient applications that isolate themselves from some common scenarios.

##### Amazon EC2 cost optimization:

Amazon EC2 pricing models include On-Demand Instances, Reserved Instances, Spot Instances, Dedicated Instances, and Dedicated Hosts.

The four pillars of cost optimization are–

* Right size
* Increase elasticity
* Optimal pricing model
* Optimize storage choices

##### Container services:

* Containers can hold everything that an application needs to run.
* Docker is a software platform that packages software into containers.
* A single application can span multiple containers.
* Amazon Elastic Container Service (Amazon ECS) orchestrates the running of Docker containers.
* Kubernetes is open-source software for container orchestration.

##### Introduction to AWS Lambda:

* AWS Lambda is a serverless compute service that lets you run your code without worrying about provisioning or managing any server
* You can call Lambda is FaaS by AWS.
* AWS Lambda is secure. It uses AWS IAM to define all the roles and security policies.
* It offers fault tolerance for both services running the code and the function

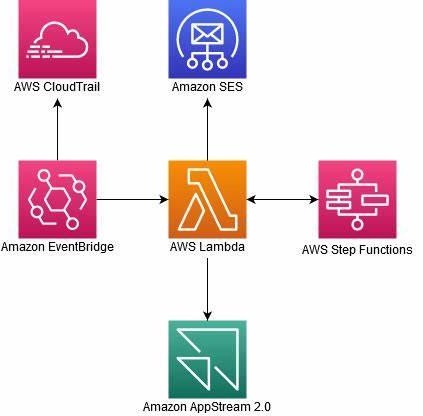


Fig 1.9 AWS LAMBDA

##### Activity: Introduction to AWS Elastic Beanstalk:

AWS Elastic Beanstalk is an AWS managed service for web applications.

Elastic beanstalk is a pre-configured EC2 server that can directly take up your application code and environment configurations and use it to automatically provision and deploy the required resources within AWS to run the web application.

**STORAGE**

## MODULE -7

##### Amazon Elastic Block Store (Amazon EBS):

* + - Amazon EBS provides block-level storage volumes for use with Amazon EC2 instances. They are analogous to virtual disks in the cloud. Amazon EBS provides three volume types: General Purpose SSD, Provisioned IOPS SSD, and magnetic.
    - The three volume types differ in performance characteristics and cost, so you can choose the right storage performance and price for the needs of your applications.
    - Additional benefits include replication in the same Availability Zone, easy and transparent encryption, elastic volumes, and backup by using snapshots.

##### Amazon Simple Storage Service (Amazon S3):

* + - Amazon S3 is a fully managed cloud storage service.
    - You can store a virtually unlimited number of objects.
    - You pay for only what you use.
    - You can access Amazon S3 at anytime from anywhere through a URL.

##### Amazon Elastic File System (Amazon EFS):

* + - Amazon EFS is a cloud-based file storage service for applications and workloads that run in the Amazon Web Services public cloud
    - Amazon EFS provides file storage over a network.
    - Fully managed service that eliminates storage administration tasks.
    - Accessible from the console, an API, or the CLI.
    - Scales up or down as files are added or removed and you pay for what you use.

##### Amazon S3 Glacier:

* + - Amazon S3 Glacier is a data archiving service that is designed for security, durability, and an extremely low cost.
    - Amazon S3 Glacier pricing is based on region.
    - Its extremely low-cost design works well for long-term archiving.
    - The service is designed to provide 11 9s of durability for objects.

## MODULE -8

**DATABASES**

##### Amazon Relational Database Service:

* + - With Amazon RDS, you can set up, operate, and scale relational databases in the cloud.
    - Features –
      * Managed service
      * Scalable
      * Automated redundancy and backup are available
      * Supported database engines:

- Amazon Aurora, PostgreSQL, MySQL, MariaDB, Oracle, Microsoft

##### Amazon DynamoDB:

Amazon DynamoDB:

* + - Runs exclusively on SSDs.
    - Supports document and key-value store models.
    - Replicates your tables automatically across your choice of AWS Regions

##### Amazon Redshift:

Redshift is a fast and powerful, fully managed, petabyte-scale data warehouse service in the cloud. Redshift consists of two types of nodes:

1. Single node
2. Multi-node

Single node: A single node stores up to 160 GB.

Multi-node: multi-node is a node that consists of more than one node. It is of two types:

-Leader Node

- Compute Node

##### Amazon Aurora:

* + - The Amazon Aurora is a relational database service offered from amazon cloud.
    - This is one of the widely used services for the data storage, for low latency and value- based data storage and processing
    - Features of Amazon Aurora
* Availability and Durability
* Performance and Scalability
* Fully Managed
* Compatible with MySql and PostgreSQL



Fig 1.10 Features of Amazon Auror

**CLOUD ARCHITECTURE**

**MODULE -9**

##### AWS Well-Architected Framework:

* + - The AWS Well-Architected Framework provides a consistent approach to evaluate cloud architectures and guidance to help implement designs.
    - AWS Well-Architected helps cloud architects build secure, high-performing, and efficient infrastructure for a variety of applications and workloads.
    - The AWS Well-Architected Framework is organized into six pillars.
    - Each pillar includes its own set of design principles and best practices.

##### Reliability and availability:

Some key takeaways from this section of the module include:

* + - Reliability is a measure of your system’s ability to provide functionality when desired by the user. It can be measured in terms of MTBF.
    - Availability is the percentage of time that a system is operating normally or correctly performing the operations expected of it.
    - Three factors that influence the availability of your applications are fault tolerance, scalability, and recoverability.

##### AWS Trusted Advisor:

AWS Trusted Advisor is an online tool that provides real time guidance to help you provision your resources following AWS best practices.

* Cost Optimization: AWS Trusted Advisor looks at your resource use and makes recommendations to help you optimize cost by eliminating unused and idle resources.
* Performance: Improve the performance of your service by checking your service limits.
* Security: Improve the security of your application by closing gaps, enabling various AWS security features.
* Fault Tolerance: Increase the availability and redundancy of your AWS application by taking advantage of automatic scaling, health check

### MODULE 10

##### AUTO SCALING AND MONITORING

* 1. **Elastic Load Balancing:**
     + Distributes incoming application or network traffic across multiple targets in a single Availability Zone or across multiple Availability Zones.
     + Scales your load balancer as traffic to your application changes over time. Types of load balancers:

1. Application load balancer
2. Classic load balancer
3. Network load balancer

##### Amazon CloudWatch:

* Amazon CloudWatch helps you monitor your AWS resources and the applications that you run on AWS in real time.
* CloudWatch enables you to –
* Collect and track standard and custom metrics.
* Set alarms
* Define rules that match changes in your AWS

##### Amazon EC2 Auto Scaling:

* + - AWS Auto Scaling monitors your applications and automatically adjusts capacity to maintain steady, predictable performance at the lowest possible cost. Using AWS Auto Scaling, it’s easy to setup application scaling for multiple resources across multiple services in minutes.
    - AWS Auto Scaling is a separate service that monitors your applications, and it automatically adjusts capacity for the following resources:

**Amazon EC2 instances and Spot Fleets Amazon ECS tasks

**Amazon DynamoDB tables and indexes Amazon Aurora Replica

# COURSE 2:

# AWS DATA ANALYTICS

# BIG DATA ANALYTICS

## BIG DATA

Analyzing large datasets requires significant compute capacity that can vary in size based on the amount of input data, the type of analysis, and the time required to complete the analysis. This characteristic of big data workloads is ideally suited to the pay-as-you-go cloud computing model, where resources can easily scale up and down based on demand. As requirements change, you can easily resize your environment (horizontally or vertically) on Amazon Web Services (AWS) to meet your needs, without having to wait for additional hardware or being required to overinvest to provision enough capacity. Also, because the services are managed by AWS, they are highly available without incurring additional costs of operations.

## BIG DATA PIPELINE

The figure above illustrates the big data pipeline. The pipeline consists of the following phases:

* 1. Data is collected (or ingested) by an appropriate tool.
  2. The data is persisted to storage.
  3. The data is processed or analyzed. The data processing or analysis solution takes the data from storage, performs operations, and then stores the data again.
  4. Analysis can be repeated to get further answers from the data.
  5. Data can be visualized with business intelligence (BI) tools to provide useful answers to business users.
  6. Insights are the output of the big data pipeline. The business owners use these insights to make critical decisions.

Some big data solutions do not require visualizations. For example, machine learning and predictive analytics create models, which are used to answer questions or provide estimates based on the data that is analyzed.The tools you deploy in your pipeline determine your time-to-answer, which is the latency between the time when your data is created and when you can get insight from it.The best way to architect a big data solution that accounts for latency is to determine how to balance throughput with cost. A higher throughput (and the subsequent reduced latency) usually results in higher costs.

## BIG DATA TOOLS

There are many open source and proprietary big data tools that are available and that are currently used in the industry. AWS has managed-service versions of many open-standard tools. All these services can be seamlessly integrated with other services of AWS to provide users with a complete end-to-end solution for various big data analytics problems.

The following table provides a mapping of popular tools to AWS services.

|  |  |  |
| --- | --- | --- |
| **Area** | **Open Source** | **AWS Services** |
| **Batch Ingest** | Apache Flume StreamSets Apache Sqoop | AWS Glue  AWS Data Pipeline  AWS Internet of Things (IoT) services |
| **Stream Ingest** | Apache Flume StreamSets | Amazon Kinesis Data Firehose |
| **Persistent Storage** | Hadoop Distributed File System (HDFS)  Relational database management system (RDBMS) | Amazon Simple Storage Service (Amazon S3)  Amazon EMR  Amazon Relational Database Service (Amazon RDS) |
| **Transient Storage** | Apache Kafka | Amazon Kinesis |
| **Batch Processing** | Apache Hive Apache Flink Apache Spark MapReduce PostgreSQL | Amazon Redshift AWS Glue  AWS Data Pipeline Amazon RDS |
| **Stream Processing** | Apache Flink | Amazon Kinesis Data Analytics |

|  |  |  |
| --- | --- | --- |
|  | Apache Spark Apache Beam |  |
| **Clients or Data Applications** | Apache Superset (BI) | Amazon QuickSight |
| **Visualize** | SAS  Tableau TIBCO tools | Amazon QuickSight  Amazon Elasticsearch Service (Amazon ES) and Kibana AWS IoT Analytics  Amazon Kinesis Data Analytics Amazon Elastic Compute Cloud (Amazon EC2) Jupyter notebooks |

Table-1(Tools of AWS services)

## BIG DATA COLLECTION

There are many ways in which the data can be collected for analysis. The following list provides more information about AWS collection services covered in this course.

* [Amazon Elastic Compute Cloud (Amazon EC2) Links to an external site.](https://aws.amazon.com/ec2/): Amazon EC2 is a compute service that is good for hosting web applications. Agents can be installed on Amazon EC2 to send clickstream data, web server access logs, error logs, etc.
* [Amazon Kinesis Links to an external site.](https://aws.amazon.com/kinesis/)(Amazon Kinesis Data Streams, Amazon Kinesis Video Streams, Amazon Kinesis Data Firehose, and Amazon Kinesis Data Analytics): Amazon Kinesis is a set of AWS services for processing big data in real time. Kinesis can process hundreds of terabytes per hour from high volumes of streaming data that comes from sources such as operating logs, financial transactions, and social media feeds
* [Internet of Things (IoT) Links to an external site.](https://aws.amazon.com/iot/): AWS offers a suite of IoT services that provide device software, control, and data services. These services enable you to connect securely to IoT devices and transfer the data at any scale. They also further provide the ability to store the data in various storage options and perform analytics.

## BIG DATA STORAGE

There are many storage options available in AWS, and all AWS storage options are made for specific purposes. The following list describes some of the more commonly used storage services for big data solutions. This course uses Amazon Simple Storage Service (Amazon S3) extensively.

* [**Amazon Simple Storage Service (Amazon S3) Links to an external site.**](https://aws.amazon.com/s3/)**:** Amazon S3 is an object storage service that offers scalability, data availability, security, and performance. This means that customers of all sizes and industries can use it to store and protect any amount of data for a range of use cases, such as websites, mobile applications, backing up and restoring, archiving, enterprise applications, IoT devices, and big data analytics. Amazon S3 provides easy-to-use management features so you can organize your data and configure finely tuned access controls that are designed to help you meet your specific business, organizational, and compliance requirements. Amazon S3 is a good choice for staging data from various sources. It is also good for creating [data lakes Links to an external site.](https://aws.amazon.com/big-data/datalakes-and-analytics/).
* [**Amazon Relational Database Service (Amazon RDS) Links to an external site.**](https://aws.amazon.com/rds/): Amazon RDS enables you to set up and operate a relational database management system (RDBMS) in the AWS Cloud with minimal effort. It automates the backup and recovery process. It also provides scalable storage and multiple read replicas for better performance. Amazon RDS is suitable for structured data storage in AWS without any management overhead.
* [**Amazon DynamoDB Links to an external site.**](https://aws.amazon.com/dynamodb/)**:** Amazon DynamoDB is a key-value and document database that delivers single-digit millisecond performance at any scale. It's a fully managed, multi-Region, multi-master database with built-in security, backup and restore capabilities, and in-memory caching for internet-scale applications.

## BIG DATA INGESTION

This stage of the data processing pipeline has some overlap with the Collection stage. Data can be collected by or ingested into AWS services in various ways. The following two managed AWS services—which can be used for ingestion—are included in this course.

* [**AWS Glue Links to an external site.**](https://aws.amazon.com/glue/)**:** AWS Glue is a fully managed extract, transform, and load (ETL) service that makes it easy for customers to prepare and load their data for analytics. ETL jobs can be created with a few clicks in the AWS Management Console. AWS Glue can discover data and store the inferred schema in the AWS Glue Data Catalog, which can then be

available for ETL. AWS Glue can also act as a remote metadata store for various AWS services like Amazon Athena, AWS Data Pipeline, etc.

34

* [**AWS Data Pipeline Links to an external site.**](https://aws.amazon.com/datapipeline/)**:** Data Pipeline is a managed service that can be used to move data between various data sources in the AWS Cloud, like Amazon S3, Amazon RDS, DynamoDB, Amazon Redshift, and Amazon EMR. It can reduce the complexities of handling data pipelines, and reliably move data from source to destination in a cost-effective way.

## BIG DATA PROCESSING AND ANALYSIS

In this stage of the data pipeline, data is processed and readied for analysis, visualization, and other downstream systems. Various managed and scalable services that are available in this space make it easy to implement. Three of the important services are described below. Amazon Elasticsearch Service and Amazon Redshift are both included in this course.

* [**Amazon EMR Links to an external site.**](https://aws.amazon.com/emr/)**(Hadoop in AWS):** Amazon EMR provides quick processing of large datasets using open source tools like Apache Spark, Apache Hive, Apache HBase, and Presto. Amazon EMR uses the scalable compute capacity of Amazon EC2 and the storage of Amazon S3. Processed data can be pushed to Amazon Redshift, DynamoDB, and Amazon RDS for further storage; or to Jupyter notebooks for analysis.
* [**Amazon Redshift Links to an external site.**](https://aws.amazon.com/redshift/)**(data warehouse):** Amazon Redshift is a petabyte- scale managed data warehouse service. Amazon Redshift makes it simple and cost-effective to analyze your data by using standard SQL and other BI tools. Amazon Redshift offers analytics features like columnar data storage, data compression, and massive parallel processing.
* [**Amazon Elasticsearch Service (Amazon ES) Links to an external site.**](https://aws.amazon.com/elasticsearch-service/)**:** Amazon ES is an enterprise-grade search service that offers access to open-source Elasticsearch application programming interfaces (APIs). Amazon ES is a managed service that only needs to be configured, and is ready for use. It can be configured to be highly scalable and available to avoid downtimes. You can use Amazon ES with Logstash and Kibana to provide complete end-to-end solutions for log analysis and visualizations in real time.

## BIG DATA VISUALIZATION

Visualization is a critical component of any big data analytics solution. Visualization provides business decision makers with graphical representations of analysis, which can make it easier to see the implications of the data. Many visualization tools also create interactive graphics to allow for what-if scenario analysis..

* [**Amazon SageMaker Links to an external site.**](https://aws.amazon.com/sagemaker/)**:** Amazon SageMaker is a fully managed service for data scientists and developers. It covers the machine learning workflow to label and prepare your data, choose an algorithm, train a model, and tune and optimize the model for deployment. Amazon SageMaker models can also make predictions and take action.
* [**Jupyter notebooks Links to an external site.**](https://jupyter.org/)**:** You can host Jupyter notebooks with Amazon SageMaker. Jupyter notebooks are an open-source tool you can use to create narratives that combine text with visualizations. There are many Jupyter notebooks available for fraud detection, forecasting, image classification, etc. In this course, learners learn to use Jupyter notebooks to create narratives that describe their analysis.
* [**AWS IoT Analytics Links to an external site.**](https://aws.amazon.com/iot-analytics/)**:** You can use the AWS IoT Analytics service to run real-time analytics directly on data sent by IoT devices. You can also run queries to send filtered data to a permanent data store or to visualization tools. It is a fully managed and highly scalable service that can be natively integrated with other services such as Amazon SageMaker Jupyter notebooks, DynamoDB, or external tools like Salesforce.
* [**Amazon Kinesis Data Analytics Links to an external site.**](https://aws.amazon.com/kinesis/data-analytics/): You can use Amazon Kinesis Data Analytics to perform real-time analytics on data in Amazon Kinesis streams. It is a fully managed, highly available, and scalable service that can handle petabytes of data in a cost- effective, pay-as-you-go model. Structured query language (SQL) users can run queries on it, and developers can use Java libraries for sophisticated transformations and analysis.
* [**Amazon QuickSight Links to an external site.**](https://aws.amazon.com/quicksight/)**:** Amazon QuickSight is a data analysis and visualization service that you can use for a wide variety of data types and sources. You can use it to automatically discover data sources in your AWS solution, or you can pull data from other data sources in your organization. You can integrate it with other AWS services, like Amazon Redshift, Amazon S3, Amazon Kinesis, or AWS IoT Analytics to create visualizations quickly in a cost-effective way. It is easy to share these visualizations with important stakeholders, which can help their decision-making process.

**AMAZON-S3**

# LAB-1

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.

Amazon S3 is an object storage service that stores data as objects within buckets. An *object* is a file and any metadata that describes the file. A *bucket* is a container for objects.

To store your data in Amazon S3, you first create a bucket and specify a bucket name and AWS Region. Then, you upload your data to that bucket as objects in Amazon S3. Each object has a *key* (or *key name*), which is the unique identifier for the object within the bucket.

S3 provides features that you can configure to support your specific use case. For example, you can use S3 Versioning to keep multiple versions of an object in the same bucket, which allows you to restore objects that are accidentally deleted or overwritten.

Buckets and the objects in them are private and can be accessed only if you explicitly grant access permissions. You can use bucket policies, AWS Identity and Access Management (IAM) policies, access control lists (ACLs), and S3 Access Points to manage access.

# LAB-2

**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. Simply point to your data in Amazon S3, define the schema, and start querying using standard SQL. Most results are delivered within seconds. With Athena, there’s no need for complex ETL jobs to prepare your data for analysis. This makes it easy for anyone with SQL skills to quickly analyze large-scale datasets.

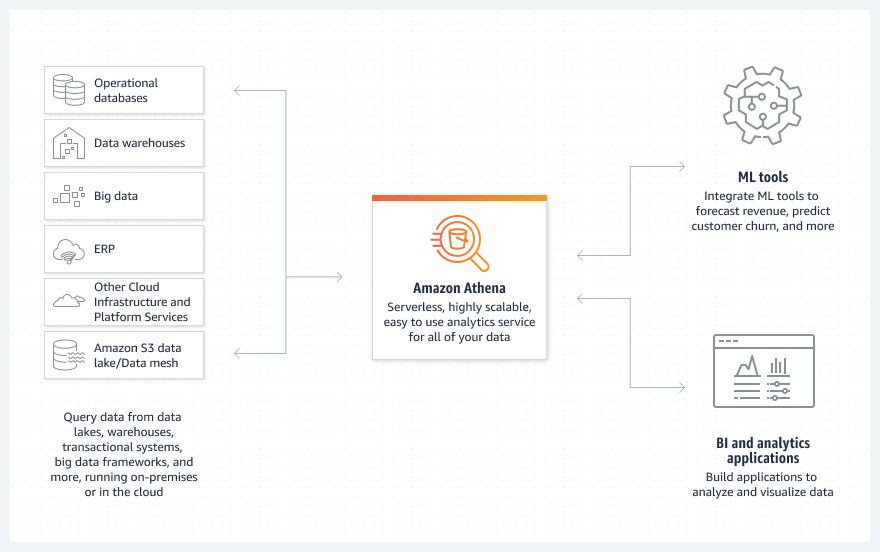
Athena is out-of-the-box integrated with [AWS Glue](https://aws.amazon.com/glue/) Data Catalog, allowing you to create a unified metadata repository across various services, crawl data sources to discover schemas and populate your Catalog with new and modified table and partition definitions, and maintain schema versioning.

Fig-2.1(Amazon Athena)

**AMAZON GLUE:**

# LAB-3

AWS Glue is a serverless data integration service that makes it easier to discover, prepare, move, and integrate data from multiple sources for analytics, machine learning (ML), and application development.

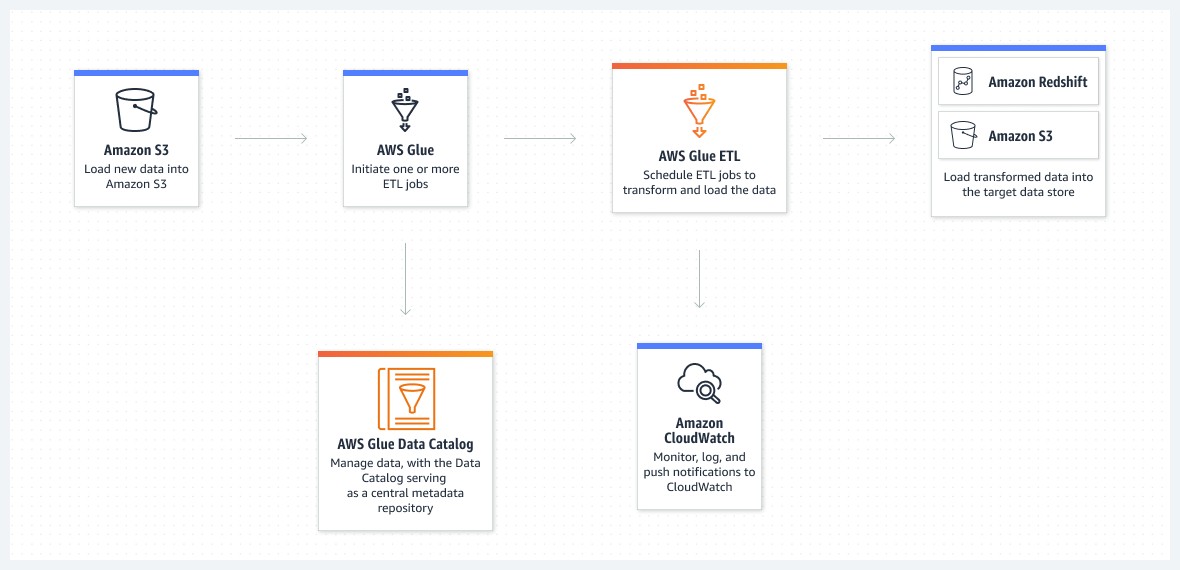
Preparing your data to obtain quality results is the first step in an analytics or ML project. AWS Glue is a serverless data integration service that makes data preparation simpler, faster, and cheaper. You can discover and connect to over 70 diverse data sources, manage your data in a centralized data catalog, and visually create, run, and monitor ETL pipelines to load data into your data lakes

Fig-3.1(Amazon Glue)

**AMAZON REDSHIFT:**

# LAB-4

Amazon Redshift uses SQL to analyze structured and semi-structured data across data warehouses, operational databases, and data lakes, using AWS-designed hardware and machine learning to deliver the best price performance at any scale.

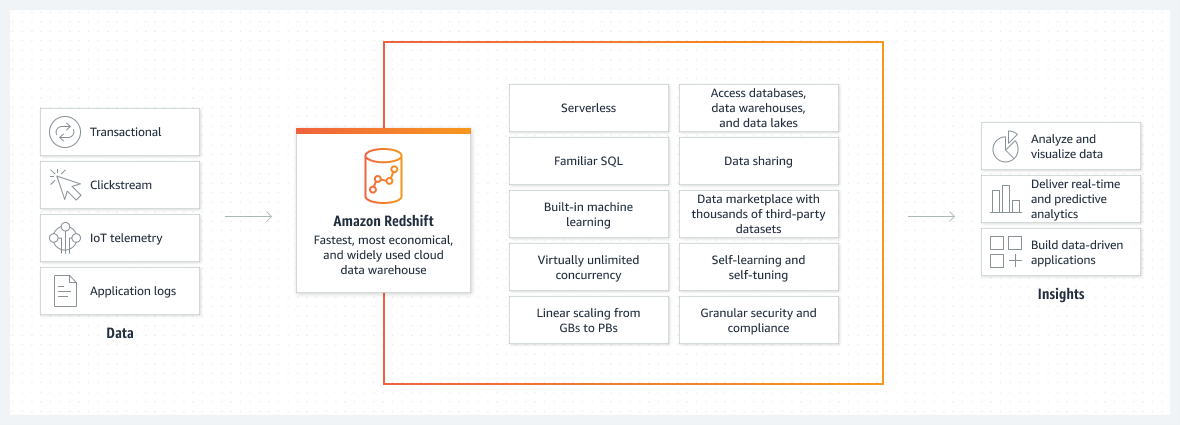


Fig-4.1(Amazon Redshift)

38

# LAB-5

39

**ANALYSE DATA WITH AMAZON SAGEMAKER:**

Amazon SageMaker is a fully managed machine learning service. With SageMaker, data scientists and developers can quickly and easily build and train machine learning models, and then directly deploy them into a production-ready hosted environment. It provides an integrated Jupyter authoring notebook instance for easy access to your data sources for exploration and analysis, so you don't have to manage servers. It also provides common machine learning algorithms that are optimized to run efficiently against extremely large data in a distributed environment. With native support for bring-your-own-algorithms and frameworks, SageMaker offers flexible distributed training options that adjust to your specific workflows. Deploy a model into a secure and scalable environment by launching it with a few clicks from SageMaker Studio or the SageMaker console. Training and hosting are billed by minutes of usage, with no minimum fees and no upfront commitments.

Creating jupyter notebook with amazon sagemaker

1. Open the notebook instance as follows:
   1. Sign in to the SageMaker console at <https://console.aws.amazon.com/sagemaker/>.
   2. On the **Notebook instances** page, open your notebook instance by choosing either **Open JupyterLab** for the JupyterLab interface or **Open Jupyter** for the classic Jupyter view.
2. Create a notebook as follows:
   1. If you opened the notebook in the JupyterLab view, on the **File** menu, choose **New**, and then choose **Notebook**. For **Select Kernel**, choose **conda\_python3**. This preinstalled environment includes the default Anaconda installation and Python 3.
   2. If you opened the notebook in the classic Jupyter view, on the **Files** tab, choose **New**, and then choose **conda\_python3**. This preinstalled environment includes the default Anaconda installation and Python 3.
3. Save the notebooks as follows:
   1. In the JupyterLab view, choose **File**, choose **Save Notebook As...**, and then rename the notebook.
   2. In the Jupyter classic view, choose **File**, choose **Save as...**, and then rename the notebook.

**LOAD DATA USING PIPELINE:**

# LAB-6

AWS Data Pipeline is a web service that you can use to automate the movement and transformation of data. With AWS Data Pipeline, you can define data-driven workflows, so that tasks can be dependent on the successful completion of previous tasks. You define the parameters of your data transformations and AWS Data Pipeline enforces the logic that you've set up.

The following components of AWS Data Pipeline work together to manage your data:

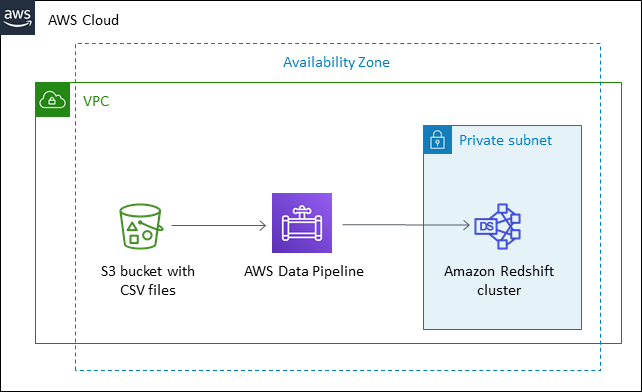
* A *pipeline definition* specifies the business logic of your data management. For more information, see [Pipeline Definition File Syntax](https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/dp-writing-pipeline-definition.html).
* A *pipeline* schedules and runs tasks by creating Amazon EC2 instances to perform the defined work activities. You upload your pipeline definition to the pipeline, and then activate the pipeline. You can edit the pipeline definition for a running pipeline and activate the pipeline again for it to take effect. You can deactivate the pipeline, modify a data source, and then activate the pipeline again. When you are finished with your pipeline, you can delete it.
* *Task Runner* polls for tasks and then performs those tasks. For example, Task Runner could copy log files to Amazon S3 and launch Amazon EMR clusters. Task Runner is installed and runs automatically on resources created by your pipeline definitions. You can write a custom task runner application, or you can use the Task Runner application that is provided by AWS Data Pipeline. For more information, see [Task Runners](https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/dp-how-remote-taskrunner-client.html).

Fig-6.1 (Loading Data)

**ANALIZE SREAMING DATA:**

# LAB-7

Amazon Kinesis Data Firehose is an extract, transform, and load (ETL) service that reliably captures, transforms, and delivers streaming data to data lakes, data stores, and analytics services

.Amazon Kinesis is a suite of services for processing streaming data. With Amazon Kinesis, you can ingest real-time data such as video, audio, website clickstreams, or application logs. You can process and analyze the data as it arrives, instead of capturing it all to storage before you begin analysis.

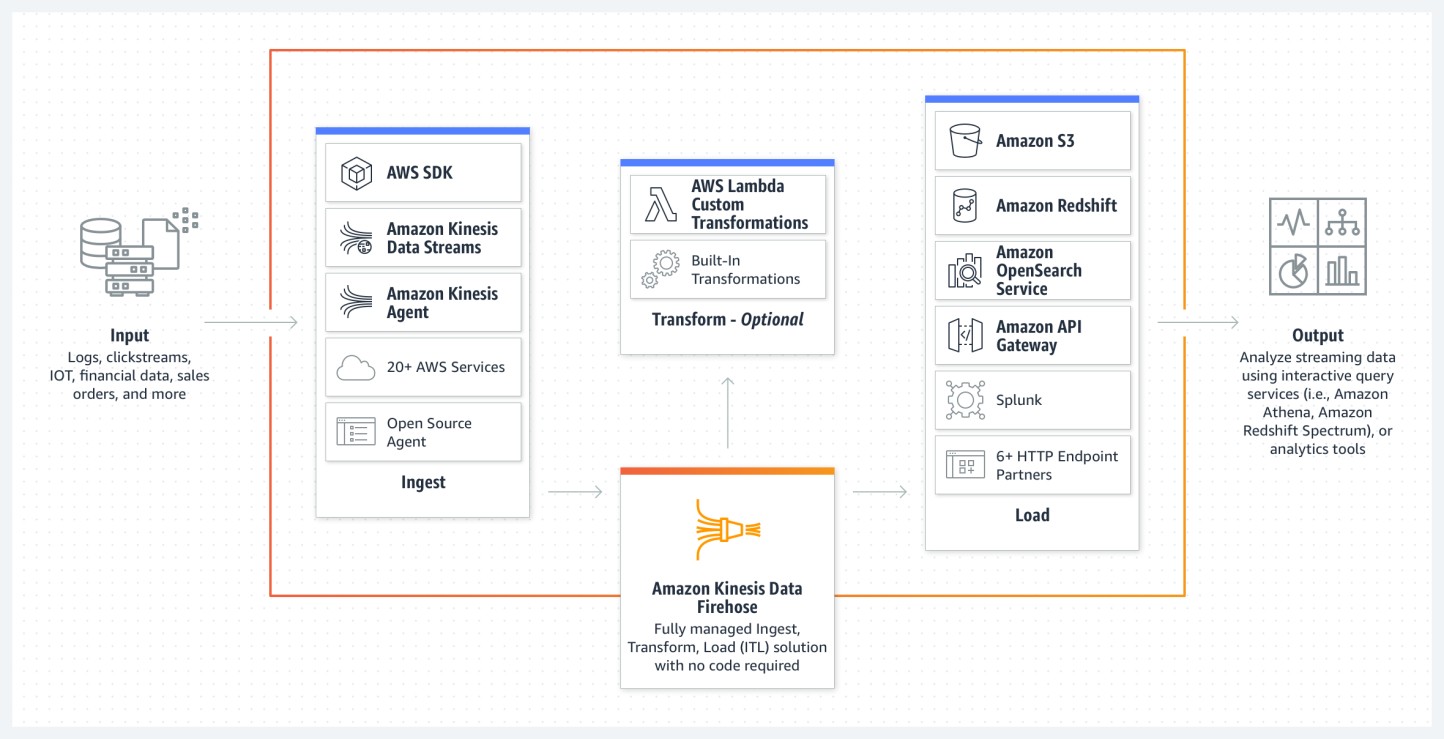


Fig-7.1 (Analyzing data)

# LAB-8

**ANALYSE IOT DATA WITH AWS IOT ANALYTICS:**

AWS IoT Analytics automates the steps required for analyzing IoT data. You can filter, transform, and enrich the data before storing it in a time-series data store. AWS IoT Core provides connectivity between IoT devices and AWS Services. IoT Core is fully-integrated with IoT Analytics.

IoT data is highly unstructured which makes it difficult to analyze with traditional analytics and business intelligence tools that are designed to process structured data. IoT data comes from devices that often record fairly noisy processes (such as temperature, motion, or sound). The data from these devices can frequently have significant gaps, corrupted messages, and false readings that must be cleaned up before analysis can occur. Also, IoT data is often only meaningful in the context of additional, third party data inputs. For example, to help farmers determine when to water their crops, vineyard irrigation systems often enrich moisture sensor data with rainfall data from the vineyard, allowing for more efficient water usage while maximizing harvest yield.

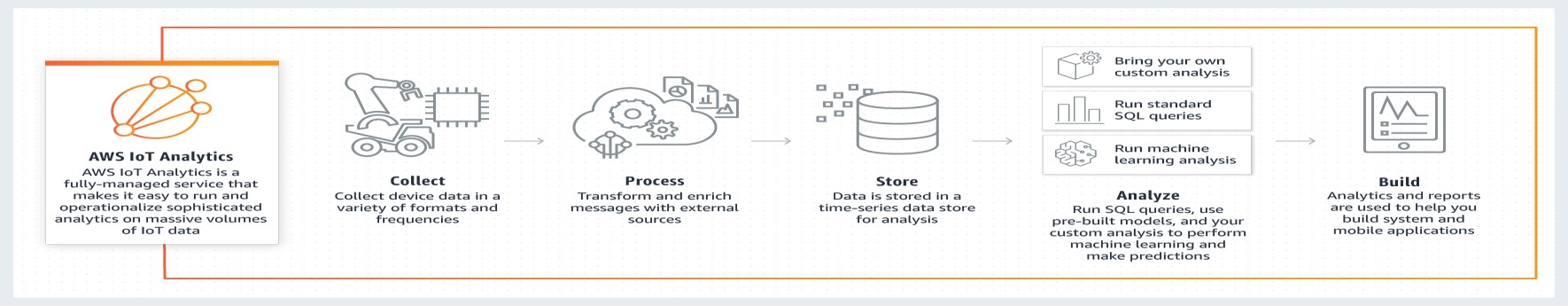


Fig-8.1 (IOT Data)

**CASE STUDY**



**Big Data Analytics Pipeline Using Modern Data Architecture**

**Problem Statement:**

For the past several years, the BMW Group has worked to stay at the forefront of the automotive industry’s digital transformation by using data and predictive analytics. In 2015 the BMW Group created a centralized, on-premises data lake that collects and combines anonymized data from sensors in vehicles, operational systems, and data warehouses to derive historical, real-time, and predictive insights. However, the company needed to more easily scale its data lake to support the growing demands of internal and external stakeholders. Because data wasn’t easily accessible—spread across myriad, siloed environments—the BMW Group’s innovation was slowed down by its own IT infrastructure and the long lead times required to support new initiatives. The BMW Group needed to develop a solution agile enough to both support the data needs of all the various internal business units and allow the company to move quickly to address the array of emerging use cases its customers demand.

**Solution Overview:**

BMW Group, facing challenges in data management and analysis, implemented a **Modern data architecture** using Amazon Web Services (AWS) to revolutionize their data-driven decision-making process. The solution consisted of several key steps and AWS services:

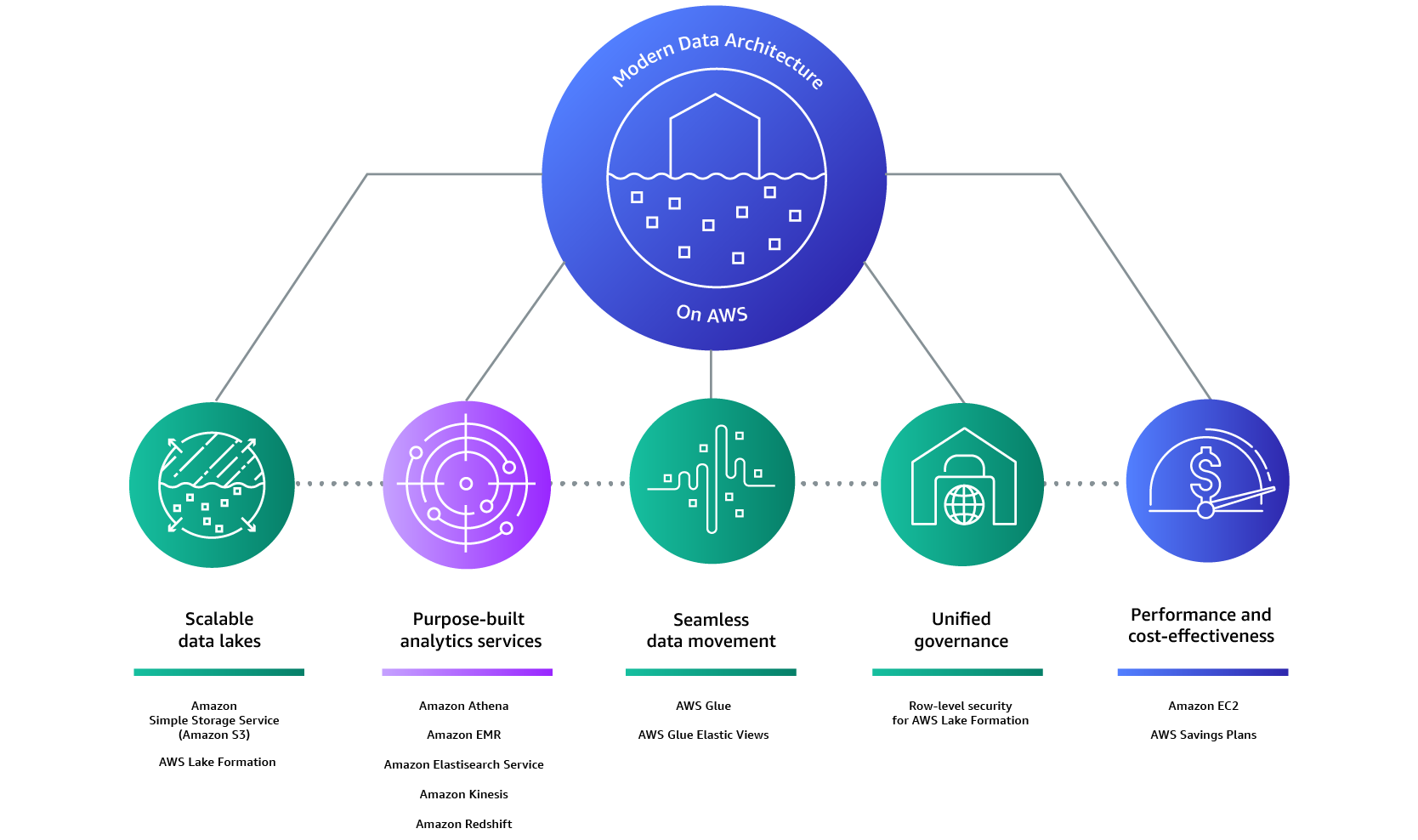
1. Data Lake Implementation with Amazon S3

2. Automated Data Crawling and Transformation with AWS Glue

3. ETL and Parquet Format Conversion using AWS Glue Studio

4. Real-time Data Querying with Amazon Athena

5. Enhanced Data Visualization via Amazon QuickSight



**Fig 1**: Modern Data Architecture on AWS

Modern data architecture pillars:

1.Scalable data lakes

2.Purpose-built analytics service

3.Unified data access

4.Unified governance

5.Performant and cost effective

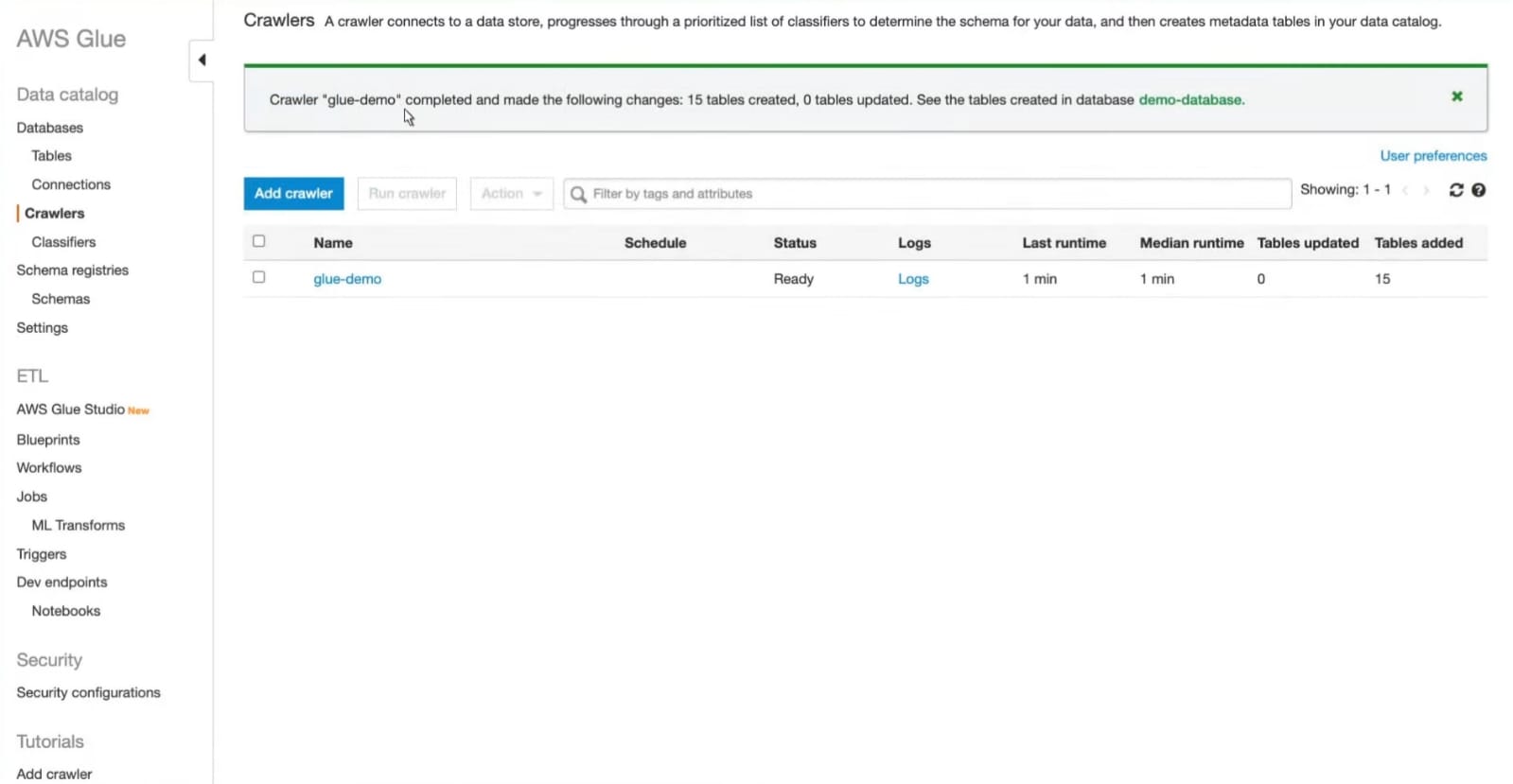
**Implementation steps:**

**1.Setup**

* The tools and services that will be utilized: Amazon S3, AWS Glue, Amazon Athena, and Amazon QuickSight.

**2.** **Add a Crawler**

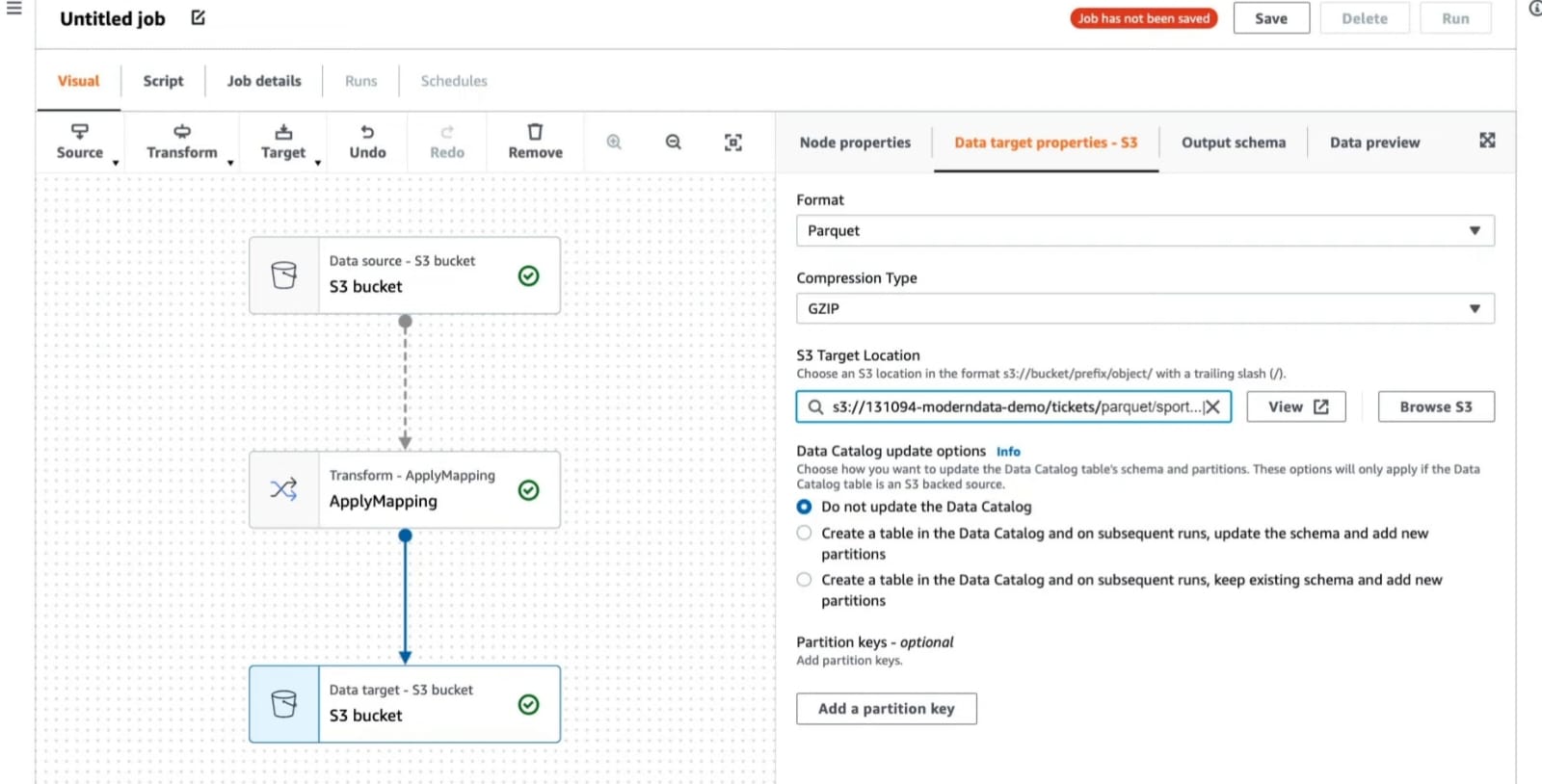
* Access AWS Glue to start populating the data lake.
* Add a crawler to extract data from an S3 bucket and load it into the data lake stored in Amazon S3.
* Name the crawler and choose default options to crawl all folders in the data store.
* Specify the path for the S3 bucket to be populated.
* Select an appropriate AWS IAM role for the crawler.
* Configure the crawler to run on demand.
* Create a new database to store the crawler's output.
* Review the crawler's specifications and complete the setup.
* Run the crawler to extract data from the S3 bucket and add tables to the database.



**Fig 2**: Successfully Crawled

**3. ETL with AWS Glue Studio**

* Access AWS Glue Studio to set up Extract, Transform, Load (ETL) jobs for data transformation.
* Select the database created by the crawler as the data source.
* Create separate ETL jobs for each table in the database.
* Adjust data types in the transform node as necessary.
* Specify the Parquet format in the data target node and select the same S3 bucket as the source.
* Name the job, choose its IAM role, and disable job bookmarks for complete dataset processing.
* Save and run the job.
* Monitor the job run status from the Monitoring tab and confirm its success.
* Visit the S3 bucket to verify the presence of new folders with Parquet files.



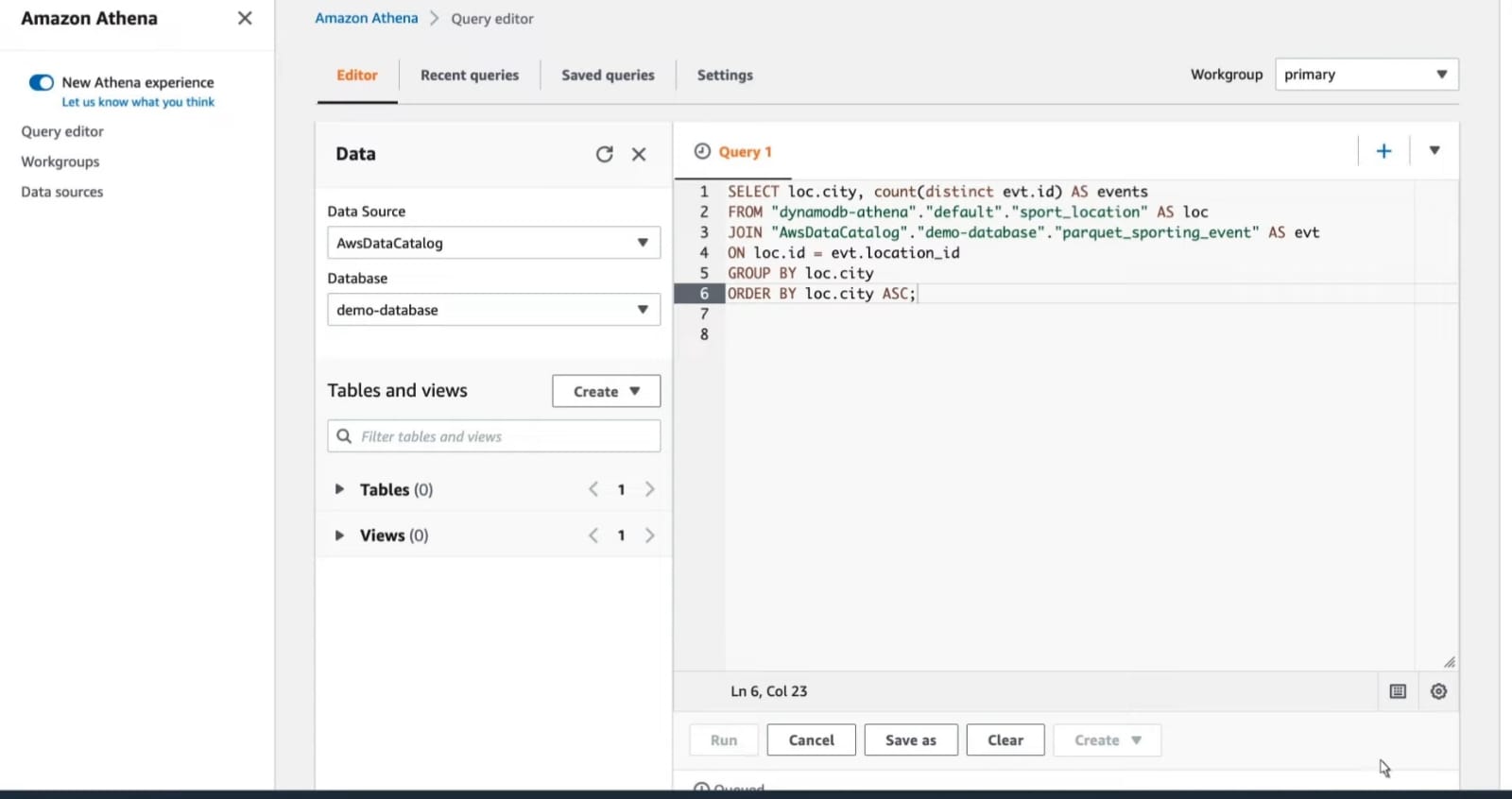
**Fig 3**: AWS Glue Studio

**4. Create a Crawler for Parquet Conversion**

* Create jobs behind the scenes to convert remaining tables into Parquet format.
* Build a new crawler to extract and save the Parquet files to the data lake.
* Name the crawler and choose similar specifications as the previous crawler, focusing on the path for Parquet files.
* Add a prefix to distinguish Parquet files.
* Complete the crawler setup and run it to extract and save Parquet files.

**5. Query Data Sources with Amazon Athena**

* Utilize Amazon Athena for interactive query services.
* Create a data source connection to Amazon DynamoDB.
* Name the data source and create a new Lambda function for the connection.
* Provide a spill bucket name for output data storage.
* Choose a name for the Athena catalog, which will be the Lambda function's name.
* Retain default settings and deploy the Lambda function.
* Check the function status from the AWS Lambda Deployments tab.
* Connect the data source in the Athena data source connection page.



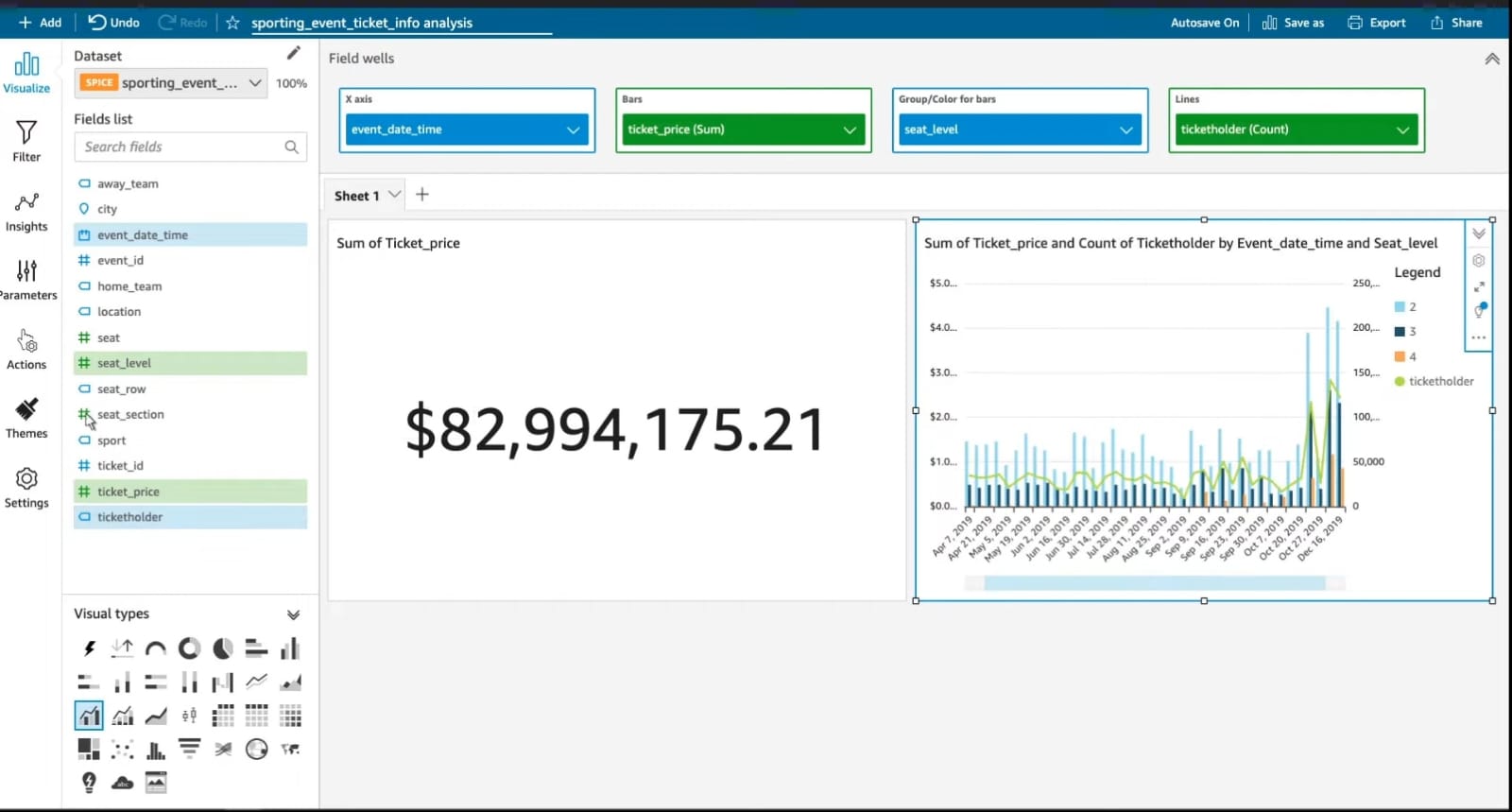
**Fig 4**: Amazon Athena

**6. Test Data Source Connection and Querying**

* Go to the query editor to test the data source connection.
* Execute queries that involve both DynamoDB and Parquet table data.
* Ensure accurate query results to confirm the data source connector's functionality.

**7.** **Visualize Data using Amazon QuickSight**

* Utilize Amazon QuickSight to build interactive BI dashboards and visualizations.
* Create a new analysis using Athena as the source for the dataset.
* Select the table with the data to be visualized.
* Import data to SPICE for optimized performance.
* Generate visualizations using different types such as currency sums and vertical bar graphs.
* Add dimensions like seat level data and change visual types as needed.
* Benefit from QuickSight Insights to generate plain language summaries of findings.



**Fig 5**: Amazon QuickSight

**AWS Tools used:**

* Amazon S3: Object storage for data lakes
* AWS Glue: Data catalog
* Amazon Athena: Interactive analytics
* Amazon QuickSight: Dashboards and visualizations

**Outcomes:**

The successful implementation of the big data analytics pipeline using modern data architecture has ushered in a new era of data-driven insights for the organization. By leveraging Amazon S3 as a data lake, data from various sources has been seamlessly integrated, eliminating silos and promoting data accessibility. The utilization of Amazon Athena enables real-time querying across multiple data stores, facilitating timely and informed decision-making. The integration of Amazon QuickSight for business intelligence visualization empowers stakeholders to effortlessly uncover valuable insights from complex data patterns. Overall, this project has streamlined data management, enhanced analytics capabilities, and provided a powerful toolset for generating meaningful insights, positioning the organization at the forefront of data-driven innovation.

**Conclusion:**

Through the adoption of a modern data architecture, BMW has successfully transformed its data management and analysis processes. By harnessing the capabilities of AWS, BMW established a robust data lake. This endeavor has empowered BMW to capitalize on data-driven insights, enabling smarter business decisions and solidifying its position as an industry leader.

# CONCLUSION

As more and more data is generated and collected, data analysis requires scalable, flexible, and high performing tools to provide insights in a timely fashion. However, organizations are facing a growing big data ecosystem where new tools emerge and become outdated very quickly. Therefore, it can be very difficult to keep pace and choose the right tools.

This whitepaper offers a first step to help you solve this challenge. With a broad set of managed services to collect, process, and analyze big data, AWS makes it easier to build, deploy, and scale big data applications. This enables you to focus on business problems instead of updating and managing these tools.

AWS provides many solutions to address your big data analytic requirements. Most big data architecture solutions use multiple AWS tools to build a complete solution. This approach helps meet stringent business requirements in the most cost-optimized, performant, and resilient way possible. The result is a flexible big data architecture that is able to scale along with your business.

# REFERENCE LINKS

Data Analytics on AWS <https://awsacademy.instructure.com/courses/25825/modules>

Amazon-S3 <https://docs.aws.amazon.com/AmazonS3/latest/userguide/Welcome.html>

Amazon Athena [https://aws.amazon.com/athena/?whats-new-cards.sort-](https://aws.amazon.com/athena/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc)

[by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc](https://aws.amazon.com/athena/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc)

Amazon Glue <https://aws.amazon.com/glue/>

Amazon Redshift <https://aws.amazon.com/redshift/>

Amazon Sagemaker <https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html>

Amazon IOT data Analysis <https://aws.amazon.com/iot-analytics/>

Case Study

<https://aws.amazon.com/solutions/case-studies/bmw-group-case-study/>