# P. E. S. College of Engineering, Mandya

## Department of Information Science and Engineering

An Internship Presentation on

# Data Engineering Virtual Internship

**Submitted by:**
**Rakesh S**
**4PS21IS045**
**Dept. of ISE,**
**PESCE, Mandya**

**Under the Guidance of**
**M R Suresh**
**Associate Professor,**
**Dept. of ISE,**
**PESCE, Mandya**

# Contents

- Introduction
- About the Course
- Key Learnings
- Hands-on Work
- Project
- Conclusion

# Introduction

- I completed a 10-week Data Engineering Virtual Internship from January to March 2025.
- The internship was organized by AICTE (All India Council for Technical Education) in collaboration with EduSkills and AWS Academy.
- This program was delivered through the National Internship Portal and was open to engineering students across India.
- The curriculum provided in-depth exposure to AWS Cloud and Data Engineering concepts, hands-on labs, and real-world projects.
- The internship enabled me to gain practical skills in cloud computing, data pipeline design, and AWS services.
- Certificate issued by AICTE, EduSkills, and AWS Academy.

# About the course

The internship combined two in-depth AWS Academy programs:

- **AWS Academy Cloud Foundations**
- **AWS Academy Data Engineering**

## Course Objectives:

- Provided a foundational understanding of cloud computing and AWS services.
- Covered essential data engineering concepts and practical skills for designing data pipelines on AWS.
- Prepared students for AWS certification exams and industry roles.

# Cloud Service Models

1. **Infrastructure as a Service (IaaS):**
   - Provides the basic building blocks for cloud IT: networking, virtual computers, and storage.
   - Offers the highest level of flexibility and management control over IT resources.
   - Most similar to traditional IT resources familiar to many IT departments and developers.
2. **Platform as a Service (PaaS):**
   - Reduces the need to manage underlying infrastructure (hardware, OS).
   - Focuses on deploying and managing applications, not servers or storage.
3. **Software as a Service (SaaS):**
   - Delivers a complete product managed by the provider, typically end-user applications.
   - Users only need to consider how to use the software, not how it is maintained or hosted (e.g., web-based email)

# Cloud Deployment Models

1. **Cloud:**
    - Applications are fully deployed and run in the cloud, either built natively or migrated from existing infrastructure.
    - Can use low-level infrastructure or higher-level managed services.
2. **Hybrid:**
    - Connects cloud-based resources with on-premises infrastructure.
    - Enables organizations to extend and integrate their IT resources.
3. **On-premises (Private Cloud):**
    - Resources are deployed in-house, often using virtualization.
    - Offers dedicated resources but lacks many cloud benefits; sometimes called private cloud

# Key Advantages of Cloud Computing

- Trade capital expense for variable expense
- Massive economies of scale
- Stop guessing capacity
- Increase speed and agility
- Stop spending on running and maintaining data centers
- Go global in minutes

# Web Services and AWS Overview

**Web Service:**

- Software accessible over the internet using standardized formats (XML, JSON) for API interactions.

**What is AWS?**

- A secure cloud platform offering a broad set of global cloud-based products (compute, storage, database, networking, etc.).
- Provides on-demand access, flexibility, and pay-as-you-go pricing.
- Services are modular and work together like building blocks

**Accessing AWS Services**

- AWS Management Console
- AWS Command Line Interface (CLI)
- AWS Software Development Kits (SDKs)

# AWS Global Infrastructure Overview

**AWS Regions**

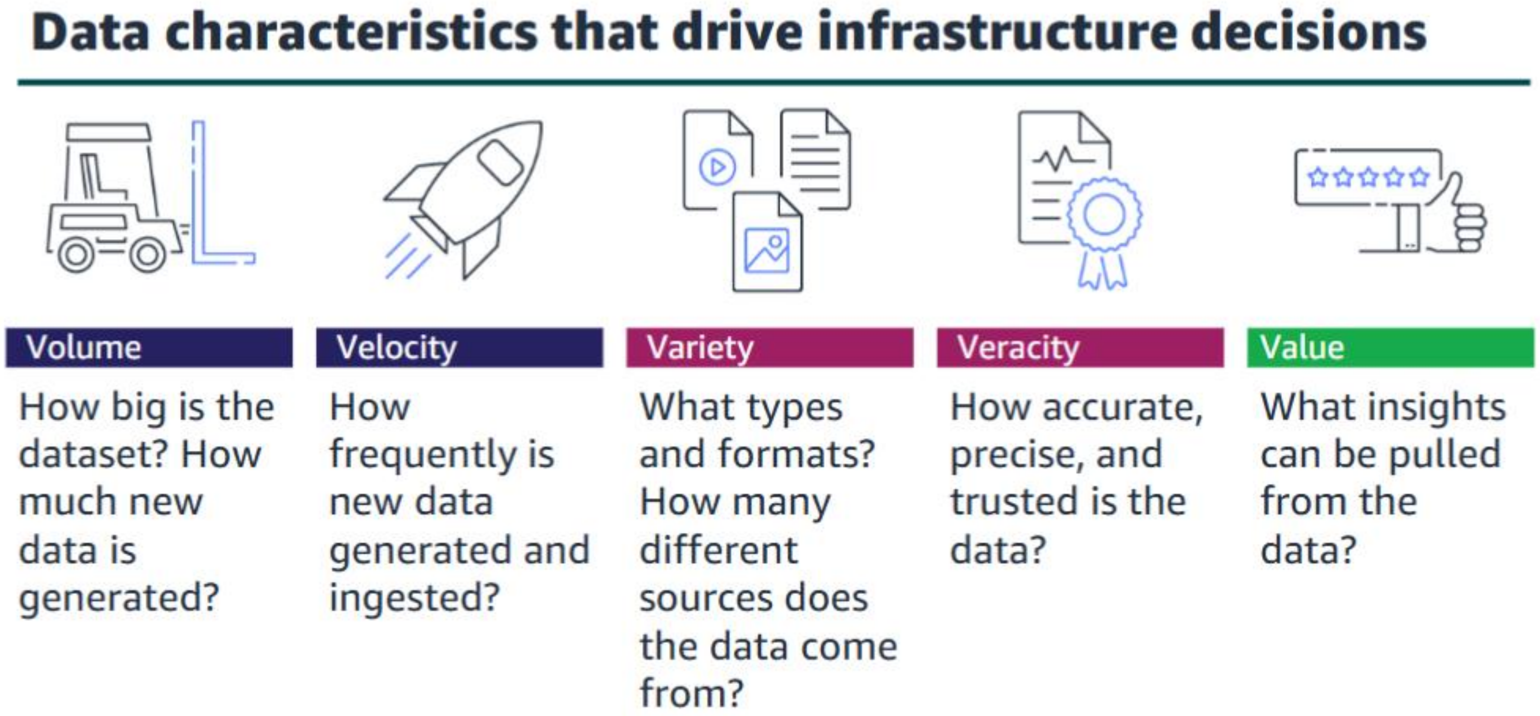
- An AWS Region is a distinct geographic area with multiple, isolated locations known as Availability Zones.
- AWS operates Regions worldwide, each designed to provide fault tolerance and stability by isolating them from one another.
- Data stored in one Region is not automatically replicated to another; cross-region replication must be configured by the user.
- Some Regions, such as AWS GovCloud (US) and China Regions, have special access or compliance requirements

**Availability Zones (AZs)**

- Each Region consists of two or more Availability Zones, which are separate physical locations with independent power, networking, and cooling.
- An Availability Zone may contain one or more data centers, and customers deploy resources at the AZ level, not the data center level.

# The Elements of Data

The five Vs of data, which map to the questions that a data engineer must answer to design a good data infrastructure.



## Data characteristics that drive infrastructure decisions

| Volume | Velocity | Variety | Veracity | Value |
|---|---|---|---|---|
| How big is the dataset? How much new data is generated? | How frequently is new data generated and ingested? | What types and formats? How many different sources does the data come from? | How accurate, precise, and trusted is the data? | What insights can be pulled from the data? |

# Ingesting and Preparing Data

**Data Pipeline Layers**

- Ingestion: Extract data from external sources and bring it into the pipeline.
- Temporary Storage: Hold data for initial processing or staging.
- Processing/Transformation: Convert data into usable formats, either before or after loading into permanent storage.

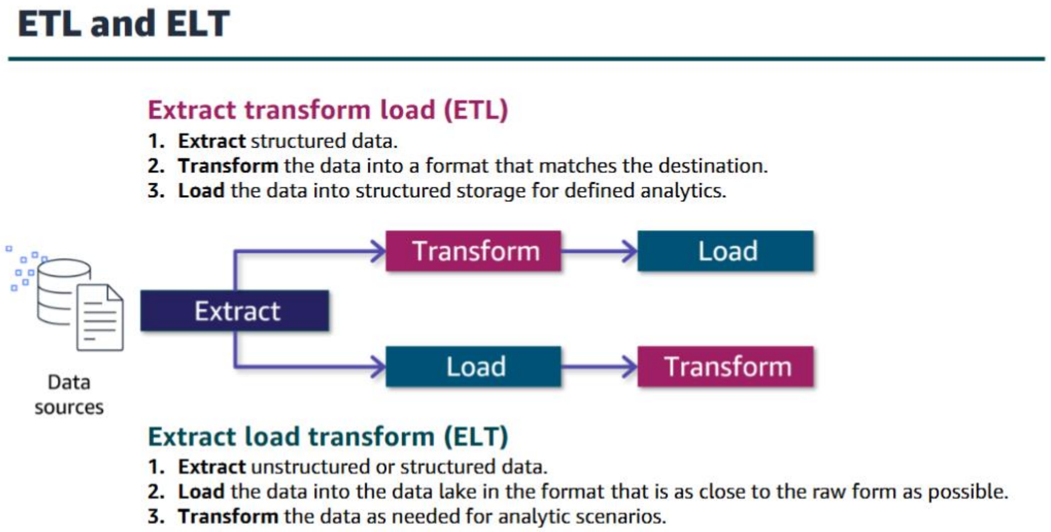
**Traditional ETL (Extract, Transform, Load)**

Best for structured data and traditional analytics.

Data is cleaned and formatted before storage, ready for immediate analysis.

**Modern ELT (Extract, Load, Transform)**

Ideal for handling large volumes of unstructured or semi-structured data. Enables flexible, on-demand transformations for diverse analytics needs.



**ETL and ELT**

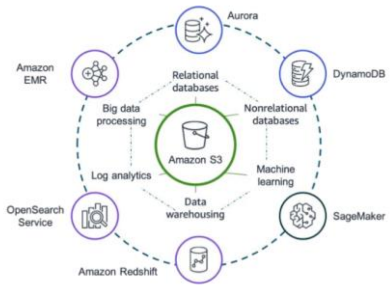
**Extract transform load (ETL)**

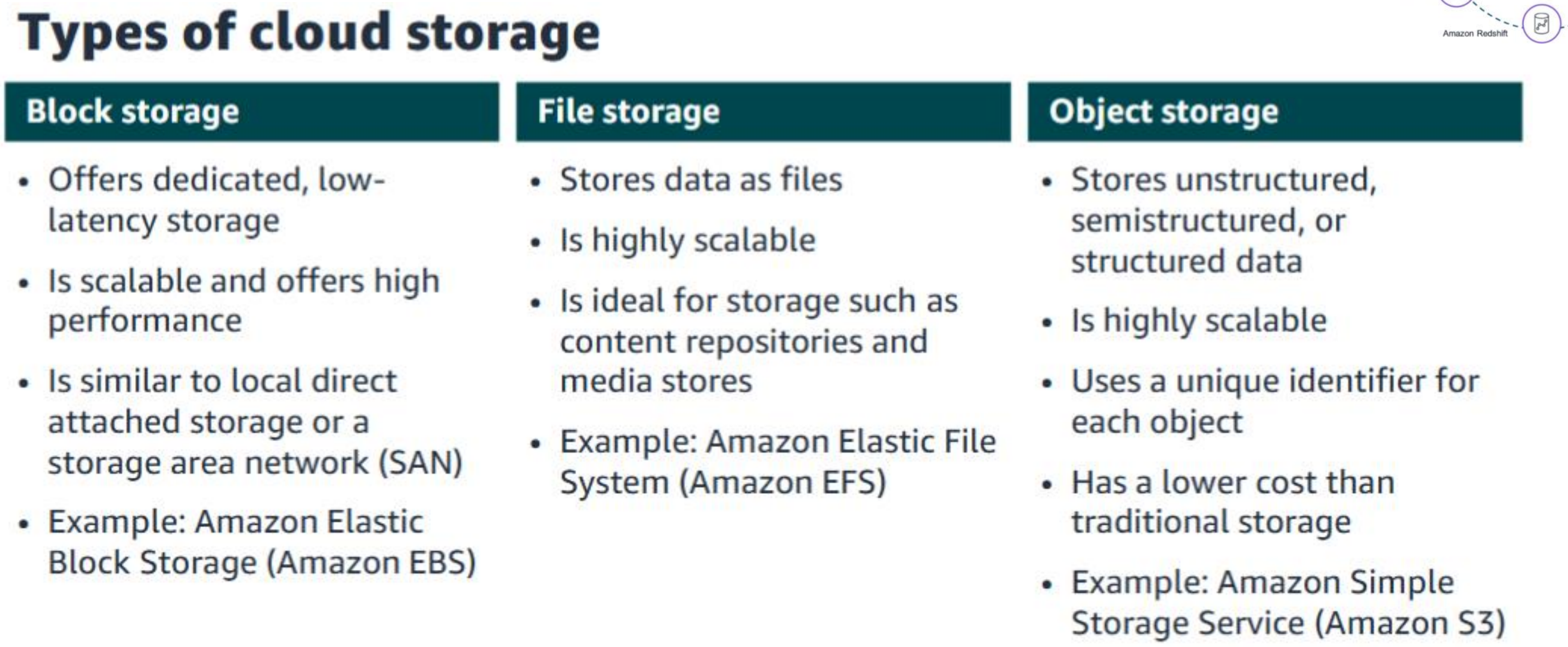
1. **Extract** structured data.
2. **Transform** the data into a format that matches the destination.
3. **Load** the data into structured storage for defined analytics.

**Extract load transform (ELT)**

1. **Extract** unstructured or structured data.
2. **Load** the data into the data lake in the format that is as close to the raw form as possible.
3. **Transform** the data as needed for analytic scenarios.

# Storing and Organizing Data





## Types of cloud storage

### Block storage

- Offers dedicated, low-latency storage
- Is scalable and offers high performance
- Is similar to local direct attached storage or a storage area network (SAN)
- Example: Amazon Elastic Block Storage (Amazon EBS)

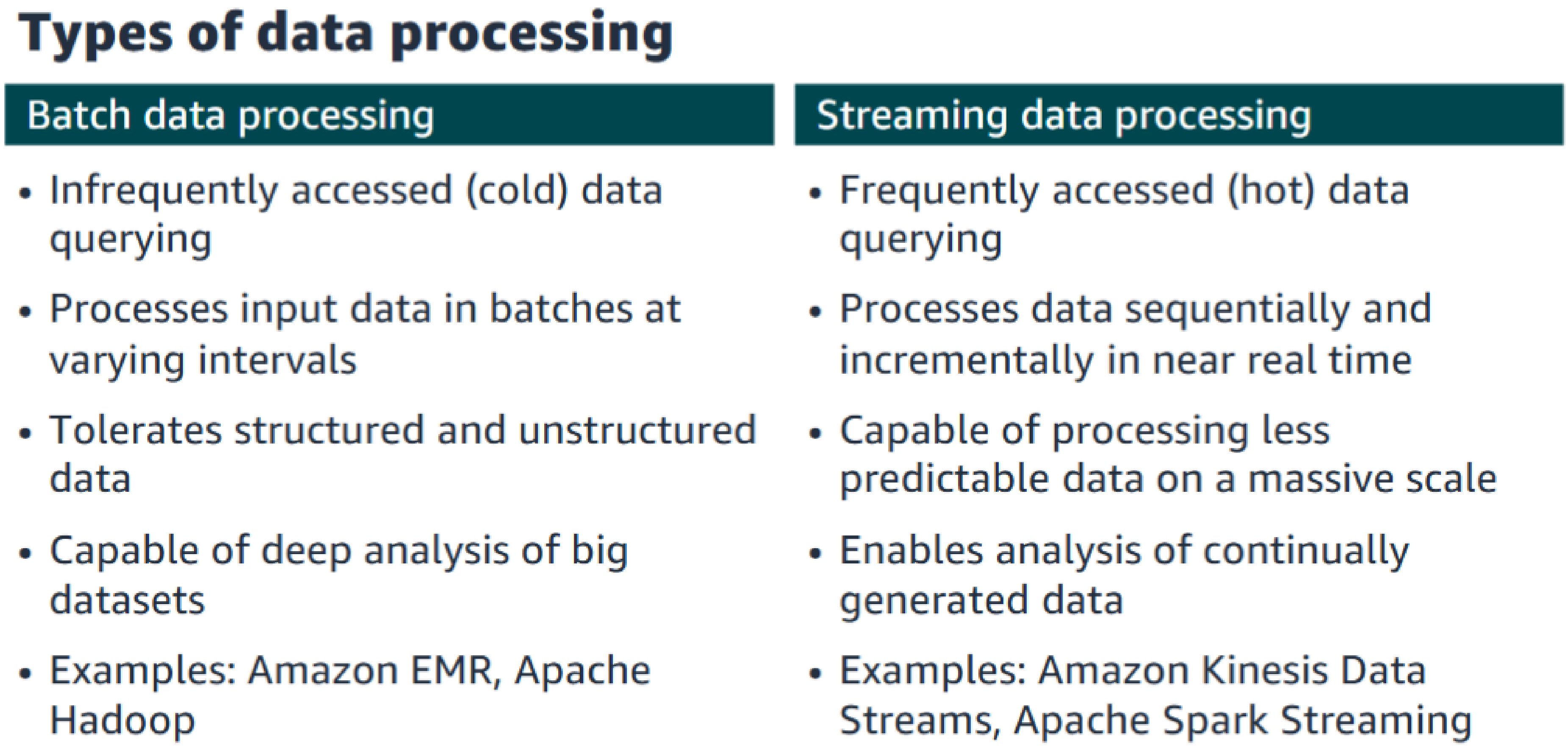
### File storage

- Stores data as files
- Is highly scalable
- Is ideal for storage such as content repositories and media stores
- Example: Amazon Elastic File System (Amazon EFS)

### Object storage

- Stores unstructured, semistructured, or structured data
- Is highly scalable
- Uses a unique identifier for each object
- Has a lower cost than traditional storage
- Example: Amazon Simple Storage Service (Amazon S3)

# Processing Big Data



## Types of data processing

| Batch data processing | Streaming data processing |
|---|---|
| • Infrequently accessed (cold) data querying | • Frequently accessed (hot) data querying |
| • Processes input data in batches at varying intervals | • Processes data sequentially and incrementally in near real time |
| • Tolerates structured and unstructured data | • Capable of processing less predictable data on a massive scale |
| • Capable of deep analysis of big datasets | • Enables analysis of continually generated data |
| • Examples: Amazon EMR, Apache Hadoop | • Examples: Amazon Kinesis Data Streams, Apache Spark Streaming |

# Processing Data for ML

**Three Layers of AWS ML Infrastructure**

## 1. Compute, Networking, and Storage Layer

1. Compute: EC2 P3/P4 (training), G4/Inf1 (inference), Elastic Inference (GPU acceleration), Elastic Fabric Adapter (low-latency networking)
2. Storage:
    1. Amazon S3: Scalable object storage for ML datasets
    2. Amazon EBS: High-performance block storage
    3. Amazon EFS/FSx: File system storage for large datasets and shared code

## 2. Framework Layer

1. Supports popular ML frameworks via Deep Learning AMIs and Deep Learning Containers.

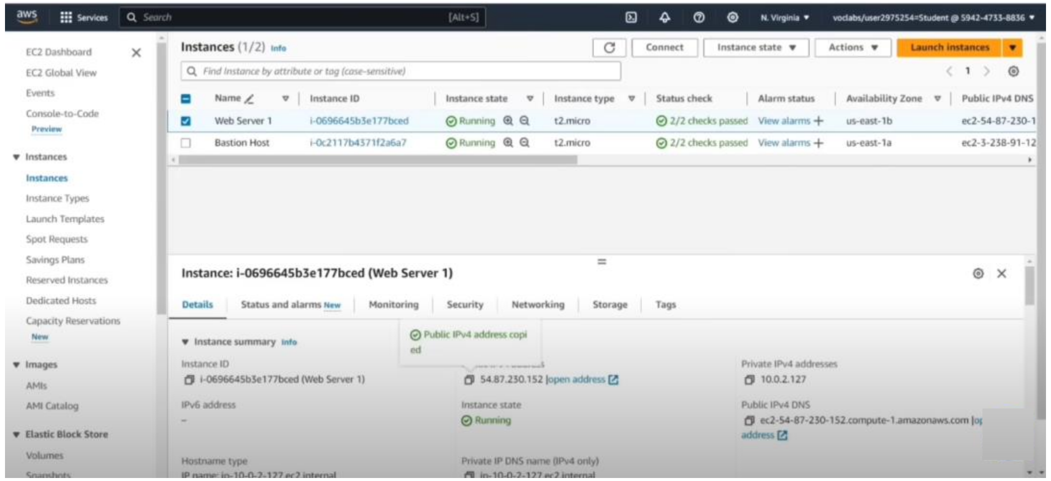
## 3. Workflow Services Layer

1. Amazon SageMaker: Integrated ML development, training, and deployment
2. Amazon EMR: Big data processing for ML
3. AWS Batch: Orchestrates scheduled ML training jobs
4. Amazon ECS/EKS: Container orchestration for scalable ML environments
5. AWS ParallelCluster: Simplifies HPC cluster deployment

# Automating the Pipeline

- Benefits of Automation
    - Improves stability, consistency, and efficiency by reducing manual intervention
    - Enables rapid response to changes and failures (e.g., automatic server replacement via CloudWatch)
    - Built-in AWS monitoring and automation tools support proactive management
- Infrastructure as Code (IaC)
    - Use AWS CloudFormation or AWS CDK to define and deploy infrastructure in code
    - Repeatability: Deploy identical environments (dev, test, prod) reliably
    - Supports consistent, scalable, and testable infrastructure for analytics and ML workloads
- Orchestration with AWS Step Functions
    - Automate and visualize complex workflows (e.g., ETL pipelines)
    - Integrate with services like Amazon Athena for hands-on lab scenarios

# Hands-on work



Instances (1/2) Info

| | Name | Instance ID | Instance state | Instance type | Status check | Alarm status | Availability Zone | Public IPv4 DNS |
|---|---|---|---|---|---|---|---|---|
| ☑ | Web Server 1 | i-0696645b3e177bced | Running | t2.micro | 2/2 checks passed | View alarms + | us-east-1b | ec2-54-87-230-1 |
| □ | Bastion Host | i-0c2117b4371f2a6a7 | Running | t2.micro | 2/2 checks passed | View alarms + | us-east-1a | ec2-3-238-91-12 |

Instance: i-0696645b3e177bced (Web Server 1)

Instance summary Info

Public IPv4 address copied

Instance ID: i-0696645b3e177bced (Web Server 1)

54.87.230.152 |open address

Private IPv4 addresses: 10.0.2.127

IPv6 address: –

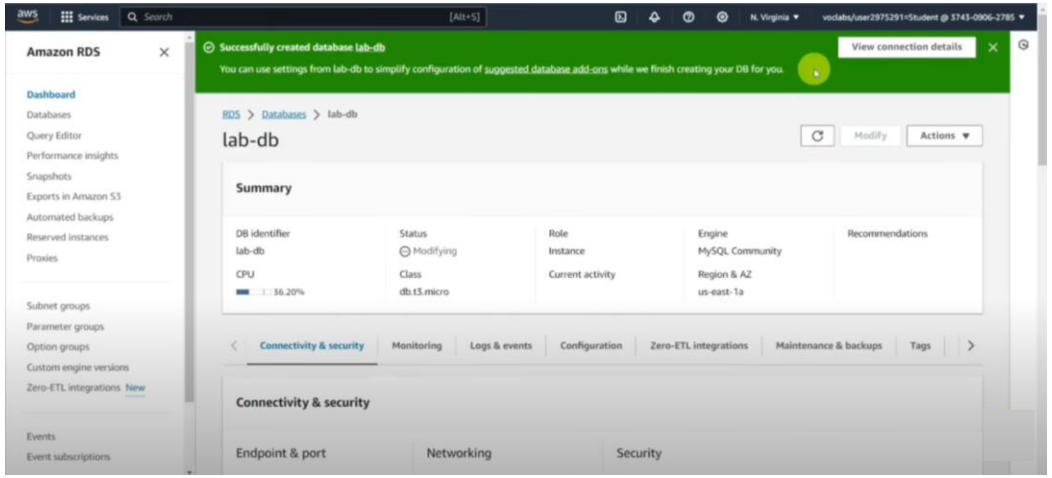
Instance state: Running

Public IPv4 DNS: ec2-54-87-230-152.compute-1.amazonaws.com |op address

Hostname type: IP name: ip-10-0-2-127.ec2.internal

Private IP DNS name (IPv4 only): ip-10-0-2-127.ec2.internal

**Lab - 2 Build your VPC and Launch a Web Server**



Successfully created database lab-db

You can use settings from lab-db to simplify configuration of suggested database add-ons while we finish creating your DB for you.

lab-db

Summary

DB identifier: lab-db

CPU: 36.20%

Status: Modifying

Class: db.t3.micro

Role: Instance

Current activity

Engine: MySQL Community

Region & AZ: us-east-1a

Recommendations

Connectivity & security

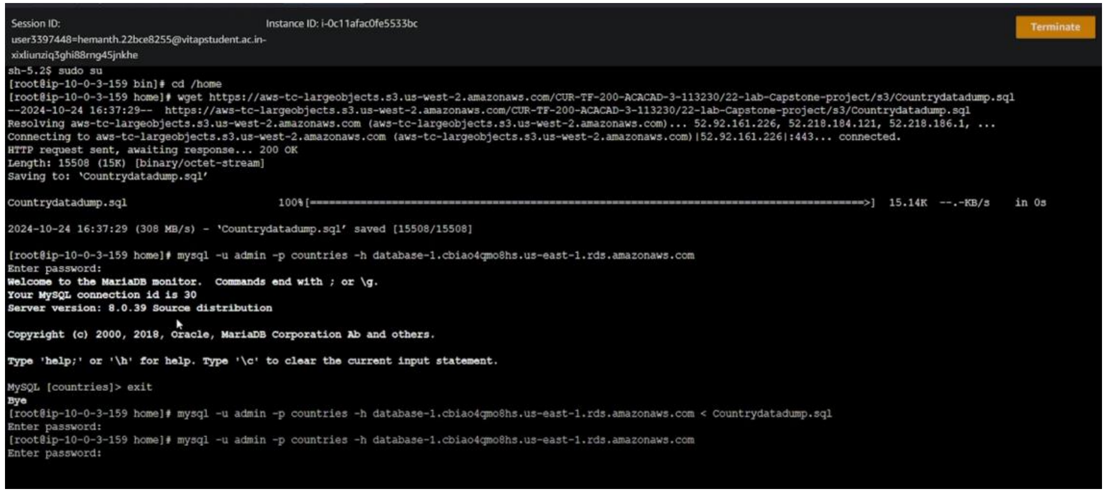
Endpoint & port

Networking

Security

**Lab - 5 Build a Database Server**

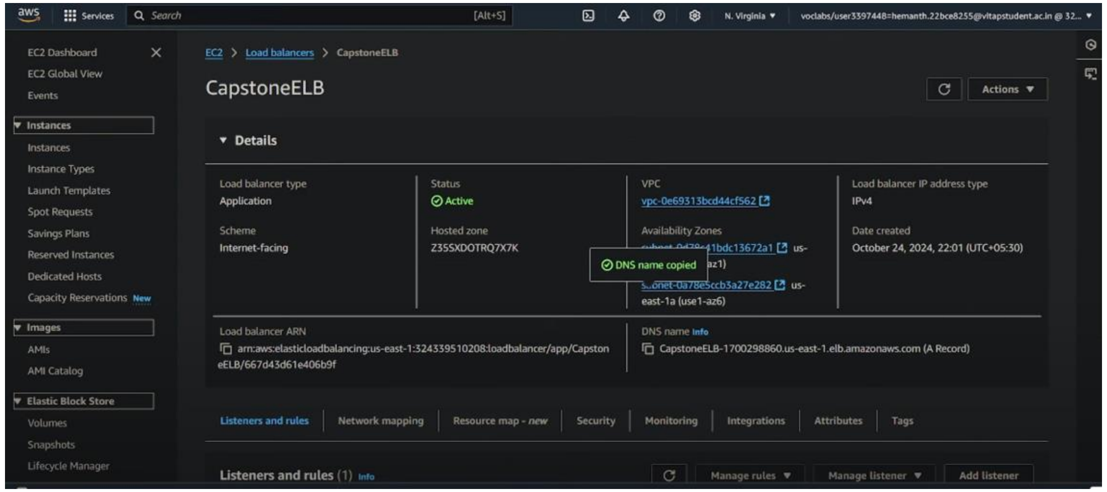
# Project: Capstone project



Session ID: user3397448=hemanth.22bce8255@vitapstudent.ac.in-xixliunziq3ghi88rng45jnkhe

Instance ID: i-0c11afac0fe5533bc

```
sh-5.2$ sudo su
[root@ip-10-0-3-159 bin]# cd /home
[root@ip-10-0-3-159 home]# wget https://aws-tc-largeobjects.s3.us-west-2.amazonaws.com/CUR-TF-200-ACACAD-3-113230/22-lab-Capstone-project/s3/Countrydatadump.sql
--2024-10-24 16:37:29--  https://aws-tc-largeobjects.s3.us-west-2.amazonaws.com/CUR-TF-200-ACACAD-3-113230/22-lab-Capstone-project/s3/Countrydatadump.sql
Resolving aws-tc-largeobjects.s3.us-west-2.amazonaws.com (aws-tc-largeobjects.s3.us-west-2.amazonaws.com)... 52.92.161.226, 52.218.184.121, 52.218.186.1, ...
Connecting to aws-tc-largeobjects.s3.us-west-2.amazonaws.com (aws-tc-largeobjects.s3.us-west-2.amazonaws.com)|52.92.161.226|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 15508 (15K) [binary/octet-stream]
Saving to: 'Countrydatadump.sql'

Countrydatadump.sql                 100%[===================================================================================>]  15.14K  --.-KB/s    in 0s

2024-10-24 16:37:29 (308 MB/s) - 'Countrydatadump.sql' saved [15508/15508]

[root@ip-10-0-3-159 home]# mysql -u admin -p countries -h database-1.cbiao4qmo8hs.us-east-1.rds.amazonaws.com
Enter password:
Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MySQL connection id is 30
Server version: 8.0.39 Source distribution

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MySQL [countries]> exit
Bye
[root@ip-10-0-3-159 home]# mysql -u admin -p countries -h database-1.cbiao4qmo8hs.us-east-1.rds.amazonaws.com < Countrydatadump.sql
Enter password:
[root@ip-10-0-3-159 home]# mysql -u admin -p countries -h database-1.cbiao4qmo8hs.us-east-1.rds.amazonaws.com
Enter password:
```



CapstoneELB

Details

Load balancer type: Application

Scheme: Internet-facing

Status: Active

Hosted zone: Z35SXDOTRQ7X7K

VPC: vpc-0e69313bcd44cf562

Availability Zones: [illegible]1bdc13672a1 us-[illegible]az1) s[illegible]onet-0a78e5ccb3a27e282 us-east-1a (use1-az6)

DNS name copied

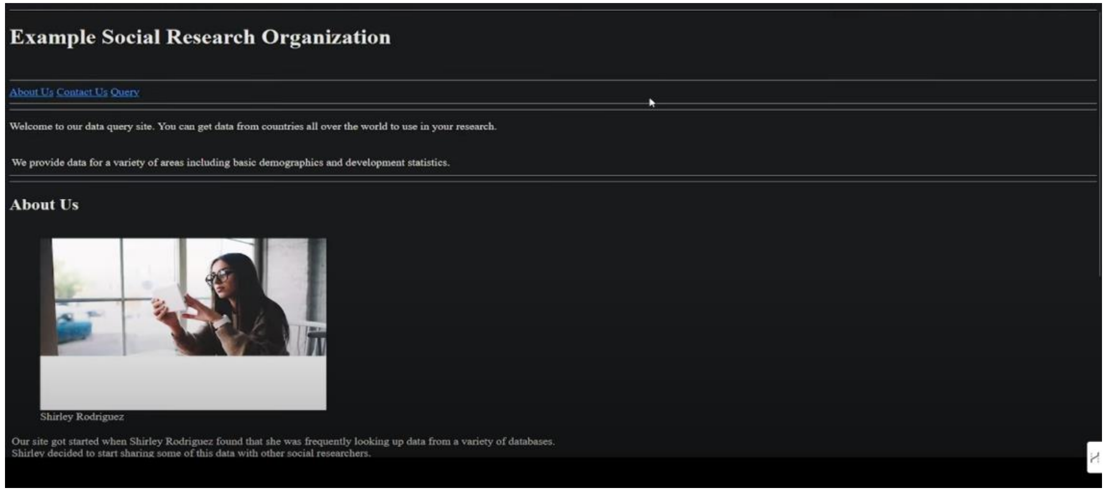
Load balancer IP address type: IPv4

Date created: October 24, 2024, 22:01 (UTC+05:30)

Load balancer ARN: arn:aws:elasticloadbalancing:us-east-1:324339510208:loadbalancer/app/CapstoneELB/667d43d61e406b9f

DNS name Info: CapstoneELB-1700298860.us-east-1.elb.amazonaws.com (A Record)

Listeners and rules (1) Info



Example Social Research Organization

About Us Contact Us Query

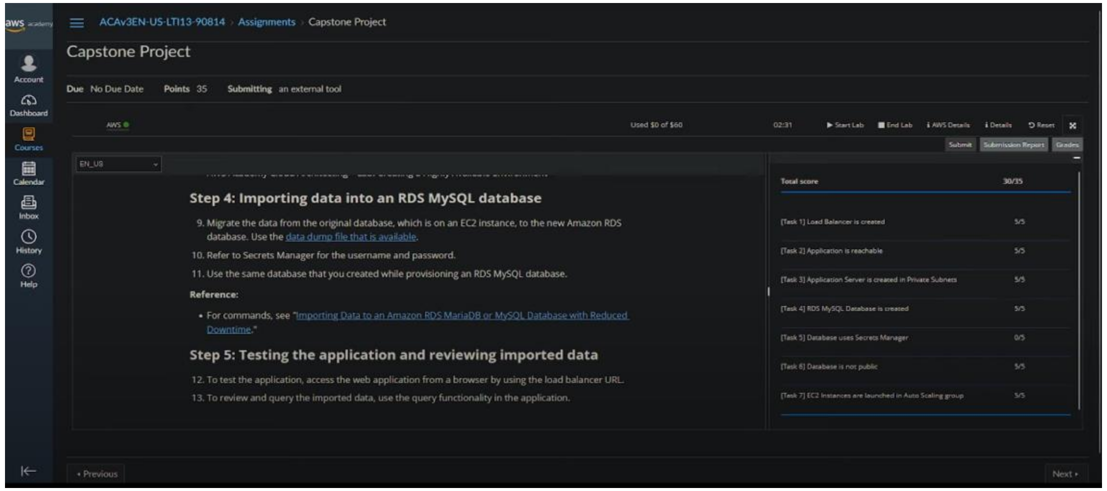
Welcome to our data query site. You can get data from countries all over the world to use in your research.

We provide data for a variety of areas including basic demographics and development statistics.

About Us

Shirley Rodriguez

Our site got started when Shirley Rodriguez found that she was frequently looking up data from a variety of databases. Shirley decided to start sharing some of this data with other social researchers.



ACAv3EN-US-LTI13-90814 › Assignments › Capstone Project

Capstone Project

Due No Due Date Points 35 Submitting an external tool

Used $0 of $60 02:31

Step 4: Importing data into an RDS MySQL database

9. Migrate the data from the original database, which is on an EC2 instance, to the new Amazon RDS database. Use the data dump file that is available.
10. Refer to Secrets Manager for the username and password.
11. Use the same database that you created while provisioning an RDS MySQL database.

Reference:

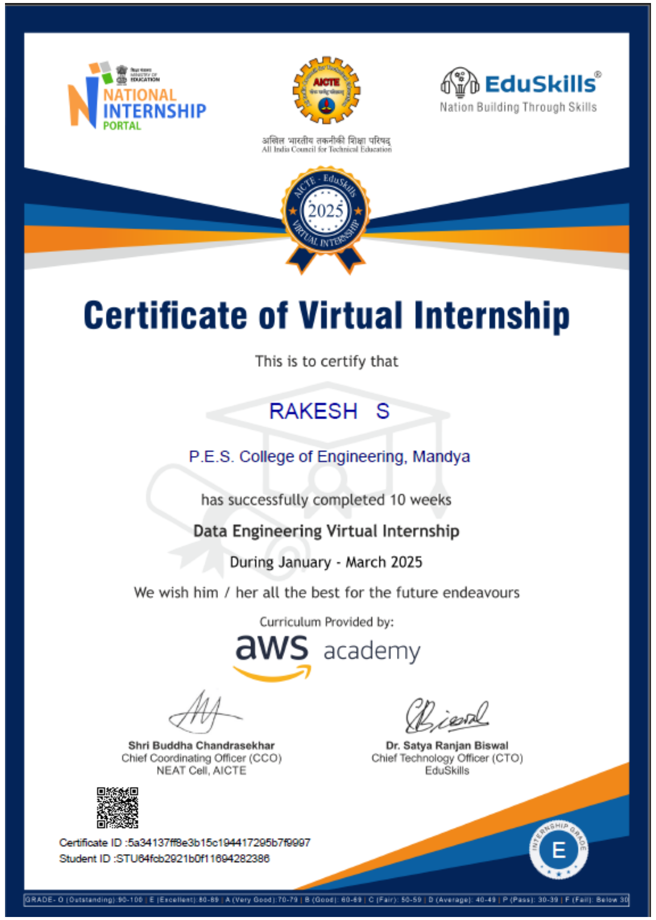
- For commands, see "Importing Data to an Amazon RDS MariaDB or MySQL Database with Reduced Downtime."

Step 5: Testing the application and reviewing imported data

12. To test the application, access the web application from a browser by using the load balancer URL.
13. To review and query the imported data, use the query functionality in the application.

| Total score | 30/35 |
|---|---|
| [Task 1] Load Balancer is created | 5/5 |
| [Task 2] Application is reachable | 5/5 |
| [Task 3] Application Server is created in Private Subnets | 5/5 |
| [Task 4] RDS MySQL Database is created | 5/5 |
| [Task 5] Database uses Secrets Manager | 0/5 |
| [Task 6] Database is not public | 5/5 |
| [Task 7] EC2 Instances are launched in Auto Scaling group | 5/5 |

# Certification



NATIONAL INTERNSHIP PORTAL

अखिल भारतीय तकनीकी शिक्षा परिषद्
All India Council for Technical Education

EduSkills®
Nation Building Through Skills

## Certificate of Virtual Internship

This is to certify that

RAKESH S

P.E.S. College of Engineering, Mandya

has successfully completed 10 weeks

**Data Engineering Virtual Internship**

During January - March 2025

We wish him / her all the best for the future endeavours

Curriculum Provided by:

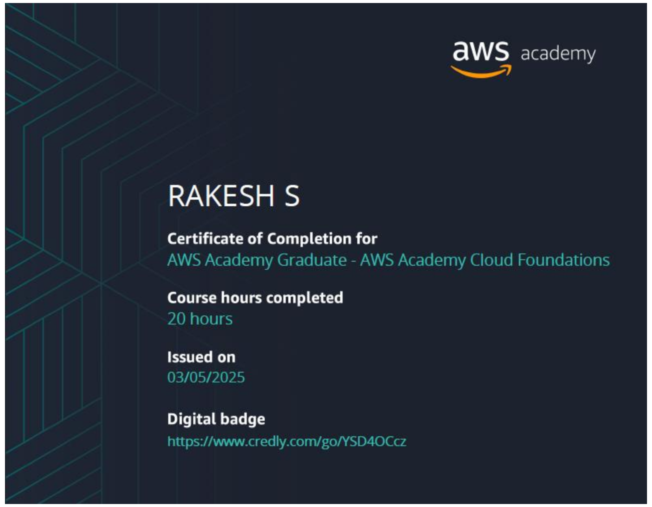
aws academy

**Shri Buddha Chandrasekhar**
Chief Coordinating Officer (CCO)
NEAT Cell, AICTE

**Dr. Satya Ranjan Biswal**
Chief Technology Officer (CTO)
EduSkills

Certificate ID :5a34137ff8e3b15c194417295b7f9997
Student ID :STU64fcb2921b0f11694282386

GRADE- O (Outstanding):90-100 | E (Excellent):80-89 | A (Very Good):70-79 | B (Good): 60-69 | C (Fair): 50-59 | D (Average): 40-49 | P (Pass): 30-39 | F (Fail): Below 30



aws academy

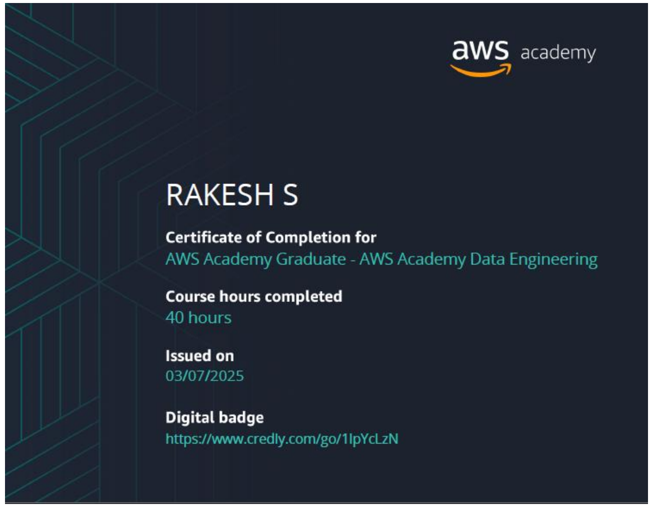
RAKESH S

**Certificate of Completion for**
AWS Academy Graduate - AWS Academy Cloud Foundations

**Course hours completed**
20 hours

**Issued on**
03/05/2025

**Digital badge**
https://www.credly.com/go/YSD4OCcz



aws academy

RAKESH S

**Certificate of Completion for**
AWS Academy Graduate - AWS Academy Data Engineering

**Course hours completed**
40 hours

**Issued on**
03/07/2025

**Digital badge**
https://www.credly.com/go/1lpYcLzN

# Thank you..