An Internship Report

On

DATA ENGINEERING VIRTUAL INTERNSHIP

Submitted in partial fulfillment of the requirements for

the award of the degree of

BACHELOR OF TECHNOLOGY

in

Computer Science and Engineering (Data Science)

by

S.N. SHARANYA LAKSHMI

(224G1A3288)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING (DATA SCIENCE)

SRINIVASA RAMANUJAN INSTITUTE OF TECHNOLOGY (AUTONOMOUS)

(Affiliated to JNTUA, accredited by NAAC with 'A' Grade, Approved by AICTE, New Delhi & Accredited by NBA (EEE, ECE & CSE))

Rotary Puram Village, B K Samudram Mandal, Ananthapuramu-515701.

2025 - 2026

SRINIVASA RAMANUJAN INSTITUTE OF TECHNOLOGY (AUTONOMOUS)

(Affiliated to JNTUA, accredited by NAAC with 'A' Grade, Approved by AICTE, New Delhi & Accredited by NBA (EEE, ECE & CSE))
Rotarypuram village, B K Samudram Mandal, Ananthapuramu-515701.

Department of Computer Science & Engineering (Data Science)

Certificate

This is to certify that the internship report entitled "Data Engineering Virtual Internship" is the bonafide work carried out by **S.N. SHARANYA LAKSHMI** bearing Roll Number 224G1A3288 in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering (Data Science)** for 10 weeks from July 2025 to September 2025.

Internship Coordinator

Dr. G. Hemanth Kumar Yadav, M. Tech., Ph.D.,
Associate Professor

Head of the Department

Dr. P. Chitralingappa, M. Tech., Ph.D.,
Associate Professor & HOD of CSD

Date:

Place: Ananthapuramu

EXTERNAL EXAMINER

PREFACE

Brief overview of the company's history:

- **Who founded it:** All India Council for Technical Education (AICTE) has initiated various activities for promoting industrial internship at the graduate level in technical institutes and Eduskills is a Non-profit organization which enables industry 4.0 ready digital workforce in India. The vision of the organization is to fill the gap between Academic and Industry by ensuring world class curriculum access to the faculties and students. Formation of the All-India Council for Technical Education (AICTE) in 19445 by the Government of India.
- **What purpose and when:** With a vision to create an industry-ready workforce who will eventually become leaders in emerging technologies, Eduskills & AICTE launches 'Virtual Internship' program on Data Engineering. This field is one of the most in-demand, and this internship will serve as a primer.

Company's Mission Statement: The main mission of these initiatives is enhancement of the employability skills of the students passing out from Technical Institutions.

Business Activities: The All India Council for Technical Education (AICTE) primarily focuses on regulating and promoting technical education in India. Its business activities include accrediting institutions, approving new courses, setting quality standards, fostering research, providing policy recommendations, and ensuring the overall development of technical education across the country.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of people who made it possible, whose constant guidance and encouragement crowned our efforts with success. It is a pleasant aspect that I have now the opportunity to express my gratitude for all of them.

It is with immense pleasure that I would like to express my indebted gratitude to my internship coordinator **Dr. G. Hemanth Kumar Yadav, Associate Professor, Department of Computer Science and Engineering (AI&ML)**, who has supported me a lot and encouraged me in every step of the internship work. I thank him for the stimulating support, constant encouragement and constructive criticism which have made possible to bring out this internship work.

I am very much thankful **to Dr. P. Chitralingappa, Associate Professor & HOD, Computer Science and Engineering (Data Science),** for his kind support and for providing necessary facilities to carry out the work.

I wish to convey my special thanks to **Dr. G. Balakrishna, Principal of Srinivasa Ramanujan Institute of Technology** for giving the required information in doing my internship. Not to forget, I thank all other faculty and non-teaching staff, and my friends who had directly or indirectly helped and supported me in completing my internship in time.

I also express our sincere thanks to the Management for providing excellent facilities and support.

Finally, I wish to convey my gratitude to my family who fostered all the requirements and facilities that I need.

S.N. Sharanya Lakshmi

(224G1A3288)

INDEX

List of Figures

List of Abbreviations

AWS	Amazon Web Services
CDN	Content Delivery Network
IAM	Identity and Access Management
EC2	Elastic Cloud Compute
VPC	Virtual Private Cloud
DNS	Domain Name System
RDS	Relational Database Service
ELB	Elastic Load Balancer

CHAPTER -1
INTRODUCTION

1.1 Introduction to AWS:

Amazon Web Services (AWS) revolutionized computing with its 2006 launch, introducing scalable and cost-effective infrastructure solutions. Initially offering Amazon S3 for scalable online storage and Amazon EC2 for flexible cloud computing, AWS provided on-demand access to storage and virtual servers, eliminating physical infrastructure needs. Addressing growing demands for flexibility, reduced costs, enhanced reliability, and collaboration, AWS has expanded to over 200 services, including database management, artificial intelligence, security, and analytics. Today, AWS is a leading cloud platform supporting businesses, governments, and organizations worldwide in their digital transformation journeys, offering a robust ecosystem for innovation and growth.

What is AWS?

Amazon Web Services (AWS) is a pioneering cloud computing platform that has transformed the IT landscape since its inception in 2006, offering a broad spectrum of cloud-based solutions including compute services, storage, networking, artificial intelligence, machine learning, analytics, databases, and IoT. By providing scalable, on-demand resources, AWS empowers businesses and individuals to deploy and manage infrastructure without physical hardware constraints, quickly adapt to changing workload requirements, and focus on innovation and growth, rather than infrastructure maintenance, all while fostering flexibility, reliability, and cost-effectiveness. It operates on a pay-as-you-go pricing model, which allows users to pay only for the resources they consume, thereby optimizing costs and eliminating the need for significant upfront investments. AWS's global infrastructure, with multiple geographic regions and Availability Zones, ensures high availability, reliability, and low latency for applications worldwide. The platform's robust security features, compliance certifications, and extensive range of tools and services make it a preferred choice for businesses of all sizes seeking to innovate, scale, and optimize.

Fig 1.1 AWS Cloud

1.2 Introduction to Cloud Computing:

Cloud computing is a revolutionary paradigm that has fundamentally changed how computing resources are delivered and managed. By leveraging the power of the internet, it allows users to access a comprehensive range of services—includingservers, storage, databases, networking, software, and analytics—without the need forphysical hardware or on-site infrastructure. This shift, which began in the mid-2000s, has provided both organizations and individuals with unprecedented flexibility, enabling them to scale resources dynamically based on demand.

The scalability and efficiency of cloud computing are unmatched, as they allow for cost-effective adjustments in resource allocation. This model not only reduces the need for substantial capital investment in physical infrastructure but also facilitates rapid deployment and innovation. Cloud computing supports a variety of service models such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), catering to diverse needs and preferences.

Importance of Cloud Computing:

Cloud computing has revolutionized the IT industry, providing vital benefits that are now indispensable for both businesses and individuals. Its cost efficiency

eliminates the necessity for large upfront investments in hardware and infrastructure, offering instead a flexible pay-as-you-go model. This approach allows organizations to scale their resources up or down with ease, adapting to current demands without hassle. Additionally, cloud computing enhances accessibility and collaboration, enabling users to access services and data from anywhere with an internet connection. This capability supports remote work and fosters global teamwork, making it an essential tool for modern enterprises.

AWS provides a robust platform for various use cases, including web and mobile applications, data analytics, machine learning, enterprise IT, DevOps, and disaster recovery. AWS offers various services like Compute Services (EC2, Lambda, Elastic Beanstalk), Storage Services (S3, EBS, Elastic File System), Database Services (RDS, DynamoDB, DocumentDB).

Types of Cloud Computing:

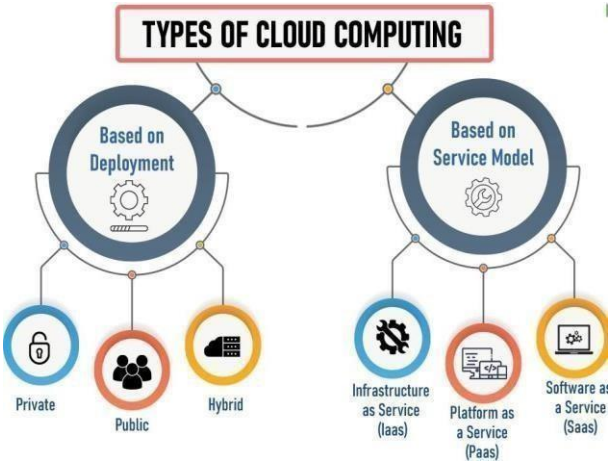

Fig 1.2 Cloud Computing

Deployment Models:

Public Cloud : Public clouds are operated by third-party providers and offer services to multiple organizations over the internet .This model offers high scalability, cost efficiency, and ease of access, making it ideal for businesses that need to quickly scale resources without significant capital investment. Examples include Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP).

Private Cloud : A private cloud is a dedicated environment used exclusively by a

single organization. It can be hosted on-premises or by a third-party provider, but it is not shared with other organizations. This model provides enhanced control, security, and customization options, making it suitable for businesses with specific compliance requirements or sensitive data. Private clouds allow organizations to tailor their infrastructure to meet specific needs while maintaining greater control over security and performance.

Hybrid Cloud: Hybrid clouds combine public and private cloud elements, allowing data and applications to be shared between them. This model offers flexibility and optimized resource use. Organizations can keep critical or sensitive workloads on the private cloud while utilizing the public cloud for less sensitive operations or to handle peak demands.

Community Cloud: Community clouds are shared among organizations with common concerns, such as regulatory compliance or security needs. They can be managed by one or more organizations or a third-party provider.Community clouds canbe managed by the participating organizations .

Service Models:

Infrastructure as a Service (IaaS):

Infrastructure as a Service (IaaS) provides virtualized computing resources over the internet, such as virtual machines, storage, and networks. In this model, the cloud provider manages the physical hardware, but users have control over the operating systems, applications, and configurations. IaaS offers the highest level of flexibility among the cloud service models, allowing organizations to scale resources up or down according to their needs. It is ideal for businesses that require control over their infrastructure but do not want to invest in physical hardware. Examples include Amazon EC2, Microsoft Azure, and Google Compute Engine.

Platform as a Service (PaaS):

Platform as a Service (PaaS) delivers a platform that allows developers to build, deploy, and manage applications without worrying about the underlying infrastructure. The cloud provider manages everything from the operating system to the servers, storage, and networking, enabling developers to focus solely on writing code and developing applications. PaaS simplifies the development process providing

pre-configured environments, development tools, and frameworks, making it easier to develop, test, and deploy applications quickly. Examples include Google App Engine, AWS Elastic Beanstalk, and Microsoft Azure App Services.

Software as a Service (SaaS):

Software as a Service (SaaS) delivers software applications over the internet, typically on a subscription basis. In this model, the cloud provider hosts and manages the software, including maintenance, updates, and security, while users access the application through a web browser. SaaS eliminates the need for local installation and maintenance, making it convenient for end-users who only need to access the software's functionality. It is widely used for applications like email, customer relationship management (CRM), and collaboration tools. Examples include Google Workspace, Microsoft Office 365, and Salesforce. SaaS offers the convenience of ready-to-use applications.

CHAPTER – 2
AWS CLOUD FOUNDATIONS

Amazon Web Services (AWS) empowers businesses to innovate and thrive in the cloud, offering a robust suite of scalable, flexible, and cost-effective computing services. The AWS Cloud Foundations framework provides a solid understanding of the essential services and concepts necessary to unlock the full potential of cloud computing, enabling organizations to build, deploy, and manage applications with ease and efficiency.The AWS Cloud Foundations encompass a broad range of fundamental services and concepts essential for leveraging the full potential of the cloud.

2.1 AWS Cloud Infrastructure overview:

The AWS Global Infrastructure provides a robust, scalable, and reliable foundation for delivering cloud services worldwide, enabling customers to deploy applications closer to users, meet regulatory requirements, and achieve high availability and reliability, ensuring business success and global reach through its secure data centers, low-latency network connectivity, strategic edge locations, and compliant architecture.

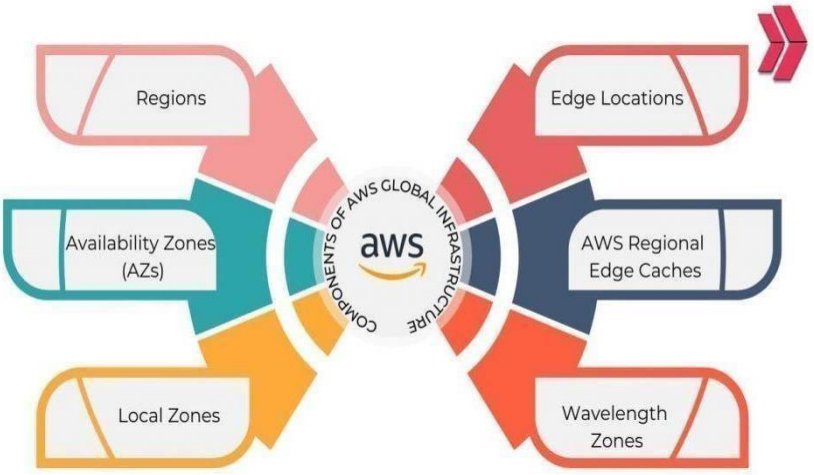

Fig 2.1 AWS Cloud Infrastructure

Regions: AWS Regions are separate geographic areas where AWS data centers are clustered. Each region is isolated from the others, providing a level of fault tolerance and stability.

Availability Zones: Each region consists of multiple, physically separated Availability Zones. AZs are designed to be isolated from failures in other AZs, offering customers the ability to run apps with high availability and fault tolerance.

Edge Locations: These are locations worldwide where AWS has deployed resources to cache copies of your data closer to end users. AWS uses edge locations for services like Amazon CloudFront and Route 53, reducing latency by serving content faster to users around the globe.

AWS Regional Edge Caches: These are a feature of Amazon CloudFront, a content delivery network (CDN) offered by AWS. They are used to cache content at edge locations, which are strategically situated around the world. This allows for faster delivery of content to users and reduces latency.

Wavelength Zones: There are specific edge locations that are designed to provide ultra-low latency and high-bandwidth access to AWS services and applications. They are typically deployed in conjunction with telecommunications providers and are connected to the AWS global network.

2.2 Cloud Economics and Billing:

Cloud economics refers to the financial management and cost optimization of cloud computing resources. It involves understanding the costs associated with cloud computing, such as usage-based pricing, tiered pricing, and discounts for committed usage. By optimizing cloud economics, organizations can maximize the value of their cloud investments and achieve cost savings. Cloud economics involves analyzing usage patterns, rightsizing resources, and selecting the most cost-effective pricing models. It also involves taking advantage of discounts, such as reserved instances and spot instances, and using cost allocation tags to track costs.

Billing refers to the process of generating and managing invoices for cloud computing services. In the context of cloud economics, billing is a critical component as it directly impacts an organization's cost management and optimization strategies.

Cloud providers offer various billing models, including:

- Usage-based billing: Charging customers based on their actual resource usage.

- Tiered pricing: Offering discounts for higher usage levels.
- Reserved Instances: Discounted pricing for committed usage.
- Spot Instances: Bid on unused capacity for discounted pricing.

Fig 2.2 Billing and Pricing

2.3 AWS Cloud Security and compliance:

AWS Cloud Security and Compliance refer to the practices and controls used to protect and secure data and applications in the AWS cloud. AWS provides a secure infrastructure and services, but it's the customer's responsibility to ensuretheir data and applications are secure and compliant with relevant regulations.

2.3.1 Security:

Security in the AWS cloud refers to the practices and controls used to protect and secure data, applications, and infrastructure from unauthorized access, use, disclosure, disruption, modification, or destruction. AWS provides a secure infrastructure and services, but it's the customer's responsibility to ensure their data and applications are secure.

Some keysecurity features and services in AWS include:

- Identity and Access Management (IAM): manages access to AWS resources
- Virtual Private Cloud (VPC): provides a secure network environment
- Security Groups: controls access to instances
- Network ACLs: controls access to subnets
- Encryption: protects data at rest and in transit

- Key Management Service (KMS): manages encryption keys
- CloudWatch: monitors and logs security-related events

2.3.2 Compliance:

Compliance in the AWS cloud refers to the process of ensuring that your cloud resources and data meet specific regulatory, industry, or organizational requirements. AWS provides various compliance programs and certifications to help customers meet these requirements.

Some key compliance features and services in AWS include:

- Compliance frameworks: AWS supports various compliance frameworks such as PCI-DSS, HIPAA/HITECH, GDPR, and more.
- Certifications: AWS holds various certifications such as SOC, ISO, and PCI-DSS
- Security and compliance controls: AWS provides security and compliance controls such as IAM, VPC, and encryption.
- Audit and logging: AWS provides audit and logging capabilities such as CloudWatch and CloudTrail.
- Compliance monitoring: AWS provides compliance monitoring capabilities such as Inspector and Config.

2.4 Networking and content delivery:

VPC (Virtual Private Cloud): VPC is a networking service that allows you to create a virtual network in the cloud. You can create subnets, route tables, and network ACLs to manage traffic flow. VPC provides a secure and isolated environment for your resources.

Route 53: Route 53 is a domain name system (DNS) service that allows you to route traffic to your resources. You can create hosted zones, record sets, and health checks to manage traffic flow. Route 53 provides high availability and durability for your DNS needs.

CloudFront: CloudFront is a content delivery network (CDN) service that allows you to distribute content to your users. You can create distributions, origins, and cache behaviors to manage content delivery. CloudFront provides high performance and low latency for your content delivery needs.

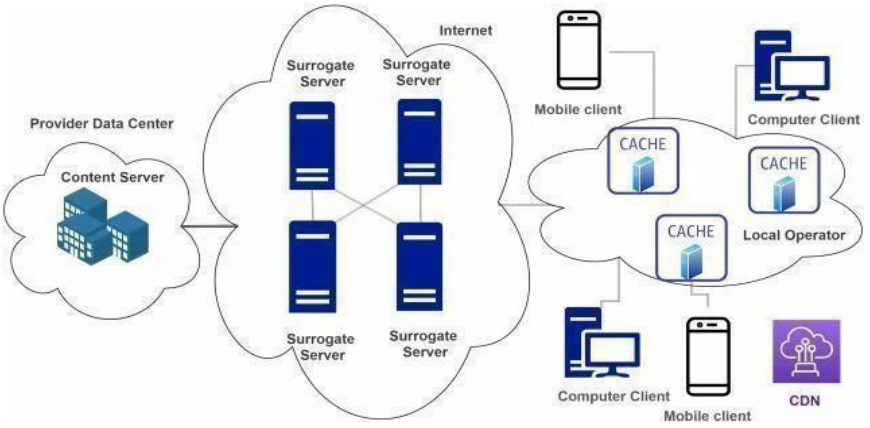

Fig 2.4 Content Delivery

AWS Networking and Content Delivery services enable secure, fast, and reliable connectivity and content distribution, featuring Amazon Virtual Private Cloud (VPC), Elastic Load Balancer (ELB), Elastic Network Interface (ENI), Direct Connect, Transit Gateway, CloudFront, Route 53, API Gateway, and Web Application Firewall (WAF).

2.5 Auto Load Balancing:

Auto Load Balancing in AWS is a feature that automatically distributes

incoming traffic across multiple targets, such as EC2 instances, containers, or IP addresses, to ensure high availability and scalability. AWS offers several types of load balancers, including Application Load Balancer (ALB), Network Load Balancer (NLB),and Classic Load Balancer (CLB), each suitable for different types of traffic and use cases. By using Auto Load Balancing, applications can ensure high availability, scalability, and improved responsiveness, as traffic is automatically redirected away from unhealthy targets and adjusted to changes in traffic volume.

To set up Auto Load Balancing, users create a load balancer, configure settings and targets, and enable health checks and Auto Scaling. Load balancing options to support different use cases and requirements. By leveraging Auto Load Balancing, users can ensure their applications are always available and responsive, even in the face of changing traffic demands.

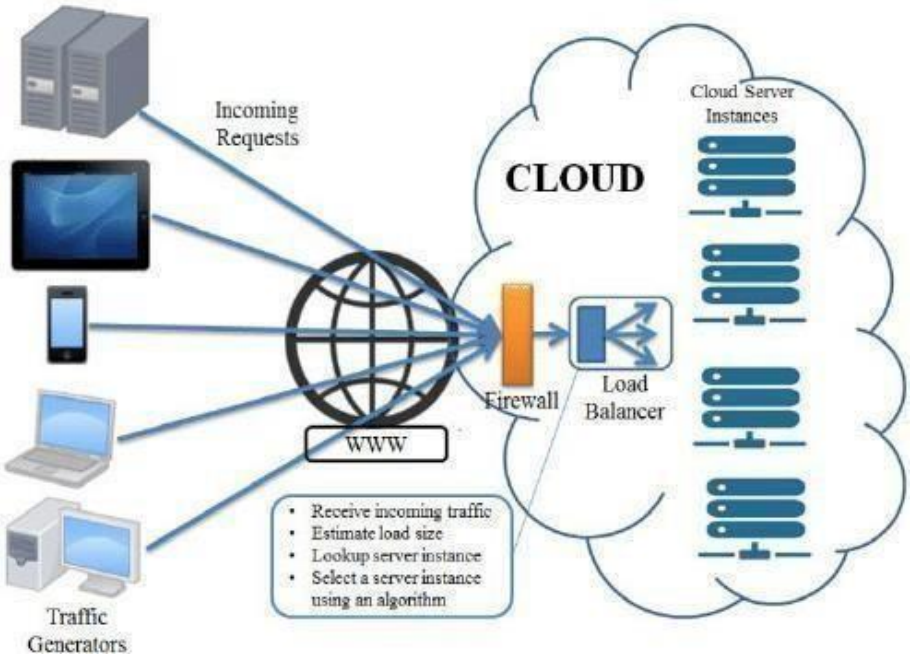

Fig 2.5 Working of Auto Load Balancing

Auto Load Balancing in AWS provides several benefits, including:

1. High Availability: Ensures that applications are always available and accessible.
2. Scalability: Automatically adjusts to changes in traffic volume.
3. Fault Tolerance: Detects and redirects traffic away from unhealthy targets.

4. Improved Responsiveness: Reduces latency and improves application responsiveness

To set up Auto Load Balancing in AWS, follow these steps:

1. Create a load balancer (ALB, NLB, or CLB).
2. Configurethe load balancer with the desired settings (e.g., protocol, port, and targets)
3. Add targets (EC2 instances, containers, or IP addresses) to the load balancer.
4. Configure health checks to monitor target health.
5. Enable Auto Scaling to adjust the number oftargets based on traffic demand.

CHAPTER 3

AWS Academy Data Engineering

3.1 Introduction to AWS Data Engineering :

This course is most aligned to a data engineer role. However, this course would also be appropriate for data analysts; data scientists; extract, transform, and load (ETL) developers; or machine learning (ML) practitioners who want to understand how the data that they use in their analysis and predictions is prepared for analysis using AWS.

3.2 Data-Driven Organizations :

The data science behind data-driven decisions falls into two main categories: data analytics, and artificial intelligence (AI) or its subfield, machine learning (ML). The primary distinction between the two is how data scientists use them to arrive at their results.

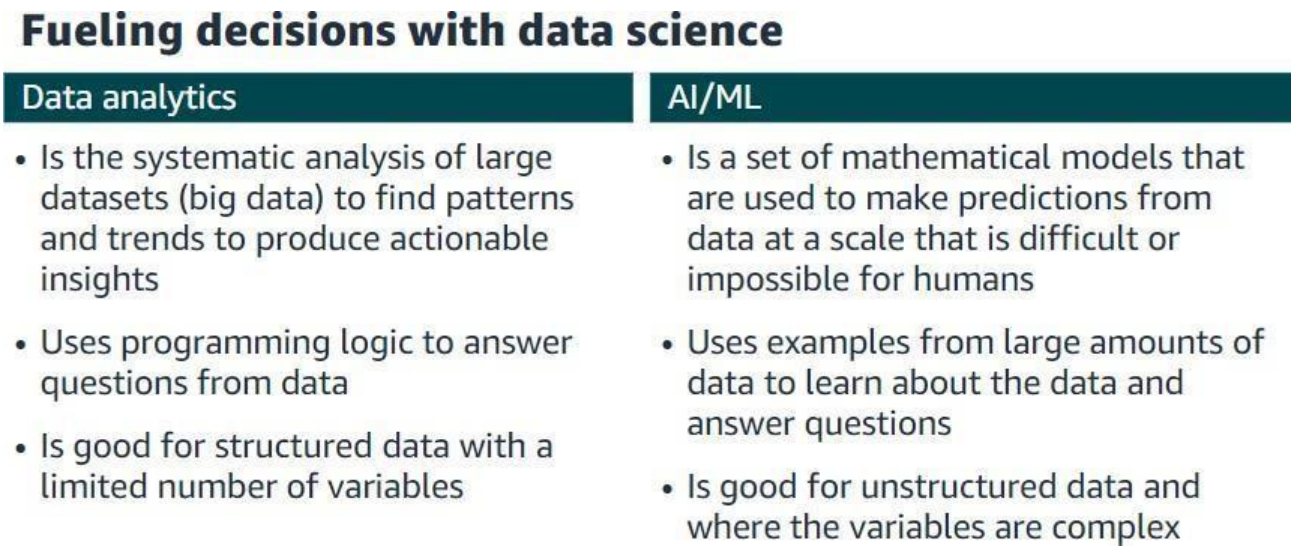

Fig 3.2 Harnessing Data Science for Informed Decisions

3.3 The Elements of Data :

AWS data engineering elements include data ingestion (Kinesis, S3), storage (S3, DynamoDB), processing (Glue, Lambda), transformation (Glue, EMR), analytics (Redshift, QuickSight), visualization (QuickSight), and governance (Lake Formation, IAM). Data types include structured, semi-structured, and unstructured, with integration via APIs, messaging, and ETL.

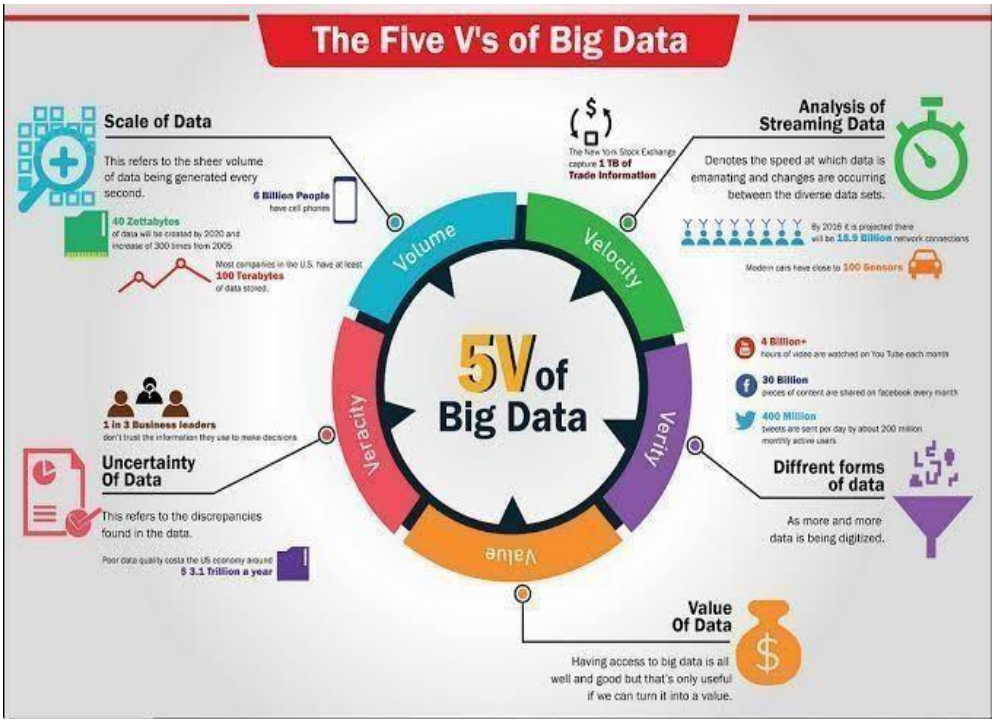

Fig.no 3.3 Five V's of Big data

The five Vs are volume, velocity, variety, veracity and value.

- Consider volume and velocity together because you will make infrastructure decisions about how to collect, store, and process data based on the combination of how much data you need to ingest and how quickly you will ingest it.
- Variety and veracity both relate to the data itself—what type of data is it and what's the quality of it. Data engineers and data scientists will transform and organize the data based on its variety and veracity to make it useful for analysis.
- Value is about ensuring that you are getting the most out of the data that you have collected. Value is also about ensuring that there is business value in the outputs

from all that collecting, storing, and processing.

- Volume and velocity requirements drive infrastructure decisions across all pipeline layers. Although they are different measures, you need to look at them together to determine how your pipeline needs to scale. You need to scale to handle ingestion and storage of the amount of data at the pace of its arrival. You also need to consider how long data should be stored to balance cost and availability.

- From the processing standpoint, you need to understand how much data must be processed and analyzed to address a singular business problem. You also need to know how quickly the data must be processed after it arrives in the pipeline. In terms of visualizing data, you need to understand how much data consumers need to access at one time and how frequently new data must be incorporated.

3.4 Matching ingestion services to variety, volume, and velocity:

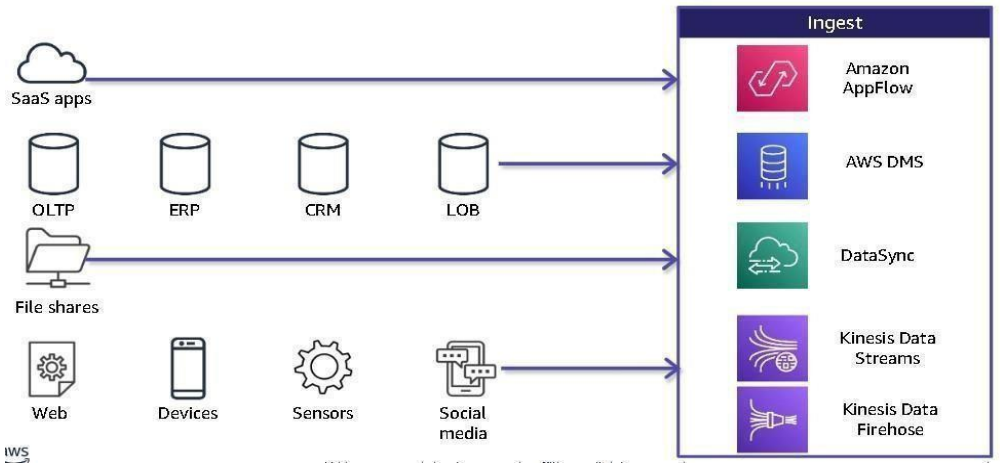

Fig No 3.4 Ingestion services to variety-volume-velocity

The ingestion layer uses individual purpose-built AWS services to match the unique connectivity, data format, data structure, and data velocity requirements of source types and to deliver them to the storage layer components. The services include the following:

• Amazon AppFlow can ingest from Software as a service (SaaS) application, such as Salesforce or Zendesk.

• AWS Database Migration Service (AWS DMS) can ingest from operational databases like online transaction processing (OLTP), enterprise resource planning (ERP), customer relationship management (CRM) and line of business (LOB) databases.

• AWS DataSync can ingest from file shares.

• Amazon Kinesis Data Streams and Amazon Kinesis Data Firehose can ingest from streaming data sources.

3.5 Pipeline Architecture

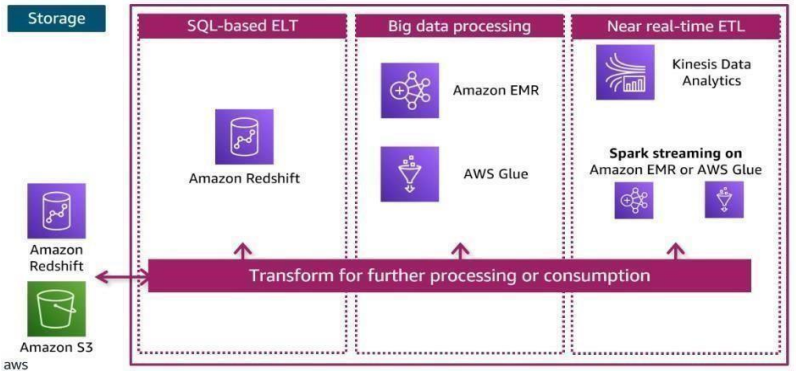

Fig No 3.5 Pipeline Architecture

Each pipeline reads data from the storage layer, processes it using temporary storage as needed, and then writes it to the appropriate location in the storage layer. The transformations are grouped into three main types, which are aligned to the use case:

• SQL-based processing using a data warehouse (in this case, Amazon RedShift)

• Big data processing using big data tools (in this case, Amazon EMR and AWS Glue)

• Near real-time processing using streaming (in this case, Amazon Kinesis Data Analytics or Spark streaming on Amazon EMR or AWS Glue) .

3.6 Stream processing pipeline

Streaming pipelines follow the same general layers as other pipelines, but there are some unique considerations. Data sources include clickstream logs, mobile apps,

existing application databases, or Internet of Things (IoT) sensors. You might want to respond to this data in real time or use it for analysis later Producers ingest records onto the stream. Producers are integrations that collect data from a source and load it onto the stream. Consumers process records. Consumers read data from the stream and perform their own processing on it. The stream itself provides a temporary but durable storage layer for the streaming solution.

In the pipeline that is depicted in this slide, Amazon CloudWatch Events is the producer that puts CloudWatch Events event data onto the stream. Kinesis Data Streams provides the storage. The data is then available to multiple consumers.

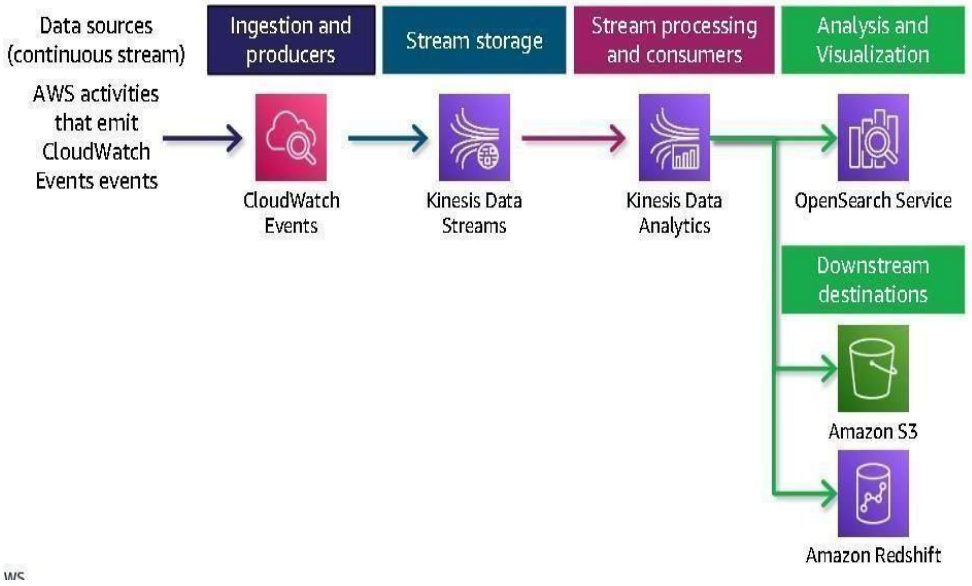

Fig No 3.6 Stream processing pipeline

With real-time streaming analytics, records on the stream are typically processed sequentially and incrementally by recording over sliding time windows. In the pipeline that is depicted on the slide, Kinesis Data Analytics is a consumer of the stream and processes streaming data by using custom applications or standard SQL. In this example, results are sent to OpenSearch Service, where they can be used to visualize real-time insights with OpenSearch Dashboards immediately.

In this scenario, Amazon S3 and Amazon Redshift also consume the data that Kinesis Data Analytics processes. These downstream destinations aren't being usedfor real-time analytics but could be used for serving applications such as one-time analytics and ML. This is an example of how the modern data architecture supportsthe goal of making ingested data available to let different consumers perform different types of analytics and run AI/ML applications.

3.7 Securing the stream processing pipeline

- Understand data classifications and their protection policies is the best practice will govern how you handle all the data that flows through your data pipeline. Reviewing organization's policies for classifying sensitive data and knowing what steps you will need to take to ensure adherence to those policies.
- Identify the source data owners and have them set the data classifications: The source data types that are listed in the operational data section of your pipeline will likely have various teams or individuals as owners. Identify who the dataset owners are, and request that they properly classify their datasets if they haven't already done so.
- Record data classifications into the Data Catalog so that analytics workloads can understand: Keep your Data Catalog up to date so that you have accurate and reliable records of data locations and classifications.
- Implement encryption policies for each class of data in the analytics workload: After you identify the source data that you will work with and have established the classification level of each dataset, you will need to implement the applicable encryption policies**.**
- For data at rest, use one of the multiple encryption options available in Amazon S3. Secure your Amazon Kinesis data streams with server-side encryption by using AWS KMS.
- Implement data retention policies for each class of data in the analytics workload: Use classification-based retention policies for your datasets. Back up and retain analytics datasets based on your organization's policies for classified data.

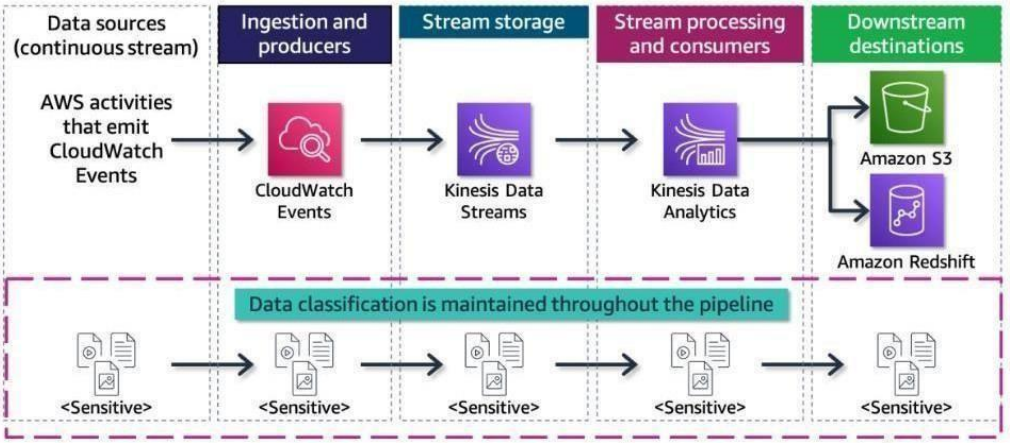

Fig No 3.7 Securing the stream processing pipeline

3.8 Processing of Big Data

In the Design Principles and Patterns for Data Pipelines module, you learned that components in the data processing layer are responsible for transforming data into a consumable state. The processing layer provides purpose - built components that enable a variety of data types, velocities, and transformations. Each component can read and write data to both Amazon S3 and Amazon Redshift in the storage layer, and all can scale to high data volumes.

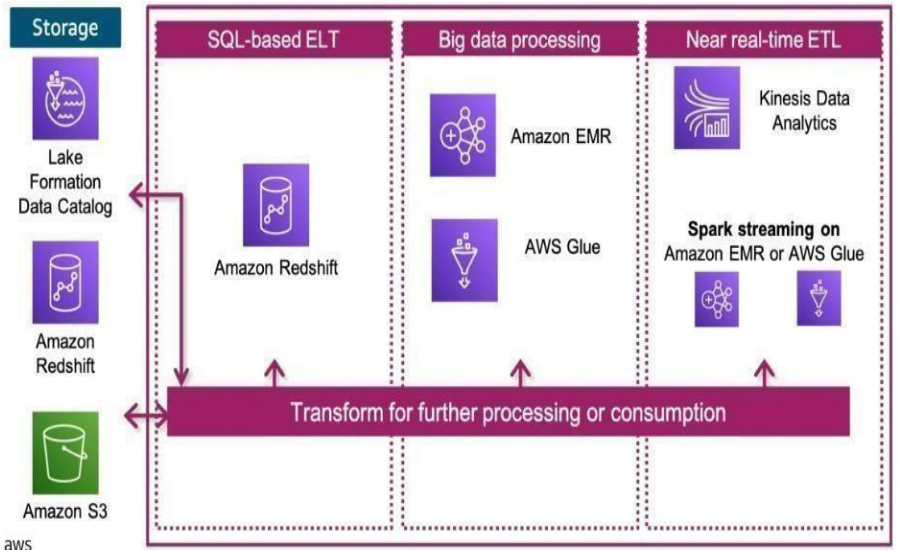

Fig No 3.8 Processing of Big Data

This architecture supports multiple ETL or ELT pipelines that perform iterative processing of data for different types of preparation and consumption. Each pipeline reads data from the storage layer, processes it using temporary storage as needed, and then writes it to the appropriate location within the storage layer.

Data stores evolved from handling only structured data to include options for handling high volumes of unstructured data. Specifically, there has been a move from structured data in data warehouses towards unstructured and semi structured data in data lakes. Tools and methods for ingesting data into a pipeline evolved along the same lines. ETL techniques have continued to evolve and are still a valuable part of data architectures for structured data.

ETL stores data that is ready to be analysed, so it can save time for an analyst. If analysts routinely perform the same transformations on data that they analyse, it might be more efficient to incorporate those transformations into the ingestion process. So, with ETL, you trade longer transformation processing time before the data is available for faster access. Another advantage of performing transformations up front is that you can filter out personally identifiable information (PII) or other sensitive data that could create a compliance risk. If you never store the sensitive data, you reduce the risk.

3.9 Processing Data for ML

This module provides a brief look at machine learning (ML) and focuses on the ML lifecycle, which guides how data engineers and data scientists collect and process data for ML models. The module takes a quick look at general ML conceptsand describes the activities that are performed in each phase of the ML lifecycle.

The Data-Driven Organizations module presents this contrast between the characteristics of traditional analytics and the field of AI/ML.

With traditional methods, developers or data scientists write programming logic to represent a set of rules, which are applied to data to get outcomes.

With machine learning, the rules are not hardcoded. ML practitioners build a mathematical function that finds patterns in historical data and uses the patterns to learn how to predict outcomes when given new data.

- An ML model is a function that combines a trained dataset and an algorithm to predict outcomes.
- The three general types of machine learning include supervised, unsupervised and reinforcement.
- Deep learning is a subcategory of machine learning that uses neural networks to develop models.
- Generative AI is a subcategory of deep learning that can generate content and is trained on large amounts of data.

3.10 Analyzing and Visualizing Data

This module offers a concise overview of machine learning (ML) with a focus on the ML lifecycle. It highlights the key stages that guide data engineers and data scientists in collecting, processing, and using data to build ML models. The module covers foundational ML concepts and provides a detailed explanation of the activities conducted in each phase of the ML lifecycle.

Data engineers focus on the infrastructure that the data passes through. This three-part framework will guide you through the questions that you need to ask and the factors that you should consider when selecting data analysis and visualization tools and services. To build the correct pipeline, it’s necessary to fully understand the desired outcomes based on business needs, the characteristics of the data, and who needs access to analyze and visualize data.

- When you select analysis and visualization tools, consider the business needs, data characteristics, and access to data.
- Consider the granularity and format of the insights based on business needs.
- Consider the volume, velocity, variety, veracity, and value of your data.
- Consider the functions of individuals who will access, analyze, and visualize the data.
- Athena provides interactive analysis using SQL, while Quick Sight is for dashboards and visualizations. OpenSearch Service is for operational analytics.
- With Athena, you can start querying data instantly and directly in Amazon S3. With Quick Sight, you can visualize data into insights quickly, and with OpenSearch Service, you can visualize data in near real time.
- Athena and Quick Sight are server less, and OpenSearch Service is a fully managed service.

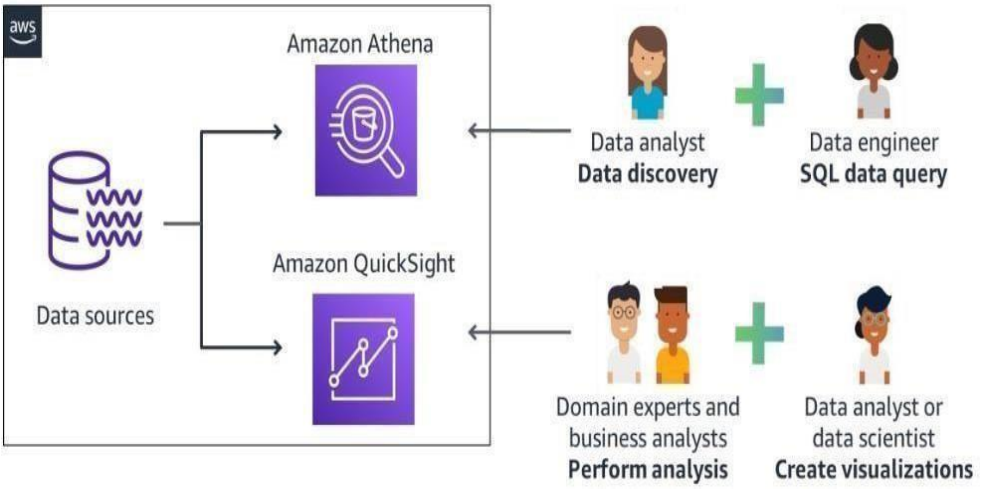

Fig No 3.10 Analyzing and Visualizing Data

Amazon Athena is an interactive query service that provides the ability to use SQL to analyze data in Amazon S3. Athena includes the following features:

- It is serverless
- Provides the ability to combine data from multiple data sources

- Can be used for one - time queries
- Can be used from your favorite business intelligence (BI) tools (such as QuickSight).
- Can update data stored in Amazon S3 with Apache Iceberg integration Amazon Athena is an interactive query service in which you can use standard SQL to query and analyze data.

3.11 Automating the Pipeline

This module discusses the benefits of automating your pipeline by using infrastructure as code and Step Functions. You will learn about flow states in Step Functions and how to use them to control and direct your workflows. You will also examine the elements of the Step Functions Workflow Studio.

In the past, the systems administrator would be manually notified when the application server crashed. The administrator then needed to manually launch a new server. Instead of following that traditional pattern, a best practice is to configure Amazon.

Cloud Watch to automatically detect a crash. When that crash is detected, the administrator is notified, and a new server with the same configuration is launched in parallel. All of these steps happen at the same time, without human intervention.

AWS offers built-in monitoring and automation tools at virtually every layer of your infrastructure. Take advantage of these resources to ensure that your infrastructure can respond quickly to changes. You can automate detection of unhealthy resources and launching replacement resources. You can even be notified when resources are changed. By removing manual processes, you can improve your system's stability and consistency, as well as the efficiency of the organization.

An elastic infrastructure can expand and contract as capacity needs change. Amazon EC2Auto Scaling automatically adds or removes EC2instances according to policies that you define, schedules, and health checks. Amazon EC2Auto Scaling provides several scaling options to best meet the needs of your applications. You

can use push-button scaling to vertically scale compute capacity for your RDS DB instance. You can use read replicas or shards to horizontally scale your RDS DB instance. Aurora Serverless scales resources automatically based on the minimum and maximum capacity specifications.

Route 53 offers a variety of routing options that can be combined with DNS failover to enable a variety of low-latency, fault-tolerant architectures. AWS Cost Explorer, AWS Budgets, AWS Cost and Usage Report, and the Cost Optimization.

CHAPTER-4

REAL TIME APPLICATIONS OF DATA ENGINEERING

Data engineering plays a crucial role in powering real-time applications across various industries. By designing, developing, and maintaining **data pipelines** that handle continuous data streams, data engineers enable immediate decision-making and operational efficiency. Real-time data engineering focuses on capturing, processing, and analyzing data as it is generated, ensuring that actionable insights are available instantly.

Instead of owning physical servers or data centers, companies can rent access to anything from applications to storage from a cloud service provider. Data Engineering has become an integral part of many industries, enabling businesses to operate more efficiently, scale rapidly, and innovate faster.

Below are some **key real-time applications** of data engineering:

Fraud Detection in Financial Services:

Real-time data pipelines analyze transactions as they occur to identify fraudulent activities. Data engineers build systems to detect anomalies using machine learning models and rule-based algorithms.

Real-Time Recommendation Systems:

E-commerce platforms (e.g., Amazon, Netflix) use real-time data processing to recommend products, movies, or content based on user behavior, preferences, and browsing patterns.

- **Enhance User Engagement**: By recommending relevant content or products in real time, users are more likely to stay on the platform longer, increasing overall engagement and reducing churn rates.
- **Dynamic Personalization**: Recommendations are updated instantly based on real-time interactions, such as adding items to the cart, watching a trailer, or liking a song. This helps provide a more personalized experience that adapts to the user's immediate actions.
- **Cross-Sell and Up-Sell Opportunities**: E-commerce platforms use real-

time recommendations to suggest complementary products (cross-sell) or premium versions (up-sell) based on a user's browsing or purchase history, driving increased revenue.

Netflix

Netflix is an entertainment platform that started in the United States, but eventually, it expanded to many countries and soon became popular. However, once Netflix confronted the scalability problem because of the sudden increase in viewers. That made Netflix choose AWS services. Netflix reports that when it started using AWS services like DynamoDB and Cassandrafor its distributed databases, it could handle the data easily. So, scalability is a great advantage of AWS. Netflix has adapted around 100,000 server instances from AWS for computing and storage databases, analytics, recommendation engines, and video transcoding as well.

Music Streaming and Personalization on Spotify:

Objective: Deliver high-quality music streaming to millions of users in real-time while providing personalized playlists, song recommendations, and ensuring seamless playback.

AWS Lambda to update user profiles and provide tailored recommendations. Music files are stored in Amazon S3 and delivered globally via Amazon CloudFront, ensuring low-latency streaming. Real-time data is continuously analyzed using Amazon DynamoDB and advanced machine learning models in Amazon SageMaker, enabling Spotify to adaptively adjust streaming quality and curate personalized playlists. Monitoring with Amazon CloudWatch ensures optimal performance, creating an engaging and uninterrupted listening experience for millions of users.

However, once Netflix confronted the scalability problem because of the sudden increase in viewers. That made Netflix choose AWS services. Spotify reports that when it started using AWS services like DynamoDB and Cassandra for its distributed databases, it could handle the data easily. So, scalability is a great advantage of AWS. Netflix has adapted around 100,000 server instances from AWS for computing and storage databases, analytics, recommendation engines, and video transcoding as well.

Healthcare Monitoring and Diagnostics:

Wearable devices and hospital monitoring systems collect patient data (e.g., heart rate, glucose levels) in real-time. Data pipelines process this data to provide early diagnostics and trigger alerts for medical professionals.

Smart Transportation Systems:

Real-time data from GPS devices, traffic cameras, and public transport systems enable smart transportation solutions. Data engineers build pipelines that process this information in real time to optimize routes, predict traffic patterns, and manage transportation services dynamically. Cities like Singapore and London utilize such systems for efficient traffic management and public transit operations.

Real-Time Weather Monitoring and Disaster Response:

Real-time data from satellites, weather stations, and IoT sensors is processed to provide accurate and timely weather forecasts. These data pipelines are critical for disaster response systems that monitor hurricanes, floods, and other natural disasters, enabling early warnings and coordinated emergency responses.

Real-time data engineering empowers organizations to respond proactively to ever-changing conditions, optimize operations in real-time, and personalize services at scale, driving competitive advantages across industries. It not only fuels immediate decision-making but also enables continuous improvement, ensuring businesses stay agile and responsive in today's fast-paced world.

CHAPTER 5
LEARNING OUTCOMES OF INTERNSHIP

- Gain a deep understanding of cloud computing principles, including its benefits, deployment models (public, private, hybrid), and service models (IaaS, PaaS, SaaS).
- Understand how to create, configure, and manage AWS resources using the AWS Management Console .
- Learn how to design and implement cloud architectures that are scalable, resilient, and cost-effective, following AWS best practices.
- **Data Engineering Concepts**: Developed a solid understanding of core data engineering concepts like ETL processes, big data processing, and real-time data streaming, using AWS services such as AWS Glue, Amazon Kinesis, and Amazon EMR.
- **Hands-on Experience with AWS Tools**: Acquired practical experience in deploying and managing cloud-based data pipelines, from data ingestion to transformation, using services like Amazon Redshift, Amazon Athena, and AWS Lambda for automation.
- Understood how AWS integrates with machine learning tools, facilitating data preparation, transformation, and training using Amazon SageMaker, and applied these concepts to real-world datasets.
- Explore emerging technologies in the cloud space, such as artificial intelligence (AI), machine learning (ML), Internet of Things (IoT), and edge computing, and how they integrate with AWS services.

CHAPTER 6
CONCLUSION

In conclusion, AWS Data Engineering Virtual internship has been a transformative experience, offering not only theoretical insights but also hands-on practice with industry-standard tools and technologies. Over the course of the internship, I gained a deep understanding of data engineering fundamentals, particularly in leveraging AWS services like Amazon S3, Redshift, Glue, and Lambda to create, maintain, and optimize data pipelines.

One of the major highlights of the internship was the opportunity to work with real-world datasets, which enhanced my understanding of the complexities involved in data ingestion, transformation, and storage at scale. Through projects like building ETL pipelines and designing efficient data architectures, I learned how to optimize workflows for performance, reliability, and cost-effectiveness.

By the end of the internship, I had significantly improved my technical proficiency with AWS and cloud computing, developed a more strategic approach to data engineering, and gained confidence in handling large, complex datasets. This experience has been instrumental in solidifying my interest in data engineering and has equipped me with the skills and mindset needed to excel in future projects and roles within the field. It equipped me with the skills and confidence to contribute meaningfully to future data-driven projects and provided a strong foundation for continued learning and growth in the field of cloud computing and data engineering.

INTERNSHIP CERTIFICATE

REFERENCES

- https://awsacademy.instructure.com/courses
- https://aws.amazon.com/training/awsacademy/
- https://www.udemy.com/course/data-engineering-using-aws-analytics-services/
- https://github.com/pablosalme/aws-data-engineering