**DATA ANALYTICS USING AWS S3 AND AMAZON QUICKSIGHT**

by

B. Preethi

(2443012)

Under the guidance of

Dr. Deepa S

****

A Project report submitted in partial fulfillment of the requirements

for the award of the degree of Bachelor of Computer Applications of

CHRIST (Deemed to be University)

May – 2025

**CERTIFICATE**

This is to certify that the report titled **DATA ANALYTICS USING AWS S3 AND AMAZON QUICKSIGHT** is a Bonafide record of work done by **B. Preethi (2443012)** of CHRIST (Deemed to be University), Bengaluru, in partial fulfillment of the requirements of II Semester Bachelor of Computer Applications during the academic year 2024-25

**Head of the Department Project Guide Valued-by**

1. Name :

Register Number :

2. Date of Exam :

Table of Contents

[Acknowledgement i](#_Toc200388903)

[Completion Certificate ii](#_Toc200388904)

[Abstract iii](#_Toc200388905)

[About Shellkode 1](#_Toc200388906)

[AWS Educate: EDINCOv1EN-US 2](#_Toc200388907)

[AWS Data Analytics: 7](#_Toc200388908)

[About Aws Educate: 8](#_Toc200388909)

[1.INTRODUCTION TO CLOUD COMPUTING 9](#_Toc200388910)

[1.1 What is cloud computing? 9](#_Toc200388911)

[1.2 Differences between traditional computing and cloud computing. 9](#_Toc200388912)

[1.3 Introduction to AWS (Amazon Web Services) 10](#_Toc200388913)

[2. AWS GLOBALINFRASTRUCTURE 11](#_Toc200388914)

[2.1 AWS infrastructure features: 11](#_Toc200388915)

[3. AWS CLOUD SECURITY 12](#_Toc200388916)

[3.1 AWS cloud security: 12](#_Toc200388917)

[3.2 IAM: 12](#_Toc200388918)

[4. NETWORKING AND CONTENT DELIVERY 13](#_Toc200388919)

[4.1 Networking basics 13](#_Toc200388920)

[4.2 Amazon VPC: 13](#_Toc200388921)

[5. COMPUTE 14](#_Toc200388922)

[5.1 Amazon Ec2 14](#_Toc200388923)

[6. STORAGE 14](#_Toc200388924)

[6.1 Amazon S3: 14](#_Toc200388925)

[7. Databases 15](#_Toc200388926)

[7.1 Relational database service: 15](#_Toc200388927)

[8. AUTOMATIC SCALINGAND MONITORING 15](#_Toc200388928)

[8.1 Amazon cloud watch: 15](#_Toc200388929)

[8.2 Amazon ec2 auto scaling: 15](#_Toc200388930)

[9. Introduction to data Engineering 15](#_Toc200388931)

[10. Analyzing and Visualizing data 17](#_Toc200388932)

[10.1 The Role of Data Analysis and Visualization: 17](#_Toc200388933)

[10.2 Strategies for Data Analysis: 18](#_Toc200388934)

[10.3 Strategies for Data Visualization: 18](#_Toc200388935)

[LABS 19](#_Toc200388936)

[Lab 1: Store data in Amazon S3 19](#_Toc200388937)

[Lab 2: Query Data in Amazon Anthena 20](#_Toc200388938)

[Lab 3: Creating an AWS Glue crawler: 21](#_Toc200388939)

[Lab 4: Visualize Data with Amazon QuickSight 22](#_Toc200388940)

[Lab 5: Automate Loading Data with the AWS Data Pipeline 23](#_Toc200388941)

[Work Samples-1 24](#_Toc200388942)

[Project Title 1: Data Visualization using AWS S3 and Amazon QuickSight 24](#_Toc200388943)

[Work Samples-2 27](#_Toc200388944)

[Project Title 2: Visualize Data with Amazon QuickSight 27](#_Toc200388945)

[Conclusion 31](#_Toc200388946)

[Appendices 32](#_Toc200388947)

[Appendix A – Screenshots of AWS Setup 32](#_Toc200388948)

[Appendix B – Sample Manifest File 34](#_Toc200388949)

[Appendix C – Data Sample (CSV Snippet) 35](#_Toc200388950)

[Appendix D – Visualization Insights 35](#_Toc200388951)

[Appendix E – QuickSight Account Setup Notes 35](#_Toc200388952)

[Appendix F – AWS IAM Role and Permission Settings 36](#_Toc200388953)

[Appendix G – Troubleshooting and Errors Encountered 36](#_Toc200388954)

[Appendix H – Glossary of Technical Terms 37](#_Toc200388955)

[Appendix I – Learning Resources and Tutorials Followed 37](#_Toc200388956)

[References 38](#_Toc200388957)

# Acknowledgement

I take this opportunity to express our deepest gratitude and appreciation to all those people who made this Internship work easier with words of encouragement, motivation, discipline, and faith by offering different places to look to expand my ideas and help me towards the successful completion of this Internship work.

First and foremost, we express our deep gratitude to Dr. Vinay M, HOD, Christ Deemed to be University for providing necessary facilities throughout the BCA program.

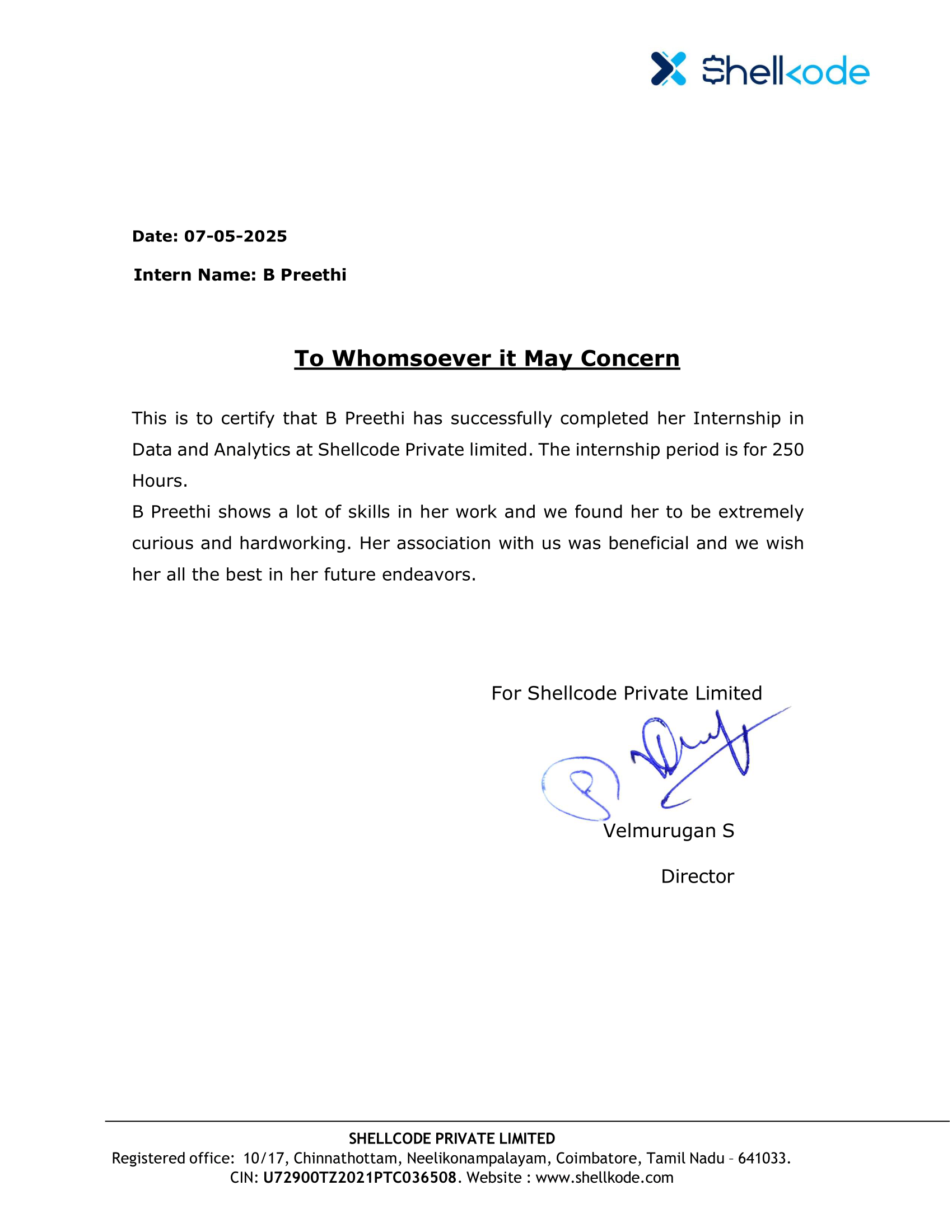
I express our sincere thanks to Dr. Deepa S, Project Guide, for her constant support and cooperation throughout the BCA program.

We express our sincere gratitude to Dr. Teena Jose, Program Coordinator, BCA, for her constant encouragement, motivation and faith by offering different places to look to expand my ideas. I would like to express our sincere gratitude to our CTO Mr. Bhuvanesh and our Internship Coordinators Mr. Prajwal & Ms. Manasi for their insightful advice, motivating suggestions, invaluable guidance, help and support in successful completion of this Internship.

We would like to take this opportunity to express our thanks to the teaching and nonteaching staff in the Department of Compute Science, Christ for their invaluable help and support.

B. Preethi

# Completion Certificate



# Abstract

The objective of this **Data Analytics Internship** provides the strong foundation about AWS fundamentals which offers common cloud concepts, Core AWS services, AWS platform, including available services and their common use cases, AWS Cloud architectural principles (at the conceptual level), account security, and compliance through Cloud Foundations.

Data engineering is the science of analysing raw data to make conclusions about that information. Data engineering relies on a variety of software tools ranging from spreadsheets, data visualization, and reporting tools, data mining programs, or opensource 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 engineering 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 engineering, and here comes the problem of storing the data for that we have learnt different tools used for data storage and how to analyse 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 QuickSight.

# About Shellkode

Shellkode is a born in the cloud company with multiple service delivery locations spread across the globe, we help organizations transform their legacy IT environment to cloud with emphasis on data and cloud-native solutions. Our aim is to enable our customers to maximize insights and make intelligent predictions by leveraging the power of cloud and data. Through our consulting, we help organizations become agile and up-to-date with the changing market landscape, thus maximizing their Return on Investment(ROI).  
  
At Shellkode, we are passionate about cloud, data, and its implication and we strongly believe that it has the power to revolutionize the way organizations do business in our current landscape. Our approach integrates emerging technologies, to build, and implement dynamic, and scalable enterprise-grade solutions that can be used to resolve some of the most complex IT challenges.

# AWS Educate: [EDINCOv1EN-US](https://awseducate.instructure.com/courses/909)

|  |  |
| --- | --- |
| **Module1** | **Introduction to AWS management console**   * Describe the purpose of the AWS Management Console. * Navigate the features and functions of the AWS Management Console. * Discuss the structure of the AWS global infrastructure. * Identify AWS payment models and service cost factors. |
| **Module2** | **Introduction to Cloud 101**   * Define cloud computing and describe its benefits * Compare and contrast cloud computing service and deployment models * Identify the AWS Global Infrastructure * Discuss the shared responsibility model * Describe the Well-Architected Framework and its role in building flexible and reliable architectures * Identify the use, features, and benefits of Amazon Web Services (AWS) core services * Practice using AWS core services in a lab environment * Describe four entry-level cloud computing careers |
| **Module3** | **Introduction to Cloud Computing**   * Define cloud computing and describe its benefits * Compare and contrast cloud computing service and deployment models * Identify the AWS Global Infrastructure * Discuss the shared responsibility model * Describe the Well-Architected Framework and its role in building flexible and reliable architectures * Identify the use, features, and benefits of Amazon Web Services (AWS) core services * Practice using AWS core services in a lab environment * Describe four entry-level cloud computing careers |
| **Module4** | **Introduction to AWS**   * Define cloud computing and describe its benefits * Compare and contrast cloud computing service and deployment models * Identify the AWS Global Infrastructure * Discuss the shared responsibility model * Describe the Well-Architected Framework and its role in building flexible and reliable architectures * Identify the use, features, and benefits of Amazon Web Services (AWS) core services * Practice using AWS core services in a lab environment * Describe four entry-level cloud computing careers |
| **Module5** | **AWS Core Services**   * Introduction to Amazon Simple Storage Service (Amazon S3) * [Introduction to Amazon Elastic Compute Cloud (Amazon EC2)](https://awseducate.instructure.com/courses/891/modules/items/15086) * [Introduction to Amazon Amazon Virtual Private Cloud (Amazon VPC)](https://awseducate.instructure.com/courses/891/modules/items/15091) * [Introduction to Amazon Relational Database Service (Amazon RDS)](https://awseducate.instructure.com/courses/891/modules/items/15089) * Introduction to AWS Identity and Access Management (IAM) * [Introduction to AWS Lambda](https://awseducate.instructure.com/courses/891/modules/items/15088) * Introduction to Amazon CloudWatch |
| **Module6** | **Getting Started with Storage**   * Discuss different types of storage solutions and their features and benefits. * Discuss the features and concepts of Amazon S3. * Describe Amazon S3 storage classes and associated use cases. * Discuss how to use Amazon S3 to create a bucket, upload objects, and work with objects. * Describe Amazon S3 configurations for cost savings and security. * Identify other AWS storage solutions and their use cases. * Use Amazon S3 to create a static website. |

|  |  |
| --- | --- |
| **Module7** | **Getting Started with Compute**   * Discuss different types of compute solutions, their features and benefits * Discuss the basic features and concepts of Amazon EC2 * Describe Amazon EC2 instance types and how to choose an instance type * Describe how to use Amazon EC2 to launch and configure an instance * Describe how to manage Amazon EC2 instances * Use Amazon EC2 to launch and manage and instance |
| **Module8** | **Introduction to Data Analytics in Sports**   * Working with Big Data Using Athena Simulation. * Investigate the different types of data visualization in gaining business insights in the sports industry * Use data models and visualization tools to support decision making. |
| **Module9** | **Visualization of Sports Big Data**   * Creating a Data Dashboard Using Amazon QuickSight Simulation * Investigate the different types of data visualization in gaining business insights in the sports industry * Use data models and visualization tools to support decision making |

# AWS Data Analytics:

|  |  |
| --- | --- |
| **Module 1** | **Introduction**  Big Data  • Big Data Pipeline  • Big Data Tools  • Big Data Collection  • Big Data Storage  • Big Data Ingestion  • Big Data Processing and Analysis  • Big Data Visualization |
| **Module 2** | **Lab 1**   * Lab 1 Introduction * Store Data in Amazon S3 |
| **Module 3** | **Lab 2**   * Lab 2 Introduction * Query Data in Amazon Athena |
| **Module 4** | **Lab 3**   * Lab 3 Introduction * Query data in Amazon S3 with Amazon Athena and AWS Glue |
| **Module 5** | **Lab 4**   * Lab 4 Introduction * Analyse Data with Amazon Quicksight |
| **Module 6** | **Lab 5**   * Lab 5 Introduction * Automate Loading Data with the AWS Data Pipeline |

# About Aws Educate:

AWS Educate is open to any individual, regardless of where they are in their education, technical experience, or career journey. Simplify your journey into cloud computing and AI with AWS Educate - your free, curated learning path to in-demand tech skills. AWS Educate offers an extensive library of free, self-paced online training that covers a range of topics - from cloud fundamentals to innovative technologies like Artificial Intelligence, machine learning, and more.



# 1.INTRODUCTION TO CLOUD COMPUTING

## What is 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 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:

* IaaS
* PaaS
* SaaS

## Differences between traditional computing and cloud computing.

Traditional Computing model

1. Infrastructure as hardware

2. Hardware Solutions:

1. Require Space, Staff, physical security, planning, Capital expenditure.
2. Have a long hardware procurement cycle
3. Require you to provision capacity by guessing theoretical maximum peaks

Cloud computing model

* Infrastructure as software
* Software solutions:

1. Are flexible
2. Can change more quickly, easily, and cost-effectively than hardware solutions
3. Eliminate the undifferentiated heavy-lifting tasks

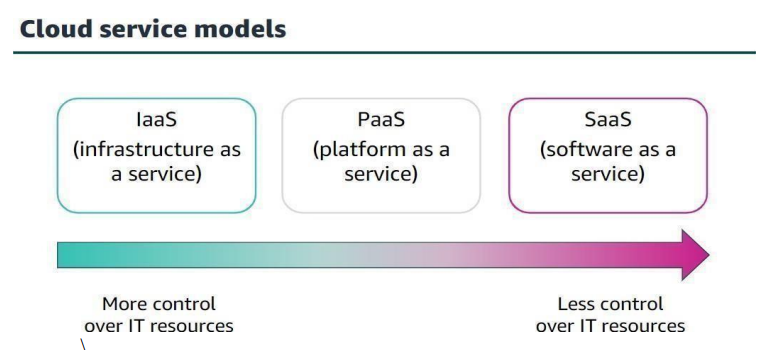


Figure 1.2.1 Cloud Service Models

Advantages of Cloud computing

* Trade capital expense for variable expense
* Benefit from massive economies of scale
* Stop guessing capacity
* Increase speed and agility Go global in minutes
* Stop spending money on running and maintaining data centers

## Introduction to AWS (Amazon Web Services)

Amazon Web Services (AWS) is a secure cloud platform that offers a broad set of global cloud based products. Because these products are delivered over the internet, you have on-demand access to the compute, storage, network, database, and other IT resources that you might need for your projects— and the tools to manage them. AWS offers flexibility. Your AWS environment can be reconfigured and updated on demand, scaled up or down automatically to meet usage patterns and optimize spending, or shut down temporarily or permanently.

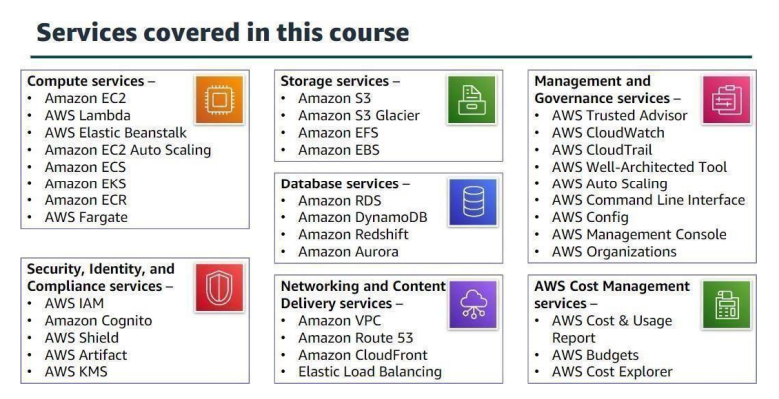
****

Figure 1.3.1 Services covered in the course

# 2. **AWS GLOBALINFRASTRUCTURE**

The AWS Global Infrastructure is designed and built to deliver a flexible, reliable, scalable, and Secure cloud computing environment with high-quality global network performance. The AWS Cloud infrastructure is built around regions.

## 2.1 **AWS infrastructure features:**

**Elasticity and scalability**

* Elastic infrastructure; dynamic adaption of capacity
* Scalable infrastructure; adapts to accommodate growth

**Fault-tolerance**

* Continues operating properly in the presence of a failure
* Built-in redundancy of components

**High availability**

* High level of operational performance
* Minimized downtime
* No human intervention

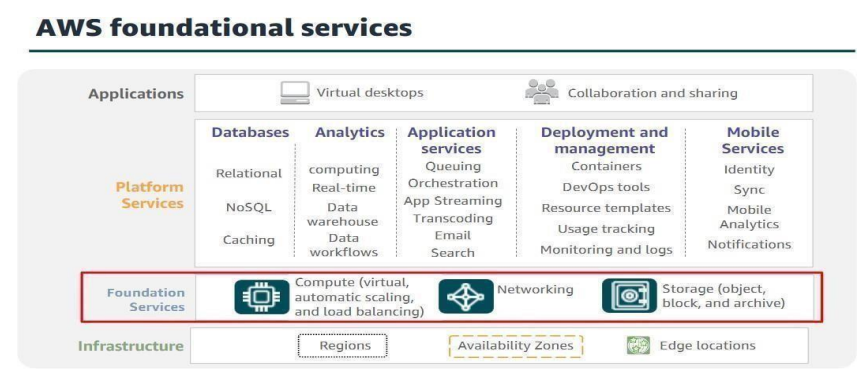


Figure 2.1.1 AWS foundational services

# 3. AWS CLOUD SECURITY

## 3.1 **AWS cloud security:**

Cloud security is a collection of security measures designed to protect cloud-based infrastructure applications and data. AWS provide security and services that help you protect your data,accounts, workloads form unauthorized access.

## 3.2 IAM:

With AWS identity and access management (IAM) you can specify who or what can access services and resources in AWS. IAM is a web service that helps securely control access to AWS resources. We use IAM control who is authenticated and authorized to use resources.

**There are two types of IAM policies:**

1. AWS managed policies
2. Customer managed policies

# 

# 4. NETWORKING AND CONTENT DELIVERY

## 4.1 Networking basics

A computer network is two or more client machines that are connected together to share resources. A network can be logically partitioned into subnets. Networking requires a networking device (such as a router or switch) to connect all the clients together and enable communication between them. Each client machine in a network has a unique Internet Protocol (IP) address that identifies it. A 32bit IP address is called an IPv4 address. A 128-bit IP address is called an IPv6 address.

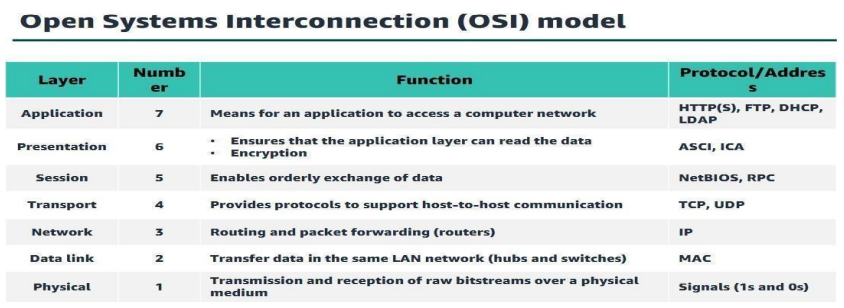


Figure 4.1.1 OSI model

## 

## 4.2 Amazon VPC:

VPC Amazon Virtual Private Cloud (Amazon VPC) is a service that lets you provision a logically isolated section of the AWS Cloud (called a virtual private cloud, or VPC) where you can launch your AWS resources.

# 

# 5. COMPUTE

## 5.1 Amazon Ec2

Amazon Elastic Compute Cloud (Amazon EC2):

* Provides virtual machines—referred to as EC2 instances—in the cloud.
* Gives you full control over the guest operating system (Windows or Linux) on each instance.
* You can launch instances of any size into an Availability Zone anywhere in the world.
* Launch instancesfrom Amazon Machine Images (AMIs).
* Launch instances with a few clicks or a line of code, and they are ready in minutes.
* You can control traffic to and from instances.

# 6. STORAGE

## 6.1 Amazon S3:

Amazon S3 is object storage that is built to store and retrieve any amount of data from anywhere: websites and mobile apps, corporate applications, and data from Internet of Things (IoT) sensors or devices. Amazon S3 is object - level storage, which means that if you want to change a part of a file, you must make the change and then re-upload the entire modified file. Amazon S3 stores data as objects within resources that are called buckets. The data that you store in Amazon S3 is not associated with any particular server, and you do not need manage any infrastructure yourself. You can put as many objects into Amazon S3 as you want. Amazon S3 holds trillions of objects and regularly peaks at millions of requests per second

# 7. Databases

## 7.1 Relational database service:

Amazon relational database service (Amazon RDS) is a collection of managed services that make it simple to set up, operate, and scale databases in the cloud. To address the challenges of running an unmanaged, standalone relational database, AWS provides a service that sets up, operates, and scales the relational database without any ongoing administration.

# 8. AUTOMATIC SCALINGAND MONITORING

## 8.1 Amazon cloud watch:

Amazon CloudWatch helps you monitor your AWS resources and the applications that you run on AWS in real time.

## 8.2 Amazon ec2 auto scaling:

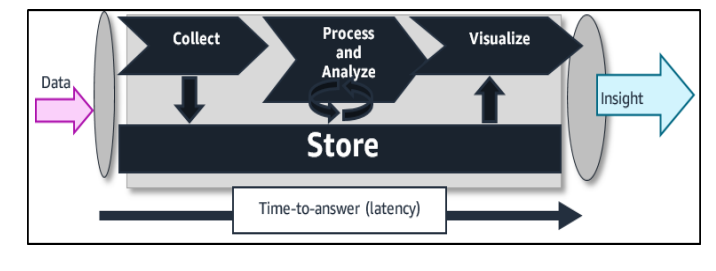
Scaling is the ability to increase or decrease the compute capacity of your application Amazon Ec2 Auto Scaling

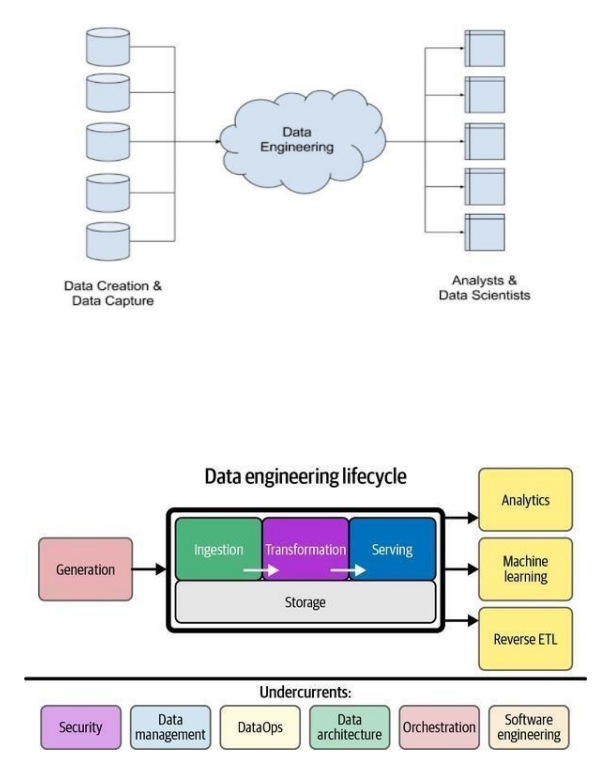
# 9. Introduction to data Engineering

Data engineering is the process of designing and building systems that let people collect and analyze raw data from multiple sources and formats. These systems empower people to find practical applications of the data, which businesses can use to thrive. Data engineering is a skill that is in increasing demand. Data engineers are the people who design the system that unifies data and can help you navigate it. Data engineers perform many different tasks including:

* Acquisition: Finding all the different data sets around the business
* Cleansing: Finding and cleaning any errors in the data
* Conversion: Giving all the data a common format
* Disambiguation: Interpreting data that could be interpreted in multiple ways
* Deduplication: Removing duplicate copies of data Once this is done, data may be stored in a central repository such as a data lake or data lakehouse.

Data engineers may also copy and move subsets of data into a data warehouse.





# 10. Analyzing and Visualizing data

10.1 The Role of Data Analysis and Visualization:

Data analysis and visualization encompass the techniques and technologies involved in exploring, summarizing, and communicating insights from raw data in a visual and intuitive manner. At its core, data analysis serves several critical functions, including descriptive analysis, diagnostic analysis, predictive analysis, and prescriptive analysis. These analyses enable organizations to uncover patterns, trends, anomalies, and relationships within their data, providing a foundation for informed decision-making and strategic planning.

## 10.2 Strategies for Data Analysis:

**1. Descriptive Analysis:** Descriptive analysis involves summarizing and aggregating data to provide a high-level overview of key metrics, trends, and distributions. This may include summary statistics, frequency distributions, histograms, or heatmaps, depending on the nature of the data and the specific analysis objectives.

**2. Diagnostic Analysis:** Diagnostic analysis focuses on understanding the root causes of observed patterns or anomalies within the data. This may involve hypothesis testing, correlation analysis, regression analysis, or causal inference techniques to identify relationships and dependencies between variables.

## 10.3 Strategies for Data Visualization:

**1. Charts and Graphs:** Charts and graphs are powerful tools for visualizing patterns, trends, and relationships within the data. This may include bar charts, line charts, scatter plots, pie charts, or box plots, each offering unique advantages for representing different types of data and analysis objectives.

**2. Dashboards:** Dashboards provide a centralized and interactive interface for visualizing and exploring data in real-time. This may include interactive charts, tables, maps, or widgets, enabling users to drill down into specific data subsets, filter data based on criteria, and gain deeper insights into key metrics and KPIs.

# LABS

## Lab 1: Store data in Amazon S3

This module presents an introduction to Amazon Simple Storage Service (Amazon S3). Companies need the ability to simply and securely collect, store, and analyze their data on a massive scale. Amazon S3 is object storage that is built to store and retrieve any amount of data from anywhere: websites and mobile apps, corporate applications, and data from Internet of Things (IoT) sensors or devices.

**Amazon S3 overview :**

* Data is stored as objects in buckets
* Virtually unlimited storage
* Single object is limited to 5 TB
* Designed for 11 9s of durability
* Granular access to bucket and objects

**Amazon S3 storage classes :**

Amazon S3 offers a range of object-level storage classes that are designed for different use cases:

* Amazon S3 Standard
* Amazon S3 Intelligent-Tiering
* Amazon S3 Standard-Infrequent Access (Amazon S3 Standard-IA)
* Amazon S3 One Zone-Infrequent Access (Amazon S3 One Zone-IA)
* Amazon S3 Glacier
* Amazon S3 Glacier Deep Archive

In this lab, Amazon S3 is used throughout the course, you must know how to create Amazon S3 buckets and load data for subsequent labs. using the AWS management console to create an Amazon S3 bucket, add an IAM user to a group that has full access to the Amazon S3 service, upload files to Amazon S3, and run simple queries on the data in Amazon S3.

## Lab 2: Query Data in Amazon Anthena

Lab 2 introduces you to Amazon Athena, which is the first analysis service can use Amazon Athena to query structured, unstructured, and semi-structured data. Amazon Athena integrates with AWS Glue, In this lab you will practice using the AWS management console to create an Amazon S3 bucket, add an IAM user to a group that has full access to the Amazon S3 service, upload files to Amazon S3, and run simple queries on the data in Amazon S3.

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.

**Benefits:**

* Start querying instantly
* Pay per query
* Open, powerful, standard
* Fast , accurate

****

Figure *2.1* Amazon Athena

## Lab 3: Creating an AWS Glue crawler:

Lab 3 introduces you to AWS Glue. Lab 3 builds on that idea to show how to use AWS Glue to infer the schema from the data. This lab includes:

* Access AWS Glue in the AWS Management Console.
* Create a crawler with AWS Glue.
* Create a database and table with AWS Glue.
* Query data in Amazon S3 with Amazon Athena and AWS Glue.

You can create a crawler by starting in the Athena console and then using the AWS Glue console in an integrated way. When you create the crawler, you specify a data location in Amazon S3 to crawl.

**To create a crawler in AWS Glue starting from the Athena console**

1. Open the Athena console at [https://console.aws.amazon.com/athena/](https://console.aws.amazon.com/athena/home).
2. In the query editor, next to **Tables and views**, choose **Create**, and then choose **AWS Glue crawler**.
3. On the **AWS Glue** console **Add crawler** page, follow the steps to create a crawler. For more information, see [Using AWS Glue Crawlers](https://docs.aws.amazon.com/athena/latest/ug/glue-best-practices.html#schema-crawlers) in this guide and [Populating the AWS Glue Data Catalog](https://docs.aws.amazon.com/glue/latest/dg/populate-data-catalog.html) in the AWS Glue Developer Guide.

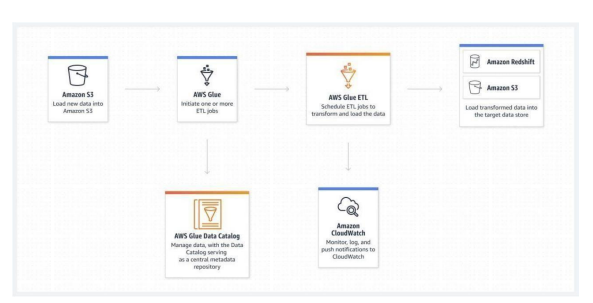
****

Figure *3*.1 Amazon Glue

## Lab 4: Visualize Data with Amazon QuickSight

Lab 4 introduces you to Amazon QuickSight, a cloud-powered business intelligence (BI) service. This lab addresses the Variety and Velocity aspects of big data problems by helping you quickly visualize insights from diverse data sources in real time. QuickSight is designed to scale automatically and enables users to create and publish interactive dashboards that can be accessed from any device.

Although the lab does not focus deeply on advanced dashboard design or embedding options, it helps you understand the overall architecture and integration of Amazon QuickSight with other AWS services like Amazon S3, Amazon Redshift, and RDS. You should also review the supported data source connectors and sharing capabilities to fully understand its integration capabilities. An architectural diagram of a basic QuickSight setup is included in the lab instructions, with additional AWS documentation provided for extended learning.

Amazon QuickSight is a fast, serverless, and scalable BI service that makes it easy to deliver insights to everyone in your organization. It allows you to build visualizations, perform ad hoc analysis, and get business insights from your data with no infrastructure management.

**Amazon QuickSight Features:**

* Fully managed, serverless BI service with no need to provision or manage servers
* Seamlessly integrates with AWS data sources like S3, Redshift, RDS, Athena, and more
* Supports a wide range of visualizations, including bar, line, pie, donut, maps, and word clouds
* SPICE engine for super-fast performance with in-memory optimized data analysis
* Built-in sharing and embedding options for dashboards across applications and portals
* Secure access control with AWS IAM integration and row-level security
* AI/ML-powered insights using ML-powered anomaly detection and forecasting
* Customizable dashboards with filters, calculated fields, and drill-down capabilities
* Pay-per-session pricing model helps reduce BI costs significantly

## Lab 5: Automate Loading Data with the AWS Data Pipeline

Lab 5 introduces you to the AWS Data Pipeline. The AWS Data Pipeline is a web service you can use to migrate and transform data. The main purpose of this lab is to provide learners with an opportunity to automate moving data and to understand how this service fits into the larger context of data analysis.

At the end of this module, learners will be able to:

* Access AWS Data Pipeline in the AWS Management Console.
* Create a data pipeline.
* Load data from Amazon S3 into Amazon Redshift with a data pipeline.
* Troubleshoot a data pipeline.
* Export data from Amazon Redshift to a Jupyter notebook.

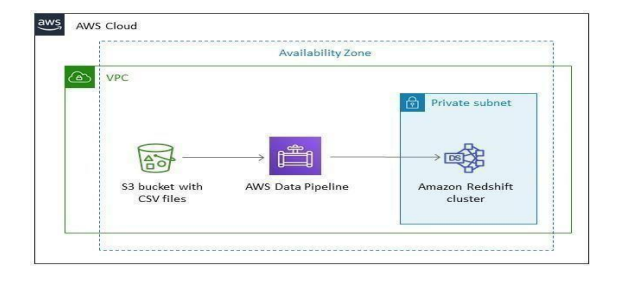
****

Figure 5.1 Loading Data

# Work Samples-1

## Project Title 1: Data Visualization using AWS S3 and Amazon QuickSight

**Overview**

As part of my internship, I completed a mini project focused on data visualization using Amazon Web Services (AWS). This project was inspired by a hands-on tutorial that demonstrates the integration of **Amazon S3** and **Amazon QuickSight** to analyse a large dataset. The dataset used in this project consists of **50,000 best-selling products on Amazon.com**, provided by **Bright Data**, a leading web data provider.

**Project Objectives**

* To strengthen foundational AWS skills through practical, real-world application.
* To understand how cloud services like Amazon S3 and QuickSight can work together for data analytics.
* To build interactive dashboards that offer insights into e-commerce trends.

**Tools and Technologies Used**

* **Amazon S3** – For securely storing the dataset files.
* **Amazon QuickSight** – For creating visual dashboards and performing data analysis.
* **CSV and JSON** – File formats used for the dataset and the manifest file.

**Steps Followed**

1. **Dataset Preparation**: Downloaded the amazonbestseller.csv and manifest.json files from the instructor’s GitHub repository
2. **Data Upload**: Uploaded both files to a newly created S3 bucket.

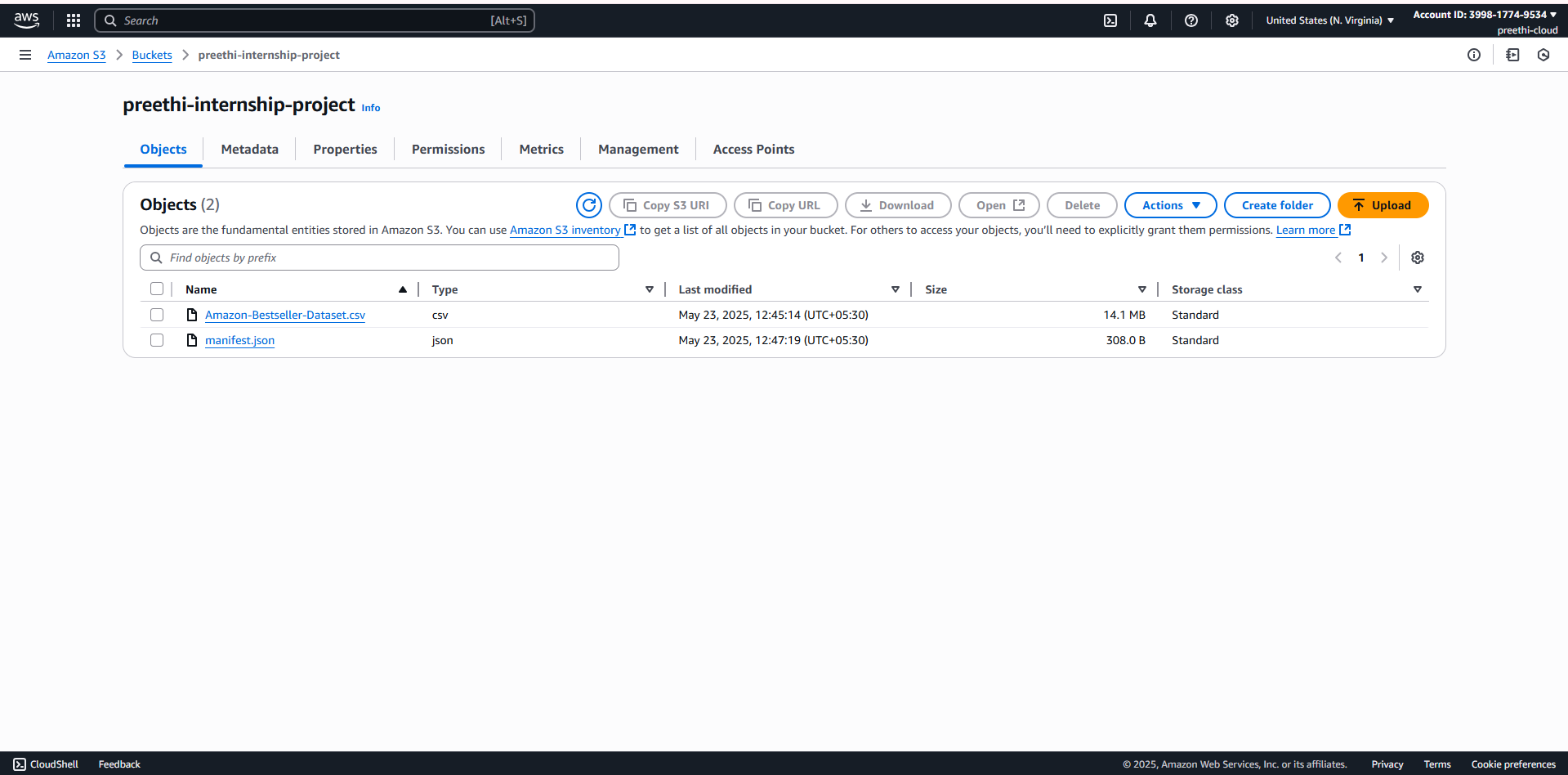


Figure 2.1: Data upload

1. **QuickSight Setup**: Configured a new QuickSight account, linked it to the S3 bucket using the manifest file.

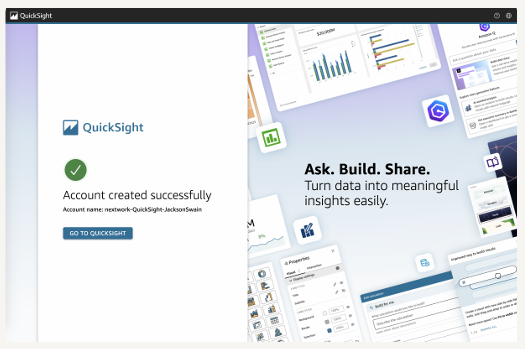


Figure 3.1: QuickSight setup

1. **Data Visualization**: Imported the dataset into QuickSight and created various visualizations including:
   * Most popular brands based on product frequency.
   * Charts comparing categories, prices, availability, and seller distribution.

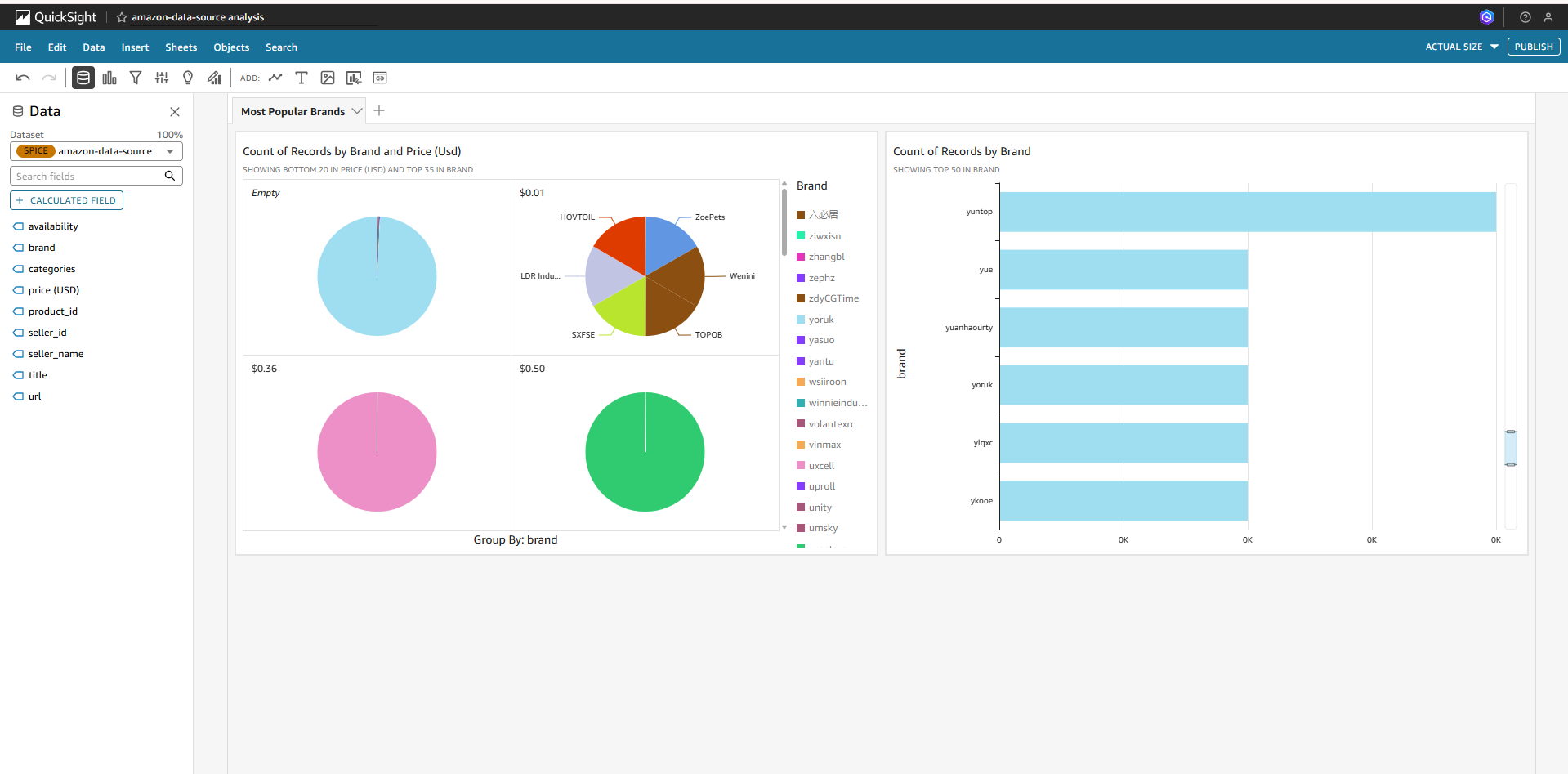


Figure : Data Visualization

**Key Learnings**

* Practical understanding of **Amazon S3 bucket creation**, object uploading, and permission settings.
* Configuring and using **Amazon QuickSight** for visual data exploration.
* Working with **real-world e-commerce data** to generate meaningful business insights.
* Experimenting with different **chart types** (bar, pie, donut, word cloud) for better storytelling through data.

**Conclusion**

This AWS mini project served as a valuable opportunity to apply cloud and analytics tools in a real-world scenario. It helped enhance my technical skills and demonstrated how large datasets can be processed and visualized efficiently using cloud-based solutions. The hands-on approach not only improved my confidence in using AWS but also prepared me for future data-driven projects.

# Work Samples-2

## Project Title 2: Visualize Data with Amazon QuickSight

**Objective**

This lab focused on the use of **Amazon QuickSight**, a serverless business intelligence (BI) tool, in combination with **Amazon S3**, to analyze and visualize large-scale datasets stored in the cloud. The aim was to gain practical experience in building interactive dashboards from structured data using AWS services.

**Introduction to Amazon QuickSight**

Amazon QuickSight is a fast, cloud-powered business analytics service that enables users to easily create and publish interactive dashboards that include machine learning (ML) insights. It integrates seamlessly with AWS data sources like Amazon S3, Redshift, and Athena, and supports data inputs from external services.

**Tools and Technologies Used**

* **Amazon S3** – For cloud storage of dataset files.
* **Amazon QuickSight** – For BI and data visualization.
* **CSV and JSON files** – Used for data content and configuration.
* **GitHub Repository** – Source for downloading project files.
* **AWS IAM and Management Console** – For service configuration and access management.

**Steps Followed**

1. **Dataset Acquisition and Storage**
   * Downloaded the dataset (Netflix-titles.csv) and its manifest file (manifest.json) from a GitHub repository.
   * Created a new Amazon S3 bucket in my AWS account and uploaded both files.
   * Edited the manifest.json file to reflect the exact location of the dataset in the S3 bucket, enabling proper mapping for QuickSight.

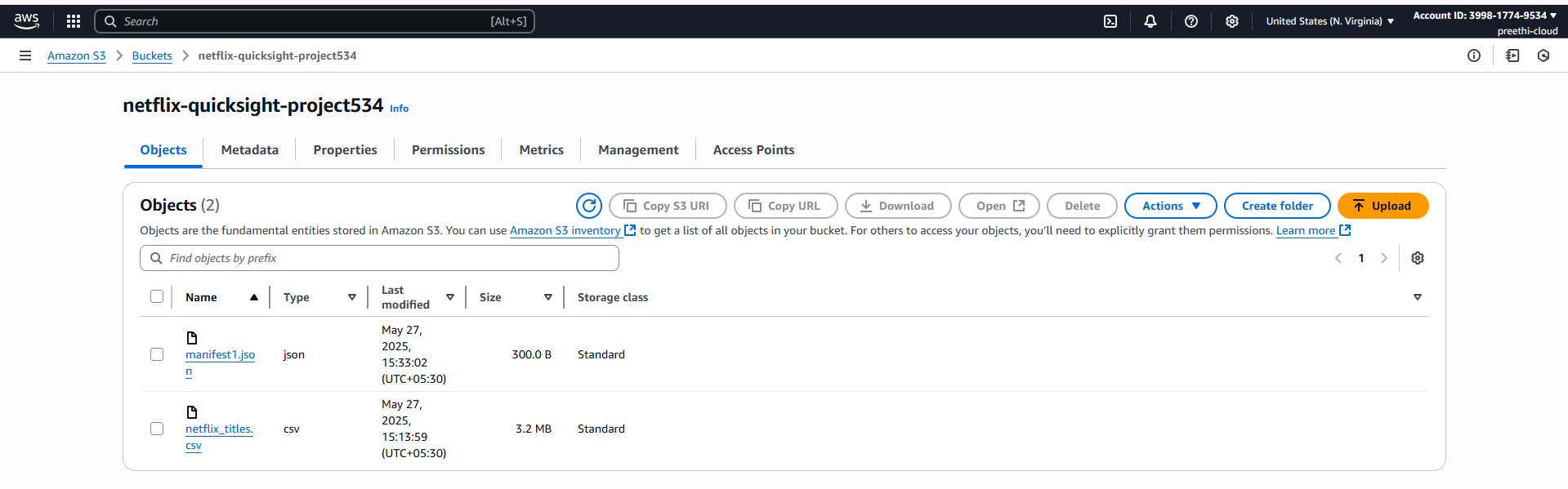


Figure 1.1: Dataset Acquisition and Storage

1. **Amazon QuickSight Setup**
   * Signed up for the **Amazon QuickSight Enterprise free trial** using my AWS account.
   * Linked QuickSight to my S3 bucket by adding a new S3 data source.
   * Uploaded the **S3 URI** of the manifest.json file which directed QuickSight to the CSV dataset.
   * Created a new dataset in QuickSight and verified that all fields (like brand, category, price, availability) were correctly loaded.

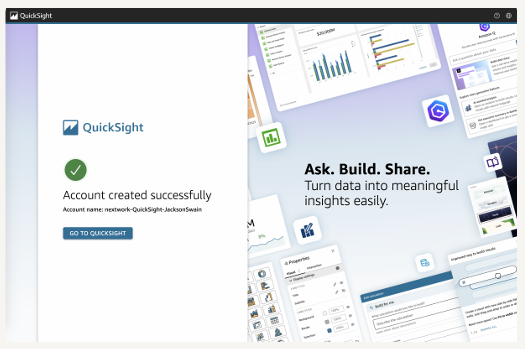


Figure 2.1: QuickSight Setup

1. **Creating Visualizations**
   * Used the drag-and-drop interface of QuickSight to design several visualizations:
     + **Bar charts** showing the most common product brands.
     + **Donut charts** analysing the distribution of categories or availability.
     + **Pie charts** and **filters** to segment and compare key metrics.
   * Explored the impact of filtering, sorting, and grouping on data interpretation.

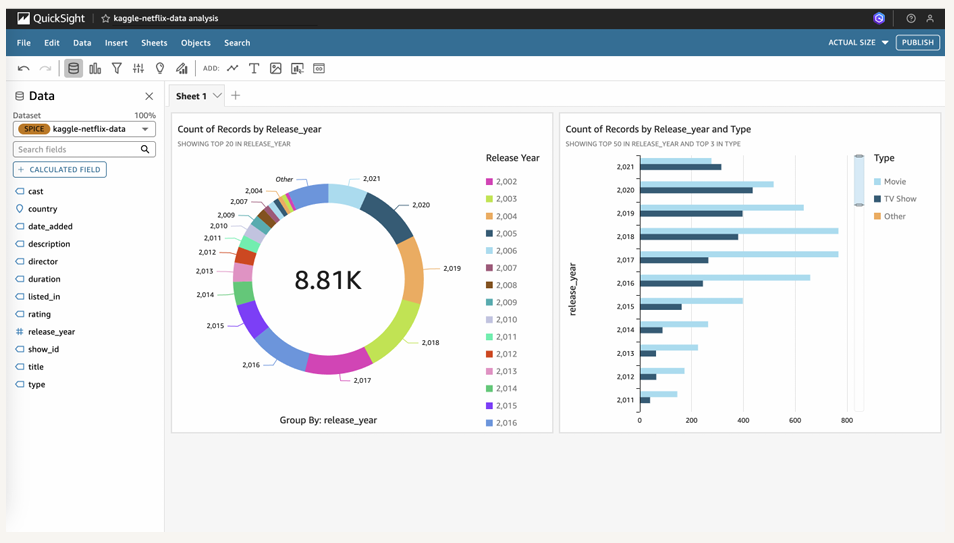


Figure 3.1: Visualizations

1. **Dashboard Development**
   * Combined visualizations into a cohesive interactive dashboard.
   * Added titles, legends, and formatting for clarity.
   * Published and exported the dashboard as a **PDF report** to preserve my work.

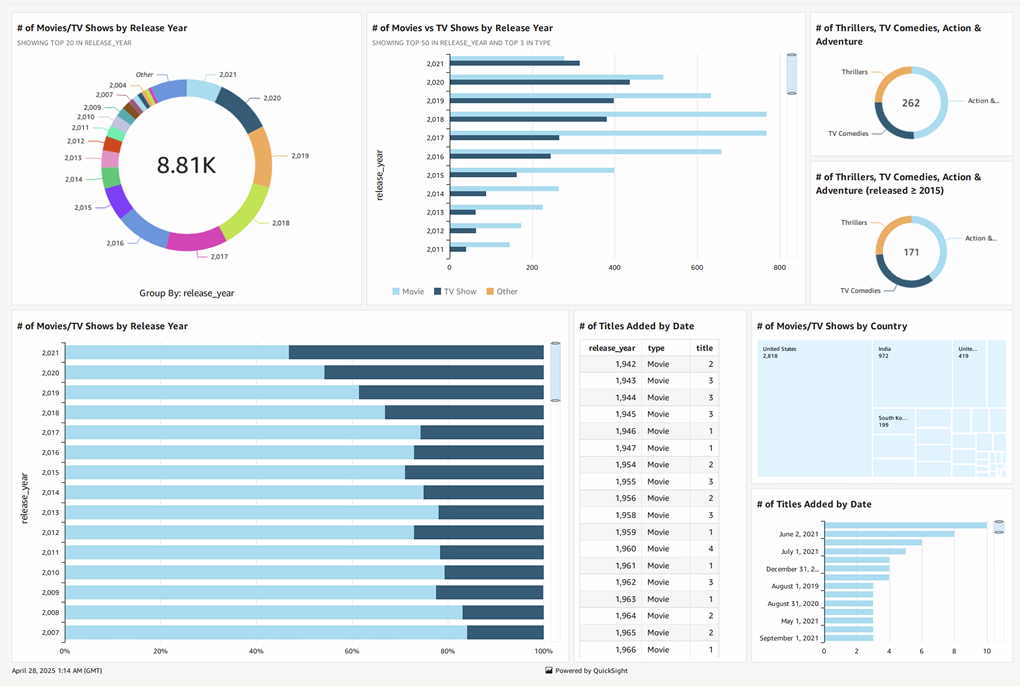


Figure : Dashboard Development

**Key Learnings**

* How to store and manage datasets in Amazon S3.
* How to configure and use **Amazon QuickSight** with S3 as a data source.
* Understanding the use of **manifest files** to guide QuickSight data imports.
* Designing and customizing dashboards using a **visual, code-free interface**.
* Gained insight into applying data analytics for e-commerce trends and product performance.

**Challenges Faced**

* **Linking S3 to QuickSight** required precise JSON configuration and correct S3 URIs. Initial errors in the path formatting led to connection failures.
* **Data rendering** in visualizations required understanding of field types and QuickSight’s auto-recognition behaviour.

**Outcome**

The successful creation of an interactive dashboard to visualize Amazon’s 50,000 best-selling products demonstrated how AWS services can be used effectively for business intelligence and data analytics. This project gave me confidence in integrating multiple cloud tools and understanding how real-world datasets can be used to drive decisions.

**Conclusion**

This lab enhanced my technical capabilities in cloud computing and data visualization using AWS tools. By following this step-by-step tutorial with my own AWS account, I built a practical understanding of using **Amazon QuickSight** for dashboard creation. I now feel prepared to apply these skills in real data-driven roles and projects.

# Conclusion

As a result, I'd like to conclude that internship played a critical part in not only expanding my theoretical but also practical knowledge. By pursuing this internship, I was able to get data analytics-based knowledge. As data analytics is a popular technology, it is both beneficial and promising in the future. Because of its established platforms such as databases and compilers, this platform is user friendly and simple to use.

# Appendices

## Appendix A – Screenshots of AWS Setup

**A.1:** amazonbestseller.csv and manifest.json files uploaded to an S3 bucket named my-quicksight-project.

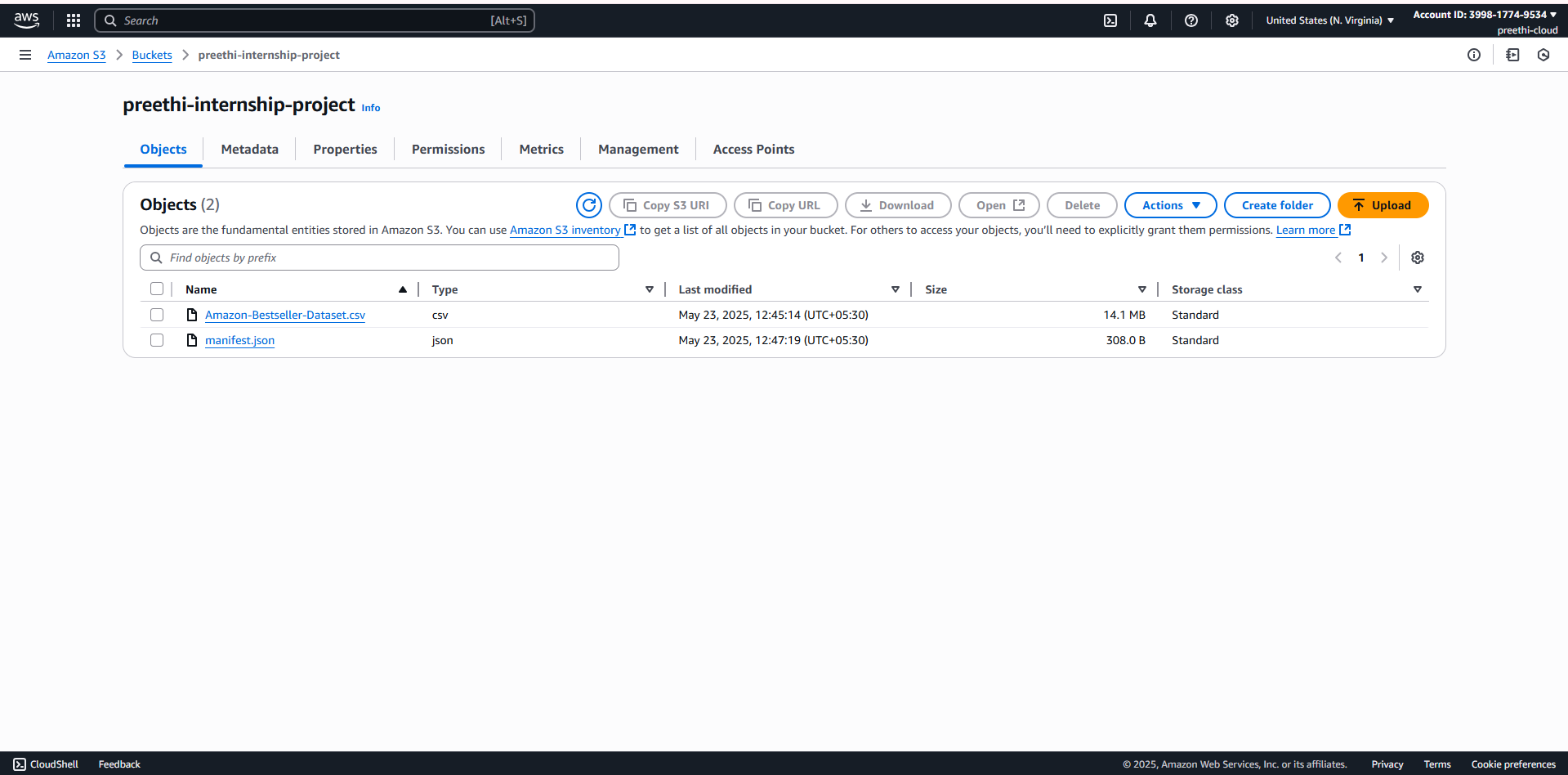


Figure A.*1*:AWS S3 bucket showing uploaded files  
A.2: Manifest file configured with the correct S3 path to enable QuickSight access.

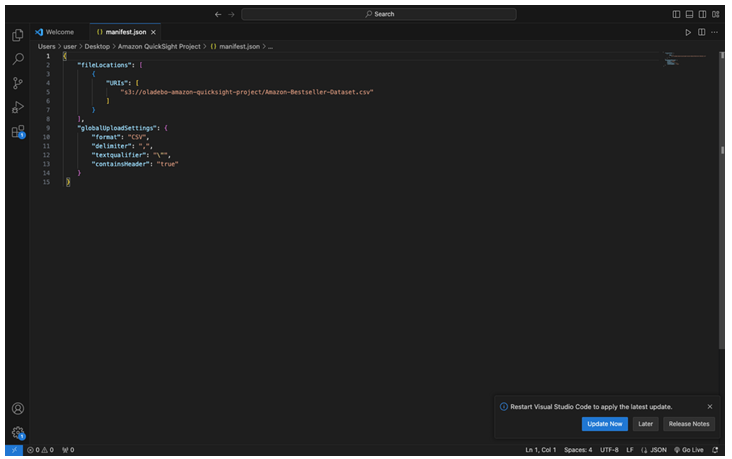


Figure A.*2*: Edited manifest.json file

**A.3:** QuickSight dashboard showing data source connection and available datasets.

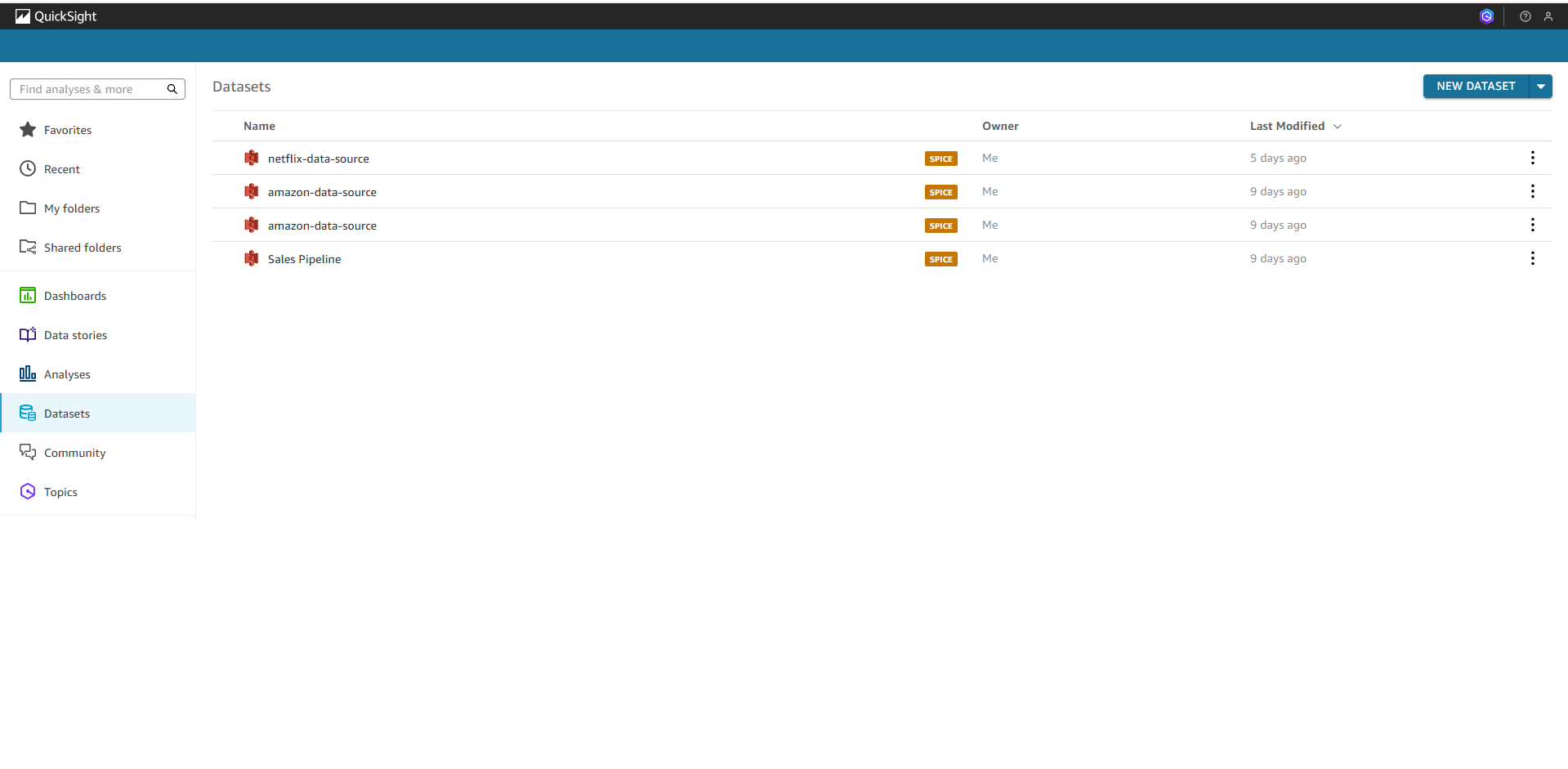


Figure A.*3*: Amazon QuickSight home screen

**A.4**: QuickSight chart showing the count of brands and price from the dataset.

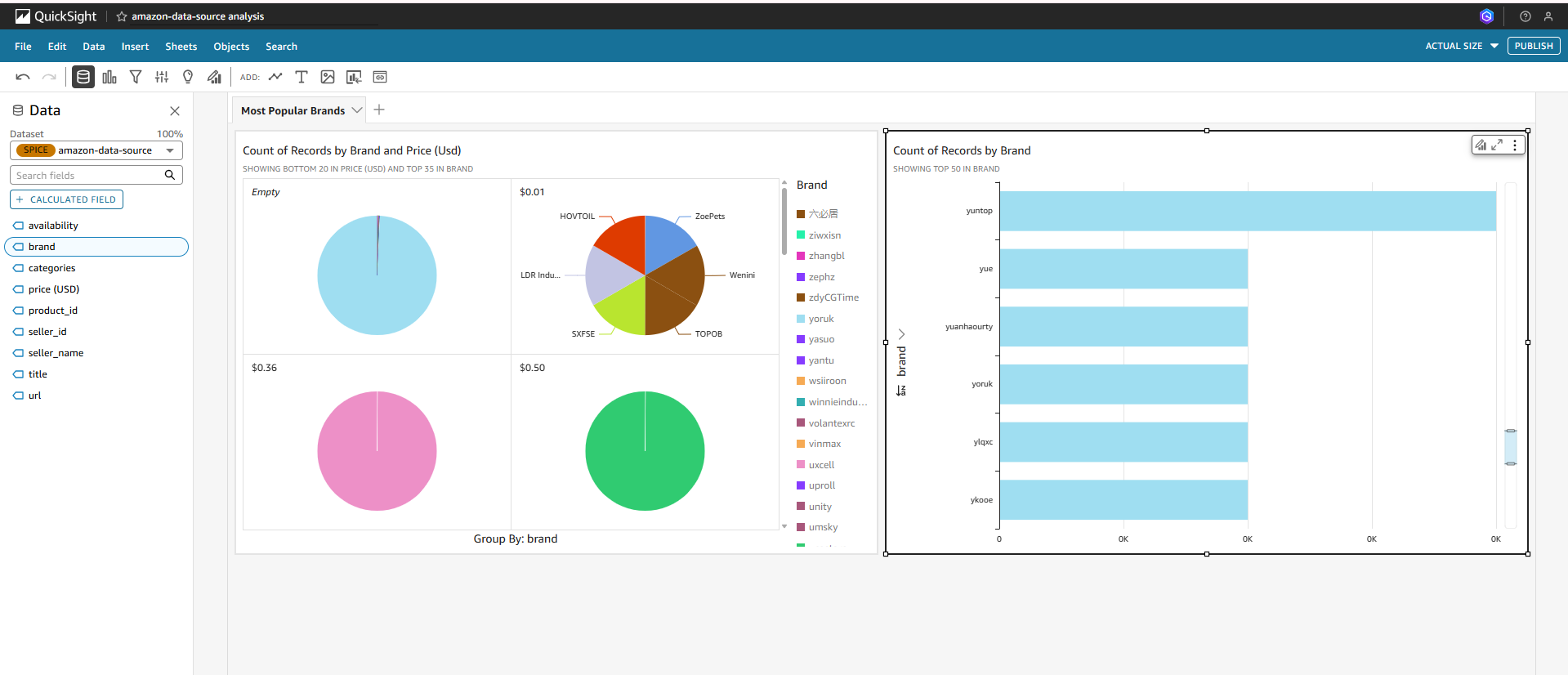


Figure A.*4*: Visualization – Most Popular Brands (Bar Chart & Pie Chart)

**A.5:** A donut chart representing the proportion of different product categories.

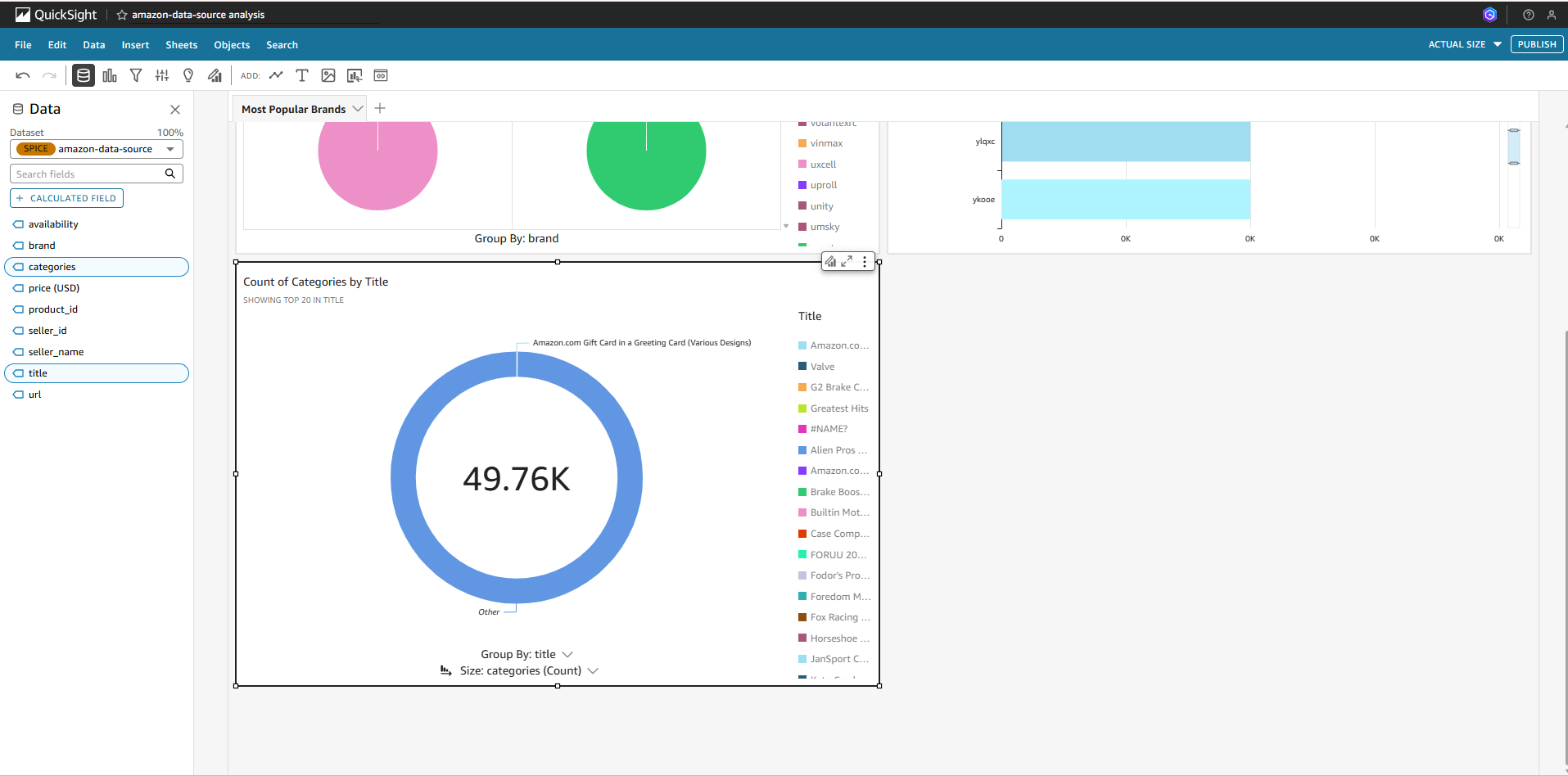


Figure A.*5*: Visualization – Category Distribution (Donut Chart)

## Appendix B – Sample Manifest File

Below is the sample manifest. json used to configure QuickSight data import:

json

{

"fileLocations": [

{

"URIs": [

"s3://preethi-internship-project/amazonbestseller.csv"

]

}

],

"globalUploadSettings": {

"format": "CSV",

"delimiter": ",",

"textqualifier": "'",

"containsHeader": "true"

}

}

## Appendix C – Data Sample (CSV Snippet)

Sample of the uploaded dataset (amazonbestseller.csv):

| **Title** | **Brand** | **Price** | **Category** | **Availability** |
| --- | --- | --- | --- | --- |
| Wireless Bluetooth Earbuds | SoundPEATS | 39.99 | Electronics | In Stock |
| 1080p Action Camera | GoPro | 199.99 | Sports & Cam | In Stock |
| 12-Pack Gel Pens | Paper Mate | 7.99 | Stationery | In Stock |
| Stainless Steel Water Bottle | Hydro Flask | 29.95 | Home & Kitchen | Out of Stock |

Table 1Sample Dataset

## Appendix D – Visualization Insights

* **Bar Chart (Most Popular Brands)**: Displays top brands with highest occurrence in the dataset.
* **Donut Chart (Product Categories)**: Helps identify dominant product categories.
* **Pie Chart (Availability)**: Compares number of in-stock vs. out-of-stock products.
* **Line Graph (Price Trends)**: Could be added to explore pricing over categories or brands.

## Appendix E – QuickSight Account Setup Notes

* **QuickSight Edition Used**: Enterprise (Free Trial)
* **Region**: US East (N. Virginia)
* **S3 Access Configuration**: Enabled while setting up QuickSight account; appropriate IAM role was created.
* **Dataset Connection**: S3 manifest JSON used to import and prepare the dataset in QuickSight.

## Appendix F – AWS IAM Role and Permission Settings

*Description*: To allow Amazon QuickSight to access the S3 bucket, an IAM role was either created or updated with the following policy permissions:

json

{

"Version": "2012-10-17",

"Statement": [

{

"Effect": "Allow",

"Action": [

"s3:GetObject",

"s3:ListBucket"

],

"Resource": [

"arn:aws:s3:::netflix-quicksight-project534",

"arn:aws:s3:::netflix-quicksight-project534/\*"

]

}

]

}

## Appendix G – Troubleshooting and Errors Encountered

* **Issue**: JSON manifest path error when linking S3 to QuickSight  
  **Resolution**: Verified S3 URI and corrected typo in bucket name.
* **Issue**: Dataset columns not loading in QuickSight  
  **Resolution**: Rechecked CSV format and re-uploaded clean version of the dataset.
* **Issue**: Dashboard export quality  
  **Resolution**: Used Quick Sight’s built-in "Export to PDF" function and adjusted visualization sizes for better output.

## Appendix H – Glossary of Technical Terms

| **Term** | **Definition** |
| --- | --- |
| **Amazon S3** | Simple Storage Service by AWS for object storage. |
| **Amazon QuickSight** | A cloud-powered BI tool for data visualization. |
| **CSV** | Comma-Separated Values, used for tabular data files. |
| **JSON** | JavaScript Object Notation, used here for manifest configuration. |
| **SPICE** | Super-fast, Parallel, In-memory Calculation Engine used by QuickSight. |
| **IAM** | Identity and Access Management – AWS's service for controlling permissions. |

Table 2 Technical term & Definition

## Appendix I – Learning Resources and Tutorials Followed

1. **YouTube Tutorial**: *Build an End-to-End Analytics Dashboard Using Amazon QuickSight* – AWS Online Tech Talks.  
   <https://www.youtube.com>
2. **Documentation**:
   * Amazon QuickSight: <https://docs.aws.amazon.com/quicksight>
   * Amazon S3: <https://docs.aws.amazon.com/s3>

# References

[1] Amazon Web Services, Inc., “Amazon QuickSight Documentation.” [Online]. Available: <https://docs.aws.amazon.com/quicksight>

[2] Amazon Web Services, Inc., “Amazon S3 Documentation.” [Online]. Available: <https://docs.aws.amazon.com/s3>

[3] Bright Data, “Public Web Data for Research and Analytics.” [Online]. Available: <https://brightdata.com>

[4] J. Swain, Visualize Data with QuickSight, NextWork Community, PDF document. [Online]. Available: Internal classroom distribution, titled legendary-aws-analytics-quicksight.pdf

[5] AWS Online Tech Talks, “Build an End-to-End Analytics Dashboard Using Amazon QuickSight,” YouTube. [Online]. Available: <https://www.youtube.com>

[6] GitHub, “Amazon Best Seller Dataset Repository.” [Online]. Available: Private Repository (used for CSV and manifest.json files)