Mohammad Idrees Bhat

Tech Skills Trainer | AI/ML Consultant

Cloud Computing

Using AWS, Azure, and GCP for data analytics
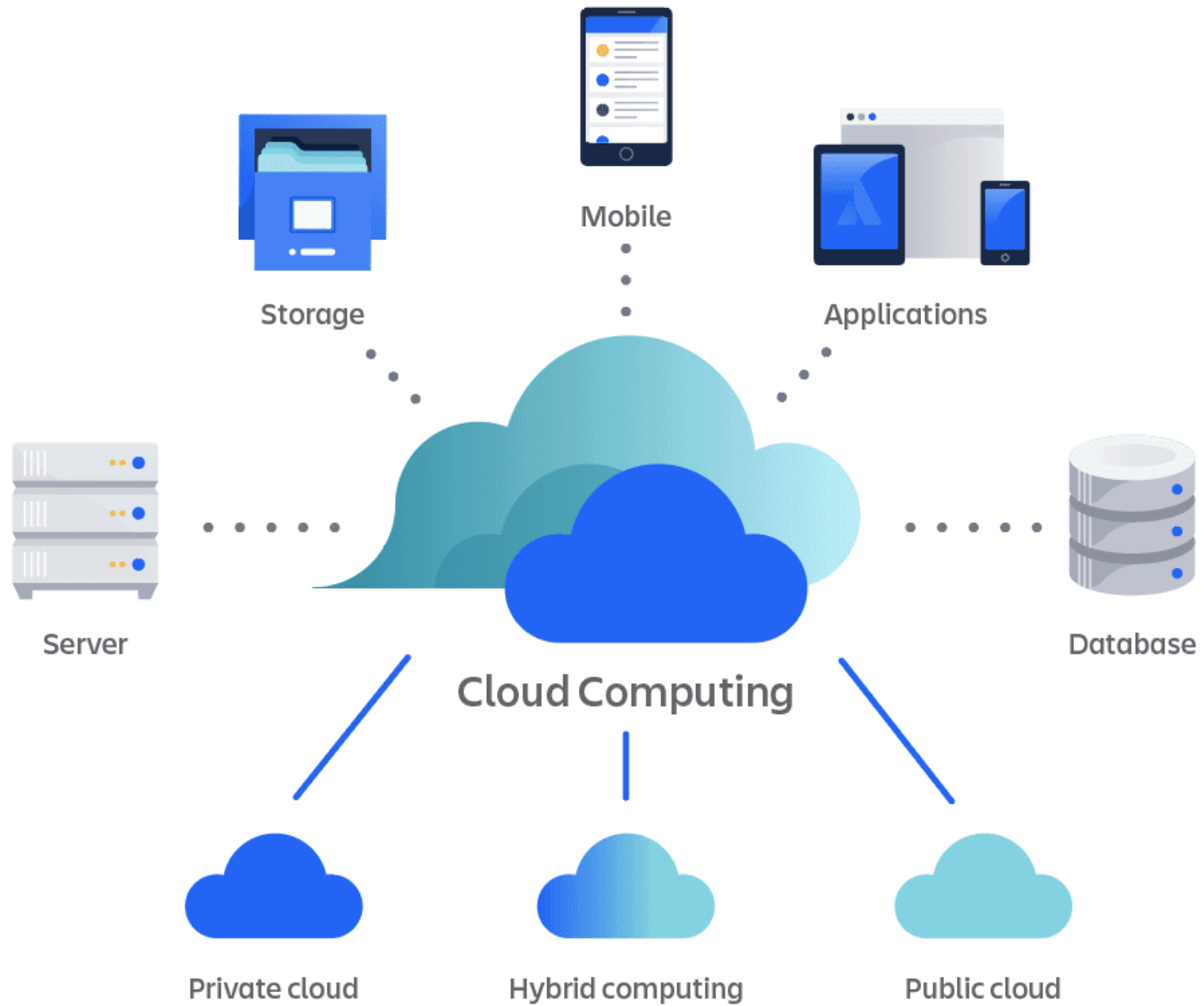
AGENDA

1. Introduction to Cloud Computing and Platforms

2. Data Storage and Analytics Services

3. Advanced Tools for Cloud Analytics

What's a hobby or activity you've always wanted to try but haven't yet?

1. Introduction to Cloud Computing and Platforms

Cloud computing refers to the delivery of computing services (such as servers, storage, databases, networking, software, and analytics) over the internet ("the cloud").

These services are offered on-demand, typically through a subscription-based model. Cloud computing enables businesses and individuals to access powerful computing resources without having to invest in physical hardware.
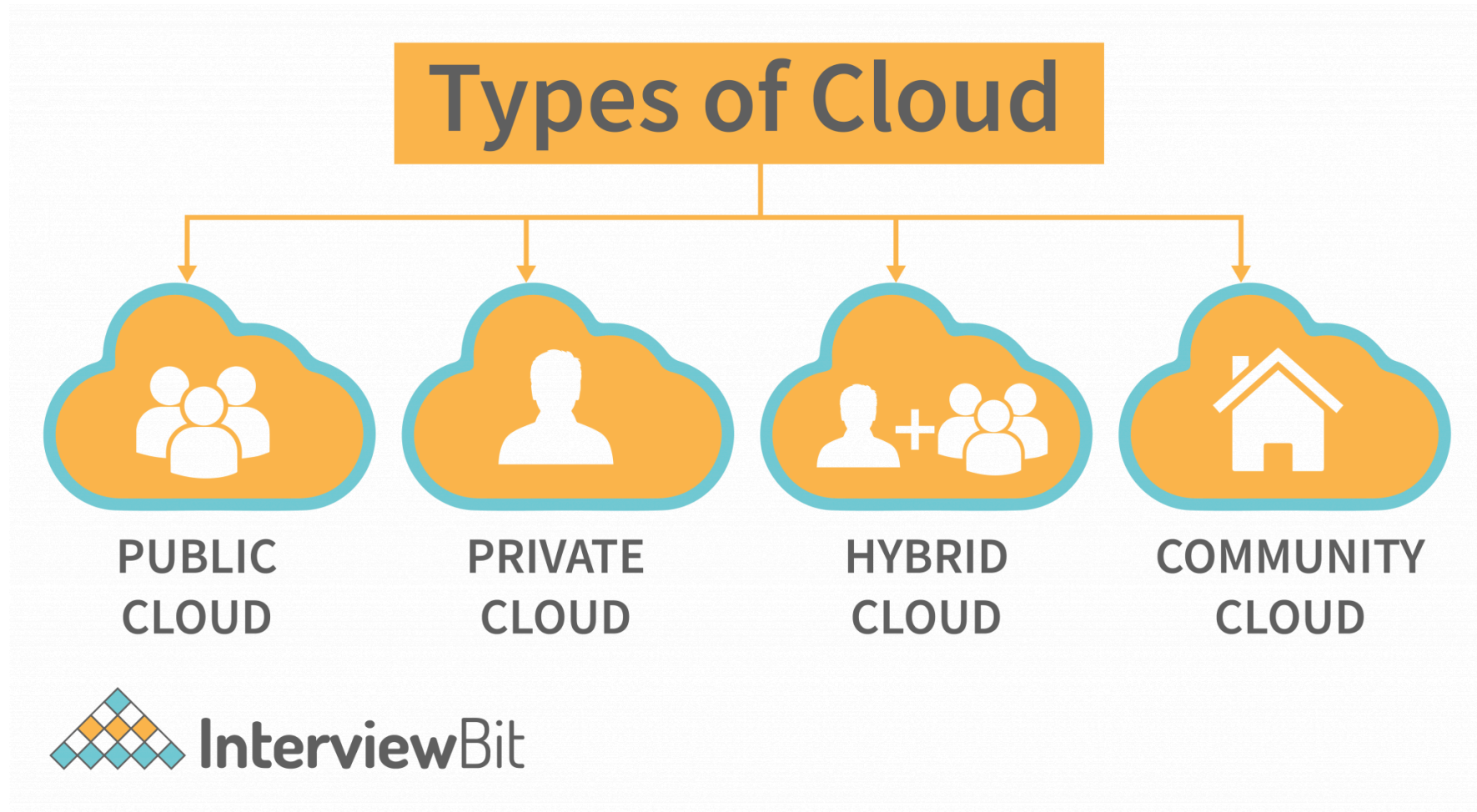
- **Public Cloud:**

The infrastructure is owned and **operated by third-party providers** (e.g., AWS, Azure, GCP) and shared among many customers.
Example: AWS, where multiple organizations use shared resources provided by Amazon.

- **Private Cloud:**

A cloud environment used **exclusively by one organization**, providing more control and security.
Example: Microsoft Azure Stack, where an organization uses Azure's private cloud services in its own data center.

- **Hybrid Cloud:**

A mix of private and public clouds, allowing data and applications to be shared between them.
Example: AWS Outposts, where an organization uses AWS services both in their private data center and on AWS's public cloud.

- **Community Cloud:**

A cloud infrastructure shared by several organizations with common concerns, allowing for resource sharing while meeting specific security, compliance, or industry requirements.
Example: A healthcare cloud, where multiple healthcare organizations share resources while adhering to privacy regulations like HIPAA.

Benefits of Cloud Computing in Data Analytics

Cloud computing offers significant advantages when it comes to processing large datasets for analytics:

1. **Scalability:**

Cloud platforms **allow you to scale resources** up or down based on demand. For example, if you're analyzing a large dataset, you can provision more compute power to handle the load, and reduce it when the task is complete.

2. **Cost Efficiency:**

With a **pay-as-you-go model**, you only pay for the computing resources you use, saving costs on infrastructure and hardware maintenance.

3. **Flexibility:**

You can **run workloads on the cloud** and **use a variety of tools and services**, from simple data storage to complex machine learning models.

    4. **Collaboration:**

Cloud platforms **enable teams to access**, share, and analyze data from anywhere, facilitating collaborative data analytics work.

Key Concepts in Cloud Computing

- **Scalability:**

The **ability to grow or shrink** your computing resources as needed. This is particularly useful in data analytics, where you may need to process large amounts of data at varying speeds.

Example: AWS Auto Scaling automatically adjusts the amount of compute resources based on traffic and demand.

Learn more about scalability: AWS Auto Scaling Overview

- **Elasticity:**

The ability of cloud resources to **automatically scale based on workload demands** . Elasticity ensures that you have the right amount of resources available at all times.

Example: If an unexpected spike in website traffic occurs, elastic cloud systems can automatically allocate more compute resources to handle the extra load.

Learn more about elasticity: Microsoft Azure Elasticity Guide

- **Pay-As-You-Go:**

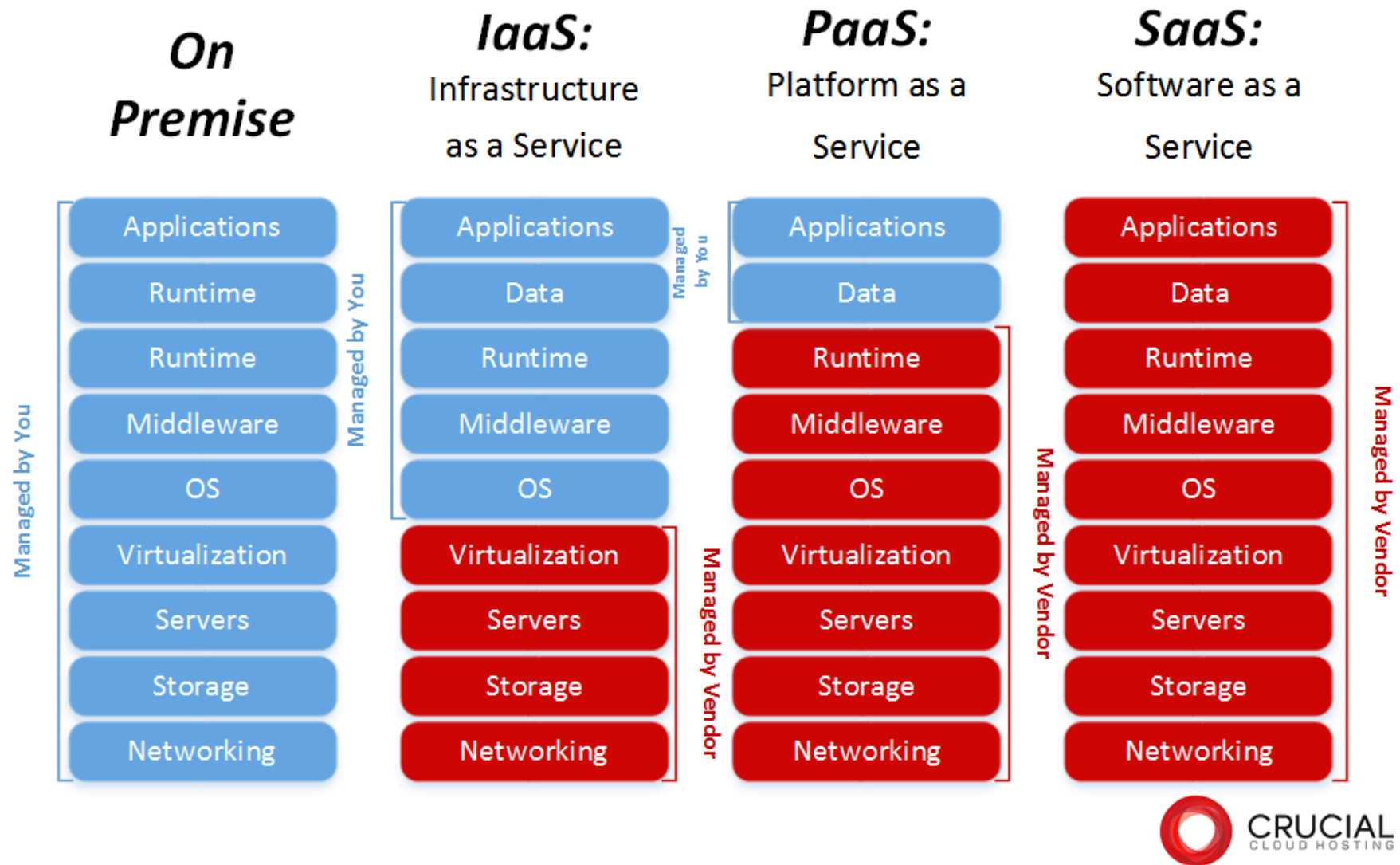With cloud services, you only pay for what you use. This pay-per-use model helps organizations save money by eliminating the need for expensive hardware investments.

Example: You can start small with limited compute power and scale your infrastructure only when necessary, avoiding unnecessary upfront costs.

Learn more about Pay-As-You-Go pricing: GCP Pricing Overview

Types of Cloud Services

On Premise

Managed by You
Applications
Runtime
Runtime
Middleware
OS
Virtualization
Servers
Storage
Networking

IaaS: Infrastructure as a Service

Managed by You

Applications
Data
Runtime
Middleware
OS

Managed by Vendor

Virtualization
Servers
Storage
Networking

PaaS: Platform as a Service

Applications
Data

Managed by Vendor

Runtime
Middleware
OS
Virtualization
Servers
Storage
Networking

SaaS: Software as a Service

Managed by Vendor

Applications
Data
Runtime
Middleware
OS
Virtualization
Servers
Storage
Networking

CRUCIAL
CLOUD HOSTING

Cloud computing offers different service models depending on how much control you need over the infrastructure, platform, and software. The three primary service models are:

1. **IaaS (Infrastructure as a Service):**

- Provides virtualized computing resources over the internet.
- You can rent IT infrastructure (like servers and storage) on a pay-per-use basis.
- Examples: AWS EC2, Azure Virtual Machines, Google Compute Engine.

Learn more about IaaS: AWS EC2 Overview

2. **PaaS (Platform as a Service):**

- Provides a platform allowing customers to develop, run, and manage applications without managing the underlying infrastructure.
- It simplifies the deployment of applications by providing tools and services for development.
- Examples: Azure App Services, AWS Elastic Beanstalk, Google App Engine.

Learn more about PaaS: Azure App Service Overview

3. **SaaS (Software as a Service):**

- Delivers software applications over the internet, typically via a subscription model.
- The provider hosts and manages the software, while users access it through a web browser.
- Examples: Google Workspace, Microsoft 365, Salesforce.

Learn more about SaaS: Salesforce Overview

Overview of Cloud Platforms

1. AWS (Amazon Web Services)

AWS is the leading cloud platform with a wide range of services for computing, storage, databases, and machine learning. It offers **scalability**, **reliability**, and **flexibility** that are ideal for businesses of all sizes.

Key Features and Services for Data Analytics:

- **Amazon S3**: Object storage for large-scale data.
- **Amazon Redshift**: Data warehouse service optimized for analytics.
- **AWS Glue**: Data integration service for ETL (extract, transform, load) operations.
- **Amazon EMR**: Managed Hadoop framework for big data processing.
- **AWS Lambda**: Serverless compute service for running code in response to events.

Pricing Model:

- **Pay-as-you-go** model based on usage.
- **Free tier**: Includes limited usage for new users.
- **On-demand pricing** and **reserved instances** are available for computing services, which provide flexibility based on workload requirements.

Real-world Use Case:

- **Netflix**: Uses AWS for streaming, storage, and data processing. AWS helps Netflix manage billions of hours of content delivered globally with seamless performance.

---

2. Microsoft Azure

Azure is known for its deep integration with Microsoft's existing products and services, making it a go-to option for businesses that rely on Microsoft tools.

Key Features and Services for Data Analytics:

- **Azure Blob Storage**: Scalable object storage for large volumes of unstructured data.
- **Azure Synapse Analytics**: Integrated analytics service for big data and data warehousing.
- **Azure Data Lake Analytics**: Big data analytics service.
- **Power BI**: Analytics and business intelligence platform for interactive data visualizations.
- **Azure Machine Learning**: A set of tools for building, training, and deploying machine learning models.

Pricing Model:

- **Pay-as-you-go**: Charges based on the amount of storage, computing, and services used.
- **Free tier**: Available for beginners to explore Azure's features with limited resources.
- **Reserved pricing**: Discounts for long-term use, ideal for predictable workloads.

Real-world Use Case:

- **Adobe**: Uses Azure for its Creative Cloud and Document Cloud services, leveraging Azure's computing and storage power for massive data processing and customer insights.

3. Google Cloud Platform (GCP)

GCP is known for its **advanced data analytics** and **machine learning** capabilities. Google's infrastructure and expertise in data-driven technologies make it a strong contender in the cloud market.

Key Features and Services for Data Analytics:

- **Google Cloud Storage**: Object storage for any amount of data.
- **BigQuery**: Serverless, highly scalable, and cost-effective data warehouse for analytics.
- **Dataflow**: Stream and batch data processing service.
- **Google AI and ML Services**: Tools for building and deploying machine learning models.
- **Google Dataproc**: Managed Hadoop and Spark for big data processing.

Pricing Model:

- **Pay-per-use**: GCP's pricing model is based on resource usage, with separate charges for storage, computing, and data processing.
- **Free tier**: Provides users with access to certain services for free with limited usage.
- **Sustained use discounts**: Automatically applied for long-running workloads.

Real-world Use Case:

- **Spotify**: Uses GCP's BigQuery and other data analytics tools to analyze large volumes of music data to personalize recommendations and improve user experience.

Comparison Table:

Feature/Service	AWS	Azure	GCP
Storage	Amazon S3	Azure Blob Storage	Google Cloud Storage
Data Analytics	Amazon Redshift, EMR, Glue	Azure Synapse Analytics, Data Lake	BigQuery, Dataflow, Dataproc
Machine Learning	SageMaker	Azure Machine Learning	AI and ML Services, TensorFlow
Pricing	Pay-as-you-go, reserved instances	Pay-as-you-go, reserved pricing	Pay-per-use, sustained use discounts

Feature/Service	AWS	Azure	GCP
Real-time Analytics	Kinesis, Lambda	Azure Stream Analytics	Dataflow, Pub/Sub
Free Tier	Yes, with limited usage	Yes, limited resources	Yes, limited usage

Choosing the Right Platform: Each platform has its strengths:

- **AWS** is great for organizations needing a comprehensive suite of services with an established ecosystem.
- **Azure** works best for companies already using Microsoft tools, providing a seamless transition to the cloud.
- **GCP** is ideal for businesses focused on cutting-edge machine learning and big data analytics, especially if they already use Google products.

Let's move towards setting up the environments now

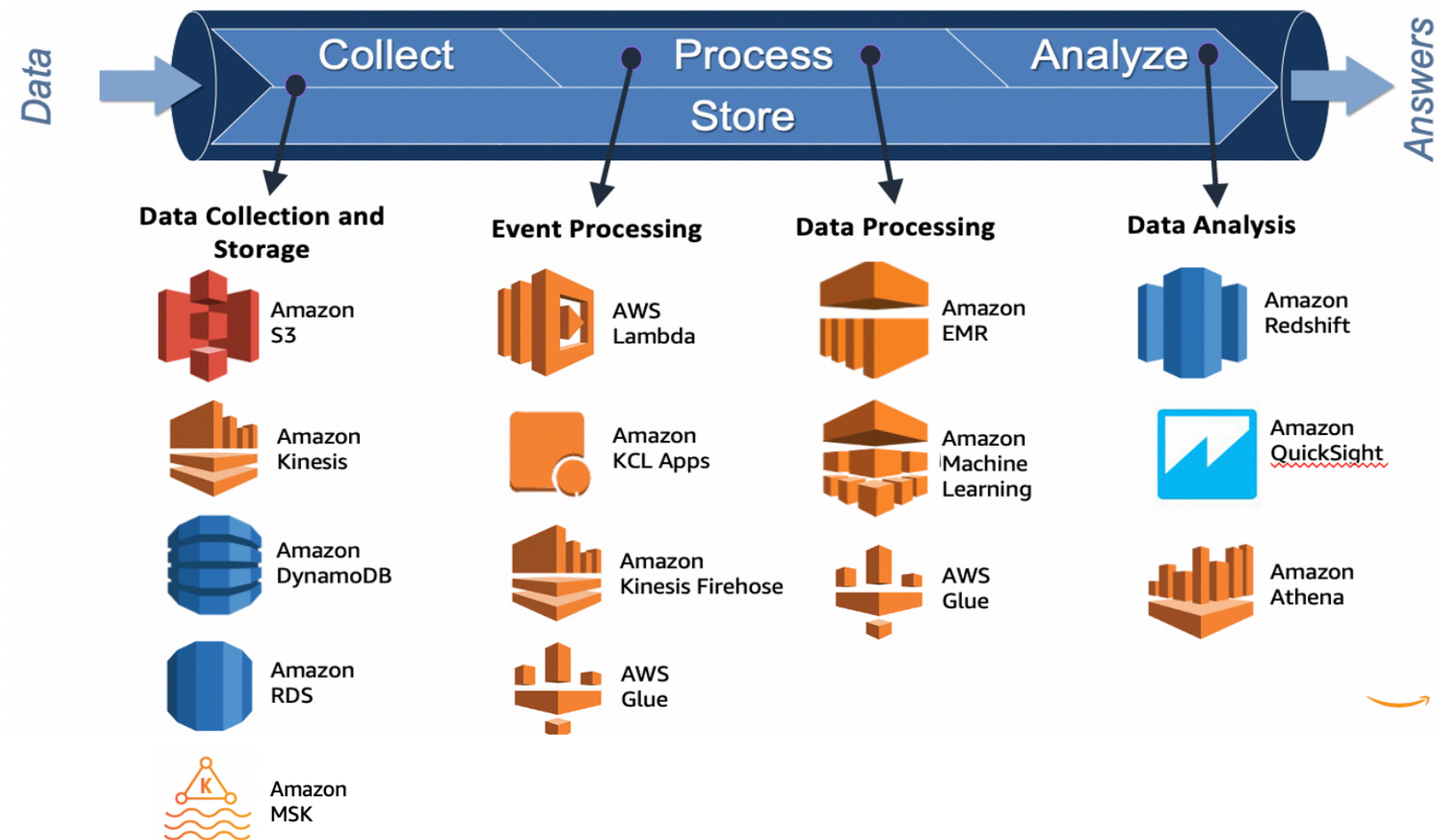## 2. Data Storage and Analytics Services

Setting Up a Cloud Environment

Each cloud platform provides a **management console** or **dashboard**, which is the central hub for managing your cloud environment.

Most cloud platforms offer **free tiers** or **credits** for new users, allowing you to explore their services without incurring any charges. These are especially helpful for beginners who want to practice without worrying about costs.

Steps to Set Up an Account on AWS, Azure, or GCP

AWS:

Data → **Collect** → **Process / Store** → **Analyze** → **Answers**

Data Collection and Storage
- Amazon S3
- Amazon Kinesis
- Amazon DynamoDB
- Amazon RDS
- Amazon MSK

Event Processing
- AWS Lambda
- Amazon KCL Apps
- Amazon Kinesis Firehose
- AWS Glue

Data Processing
- Amazon EMR
- Amazon Machine Learning
- AWS Glue

Data Analysis
- Amazon Redshift
- Amazon QuickSight
- Amazon Athena

1. Visit the AWS website: https://aws.amazon.com.

2. Click on **Create a Free Account**.

3. Sign in with your **Amazon account** (or create one if you don't have one).

4. Provide your **payment information** (AWS offers a free tier, but they still require payment details for verification purposes).

5. Choose your **support plan** (Basic Support is free).

6. Complete the **identity verification** via text message or phone call.

7. Once your account is set up, you will be directed to the **AWS Management Console**, where you can start creating and managing cloud resources.

- **AWS Management Console:**
    - The AWS Console is a web-based interface for managing AWS services and resources.
    - It provides a **user-friendly interface** where you can launch new instances, manage databases, monitor usage, and access detailed analytics for your cloud resources.
    - The **AWS Resource Groups** feature lets you organize resources into logical groupings for easier management.

AWS Free Tier:

- **Free for 12 months** after signing up (some services are always free).
- **Services included**: EC2 (750 hours/month of t2.micro instances), S3 (5GB standard storage), RDS (750 hours/month of db.t2.micro instances), and more.
- Check AWS Free Tier details: AWS Free Tier

Azure:

1. Go to the Azure website: https://azure.microsoft.com.
2. Click on **Start free** to create a new Azure account.
3. Sign in with a **Microsoft account** (or create one if needed).
4. Enter your **payment information** (Azure offers a **free tier** with limited services and a credit for new users).
5. Verify your **identity** through a phone number and credit card (you won't be charged unless you upgrade).
6. After setup, you will be taken to the **Azure Portal**, where you can manage your cloud resources and services.

- **Azure Portal:**
    - The **Azure Portal** is a web-based application that provides a comprehensive view of your resources in the Azure cloud.
    - You can monitor, configure, and manage all your Azure services from a single dashboard.
    - The portal offers **customizable dashboards**, allowing you to track important metrics and manage your resources efficiently.

Azure Free Account:

- **$200 in credits** for the first 30 days, plus free access to over 25 services for 12 months.
- **Free services**: Virtual machines, storage, and databases.

- Check Azure Free Account details: Azure Free Account

Google Cloud Platform (GCP): Learn More about GCP

1. Visit the Google Cloud website: https://cloud.google.com.
2. Click on **Get Started for Free**.
3. Sign in with your **Google account** (or create one if you don't already have one).
4. Enter your **payment information** (GCP offers $300 in free credits for the first 90 days, even if you don't enter a billing account, you won't be charged without explicit consent).
5. Verify your **identity** via phone or email.
6. Once your account is set up, you'll be directed to the **Google Cloud Console** to start using the services.

- **Google Cloud Console:**
  - The **Google Cloud Console** provides a **centralized interface** to manage and configure Google Cloud resources.
  - It offers a wide range of services, including computing, storage, and machine learning.
  - Google Cloud also provides **easy access** to billing, usage reports, and project management tools.

Google Cloud Free Tier:

- **$300 in credits** for new users for the first 90 days.
- **Always Free**: Limited usage of Google Cloud services, including 1 f1-micro instance per month, 5GB of storage, and more.
- Check GCP Free Tier details: Google Cloud Free Tier

Data Storage in the Cloud

Each major cloud provider offers a storage service designed for a wide variety of use cases:

- **AWS S3 (Simple Storage Service):**

    - A highly scalable and durable object storage service.
    - Ideal for **storing unstructured data** such as images, videos, and backups.
    - Features include **lifecycle management**, versioning, and fine-grained access control.
    - Pricing is based on storage type (e.g., Standard, Intelligent-Tiering) and data retrieval frequency.
- **Azure Blob Storage:**

    - A **high-performance object storage** service optimized for large-scale unstructured data.
    - Offers integration with other Azure services, such as machine learning and analytics.
    - Tiers include **Hot, Cool, and Archive**, catering to different storage needs based on access frequency.
    - Supports features like **data encryption**, redundancy options, and shared access signatures for secure file sharing.
- **GCP Cloud Storage:**

    - A unified, object storage service with a focus on global accessibility and performance.
    - Storage classes include **Standard, Nearline, Coldline, and Archive**, optimized for cost and access speed.
    - Key features include **bucket-level management**, cross-region replication, and integrated analytics tools.
    - Simplifies storage configuration with **user-friendly tools and APIs**.

---

3. Hands-On Activity: Uploading and Managing Files in Cloud Storage

For this activity, we'll use a free-tier account on one of the platforms to upload and manage files. If you don't have an account, refer to the earlier section on account setup.

Step-by-Step Instructions:

AWS S3:

1. Log into the AWS Management Console.
2. Navigate to **S3** and create a new bucket.
3. Configure bucket settings (e.g., name, region, access permissions).
4. Click on the bucket and select **Upload** to add files.
5. Explore options like versioning, public access settings, and lifecycle rules.

Azure Blob Storage:

1. Log into the Azure Portal.
2. Create a new **Storage Account**.
3. Navigate to **Containers** and create a container.
4. Upload files to the container using the **Blob Storage Explorer**.
5. Review the access policies and try generating a **shared access signature (SAS)**.

GCP Cloud Storage:

1. Log into the Google Cloud Console.
2. Navigate to **Storage** and create a bucket.
3. Set bucket-level options such as region and storage class.
4. Use the **Upload Files** button to add data to the bucket.
5. Experiment with managing bucket permissions and creating lifecycle rules.

Data Analytics Services in the Cloud

Cloud platforms provide powerful analytics services that enable users to process, analyze, and visualize large datasets efficiently.

Here's an overview of the key analytics services offered by AWS, Azure, and GCP, along with an explanation of their capabilities.

1. Key Analytics Services

- **AWS Analytics Services:**

  - **Athena:**

    - A serverless query service that allows users to analyze data stored in S3 using SQL.
    - Ideal for ad-hoc queries and quick data insights without setting up infrastructure.
    - Example use case: Running analytics on access logs stored in S3.

  - **Glue:**

    - A fully managed **ETL (Extract, Transform, Load)** service for data preparation.
    - Automates schema discovery and transforms data for analytics applications.
    - Example use case: Preparing raw data in S3 for analysis in Redshift.

  - **Redshift:**

    - A fully managed, high-performance **data warehouse** for large-scale analytics.
    - Integrates with AWS services for seamless data movement.
    - Example use case: Storing and analyzing terabytes of e-commerce transaction data.

- **Azure Analytics Services:**

  - **Synapse Analytics:**

    - Combines big data and data warehousing into a single service.
    - Supports SQL-based and Spark-based analytics workflows.
    - Example use case: Building a data pipeline for a retail company's sales data.

  - **Data Factory:**

    - A cloud-based **data integration service** that orchestrates and automates data movement and transformation.
    - Example use case: Extracting data from on-premises databases to Azure Blob Storage for further analysis.

- **GCP Analytics Services:**

    - **BigQuery:**

        - A serverless, highly scalable **data warehouse** optimized for massive-scale analytics.
        - Supports standard SQL queries and integrates with popular visualization tools.
        - Example use case: Analyzing customer behavior data from a marketing campaign.
    - **Dataflow:**

        - A fully managed service for **streaming and batch data processing**.
        - Ideal for real-time analytics and ETL workflows.
        - Example use case: Processing IoT sensor data for predictive maintenance.

---

2. How These Services Process Large Datasets Efficiently

Cloud analytics services are optimized for handling large datasets with speed and reliability. Here's how they achieve this:

- **Scalability:**

    - Services like BigQuery and Redshift dynamically allocate resources to process massive datasets, ensuring performance remains consistent regardless of data size.
- **Distributed Processing:**

    - Tools like AWS Glue and Azure Synapse use distributed architectures to split tasks across multiple nodes, accelerating data transformations and computations.
- **Serverless Architecture:**

    - Services like Athena and BigQuery eliminate the need for infrastructure management, automatically scaling resources based on query demands.
- **Integration with Storage Services:**

- These tools seamlessly integrate with cloud storage options (e.g., AWS S3, Azure Blob Storage, GCP Cloud Storage) to ensure smooth data retrieval and processing.

3. Advanced Tools for Cloud Analytics

1. Integrating Big Data tools like Apache Spark with cloud platforms.
2. Addressing challenges and discussing best practices in cloud analytics.
3. Recap, group discussion, and Q&A to consolidate learning.

_____END_____

THANK YOU!

Live Exercise Solutions

```
In [ ]: solutions
```

Programming Interveiw Questions

1. topic:
   - question

```
In [ ]:
```

Mohammad Idrees Bhat

Tech Skills Trainer | AI/ML Consultant