Big Data on Amazon AWS

Introduction and Ingestion Layer – [Day 1]

CARLOS BARBOSA
Head of Content & Professor

WHOAMI?

$ carlosbarbosa

Professional

- Head of DataOps – A3Data
- Head of Content – Engenharia de Dados
- Tech Lead ED – A3Data
- Data Engineer – A3Data
- Data Engineer – Indra Company
- Data Scientist - Fiabilité
- Data Analyst - Fiabilité
- Data Analyst Intern – Armazém PB
- DBA – Fábrica de Software

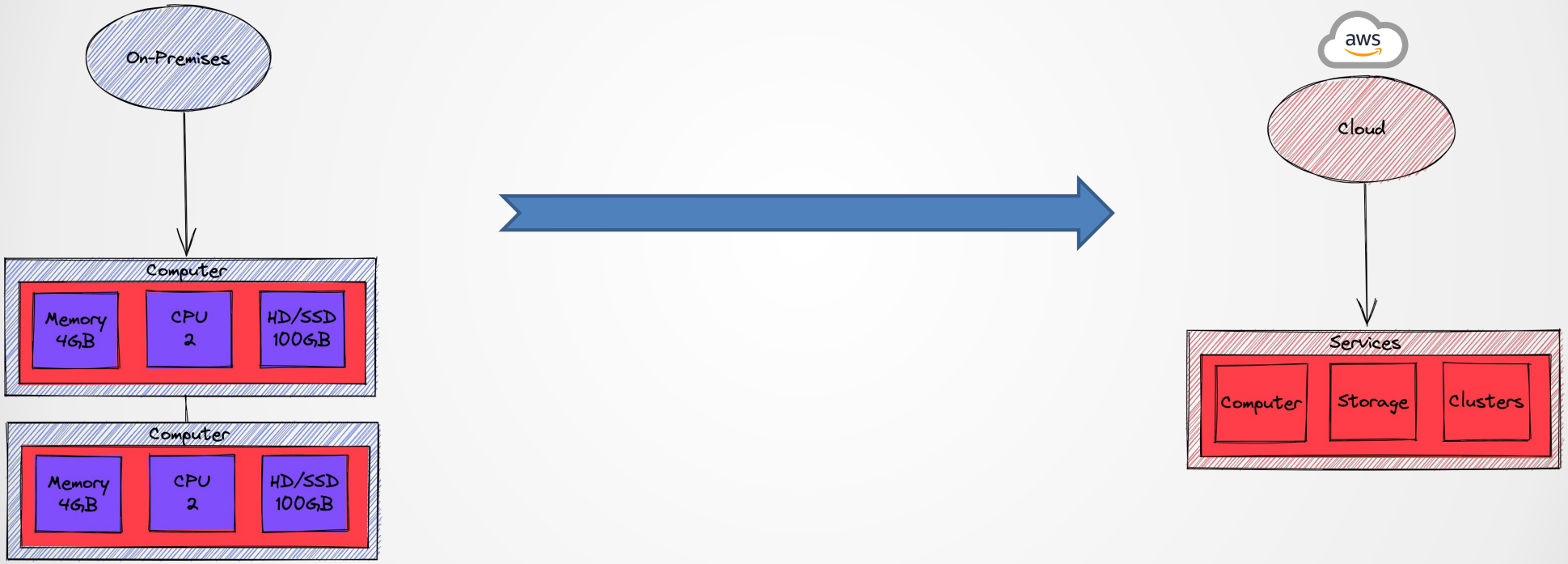
Personal

- 24 Years OLD
- Gamer
- Traveler
- Dog Lover (Spark) 🐶

What's Cloud Computing?

Simply put, cloud computing is the delivery of computing services including servers, storage, databases, networking, software, analytics, and intelligence.

"You don't generate your own electricity. Why generate your own computing?"- Jeff Bezos, Amazon.



Problems

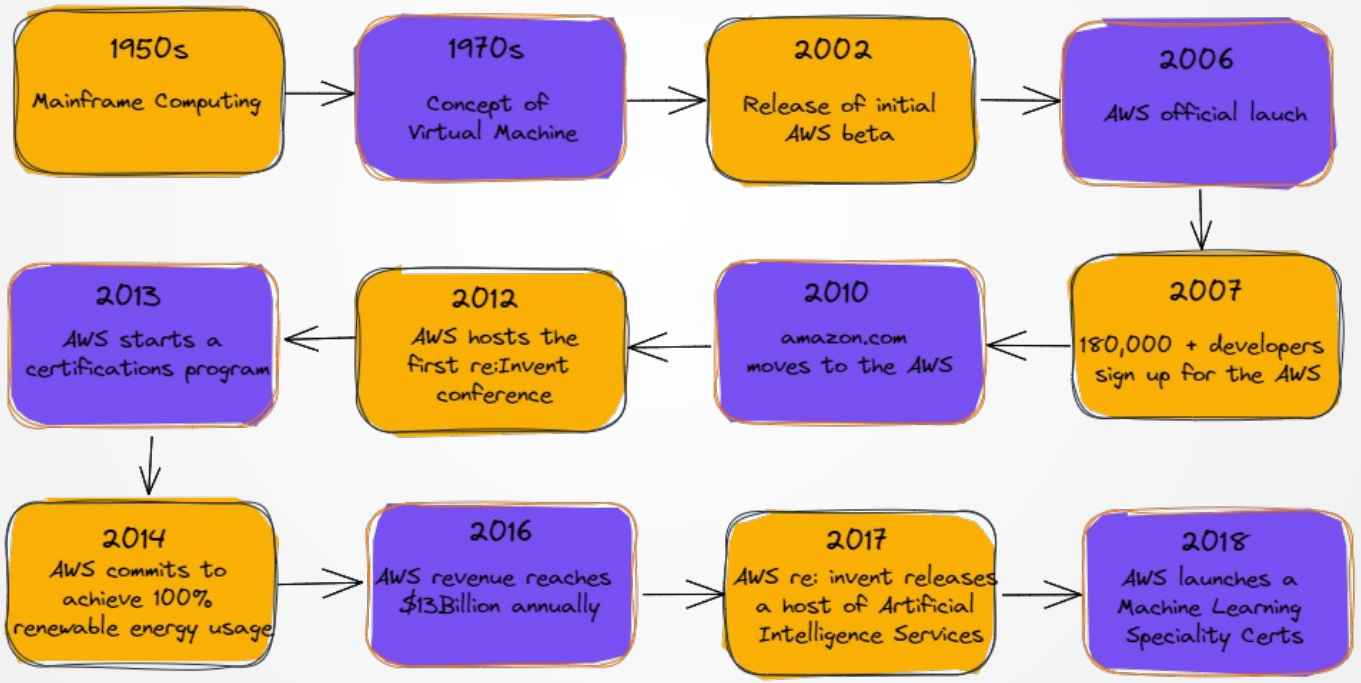
- Investment capital for hardware
- Needs physical space and ambiance
- Data center operation and maintenance costs
- Slows down the speed of value deliveries
- Too large team to maintain solutions

Solution

- On-demand cost
- Great economy of scale
- Scalability and Flexibility
- Increases development speed
- Geographically separated data centers
- No operating and maintenance costs for Data Centers

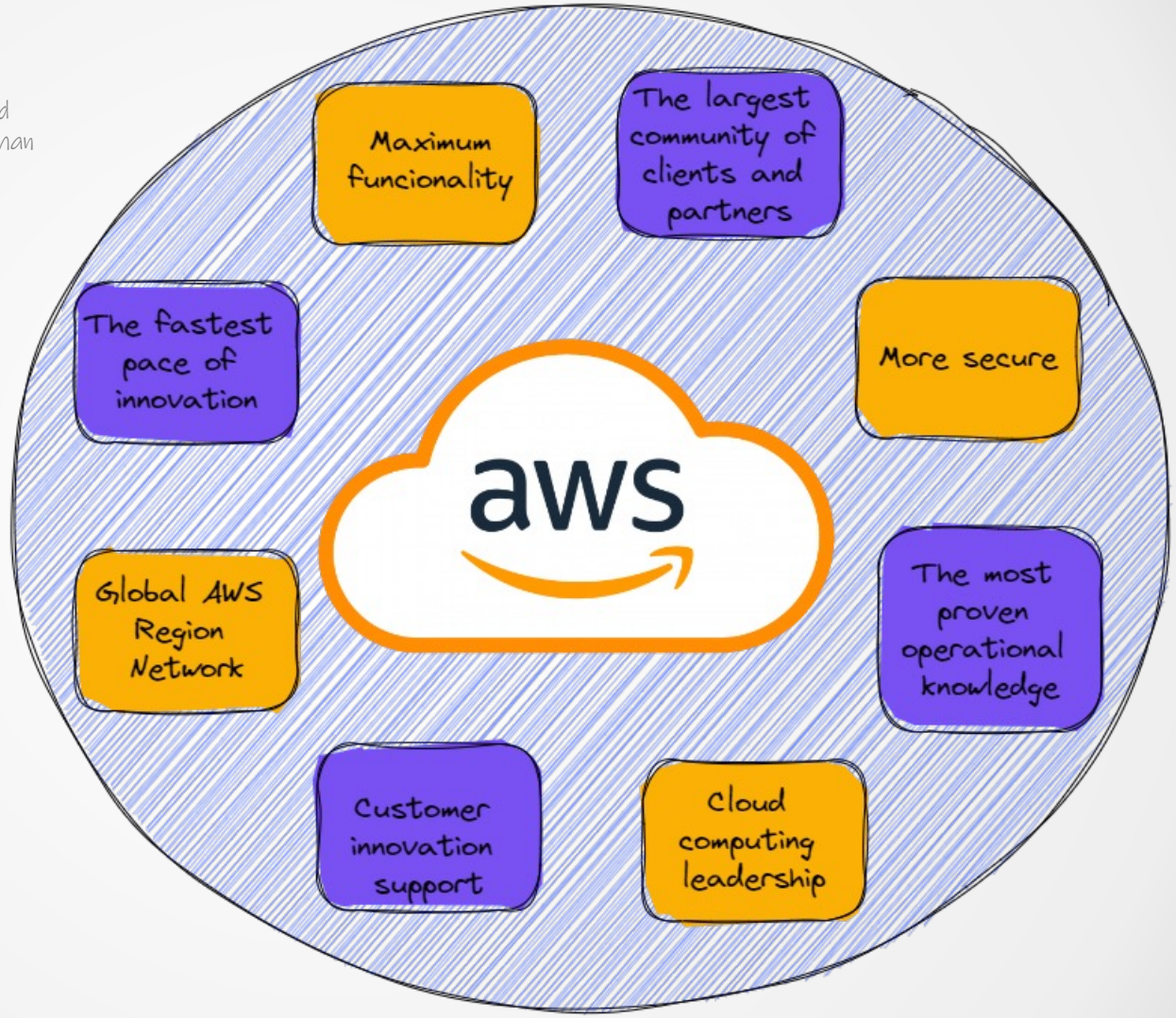
History

Breaf evolution of Cloud computing and Amazon Web Services through history



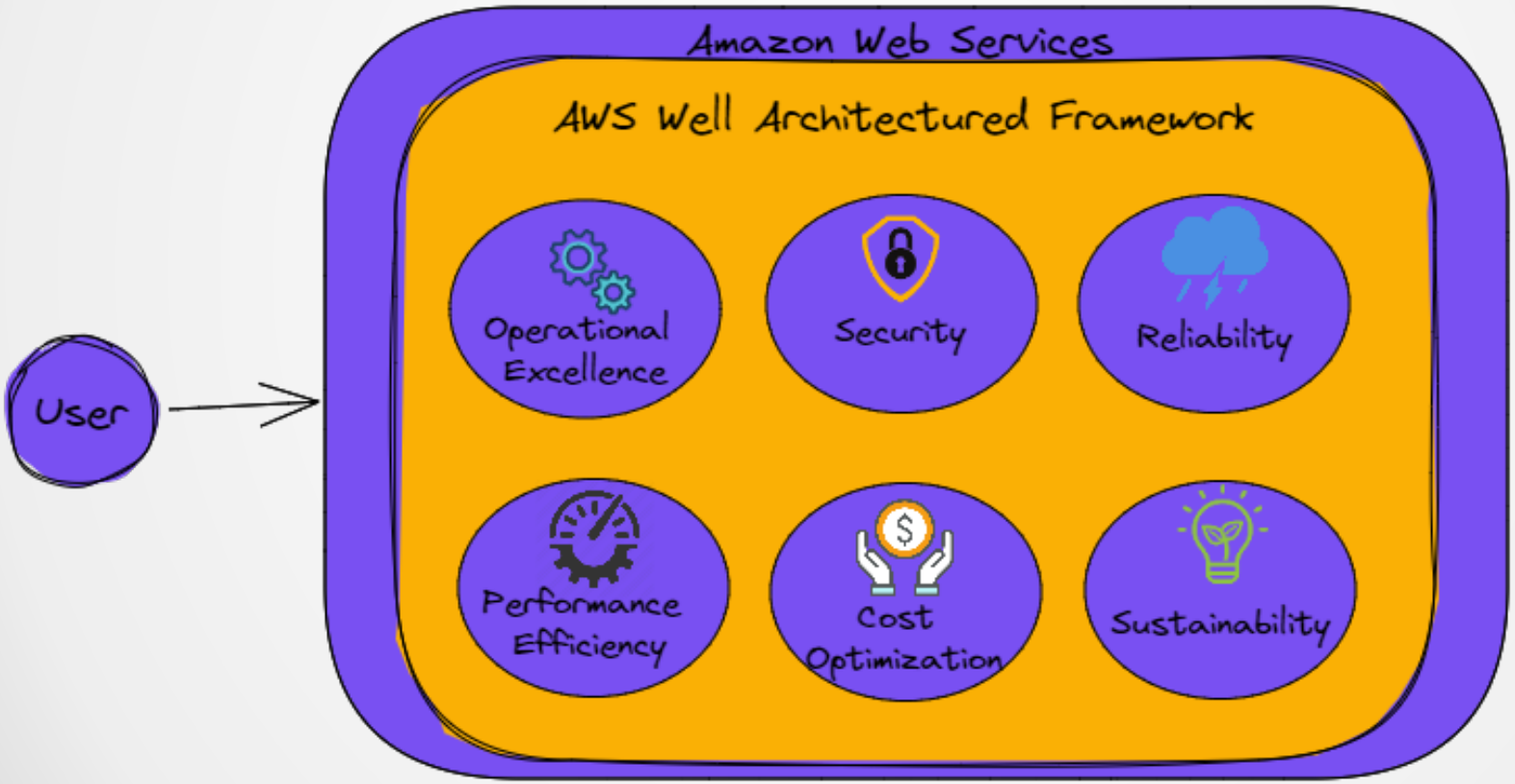
Amazon Web Services

Amazon Web Services (AWS) is the world's most adopted and most comprehensive cloud platform, offering more than 200 full services from data centers around the world.



Well-Architecture Framework

The AWS Well-Architected Framework helps you understand the pros and cons of decisions you make while building systems on AWS. By using the Framework you will learn architectural best practices for designing and operating reliable, secure, efficient, and cost-effective systems in the cloud.



Well-Architecture Framework

The AWS Well-Architected Framework helps you understand the pros and cons of decisions you make while building systems on AWS. By using the Framework you will learn architectural best practices for designing and operating reliable, secure, efficient, and cost-effective systems in the cloud.

Operational Excellence

- Perform operations as code
- Make frequent, small, reversible change
- Refine operations procedures frequently
- Anticipate failure
- Learn from all operational failures

Security

- Implement a strong identity foundation
- Enable traceability
- Apply security at all layers
- Automate security best practices
- Protect data in transit and at rest
- Keep people away from data
- Prepare for security events

Reliability

- Automatically recover from failure
- Test recovery procedures
- Scale horizontally to increase aggregate workload availability
- Stop guessing capacity
- Manage change in automation

Performance Efficiency

- Democratize advanced technologies
- Go global in minutes
- Use serverless architectures
- Experiment more often
- Consider mechanical sympathy

Cost Optimization

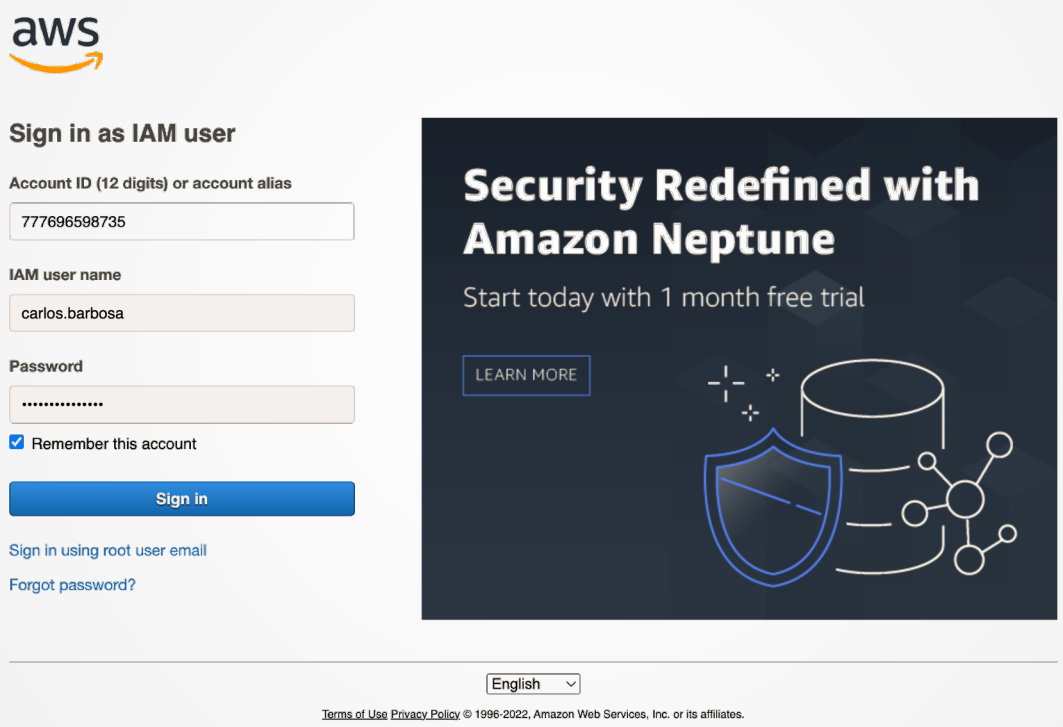
- Implement Cloud Financial Management
- Adopt a consumption model
- Measure overall efficiency
- Stop spending money on undifferentiated heavy lifting
- Analyze and attribute expenditure

Sustainability

- Understand your impact
- Establish sustainability goals
- Maximize utilization
- Anticipate and adopt new, more efficient hardware and software offerings
- Use managed services
- Reduce the downstream impact of your cloud workloads

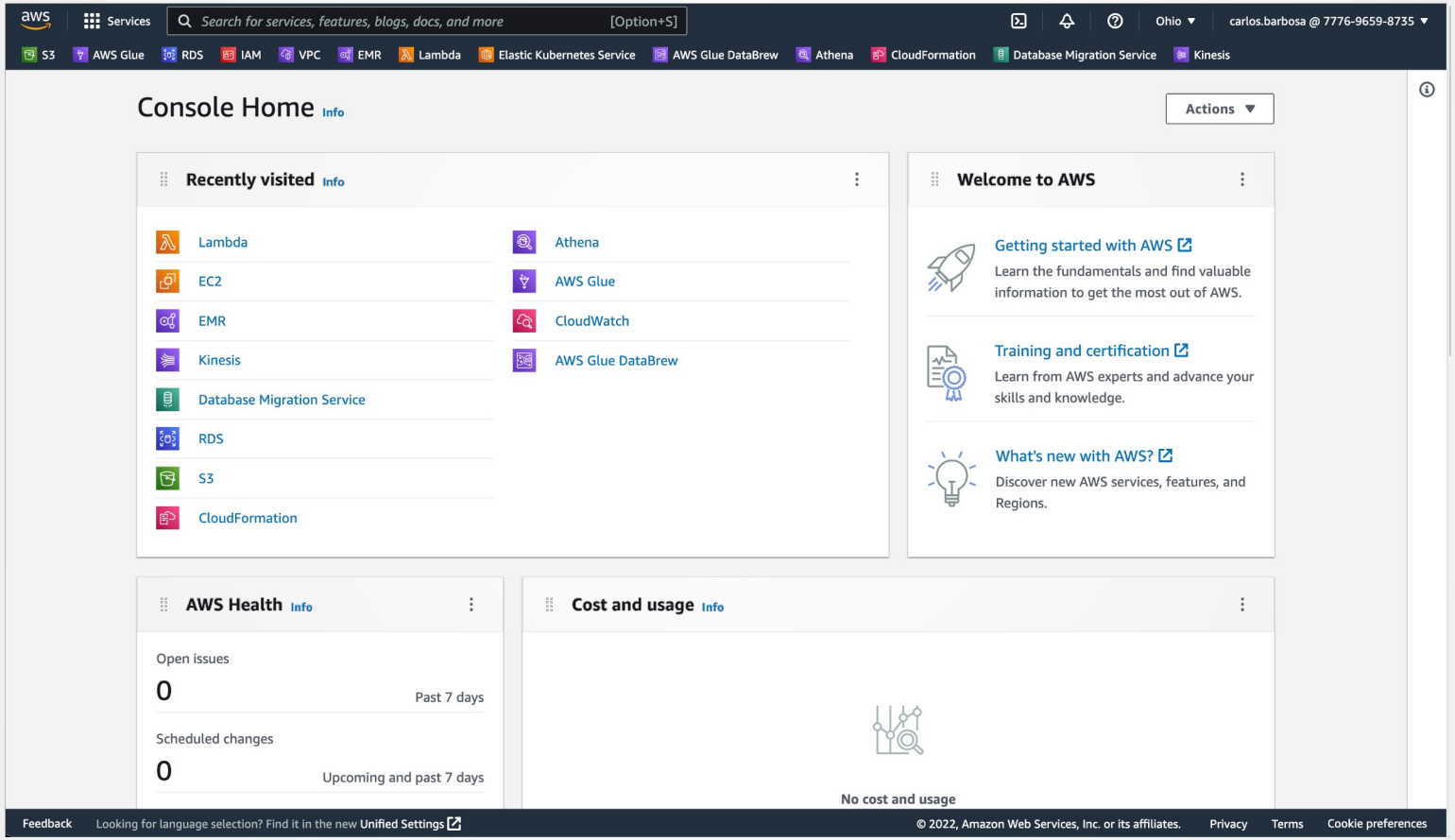
AWS Management Console

Everything you need to access and manage the AWS Cloud — in one web interface



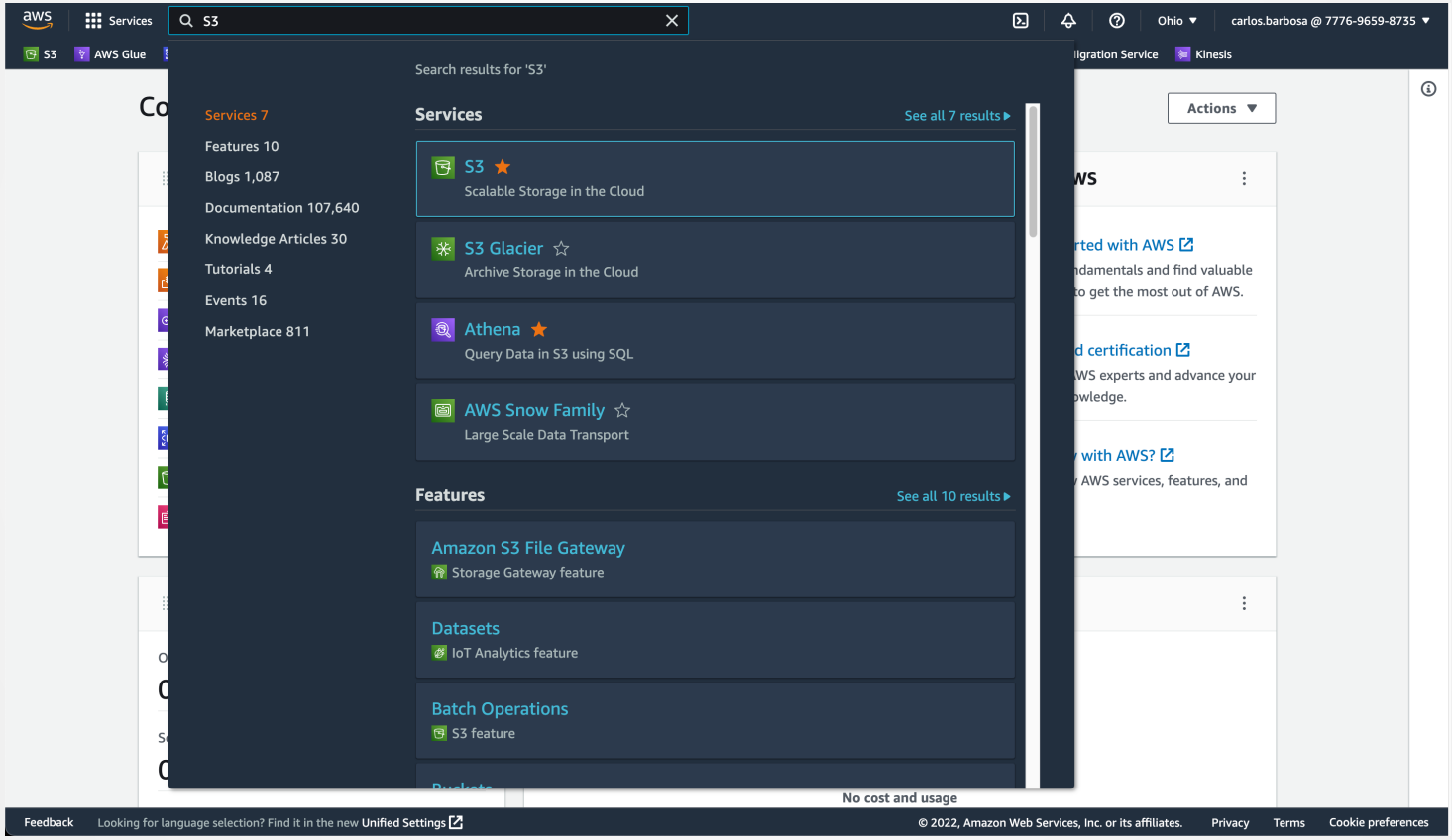
AWS Management Console

Everything you need to access and manage the AWS Cloud — in one web interface



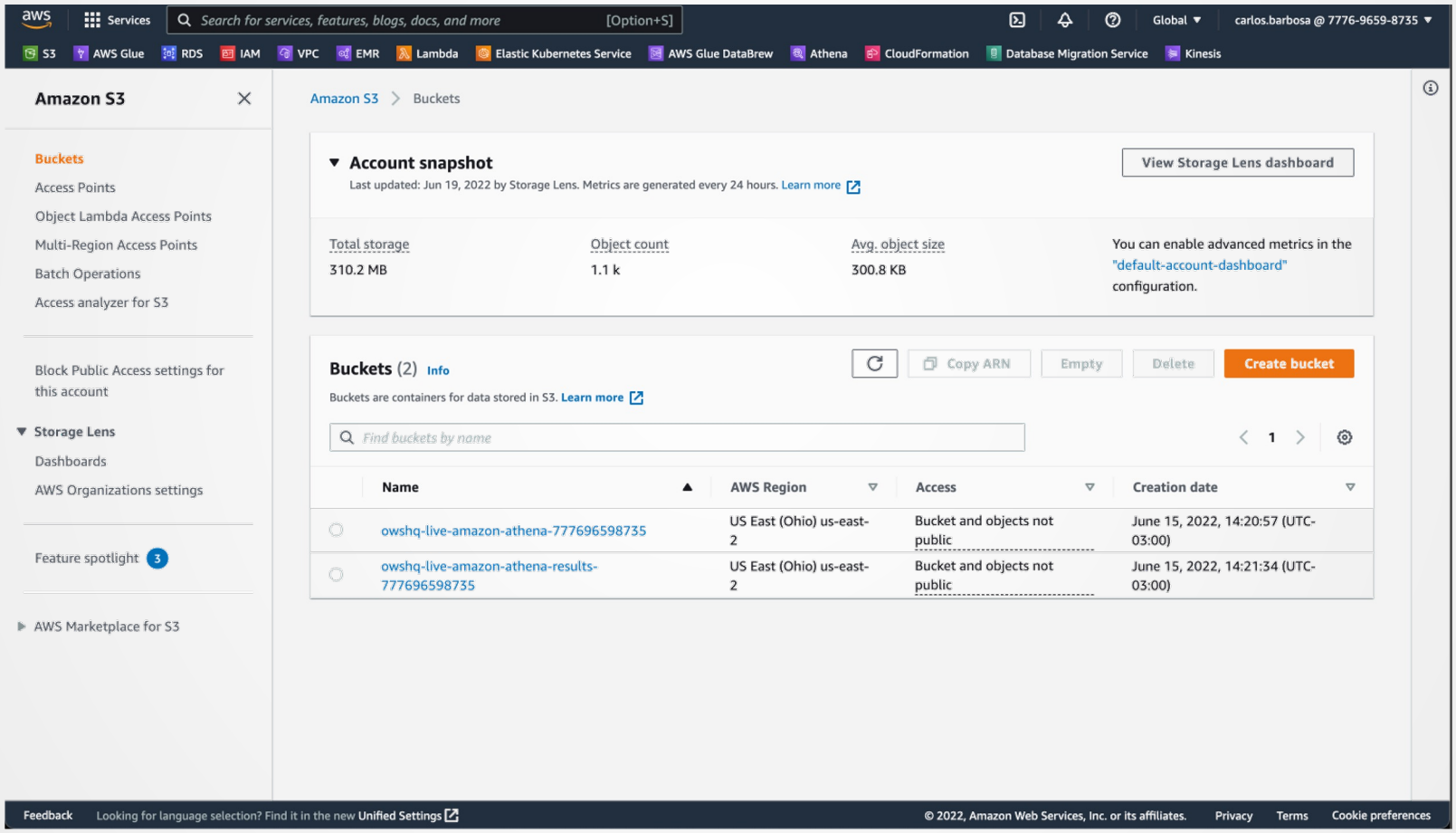
AWS Management Console

Everything you need to access and manage the AWS Cloud — in one web interface



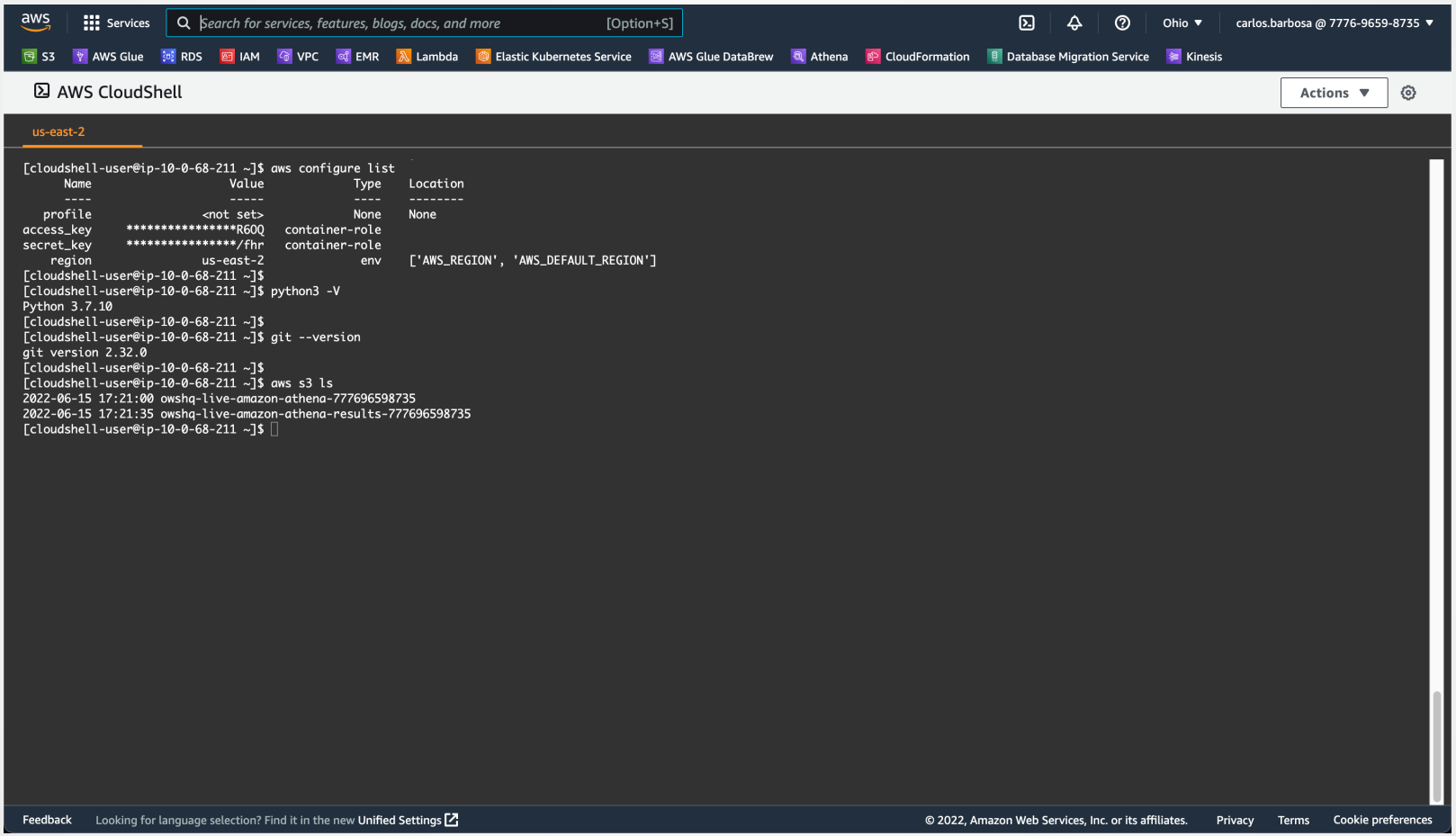
AWS Management Console

Everything you need to access and manage the AWS Cloud — in one web interface



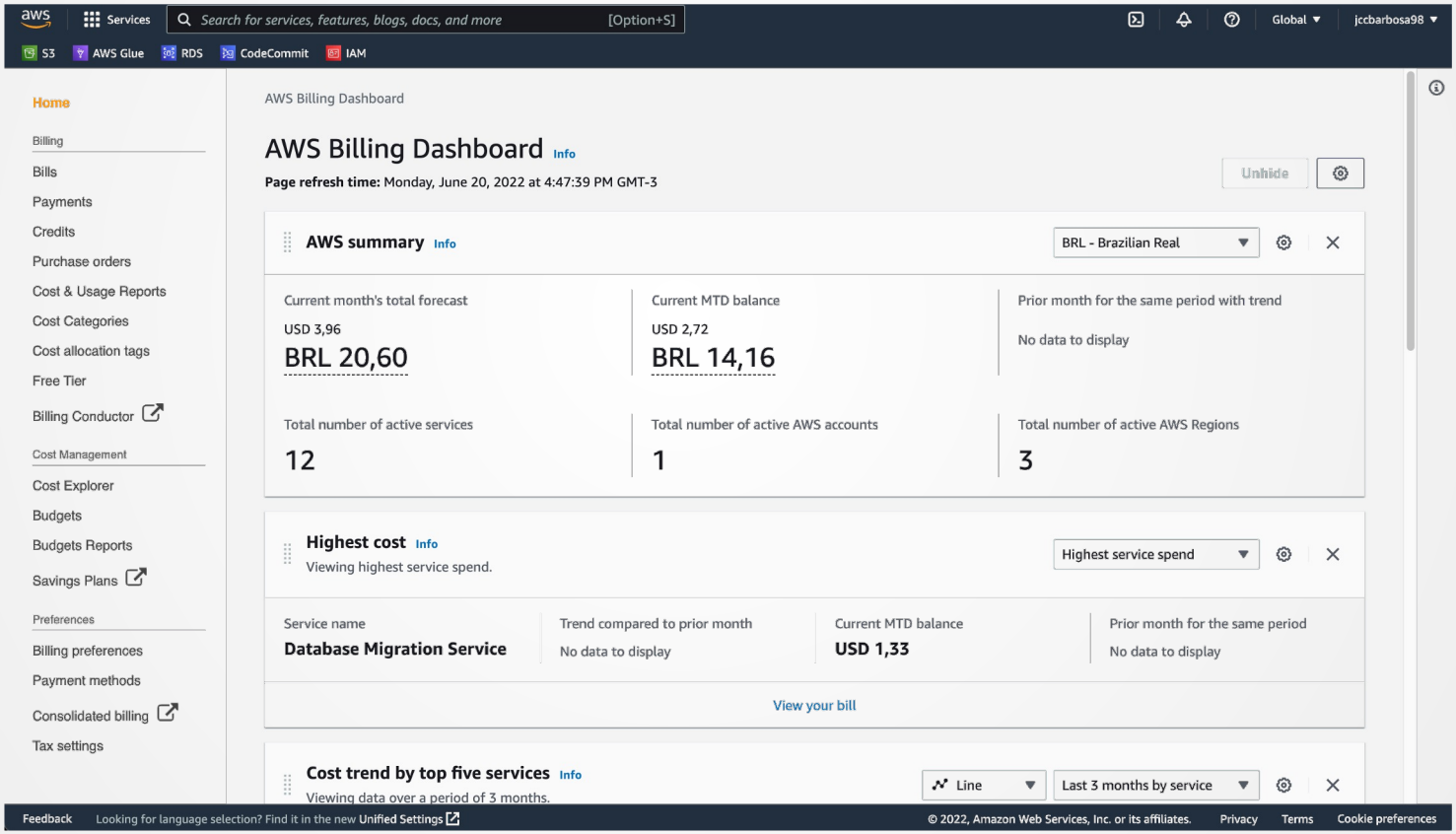
AWS CloudShell

AWS CloudShell is a browser-based shell that makes it easy to securely manage, explore, and interact with your AWS resources.



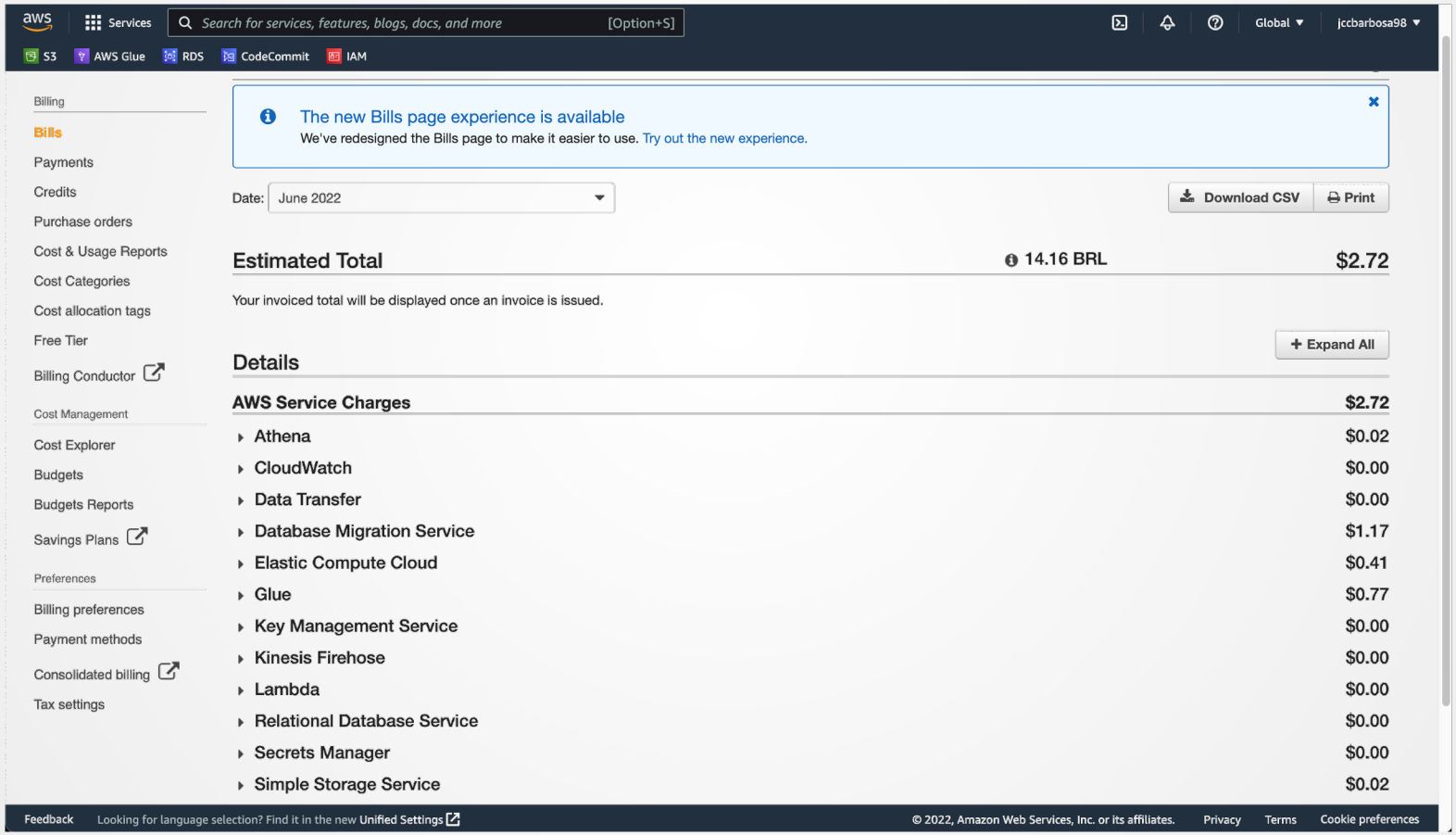
AWS Billing

The AWS Billing console contains features to pay your AWS bills, organize and report your AWS cost and usage, and manage your consolidated billing if you're a part of AWS Organizations.



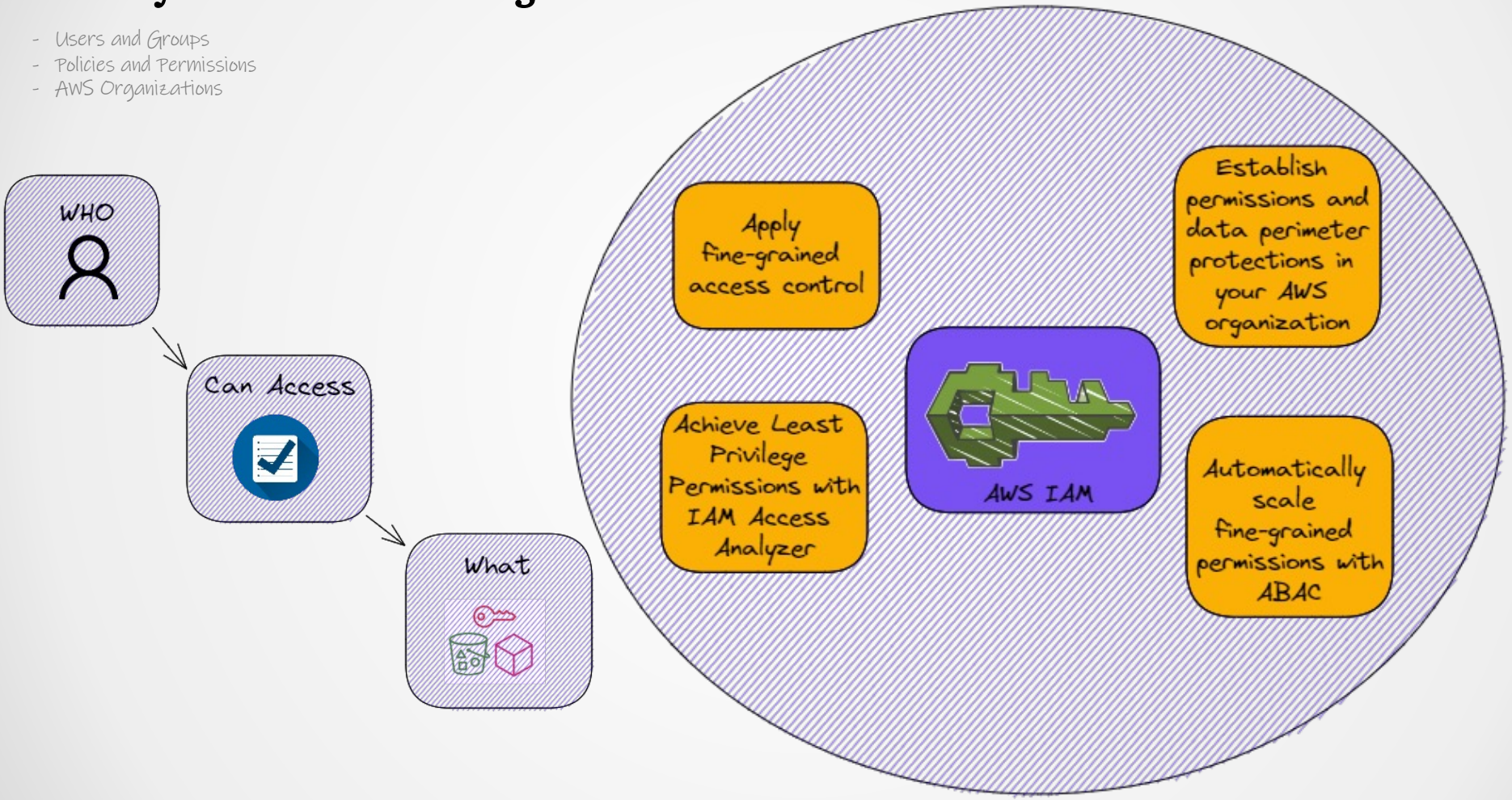
AWS Billing

The AWS Billing console contains features to pay your AWS bills, organize and report your AWS cost and usage, and manage your consolidated billing if you're a part of AWS Organizations.



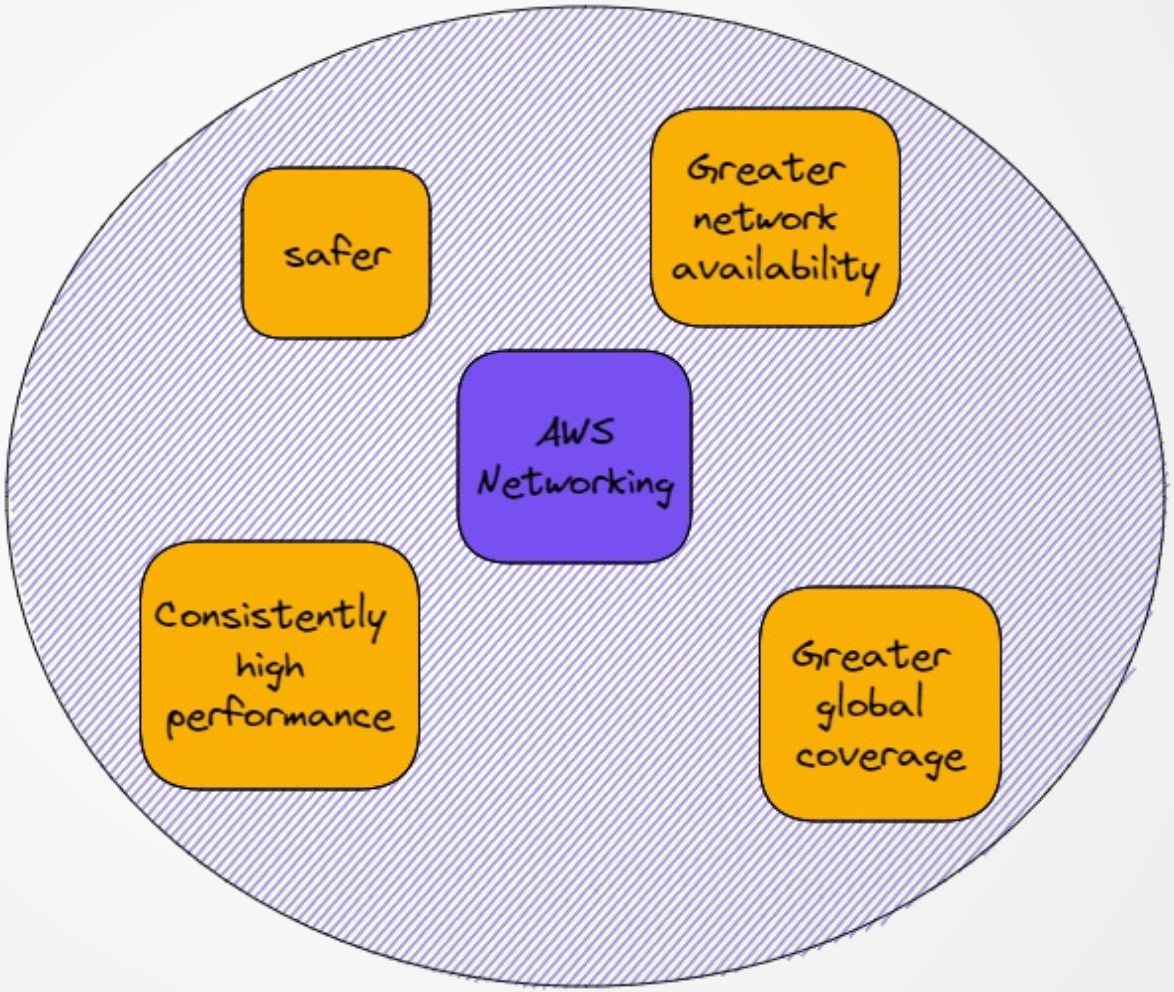
Identity and Access Management - IAM

- Users and Groups
- Policies and Permissions
- AWS Organizations



Fundamentals of Networking in AWS

- Amazon Virtual Private Cloud
- AWS Elastic Load Balancing
- Subnets
- Security Group



Fundamentals of Networking in AWS

- Amazon Virtual Private Cloud
- AWS Elastic Load Balancing
- Subnets
- Security Group

Amazon Virtual Private Cloud

- VPC to VPC peering
- VPC to On-premise data center
- Branch location to VPC connectivity
- Remote User to VPC based application
- Multicloud Peering (AWS VPC to Azure VNET or Google Cloud VPC)
- VPC to an Internet resource (VPC egress traffic)

Security Group

- Acts as a virtual firewall, controlling the traffic that is allowed to reach and leave the resources that it is associated with
- Can associate a security group only with resources in the VPC for which it is created
- Can add rules that control the traffic based on protocols and port numbers
- You might set up network ACLs with rules similar to your security groups in order to add an additional layer of security to your VPC

AWS Elastic Load Balancing

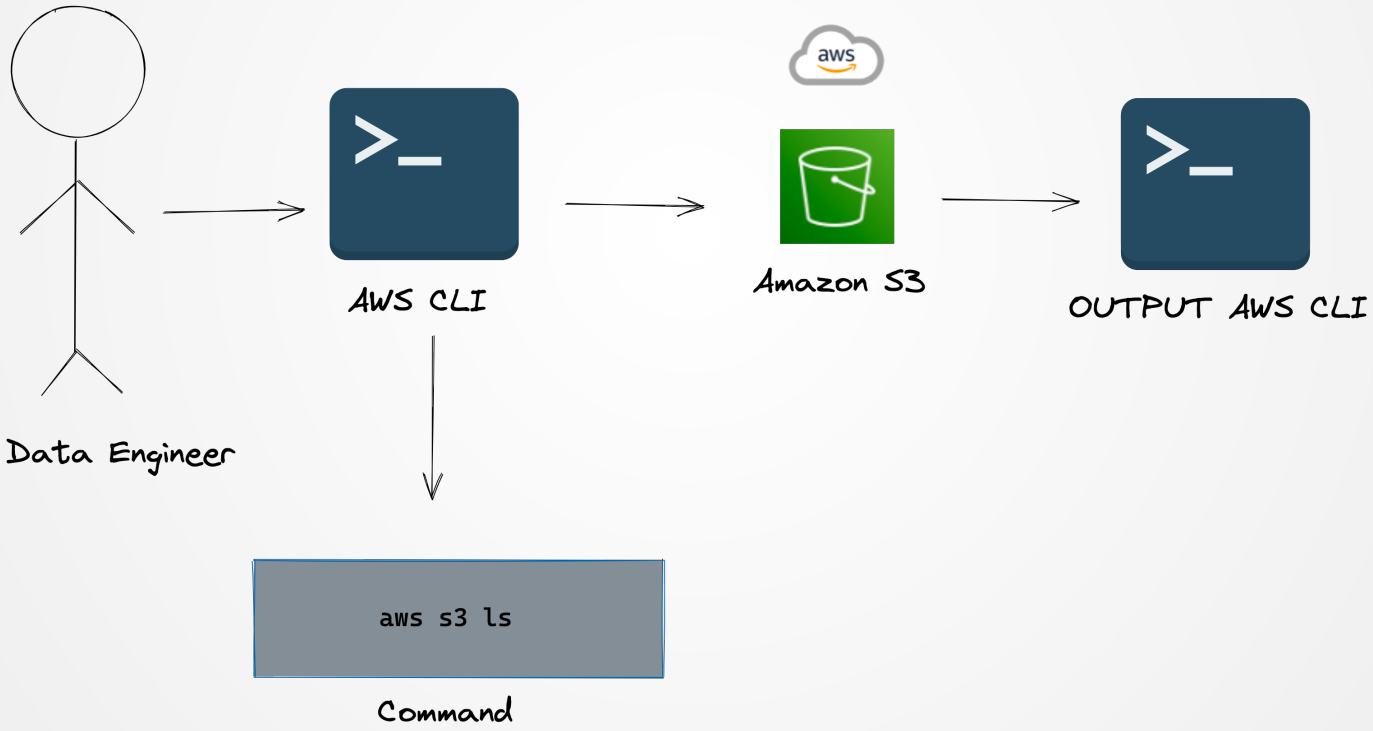
- Distributes workloads across multiple compute resources
- Add and remove compute resources from your load balancer as your needs change
- Configure health checks, which monitor the health of the compute resources, so that the load balancer sends requests only to the healthy ones

Subnets

- Public subnet: The subnet traffic is routed to the public internet through an internet gateway or an egress-only internet gateway
- Private subnet: The subnet traffic can't reach the public internet through an internet gateway or egress-only internet gateway. Access to the public internet requires a NAT device.
- VPN-only subnet: The subnet traffic is routed to a Site-to-Site VPN connection through a virtual private gateway. The subnet traffic can't reach the public internet through an internet gateway

AWS CLI

The AWS Command Line Interface (AWS CLI) is an open source tool that enables you to interact with AWS services using commands in your command-line shell.



AWS SDK for Python – Boto3

the AWS SDK for Python. Boto3 makes it easy to integrate your Python application, library, or script with AWS services including Amazon S3, Amazon EC2, Amazon DynamoDB, and more.

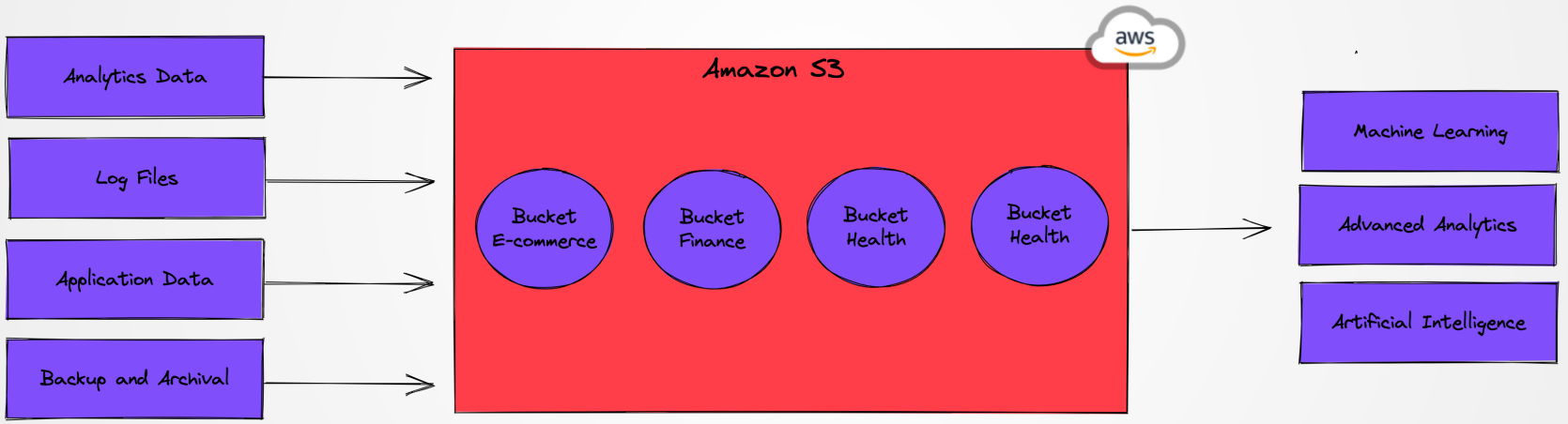
```python
import boto3

ec2 = boto3.client('ec2', region_name='us-east-2')

for instances in ec2.instances.all():
    if instances.state['Name'] == 'stopped':
        instances.start()
```

Amazon Simple Storage Service

Amazon Simple Storage Service (Amazon S3) is an object storage service offering industry-leading scalability, data availability, security, and performance.



Amazon Simple Storage Service 

Amazon Simple Storage Service (Amazon S3) is an object storage service offering industry-leading scalability, data availability, security, and performance.

S3 Standard

- Low latency and high throughput performance
- Designed for durability of 99.999999999% of objects across multiple Availability Zones
- Resilient against events that impact an entire Availability Zone
- Designed for 99.99% availability over a given year
- Backed with the Amazon S3 Service Level Agreement for availability
- Supports SSL for data in transit and encryption of data at rest
- S3 Lifecycle management for automatic migration of objects to other S3 Storage Classes

S3 One Zone-IA

- Same low latency and high throughput performance of S3 Standard
- Designed for durability of 99.999999999% of objects in a single Availability Zonet
- Designed for 99.5% availability over a given year
- Backed with the Amazon S3 Service Level Agreement for availability
- Supports SSL for data in transit and encryption of data at rest
- S3 Lifecycle management for automatic migration of objects to other S3 Storage Classes

S3 Intelligent-Tiering

- Frequent, Infrequent, and Archive Instant Access tiers have the same low-latency and high-throughput performance of S3 Standard
- The Infrequent Access tier saves up to 40% on storage costs
- The Archive Instant Access tier saves up to 68% on storage costs
- Opt-in asynchronous archive capabilities for objects that become rarely accessed
- Deep Archive Access tier has the same performance as Glacier Deep Archive and saves up to 95% for rarely accessed objects
- Designed for durability of 99.999999999% of objects across multiple Availability Zones and for 99.9% availability over a given year
- Backed with the Amazon S3 Service Level Agreement for availability
- Small monthly monitoring and auto tiering charge
- No operational overhead, no lifecycle charges, no retrieval charges, and no minimum storage duration
- Objects smaller than 128KB can be stored in S3 Intelligent-Tiering but will always be charged at the Frequent Access tier rates, and are not charged the monitoring and automation charge.

S3 Glacier Instant Retrieval

- Data retrieval in milliseconds with the same performance as S3 Standard
- Designed for durability of 99.999999999% of objects across multiple Availability Zones
- Data is resilient in the event of the destruction of one entire Availability Zone
- Designed for 99.9% data availability in a given year
- 128 KB minimum object size
- Backed with the Amazon S3 Service Level Agreement for availability
- S3 PUT API for direct uploads to S3 Glacier Instant Retrieval, and S3 Lifecycle management for automatic migration of objects

Amazon Simple Storage Service

Amazon Simple Storage Service (Amazon S3) is an object storage service offering industry-leading scalability, data availability, security, and performance.

S3 Glacier Flexible Retrieval

- Designed for durability of 99.999999999% of objects across multiple Availability Zones
- Data is resilient in the event of one entire Availability Zone destruction
- Supports SSL for data in transit and encryption of data at rest
- Ideal for backup and disaster recovery use cases when large sets of data occasionally need to be retrieved in minutes, without concern for costs
- Configurable retrieval times, from minutes to hours, with free bulk retrievals
- S3 PUT API for direct uploads to S3 Glacier Flexible Retrieval, and S3 Lifecycle management for automatic migration of objects

S3 Glacier Deep Archive

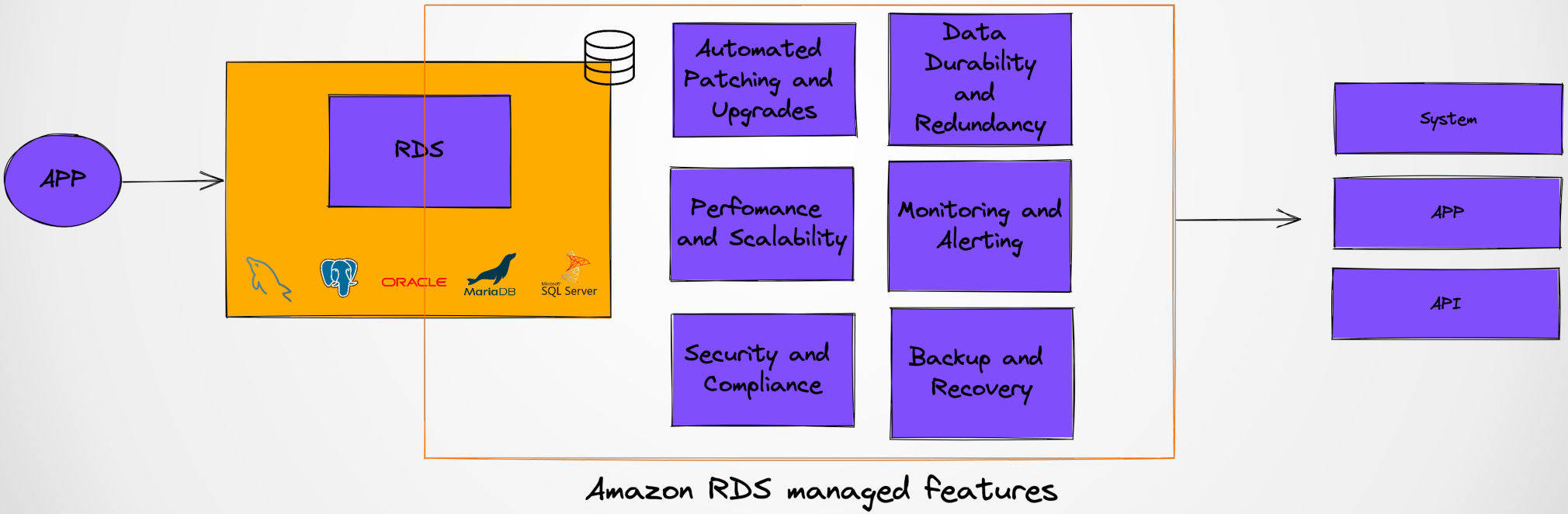
- Designed for durability of 99.999999999% of objects across multiple Availability Zones
- Lowest cost storage class designed for long-term retention of data that will be retained for 7-10 years
- Ideal alternative to magnetic tape libraries
- Retrieval time within 12 hours
- S3 PUT API for direct uploads to S3 Glacier Deep Archive, and S3 Lifecycle management for automatic migration of objects

S3 on Outposts

- S3 Object compatibility and bucket management through the S3 SDK
- Designed to durably and redundantly store data on your Outposts
- Encryption using SSE-S3 and SSE-C
- Authentication and authorization using IAM, and S3 Access Points
- Transfer data to AWS Regions using AWS DataSync
- S3 Lifecycle expiration actions

Amazon Relational Database Service

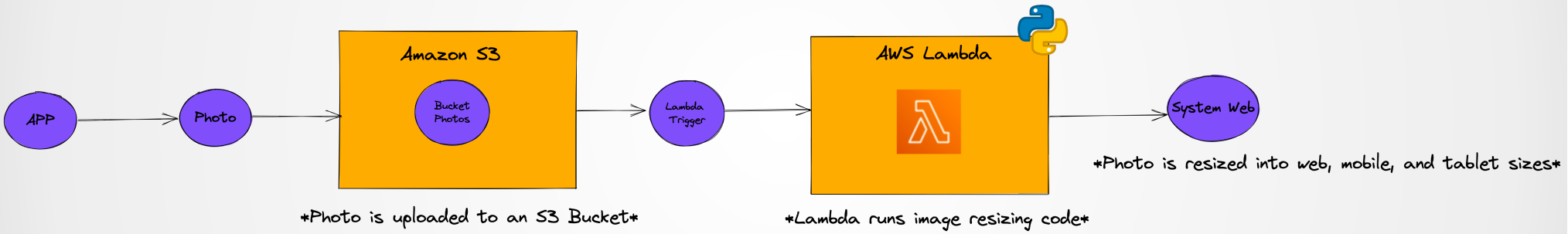
Amazon Relational Database Service (Amazon RDS) is a collection of managed services that makes it simple to set up, operate, and scale databases in the cloud. Choose from seven popular engines Amazon Aurora with MySQL compatibility, Amazon Aurora with PostgreSQL compatibility, MySQL, MariaDB, PostgreSQL, Oracle, and SQL Server and deploy on-premises with Amazon RDS on AWS Outposts.



Amazon RDS managed features

AWS Lambda

AWS Lambda is a serverless, event-driven compute service that lets you run code for virtually any type of application or backend service without provisioning or managing servers. You can trigger Lambda from over 200 AWS services and software as a service (SaaS) applications, and only pay for what you use.



AWS Lambda

AWS Lambda is a serverless, event-driven compute service that lets you run code for virtually any type of application or backend service without provisioning or managing servers. You can trigger Lambda from over 200 AWS services and software as a service (SaaS) applications, and only pay for what you use.

Bring your own code

- Lambda natively supports Java, Go, PowerShell, Node.js, C#, Python, and Ruby code, and provides a Runtime API allowing you to use any additional programming languages to author your functions.

Build custom backend services

- You can use AWS Lambda to create new backend application services triggered on demand using the Lambda application programming interface (API) or custom API endpoints built using Amazon API Gateway.

Completely automated administration

- AWS Lambda manages all the infrastructure to run your code on highly available, fault tolerant infrastructure, freeing you to focus on building differentiated backend services.
- With Lambda, you never have to update the underlying operating system (OS) when a patch is released, or worry about resizing or adding new servers as your usage grows.
- AWS Lambda seamlessly deploys your code, handles all the administration, maintenance, and security patches, and provides built-in logging and monitoring through Amazon CloudWatch.

Build custom backend services

- AWS Lambda maintains compute capacity across multiple Availability Zones (AZs) in each AWS Region to help protect your code against individual machine or data center facility failures.
- Both AWS Lambda and the functions running on the service deliver predictable and reliable operational performance.
- AWS Lambda is designed to provide high availability for both the service itself and the functions it operates. There are no maintenance windows or scheduled downtimes.

Automatic scaling

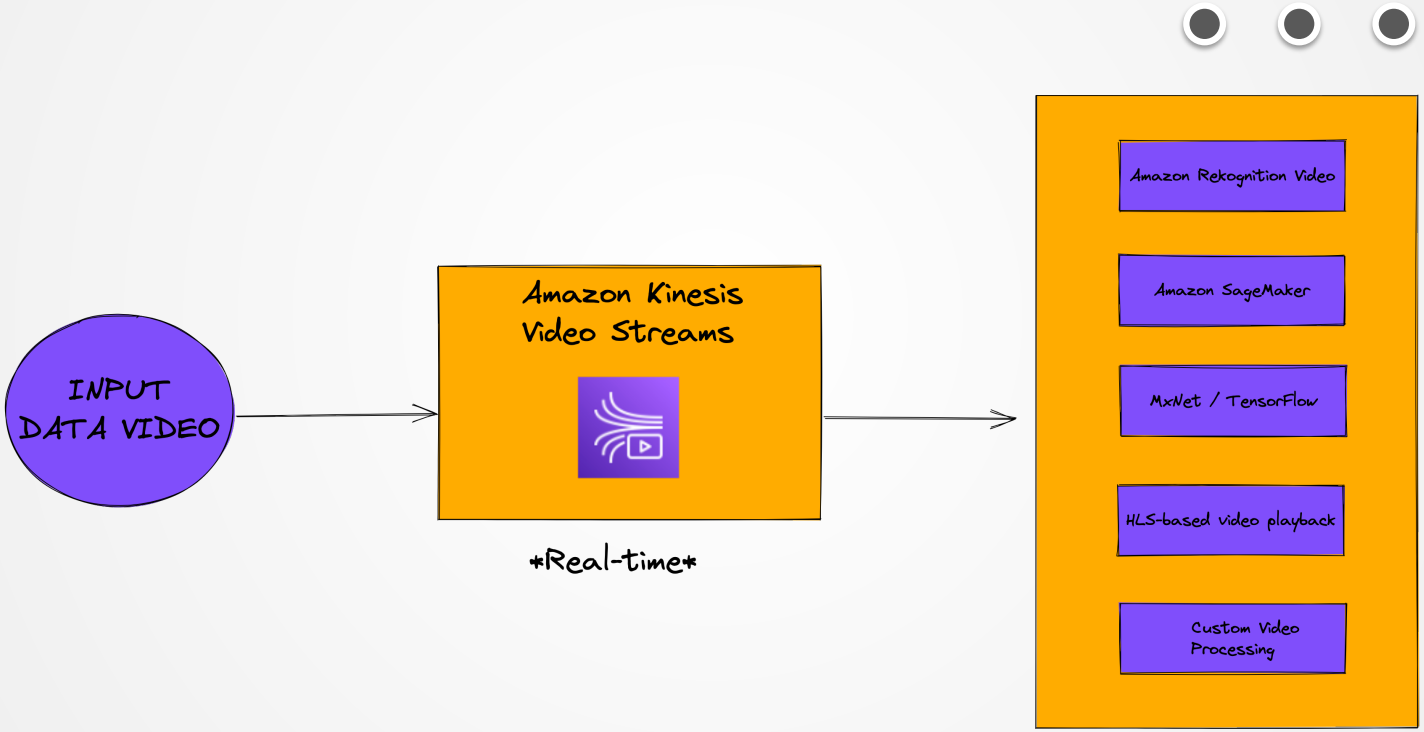
- AWS Lambda invokes your code only when needed, and automatically scales to support the rate of incoming requests without any manual configuration.
- There is no limit to the number of requests your code can handle. AWS Lambda typically starts running your code within milliseconds of an event.
- Since Lambda scales automatically, the performance remains consistently high as the event frequency increases.
- Since your code is stateless, Lambda can start as many instances as needed without lengthy deployment and configuration delays.

Only pay for what you use

- With AWS Lambda, you pay for execution duration rather than server unit.
- When using Lambda functions, you only pay for requests served and the compute time required to run your code.
- Billing is metered in increments of one millisecond, enabling easy and cost-effective automatic scaling from a few requests per day to thousands per second.
- With Provisioned Concurrency, you pay for the amount of concurrency you configure and the duration that you configure it.
- When Provisioned Concurrency is enabled and your function is executed, you also pay for requests and execution duration.

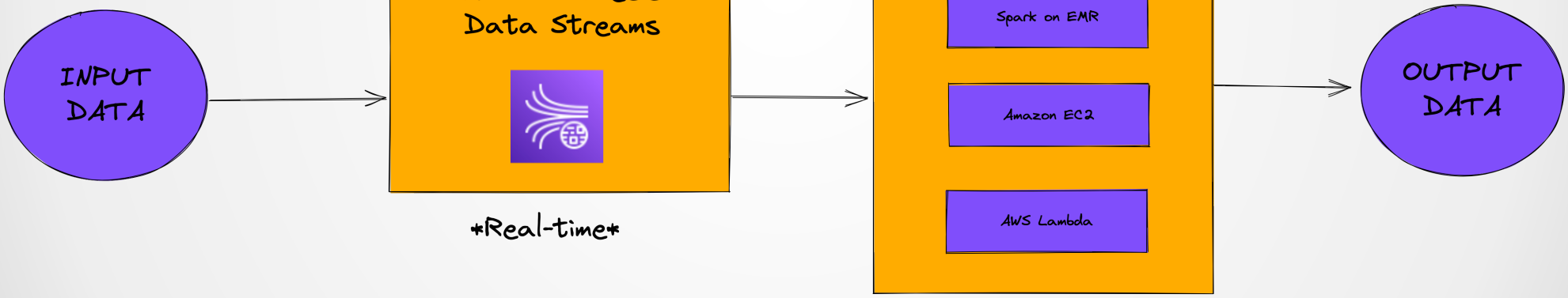
Amazon Kinesis Video Streams

Amazon Kinesis makes it easy to collect, process, and analyze real-time, streaming data so you can get timely insights and react quickly to new information.



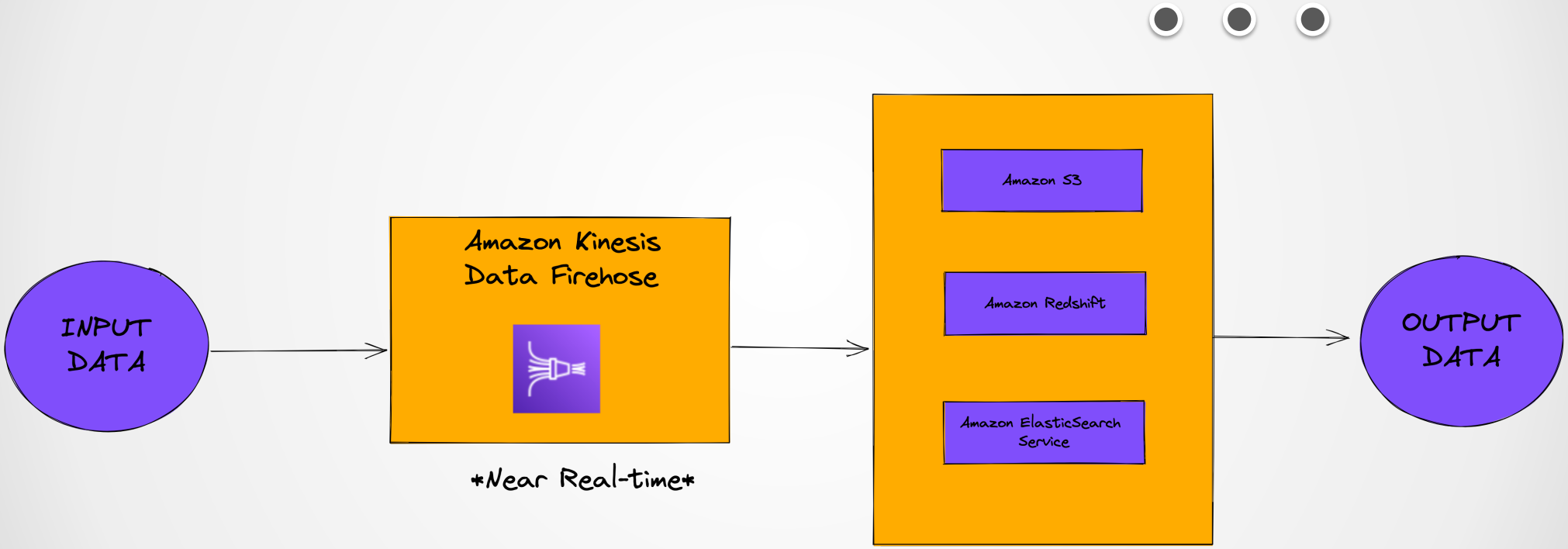
Amazon Kinesis Data Streams

Amazon Kinesis makes it easy to collect, process, and analyze real-time, streaming data so you can get timely insights and react quickly to new information.



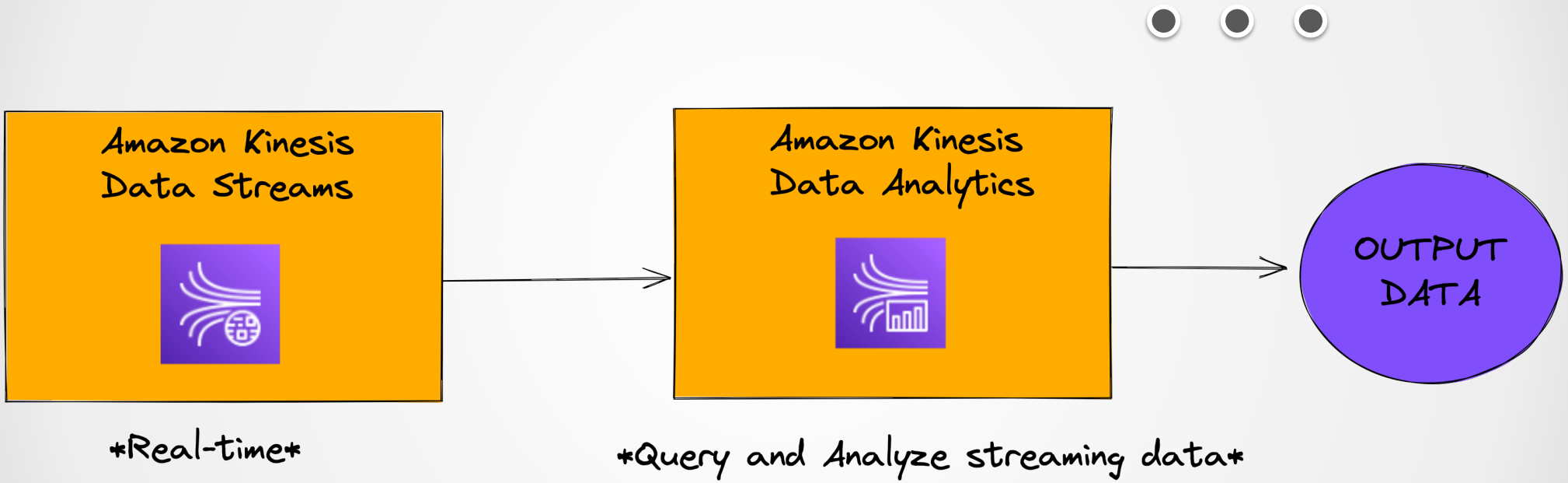
Amazon Kinesis Data Firehose

Amazon Kinesis makes it easy to collect, process, and analyze real-time, streaming data so you can get timely insights and react quickly to new information.



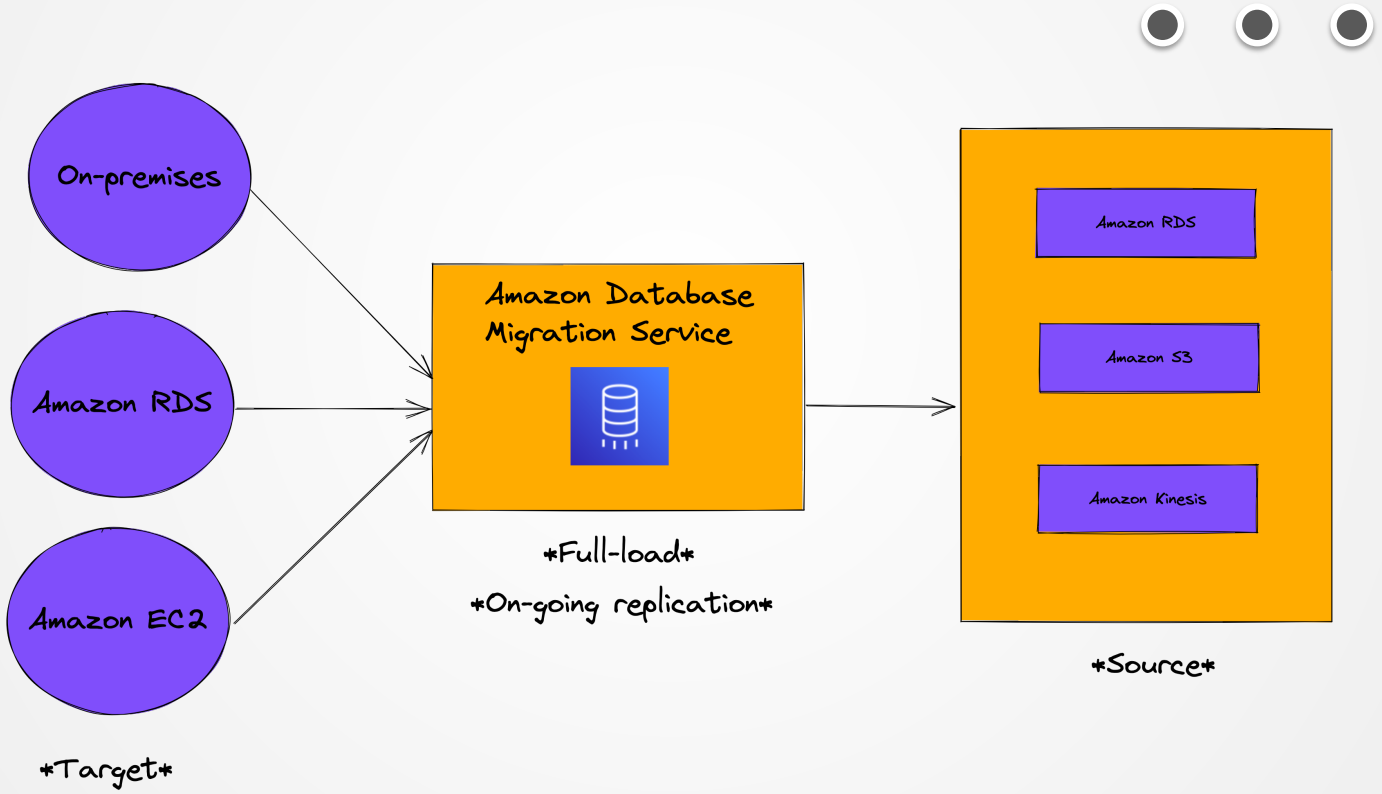
Amazon Kinesis Data Analytics

Amazon Kinesis makes it easy to collect, process, and analyze real-time, streaming data so you can get timely insights and react quickly to new information.



Amazon Database Migration Service

AWS Database Migration Service (AWS DMS) helps you migrate databases to AWS quickly and securely.



Hands on [Day 1]

Use case using AWS Lambda, Amazon RDS, Amazon S3, Kinesis Firehose, Amazon DMS technologies.