

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 - Fiabilidade
- Data Analyst Intern - Armazém PB
- DBA - Fábrica de Software

Personal

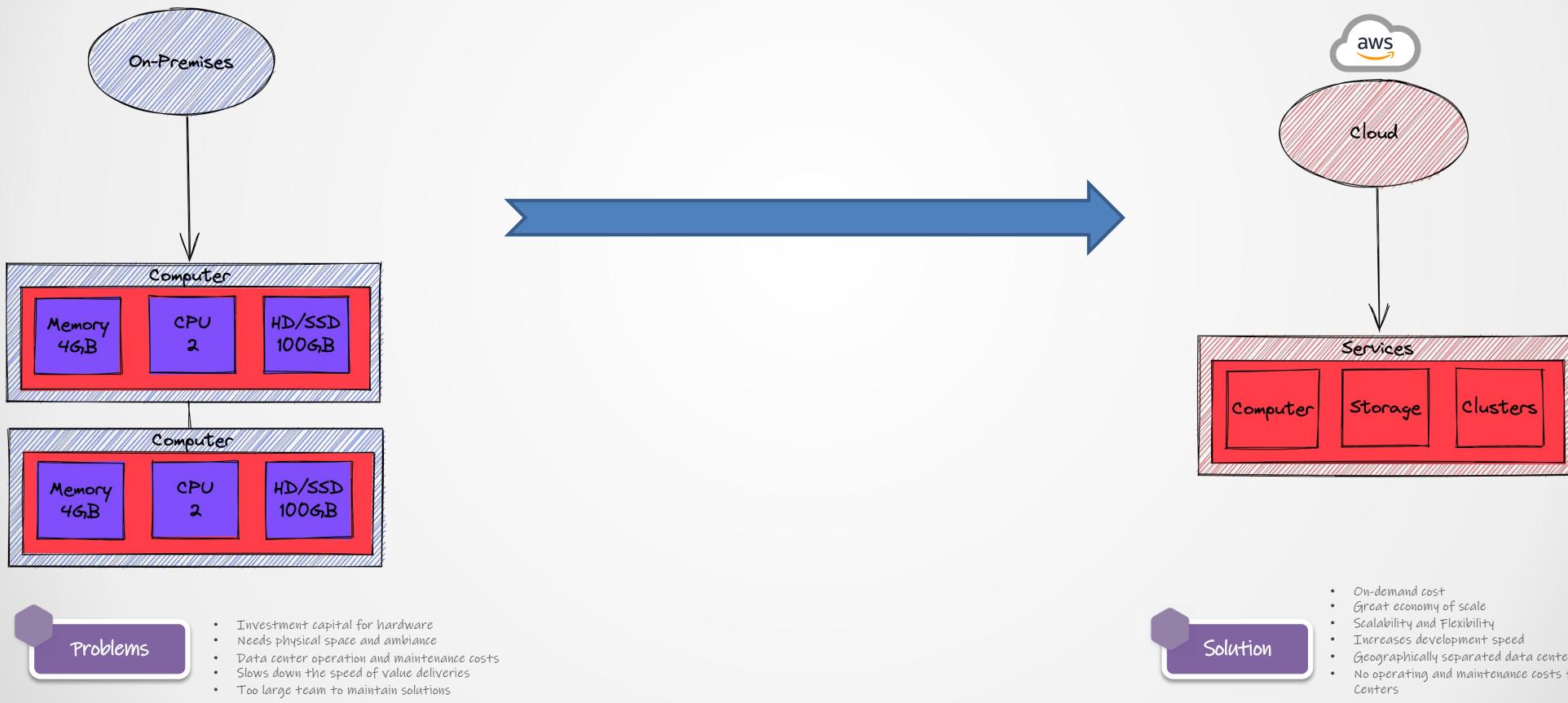
- 24 Years OLD
- Gamer
- Traveler
- Dog Lover (Spark) 🐶



What's Cloud Computing?

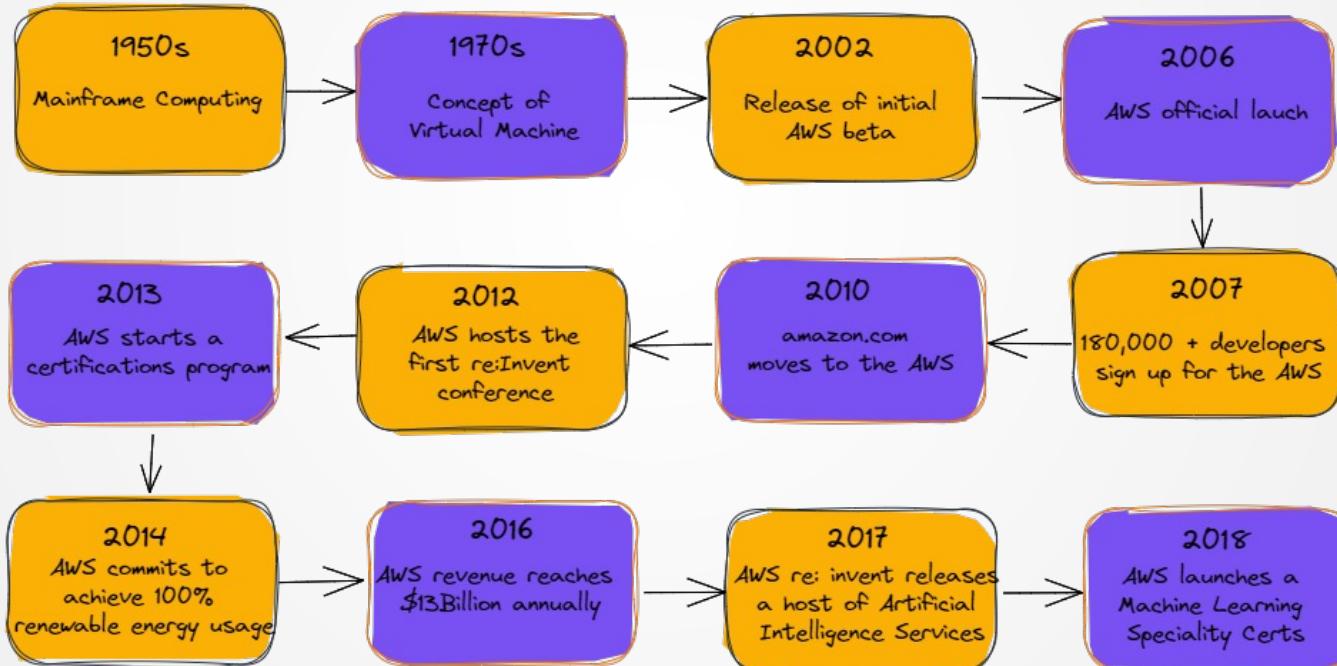
Simply put, cloud computing is the delivery of computing services including servers, storage, databases, networking, software, analytics, and intelligence.

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



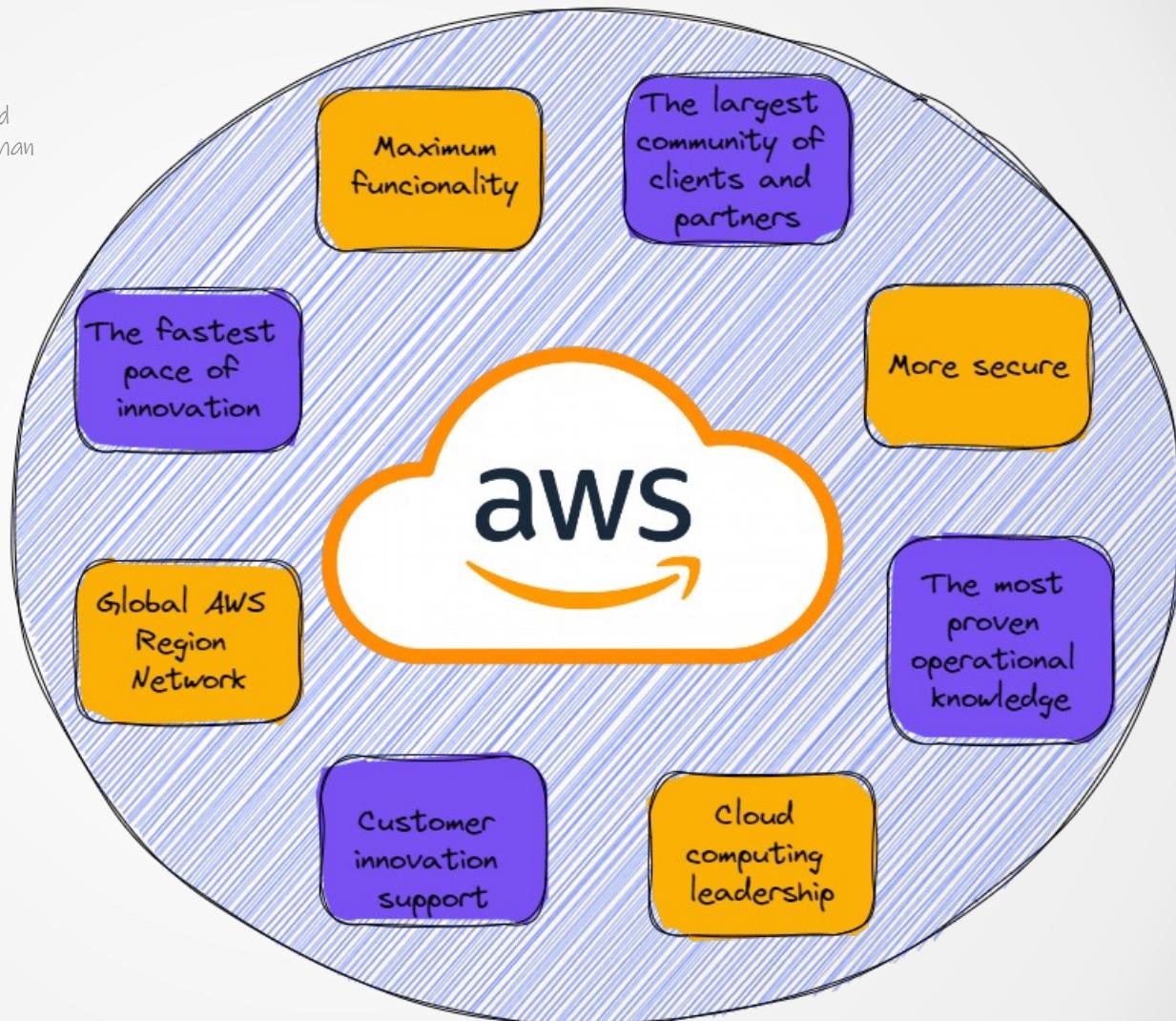
History

Breaf evolution of Cloud computing and Amazon Web Services through history



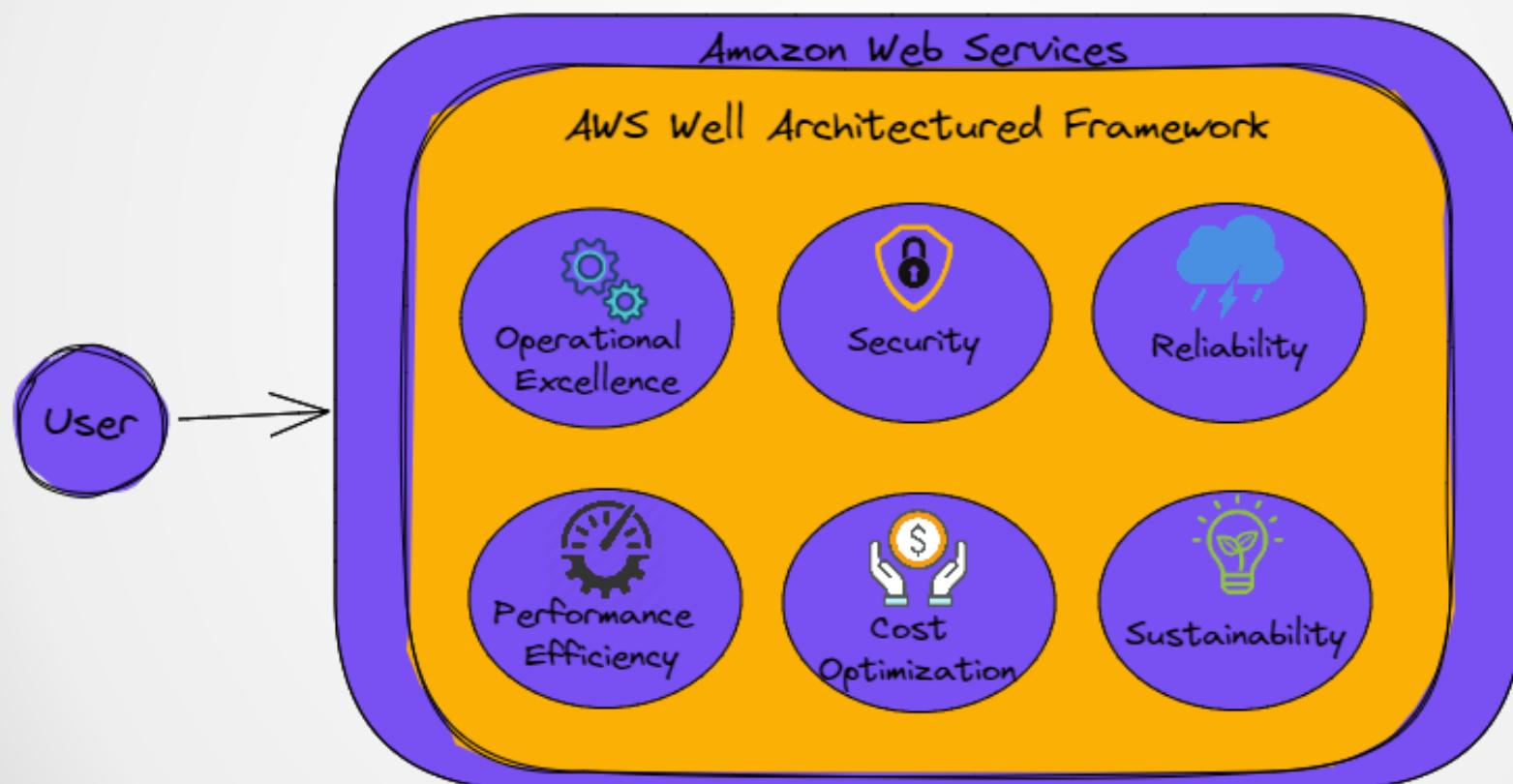
Amazon Web Services

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



Sign in as IAM user

Account ID (12 digits) or account alias

777696598735

IAM user name

carlos.barbosa

Password

.....

Remember this account

Sign in

[Sign in using root user email](#)

[Forgot password?](#)

Security Redefined with Amazon Neptune

Start today with 1 month free trial

[LEARN MORE](#)



English ▾

[Terms of Use](#) [Privacy Policy](#) © 1996-2022, Amazon Web Services, Inc. or its affiliates.

AWS Management Console

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

The screenshot shows the AWS Management Console Home page. At the top, there's a navigation bar with the AWS logo, a "Services" dropdown, a search bar ("Search for services, features, blogs, docs, and more"), and user information ("carlos.barbosa @ 7776-9659-8735"). Below the navigation bar, the main content area is divided into several sections:

- Recently visited:** A grid of recent service links:
 - Lambda
 - Athena
 - EC2
 - AWS Glue
 - EMR
 - CloudWatch
 - Kinesis
 - AWS Glue DataBrew
 - RDS
 - S3
 - CloudFormation
- Welcome to AWS:** A section with three cards:
 - Getting started with AWS**: Learn the fundamentals and find valuable information to get the most out of AWS.
 - Training and certification**: Learn from AWS experts and advance your skills and knowledge.
 - What's new with AWS?**: Discover new AWS services, features, and Regions.
- AWS Health:** A summary of open issues and scheduled changes.

Open issues	Past 7 days
0	

Scheduled changes	Upcoming and past 7 days
0	
- Cost and usage:** A summary showing "No cost and usage".

At the bottom of the page, there are footer links: Feedback, Looking for language selection? Find it in the new Unified Settings, © 2022, Amazon Web Services, Inc. or its affiliates., Privacy, Terms, and Cookie preferences.

AWS Management Console

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

The screenshot shows the AWS Management Console search results for the term 'S3'. The search bar at the top contains 'S3'. Below it, the results are categorized into 'Services' and 'Features'.

Services

- S3 ★ Scalable Storage in the Cloud
- S3 Glacier ☆ Archive Storage in the Cloud
- Athena ★ Query Data in S3 using SQL
- AWS Snow Family ☆ Large Scale Data Transport

Features

- Amazon S3 File Gateway Storage Gateway feature
- Datasets IoT Analytics feature
- Batch Operations S3 feature

At the bottom right of the search results, there is a note: "No cost and usage".

On the far right, there is a sidebar with sections like "AWS Fundamentals", "AWS certification", and "What's new with AWS?".

At the very bottom of the page, there are links for "Feedback", "Unified Settings", "© 2022, Amazon Web Services, Inc. or its affiliates.", "Privacy", "Terms", and "Cookie preferences".

AWS Management Console

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

The screenshot shows the AWS Management Console with the Amazon S3 service selected. The left sidebar contains navigation links for Buckets, Storage Lens, and Marketplace. The main content area displays an account snapshot with total storage of 310.2 MB, object count of 1.1 k, and average object size of 300.8 KB. It also includes a link to enable advanced metrics. Below this is a table listing two buckets: "owshq-live-amazon-athena-777696598735" and "owshq-live-amazon-athena-results-777696598735", both created on June 15, 2022.

Name	AWS Region	Access	Creation date
owshq-live-amazon-athena-777696598735	US East (Ohio) us-east-2	Bucket and objects not public	June 15, 2022, 14:20:57 (UTC-03:00)
owshq-live-amazon-athena-results-777696598735	US East (Ohio) us-east-2	Bucket and objects not public	June 15, 2022, 14:21:34 (UTC-03:00)

Buckets (2) Info
Buckets are containers for data stored in S3. [Learn more](#)

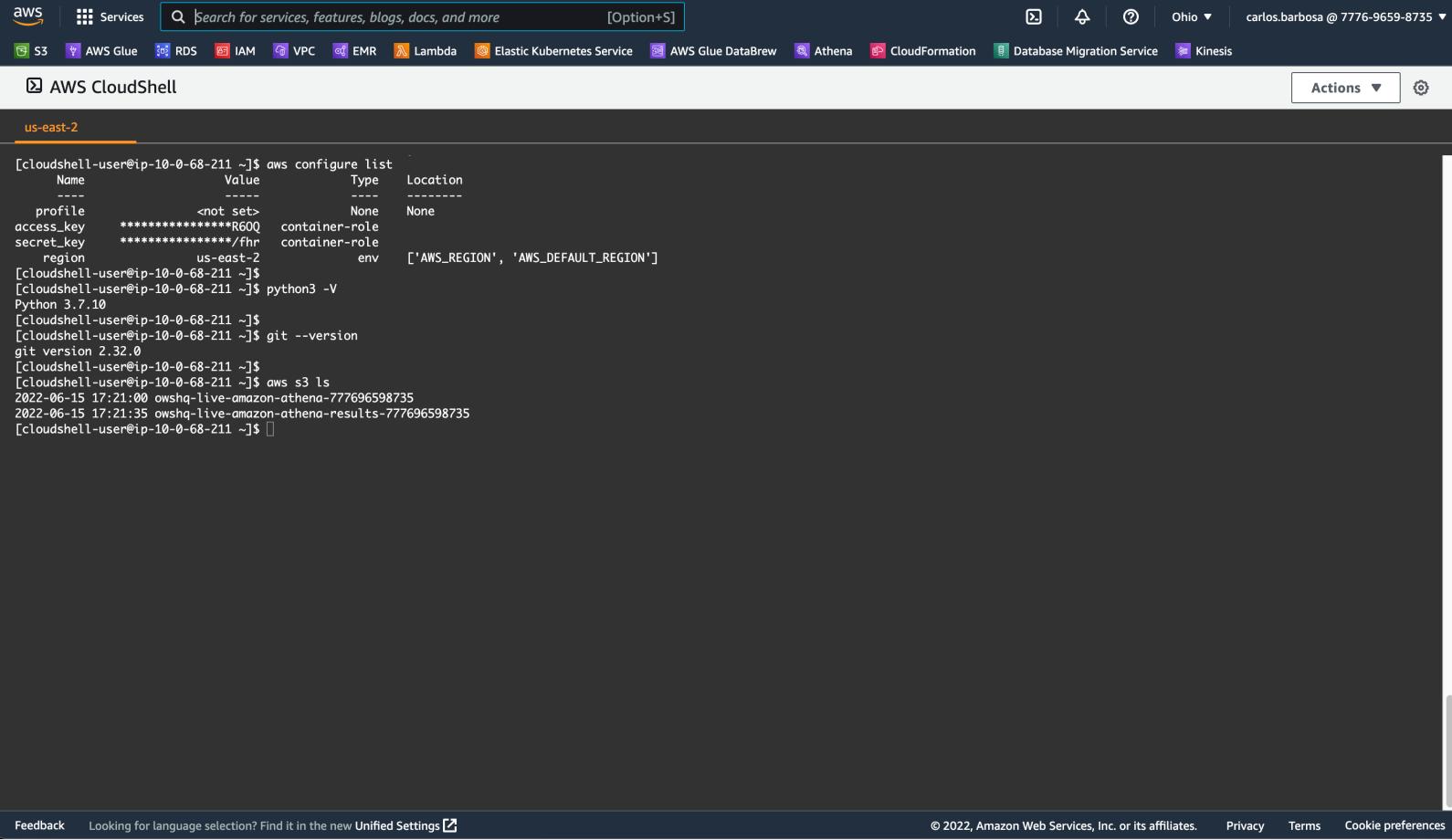
Find buckets by name

Feedback Looking for language selection? Find it in the new [Unified Settings](#)

© 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

AWS CloudShell

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



The screenshot shows the AWS CloudShell interface. At the top, there's a navigation bar with the AWS logo, a "Services" dropdown, a search bar ("Search for services, features, blogs, docs, and more"), and a "[Option+S]" keybinding. To the right are icons for S3, AWS Glue, RDS, IAM, VPC, EMR, Lambda, Elastic Kubernetes Service, AWS Glue DataBrew, Athena, CloudFormation, Database Migration Service, and Kinesis. The region is set to "Ohio" (indicated by a "▼" icon). On the far right, there are "Actions" and settings icons.

The main area is a terminal window titled "AWS CloudShell" with the region "us-east-2" selected. The terminal output shows a user session:

```
[cloudshell-user@ip-10-0-68-211 ~]$ aws configure list
Name          Value        Type    Location
----          ----        --      -----
profile       <not set>    None   None
access_key    ****REDACTED****R6Q
secret_key    ****REDACTED****/fhr
region        us-east-2   env    ['AWS_REGION', 'AWS_DEFAULT_REGION']
[cloudshell-user@ip-10-0-68-211 ~]$
[cloudshell-user@ip-10-0-68-211 ~]$ python3 -V
Python 3.7.10
[cloudshell-user@ip-10-0-68-211 ~]$
[cloudshell-user@ip-10-0-68-211 ~]$ git --version
git version 2.32.0
[cloudshell-user@ip-10-0-68-211 ~]$
[cloudshell-user@ip-10-0-68-211 ~]$ aws s3 ls
2022-06-15 17:21:00 owhq-live-amazon-athena-777696598735
2022-06-15 17:21:35 owhq-live-amazon-athena-results-777696598735
[cloudshell-user@ip-10-0-68-211 ~]$ ]
```

At the bottom of the terminal window, there's a scroll bar. The footer contains links for "Feedback", "Unified Settings", "Privacy", "Terms", and "Cookie preferences".

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.

The screenshot shows the AWS Billing Dashboard. On the left, there's a sidebar with navigation links for Home, Billing, Bills, Payments, Credits, Purchase orders, Cost & Usage Reports, Cost Categories, Cost allocation tags, Free Tier, Billing Conductor, Cost Management, Cost Explorer, Budgets, Budgets Reports, Savings Plans, Preferences, Billing preferences, Payment methods, Consolidated billing, and Tax settings. The main content area displays the AWS Billing Dashboard with the following data:

AWS summary		Prior month for the same period with trend
Current month's total forecast USD 3,96 BRL 20,60	Current MTD balance USD 2,72 BRL 14,16	No data to display
Total number of active services 12	Total number of active AWS accounts 1	Total number of active AWS Regions 3

Highest cost (Info) - Viewing highest service spend.

Service name	Trend compared to prior month	Current MTD balance	Prior month for the same period
Database Migration Service	No data to display	USD 1,33	No data to display

Cost trend by top five services (Info) - Viewing data over a period of 3 months.

Feedback: Looking for language selection? Find it in the new Unified Settings.

© 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

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.

The screenshot shows the AWS Billing console interface. On the left, there's a sidebar with a tree view of billing-related services: Billing, Bills (which is selected and highlighted in orange), Payments, Credits, Purchase orders, Cost & Usage Reports, Cost Categories, Cost allocation tags, Free Tier, Billing Conductor (with a plus icon), Cost Management, Cost Explorer, Budgets, Budgets Reports, Savings Plans (with a plus icon), Preferences, Billing preferences, Payment methods, Consolidated billing (with a plus icon), and Tax settings. A search bar at the top has placeholder text "Search for services, features, blogs, docs, and more" and a keyboard shortcut "[Option+S]". To the right of the search bar are icons for Help, Global, and a user profile (jcbarbosa98).

The main content area displays a message: "The new Bills page experience is available. We've redesigned the Bills page to make it easier to use. Try out the new experience." Below this, there's a date selector set to "June 2022" and buttons for "Download CSV" and "Print".

A section titled "Estimated Total" shows a breakdown of charges:

14.16 BRL	\$2.72
Your invoiced total will be displayed once an invoice is issued.	

Details

+ Expand All

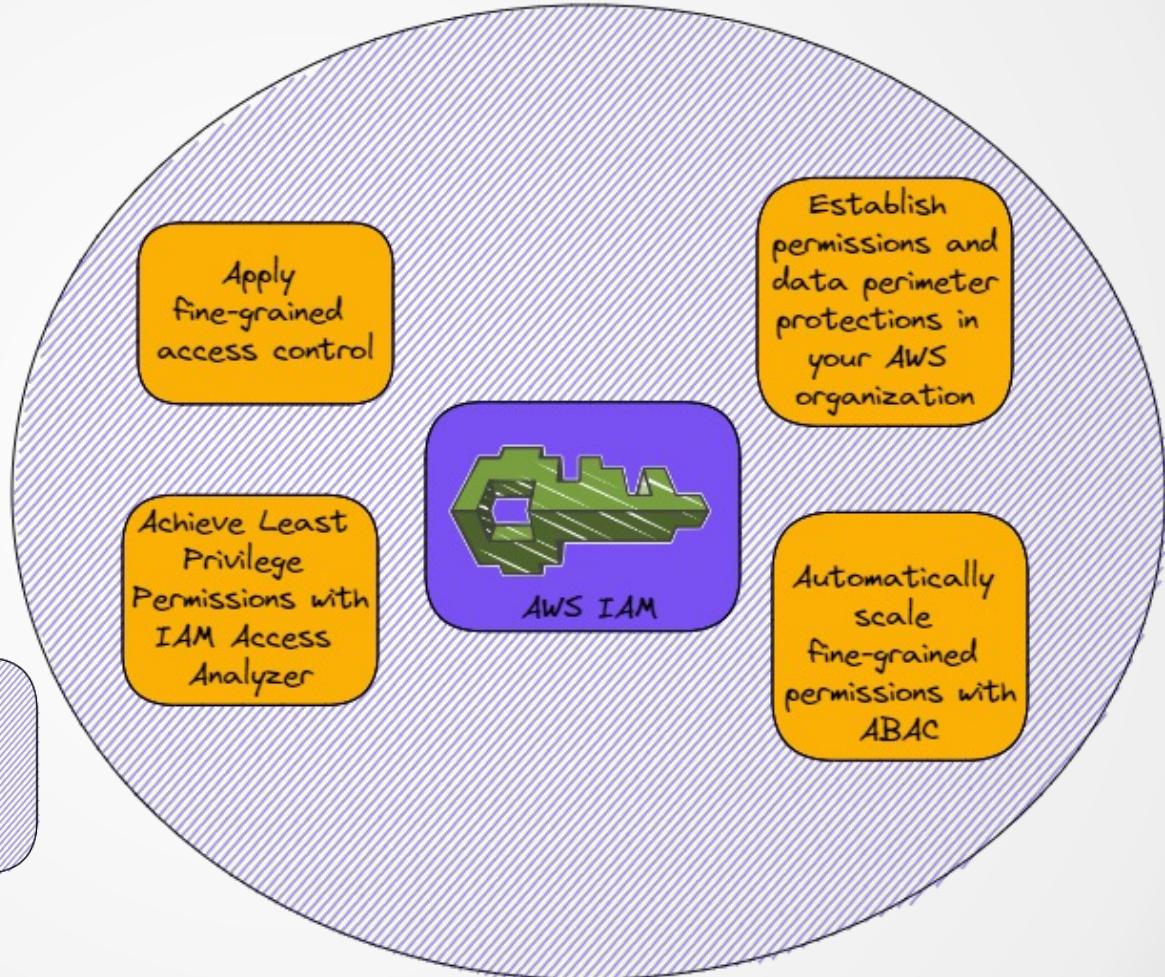
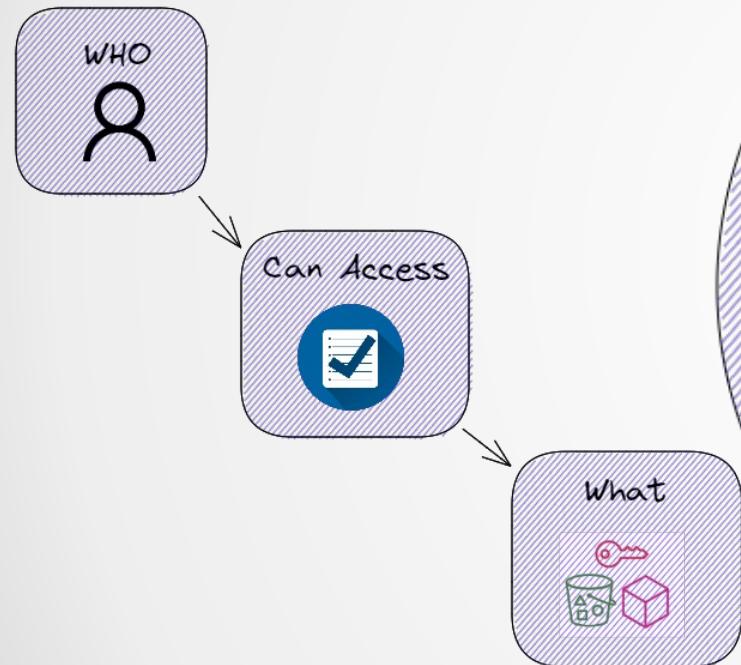
AWS Service Charges	\$2.72
▶ Athena	\$0.02
▶ CloudWatch	\$0.00
▶ Data Transfer	\$0.00
▶ Database Migration Service	\$1.17
▶ Elastic Compute Cloud	\$0.41
▶ Glue	\$0.77
▶ Key Management Service	\$0.00
▶ Kinesis Firehose	\$0.00
▶ Lambda	\$0.00
▶ Relational Database Service	\$0.00
▶ Secrets Manager	\$0.00
▶ Simple Storage Service	\$0.02

Feedback Looking for language selection? Find it in the new Unified Settings [\[?\]](#)

© 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

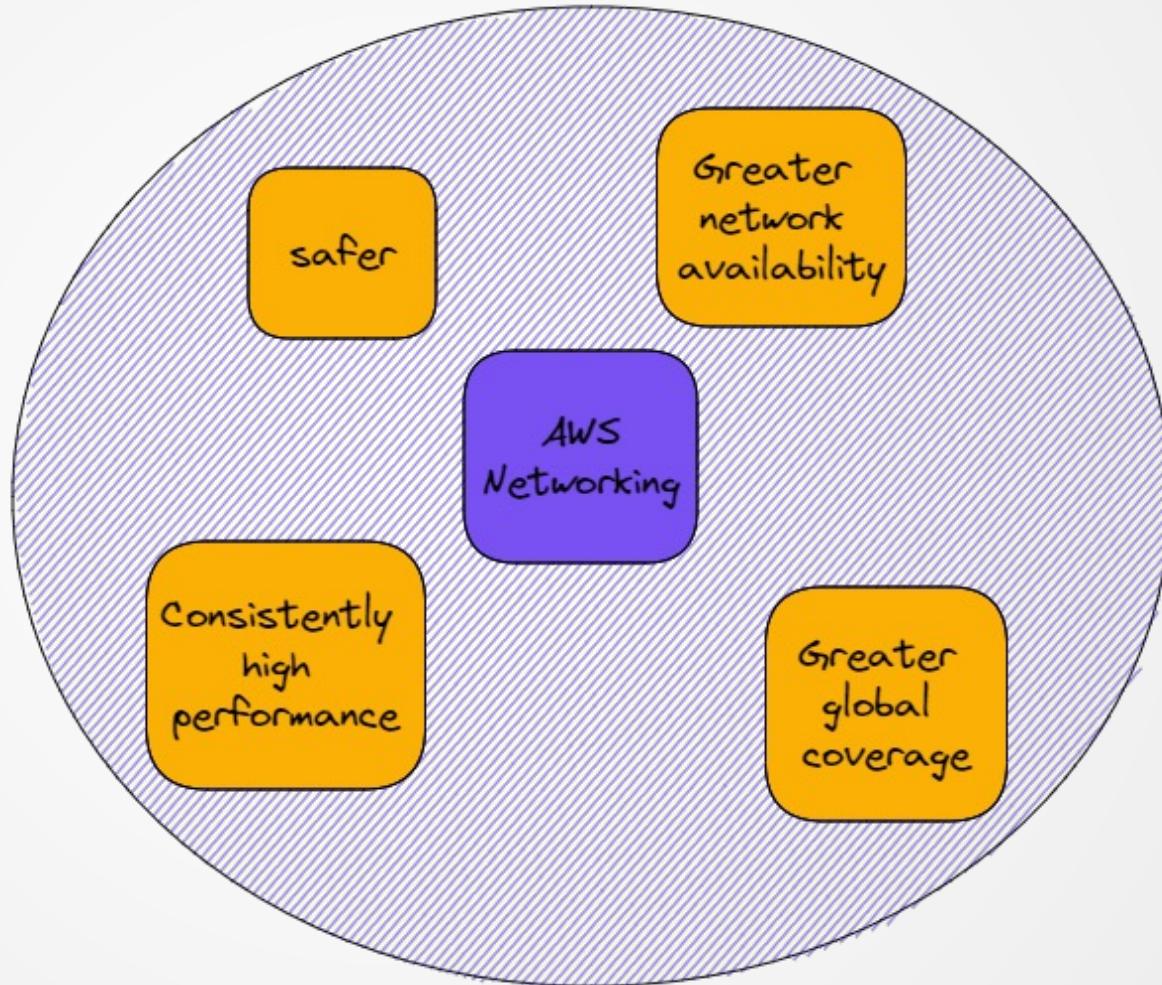
Identity and Access Management - IAM

- Users and Groups
- Policies and Permissions
- AWS Organizations



Fundamentals of Networking in AWS

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



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)

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

Security Group

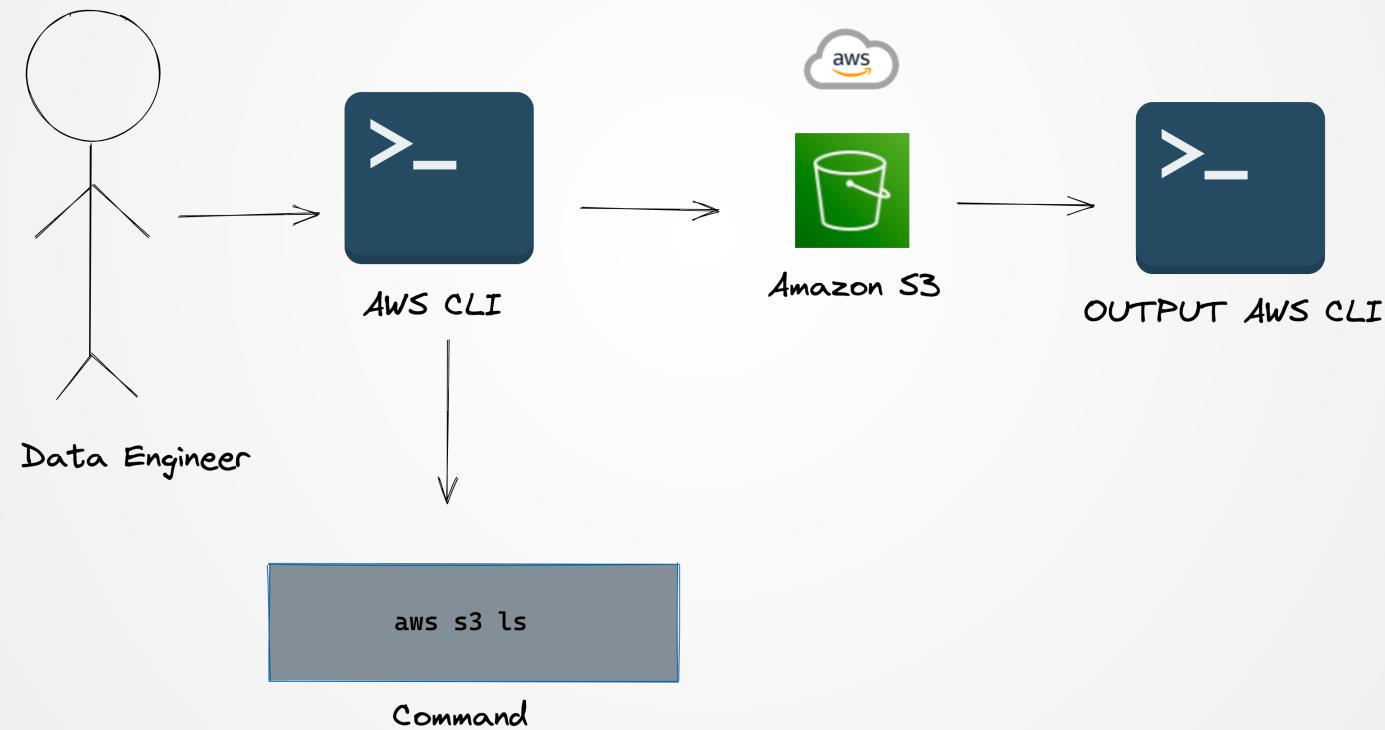
- Acts as a virtual firewall, controlling the traffic that is allowed to reach and leave the resources that it is associated with
- Can associate a security group only with resources in the VPC for which it is created
- Can add rules that control the traffic based on protocols and port numbers
- You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

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.



```
import boto3

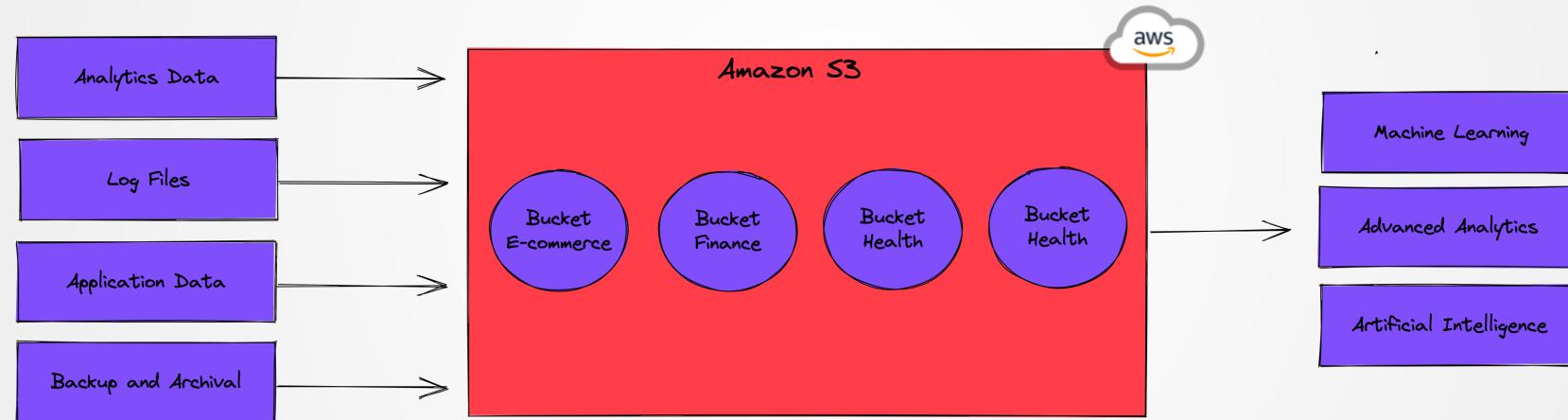
ec2 = boto3.client('ec2', region_name='us-east-2')

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 99.99999999% 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 99.99999999% of objects in a single Availability Zone
- Designed for 99.9% 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 99.99999999% 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 99.99999999% 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

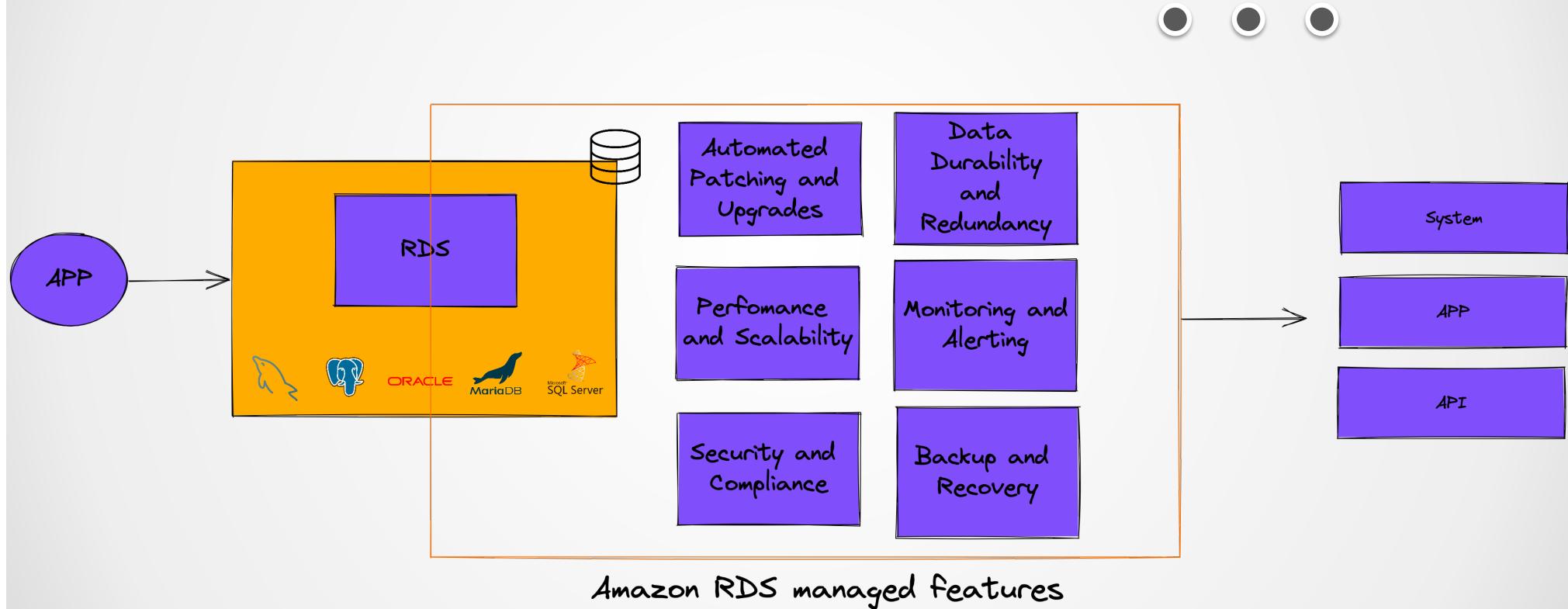
S3 on Outposts

- 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

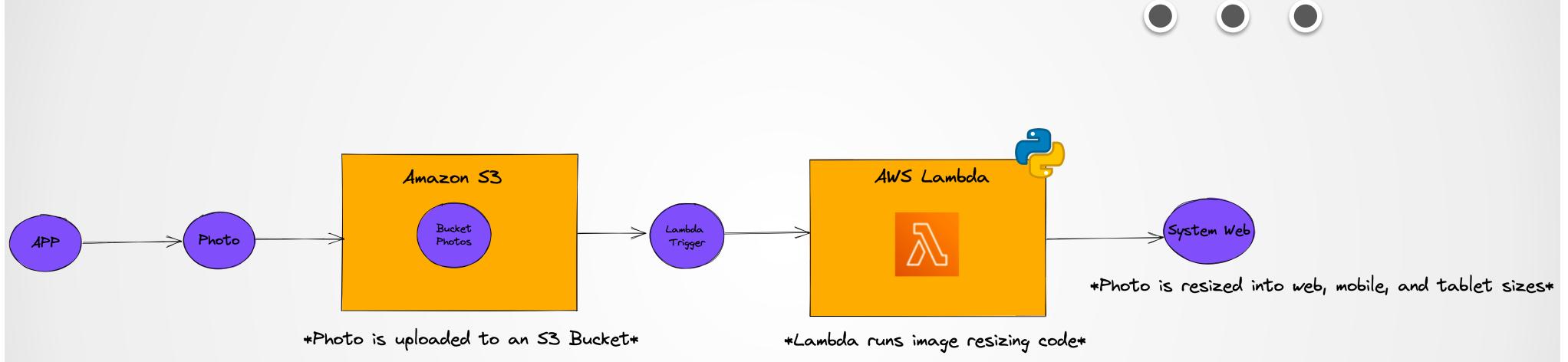


Amazon Relational Database Service (Amazon RDS) is a collection of managed services that makes it simple to set up, operate, and scale databases in the cloud. Choose from seven popular engines [Amazon Aurora with MySQL compatibility](#), [Amazon Aurora with PostgreSQL compatibility](#), [MySQL](#), [MariaDB](#),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



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.

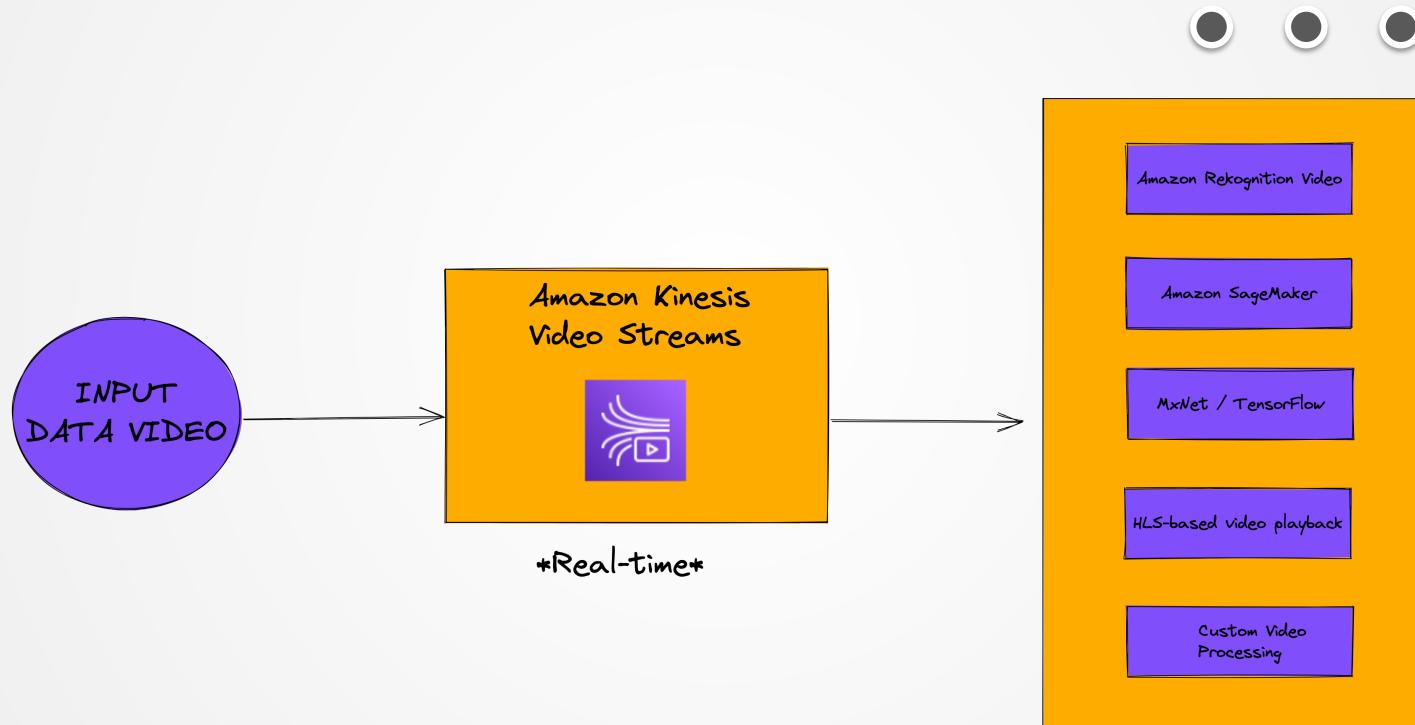
Only pay for what you use

- 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



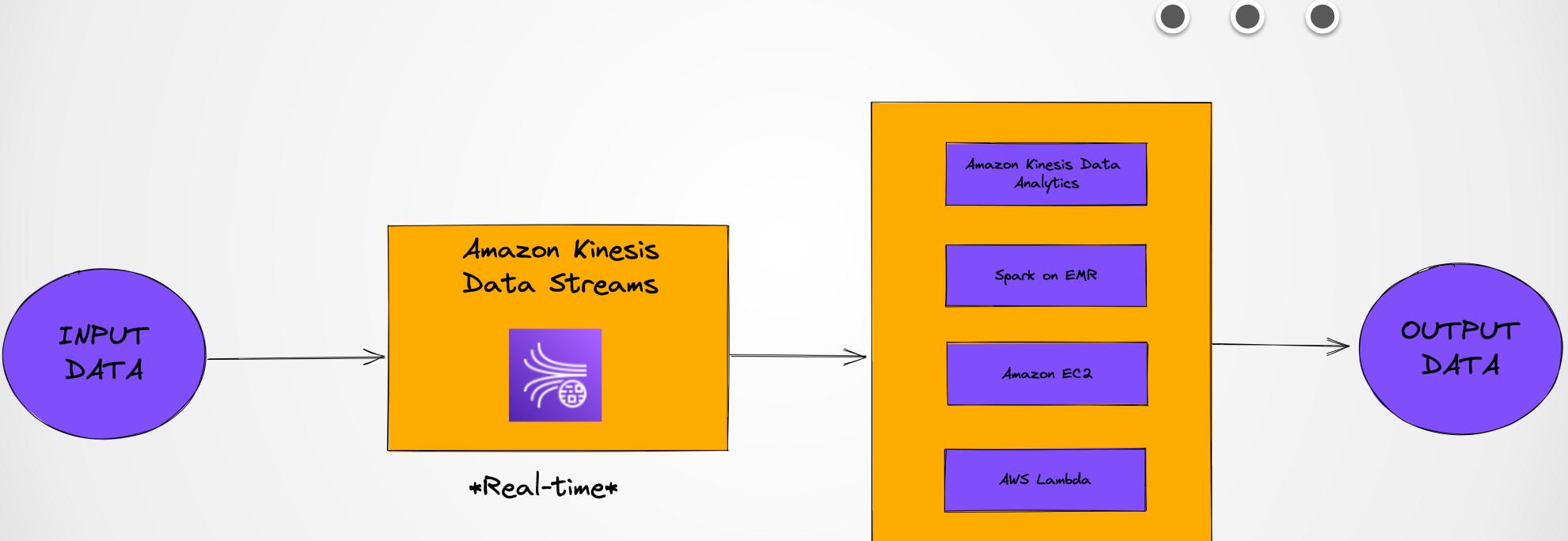
Amazon Kinesis makes it easy to collect, process, and analyze real-time, streaming data so you can get timely insights and react quickly to new information.



Amazon Kinesis Data Streams



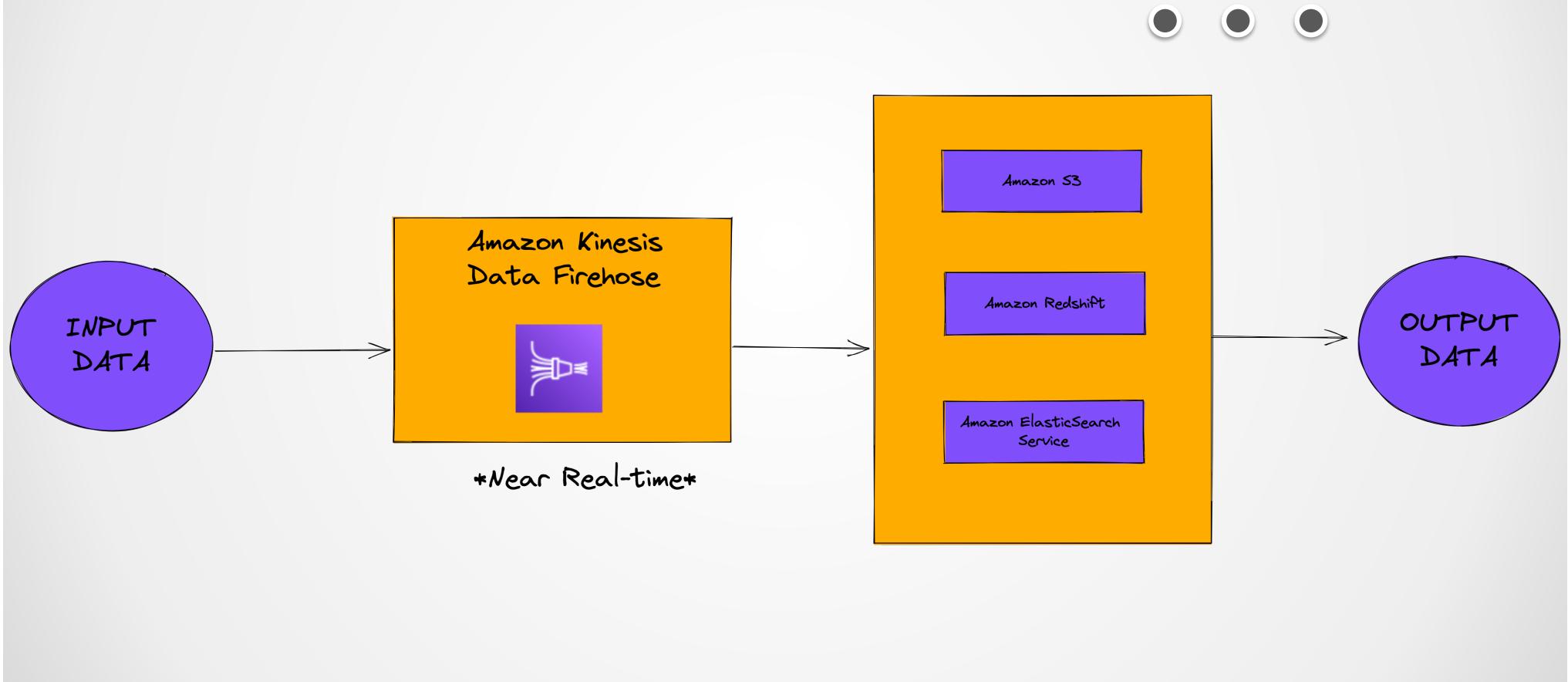
Amazon Kinesis makes it easy to collect, process, and analyze real-time, streaming data so you can get timely insights and react quickly to new information.



Amazon Kinesis Data Firehose



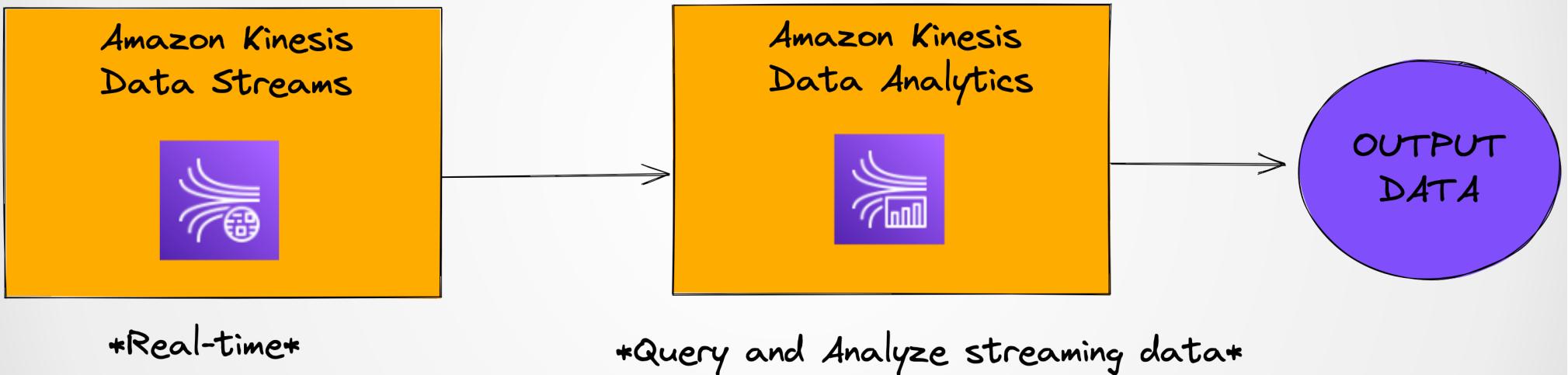
Amazon Kinesis makes it easy to collect, process, and analyze real-time, streaming data so you can get timely insights and react quickly to new information.



Amazon Kinesis Data Analytics



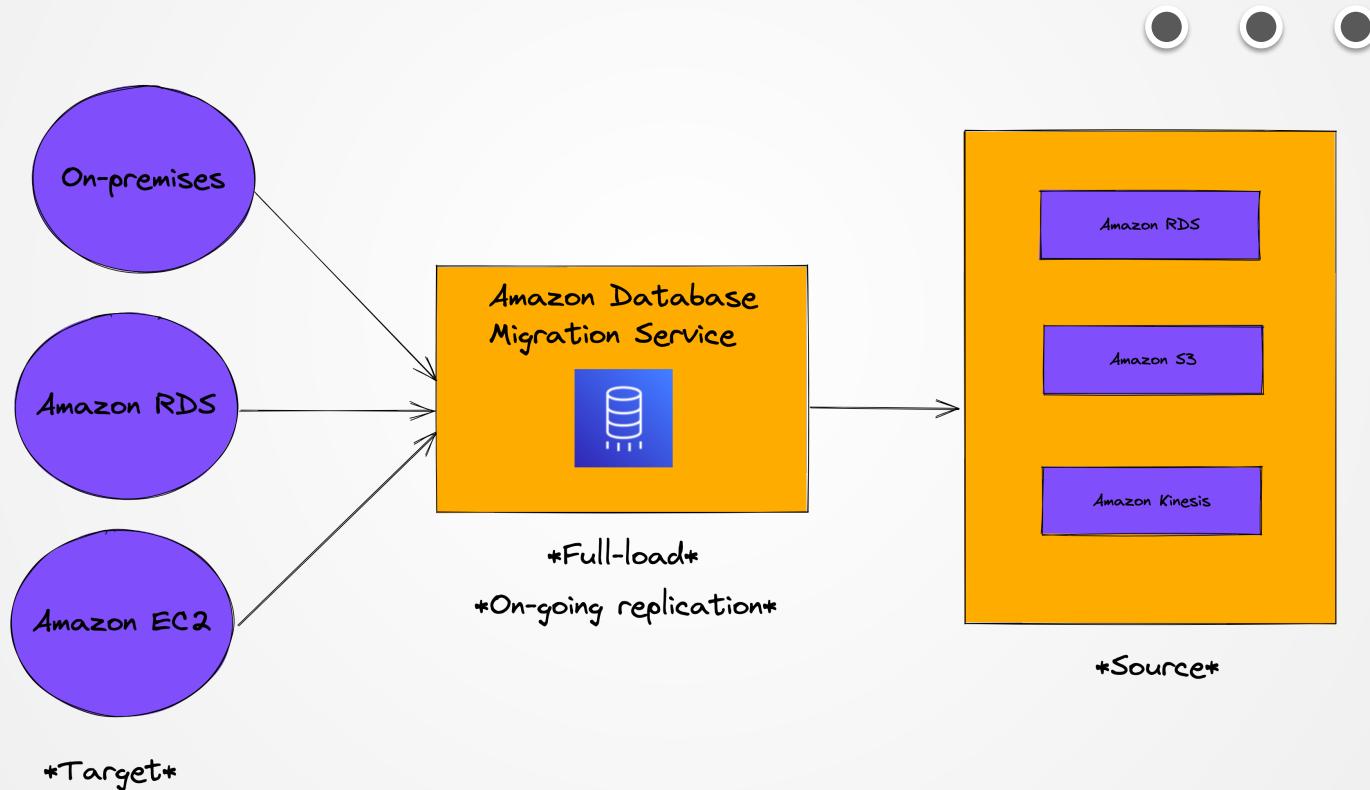
Amazon Kinesis makes it easy to collect, process, and analyze real-time, streaming data so you can get timely insights and react quickly to new information.



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.

