

# Introduction Amazon Web Services (AWS)



Dr. Villanes

#StoryTime

## General Configuration

Cluster name

☒ Logging ⓘ

S3 folder  📁

Launch mode ☒ Cluster ⓘ ☐ Step execution ⓘ

## Software configuration

Release  ⓘ

- Applications
- ☒ Core Hadoop: Hadoop 2.10.0, Hive 2.3.7, Hue 4.7.1, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2
  - ☐ HBase: HBase 1.4.13, Hadoop 2.10.0, Hive 2.3.7, Hue 4.7.1, Phoenix 4.14.3, and ZooKeeper 3.4.14
  - ☐ Presto: Presto 0.238.3 with Hadoop 2.10.0 HDFS and Hive 2.3.7 Metastore
  - ☐ Spark: Spark 2.4.6 on Hadoop 2.10.0 YARN and Zeppelin 0.8.2

☐ Use AWS Glue Data Catalog for table metadata ⓘ

## Hardware configuration

Instance type

The selected instance type adds 64 GiB of GP2 EBS storage per instance by default. [Learn more](#) 📄

Number of instances  (1 master and 2 core nodes)

## General Configuration

Cluster name

☒ Logging ⓘ

S3 folder  📁

Launch mode ☒ Cluster ⓘ ☐ Step execution ⓘ

## Software configuration

Release  ⓘ

- Applications
- ☒ Core Hadoop: Hadoop 2.10.0, Hive 2.3.7, Hue 4.7.1, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2
  - ☐ HBase: HBase 1.4.13, Hadoop 2.10.0, Hive 2.3.7, Hue 4.7.1, Phoenix 4.14.3, and ZooKeeper 3.4.14
  - ☐ Presto: Presto 0.238.3 with Hadoop 2.10.0 HDFS and Hive 2.3.7 Metastore
  - ☐ Spark: Spark 2.4.6 on Hadoop 2.10.0 YARN and Zeppelin 0.8.2

☐ Use AWS Glue Data Catalog for table metadata ⓘ

## Hardware configuration

Instance type

The selected instance type adds 64 GiB of GP2 EBS storage per instance by default. [Learn more](#) 🔗

Number of instances  (1 master and 999 core nodes)

Create cluster

# Class roadmap – Big Data

- Fall semester (two classes)
  - Why do we need Cloud Computing?
  - Introduction to AWS
  - Labs: AWS S3, AWS EC2, AWS RDS
  - Objective: give you the tools you need to get the AWS Cloud Practitioner Certification
- Spring semester (four classes)
  - Hadoop Hive & Spark
  - AWS EMR
    - AWS EMR is not a big part of the certification
- Advanced Big Data (six classes) → Dan Zaratsian (MSA alum - working at Google)

# Topics covered today

- What is Cloud Computing?
- Cloud Computing Vendors
- AWS
- AWS Cloud Practitioner Certification
- AWS S3
- AWS EC2
- Lab #1 AWS S3
- Lab #2 AWS EC2

# 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**.

Cloud computing gives you **access to servers, storage, databases, and a broad set of application services over the Internet**. A cloud services provider such as Amazon Web Services, **owns and maintains the network-connected hardware required for these application services, while you provision and use what you need via a web application**.

# Advantages of Cloud Computing



Trade capital expense  
for variable expense.



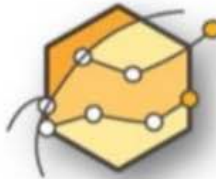
Increase speed and  
agility.



Benefit from massive  
economies of scale.



Stop spending money on  
running and maintaining data  
centers.



Stop guessing  
capacity. (Elasticity)



Go global in minutes.



# Gartner's Infrastructure as a Service (IaaS) Magic Quadrant

Top players:

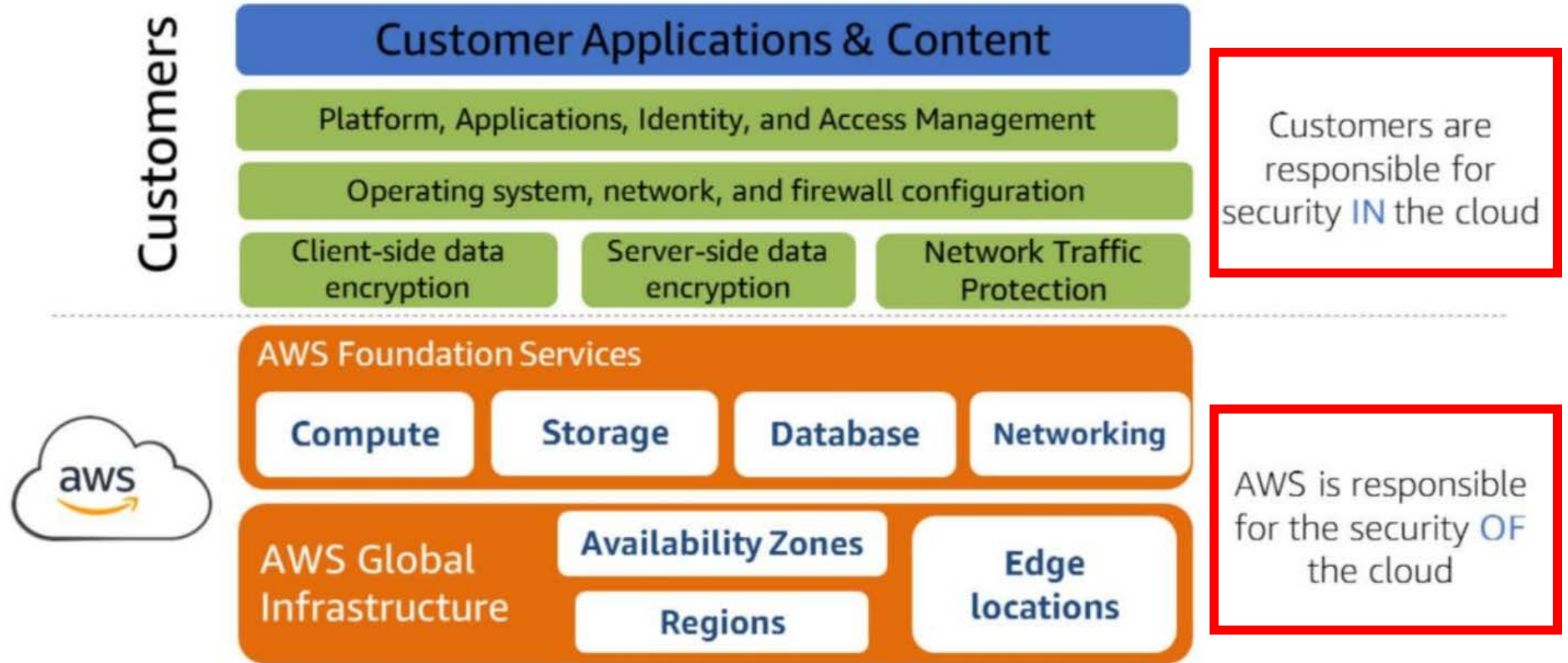


# What is Amazon Web Services (AWS)?

AWS is a collection of remote computing **services** called **web services**. These web services make up a cloud computing platform offered via the internet. AWS delivers web-based cloud **services for storage, computing, networking, databases, and more.**



# AWS Shared Responsibility Model



# AWS Global Infrastructure

- **Regions:**

- Geographic locations
- Consist of at least two Availability Zones (AZ)

- **Availability Zones:**

- Clusters of data centers
- Isolated from failures in other Availability Zones
- Availability Zones in a region are connected through low-latency links

- **Edge Locations:**

- 200+ Edge locations
- Local points of presence that support AWS services like AWS Route 53, AWS Cloud Front, AWS Shield

# AWS Global Infrastructure Map



## 24 Launched Regions

Each with multiple Availability Zones (AZ's)

## 77 Availability Zones

## 220+ Points of Presence

210+ Edge Locations and 12 Regional Edge Caches

**Interactive map:** <https://aws.amazon.com/about-aws/global-infrastructure/>

# AWS Services and Features



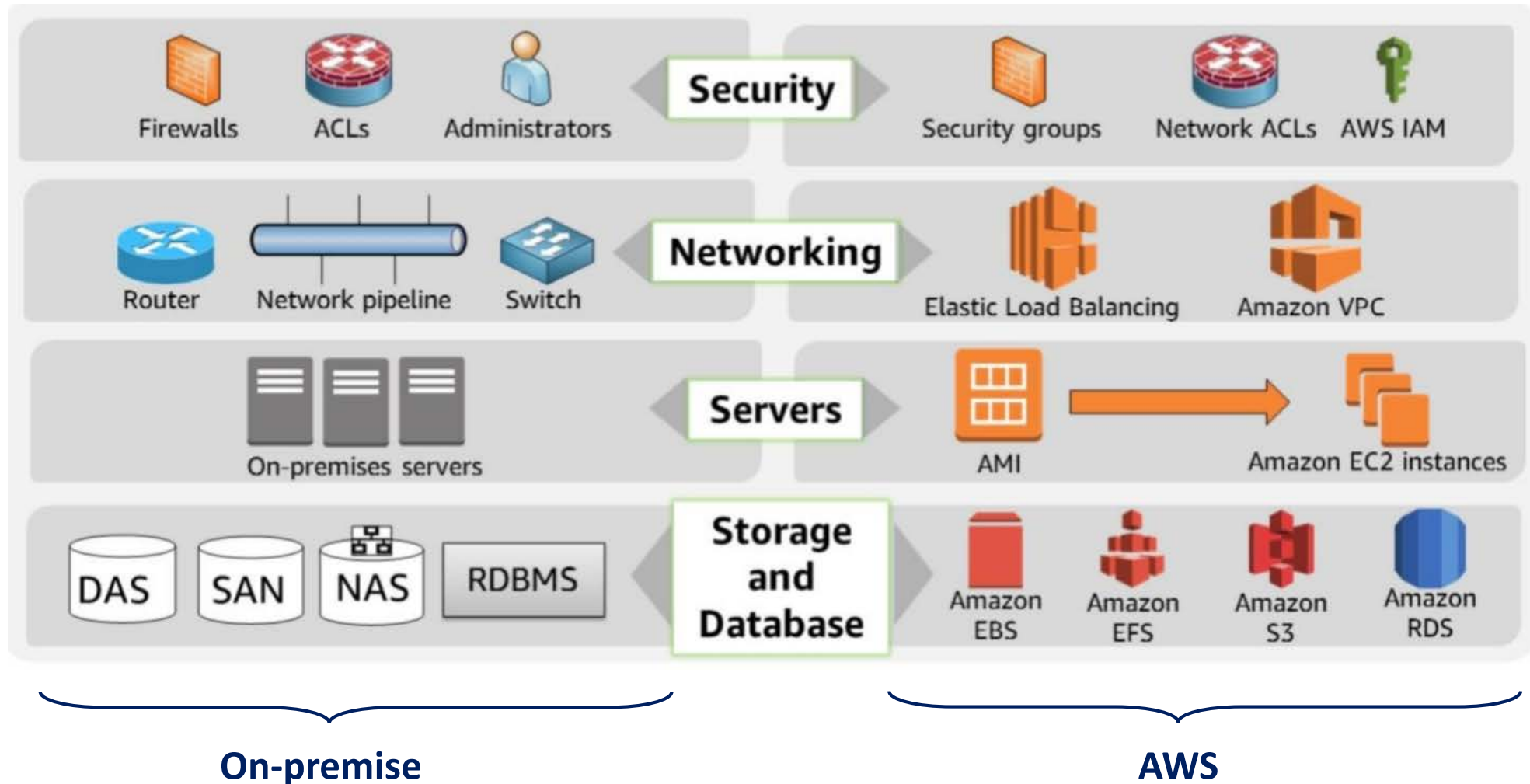


# Top 25 AWS Services




1. Amazon EC2 (Elastic Compute Cloud)
2. Amazon RDS (Relational Database Services)
3. Amazon S3 (Simple Storage Service)
4. Amazon Lambda
5. Amazon CloudFront
6. Amazon Glacier
7. Amazon SNS (Simple Notification Service)
8. Amazon EBS (Elastic Block Store)
9. Amazon VPC (Virtual Private Cloud)
10. Amazon Kinesis
11. Amazon Auto-scaling
12. Amazon IAM (Identity and Access Management)
13. Amazon SQS (Simple Queue Service)
14. Amazon Elastic Beanstalk
15. Dynamo DB
16. Amazon ElastiCache
- Business Continuity Plan
17. Amazon Redshift
18. Amazon Sagemaker
19. Amazon Lightsail
20. Amazon EFS (Elastic File System)
21. Amazon Cloudwatch
22. Amazon Chime
23. Amazon Cloud Directory
24. Amazon Cognito
25. Amazon Inspector

# AWS Core Infrastructure and Services









































# AWS Foundational Services

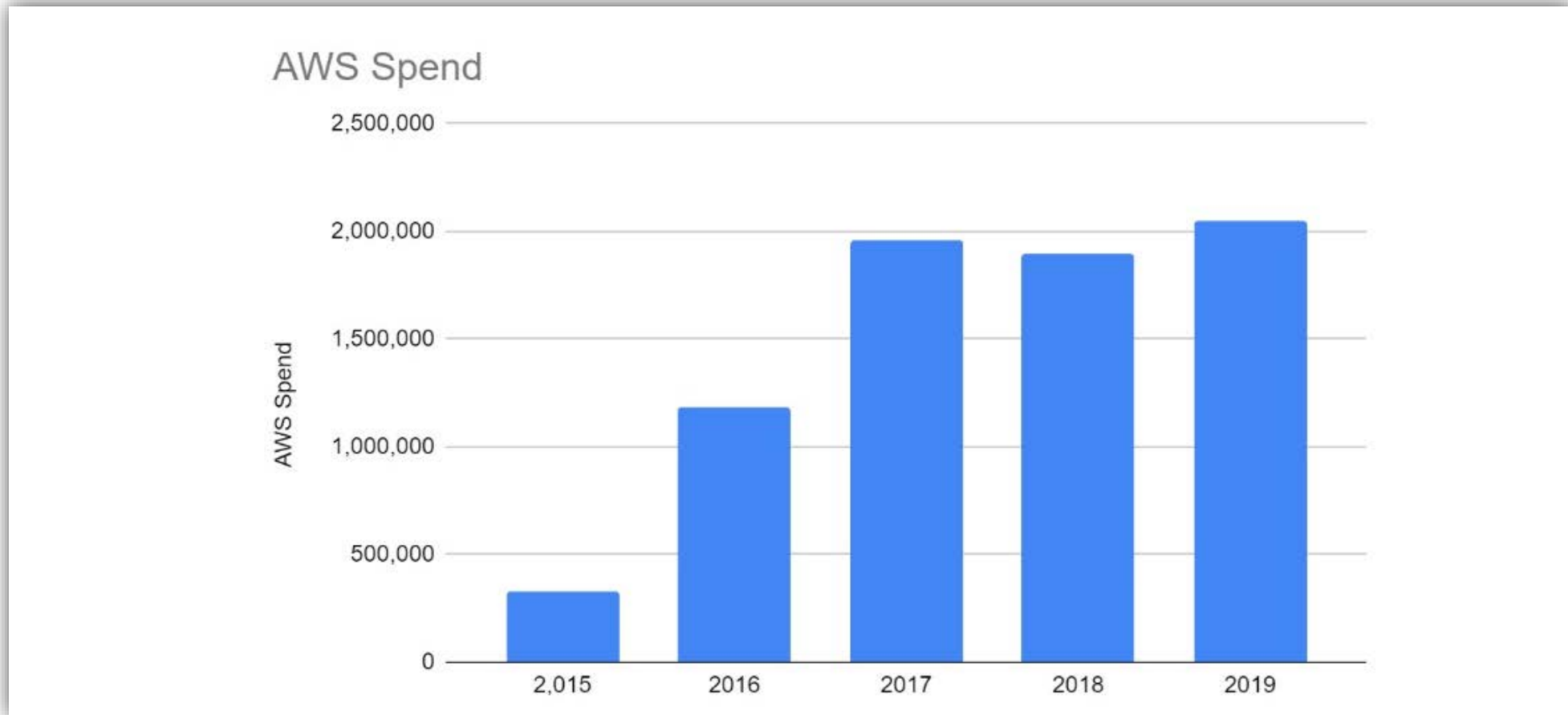
Compute	Network	Storage	Security & Identity	Applications
 Amazon EC2	 Amazon CloudFront	 Amazon EFS	 Amazon Inspector	 Amazon WorkDocs
 Amazon Elastic Container Registry	 Amazon Route 53	 Amazon S3 Glacier	 AWS Artifact	 Amazon WorkMail
 Amazon Elastic Container Service	 Amazon VPC	 Amazon S3	 AWS Certificate Manager	 Amazon AppStream 2.0
 Amazon Lightsail	 AWS Direct Connect	 AWS Snowball	 AWS CloudHSM	 Amazon WorkSpaces
 AWS Batch	 Elastic Load Balancing	 AWS Storage Gateway	 AWS Directory Service	
 AWS Elastic Beanstalk			 AWS IAM	
 AWS Lambda			 AWS KMS	
			 AWS Organizations	
			 AWS Shield	
			 AWS WAF	

# AWS Platform Services

Databases	Analytics	Application Services	Management Tools	Developer Tools	Mobile Services	Internet of Things
 Amazon DynamoDB	 Amazon Athena	 Amazon API Gateway	 Amazon CloudWatch	 AWS CodeBuild	 AWS Mobile Hub	 AWS IoT
 Amazon ElastiCache	 Amazon CloudSearch	 Amazon AppStream 2.0	 AWS CloudFormation	 AWS CodeCommit	 Amazon Cognito	 AWS IoT Greengrass
 Amazon RDS	 Amazon EMR	 Amazon Elastic Transcoder	 AWS CloudTrail	 AWS CodeDeploy	 Amazon Mobile Analytics	
 Amazon Redshift	 Amazon Elasticsearch Service	 Amazon SWF	 AWS Config	 AWS CodePipeline	 Amazon Pinpoint	
	 Amazon Kinesis	 AWS Step Functions	 AWS Managed Services	 AWS X-Ray	 AWS Device Farm	
	 Amazon QuickSight		 AWS OpsWorks	 AWS CodeStar		
			 AWS Service Catalog			
			 AWS Trusted Advisor			

# Are companies using AWS? **Yes**

Annual AWS Spend for a publicly traded e-commerce company



Let's talk about the AWS Certification...

# AWS Cloud Practitioner Certification

- Define **what the AWS Cloud is and the basic global infrastructure**
- Describe basic **AWS Cloud architectural principles**
- Describe the **AWS Cloud value proposition**
- Describe **key services on the AWS platform and their common use cases** (for example, compute and analytics)
- Describe **basic security and compliance aspects of the AWS platform and the shared security model**
- Define the **billing, account management, and pricing models**
- Identify **sources of documentation or technical assistance** (for example, whitepapers or support tickets)
- Describe **basic/core characteristics of deploying and operating in the AWS Cloud**

## Exam Details



### Format

Multiple choice, multiple answer



### Type

Foundational



### Delivery Method

Testing center or online proctored exam



### Time

90 minutes to complete the exam



### Cost

100 USD (Practice Exam: 20 USD)


































### Language

Available in English, Indonesian (Bahasa), Japanese, Korean, and Simplified Chinese

<https://aws.amazon.com/certification/certified-cloud-practitioner/>

# AWS Foundational Services

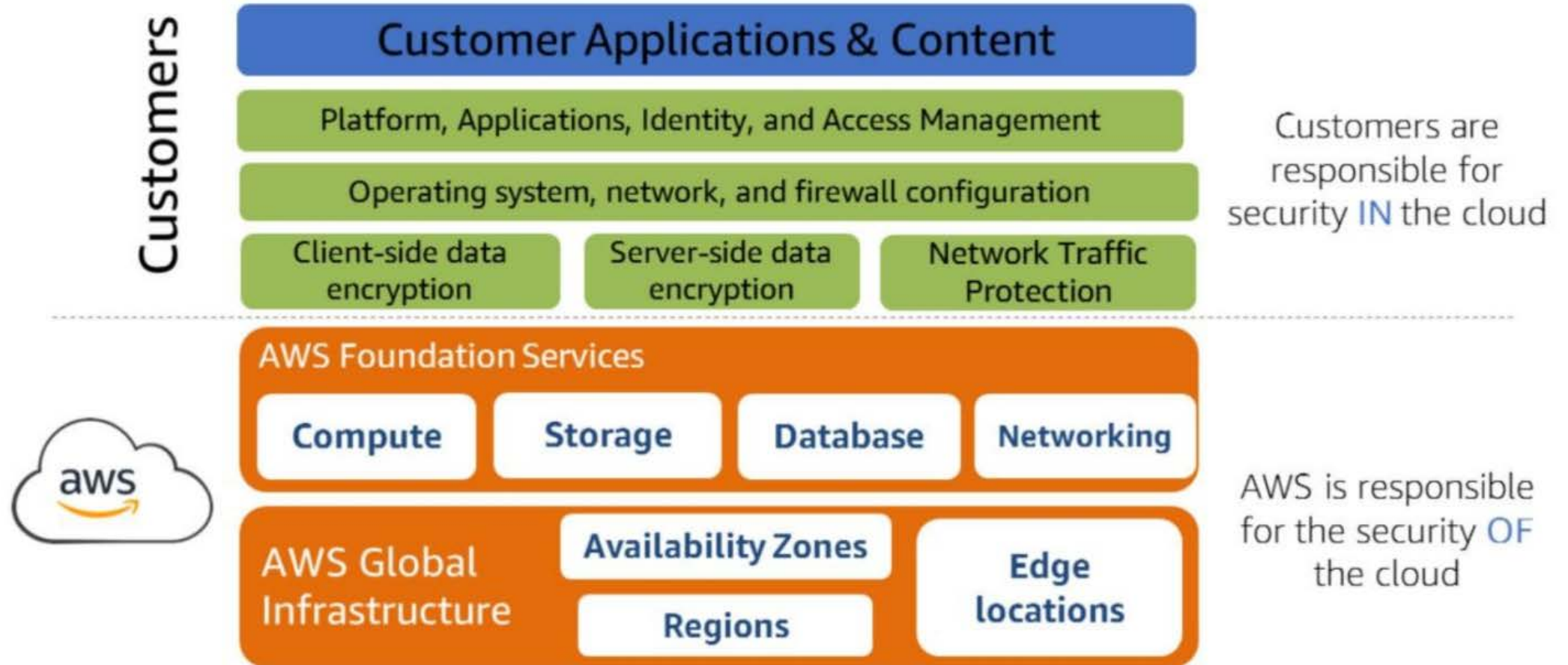
Compute	Network	Storage	Security & Identity	Applications
 Amazon EC2	 Amazon CloudFront	 Amazon EFS	 Amazon Inspector	 Amazon WorkDocs
 Amazon Elastic Container Registry	 Amazon Route 53	 Amazon S3 Glacier	 AWS Artifact	 Amazon WorkMail
 Amazon Elastic Container Service	 Amazon VPC	 Amazon S3	 AWS Certificate Manager	 Amazon AppStream 2.0
 Amazon Lightsail	 AWS Direct Connect	 AWS Snowball	 AWS CloudHSM	 Amazon WorkSpaces
 AWS Batch	 Elastic Load Balancing	 AWS Storage Gateway	 AWS Directory Service	
 AWS Elastic Beanstalk			 AWS IAM	
 AWS Lambda			 AWS KMS	
			 AWS Organizations	
			 AWS Shield	
			 AWS WAF	

# AWS Platform Services

Databases	Analytics	Application Services	Management Tools	Developer Tools	Mobile Services	Internet of Things
 Amazon DynamoDB	 Amazon Athena	 Amazon API Gateway	 Amazon CloudWatch	 AWS CodeBuild	 AWS Mobile Hub	 AWS IoT
 Amazon ElastiCache	 Amazon CloudSearch	 Amazon AppStream 2.0	 AWS CloudFormation	 AWS CodeCommit	 Amazon Cognito	 AWS IoT Greengrass
 Amazon RDS	 Amazon EMR	 Amazon Elastic Transcoder	 AWS CloudTrail	 AWS CodeDeploy	 Amazon Mobile Analytics	
 Amazon Redshift	 Amazon Elasticsearch Service	 Amazon SWF	 AWS Config	 AWS CodePipeline	 Amazon Pinpoint	
	 Amazon Kinesis	 AWS Step Functions	 AWS Managed Services	 AWS X-Ray	 AWS Device Farm	
	 Amazon QuickSight		 AWS OpsWorks	 AWS CodeStar		
			 AWS Service Catalog			
			 AWS Trusted Advisor			



# AWS Shared Responsibility Model





	Developer	Business	Enterprise
	<i>Recommended if you are experimenting or testing in AWS.</i>	<i>Recommended if you have production workloads in AWS.</i>	<i>Recommended if you have business and/or mission critical workloads in AWS.</i>
<b>AWS Trusted Advisor Best Practice Checks</b>	7 Core <a href="#">checks</a>	Full set of <a href="#">checks</a>	Full set of <a href="#">checks</a>
<b>Enhanced Technical Support</b>	Business hours** email access to Support Engineers  Unlimited cases / 1 primary contact	24x7 phone, email, and chat access to Support Engineers  Unlimited cases / unlimited contacts (IAM supported)	24x7 phone, email, and chat access to Support Engineers  Unlimited cases / unlimited contacts (IAM supported)
<b>Case Severity / Response Times*</b>	General guidance: < 24 business hours  System impaired: < 12 business hours	General guidance: < 24 hours  System impaired: < 12 hours  Production system impaired: < 4 hours  Production system down: < 1 hour	General guidance: < 24 hours  System impaired: < 12 hours  Production system impaired: < 4 hours  Production system down: < 1 hour  Business-critical system down: < 15 minutes
<b>Architectural Guidance</b>	General	Contextual to your use-cases	Consultative review and guidance based on your applications
<b>Programmatic Case Management</b>		AWS Support API	AWS Support API
<b>Third-Party Software Support</b>		Interoperability & configuration guidance and troubleshooting	Interoperability & configuration guidance and troubleshooting
<b>Proactive Programs</b>		Access to Infrastructure Event Management for additional fee.	Infrastructure Event Management Well-Architected Reviews Operations Reviews Technical Account Manager (TAM) coordinates access to programs and other AWS experts as needed.
<b>Technical Account Management</b>			Designated Technical Account Manager (TAM) to proactively monitor your environment and assist with optimization.
<b>Training</b>			Access to online self-paced labs
<b>Account Assistance</b>			Concierge Support Team



### Simple Monthly Calculator

Whether you are running a single instance or dozens of individual services, you can estimate your monthly bill using your monthly bill using AWS Simple Monthly Calculator. The calculator allows you to estimate individual or multiple prices and use templates to appraise complete solutions.

[AWS Simple Monthly Calculator](#)



### TCO Calculator

The AWS TCO calculator gives you the option to evaluate the savings from using AWS and comparing against on premises and co-location environments. The TCO calculator matches your current infrastructure to the most cost effective AWS offering. This tool takes into consideration all the costs to run a solution, including physical facilities, power and cooling, providing a realistic end-to-end comparison of your costs.

[AWS TCO Calculator](#)

**For updated pricing guide:** <https://aws.amazon.com/premiumsupport/plans/>

I will walk you through on how to prepare for the AWS Cloud Practitioner Certification on Thursday...

Let's go back to the Foundation Services:  
**AWS S3 and AWS EC2**

# AWS Foundational Services

Compute	Network	Storage	Security & Identity	Applications
 Amazon EC2	 Amazon CloudFront	 Amazon EFS	 Amazon Inspector	 Amazon WorkDocs
 Amazon Elastic Container Registry	 Amazon Route 53	 Amazon S3 Glacier	 AWS Artifact	 Amazon WorkMail
 Amazon Elastic Container Service	 Amazon VPC	 Amazon S3	 AWS Certificate Manager	 Amazon AppStream 2.0
 Amazon Lightsail	 AWS Direct Connect	 AWS Snowball	 AWS CloudHSM	 Amazon WorkSpaces
 AWS Batch	 Elastic Load Balancing	 AWS Storage Gateway	 AWS Directory Service	
 AWS Elastic Beanstalk			 AWS IAM	
 AWS Lambda			 AWS KMS	
			 AWS Organizations	
			 AWS Shield	
			 AWS WAF	

# Amazon Simple Storage Service (S3)



Amazon S3

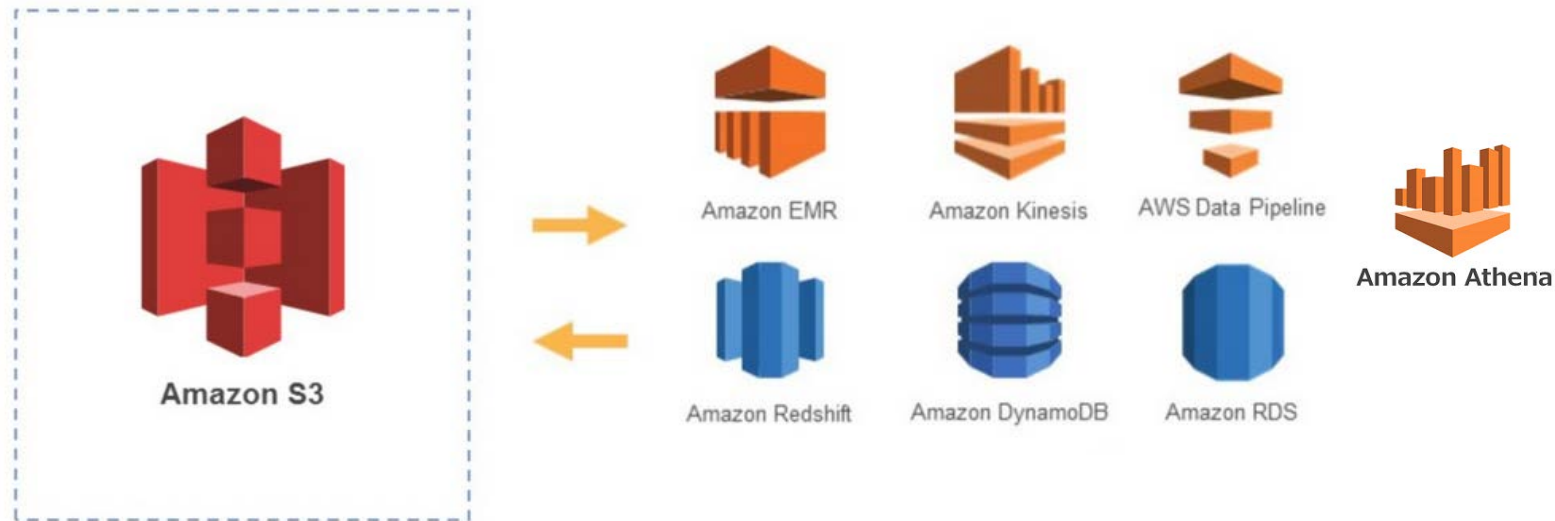
- S3 is an online storage solution from Amazon
- Objects are stored in **buckets**.
- Natively online, **HTTP access**.
- Every object in Amazon S3 can be **uniquely addressed through the combination of the web service endpoint, bucket name, key, and optionally, a version**.
- **Store and retrieve any amount of data**, any time, from anywhere on the web.
- Amazon S3 is **highly scalable, reliable, low cost, and designed for durability**.
- Data **can be stored as-is**: there is no need to convert it to a predefined schema.

# Amazon Simple Storage Service (S3) - Facts



- You can have **up to 100 buckets** in each account
- Can store an **unlimited** number of objects in a bucket
- Objects can be up to **5 TB**; no bucket size limit
- Designed for **99.999999999% (11 9's) of durability**, and **99.99% availability** in a given year
- You can **control access** to the bucket and its objects
- **Pay** only for **what you use / No fee**
- Data **can be stored as-is**: there is no need to convert it to a predefined schema.
- Allows organizations to **use S3 as their data lake for analytics**.

# AWS Recommendation for Big Data: Aggregate ALL data in an AWS S3



Amazon S3: durable, available, high performance, easy to use, scalable, integrated  
Direct integration with EMR: read/write from/to S3  
AWS S3: Scalability on demand

Demo AWS S3



## Amazon S3



📘 Unsure of your S3 access patterns? Automate storage cost savings with S3 Intelligent-Tiering.

[Learn more](#)

## Buckets

[Access points](#)[Batch Operations](#)[Access analyzer for S3](#)[Account settings for Block Public Access](#)[Feature spotlight](#) 3

## Amazon S3

## Buckets (32)

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

[Copy ARN](#)[Empty](#)[Delete](#)[Create bucket](#)

&lt; 1 &gt;



	Name ▲	Region ▼	Access ▼	Creation date ▼
<input type="radio"/>	<a href="#">airlines-ratings</a>	US West (Oregon) us-west-2	<a href="#">Bucket and objects not public</a>	November 21, 2019, 16:27 (UTC-05:00)
<input type="radio"/>	<a href="#">airlines-ratings2</a>	US East (Ohio) us-east-2	<a href="#">Bucket and objects not public</a>	November 22, 2019, 15:07 (UTC-05:00)
<input type="radio"/>	<a href="#">aws-emr-resources-799535928231-us-east-2</a>	US East (Ohio) us-east-2	<a href="#">Objects can be public</a>	January 27, 2020, 09:05 (UTC-05:00)
<input type="radio"/>	<a href="#">aws-emr-resources-799535928231-us-west-1</a>	US West (N. California) us-west-1	<a href="#">Objects can be public</a>	April 22, 2020, 13:50 (UTC-04:00)
<input type="radio"/>	<a href="#">aws-emr-resources-799535928231-us-west-2</a>	US West (Oregon) us-west-2	<a href="#">Objects can be public</a>	March 8, 2019, 11:26 (UTC-05:00)
<input type="radio"/>	<a href="#">aws-logs-799535928231-us-east-1</a>	US East (N. Virginia) us-east-1	<a href="#">Objects can be public</a>	February 10, 2018, 15:11 (UTC-05:00)
<input type="radio"/>	<a href="#">aws-logs-799535928231-us-east-2</a>	US East (Ohio) us-east-2	<a href="#">Objects can be public</a>	January 29, 2018, 14:17 (UTC-05:00)
<input type="radio"/>	<a href="#">aws-logs-799535928231-us-west-1</a>	US West (N. California) us-west-1	<a href="#">Objects can be public</a>	March 14, 2018, 11:00 (UTC-04:00)
<input type="radio"/>	<a href="#">aws-logs-799535928231-us-west-2</a>	US West (Oregon) us-west-2	<a href="#">Objects can be public</a>	March 7, 2018, 11:28 (UTC-05:00)
<input type="radio"/>	<a href="#">flu-sentiment</a>	US West (Oregon) us-west-2	<a href="#">Bucket and objects not public</a>	April 4, 2019, 15:14 (UTC-04:00)
<input type="radio"/>	<a href="#">hivedata432</a>	US East (N. Virginia) us-east-1	<a href="#">Objects can be public</a>	January 9, 2020, 13:42 (UTC-05:00)





## Amazon S3



## Buckets

Access points

Batch Operations

Access analyzer for S3

Account settings for Block Public Access

Feature spotlight **3**

Amazon S3 &gt; airlines-ratings &gt; data/

## data/

## Folder overview

Region

US West (Oregon) us-west-2

S3 URI

s3://airlines-ratings/data/

Amazon resource name (ARN)

arn:aws:s3:::airlines-ratings/data/

Drag and drop files and folders you want to upload here, or choose **Upload**.

## Objects (1)

Objects are the fundamental entities stored in Amazon S3. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#) 

Delete

Actions ▼

Create folder

Upload

Find objects by prefix

&lt; 1 &gt;

<input type="checkbox"/>	Name ▲	Type ▼	Last modified ▼	Size ▼	Storage class ▼
<input type="checkbox"/>	airlines_ratings.csv	csv	November 21, 2019, 16:40 (UTC-05:00)	51.2 KB	Standard

# Amazon Elastic Compute Cloud (EC2)



- An Amazon EC2 instance is a **virtualized computer** that runs in an Amazon's data center
- Allows users to **rent** virtual computers on which they run computer applications
- Virtual computing environments, known as instances
- Central part of AWS: provides **scalable** deployment
- ***Elastic***: a user can create, launch, and terminate instances as needed, paying by the second for active servers.
- Various configurations of CPU, memory, storage, and networking capacity for your instances, known as ***instance types***
- Secure login information for your instances using ***key pairs***

# Amazon Elastic Compute Cloud (EC2) - Facts



- **Pay only for capacity that you actually use**
- Choose **Linux or Windows**
- Consider the following when choosing your instances:
  - Core count
  - Memory size
  - Storage size and type
  - Network performance
  - CPU technologies
- Launch an instance from a **pre-configured Amazon Machine Image (AMI) – template for root volume for the instance (O.S., applications); launch permissions; volumes to attach.**

Demo AWS EC2





Services ▼



Andrea-AWS ▼

Oregon ▼

Support ▼

 New EC2 Experience  
Tell us what you think 

EC2 Dashboard NewEvents New

Tags

Limits

## ▼ Instances

**Instances** New

Instance Types

Launch Templates

Spot Requests

Savings Plans

Reserved Instances

Dedicated Hosts New

Scheduled Instances

Capacity Reservations

## ▼ Images

AMIs

## ▼ Elastic Block Store

Volumes

Instances (1/1) [Info](#)**Connect**





Instance state ▼

Actions ▼

**Launch instances** ▼ *Filter instances*

&lt; 1 &gt;



<input checked="" type="checkbox"/>	Name ▼	Instance ID	Instance state ▼	Instance type ▼	Status check	Alarm Status	Availability zone ▼	Public IPv4 DNS
<input checked="" type="checkbox"/>	-	i-0a2b0471c226ae4b0	 Running 	t2.micro	 2/2 checks ...	No alarms 	us-west-2a	ec2-34-222-37-86.

Instance: i-0a2b0471c226ae4b0

**Details**

Security

Networking

Storage

Status Checks

Monitoring

Tags

EC2 > Instances > i-0a2b0471c226ae4b0 > Connect to instance

## Connect to instance [Info](#)

Connect to your instance i-0a2b0471c226ae4b0 using any of these options

Session Manager

**RDP client**

You can connect to your Windows instance using a remote desktop client of your choice, and by downloading and running the RDP shortcut file below:

**Download remote desktop file**

When prompted, connect to your instance using the following details:

Public DNS

ec2-34-222-37-86.us-west-2.compute.amazonaws.com

User name

Administrator

Password [Get password](#)

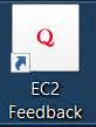
If you've joined your instance to a directory, you can use your directory credentials to connect to your instance.

Cancel

ec2-34-222-37-86 - ec2-34-222-37-86.us-west-2.compute.amazonaws.com - Remote Desktop Connection



Private IP Address: 172.31.46.250  
Instance Size: t2.micro  
Availability Zone: us-west-2a  
Architecture: AMD64  
Total Memory: 1024 MB  
Network Performance: Low to Moderate



5:43 PM  
11/1/2020



Services ▾



Andrea-AWS ▾

Oregon ▾

Support ▾

New EC2 Experience  
Tell us what you think

EC2 Dashboard New

Events New

Tags

Limits

▼ Instances

**Instances** New

Instance Types

Launch Templates

Spot Requests

Savings Plans

Reserved Instances

Dedicated Hosts New

Scheduled Instances

Capacity Reservations

▼ Images

AMIs

▼ Elastic Block Store

Volumes

## Instances (1) [Info](#)



Connect

Instance state ▾

Actions ▾

Launch instances ▾

Filter instances

< 1 >



<input type="checkbox"/>	Name ▾	Instance ID	Instance state ▾	Instance type ▾	Status check	Alarm Status	Availability zone ▾	Public IPv4 DNS
<input type="checkbox"/>	-	i-0a2b0471c226ae4b0	Termina...	t2.micro	-	No alarms +	us-west-2a	-

Select an instance above





Lab #1 – AWS S3

Lab #2 – AWS EC2