



# AI MULTI-TENANT PDF/BOOK SUMMARIZATION USING AMAZON BEDROCK- 3 TIER ARCHITECTURE

PRIYANKA RAJAGOPAL

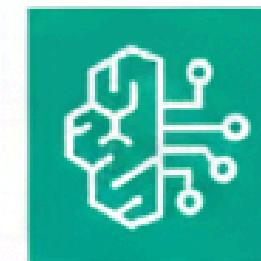
[GitHub](#)



# PROJECT OVERVIEW

- This project presents an AI multi-tenant PDF summarization system using **Amazon Bedrock**, delivering secure, scalable, and automated document intelligence in the cloud. Users upload PDF documents through a public endpoint and receive concise summaries generated by Amazon Bedrock's **Titan Text foundation model**, demonstrating the use of AI and machine learning services.
- The solution is built entirely on AWS and follows a **3-tier, high-availability, multi-AZ architecture**, with an **Application Load Balancer** as the presentation tier, EC2 instances in an **Auto Scaling Group** as the application tier, and **Amazon S3** and **Amazon RDS** (Aurora MySQL) as the data tier. Strong security is enforced using **IAM** least privilege access, **KMS** encryption for S3, private subnet isolation, and **CloudWatch** for monitoring and observability.

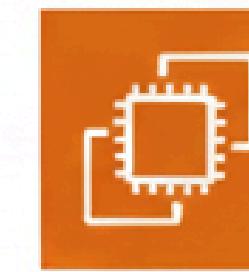
# SERVICES USED



Amazon  
Bedrock



VPC



Amazon EC2



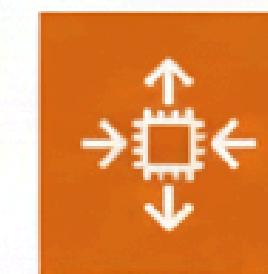
Amazon RDS



Amazon S3



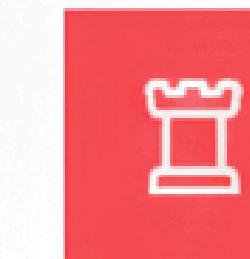
AWS Key  
Management  
Service



Auto Scaling



Application  
Load  
Balancer



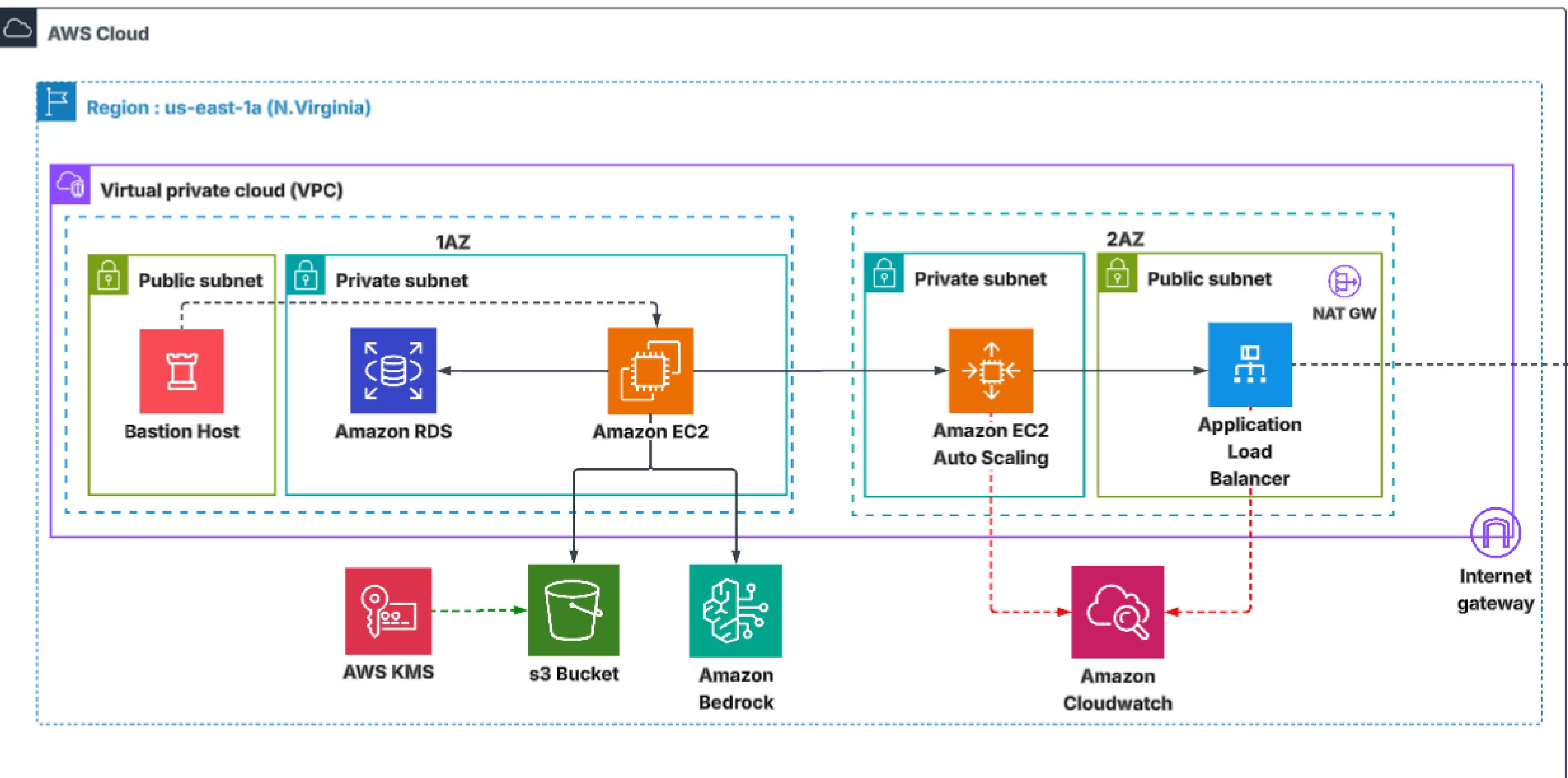
Bastion Host



Amazon  
CloudWatch



# ARCHITECTURE





Search

[Alt+S]



United States (N. Virginia) ▾

priyanka-tenant

≡ VPC &gt; Your VPCs &gt; Create VPC

i 🔍 🌐

## VPC settings

### Resources to create [Info](#)

Create only the VPC resource or the VPC and other networking resources.

 VPC only VPC and more

### Name tag - *optional*

Creates a tag with a key of 'Name' and a value that you specify.

tenant-VPC

### IPv4 CIDR block [Info](#)

 IPv4 CIDR manual input IPAM-allocated IPv4 CIDR block

### IPv4 CIDR

10.0.0.0/16

CIDR block size must be between /16 and /28.

### IPv6 CIDR block [Info](#)

 No IPv6 CIDR block IPAM-allocated IPv6 CIDR block Amazon-provided IPv6 CIDR block IPv6 CIDR owned by me

### Tenancy [Info](#)

Default



## Tags

Created VPC with CIDR 10.0.0.0/16

## VPC &gt; Subnets



## VPC dashboard &lt;

AWS Global View ↗

Filter by VPC ▼

## ▼ Virtual private cloud

Your VPCs

## Subnets

Route tables

Internet gateways

Egress-only internet gateways

Carrier gateways

DHCP option sets

Elastic IPs

Managed prefix lists

NAT gateways

Peering connections

Route servers [New](#)

## ▼ Security

Network ACLs

Security groups

## ▼ PrivateLink and Lattice

✓ You have successfully created 4 subnets: subnet-0353a800c2b8b15ea, subnet-0708d50287b369059, subnet-0112df5af4d19c73c, subnet-02a25b96005128415

Last updated  
less than a minute ago

Actions ▾

Create subnet

Subnets (4) [Info](#)

Find subnets by attribute or tag

Subnet ID : subnet-0353a800c2b8b15ea

Subnet ID : subnet-0708d50287b369059

Subnet ID : subnet-0112df5af4d19c73c

Show more (+1)

Clear filters

&lt; 1 &gt; |

| <input type="checkbox"/> | Name         | Subnet ID                                | State     | VPC  | Block Public |
|--------------------------|--------------|--|-----------|--|--------------|
| <input type="checkbox"/> | Pri Sub - 1a | <a href="#">subnet-0112df5af4d19c73c</a> | Available | <a href="#">vpc-0dea9b2b902b494b5   ten...</a> | Off          |
| <input type="checkbox"/> | Pub Sub - 1b | <a href="#">subnet-0708d50287b369059</a> | Available | <a href="#">vpc-0dea9b2b902b494b5   ten...</a> | Off          |
| <input type="checkbox"/> | Pri Sub - 1b | <a href="#">subnet-02a25b96005128415</a> | Available | <a href="#">vpc-0dea9b2b902b494b5   ten...</a> | Off          |
| <input type="checkbox"/> | Pub Sub - 1a | <a href="#">subnet-0353a800c2b8b15ea</a> | Available | <a href="#">vpc-0dea9b2b902b494b5   ten...</a> | Off          |

## Select a subnet



Created 4 Subnets in 2 Availability zones, 2 Public Subnets and 2 Private Subnets.

Search [Alt+S] United States (N. Virginia) priyanka-tenant

VPC > Internet gateways > Create internet gateway

### Create internet gateway Info

An internet gateway is a virtual router that connects a VPC to the internet. To create a new internet gateway specify the name for the gateway below.

#### Internet gateway settings

**Name tag**  
Creates a tag with a key of 'Name' and a value that you specify.

tenant-IGW

**Tags - optional**  
A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

| Key                               | Value - optional   |
|-----------------------------------|--|
| <input type="text" value="Name"/> | <input type="text" value="tenant-IGW"/> <span>X</span> <span>Remove</span> |

**Add new tag**  
You can add 49 more tags.

### Create NAT gateway Info

A highly available, managed Network Address Translation (NAT) service that instances in private subnets can use to connect to services in other VPCs, on-premises networks, or the internet.

#### NAT gateway settings

**Name - optional**  
Create a tag with a key of 'Name' and a value that you specify.

tenant-NGW  
The name can be up to 256 characters long.

**Subnet**  
Select a subnet in which to create the NAT gateway.

subnet-0353a800c2b8b15ea (Pub Sub - 1a)

**Connectivity type**  
Select a connectivity type for the NAT gateway.

Public  Private

**Elastic IP allocation ID** Info  
Assign an Elastic IP address to the NAT gateway.

eipalloc-0d6edeadd8ef740a3 X Allocate Elastic IP

Created and Attached Internet gateway ,Allocated Elastic IP and created NAT Gateway.

Search [Alt+S] United States (N. Virginia) priyanka-tenant

VPC > Route tables > Create route table

### Create route table Info

A route table specifies how packets are forwarded between the subnets within your VPC, the internet, and your VPN connection.

#### Route table settings

Name - *optional*  
Create a tag with a key of 'Name' and a value that you specify.

VPC  
The VPC to use for this route table.

#### Tags

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

| Key  | Value - <i>optional</i>   |
|--|---|
| <input type="text" value="Name"/> <input type="button" value="X"/> | <input type="text" value="tenant-pub-RT"/> <input type="button" value="X"/> <input type="button" value="Remove"/> |
| <input type="button" value="Add new tag"/>                         |   |

You can add 49 more tags.

Search [Alt+S] United States (N. Virginia) priyanka-tenant

VPC > Route tables > rtb-01955154a5c02d27d > Edit subnet associations

### Edit subnet associations

Change which subnets are associated with this route table.

#### Available subnets (2/4)

| <input type="checkbox"/>            | Name         | Subnet ID                                | IPv4 CIDR   | IPv6 CIDR | Route table ID               |
|-------------------------------------|--------------|--|-------------|-----------|------------------------------|
| <input type="checkbox"/>            | Pri Sub - 1a | <a href="#">subnet-0112df5af4d19c73c</a> | 10.0.3.0/24 | -         | Main (rtb-0f604dedd75637708) |
| <input checked="" type="checkbox"/> | Pub Sub - 1b | <a href="#">subnet-0708d50287b369059</a> | 10.0.2.0/24 | -         | Main (rtb-0f604dedd75637708) |
| <input type="checkbox"/>            | Pri Sub - 1b | <a href="#">subnet-02a25b96005128415</a> | 10.0.4.0/24 | -         | Main (rtb-0f604dedd75637708) |
| <input checked="" type="checkbox"/> | Pub Sub - 1a | <a href="#">subnet-0353a800c2b8b15ea</a> | 10.0.1.0/24 | -         | Main (rtb-0f604dedd75637708) |

#### Selected subnets

[subnet-0708d50287b369059 / Pub Sub - 1b](#)  [subnet-0353a800c2b8b15ea / Pub Sub - 1a](#)

Created 2 Route tables and associated with the Public and Private Subnets.

The screenshot shows the AWS VPC Route Tables interface. A route table has two entries:

- A route from 10.0.0.0/16 to the target "local" with the status "Active".
- A route from 0.0.0.0/0 to the target "Internet Gateway" with the status "Active". The target dropdown shows "igw-0f6efafa4ba713d70" selected.

Buttons at the bottom include "Cancel", "Preview", and "Save changes".

The screenshot shows the AWS VPC Route Tables interface. A route table has two entries:

- A route from 10.0.0.0/16 to the target "local" with the status "Active".
- A route from 0.0.0.0/0 to the target "NAT Gateway" with the status "Active". The target dropdown shows "nat-06aeb1697ccf36a4d" selected.

Buttons at the bottom include "Cancel", "Preview", and "Save changes".

Public subnets route 0.0.0.0/0 to the Internet Gateway (IGW) to allow internet access for the ALB and Bastion Host. Private subnets route 0.0.0.0/0 to the NAT Gateway to enable secure outbound access without public exposure.

**Permissions defined in this policy**

Permissions defined in this policy document specify which actions are allowed or denied. To define permissions for an IAM identity (user, user group, or role), attach a policy to it

Search

**Allow (8 of 454 services)** Show remaining 446 services

| Service          | Access level                       | Resource                       | Request condition |
|------------------|------------------------------------|--------------------------------|-------------------|
| Bedrock          | Limited: Read                      | region  string like  us-east-1 | None              |
| CloudWatch       | Limited: List, Read, Write         | All resources                  | None              |
| CloudWatch Logs  | Limited: Write                     | All resources                  | None              |
| EC2              | Full: Tagging Limited: List, Write | All resources                  | None              |
| EC2 Auto Scaling | Limited: List                      | All resources                  | None              |
| Marketplace      | Limited: List, Write               | All resources                  | None              |
| RDS              | Limited: List                      | All resources                  | None              |
| S3               | Limited: List, Read, Write         | Multiple                       | None              |

Cancel

Previous

Create policy

Created IAM roles for the services with the Least Privilege principle.

Search [Alt+S] United States (N. Virginia) priyanka-tenant

KMS > Customer managed keys > Create key

**i Introducing the new Create key experience**  
We've improved the create key experience with an enhanced policy editor. [Let us know what you think](#) or you can [use the old experience](#).

Step 1  
**Configure key**

Step 2  
Add labels

Step 3 - optional  
Define key administrative permissions

Step 4 - optional  
Define key usage permissions

Step 5 - optional  
Edit key policy

Step 6  
Review

## Configure key

**Key type** [Help me choose](#)

**Symmetric**  
A single key used for encrypting and decrypting data or generating and verifying HMAC codes

**Asymmetric**  
A public and private key pair used for encrypting and decrypting data, signing and verifying messages, or deriving shared secrets

**Key usage** [Help me choose](#)

**Encrypt and decrypt**  
Use the key only to encrypt and decrypt data.

**Generate and verify MAC**  
Use the key only to generate and verify hash-based message authentication codes (HMAC).

Search [Alt+S] United States (N. Virginia) priyanka-tenant

KMS > Customer managed keys > Create key

**i Introducing the new Create key experience**  
We've improved the create key experience with an enhanced policy editor. [Let us know what you think](#) or you can [use the old experience](#).

Step 1  
Configure key

Step 2  
**Add labels**

Step 3 - optional  
Define key administrative permissions

Step 4 - optional  
Define key usage permissions

Step 5 - optional  
Edit key policy

Step 6  
Review

## Add labels

**Alias**  
You can change the alias at any time. [Learn more](#)

Alias  
tenant-KMS

**Description** - optional  
You can change the description at any time.

Description  
Key for encrypting S3 and RDS data in tenant project

Created KMS encryption key for S3.

Search [Alt+S]

Amazon S3 > Buckets > Create bucket

AWS Region: US East (N. Virginia) us-east-1

Bucket type: General purpose (selected)

General purpose: Recommended for most use cases and access patterns. General purpose buckets are the original S3 bucket type. They allow a mix of storage classes that redundantly store objects across multiple Availability Zones.

Bucket name: tenant-s3

Bucket names must be 3 to 63 characters and unique within the global namespace. Bucket names must also begin and end with a letter or number.

Copy settings from existing bucket - optional: Only the bucket settings in the following configuration are copied.

Choose bucket

Format: s3://bucket/prefix

Object Ownership: ACLs disabled (recommended) (selected)

ACLs disabled (recommended): All objects in this bucket are owned by this account. Access to this bucket and its objects is specified using only policies.

ACLs enabled: Objects in this bucket can be owned by other AWS accounts. Access to this bucket and its objects can be specified using policies.

Search [Alt+S]

Amazon S3 > Buckets > Create bucket

Encryption type: Server-side encryption with AWS Key Management Service keys (SSE-KMS) (selected)

Secure your objects with two separate layers of encryption. For details on pricing, see DSSE-KMS pricing on the Storage tab of the Amazon S3 console.

Server-side encryption with Amazon S3 managed keys (SSE-S3)

Server-side encryption with AWS Key Management Service keys (SSE-KMS) (selected)

Dual-layer server-side encryption with AWS Key Management Service keys (DSSE-KMS)

AWS KMS key: Choose from your AWS KMS keys (selected)

Enter AWS KMS key ARN

Available AWS KMS keys: arn:aws:kms:us-east-1:058707557212:key/daf7d023-8b5c-48a4-ba55-ce67fc0575c8

Bucket Key: Using an S3 Bucket Key for SSE-KMS reduces encryption costs by lowering calls to AWS KMS. S3 Bucket Keys aren't supported for DSSE-KMS.

Disable (radio button)

Enable (selected)

Advanced settings

After creating the bucket, you can upload files and folders to the bucket, and configure additional bucket settings.

Created S3 Bucket for the Raw storage of PDF/Book when uploaded from ALB and attached the KMS encryption key SSE-KMS.

Screenshot of the AWS RDS 'Create database' wizard:

**Create database** Info

**Choose a database creation method**

- Standard create  
You set all of the configuration options, including ones for availability, security, backups, and maintenance.
- Easy create  
Use recommended best-practice configurations. Some configuration is automated.

**Engine options**

**Engine type** Info

- Aurora (MySQL Compatible) 
- Aurora (PostgreSQL Compatible) 
- MySQL 
- PostgreSQL 
- MariaDB 
- Oracle 
- Microsoft SQL Server
- IBM Db2

Aurora and RDS > Databases > Create database

**Credentials Settings**

**Master username** Info  
Type a login ID for the master user of your DB instance.  
  
1 to 32 alphanumeric characters. The first character must be a letter.

**Credentials management**  
You can use AWS Secrets Manager or manage your master user credentials.

- Managed in AWS Secrets Manager - *most secure*  
RDS generates a password for you and manages it throughout its lifecycle using AWS Secrets Manager.
- Self managed  
Create your own password or have RDS create a password

**Master password** Info  
  
Password strength: **Very strong**   
Minimum constraints: At least 8 printable ASCII characters. Can't contain any of the following symbols: / " @

**Confirm master password** Info

**Cluster storage configuration** Info  
Choose the storage configuration for the Aurora DB cluster that best fits your application's price predictability and price performance needs.

**Configuration options**  
Database instance, storage, and I/O charges vary depending on the configuration. [Learn more](#) 

Created RDS database instance and placed database inside Private Subnets .It is secured with Strong credentials.

## sg-0f2b7953b093929e3 - EC2 SG

Actions ▾

### Details

|                     |              |                     |                      |                      |                        |        |                       |
|---------------------|--------------|---------------------|----------------------|----------------------|------------------------|--------|-----------------------|
| Security group name | EC2 SG       | Security group ID   | sg-0f2b7953b093929e3 | Description          | Security group for EC2 | VPC ID | vpc-0dea9b2b902b494b5 |
| Owner               | 058707557212 | Inbound rules count | 2 Permission entries | Outbound rules count | 2 Permission entries   |        |                       |



EC2 > Security Groups > sg-0f2b7953b093929e3 - EC2 SG > Edit inbound rules

### Edit inbound rules Info

Inbound rules control the incoming traffic that's allowed to reach the instance.

#### Inbound rules Info

| Security group rule ID | Type <small>Info</small> | Protocol <small>Info</small> | Port range <small>Info</small> | Source <small>Info</small> | Description - optional <small>Info</small> |        |
|------------------------|--------------------------|------------------------------|--------------------------------|----------------------------|--|--------|
| sgr-03a7024d0836af5fe  | HTTP                     | TCP                          | 80                             | Cust... ▾                  | sg-0430e64c4014b23e2                       | Delete |
| sgr-08c4f274c36bf7af8  | SSH                      | TCP                          | 22                             | Cust... ▾                  | sg-01b22db15348c9335                       | Delete |



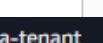
EC2 > Security Groups > sg-0f2b7953b093929e3 - EC2 SG > Edit outbound rules

### Edit outbound rules Info

Outbound rules control the outgoing traffic that's allowed to leave the instance.

#### Outbound rules Info

| Security group rule ID | Type <small>Info</small> | Protocol <small>Info</small> | Port range <small>Info</small> | Destination <small>Info</small> | Description - optional <small>Info</small> |        |
|------------------------|--------------------------|------------------------------|--------------------------------|---------------------------------|--|--------|
| sgr-0444e44497ab83015  | HTTPS                    | TCP                          | 443                            | Cust... ▾                       | 0.0.0.0/0                                  | Delete |
| sgr-0287ad3185ba1e7cd  | MySQL/Aurora             | TCP                          | 3306                           | Cust... ▾                       | sg-0a4115a558c7324cc                       | Delete |



Configured Security groups for Ec2,Basion Host,ALB and RDS.

**Name and tags**

Name: tenant-bastion-host

**Application and OS Images (Amazon Machine Image)**

An AMI contains the operating system, application server, and applications for your instance. If you don't see a suitable AMI below, use the search field or choose [Browse more AMIs](#).

**Software Image (AMI)**

Amazon Linux 2023 AMI 2023.03-1 (ami-0341d95f75f311023)

**Virtual server type**

t2.micro

**Firewall (security group)**

New security group

**Storage (volumes)**

1 volume(s) - 8 GiB

**Free tier:** In your first month, you can launch up to 10 t2.micro instances per month (or t3.micro where t3 is available).

**Common security groups**

Select security groups

**Number of instances**

1

**Software Image (AMI)**

Amazon Linux 2023 AMI 2023.03-1 (ami-0341d95f75f311023)

**Virtual server type (instance type)**

t2.micro

**Firewall (security group)**

EC2 SG

**Storage (volumes)**

1 volume(s) - 8 GiB

**Free tier:** In your first month, you can launch up to 10 t2.micro instances per month (or t3.micro where t3 is available).

**Common security groups**

Select security groups

**Compare security group rules**

Security groups that you add or remove here will be added to or removed from all your network interfaces.

Launched Bastion Host with the instance type T2 micro, where Bastion Host is kept in Public Subnet and attached the security group created.

The screenshot shows the AWS EC2 'Launch an instance' wizard at the 'Network settings' step. The left sidebar lists 'Name and tags', 'Application and OS Images (Amazon Machine Image)', and 'Quick Start' options. The main area displays network configuration details:

- VPC - required**: tenant-VPC (vpc-0dea9b2b902b494b5)
- Subnet**: subnet-0112df5af4d19c73c (Pri Sub - 1a)
- Auto-assign public IP**: Disable
- Firewall (security groups)**: Select existing security group (EC2 SG sg-0f2b7953b093929e3)

A summary panel on the right shows 1 instance and provides links for 'Create new subnet' and 'Compare security group rules'.

Another instance is launched with the instance type T2 micro, which is placed in Private Subnet and attached the respective security group.

The screenshot shows two overlapping AWS EC2 interface windows. The left window is titled 'Create launch template' and contains fields for 'Launch template name and description' (name: 'tenant-launch-template'), 'Template version description' (description: 'Tenant web server'), and 'Auto Scaling guidance' (checkbox selected for 'Provide guidance to help me set up a template that I can use with EC2 Auto Scaling'). The right window is titled 'Search results' and shows the 'Instance type' section, listing the 't2.micro' instance type with its details: Family: t2, 1 vCPU, 1 GiB Memory, Current generation: true. It also includes a note about additional costs for pre-installed software and a 'Key pair (login)' section where 'tenant-linux-key' is selected.

EC2 > Launch templates > Create launch template

## Create launch template

Creating a launch template allows you to create a saved instance configuration that can be reused, shared and launched at a later time. Templates can have multiple versions.

### Launch template name and description

Launch template name - required

tenant-launch-template

Must be unique to this account. Max 128 chars. No spaces or special characters like '&', '\*', '@'.

### Template version description

Tenant web server

Max 255 chars

#### Auto Scaling guidance

Select this if you intend to use this template with EC2 Auto Scaling

Provide guidance to help me set up a template that I can use with EC2 Auto Scaling

### Template tags

No template tags are currently applied to this template. Add a template tag to apply it to the launch template.

Add new tag

You can add up to 50 more tags.

### Search results

#### Instance type

t2.micro

Family: t2 1 vCPU 1 GiB Memory Current generation: true

On-Demand Windows base pricing: 0.0162 USD per Hour

On-Demand Ubuntu Pro base pricing: 0.0134 USD per Hour

On-Demand SUSE base pricing: 0.0116 USD per Hour

On-Demand RHEL base pricing: 0.026 USD per Hour

On-Demand Linux base pricing: 0.0116 USD per Hour

All generations

Compare instance type

Additional costs apply for AMIs with pre-installed software

#### Key pair (login)

You can use a key pair to securely connect to your instance. Ensure that you have access to the selected key pair before you launch the instance.

Key pair name

tenant-linux-key

Create new key pair

Created Launch template with Amazon Linux with t2 micro with the key pair and attaching the respective security group.

EC2 > Target groups > Create target group

0 selected

**Ports for the selected instances**  
Ports for routing traffic to the selected instances.

80

1-65535 (separate multiple ports with commas)

Include as pending below

1 selection is now pending below. Include more or register targets when ready.

### Review targets

Targets (1)

Filter targets

Show only pending

< 1 > |

| Instance ID                         | Name       | Port | State   | Security groups   | Zone       | Private IPv4 address |
|-------------------------------------|------------|------|---------|-------------------|------------|----------------------|
| <a href="#">i-0bb7e45fd7238748a</a> | tenant-EC2 | 80   | Running | ec2-rds-1, EC2 SG | us-east-1a | 10.0.3.59            |

Remove all pending

1 pending

Cancel Previous Create target group

The screenshot shows the AWS Lambda console interface. At the top, there's a search bar and a navigation menu. Below it, a large orange button says 'Create Function'. To its right, there's a dropdown for 'Region' set to 'us-east-1' and a 'Create New' button. The main area has a heading 'HelloWorld (1 version)' and a table with columns: 'Function name', 'Last modified', 'Status', 'Code size', 'Memory size', 'Runtime', and 'Last triggered'. A single row is shown for the function 'HelloWorld'. At the bottom, there are tabs for 'Overview', 'Configuration', 'Logs', and 'Actions'.

A Target Group is created with the port 80 ,target groups acts as the bridge between the Load Balancer and the EC2 instances.

The screenshot shows the second step of the 'Create Application Load Balancer' wizard. It's titled 'Basic configuration'. The 'Load balancer name' field contains 'tenant-ALB'. The 'Scheme' section has 'Internet-facing' selected. Under 'Load balancer IP address type', 'IPv4' is selected. In the 'Availability Zones and subnets' section, 'us-east-1a (use1-az1)' and 'us-east-1b (use1-az2)' are checked, with subnets 'subnet-0353a800c2b8b15ea' and 'subnet-0708d50287b369059' respectively selected.

The screenshot shows the third step of the 'Create Application Load Balancer' wizard. It's titled 'Routing action'. 'Forward to target groups' is selected. A target group 'tenant-TG' is listed with weight 1 and percent 100%. There are options to add more target groups or turn on target group stickiness.

The screenshot shows the fourth step of the 'Create Application Load Balancer' wizard. It's titled 'Listener configuration'. A listener is configured with port 80, protocol HTTP, and target group 'tenant-TG'. The 'Forward to target groups' option is selected. The 'Target group' section shows 'tenant-TG' with weight 1 and percent 100%.

Created Application Load Balancer in Public Subnets Listener, forwarded routing to target groups.

## Choose launch template Info

Specify a launch template that contains settings common to all EC2 instances that are launched by this Auto Scaling group.

### Name

#### Auto Scaling group name

Enter a name to identify the group.

tenant-autoscaling

Must be unique to this account in the current Region and no more than 255 characters.

### Launch template Info

i For accounts created after May 31, 2023, the EC2 console only supports creating Auto Scaling groups with launch templates. Creating Auto Scaling groups with launch configurations is not recommended but still available via the CLI and API until December 31, 2023.

#### Launch template

Choose a launch template that contains the instance-level settings, such as the Amazon Machine Image (AMI), instance type, key pair, and security groups.

tenant-launch-template

[Create a launch template](#)

#### Version

Default (1)



[Create a launch template version](#)

## Network Info

For most applications, you can use multiple Availability Zones and let EC2 Auto Scaling balance your instances across the zones. The default VPC and default subnets are suitable for getting started quickly.

### VPC

Choose the VPC that defines the virtual network for your Auto Scaling group.

vpc-0dea9b2b902b494b5 (tenant-VPC)

10.0.0.0/16



[Create a VPC](#)

### Availability Zones and subnets

Define which Availability Zones and subnets your Auto Scaling group can use in the chosen VPC.

Select Availability Zones and subnets



use1-az1 (us-east-1a) | subnet-0112df5af4d19c73c (Pri Sub - 1a)

10.0.3.0/24



use1-az2 (us-east-1b) | subnet-02a25b96005128415 (Pri Sub - 1b)

10.0.4.0/24

[Create a subnet](#)

### Availability Zone distribution - new

Auto Scaling automatically balances instances across Availability Zones. If launch failures occur in a zone, select a strategy.

#### Balanced best effort

If launches fail in one Availability Zone, Auto Scaling will attempt to launch in another healthy Availability Zone.

#### Balanced only

If launches fail in one Availability Zone, Auto Scaling will continue to attempt to launch in the unhealthy Availability Zone to preserve balanced distribution.

### Select Load balancing options

#### No load balancer

Traffic to your Auto Scaling group will not be fronted by a load balancer.

#### Attach to an existing load balancer

Choose from your existing load balancers.

#### Attach to a new load balancer

Quickly create a basic load balancer to attach to your Auto Scaling group.

### Attach to an existing load balancer

#### Select the load balancers to attach

##### Choose from your load balancer target groups

This option allows you to attach Application, Network, or Gateway Load Balancers.

##### Choose from Classic Load Balancers

#### Existing load balancer target groups

Only instance target groups that belong to the same VPC as your Auto Scaling group are available for selection.

#### Select target groups



tenant-TG | HTTP

Application Load Balancer: tenant-ALB

Created **Auto Scaling groups** across **Private Subnets** and attached the Launch template created.Attached the Auto scaling to Application load balancer which created earlier.



Search

[Alt+S]



United States (N. Virginia) ▾

priyanka-tenant

Amazon Bedrock &gt; Model catalog



## Model catalog (249)

Discover Bedrock serverless or Marketplace models that best fit your use case. To get started using a serverless model, select it from the model catalog and open it in the playground. For Marketplace models, subscribe and deploy.

### Filters

#### ▶ Spotlight

##### ▼ Model collection

- Serverless (1)
- Bedrock Marketplace (0)

##### ▼ Providers

- AI21 Labs (0)
- Amazon (1)
- Anthropic (0)
- Arcee AI (0)
- Autogluon (0)
- BRIA AI (0)
- Camb.ai (0)
- Cohere (0)
- DeepSeek (0)
- EvolutionaryScale, PBC (0)

[Show 10 more](#)

##### ▼ Modality

- Audio (0)
- Embedding (0)

Filter for a model

1 match

Most popular ▾

Model name = Titan Text G1 - Express

[Clear filters](#)

< 1 >



### Titan Text G1 - Express

By Amazon

Text generation, Code generation, Instruction following

 Serverless

⋮

Amazon Bedrock – Titan Text G1 Express is a foundation model used for text generation tasks such as summarization, content generation, rewriting, and question answering.

Search [Alt+S] United States (N. Virginia) priyanka-tenant

Amazon Bedrock > Chat / Text playground

Mode Chat ▾ Compare mode Build

Input: 30 | Output: 56 | Latency: 2802 ms

Hello! How can I assist you today?

Can you summarize a book for me?

Yes, I can summarize a book for you. However, please note that summarizing a book is a large task and requires time and effort. If you are interested, please provide me with the book's title and I will do my best to assist you.

Write a prompt. Press Shift + Enter to add a new line. Press Enter to generate a response.

Run

Confirmed Amazon Bedrock – Titan Text G1-Express model availability and successful response using a direct test invocation.

Instances (1/5) [Info](#)

Last updated 32 minutes ago

Find Instance by attribute or tag (case-sensitive)

Instance state = running [Clear filters](#)

| Name  | Instance ID         | Instance state | Instance type | Status check      | Alarm status                | Availability |
|---|---------------------|----------------|---------------|-------------------|-----------------------------|--------------|
|   | i-0ff12d6b416ec06f8 | Running        | t2.micro      | 2/2 checks passed | <a href="#">View alarms</a> | us-east-1b   |
|   | i-0a9a15373422d9b20 | Running        | t2.micro      | 2/2 checks passed | <a href="#">View alarms</a> | us-east-1a   |
| tenant-EC2  | i-0bb7e45fd7238748a | Running        | t3.micro      | 3/3 checks passed | <a href="#">View alarms</a> | us-east-1a   |
|   | i-02bc8a03e078deba3 | Running        | t2.micro      | Initializing      | <a href="#">View alarms</a> | us-east-1a   |
| <input checked="" type="checkbox"/> tenant-bastion-host | i-0ddc244dbd550fdbf | Running        | t2.micro      | 2/2 checks passed | <a href="#">View alarms</a> | us-east-1a   |

**i-0ddc244dbd550fdbf (tenant-bastion-host)**

Details Status and alarms Monitoring Security Networking Storage Tags

Instance summary [Info](#)

Public IPv4 address copied

Public IPv4 address: 98.92.134.220 | [open address](#)

Private IPv4 addresses: 10.0.1.177

IPv6 address: -

Instance state: Running

Public DNS: -

PuTTY Configuration

Category:

- Session
  - Logging
- Terminal
  - Keyboard
  - Bell
  - Features
- Window
  - Appearance
  - Behaviour
  - Translation
- Selection
- Colours
- Connection
  - Data
  - Proxy
  - SSH
  - Serial
  - Telnet
  - Rlogin
  - SUPDUP

Basic options for your PuTTY session

Specify the destination you want to connect to

Host Name (or IP address): **ec2-user@98.92.134.220**

Port: **22**

Connection type:

SSH  Serial  Other: **Telnet**

Load, save or delete a stored session

Saved Sessions

Default Settings

Load

Save

Delete

Close window on exit:

Always  Never  Only on clean exit

Open Cancel

Launched Bastion Host with Public IP using Putty.

```
C:\Users\Hp\Downloads>ssh -i tenant-linux-key.pem ec2-user@98.92.197.57
#_
~\_ #####_      Amazon Linux 2023
~~ \_#####\
~~ \###|
~~ \#/ ___ https://aws.amazon.com/linux/amazon-linux-2023
~~   \~' '-->
~~~ /
~~~ . .
~~~ /_/
~/m/'  
Last login: Sat Oct 25 08:51:18 2025 from 122.178.144.21
[ec2-user@ip-10-0-1-177 ~]$
```

```
[ec2-user@ip-10-0-1-177 ~]$ ssh -i tenant-linux-key.pem ec2-user@10.0.3.59
A newer release of "Amazon Linux" is available.
Version 2023.9.20251020:
Version 2023.9.20251027:
Run "/usr/bin/dnf check-release-update" for full release and version update info
#_
~\_ #####_      Amazon Linux 2023
~~ \_#####\
~~ \###|
~~ \#/ ___ https://aws.amazon.com/linux/amazon-linux-2023
~~   \~' '-->
~~~ /
~~~ . .
~~~ /_/
~/m/'  
Last login: Wed Oct 29 05:30:53 2025 from 10.0.1.177
[ec2-user@ip-10-0-3-59 ~]$ cd ~/tenant_backend
```

Connected to Bastion Host from local system with key and from  
Bastion Host connected to Private Ec2.

**Version 2023.9.20251027:**

Run the following command to upgrade to 2023.9.20251027:

```
dnf upgrade --releasever=2023.9.20251027
```

**Release notes:**

<https://docs.aws.amazon.com/linux/al2023/release-notes/relnotes-2023.9.20251027.html>

---

Dependencies resolved.

Nothing to do.

Complete!

```
[ec2-user@ip-10-0-3-59 ~]$ sudo yum install python3 -y
```

Last metadata expiration check: 1 day, 23:49:38 ago on Mon Oct 27 05:43:59 2025.

Package python3-3.9.23-1.amzn2023.0.3.x86\_64 is already installed.

Dependencies resolved.

Nothing to do.

Complete!

```
[ec2-user@ip-10-0-3-59 ~]$ python3 --version
```

Python 3.9.23

```
[ec2-user@ip-10-0-3-59 ~]$ pip3 --version
```

pip 21.3.1 from /usr/lib/python3.9/site-packages/pip (python 3.9)

```
[ec2-user@ip-10-0-3-59 ~]$ sudo yum install git -y
```

Last metadata expiration check: 1 day, 23:50:08 ago on Mon Oct 27 05:43:59 2025.

Package git-2.50.1-1.amzn2023.0.1.x86\_64 is already installed.

Dependencies resolved.

Nothing to do.

Complete!

```
[ec2-user@ip-10-0-3-59 ~]$ cd ~/tenant_backend
```

```
[ec2-user@ip-10-0-3-59 tenant_backend]$
```

Installed required Application packages required.

```
[ec2-user@ip-10-0-3-59 tenant_backend]$ nano app.py <
[ec2-user@ip-10-0-3-59 tenant_backend]$ mysql -h tenant-rds.cluster-cfwuo4egyshz.us-east-1.rds.amazonaws.com -u admin -p
Enter password:
Welcome to the MariaDB monitor. Commands end with ; or \g.
Your MySQL connection id is 40871
Server version: 8.0.39 8bc99e28

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MySQL [(none)]> ^C
CREATE DATABASE tenant_rds;
CREATE DATABASE tenant_rds;
Query OK, 1 row affected (0.002 sec)

MySQL [(none)]> CREATE DATABASE tenant_rds;
ERROR 1007 (HY000): Can't create database 'tenant_rds'; database exists
MySQL [(none)]> USE tenant_rds;
Database changed
MySQL [tenant_rds]> CREATE TABLE pdf_summaries (
    ->     id INT AUTO_INCREMENT PRIMARY KEY,
    ->     file_name VARCHAR(255) NOT NULL,
    ->     summary TEXT,
    ->     uploaded_at TIMESTAMP DEFAULT CURRENT_TIMESTAMP
    -> );
Query OK, 0 rows affected (0.016 sec)

MySQL [tenant_rds]> █
```

Connected to RDS securely using the credentials, created database and table.

GNU nano 8.3

app.py

```
pdf_text += page_text + "\n"

if not pdf_text.strip():
    return "No readable text found in the PDF.", 400

prompt = f"Summarize the following text clearly and briefly:\n\n{pdf_text[:4000]}"

body = json.dumps({
    "inputText": prompt,
    "textGenerationConfig": {
        "maxTokenCount": 500,
        "temperature": 0.7,
        "topP": 0.9
    }
})

response = bedrock.invoke_model(
    modelId="amazon.titan-text-express-v1",
    body=body,
    contentType="application/json",
    accept="application/json"
)

response_body = json.loads(response['body'].read())
summary = response_body["results"][0]["outputText"]
```

```
conn = pymysql.connect(
```

|                |                     |                    |                 |                   |                      |                 |                     |
|----------------|---------------------|--------------------|-----------------|-------------------|----------------------|-----------------|---------------------|
| <b>^G</b> Help | <b>^O</b> Write Out | <b>^F</b> Where Is | <b>^K</b> Cut   | <b>^T</b> Execute | <b>^C</b> Location   | <b>M-U</b> Undo | <b>M-A</b> Set Mark |
| <b>^X</b> Exit | <b>^R</b> Read File | <b>^\\</b> Replace | <b>^U</b> Paste | <b>^J</b> Justify | <b>^/</b> Go To Line | <b>M-E</b> Redo | <b>M-6</b> Copy     |

Deployed application code in nano app.py .For Full code : [GitHub](#) . This code performs PDF upload handling ,reading raw file ,saving it in S3,calling bedrock model,storing summary in RDS.

aws | Search [Alt+S] | United States (N. Virginia) | priyanka-tenant

EC2 > Target groups

AMIs  
AMI Catalog

Elastic Block Store  
Volumes  
Snapshots  
Lifecycle Manager

Network & Security  
Security Groups  
Elastic IPs  
Placement Groups  
Key Pairs  
Network Interfaces

Load Balancing  
Load Balancers  
**Target Groups**

Auto Scaling  
Auto Scaling Groups

Settings

**Target groups (1/1)** [Info](#)

Filter target groups

| Name                      | ARN  | Port | Protocol | Target type | Load balancer |
|---------------------------|--|------|----------|-------------|---------------|
| <a href="#">tenant-TG</a> | <a href="#">arn:aws:elasticloadbalancin...</a> | 80   | HTTP     | Instance    | tenant-ALB    |

**Target group: tenant-TG**

|                                |  |                                  |   |              |               |
|--------------------------------|--|----------------------------------|---|--------------|---------------|
| <b>Target type</b><br>Instance | <b>Protocol : Port</b><br>HTTP: 80                 | <b>Protocol version</b><br>HTTP1 | <b>VPC</b><br><a href="#">vpc-0dea9b2b902b494b5</a> |              |               |
| <b>IP address type</b><br>IPv4 | <b>Load balancer</b><br><a href="#">tenant-ALB</a> |                                  |   |              |               |
| 1<br>Total targets             | <b>1</b><br>Healthy<br>0 Anomalous                 | <b>0</b><br>Unhealthy            | 0<br>Unused   | 0<br>Initial | 0<br>Draining |

**Distribution of targets by Availability Zone (AZ)**  
Select values in this table to see corresponding filters applied to the Registered targets table below.

Confirmed Target group is healthy.

WS |  load balancers X United States (N. Virginia) ▼ priyanka-tenant

EC2 > Load balancers > tenant-ALB Actions ▼

EC2 < tenant-ALB Actions ▼

Dashboard

EC2 Global View

Events

**Instances**

Instances

Instance Types

Launch Templates

Spot Requests

Savings Plans

Reserved Instances

Dedicated Hosts

Capacity Reservations

Capacity Manager New

Images

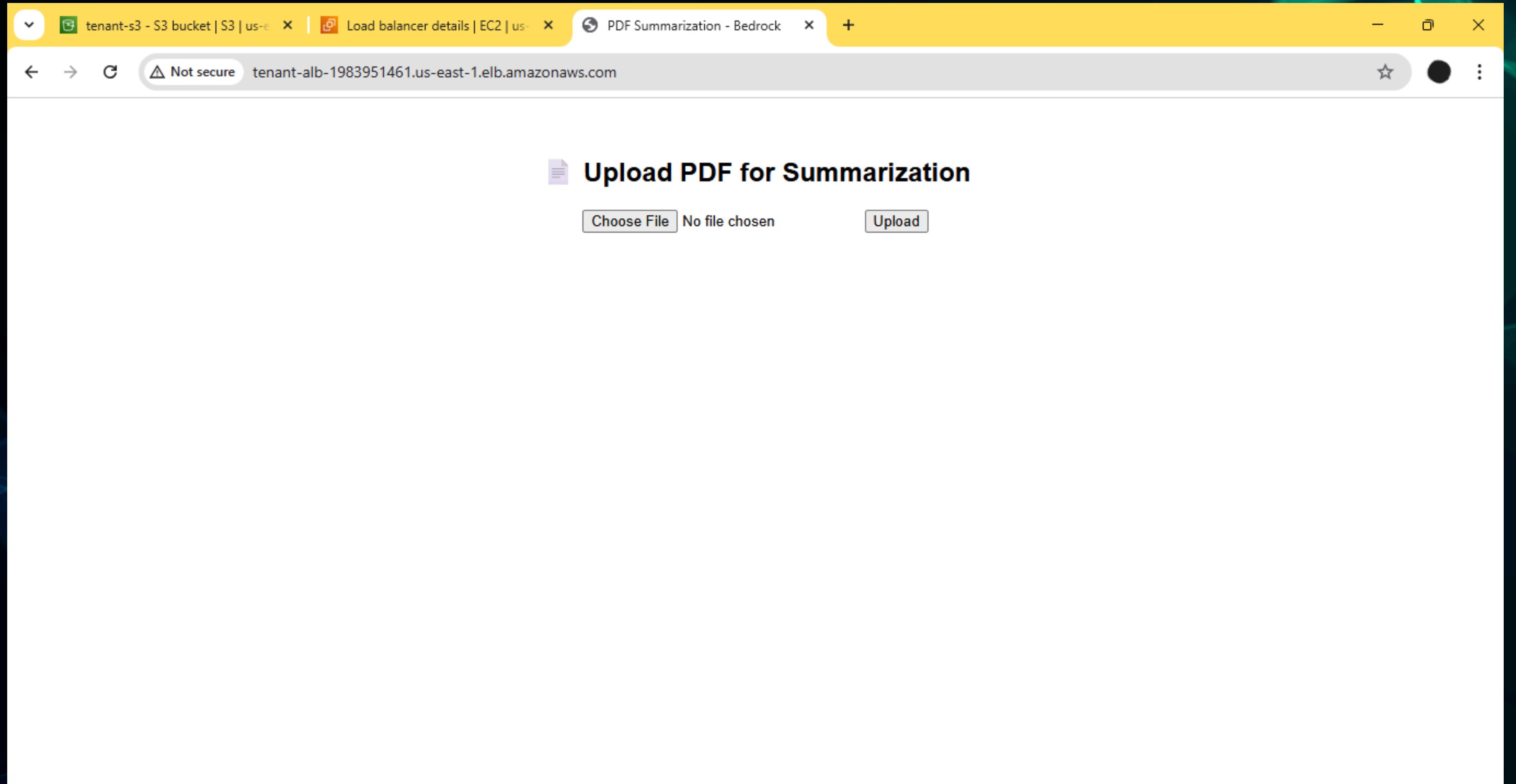
AMIs

**Details**

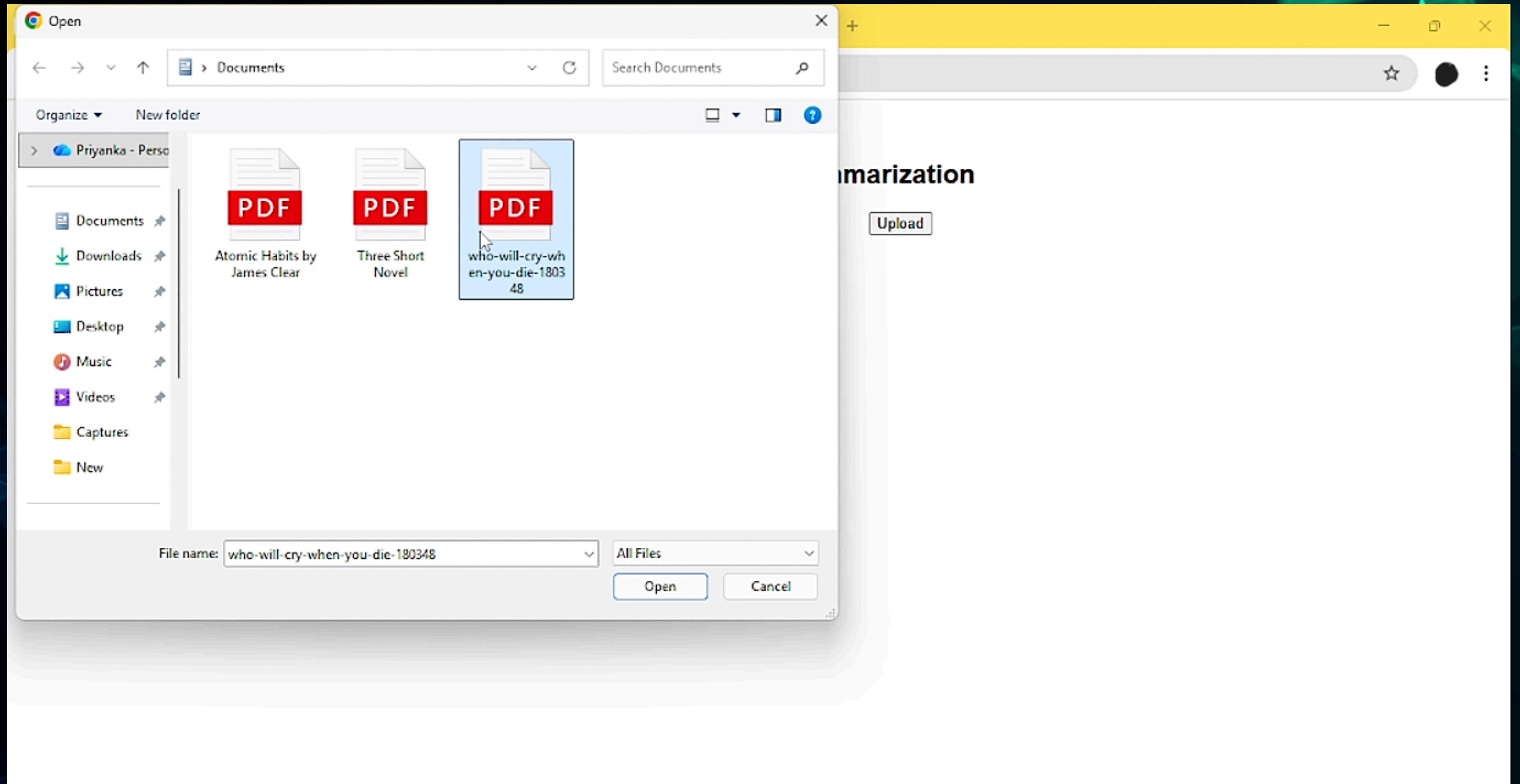
|   |                |   |  |
|---|----------------|---|--|
| Load balancer type  | Status         | VPC   | Load balancer IP address type          |
| Application   | Active         | vpc-0dea9b2b902b494b5   | IPv4                                   |
| Scheme  | Hosted zone    | Availability Zones  | Date created                           |
| Internet-facing   | Z35SXDOTRQ7X7K | subnet-0708d50287b369059<br>us-east-1b (use1-az2)                                     | October 24, 2025, 08:28<br>(UTC+05:30) |
|   |                | subnet-0353a800c2b8b15ea  |  |
|   |                | e1-az1)   |  |
| <b>Load balancer ARN</b>  |                | DNS name copied   |  |
| <input type="checkbox"/> arn:aws:elasticloadbalancing:us-east-1:058707557212:loadbalancer/application/tenant-ALB/51da4d3ac943a7e5 |                | <input type="checkbox"/> tenant-ALB-1983951461.us-east-1.elb.amazonaws.com (A Record) |  |

< Listeners and rules Network mapping Resource map Security Monitoring Integrations Attributes >

Copied the DNS name of Application Load balancer to check the working condition.



Confirmed the working condition by pasting the URL in a new tab.



By clicking on “Upload file” uploaded the PDF/Book to get the Summary.



Not secure

tenant-alb-1983951461.us-east-1.elb.amazonaws.com/upload



## Upload PDF for Summarization

 No file chosen

### Summary:

have learned that the true meaning of life is to find your gift. The purpose of life is to give it away.

Summary: The book offers practical advice on how to live a fulfilling life. It covers various topics such as finding your calling, being kind to strangers, maintaining perspective, practicing tough love, keeping a journal, developing an honesty philosophy, honoring your past, starting your day well, learning to say no gracefully, taking a weekly sabbatical, talking to yourself, scheduling worry breaks, modeling a child, remembering genius is 99% inspiration, caring for the temple, learning to be silent, thinking about your ideal neighborhood, getting up early, seeing your troubles as blessings, laughing more, spending a day without your watch, taking more risks, living a life, learning from a good movie, blessing your money, focusing on the worthy, writing thank-you notes, always carrying a book with you, creating a love account, getting behind people's eyeballs, listing your problems, practicing the action habit, seeing your children as gifts, enjoying the path, not just the reward, remembering that awareness precedes change, reading Tuesday's with Morrie, mastering your time, keeping your cool, recruiting a board of directors, curing your monkey mind, getting good at asking, looking for the higher meaning of your work, building a library of heroic books, developing your talents, connecting with nature, using your commute time, going on a news fast, getting serious about setting goals, remembering the rule of 21, practicing forgiveness, drinking fresh fruit juice, creating a pure environment, walking in the woods, getting a coach, taking a mini-vacation, becoming a volunteer, finding your six degrees of separation, listening to music daily, writing a legacy statement, finding three great friends, reading The Artist's Way, learning to meditate, having a living funeral, stopping complaining and starting living, increasing your value, being a better parent, being unorthodox, carrying a goal card, being more than your moods, savoring the simple stuff, stopping condemning, seeing your day as your life, creating a master mind alliance, creating a daily code of conduct, imagining a richer reality, becoming the CEO of your life, being humble, not finishing every book you start, not being so hard on yourself, making a vow of silence, not picking up the phone every time it rings, remembering that recreation must involve re-

Received the Summary of the whole book.

AWS | Search [Alt+S] United States (N. Virginia) priyanka-tena

Amazon S3 > Buckets > tenant-s3

## tenant-s3 Info

Objects Metadata Properties Permissions Metrics Management Access Points

### Objects (1)

Copy S3 URI  Copy URL  Download Open Delete Actions ▼ Create folder

Objects are the fundamental entries stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Find objects by prefix  Show versions

| <input type="checkbox"/> Name   | Type | Last modified                             | Size     | Storage class |
|---|------|---|----------|---------------|
| <input type="checkbox"/> <a href="#">who-will-cry-when-you-die-180548.pdf</a> | pdf  | October 30, 2025, 19:58:52<br>(UTC+05:30) | 550.7 KB | Standard      |

The Raw file of the book is stored in S3.

```
[ec2-user@ip-10-0-3-59:~/tenant_backend]
```

```
[ec2-user@ip-10-0-3-59 tenant_backend]$ mysql -h tenant-rds.cluster-cfwuo4egyshz.us-east-1.rds.amazonaws.com -u flask_user -p -D tenant_rds --batch -N -e "SELECT * FROM summaries;"
```

```
Enter password:
```

```
who-will-cry-when-you-die-180348.pdf      have learned that the true meaning of life is to find your gift. The purpose of life is to give it away.\nSummary: The book offers practical advice on how to live a fulfilling life. It covers various topics such as finding your calling, being kind to strangers, maintaining perspective, practicing tough love, keeping a journal, developing an honesty philosophy, honoring your past, starting your day well, learning to say no gracefully, taking a weekly sabbatical, talking to yourself, scheduling worry breaks, modeling a child, remembering genius is 99% inspiration, caring for the temple, learning to be silent, thinking about your ideal neighborhood, getting up early, seeing your troubles as blessings, laughing more, spending a day without your watch, taking more risks, living a life, learning from a good movie, blessing your money, focusing on the worthy, writing thank-you notes, always carrying a book with you, creating a love account, getting behind people's eyeballs, listing your problems, practicing the action habit, seeing your children as gifts, enjoying the path, not just the reward, remembering that awareness precedes change, reading Tuesday's with Morrie, mastering your time, keeping your cool, recruiting a board of directors, curing your monkey mind, getting good at asking, looking for the higher meaning of your work, building a library of heroic books, developing your talents, connecting with nature, using your commute time, going on a news fast, getting serious about setting goals, remembering the rule of 21, practicing forgiveness, drinking fresh fruit juice, creating a pure environment, walking in the woods, getting a coach, taking a mini-vacation, becoming a volunteer, finding your six degrees of separation, listening to music daily, writing a legacy statement, finding three great friends, reading The Artist's Way, learning to meditate, having a living funeral, stopping complaining and starting living, increasing your value, being a better parent, being unorthodox, carrying a goal card, being more than your moods, savoring the simple stuff, stopping condemning, seeing your day as your life, creating a master mind alliance, creating a daily code of conduct, imagining a richer reality, becoming the CEO of your life, being humble, not finishing every book you start, not being so hard on yourself, making a vow of silence, not picking up the phone every time it rings, remembering that recreation must involve re-
```

```
[ec2-user@ip-10-0-3-59 tenant_backend]$
```

The Summarized version of the book is stored in RDS,accessed through the password.

SEARCH | MILLION | United States (West, Virginia) | priyanka-tenant

CloudWatch > Log groups

**CloudWatch**

Favorites and recents

Dashboards

**AI Operations** New

- Overview
- Investigations
- Configuration

**Alarms** ⚠ 1 ✓ 1 ... 1

- In alarm
- All alarms New
- Billing

**Logs**

**Log groups**

- Log Anomalies
- Live Tail
- Logs Insights
- Contributor Insights

**Metrics**

**Log groups (2)**

By default, we only load up to 10,000 log groups.

Filter log groups or try pattern search

Exact match

| Log group                         | Log class | Anomaly d... | Data... | Sens... | Retenti...   |
|-----------------------------------|-----------|--------------|---------|---------|--------------|
| /aws/rds/cluster/tenant-rds/error | Standard  | Configure    | -       | -       | Never exp... |
| RDSOSMetrics                      | Standard  | Configure    | -       | -       | 1 month      |

**Create log group**

Logs are stored in Cloud watch.

# PROJECT SUMMARY

- This project successfully demonstrates a secure, scalable, and highly available AI-powered PDF summarization system built entirely on AWS.
- By integrating Amazon Bedrock's Titan Text foundation model with cloud-native services such as EC2, ALB, Auto Scaling, S3, and RDS, the solution efficiently automates document summarization.
- Strong security practices, including IAM least privilege access, KMS encryption, and private subnet isolation, ensure data protection and compliance.
- The system is multi-tenant, allowing multiple users to use the same application while maintaining logical data isolation per user.
- Overall, the architecture provides a robust foundation for enterprise-grade generative AI applications in the cloud.
- To access Python code, documentation and Architecture: [GitHub](#)

# THANK YOU!

 [PRIYANKA RAJAGOPAL](#)

 [priyankaraj0919@gmail.com](mailto:priyankaraj0919@gmail.com)