Fruits!

Déployez un modèle dans le cloud

Data Science | Projet 8

Fırat Yasar
20/02/2022

Sommaire

Présentation

- Présentation de la problématique
- Découverte du jeu de données

Architecture Big Data

- Qu’est-ce que le Big Data ?
- Traitement du Big Data
- Automatisation du calcul distribué sur des données massives
- Hadoop MapReduce
- Spark

Amazon Web Services

- La console d’AWS
- Les services utilisés
- Création d’un cluster via Databricks
- Le notebook

La chaîne de traitement

- Le pipeline

Annexes

Présentation

Présentation de la problématique

START-UP !

L’objectif :

Notre start-up **Fruits** fait partie de l’**AgriTech de France** et cherche à proposer des solutions innovantes pour la récolte des fruits. Le but est de mettre à disposition du grand public une application mobile qui permettrait aux utilisateurs de prendre en photo un fruit et d'obtenir des informations sur ce fruit.

La mission:

Mettre en œuvre une première version du **moteur de classification** des images de fruits.

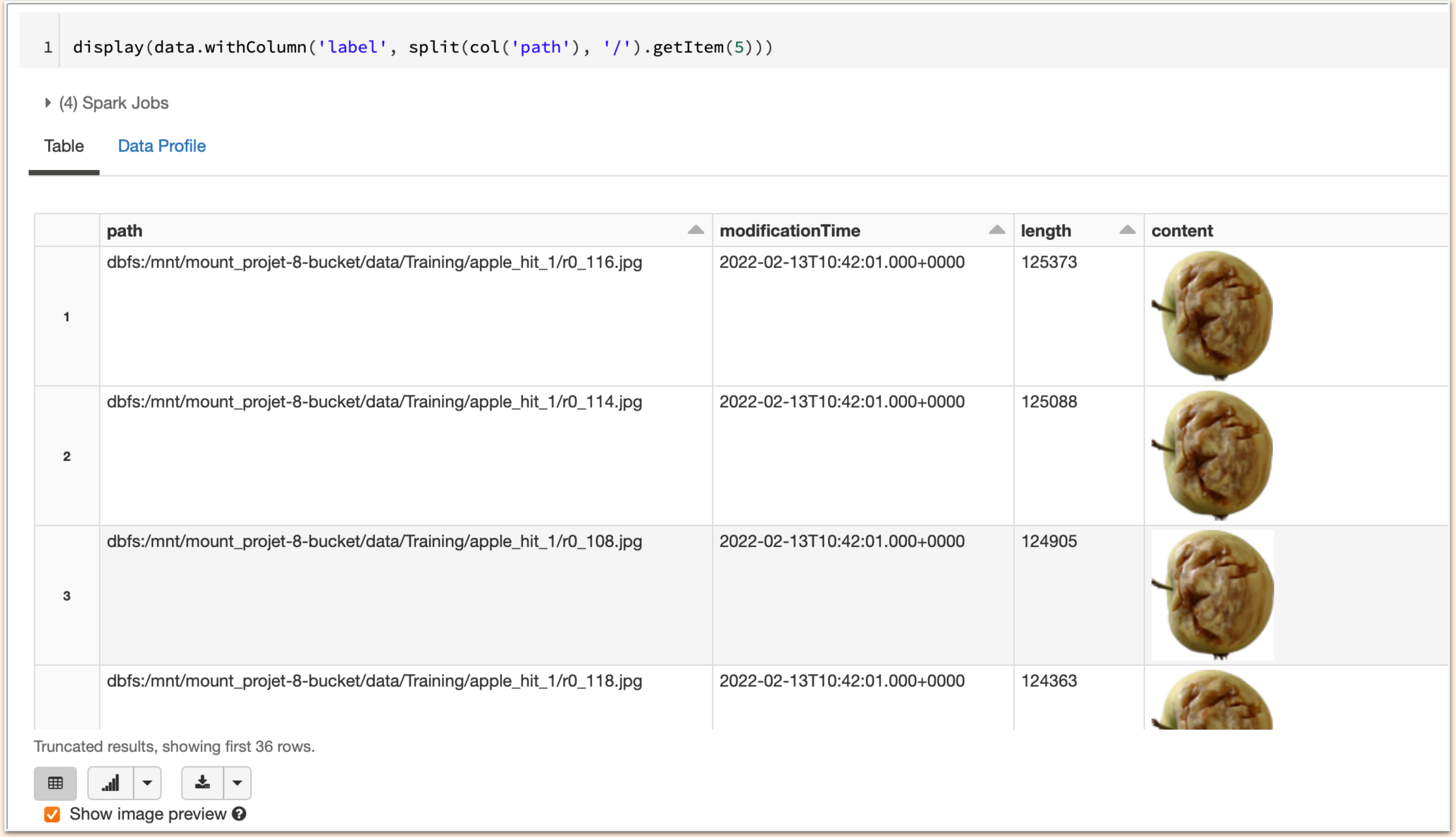
Construire dans un environnement **Big Data** une première chaîne de traitement des données qui comprendra le preprocessing et une étape de réduction de dimension.

La base de données :

Pour cela, nous avons à notre disposition la base de données :

https://www.kaggle.com/moltean/fruits

Découverte du jeu de données



```
display(data.withColumn('label', split(col('path'), '/').getItem(5)))
```

	path	modificationTime	length	content
1	dbfs:/mnt/mount_projet-8-bucket/data/Training/apple_hit_1/r0_116.jpg	2022-02-13T10:42:01.000+0000	125373	
2	dbfs:/mnt/mount_projet-8-bucket/data/Training/apple_hit_1/r0_114.jpg	2022-02-13T10:42:01.000+0000	125088	
3	dbfs:/mnt/mount_projet-8-bucket/data/Training/apple_hit_1/r0_108.jpg	2022-02-13T10:42:01.000+0000	124905	
	dbfs:/mnt/mount_projet-8-bucket/data/Training/apple_hit_1/r0_118.jpg	2022-02-13T10:42:01.000+0000	124363	

Truncated results, showing first 36 rows.

Data entraînement
67 692 images

Data test
22 688 images

une image = un fruit ou un légume

Taille d'une image : 100 x 100 pixels

Nombre de classes : 131

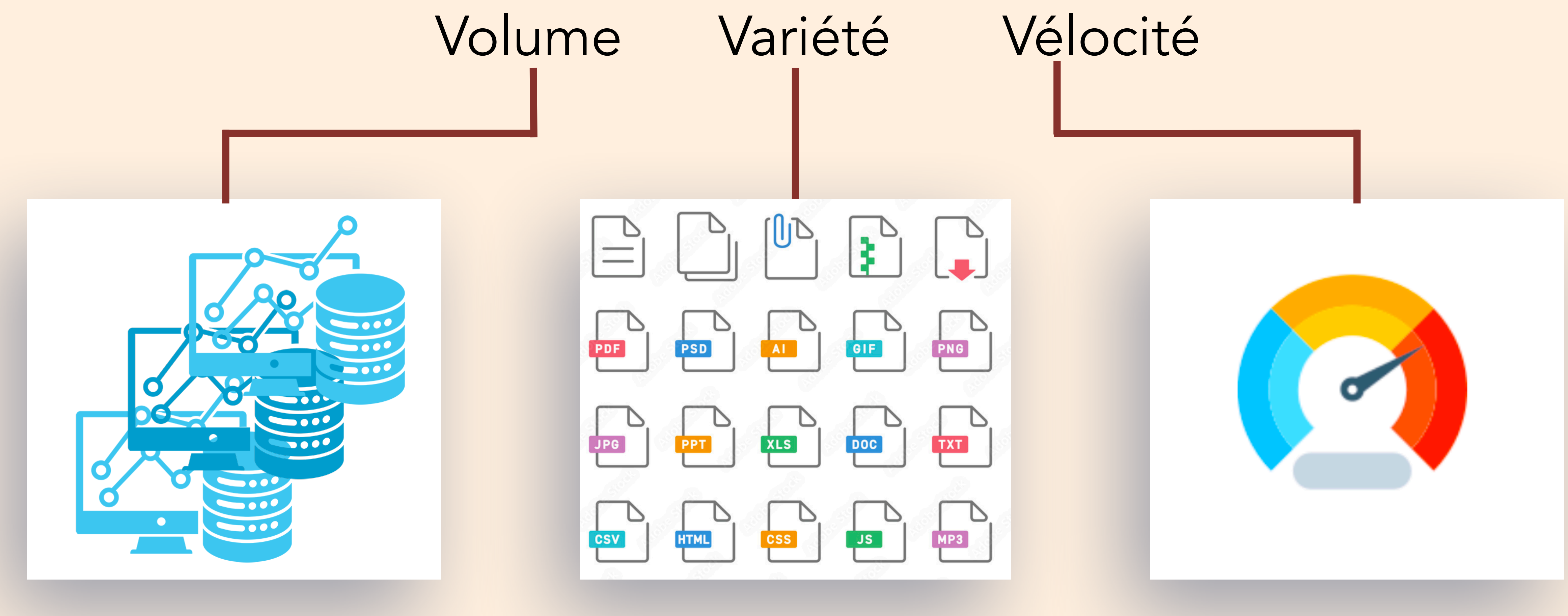
Architecture Big Data

Qu’est-ce que le Big Data ?

Définition :

Le **big data,** ou les données massives, désigne les ressources d’informations dont les caractéristiques en termes de **volume**, de **vélocité** et de **variété** imposent l’utilisation de technologies et de méthodes analytiques particulières pour générer de la valeur et qui dépassent en général les capacités d'une seule et unique machine et nécessitent des traitements parallélisés.

Les trois “V” de big data :



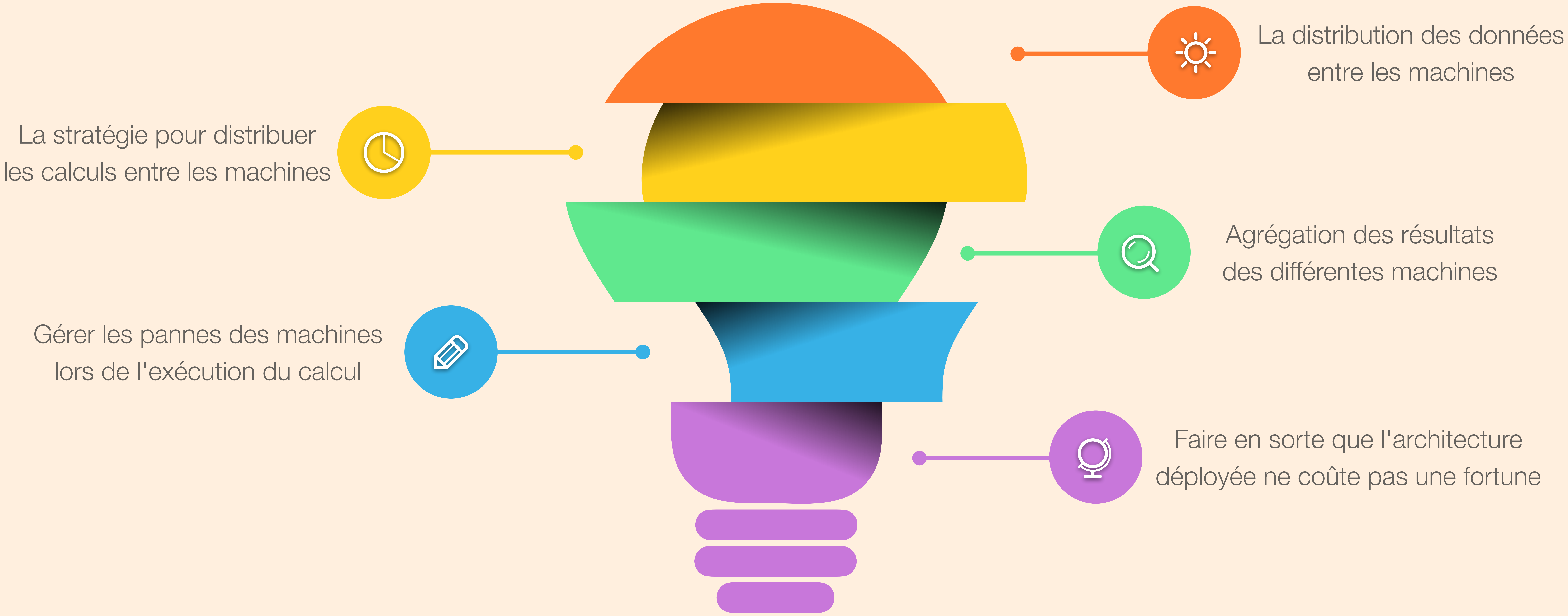
Volume

Variété

Vélocité

Traitement du Big Data

Traitement du Big Data



La distribution des données entre les machines

La stratégie pour distribuer les calculs entre les machines

Agrégation des résultats des différentes machines

Gérer les pannes des machines lors de l'exécution du calcul

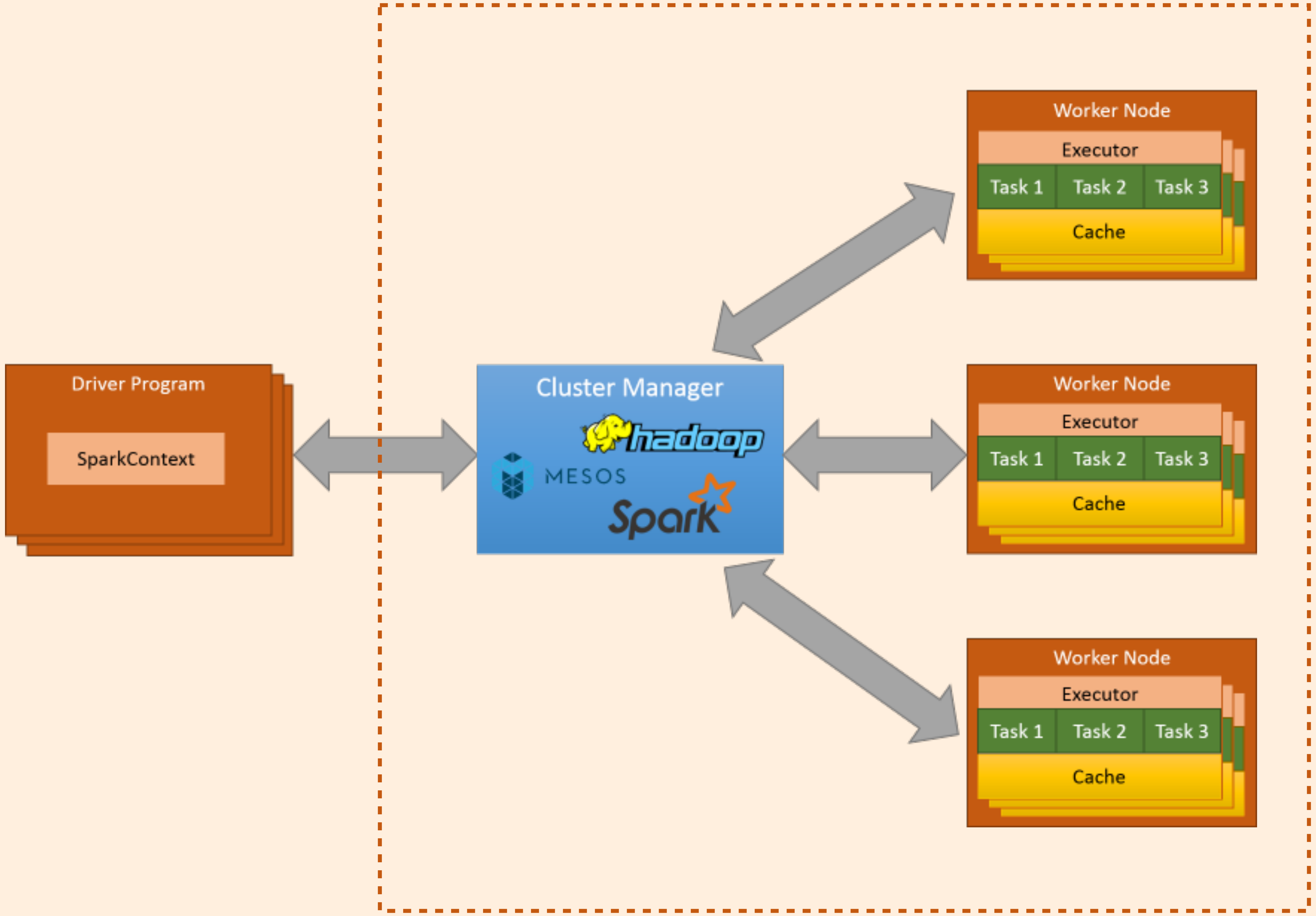
Faire en sorte que l'architecture déployée ne coûte pas une fortune

Automatisation du calcul distribué sur des données massives

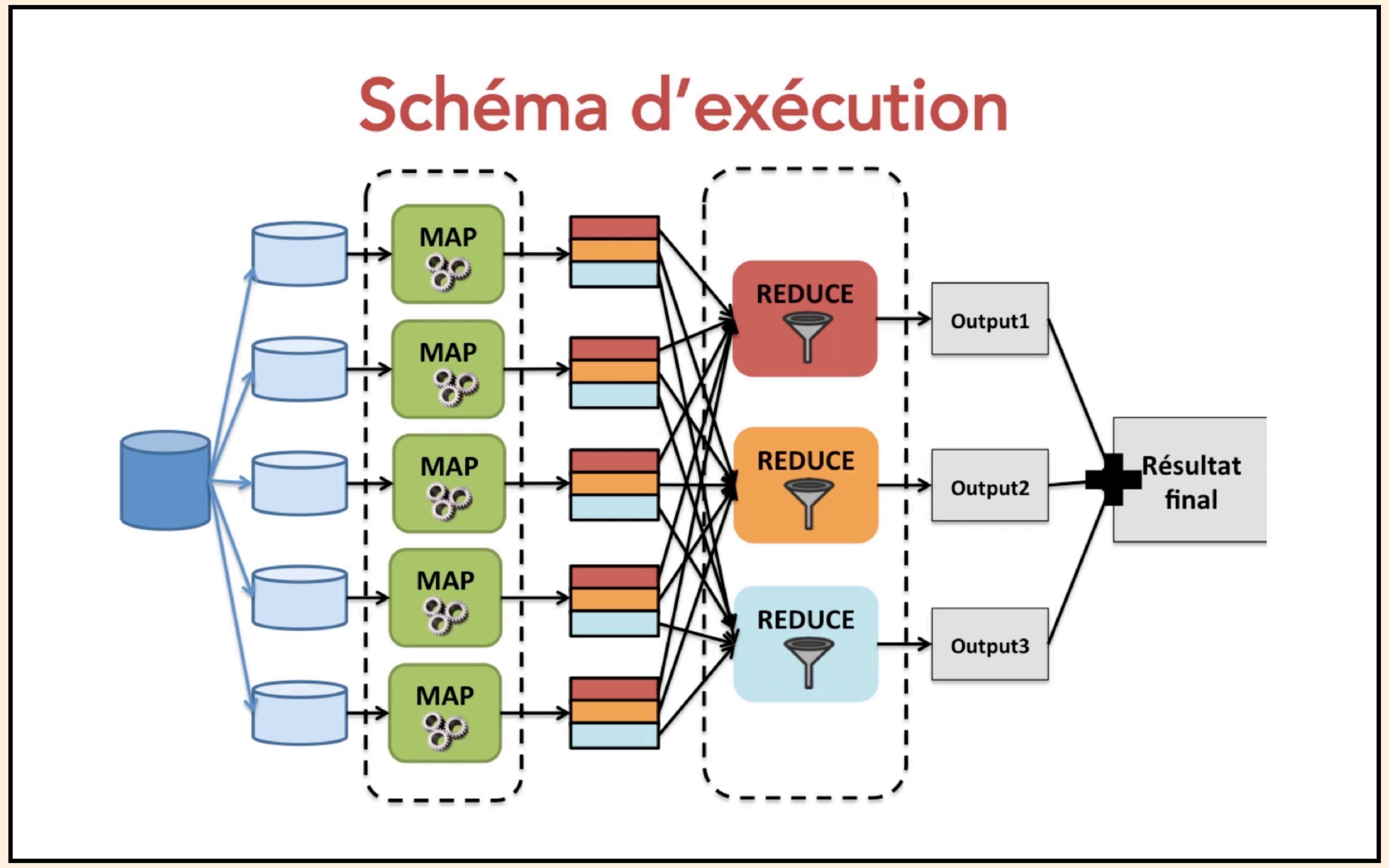
- **Le calcul distribué :**

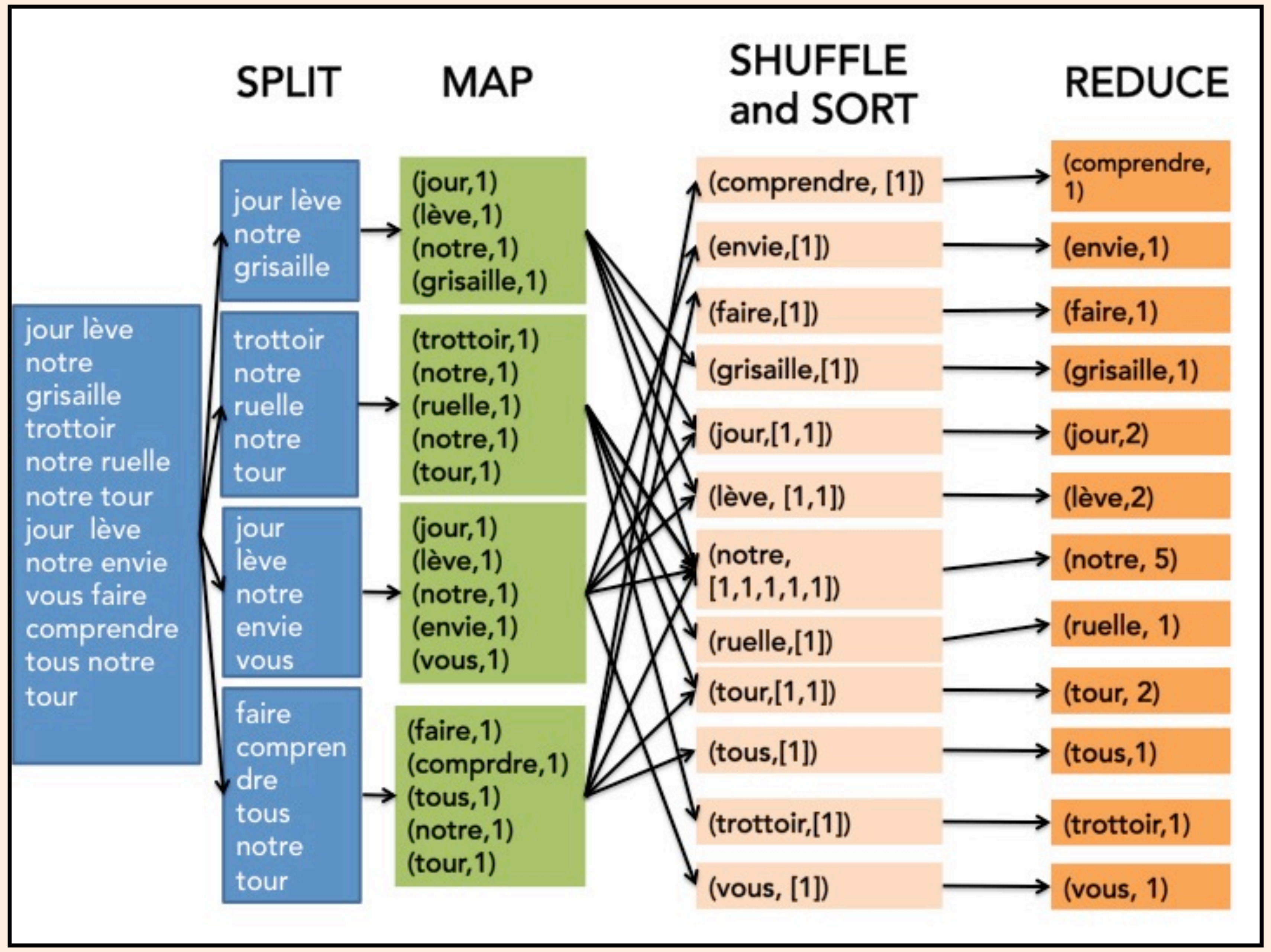
Les nœuds sur lesquels les calculs sont exécutés sont distants, autonomes et ne partagent pas de ressources ; la communication entre les nœuds s'effectue grâce à l'envoi de messages, au sein d'**un cluster.**

SparkContext est le point d'accès à toutes les fonctionnalités de Spark.



Hadoop MapReduce




SPLIT
MAP
SHUFFLE and SORT
REDUCE
jour lève notre grisaille trottoir notre ruelle notre tour jour lève notre envie vous faire comprendre tous notre tour
jour lève notre grisaille
trottoir notre ruelle notre tour
jour lève notre envie vous
faire compren dre tous notre tour
(jour,1) (lève,1) (notre,1) (grisaille,1)
(trottoir,1) (notre,1) (ruelle,1) (notre,1) (tour,1)
(jour,1) (lève,1) (notre,1) (envie,1) (vous,1)
(faire,1) (comprdre,1) (tous,1) (notre,1) (tour,1)
(comprendre, [1])
(envie,[1])
(faire,[1])
(grisaille,[1])
(jour,[1,1])
(lève, [1,1])
(notre, [1,1,1,1,1])
(ruelle,[1])
(tour,[1,1])
(tous,[1])
(trottoir,[1])
(vous, [1])
(comprendre, 1)
(envie,1)
(faire,1)
(grisaille,1)
(jour,2)
(lève,2)
(notre, 5)
(ruelle, 1)
(tour, 2)
(tous,1)
(trottoir,1)
(vous, 1)

- **Le fonctionnement général de MapReduce :**

1. L'ensemble des données à traiter est découpé en plusieurs lots ou sous-ensembles.

2. Dans une première étape, l'étape `MAP`, l'opération map, spécifiée pour notre problème, est appliquée à chaque lot. Cette opération transforme la paire `(clé, valeur)` représentant le lot en une liste de nouvelles paires (clé, valeur) constituant ainsi des résultats intermédiaires du traitement à effectuer sur les données complètes.

3. Avant d'être envoyés à l'étape REDUCE, les résultats intermédiaires sont regroupés et triés par clé. C'est l'étape de `SHUFFLE` and `SORT`.

4. Enfin, l'étape `REDUCE` consiste à appliquer l'opération reduce, spécifiée pour notre problème, à chaque clé. Elle agrège tous les résultats intermédiaires associés à une même clé et renvoie donc pour chaque clé une valeur unique.

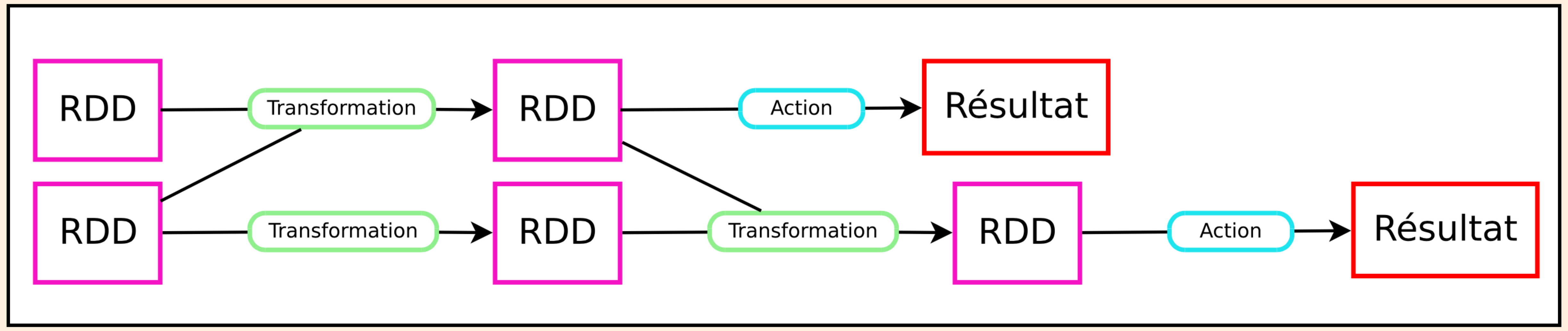
Spark

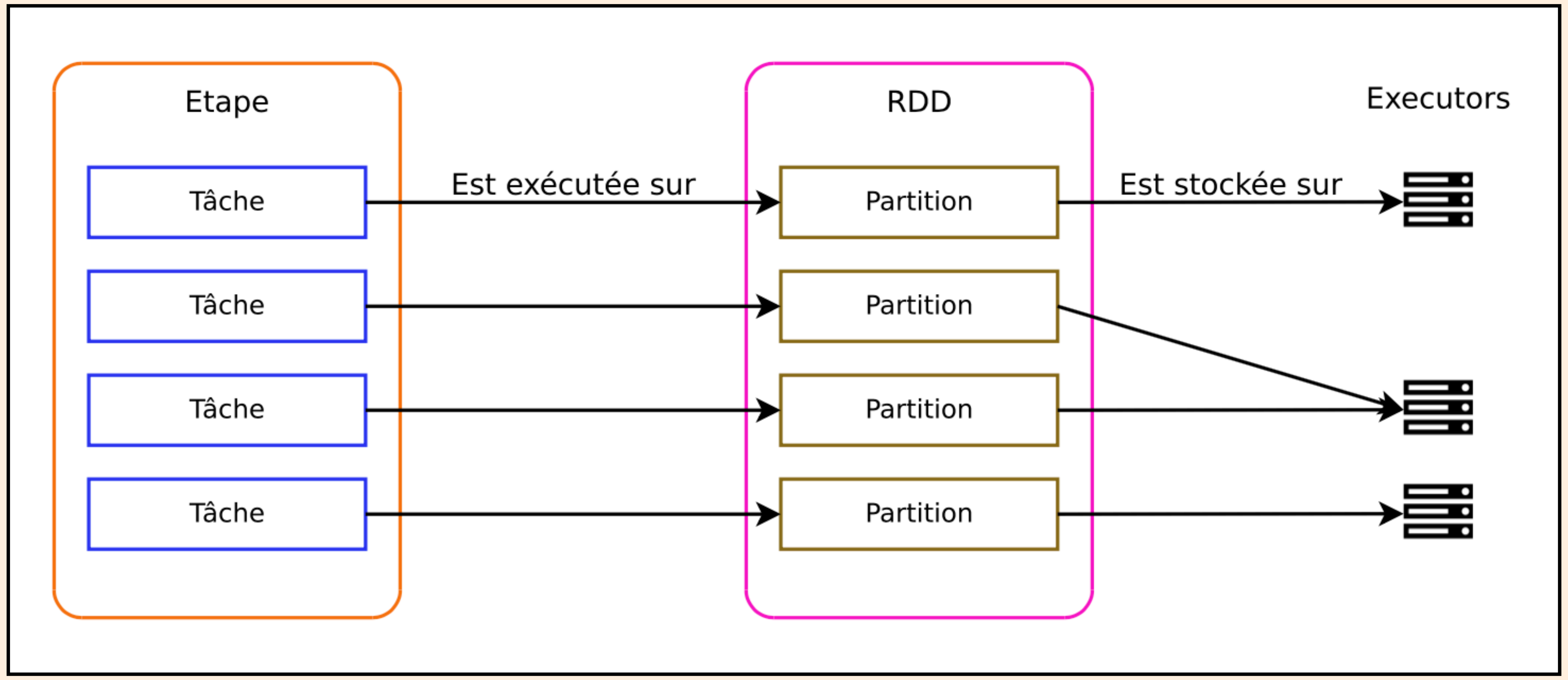
Le fonctionnement général de Spark :

- Utilisation des RDD (**R**esilient **D**istributed **D**ataset). RDD supporte deux type d'opérations : transformation et action.
  - Transformation : consiste à appliquer une fonction sur *n* RDD et à retourner un nouveau RDD (map, filter, join etc.).
  - Action : consiste à appliquer une fonction sur *n* RDD et à retourner une valeur (collect, show, count etc.).
- Les transformations évitent les calculs inutiles. Spark exécute les expressions uniquement lorsqu'elles sont nécessaires.
- Les RDD sont exécutés en mémoire de façon complètement tolérante aux pannes.

***

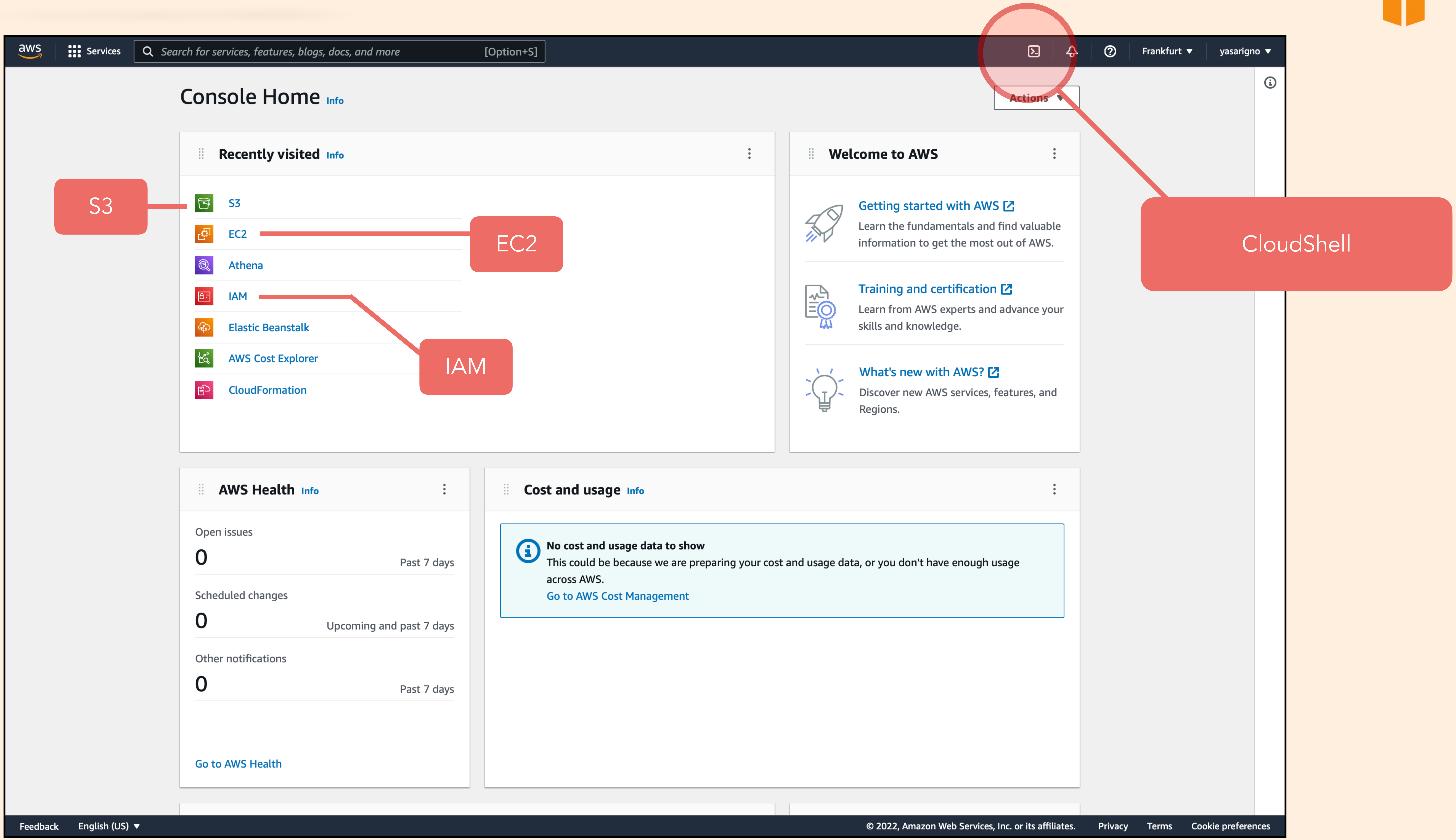
- Spark est basé sur la programmation fonctionnelle, **SCALA**





Amazon Web Services

La console d’AWS



Console Home Info

Recently visited Info

- S3
- EC2
- Athena
- IAM
- Elastic Beanstalk
- AWS Cost Explorer
- CloudFormation

Welcome to AWS

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 Info

Open issues
0 Past 7 days

Scheduled changes
0 Upcoming and past 7 days

Other notifications
0 Past 7 days

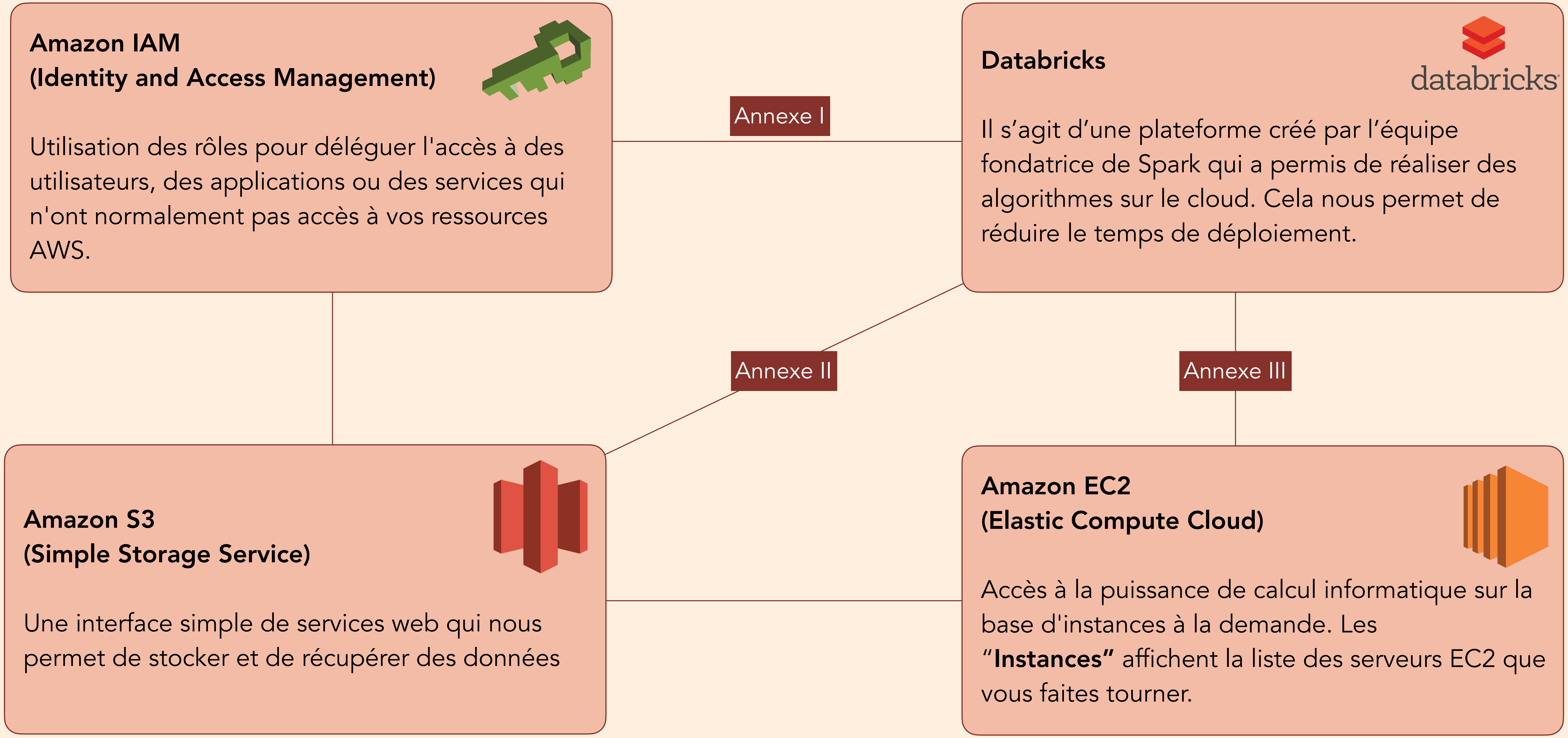
Go to AWS Health

Cost and usage Info

No cost and usage data to show
This could be because we are preparing your cost and usage data, or you don't have enough usage across AWS.
Go to AWS Cost Management

© 2022, Amazon Web Services, Inc. or its affiliates.

Les services utilisés



Amazon IAM
(Identity and Access Management)

Utilisation des rôles pour déléguer l'accès à des utilisateurs, des applications ou des services qui n'ont normalement pas accès à vos ressources AWS.

Annexe I

Databricks

Il s'agit d'une plateforme créé par l'équipe fondatrice de Spark qui a permis de réaliser des algorithmes sur le cloud. Cela nous permet de réduire le temps de déploiement.

Annexe II

Annexe III

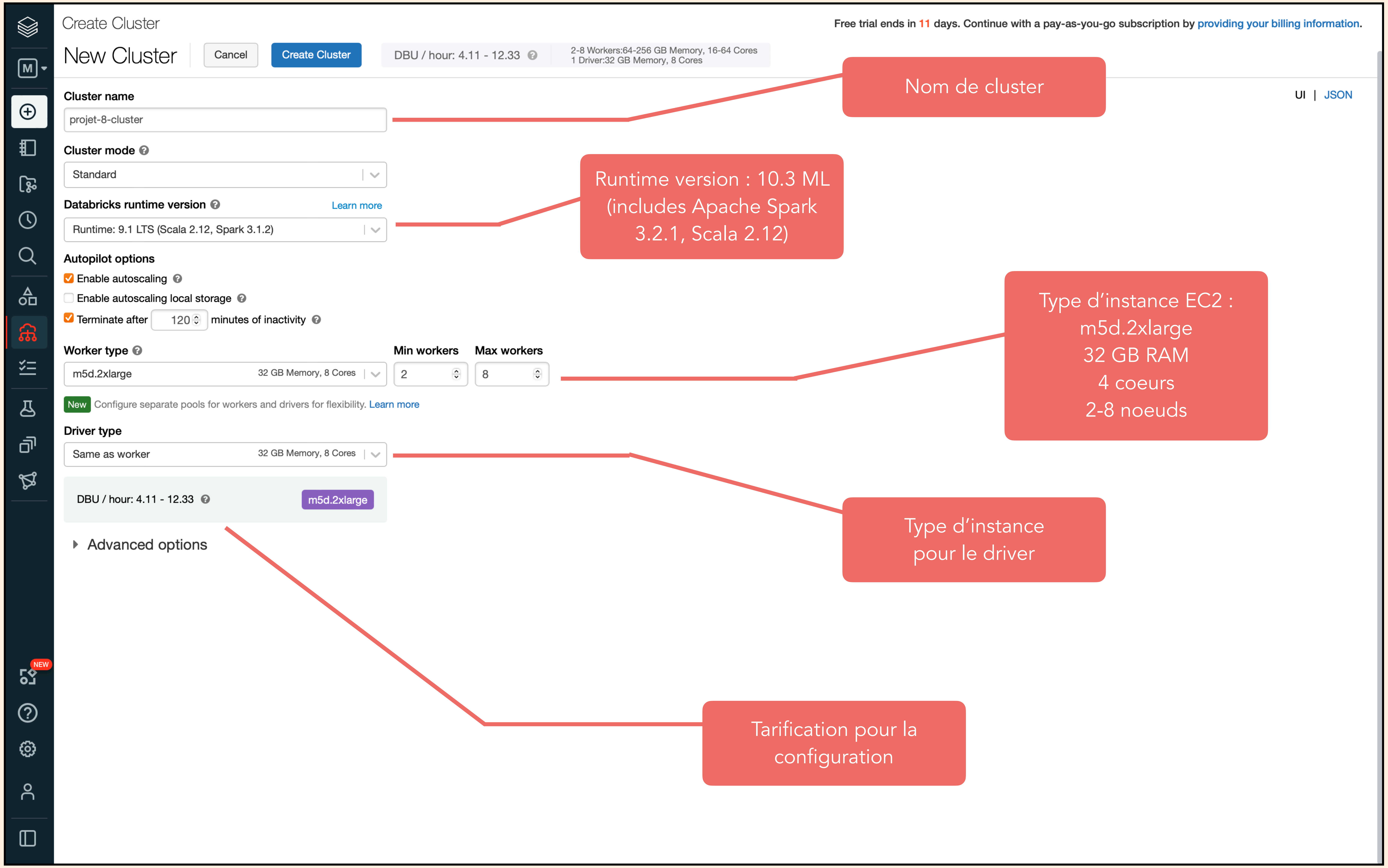
Amazon S3
(Simple Storage Service)

Une interface simple de services web qui nous permet de stocker et de récupérer des données

Amazon EC2
(Elastic Compute Cloud)

Accès à la puissance de calcul informatique sur la base d'instances à la demande. Les "**Instances**" affichent la liste des serveurs EC2 que vous faites tourner.

Création d'un cluster via Databricks



Create Cluster

Free trial ends in 11 days. Continue with a pay-as-you-go subscription by providing your billing information.

New Cluster

DBU / hour: 4.11 - 12.33

2-8 Workers:64-256 GB Memory, 16-64 Cores
1 Driver:32 GB Memory, 8 Cores

Cluster name: projet-8-cluster

Cluster mode: Standard

Databricks runtime version: Runtime: 9.1 LTS (Scala 2.12, Spark 3.1.2)

Autopilot options

- ☑ Enable autoscaling
- □ Enable autoscaling local storage
- ☑ Terminate after 120 minutes of inactivity

Worker type: m5d.2xlarge — 32 GB Memory, 8 Cores

Min workers: 2

Max workers: 8

New Configure separate pools for workers and drivers for flexibility. Learn more

Driver type: Same as worker — 32 GB Memory, 8 Cores

DBU / hour: 4.11 - 12.33 m5d.2xlarge

Nom de cluster

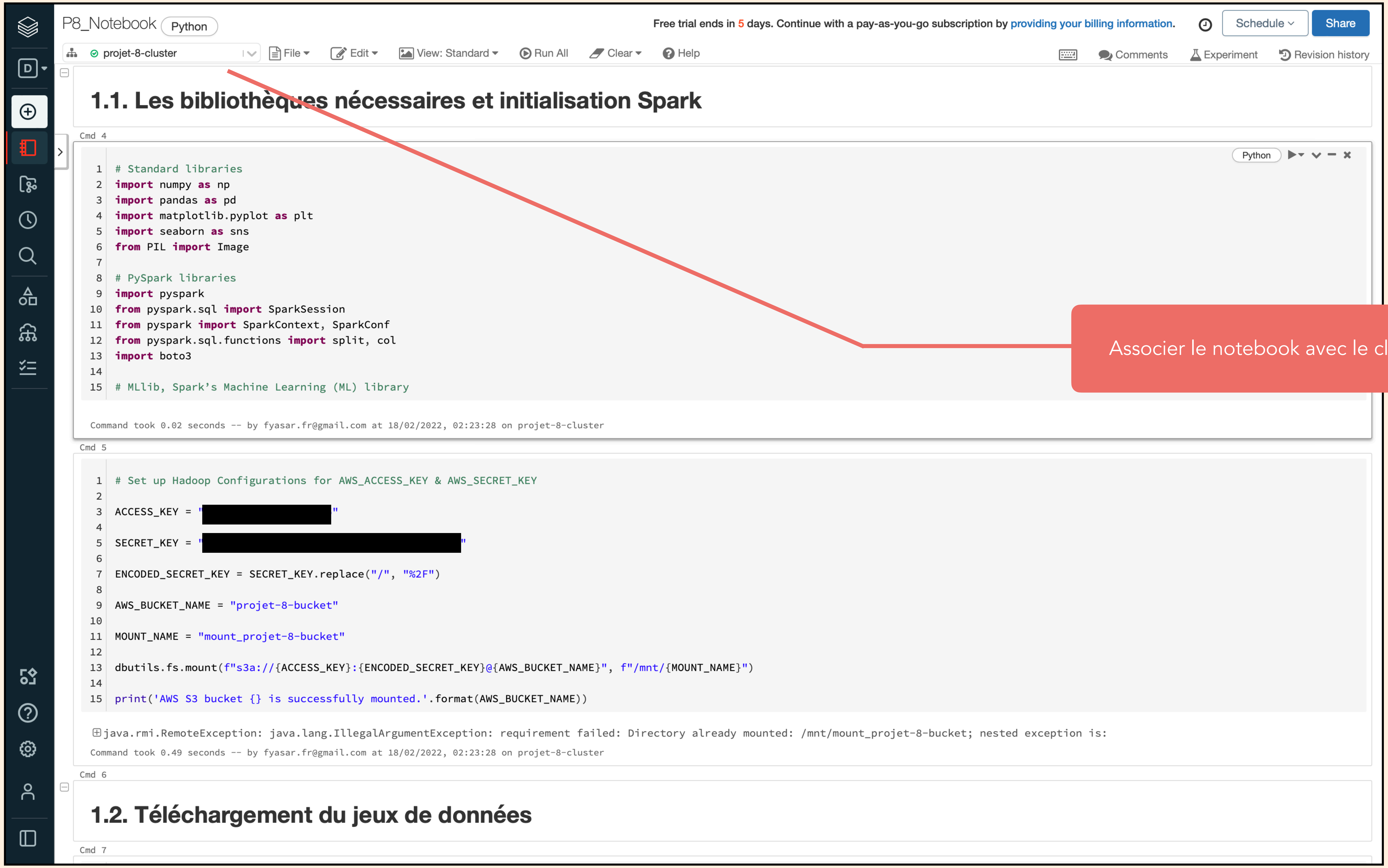
Runtime version : 10.3 ML (includes Apache Spark 3.2.1, Scala 2.12)

Type d'instance EC2 :
m5d.2xlarge
32 GB RAM
4 coeurs
2-8 noeuds

Type d'instance pour le driver

Tarification pour la configuration

Le notebook



1.1. Les bibliothèques nécessaires et initialisation Spark

Cmd 4

```
# Standard libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from PIL import Image

# PySpark libraries
import pyspark
from pyspark.sql import SparkSession
from pyspark import SparkContext, SparkConf
from pyspark.sql.functions import split, col
import boto3

# MLlib, Spark's Machine Learning (ML) library
```

```
Command took 0.02 seconds -- by fyasar.fr@gmail.com at 18/02/2022, 02:23:28 on projet-8-cluster
```

Cmd 5

```
# Set up Hadoop Configurations for AWS_ACCESS_KEY & AWS_SECRET_KEY

ACCESS_KEY = "[redacted]"

SECRET_KEY = "[redacted]"

ENCODED_SECRET_KEY = SECRET_KEY.replace("/", "%2F")

AWS_BUCKET_NAME = "projet-8-bucket"

MOUNT_NAME = "mount_projet-8-bucket"

dbutils.fs.mount(f"s3a://{ACCESS_KEY}:{ENCODED_SECRET_KEY}@{AWS_BUCKET_NAME}", f"/mnt/{MOUNT_NAME}")

print('AWS S3 bucket {} is successfully mounted.'.format(AWS_BUCKET_NAME))
```

```
⊞java.rmi.RemoteException: java.lang.IllegalArgumentException: requirement failed: Directory already mounted: /mnt/mount_projet-8-bucket; nested exception is:
Command took 0.49 seconds -- by fyasar.fr@gmail.com at 18/02/2022, 02:23:28 on projet-8-cluster
```

Cmd 6

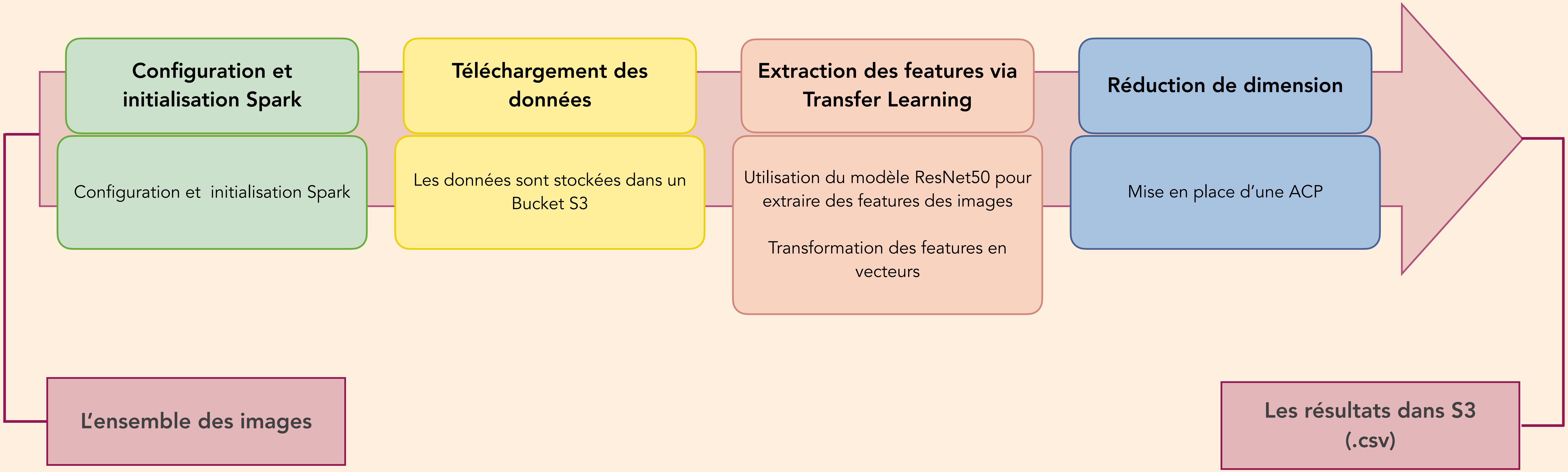
1.2. Téléchargement du jeux de données

Cmd 7

Associer le notebook avec le cluster configuré

La chaîne de traitement

Le pipeline



Configuration et initialisation Spark

Configuration et initialisation Spark

Téléchargement des données

Les données sont stockées dans un Bucket S3

Extraction des features via Transfer Learning

Utilisation du modèle ResNet50 pour extraire des features des images

Transformation des features en vecteurs

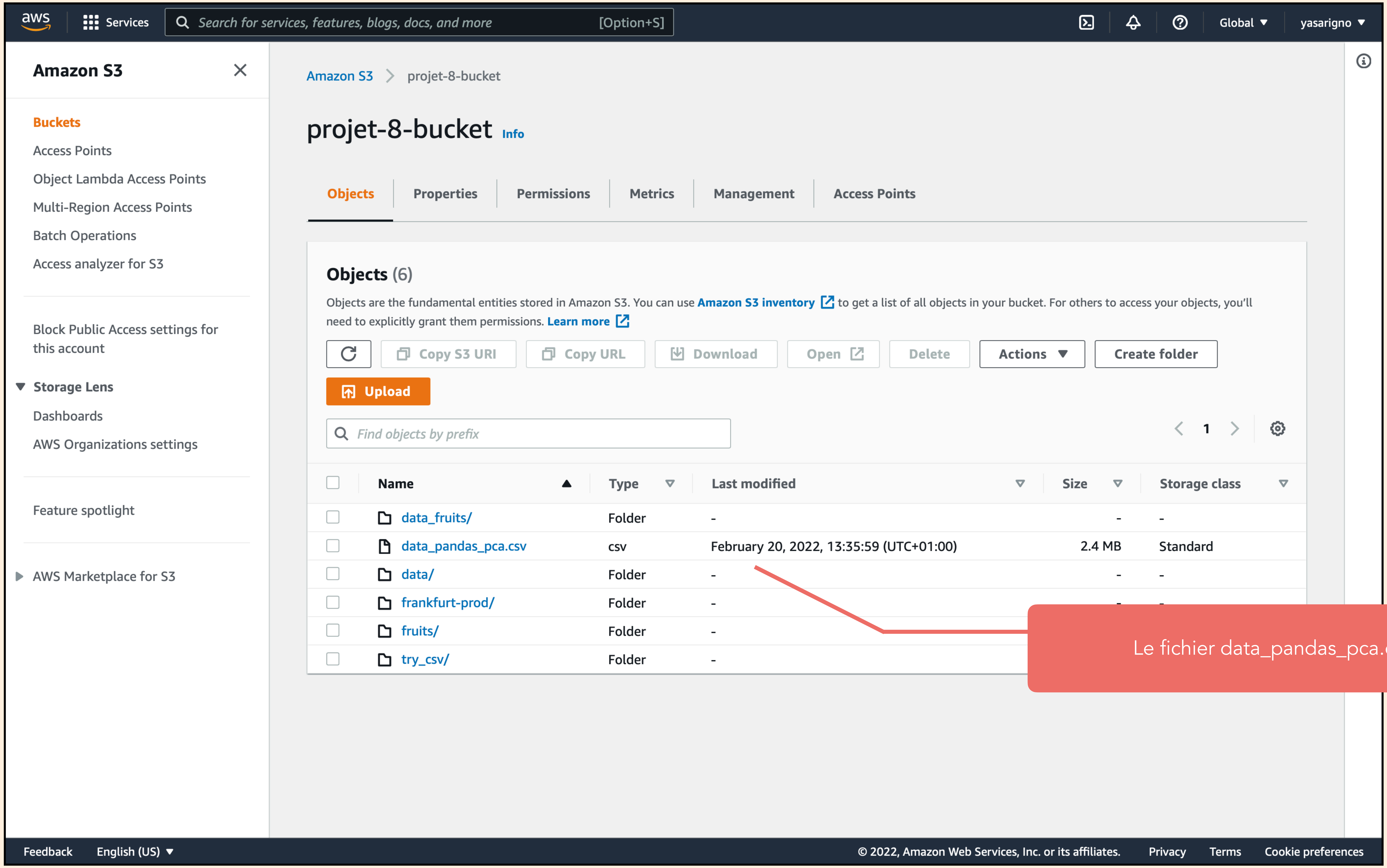
Réduction de dimension

Mise en place d'une ACP

L'ensemble des images

Les résultats dans S3 (.csv)

Les résultats



projet-8-bucket Info

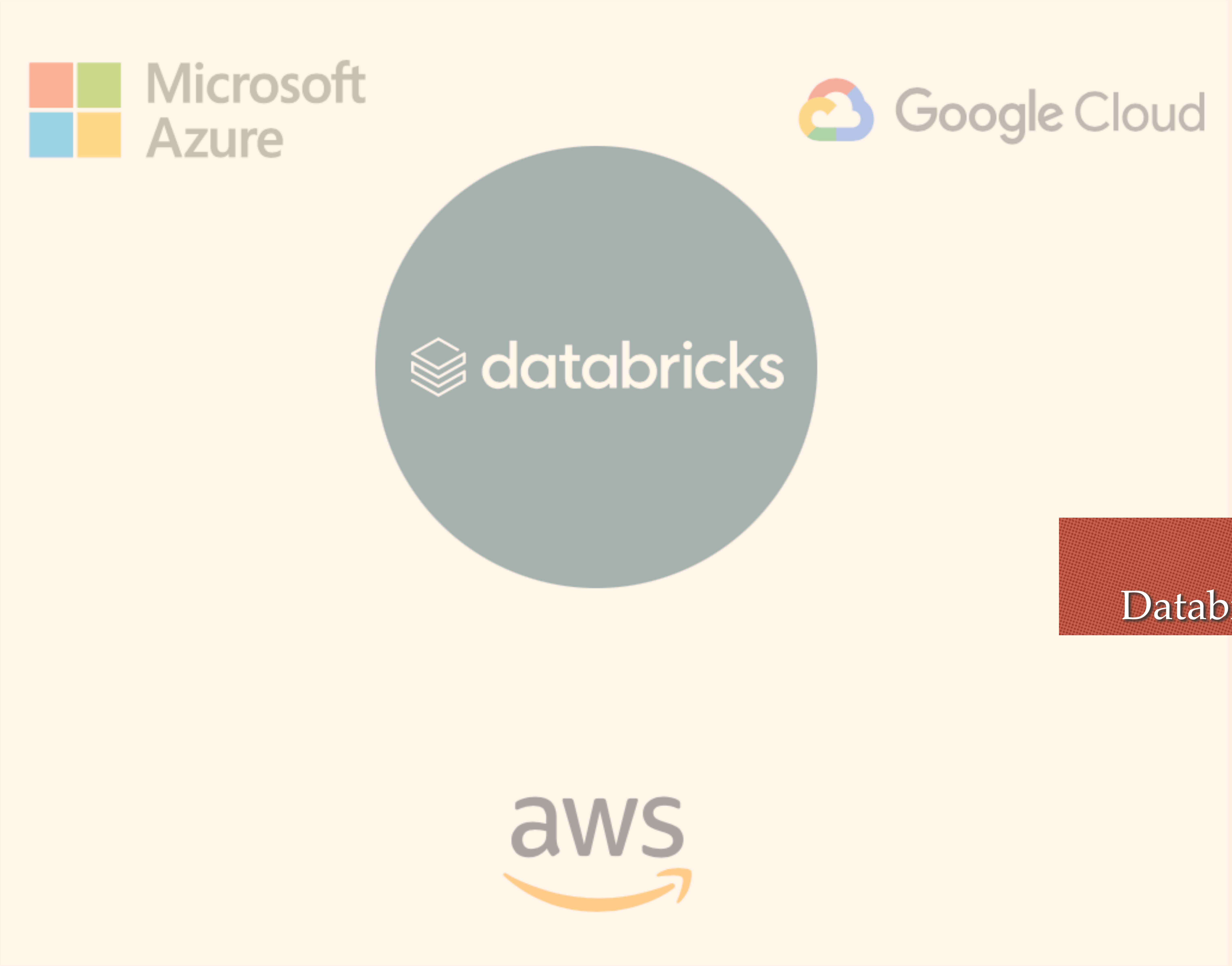
Objects (6)

Objects are the fundamental entities 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**

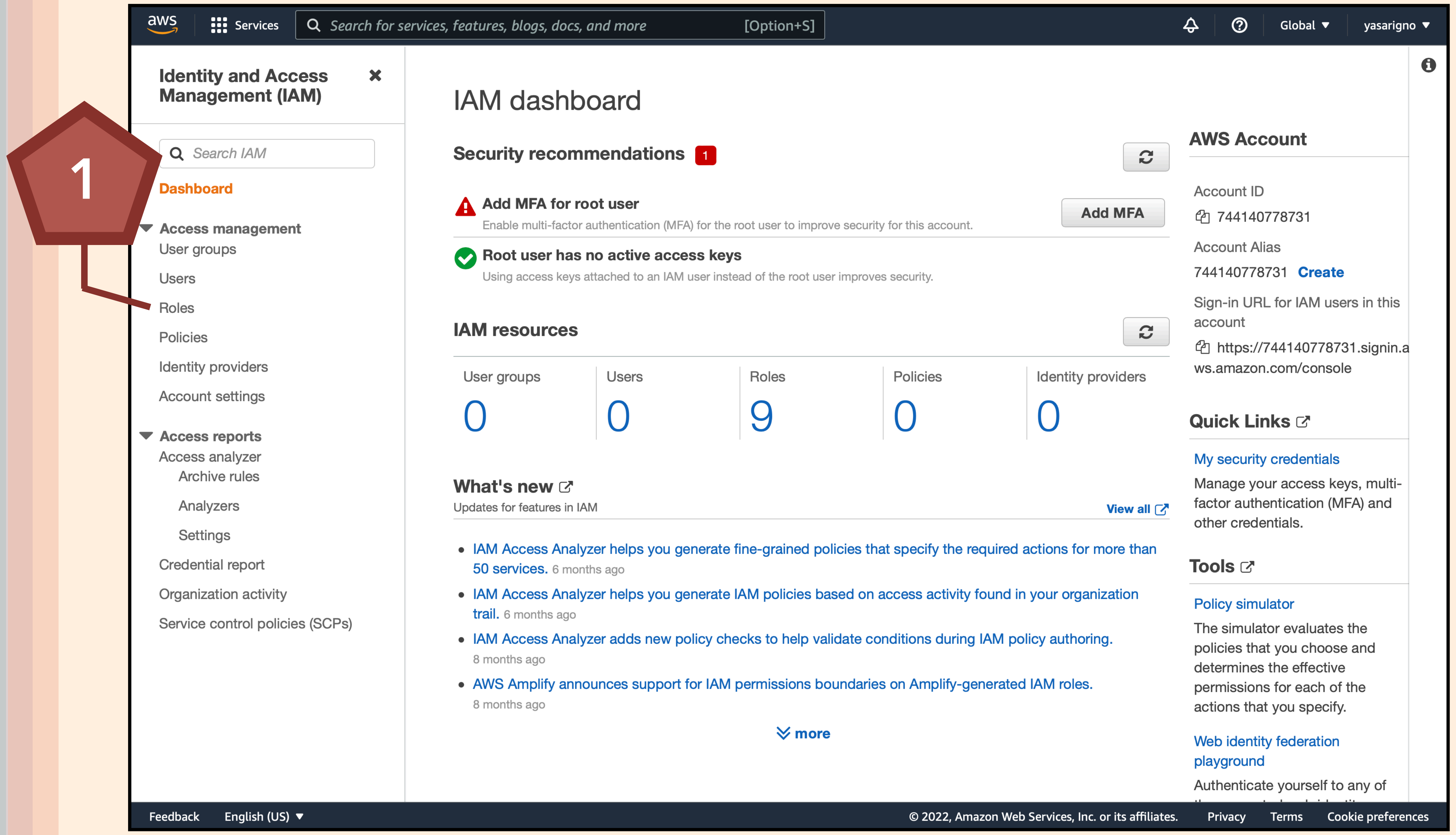
□	Name	Type	Last modified	Size	Storage class
□	data_fruits/	Folder	-	-	-
□	data_pandas_pca.csv	csv	February 20, 2022, 13:35:59 (UTC+01:00)	2.4 MB	Standard
□	data/	Folder	-	-	-
□	frankfurt-prod/	Folder	-	-	-
□	fruits/	Folder	-		
□	try_csv/	Folder	-		

Le fichier data_pandas_pca.csv

Merci de votre attention



Annexe I
Databricks - AWS : Credential configuration



1
aws
Services
Search for services, features, blogs, docs, and more [Option+S]
Global ▼
yasarigno ▼

Identity and Access Management (IAM)

Search IAM

Dashboard

- Access management
  - User groups
  - Users
  - Roles
  - Policies
  - Identity providers
  - Account settings
- Access reports
  - Access analyzer
    - Archive rules
    - Analyzers
    - Settings
  - Credential report
  - Organization activity
  - Service control policies (SCPs)

IAM dashboard

Security recommendations 1

Add MFA for root user — Enable multi-factor authentication (MFA) for the root user to improve security for this account. [Add MFA]

Root user has no active access keys — Using access keys attached to an IAM user instead of the root user improves security.

IAM resources

User groups	Users	Roles	Policies	Identity providers
0	0	9	0	0

What's new

Updates for features in IAM — View all

- IAM Access Analyzer helps you generate fine-grained policies that specify the required actions for more than 50 services. 6 months ago
- IAM Access Analyzer helps you generate IAM policies based on access activity found in your organization trail. 6 months ago
- IAM Access Analyzer adds new policy checks to help validate conditions during IAM policy authoring. 8 months ago
- AWS Amplify announces support for IAM permissions boundaries on Amplify-generated IAM roles. 8 months ago

more

AWS Account

Account ID
744140778731

Account Alias
744140778731 Create

Sign-in URL for IAM users in this account
https://744140778731.signin.aws.amazon.com/console

Quick Links

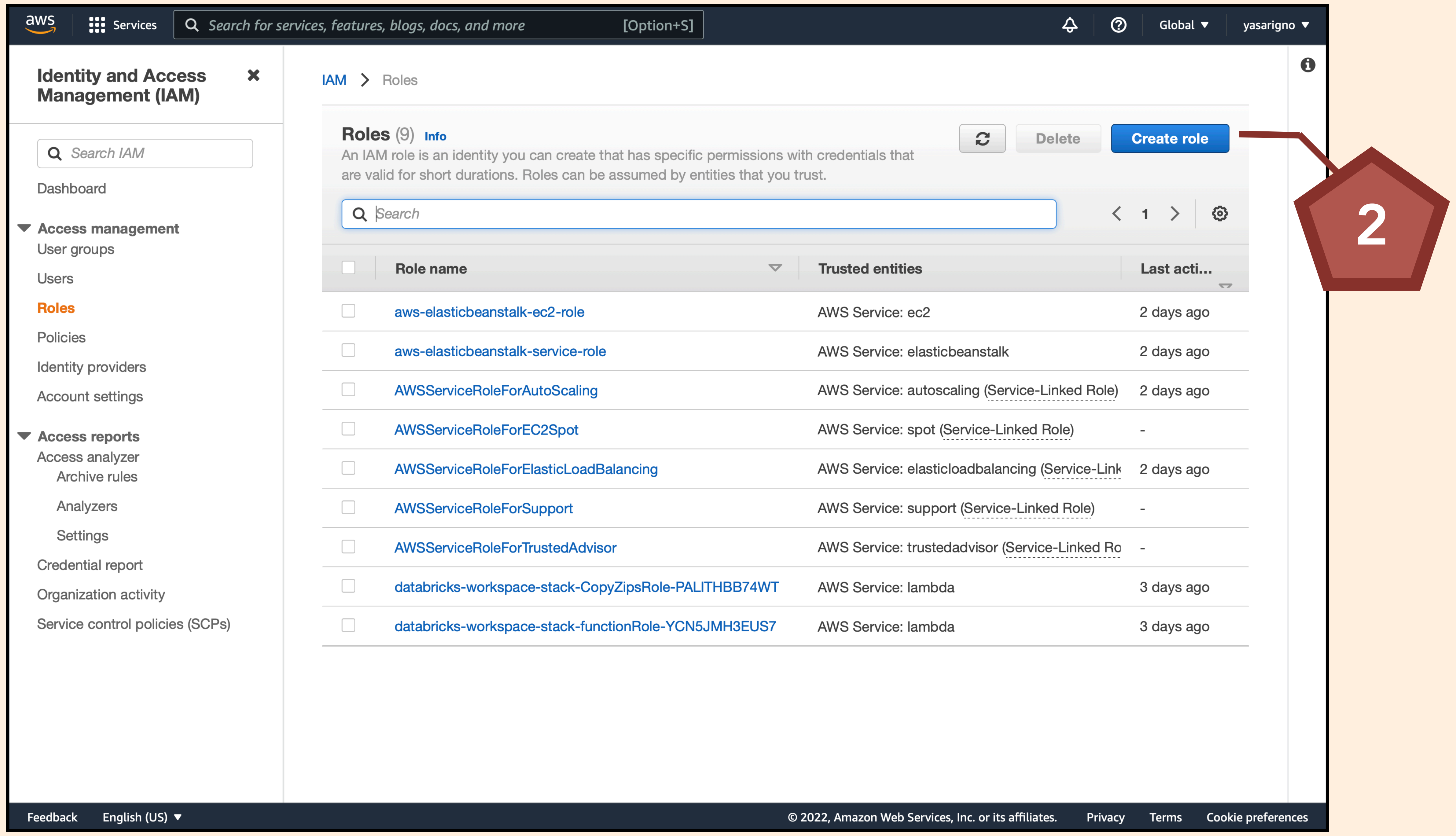
My security credentials
Manage your access keys, multi-factor authentication (MFA) and other credentials.

Tools

Policy simulator
The simulator evaluates the policies that you choose and determines the effective permissions for each of the actions that you specify.

Web identity federation playground
Authenticate yourself to any of

1. IAM Dashboard → Access management → Roles



2
aws
Services
Search for services, features, blogs, docs, and more [Option+S]
Global ▼
yasarigno ▼

Identity and Access Management (IAM)

Search IAM

Dashboard

- Access management
  - User groups
  - Users
  - Roles
  - Policies
  - Identity providers
  - Account settings
- Access reports
  - Access analyzer
    - Archive rules
    - Analyzers
    - Settings
  - Credential report
  - Organization activity
  - Service control policies (SCPs)

IAM > Roles

Roles (9) Info [Delete] [Create role]

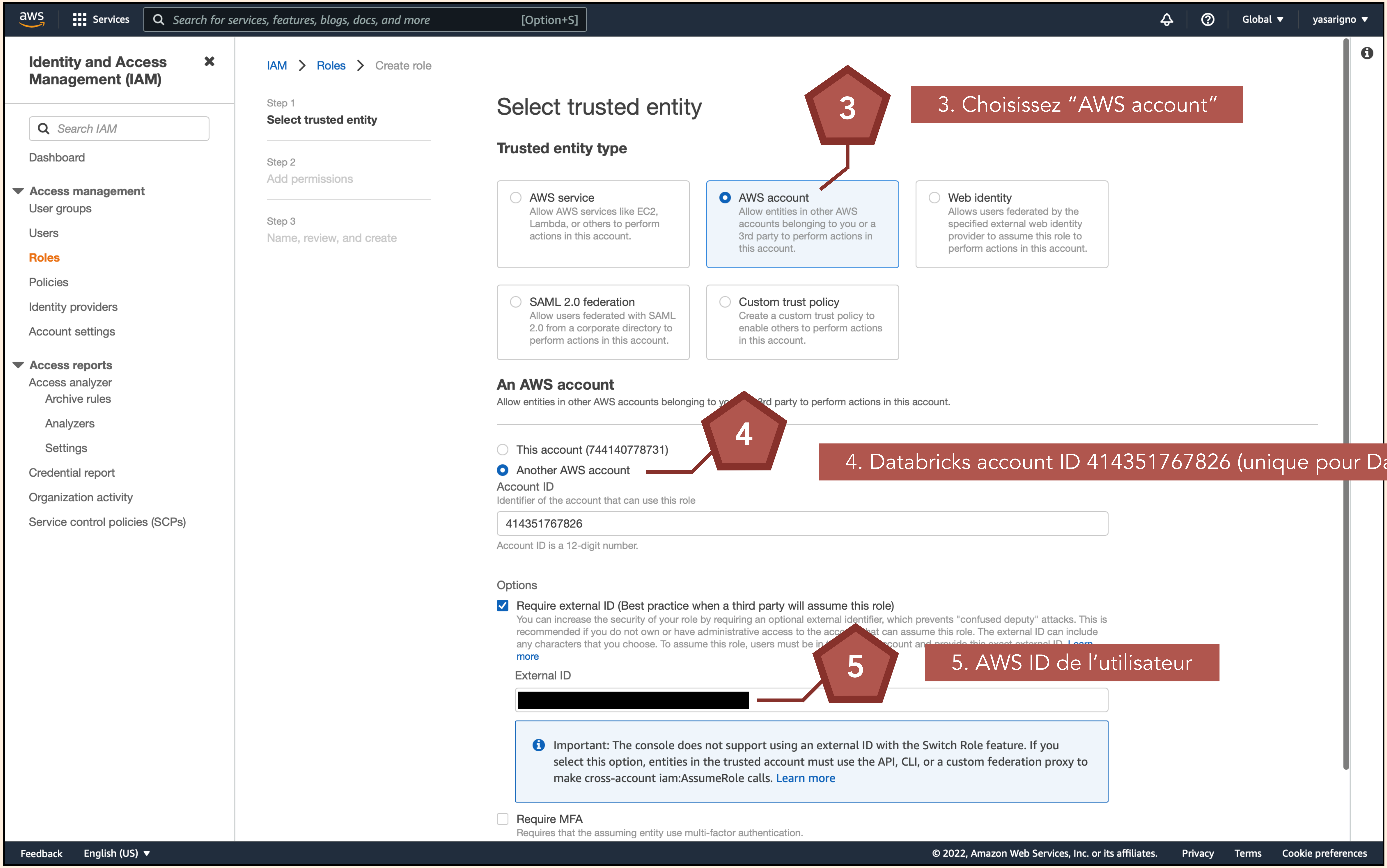
An IAM role is an identity you can create that has specific permissions with credentials that are valid for short durations. Roles can be assumed by entities that you trust.

Search

□	Role name	Trusted entities	Last acti...
□	aws-elasticbeanstalk-ec2-role	AWS Service: ec2	2 days ago
□	aws-elasticbeanstalk-service-role	AWS Service: elasticbeanstalk	2 days ago
□	AWSServiceRoleForAutoScaling	AWS Service: autoscaling (Service-Linked Role)	2 days ago
□	AWSServiceRoleForEC2Spot	AWS Service: spot (Service-Linked Role)	-
□	AWSServiceRoleForElasticLoadBalancing	AWS Service: elasticloadbalancing (Service-Link	2 days ago
□	AWSServiceRoleForSupport	AWS Service: support (Service-Linked Role)	-
□	AWSServiceRoleForTrustedAdvisor	AWS Service: trustedadvisor (Service-Linked Rc	-
□	databricks-workspace-stack-CopyZipsRole-PALITHBB74WT	AWS Service: lambda	3 days ago
□	databricks-workspace-stack-functionRole-YCN5JMH3EUS7	AWS Service: lambda	3 days ago

2. Create role

https://docs.databricks.com/security/credential-passthrough/iam-passthrough.html



Select trusted entity

Trusted entity type

○ AWS service
Allow AWS services like EC2, Lambda, or others to perform actions in this account.

● AWS account
Allow entities in other AWS accounts belonging to you or a 3rd party to perform actions in this account.

○ Web identity
Allows users federated by the specified external web identity provider to assume this role to perform actions in this account.

○ SAML 2.0 federation
Allow users federated with SAML 2.0 from a corporate directory to perform actions in this account.

○ Custom trust policy
Create a custom trust policy to enable others to perform actions in this account.

An AWS account

Allow entities in other AWS accounts belonging to yo 3rd party to perform actions in this account.

○ This account (744140778731)

● Another AWS account

Account ID
Identifier of the account that can use this role

414351767826

Account ID is a 12-digit number.

Options

☑ Require external ID (Best practice when a third party will assume this role)
You can increase the security of your role by requiring an optional external identifier, which prevents "confused deputy" attacks. This is recommended if you do not own or have administrative access to the acc hat can assume this role. The external ID can include any characters that you choose. To assume this role, users must be in t ccount and provide this exact external ID. Learn more

External ID

[redacted]

Important: The console does not support using an external ID with the Switch Role feature. If you select this option, entities in the trusted account must use the API, CLI, or a custom federation proxy to make cross-account iam:AssumeRole calls. Learn more

☐ Require MFA
Requires that the assuming entity use multi-factor authentication.

3

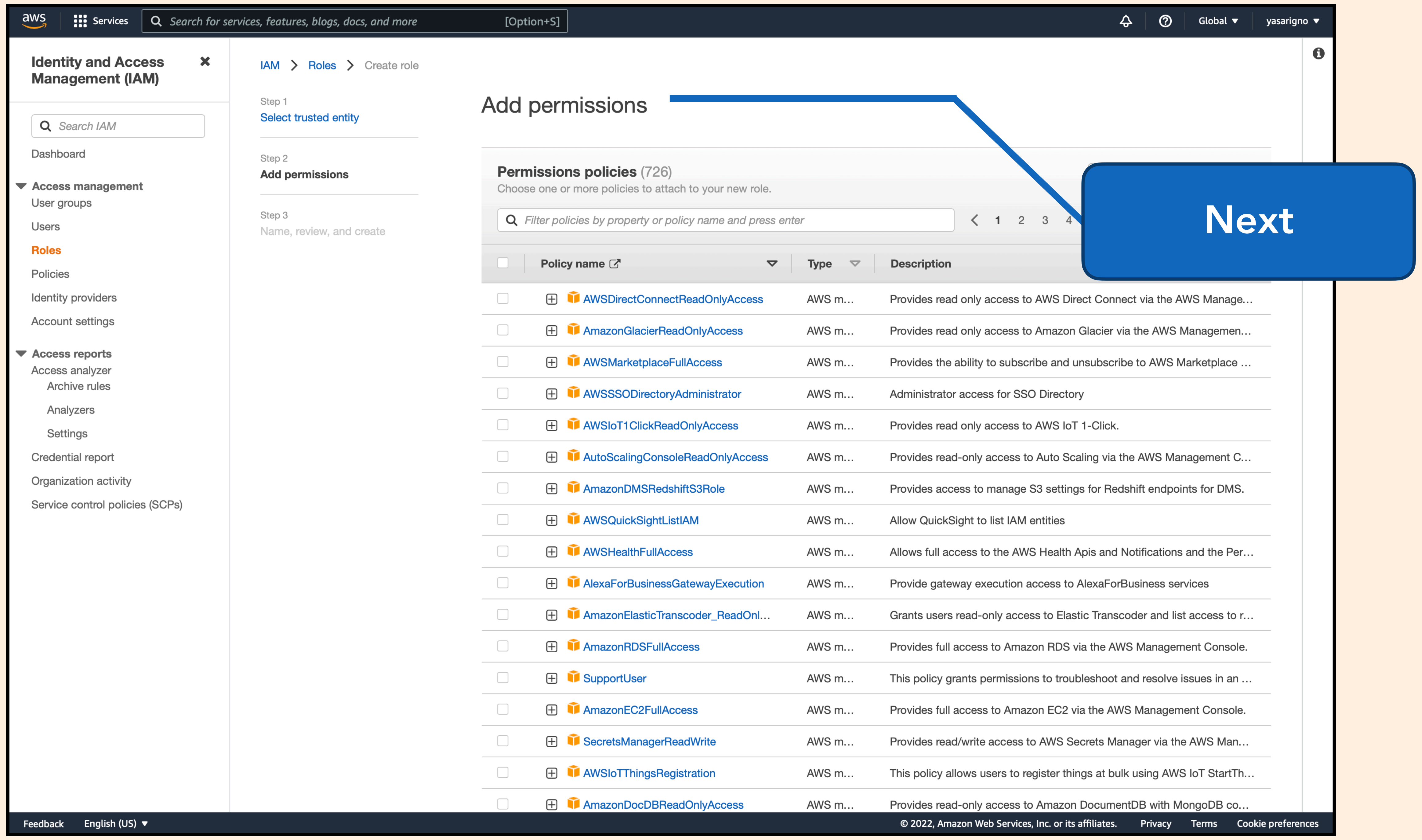
3. Choisissez “AWS account”

4

4. Databricks account ID 414351767826 (unique pour Databricks)

5

5. AWS ID de l’utilisateur



Add permissions

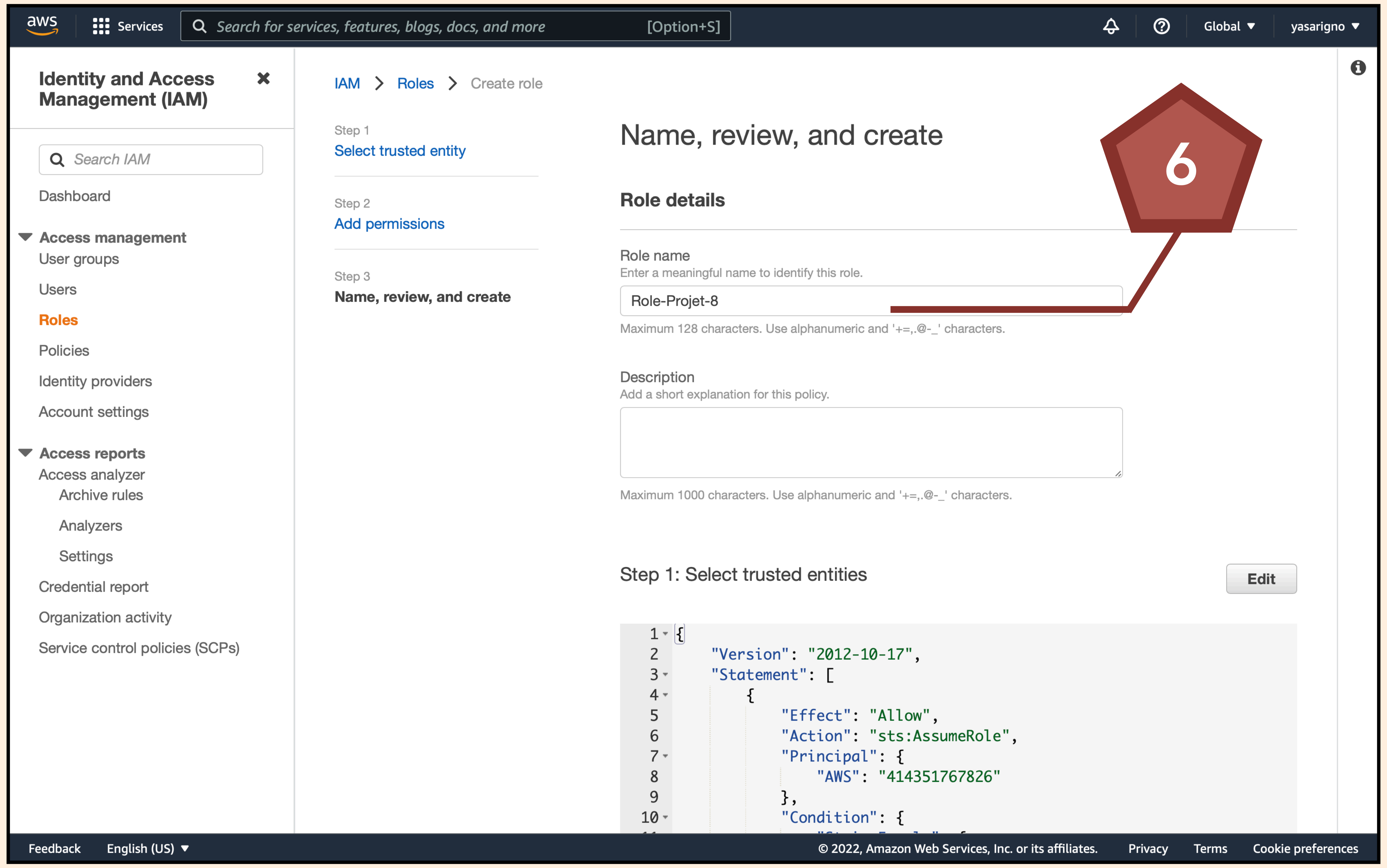
Next

Permissions policies (726)

Choose one or more policies to attach to your new role.

Filter policies by property or policy name and press enter

□	Policy name	Type	Description
□	AWSDirectConnectReadOnlyAccess	AWS m...	Provides read only access to AWS Direct Connect via the AWS Manage...
□	AmazonGlacierReadOnlyAccess	AWS m...	Provides read only access to Amazon Glacier via the AWS Managemen...
□	AWSMarketplaceFullAccess	AWS m...	Provides the ability to subscribe and unsubscribe to AWS Marketplace ...
□	AWSSSODirectoryAdministrator	AWS m...	Administrator access for SSO Directory
□	AWSIoT1ClickReadOnlyAccess	AWS m...	Provides read only access to AWS IoT 1-Click.
□	AutoScalingConsoleReadOnlyAccess	AWS m...	Provides read-only access to Auto Scaling via the AWS Management C...
□	AmazonDMSRedshiftS3Role	AWS m...	Provides access to manage S3 settings for Redshift endpoints for DMS.
□	AWSQuickSightListIAM	AWS m...	Allow QuickSight to list IAM entities
□	AWSHealthFullAccess	AWS m...	Allows full access to the AWS Health Apis and Notifications and the Per...
□	AlexaForBusinessGatewayExecution	AWS m...	Provide gateway execution access to AlexaForBusiness services
□	AmazonElasticTranscoder_ReadOnl...	AWS m...	Grants users read-only access to Elastic Transcoder and list access to r...
□	AmazonRDSFullAccess	AWS m...	Provides full access to Amazon RDS via the AWS Management Console.
□	SupportUser	AWS m...	This policy grants permissions to troubleshoot and resolve issues in an ...
□	AmazonEC2FullAccess	AWS m...	Provides full access to Amazon EC2 via the AWS Management Console.
□	SecretsManagerReadWrite	AWS m...	Provides read/write access to AWS Secrets Manager via the AWS Man...
□	AWSIoTThingsRegistration	AWS m...	This policy allows users to register things at bulk using AWS IoT StartTh...
□	AmazonDocDBReadOnlyAccess	AWS m...	Provides read-only access to Amazon DocumentDB with MongoDB co...



6

Name, review, and create

Role details

Role name

Enter a meaningful name to identify this role.

Role-Projet-8

Maximum 128 characters. Use alphanumeric and '+=,.@-_' characters.

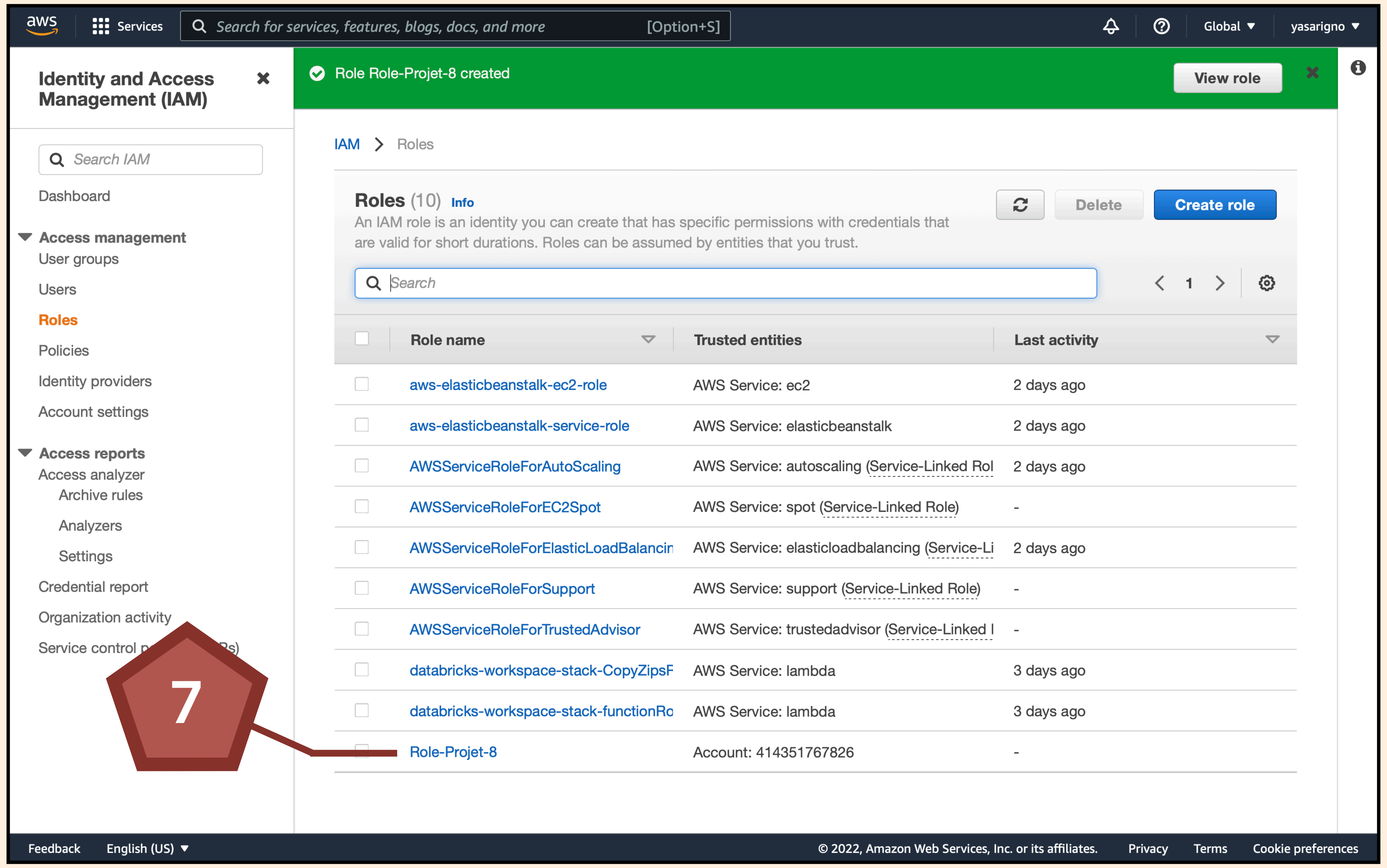
Description

Add a short explanation for this policy.

Maximum 1000 characters. Use alphanumeric and '+=,.@-_' characters.

Step 1: Select trusted entities

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "sts:AssumeRole",
            "Principal": {
                "AWS": "414351767826"
            },
            "Condition": {
```



Role Role-Projet-8 created

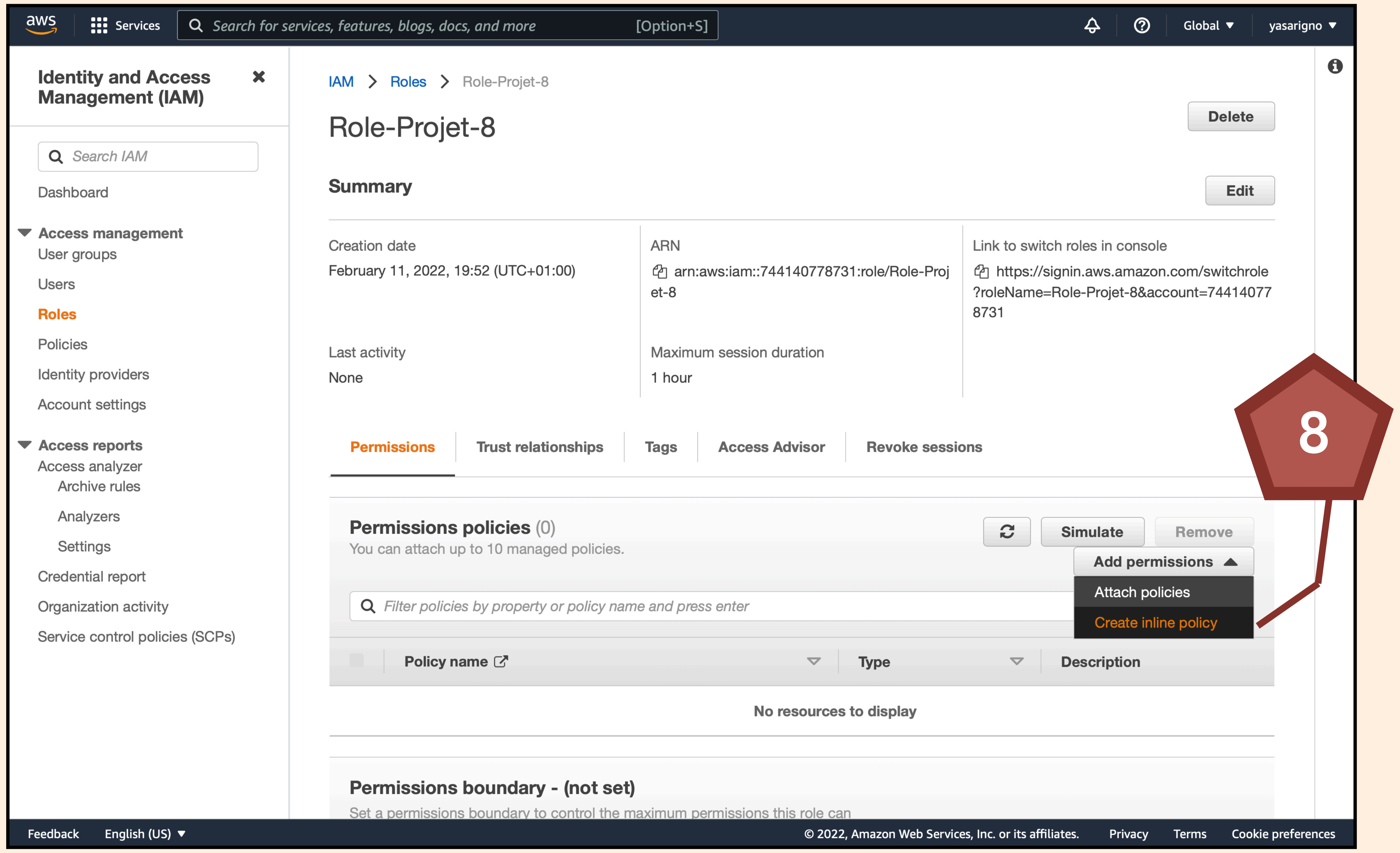
7

Roles (10)

An IAM role is an identity you can create that has specific permissions with credentials that are valid for short durations. Roles can be assumed by entities that you trust.

□	Role name	Trusted entities	Last activity
□	aws-elasticbeanstalk-ec2-role	AWS Service: ec2	2 days ago
□	aws-elasticbeanstalk-service-role	AWS Service: elasticbeanstalk	2 days ago
□	AWSServiceRoleForAutoScaling	AWS Service: autoscaling (Service-Linked Rol	2 days ago
□	AWSServiceRoleForEC2Spot	AWS Service: spot (Service-Linked Role)	-
□	AWSServiceRoleForElasticLoadBalancir	AWS Service: elasticloadbalancing (Service-Li	2 days ago
□	AWSServiceRoleForSupport	AWS Service: support (Service-Linked Role)	-
□	AWSServiceRoleForTrustedAdvisor	AWS Service: trustedadvisor (Service-Linked I	-
□	databricks-workspace-stack-CopyZipsF	AWS Service: lambda	3 days ago
□	databricks-workspace-stack-functionRc	AWS Service: lambda	3 days ago
□	Role-Projet-8	Account: 414351767826	-

Databricks VPC



Role-Projet-8

Summary

Creation date	ARN	Link to switch roles in console
February 11, 2022, 19:52 (UTC+01:00)	arn:aws:iam::744140778731:role/Role-Projet-8	https://signin.aws.amazon.com/switchrole?roleName=Role-Projet-8&account=744140778731
Last activity	**Maximum session duration**	
None	1 hour	

Permissions policies (0)

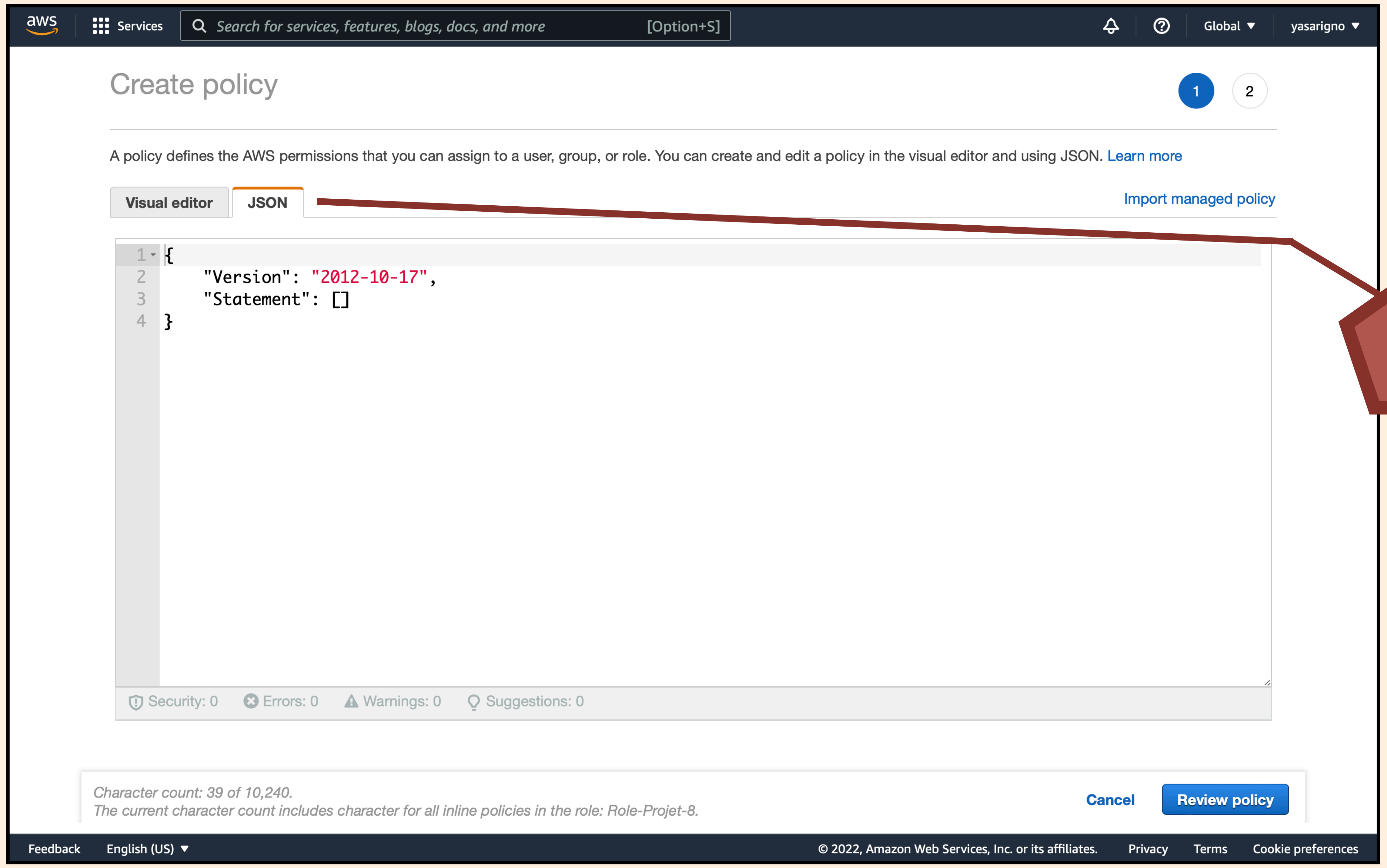
You can attach up to 10 managed policies.

No resources to display

Permissions boundary - (not set)

Set a permissions boundary to control the maximum permissions this role can

8



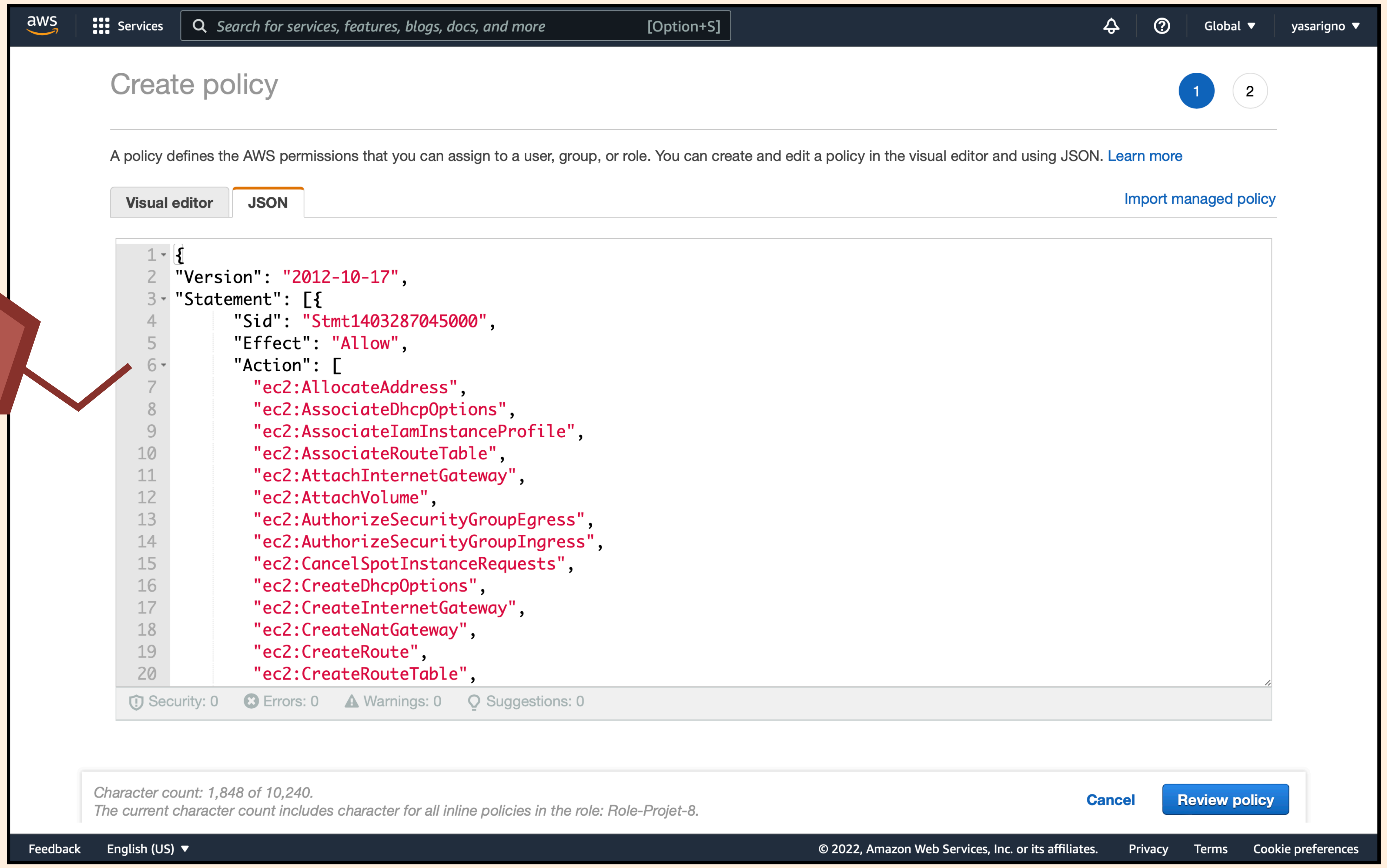
Create policy

A policy defines the AWS permissions that you can assign to a user, group, or role. You can create and edit a policy in the visual editor and using JSON. Learn more

```
{
    "Version": "2012-10-17",
    "Statement": []
}
```

Character count: 39 of 10,240.
The current character count includes character for all inline policies in the role: Role-Projet-8.

9

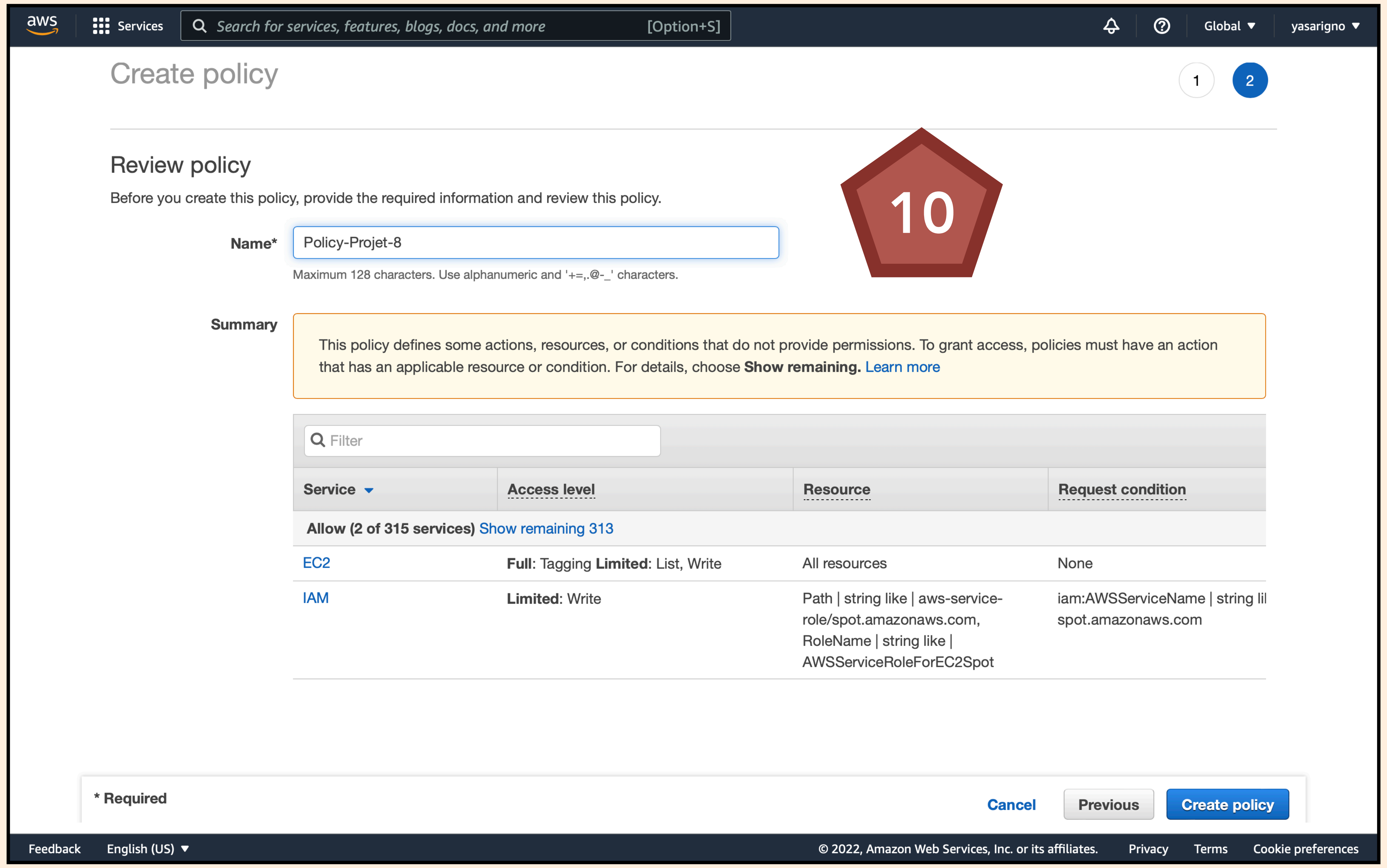


Create policy

A policy defines the AWS permissions that you can assign to a user, group, or role. You can create and edit a policy in the visual editor and using JSON. Learn more

```
{
"Version": "2012-10-17",
"Statement": [{
      "Sid": "Stmt1403287045000",
      "Effect": "Allow",
      "Action": [
        "ec2:AllocateAddress",
        "ec2:AssociateDhcpOptions",
        "ec2:AssociateIamInstanceProfile",
        "ec2:AssociateRouteTable",
        "ec2:AttachInternetGateway",
        "ec2:AttachVolume",
        "ec2:AuthorizeSecurityGroupEgress",
        "ec2:AuthorizeSecurityGroupIngress",
        "ec2:CancelSpotInstanceRequests",
        "ec2:CreateDhcpOptions",
        "ec2:CreateInternetGateway",
        "ec2:CreateNatGateway",
        "ec2:CreateRoute",
        "ec2:CreateRouteTable",
```

Character count: 1,848 of 10,240.
The current character count includes character for all inline policies in the role: Role-Projet-8.



Create policy

Review policy

Before you create this policy, provide the required information and review this policy.

Name* Policy-Projet-8

Maximum 128 characters. Use alphanumeric and '+=,.@-_' characters.

Summary

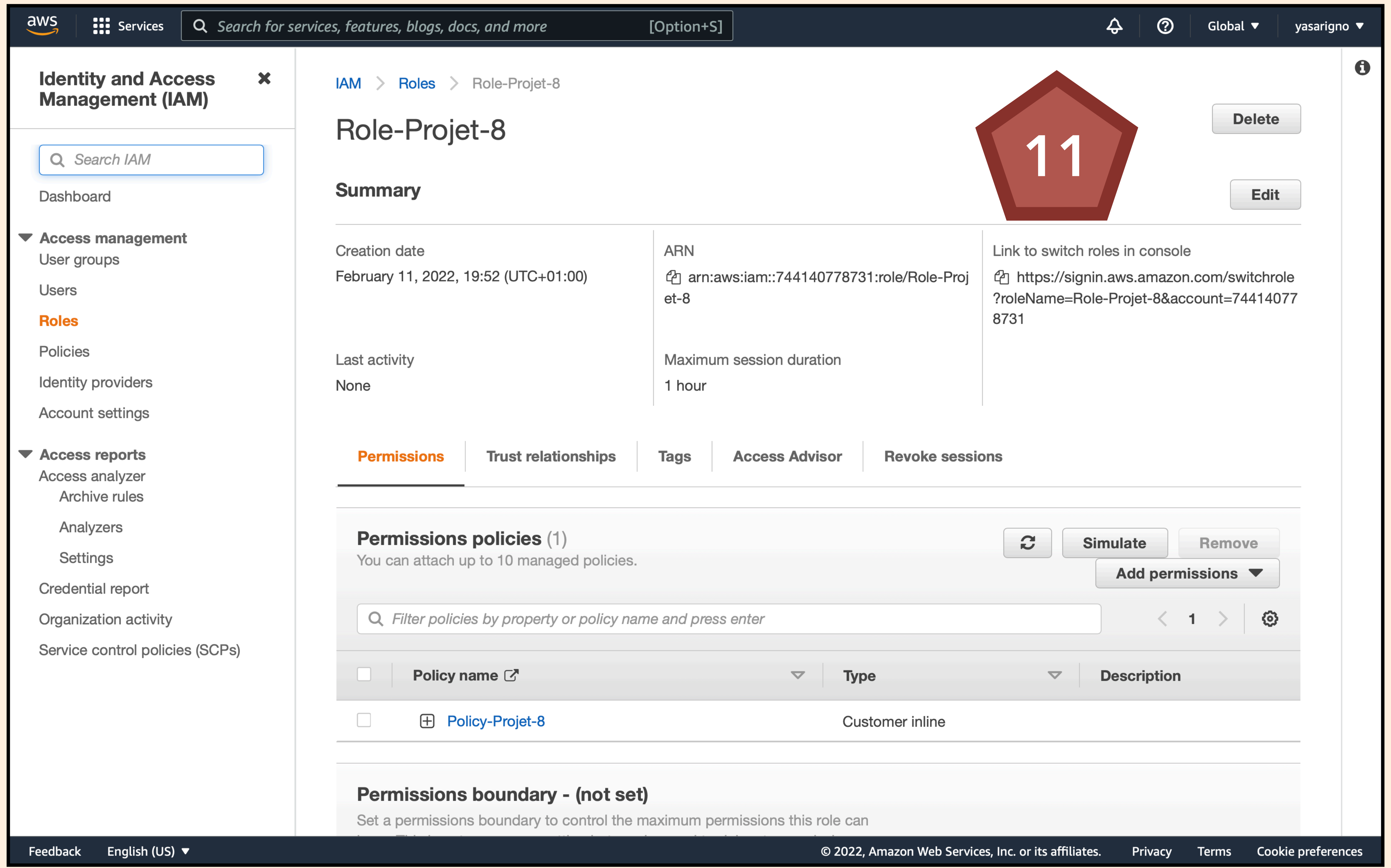
This policy defines some actions, resources, or conditions that do not provide permissions. To grant access, policies must have an action that has an applicable resource or condition. For details, choose **Show remaining.** Learn more

Filter

Service	Access level	Resource	Request condition
Allow (2 of 315 services) Show remaining 313			
EC2	**Full**: Tagging **Limited**: List, Write	All resources	None
IAM	**Limited**: Write	Path | string like | aws-service-role/spot.amazonaws.com, RoleName | string like | AWSServiceRoleForEC2Spot	iam:AWSServiceName | string lil spot.amazonaws.com

* Required

© 2022, Amazon Web Services, Inc. or its affiliates.



IAM > Roles > Role-Projet-8

Role-Projet-8

Summary

Creation date
February 11, 2022, 19:52 (UTC+01:00)

ARN
arn:aws:iam::744140778731:role/Role-Projet-8

Link to switch roles in console
https://signin.aws.amazon.com/switchrole?roleName=Role-Projet-8&account=744140778731

Last activity
None

Maximum session duration
1 hour

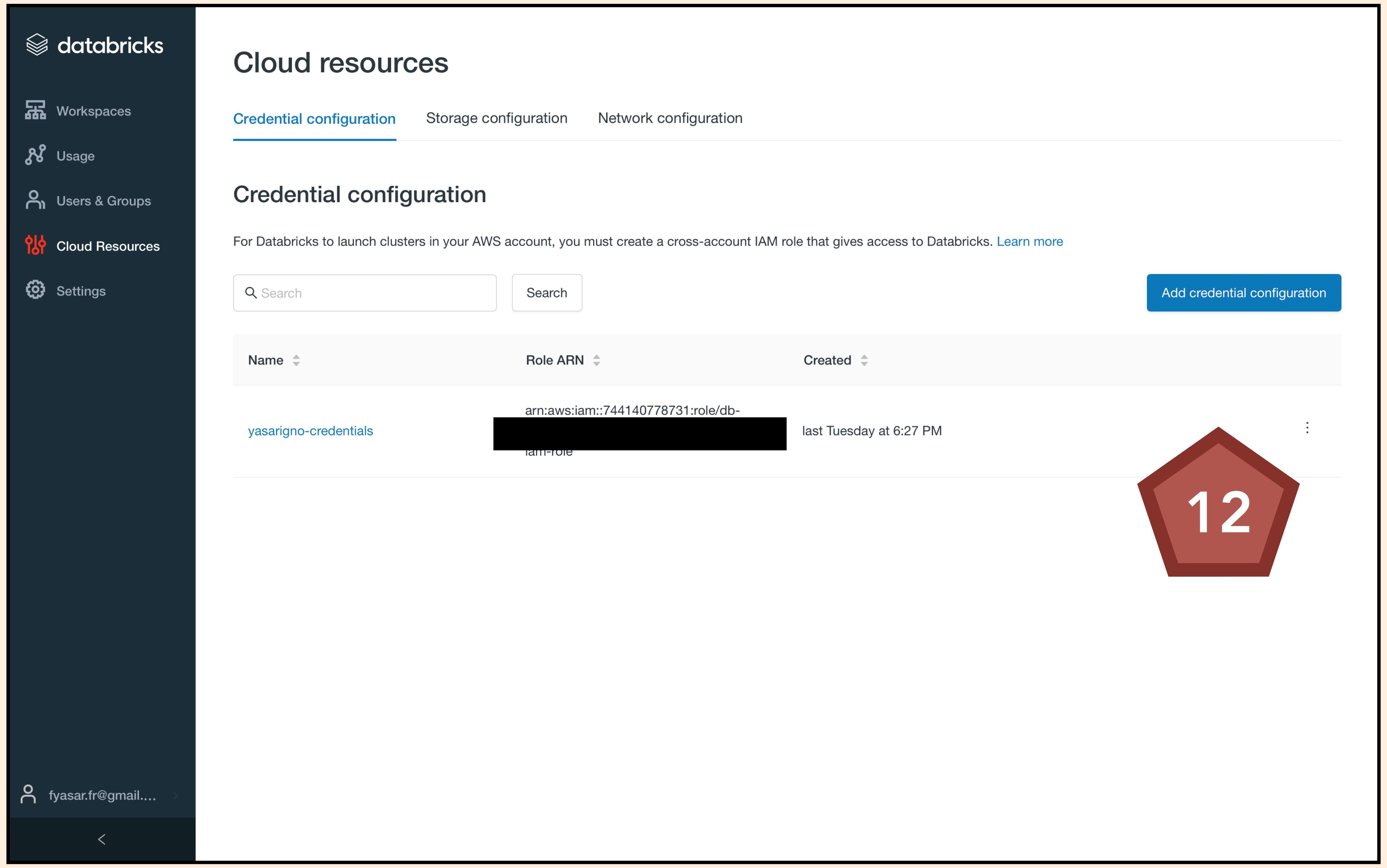
Permissions policies (1)
You can attach up to 10 managed policies.

Filter policies by property or policy name and press enter

Policy name	Type	Description
Policy-Projet-8	Customer inline	

Permissions boundary - (not set)
Set a permissions boundary to control the maximum permissions this role can

© 2022, Amazon Web Services, Inc. or its affiliates.



Cloud resources

Credential configuration | Storage configuration | Network configuration

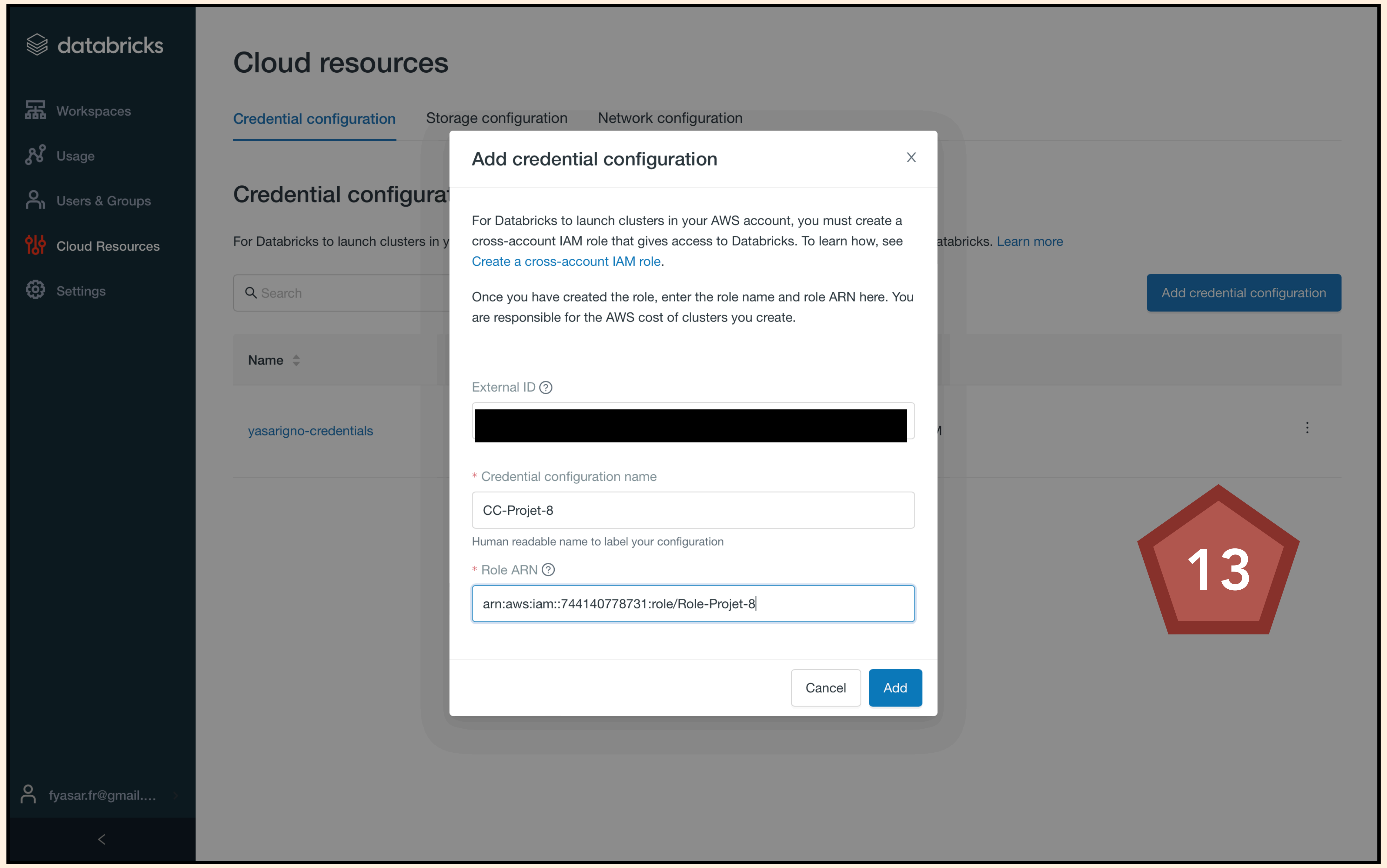
Credential configuration

For Databricks to launch clusters in your AWS account, you must create a cross-account IAM role that gives access to Databricks. Learn more

Search | Add credential configuration

Name	Role ARN	Created
yasarigno-credentials	arn:aws:iam::744140778731:role/db- [redacted]	last Tuesday at 6:27 PM

12



Add credential configuration

For Databricks to launch clusters in your AWS account, you must create a cross-account IAM role that gives access to Databricks. To learn how, see Create a cross-account IAM role.

Once you have created the role, enter the role name and role ARN here. You are responsible for the AWS cost of clusters you create.

External ID: [redacted]

* Credential configuration name: CC-Projet-8

Human readable name to label your configuration

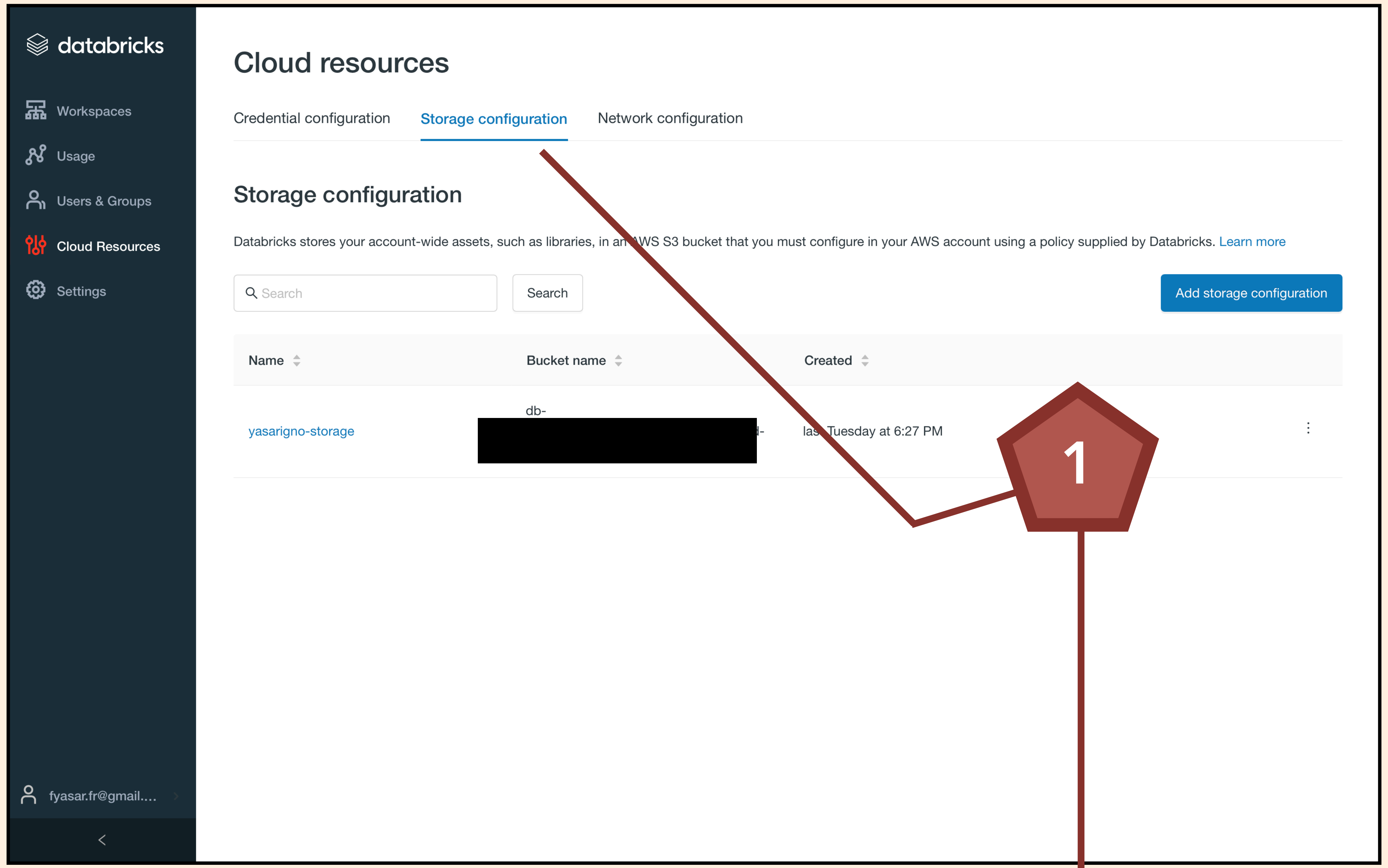
* Role ARN: arn:aws:iam::744140778731:role/Role-Projet-8

Cancel | Add

13

Annexe II

Databricks - AWS : Storage configuration



Cloud resources

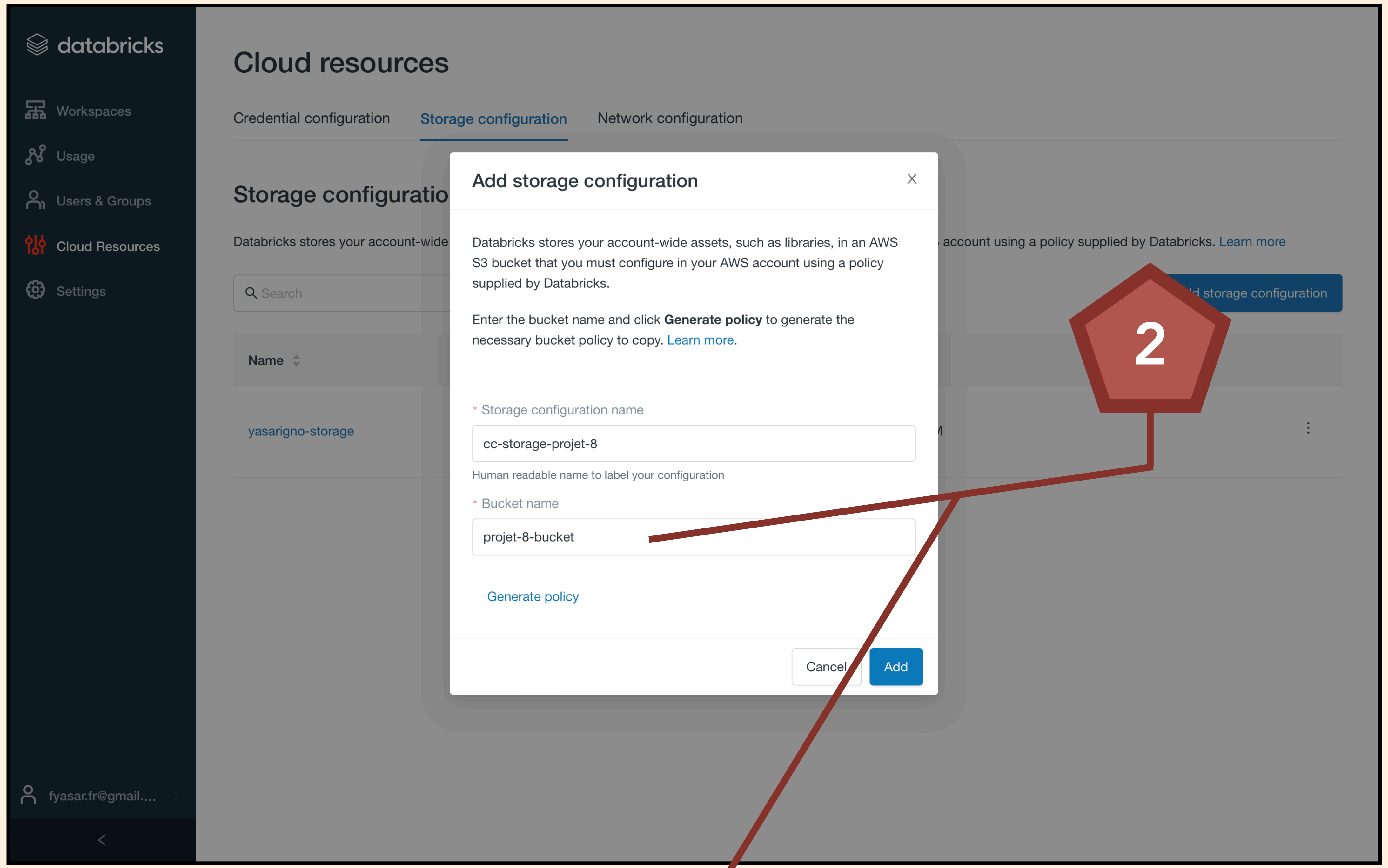
Credential configuration | Storage configuration | Network configuration

Storage configuration

Databricks stores your account-wide assets, such as libraries, in an AWS S3 bucket that you must configure in your AWS account using a policy supplied by Databricks. Learn more

Search | Search | Add storage configuration

Name	Bucket name	Created
yasarigno-storage	db-[redacted]-	last Tuesday at 6:27 PM



Add storage configuration

Databricks stores your account-wide assets, such as libraries, in an AWS S3 bucket that you must configure in your AWS account using a policy supplied by Databricks.

Enter the bucket name and click **Generate policy** to generate the necessary bucket policy to copy. Learn more.

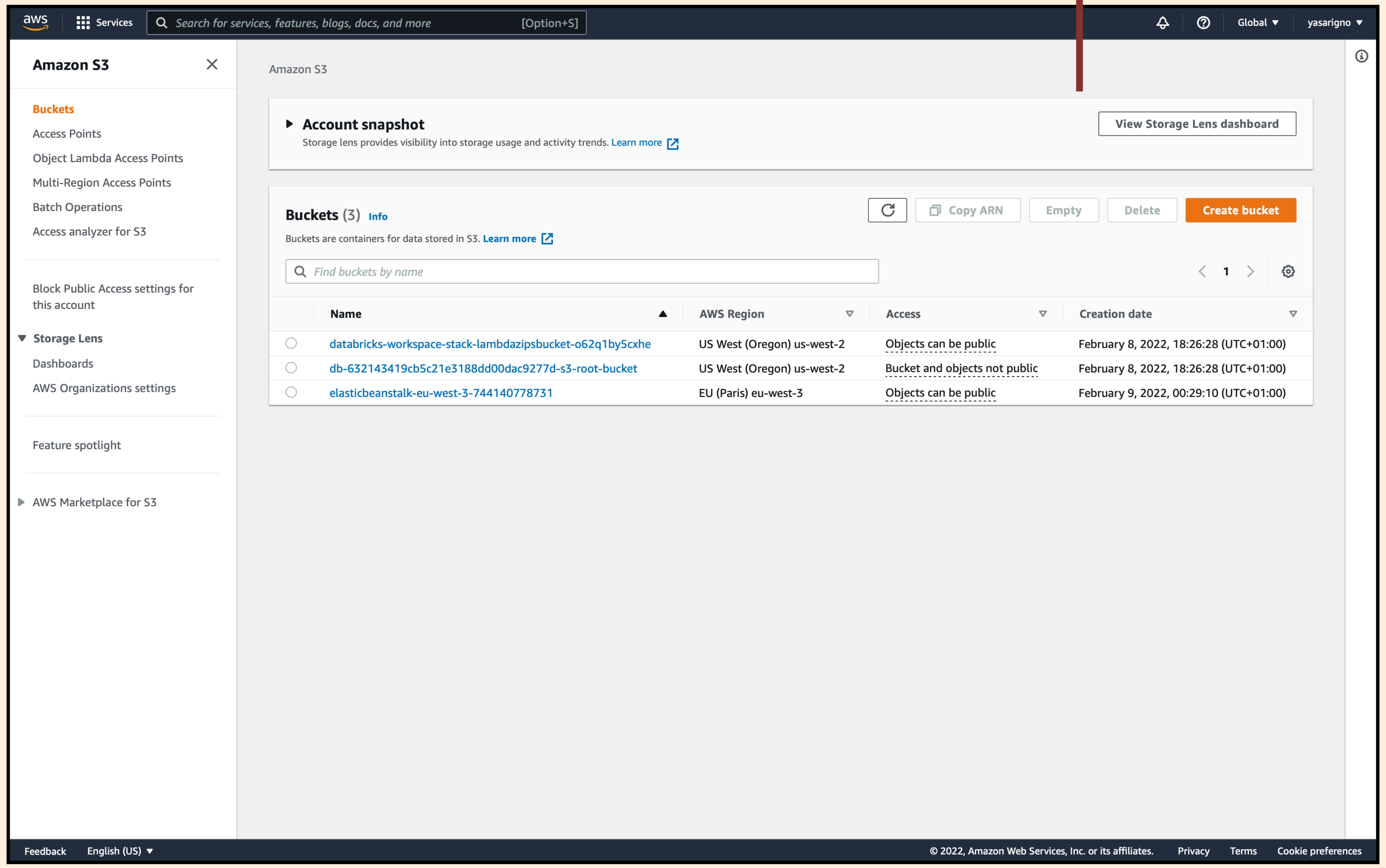
* Storage configuration name: cc-storage-projet-8

Human readable name to label your configuration

* Bucket name: projet-8-bucket

Generate policy

Cancel | Add


aws
Services
Search for services, features, blogs, docs, and more [Option+S]
Global
yasarigno
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
AWS Marketplace for S3
Feedback
English (US)
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

Amazon S3

Account snapshot

Storage lens provides visibility into storage usage and activity trends. Learn more

View Storage Lens dashboard

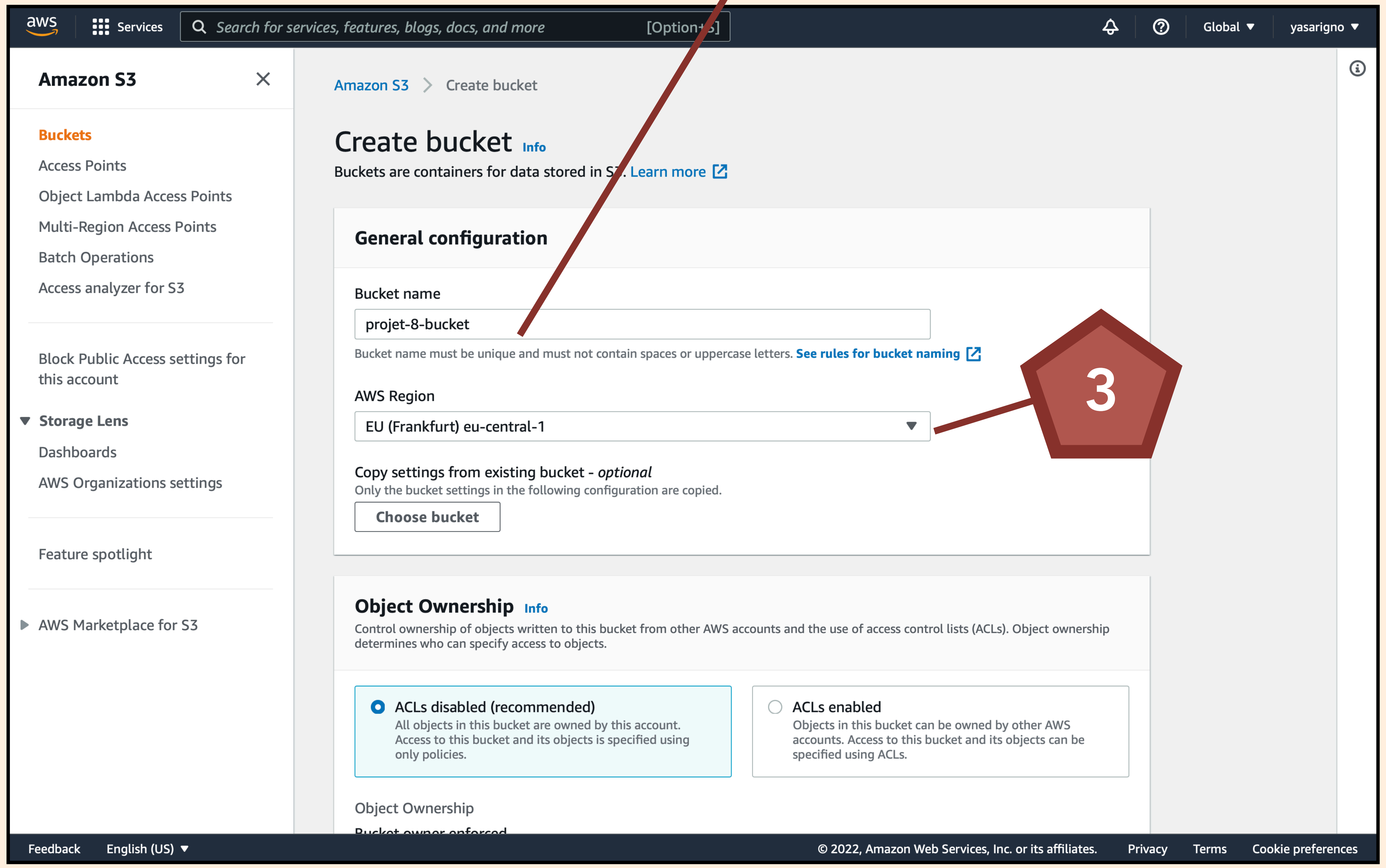
Buckets (3) Info

Copy ARN | Empty | Delete | Create bucket

Buckets are containers for data stored in S3. Learn more

Find buckets by name

Name	AWS Region	Access	Creation date
databricks-workspace-stack-lambdazipsbucket-o62q1by5cxhe	US West (Oregon) us-west-2	Objects can be public	February 8, 2022, 18:26:28 (UTC+01:00)
db-632143419cb5c21e3188dd00dac9277d-s3-root-bucket	US West (Oregon) us-west-2	Bucket and objects not public	February 8, 2022, 18:26:28 (UTC+01:00)
elasticbeanstalk-eu-west-3-744140778731	EU (Paris) eu-west-3	Objects can be public	February 9, 2022, 00:29:10 (UTC+01:00)


aws
Services
Search for services, features, blogs, docs, and more [Option+S]
Global
yasarigno
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
AWS Marketplace for S3
Feedback
English (US)
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
3

Amazon S3 > Create bucket

Create bucket Info

Buckets are containers for data stored in S3. Learn more

General configuration

Bucket name: projet-8-bucket

Bucket name must be unique and must not contain spaces or uppercase letters. See rules for bucket naming

AWS Region: EU (Frankfurt) eu-central-1

Copy settings from existing bucket - *optional*

Only the bucket settings in the following configuration are copied.

Choose bucket

Object Ownership Info

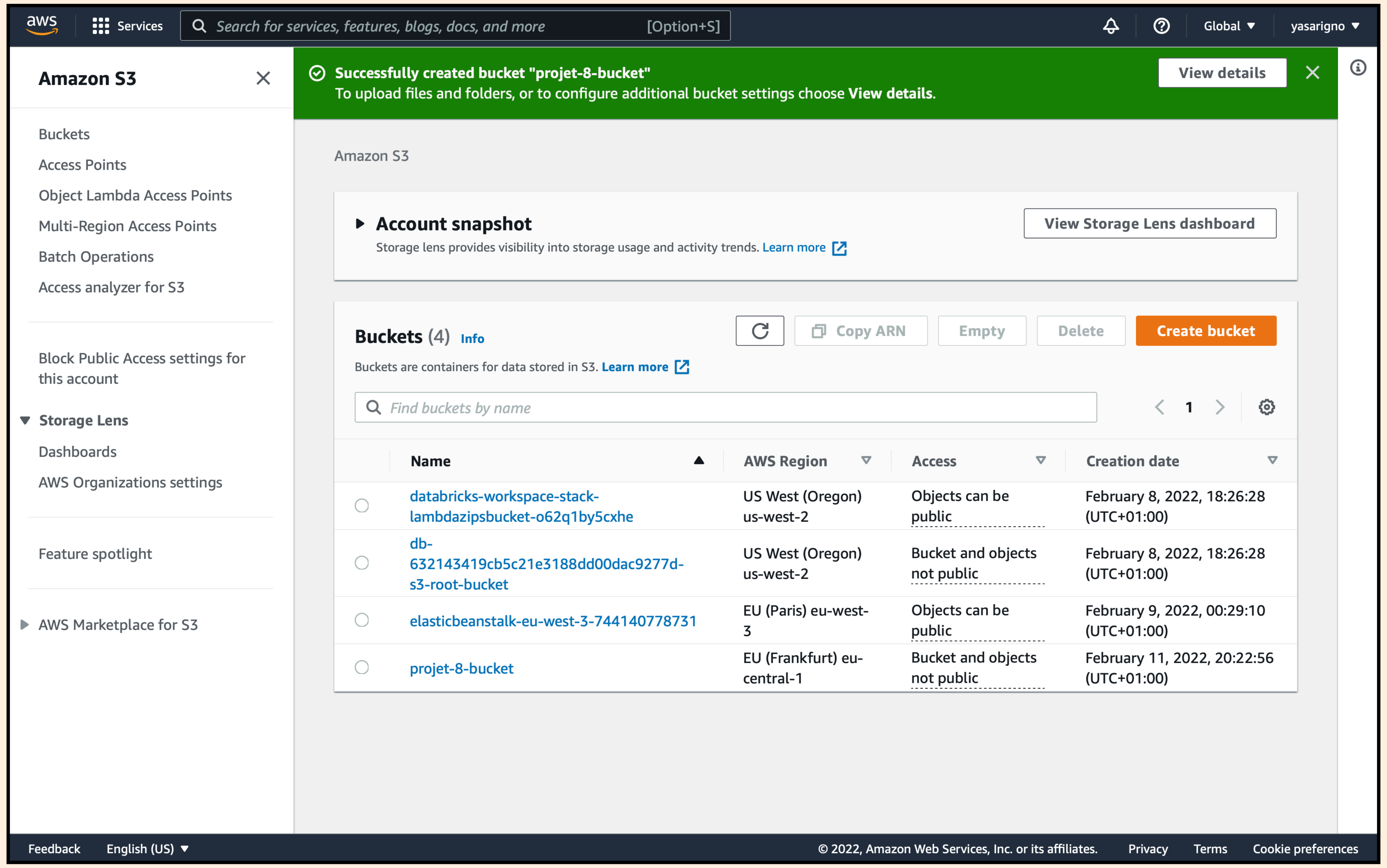
Control ownership of objects written to this bucket from other AWS accounts and the use of access control lists (ACLs). Object ownership determines who can specify access to objects.

● 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 ACLs.

Object Ownership

Bucket owner enforced


Services
Search for services, features, blogs, docs, and more [Option+S]
Global
yasarigno
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
AWS Marketplace for S3
View details
View Storage Lens dashboard
Copy ARN
Empty
Delete
Create bucket
Find buckets by name
Feedback
English (US)
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

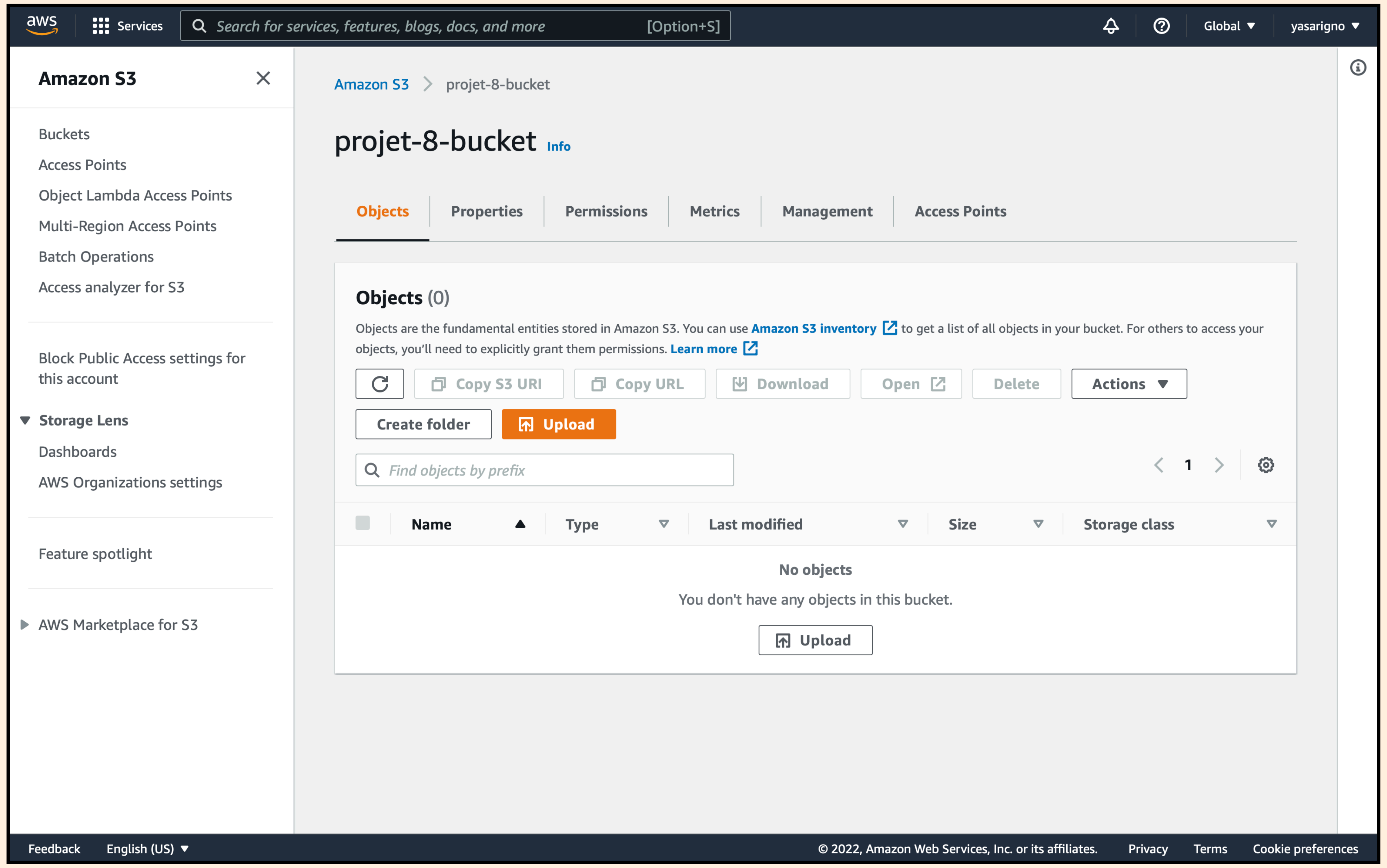
Successfully created bucket "projet-8-bucket"
To upload files and folders, or to configure additional bucket settings choose **View details**.

Amazon S3

Account snapshot
Storage lens provides visibility into storage usage and activity trends. Learn more

Buckets (4) Info
Buckets are containers for data stored in S3. Learn more

	Name	AWS Region	Access	Creation date
○	databricks-workspace-stack-lambdazipsbucket-o62q1by5cxhe	US West (Oregon) us-west-2	Objects can be public	February 8, 2022, 18:26:28 (UTC+01:00)
○	db-632143419cb5c21e3188dd00dac9277d-s3-root-bucket	US West (Oregon) us-west-2	Bucket and objects not public	February 8, 2022, 18:26:28 (UTC+01:00)
○	elasticbeanstalk-eu-west-3-744140778731	EU (Paris) eu-west-3	Objects can be public	February 9, 2022, 00:29:10 (UTC+01:00)
○	projet-8-bucket	EU (Frankfurt) eu-central-1	Bucket and objects not public	February 11, 2022, 20:22:56 (UTC+01:00)



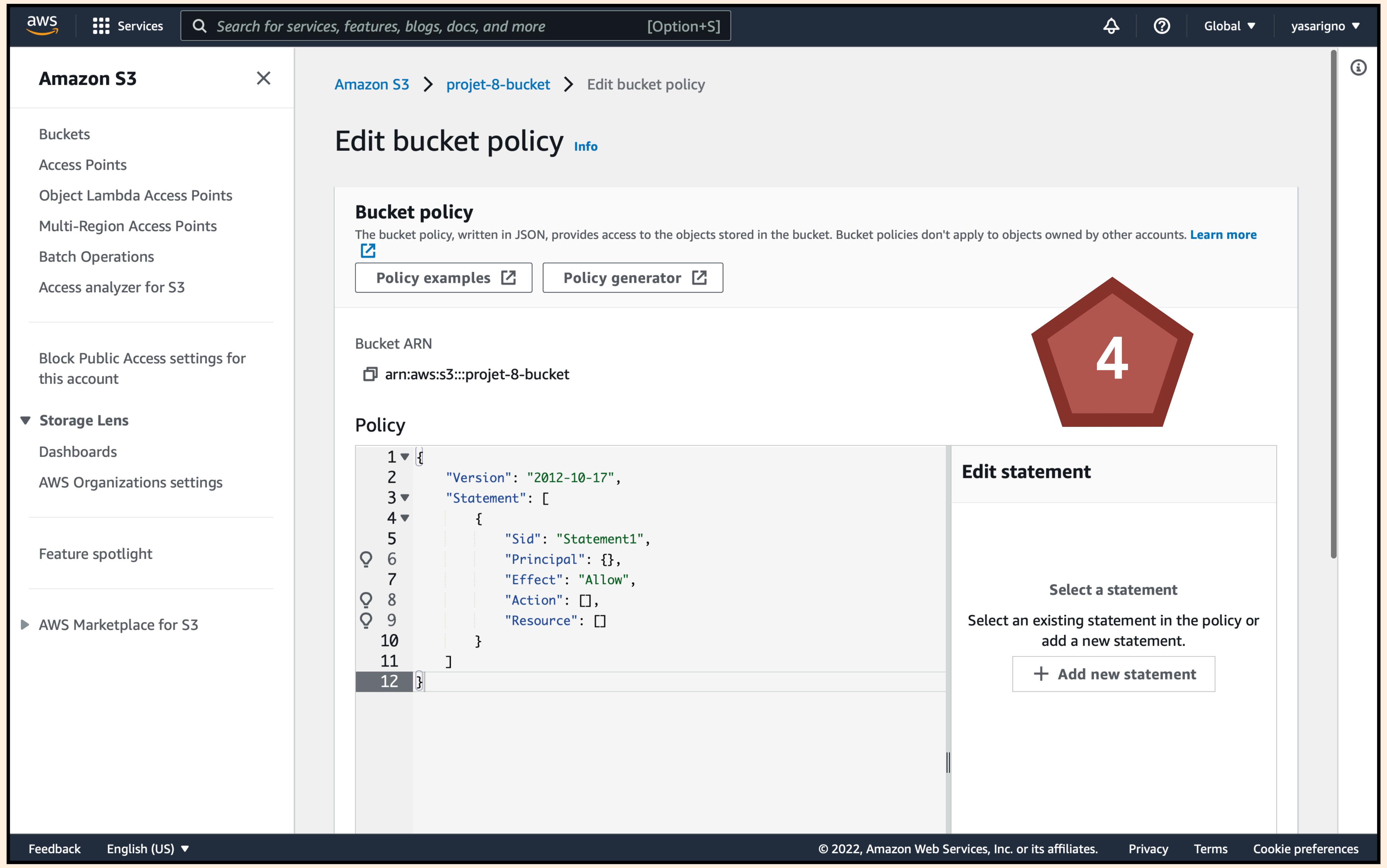
Amazon S3 > projet-8-bucket

projet-8-bucket Info

Objects (0)
Objects are the fundamental entities 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

	Name	Type	Last modified	Size	Storage class

No objects
You don't have any objects in this bucket.
Upload



Amazon S3 > projet-8-bucket > Edit bucket policy

Edit bucket policy Info

Bucket policy
The bucket policy, written in JSON, provides access to the objects stored in the bucket. Bucket policies don't apply to objects owned by other accounts. Learn more

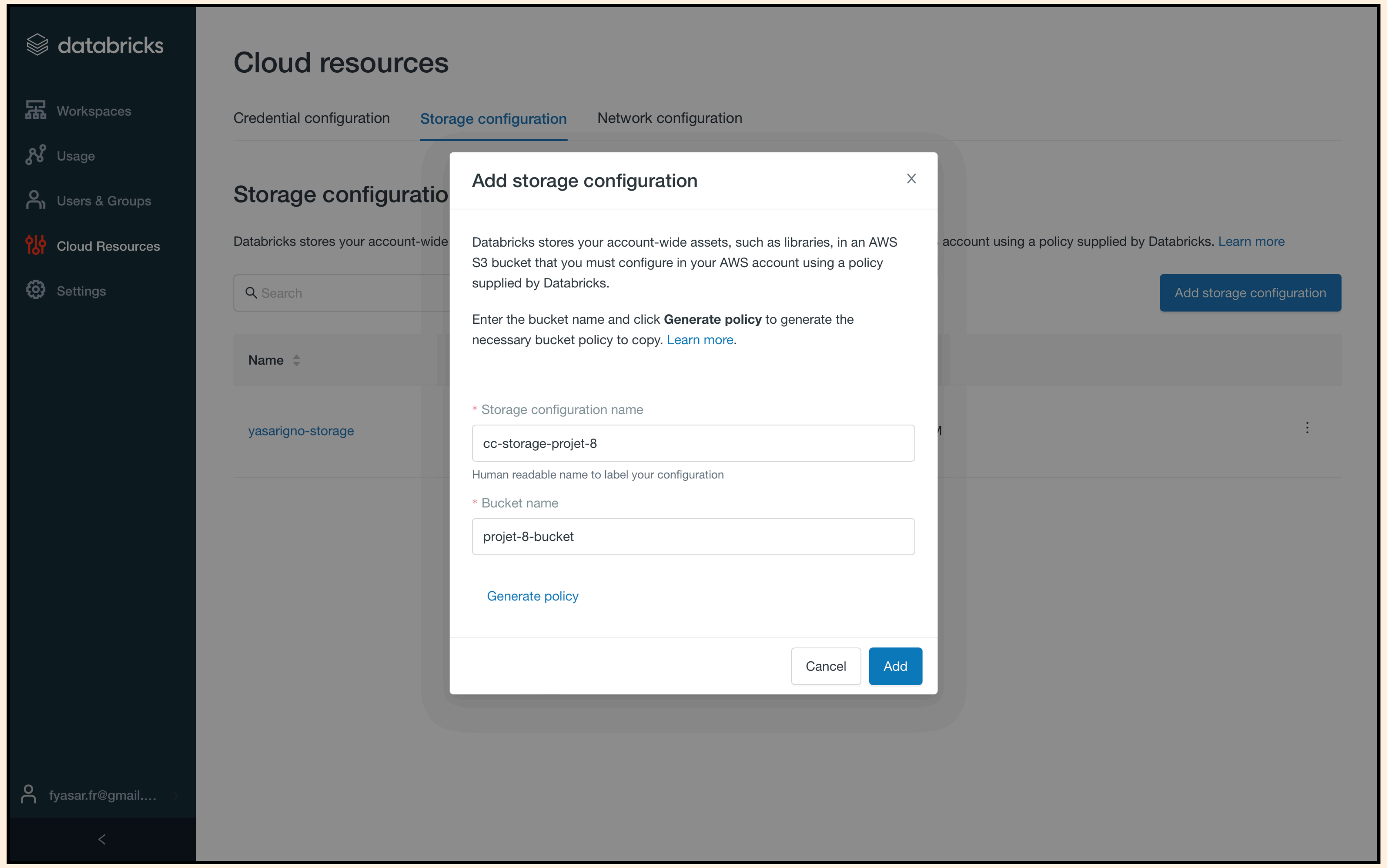
Bucket ARN
arn:aws:s3:::projet-8-bucket

Policy

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "Statement1",
            "Principal": {},
            "Effect": "Allow",
            "Action": [],
            "Resource": []
        }
    ]
}
```

Edit statement

Select a statement
Select an existing statement in the policy or add a new statement.



Add storage configuration

Databricks stores your account-wide assets, such as libraries, in an AWS S3 bucket that you must configure in your AWS account using a policy supplied by Databricks.

Enter the bucket name and click **Generate policy** to generate the necessary bucket policy to copy. Learn more.

* Storage configuration name: cc-storage-projet-8

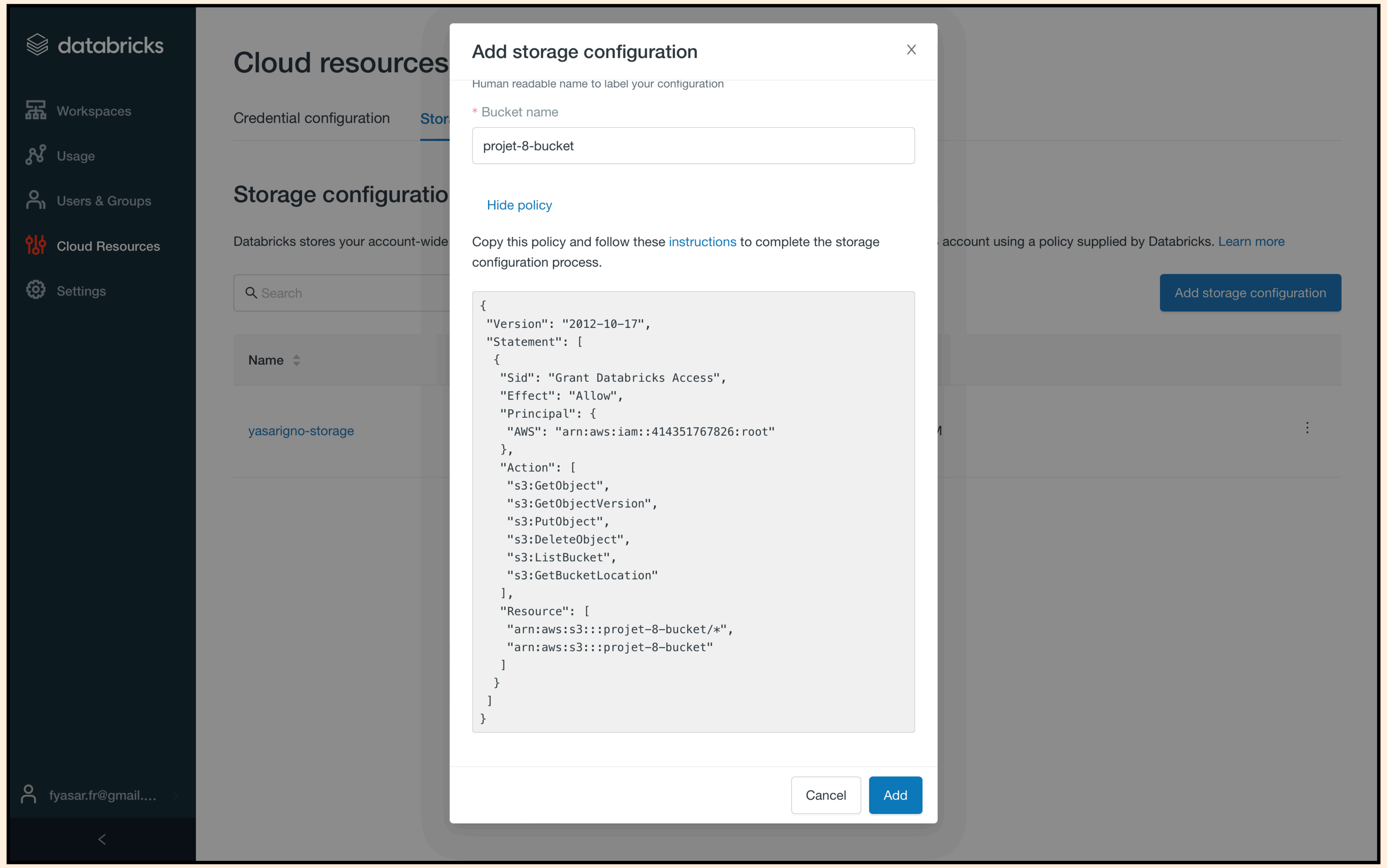
Human readable name to label your configuration

* Bucket name: projet-8-bucket

Generate policy

Cancel | Add

3



Add storage configuration

Human readable name to label your configuration

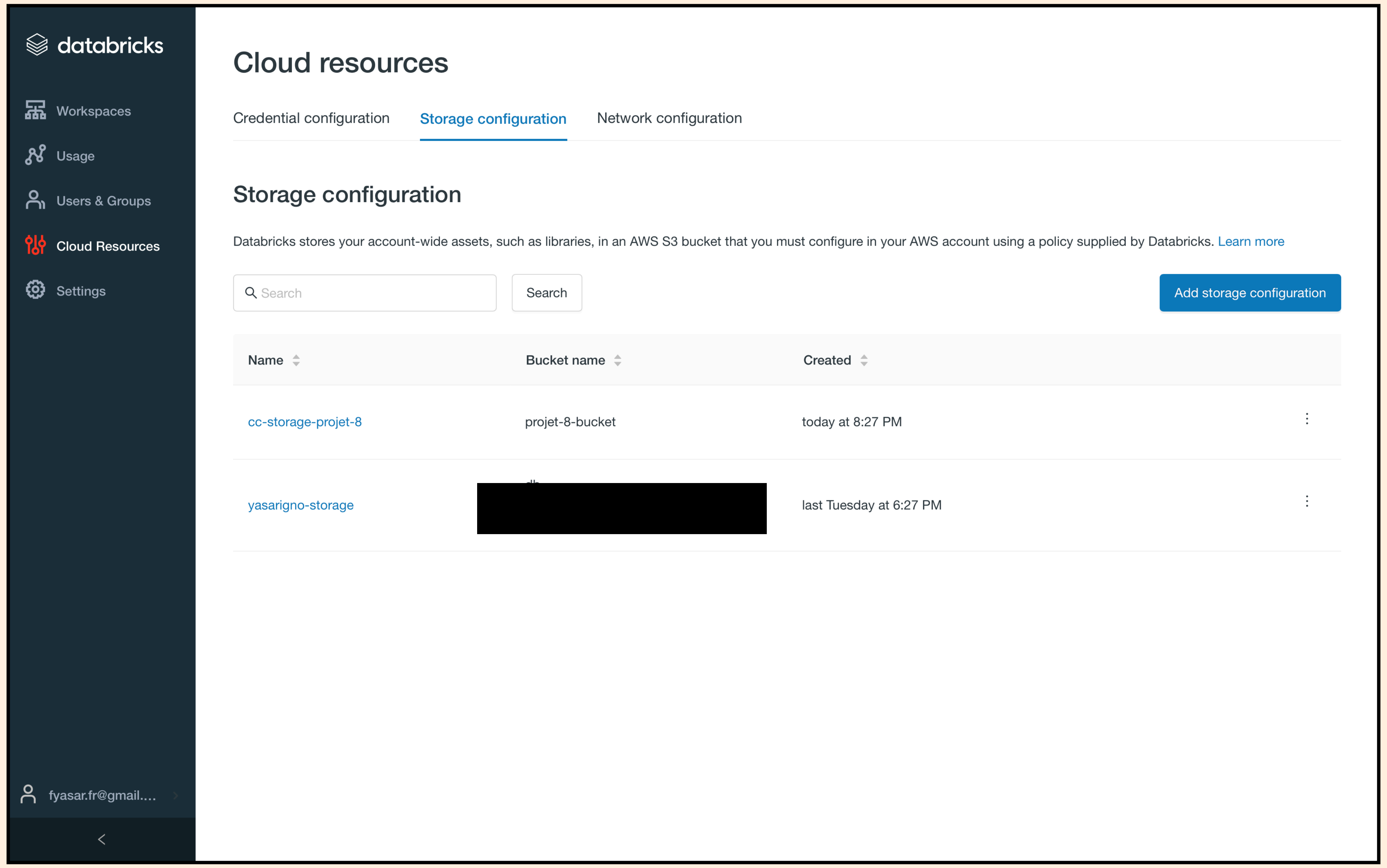
* Bucket name: projet-8-bucket

Hide policy

Copy this policy and follow these instructions to complete the storage configuration process.

```
{
 "Version": "2012-10-17",
 "Statement": [
  {
   "Sid": "Grant Databricks Access",
   "Effect": "Allow",
   "Principal": {
    "AWS": "arn:aws:iam::414351767826:root"
   },
   "Action": [
    "s3:GetObject",
    "s3:GetObjectVersion",
    "s3:PutObject",
    "s3:DeleteObject",
    "s3:ListBucket",
    "s3:GetBucketLocation"
   ],
   "Resource": [
    "arn:aws:s3:::projet-8-bucket/*",
    "arn:aws:s3:::projet-8-bucket"
   ]
  }
 ]
}
```

Cancel | Add

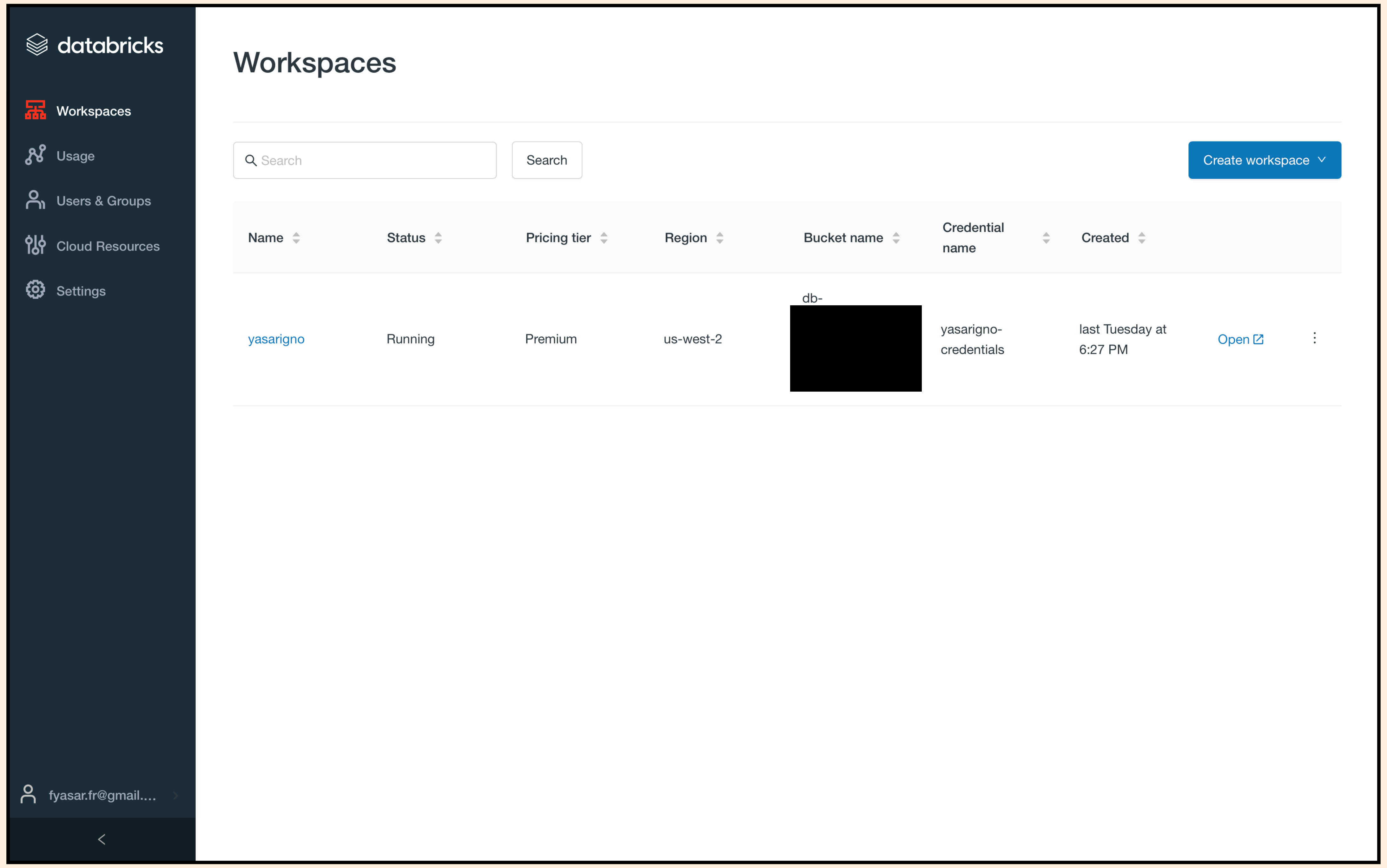


Cloud resources

Storage configuration

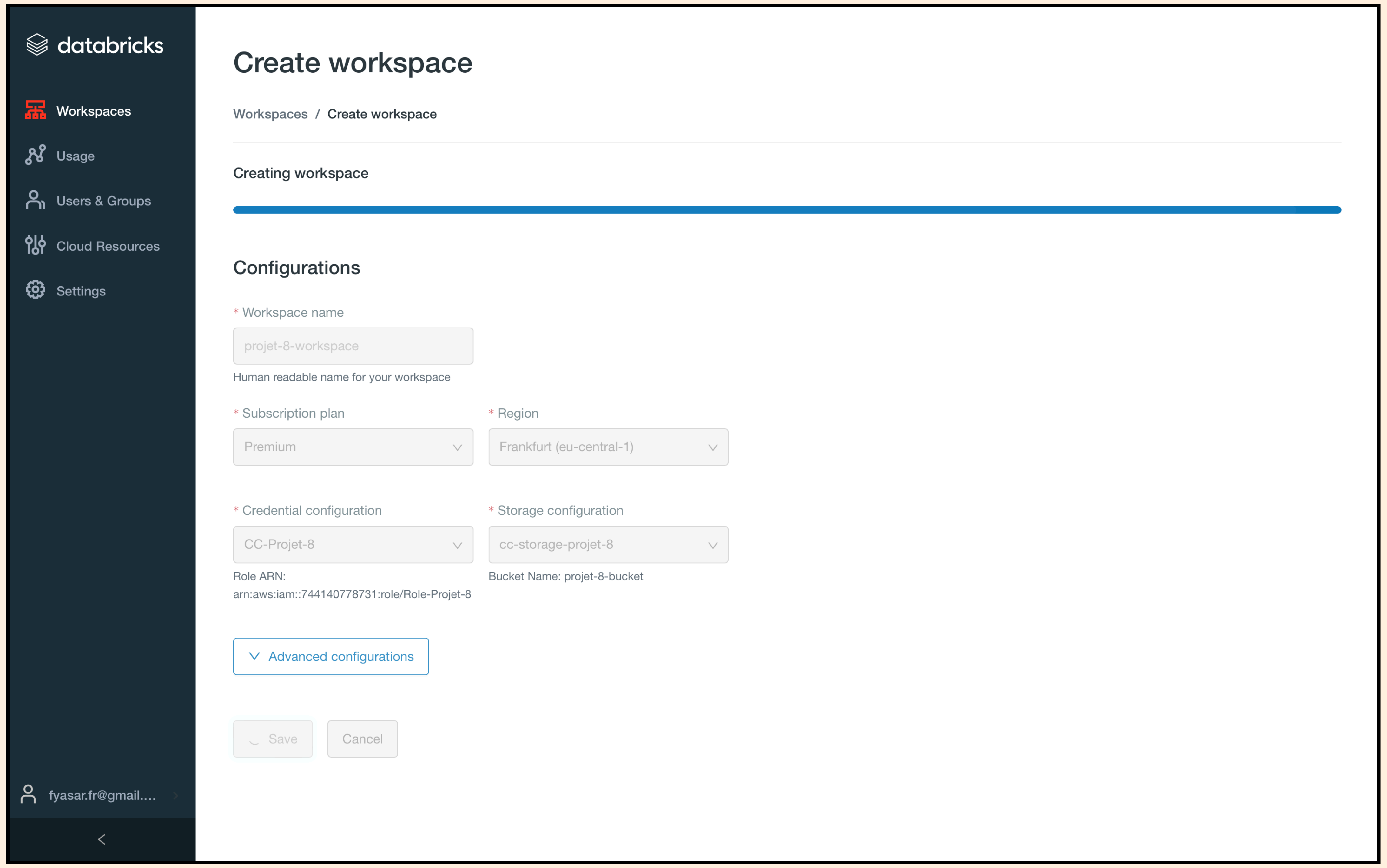
Databricks stores your account-wide assets, such as libraries, in an AWS S3 bucket that you must configure in your AWS account using a policy supplied by Databricks. Learn more

Name	Bucket name	Created
cc-storage-projet-8	projet-8-bucket	today at 8:27 PM
yasarigno-storage	[redacted]	last Tuesday at 6:27 PM



Workspaces

Name	Status	Pricing tier	Region	Bucket name	Credential name	Created	
yasarigno	Running	Premium	us-west-2	db- [redacted]	yasarigno-credentials	last Tuesday at 6:27 PM	Open



Create workspace

Workspaces / Create workspace

Creating workspace

Configurations

* Workspace name: projet-8-workspace

Human readable name for your workspace

* Subscription plan: Premium

* Region: Frankfurt (eu-central-1)

* Credential configuration: CC-Projet-8

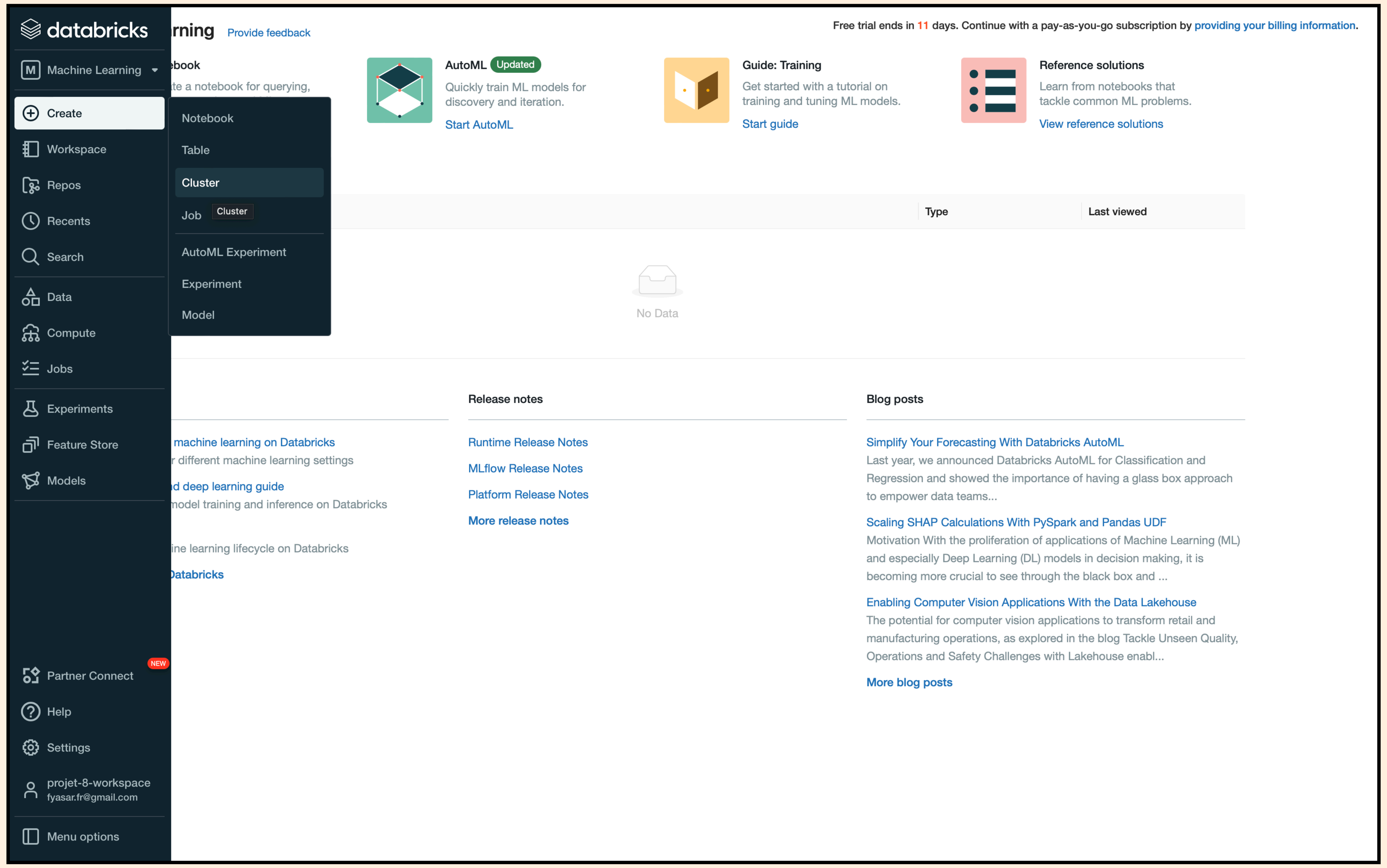
Role ARN: arn:aws:iam::744140778731:role/Role-Projet-8

* Storage configuration: cc-storage-projet-8

Bucket Name: projet-8-bucket

Annexe III

Databricks - AWS : Création d’un cluster et d’un notebook



rning Provide feedback

Free trial ends in 11 days. Continue with a pay-as-you-go subscription by providing your billing information.

book
te a notebook for querying,

AutoML Updated
Quickly train ML models for discovery and iteration.
Start AutoML

Guide: Training
Get started with a tutorial on training and tuning ML models.
Start guide

Reference solutions
Learn from notebooks that tackle common ML problems.
View reference solutions

machine learning on Databricks
r different machine learning settings
d deep learning guide
model training and inference on Databricks
ine learning lifecycle on Databricks
Databricks

Release notes
Runtime Release Notes
MLflow Release Notes
Platform Release Notes
More release notes

Blog posts

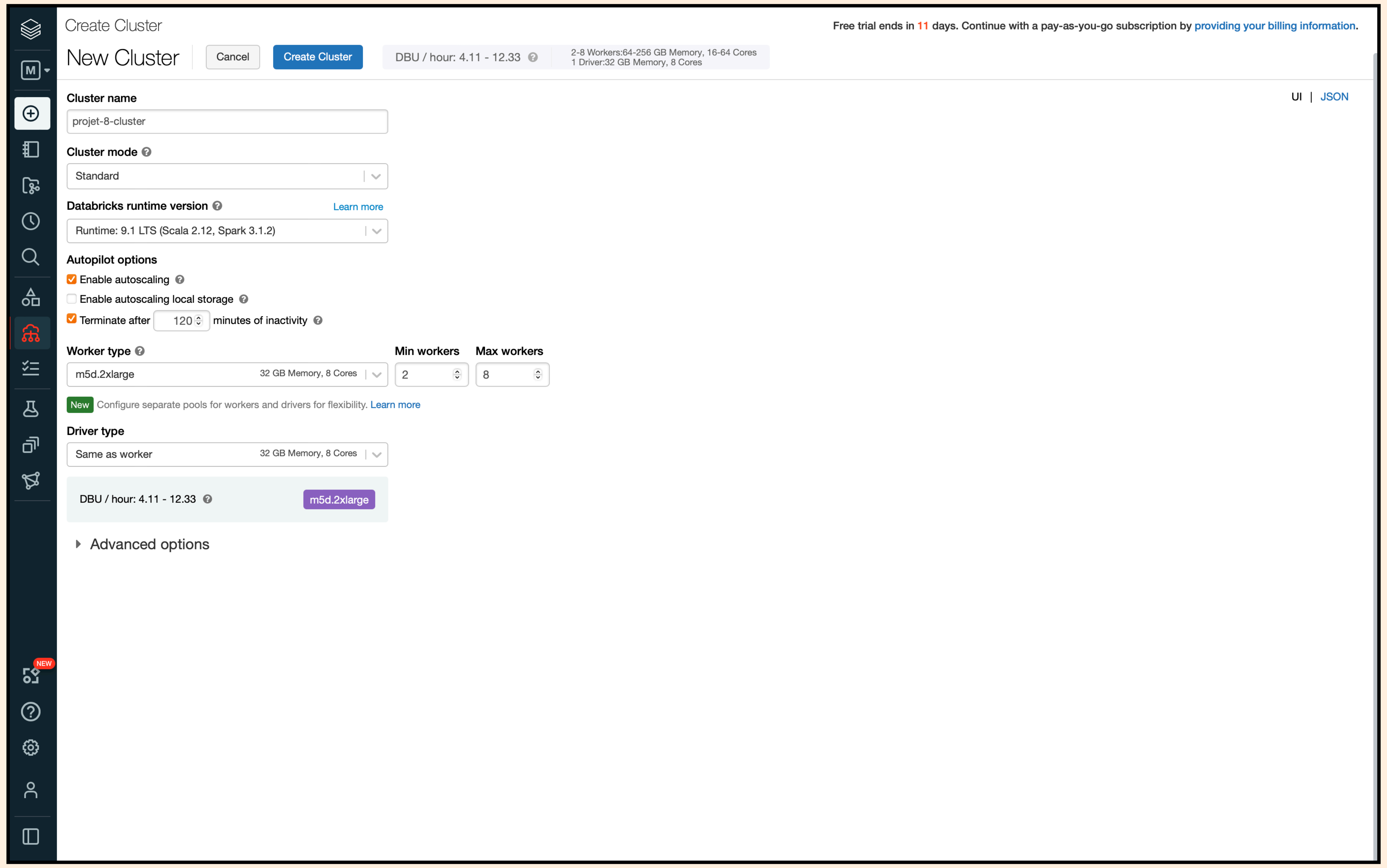
Simplify Your Forecasting With Databricks AutoML
Last year, we announced Databricks AutoML for Classification and Regression and showed the importance of having a glass box approach to empower data teams...

Scaling SHAP Calculations With PySpark and Pandas UDF
Motivation With the proliferation of applications of Machine Learning (ML) and especially Deep Learning (DL) models in decision making, it is becoming more crucial to see through the black box and ...

Enabling Computer Vision Applications With the Data Lakehouse
The potential for computer vision applications to transform retail and manufacturing operations, as explored in the blog Tackle Unseen Quality, Operations and Safety Challenges with Lakehouse enabl...

More blog posts

projet-8-workspace
fyasar.fr@gmail.com



Create Cluster

Free trial ends in 11 days. Continue with a pay-as-you-go subscription by providing your billing information.

New Cluster

DBU / hour: 4.11 - 12.33
2-8 Workers:64-256 GB Memory, 16-64 Cores
1 Driver:32 GB Memory, 8 Cores

Cluster name: projet-8-cluster

Cluster mode: Standard

Databricks runtime version: Runtime: 9.1 LTS (Scala 2.12, Spark 3.1.2)

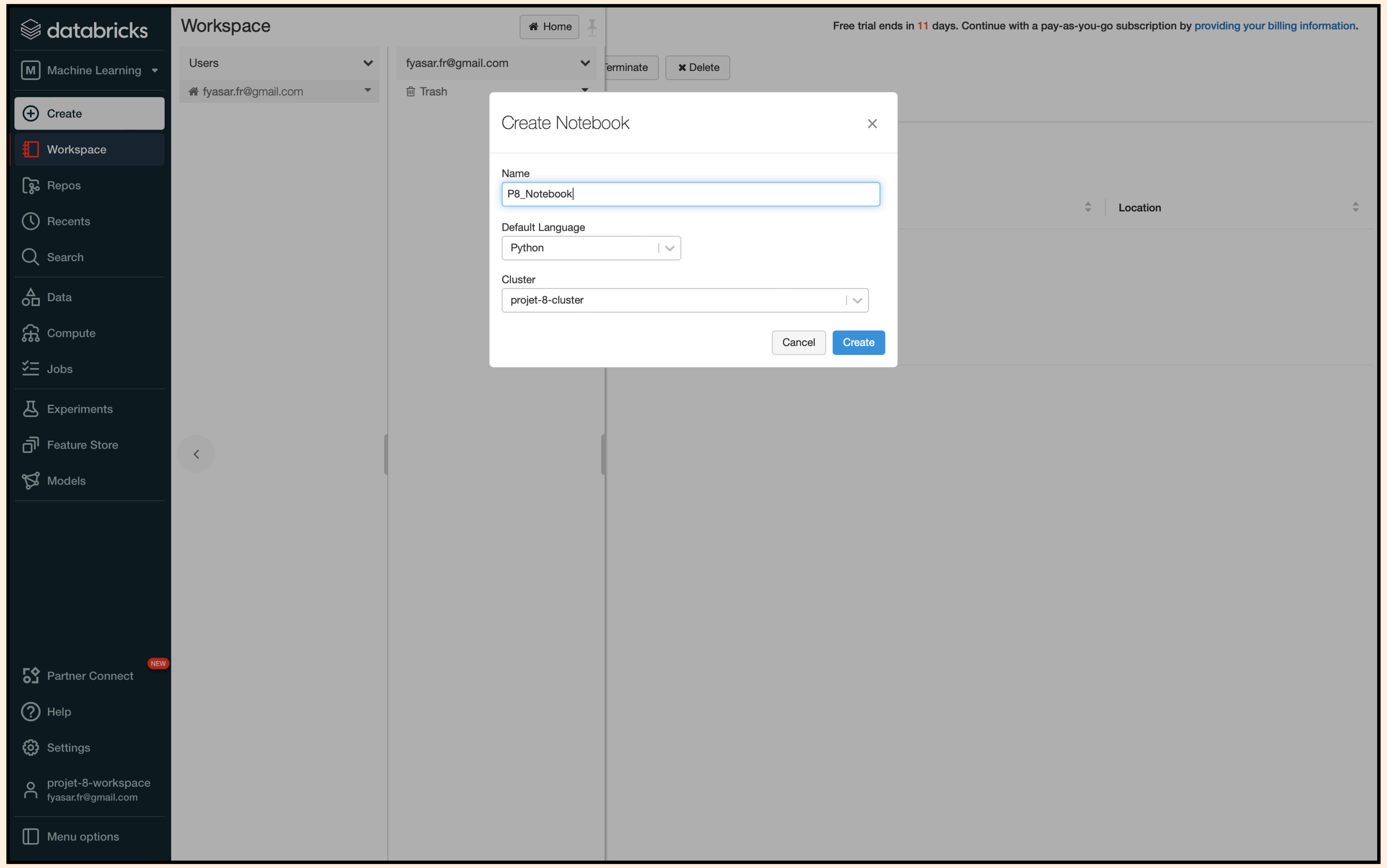
Autopilot options
☑ Enable autoscaling
□ Enable autoscaling local storage
☑ Terminate after 120 minutes of inactivity

Worker type	Min workers	Max workers
m5d.2xlarge 32 GB Memory, 8 Cores	2	8

New Configure separate pools for workers and drivers for flexibility. Learn more

Driver type: Same as worker 32 GB Memory, 8 Cores

DBU / hour: 4.11 - 12.33 m5d.2xlarge



Workspace

Free trial ends in 11 days. Continue with a pay-as-you-go subscription by providing your billing information.

fyasar.fr@gmail.com

Create Notebook

Name: P8_Notebook

Default Language: Python

Cluster: projet-8-cluster

projet-8-workspace
fyasar.fr@gmail.com