Practical - 2

Aim :- Implementation of a RAG (Retrieval-Augmented Generation) using AWS

STEPS:

Step 1: Setup AWS Services

1. **Create an AWS Account**: Sign up for AWS using a debit/credit card, email, and phone number.
2. **Enable Amazon Bedrock**: Navigate to AWS Bedrock and request access.
3. **Select LLM for Generation**: Choose the Meta model (Llama 3 8B Instruct) for text generation.
4. **Enable Amazon Titan Embeddings**: Use Titan Embeddings to generate vector representations of documents.
5. **Set Up Amazon S3**: Store raw text documents and datasets in an S3 bucket.

Step 2: Data Preprocessing & Embeddings

6. **Upload Documents to S3**: Store knowledge base files (PDFs, TXT, etc.) in an S3 bucket.
7. **Generate Embeddings**: Use Amazon Titan Embeddings to convert documents into vector representations.
8. **Store Embeddings in Amazon OpenSearch**: Use OpenSearch for vector-based similarity search.

Step 3: Retrieval System & Model Integration

9. **Implement a Search Pipeline**: Use OpenSearch's kNN search to retrieve relevant documents.
10. **Context-based Generation**: Pass retrieved results to Amazon Bedrock's Meta model to generate responses.

Step 4: API & Frontend Deployment

11. **Develop an API**: Use AWS Lambda & API Gateway to create a chatbot backend.
12. **Deploy the Chatbot Frontend**: Use AWS Amplify or Amazon S3 with CloudFront to host the chatbot UI.

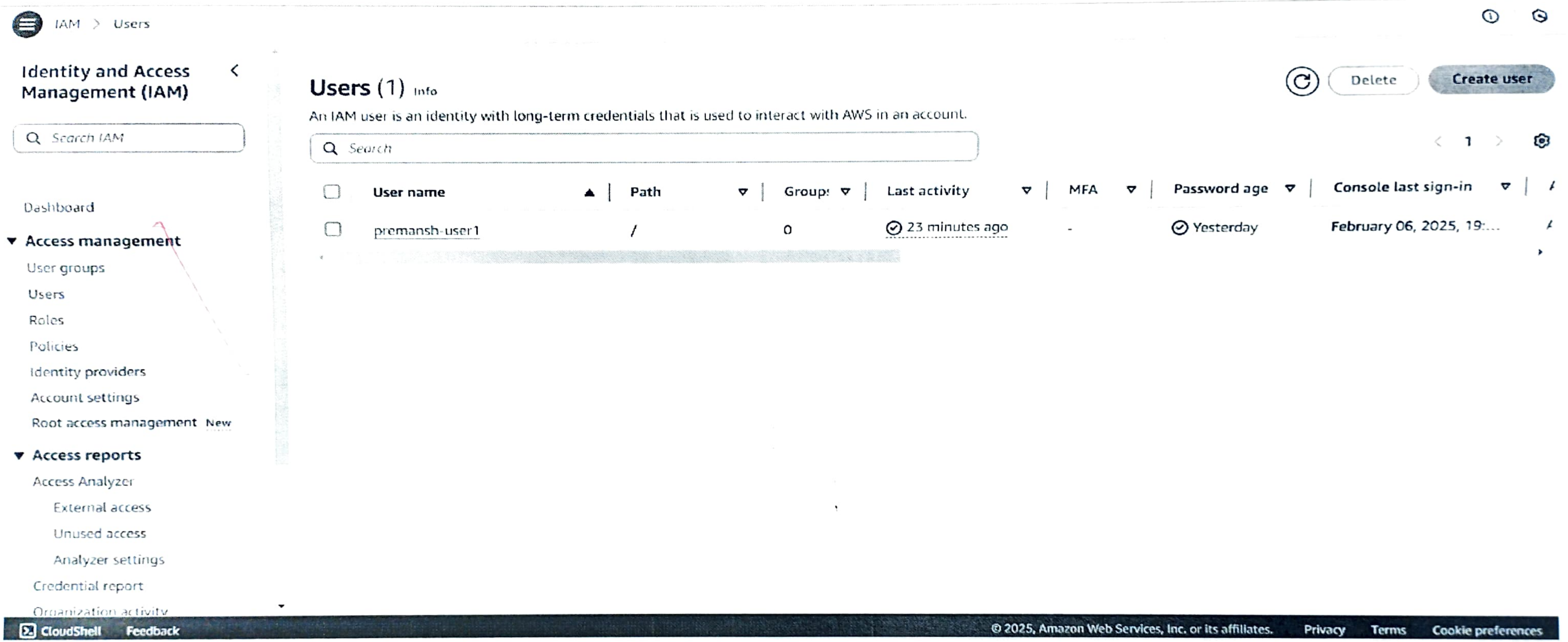
Step 5: AWS Configuration & Deployment

13. **Configure AWS CLI**: Run the following command and enter your AWS credentials:

```
aws configure
```

- Enter **Access Key ID** and **Secret Access Key**.
- Set default **region** (e.g., us-west-2).
- Select output format as json.

14. **Test the Chatbot**: Deploy the API and frontend, then run queries to test retrieval-augmented responses.

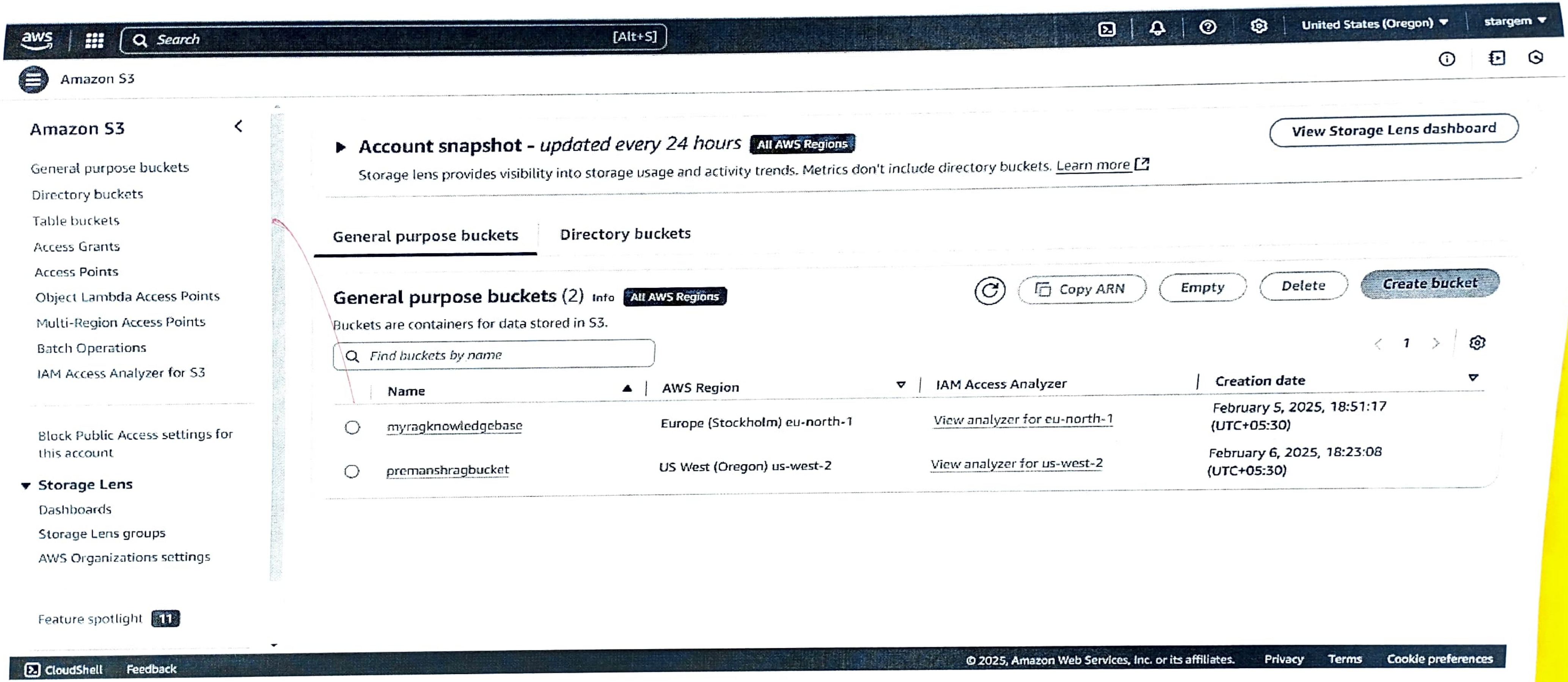


Users (1) Info

An IAM user is an identity with long-term credentials that is used to interact with AWS in an account.

□	User name	Path	Group:	Last activity	MFA	Password age	Console last sign-in
□	premansh-user1	/	0	✓ 23 minutes ago	-	✓ Yesterday	February 06, 2025, 19:...

© 2025, Amazon Web Services, Inc. or its affiliates.



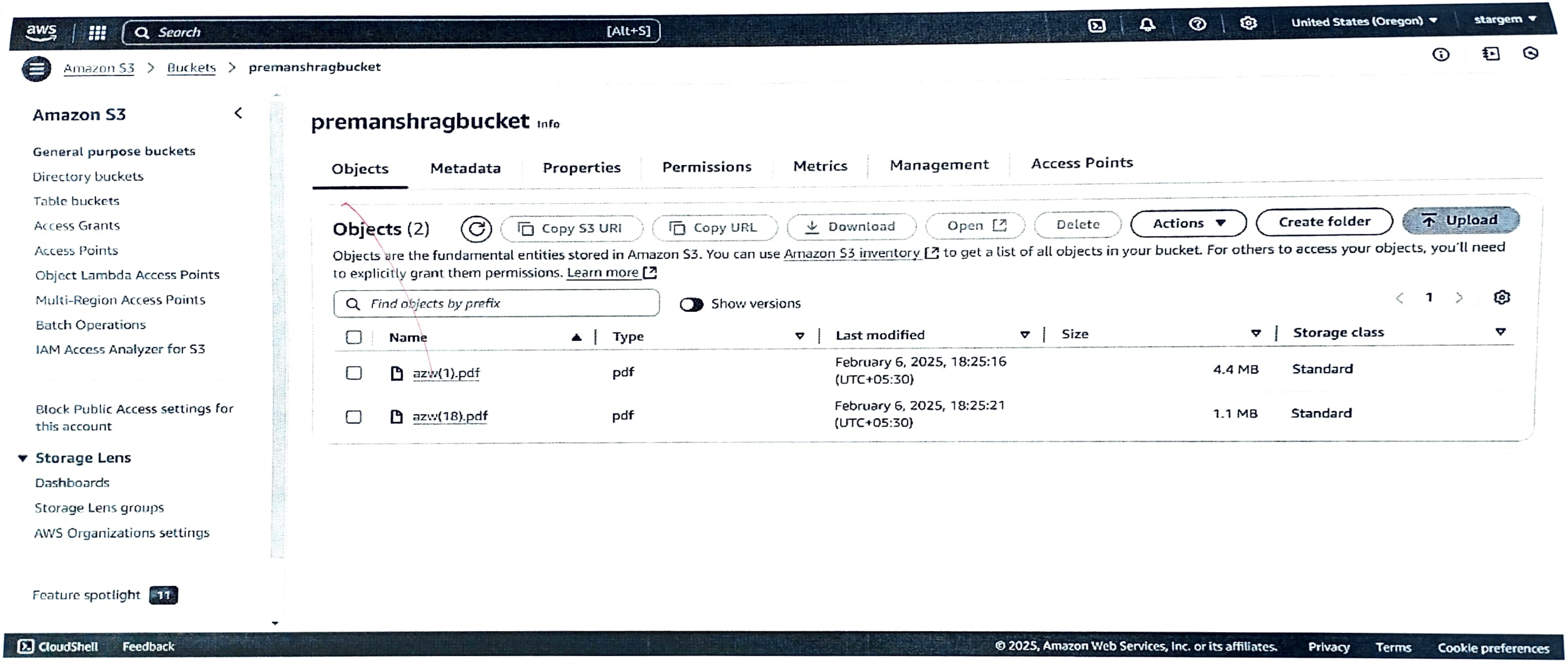
▸ **Account snapshot - *updated every 24 hours*** All AWS Regions

Storage lens provides visibility into storage usage and activity trends. Metrics don't include directory buckets. Learn more

General purpose buckets (2)

Buckets are containers for data stored in S3.

	Name ▲	AWS Region ▽	IAM Access Analyzer	Creation date ▽
○	myragknowledgebase	Europe (Stockholm) eu-north-1	View analyzer for eu-north-1	February 5, 2025, 18:51:17 (UTC+05:30)
○	premanshragbucket	US West (Oregon) us-west-2	View analyzer for us-west-2	February 6, 2025, 18:23:08 (UTC+05:30)



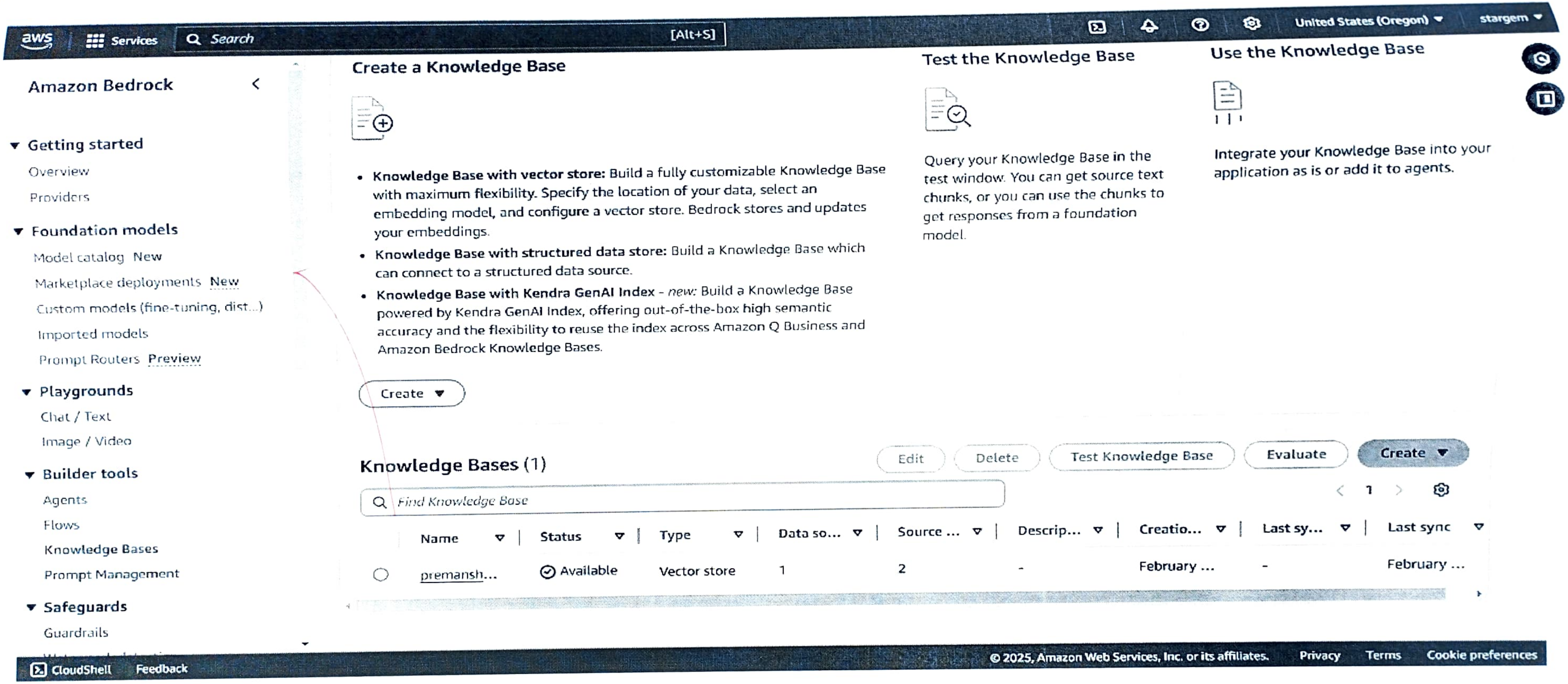
premanshragbucket Info

Objects (2)

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
□	azw(1).pdf	pdf	February 6, 2025, 18:25:16 (UTC+05:30)	4.4 MB	Standard
□	azw(18).pdf	pdf	February 6, 2025, 18:25:21 (UTC+05:30)	1.1 MB	Standard

© 2025, Amazon Web Services, Inc. or its affiliates.



Create a Knowledge Base

- **Knowledge Base with vector store:** Build a fully customizable Knowledge Base with maximum flexibility. Specify the location of your data, select an embedding model, and configure a vector store. Bedrock stores and updates your embeddings.
- **Knowledge Base with structured data store:** Build a Knowledge Base which can connect to a structured data source.
- **Knowledge Base with Kendra GenAI Index** - *new:* Build a Knowledge Base powered by Kendra GenAI Index, offering out-of-the-box high semantic accuracy and the flexibility to reuse the index across Amazon Q Business and Amazon Bedrock Knowledge Bases.

Test the Knowledge Base

Query your Knowledge Base in the test window. You can get source text chunks, or you can use the chunks to get responses from a foundation model.

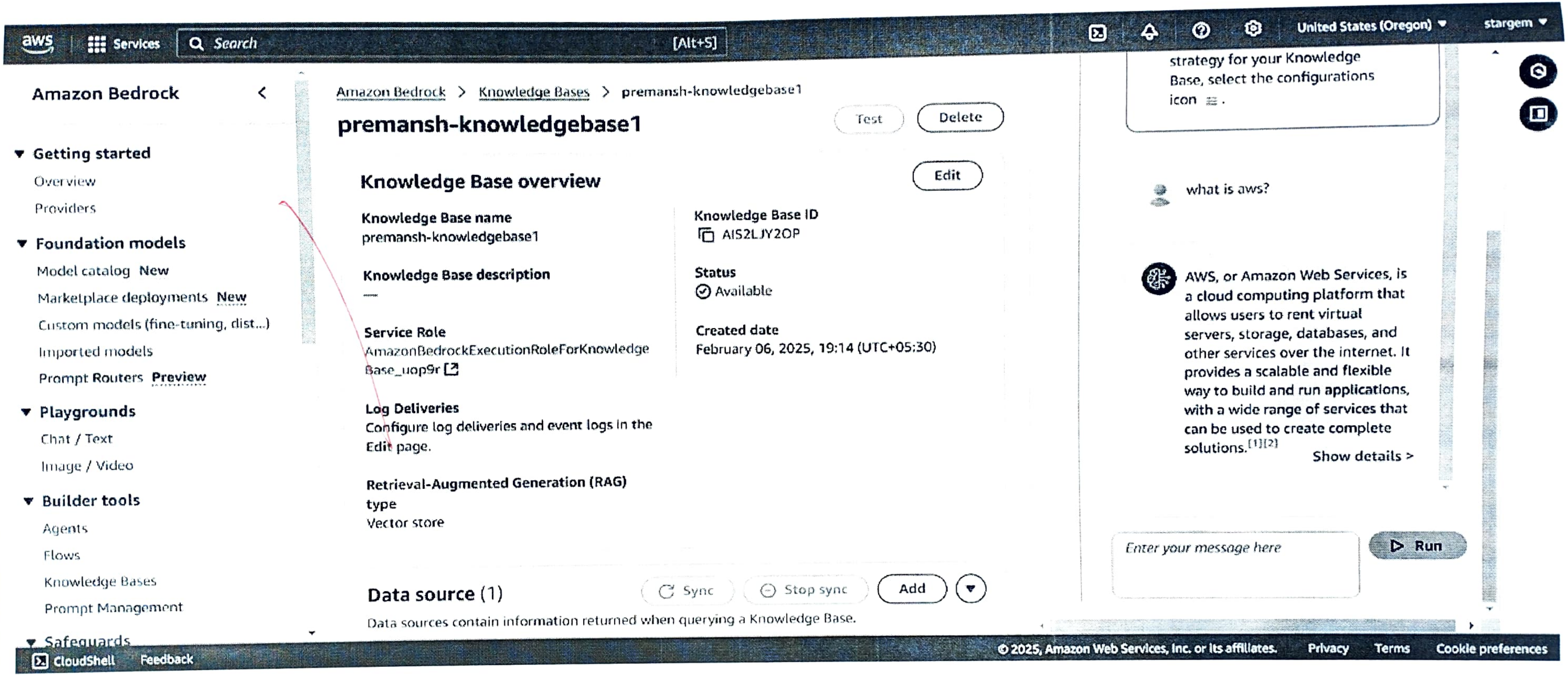
Use the Knowledge Base

Integrate your Knowledge Base into your application as is or add it to agents.

Knowledge Bases (1)

	Name	Status	Type	Data so...	Source ...	Descrip...	Creatio...	Last sy...	Last sync
○	premansh...	Available	Vector store	1	2	-	February ...	-	February ...

© 2025, Amazon Web Services, Inc. or its affiliates.



premansh-knowledgebase1

Knowledge Base overview

Knowledge Base name
premansh-knowledgebase1

Knowledge Base description
—

Service Role
AmazonBedrockExecutionRoleForKnowledgeBase_uop9r

Log Deliveries
Configure log deliveries and event logs in the Edit page.

Retrieval-Augmented Generation (RAG) type
Vector store

Knowledge Base ID
AIS2LJY2OP

Status
Available

Created date
February 06, 2025, 19:14 (UTC+05:30)

Data source (1)

Data sources contain information returned when querying a Knowledge Base.

strategy for your Knowledge Base, select the configurations icon ≡ .

what is aws?

AWS, or Amazon Web Services, is a cloud computing platform that allows users to rent virtual servers, storage, databases, and other services over the internet. It provides a scalable and flexible way to build and run applications, with a wide range of services that can be used to create complete solutions.[1][2]

Show details >

Enter your message here