Private AI Document Search by Code Creator

Customer User Instructions

What this AMI provides

A private document intelligence, vector search, and MCP-ready retrieval server running inside your AWS account. Upload documents, convert them with Docling, index them into Qdrant, and retrieve source-backed private passages from a browser interface.

Item	Value
Product	Private AI Document Search by Code Creator
Default login user	admin
Linux user	ubuntu
Main web port	80 TCP
SSH port	22 TCP
Recommended standard instance	m5.large or m6i.large
Recommended storage	50 GB or larger
First boot warm-up	**Allow 3 to 5 minutes before uploading documents**

Important: This product works in private retrieval mode out of the box. It retrieves relevant document passages without requiring an external LLM key. Optional generated summaries and polished answers can be enabled later by connecting an external or local model provider.

Contents

1. Product overview

Private AI Document Search by Code Creator is a self-hosted AWS Marketplace AMI for private document search and retrieval. It is designed for teams that want to build a searchable private knowledge base inside their own AWS account.

The server combines a simple browser interface with Docling document conversion, Qdrant vector search, and private retrieval workflows. It can be used for policies, manuals, technical documents, support knowledge, business records, research files, contracts, and other internal documents.

Plain English summary

Upload documents once. The server converts and indexes them. After that, you can search or ask questions over the indexed documents without uploading the same file again.

2. Launch requirements

Use the default Ubuntu AMI SSH username:

ubuntu

Example SSH command:

ssh -i /path/to/your-key.pem ubuntu@<public-ip-address>

Replace /path/to/your-key.pem with your EC2 key pair file and replace <public-ip-address> with the public IP address shown in the EC2 console.

This AMI uses the Ubuntu default login username: ubuntu.

Requirement	Recommendation
Operating system user	Connect with the ubuntu user.
SSH access	Use your Amazon EC2 private key.
Security group	Open TCP 22 for SSH and TCP 80 for the web application.
Storage	Use at least 30 GB. Use 50 GB or more for normal document libraries.
Browser	Use a modern browser such as Chrome, Edge, Firefox, or Safari.

***First boot timing

After launching the instance, wait approximately 3 to 5 minutes before uploading documents. If conversion is not ready immediately, wait a few minutes and refresh the page.

3. First login

After the instance is running, connect by SSH and view the first login file:

```
cat /home/ubuntu/FIRST_LOGIN.txt
```

The first login file shows the public web URL, generated administrator password, Docling UI URL, health check URL, and helper commands.

Open the main application URL in your browser:

```
http://<public-ip>/
```

Use the generated credentials shown in the first login file:

Field	Value
Username	admin
Password	Generated on first boot and shown in /home/ubuntu/FIRST_LOGIN.txt

4. Using the document search application

The home page is divided into three main areas:

Area	Purpose
Upload and index documents	Upload PDF, Word, text, Markdown, or similar document files. The server converts and indexes the text.
Search knowledge base	Search indexed documents using private semantic search.
Ask / retrieve context	Ask a question and retrieve relevant private passages from indexed documents.

Upload workflow

1. Open the main application page.
2. Choose a document file in the Upload and Index section.
3. Click Upload, Convert, and Index.
4. Wait for the upload and conversion to complete. Large PDFs may take several minutes.
5. Confirm that the document appears in the Indexed Document Library.

The browser file picker may clear

After upload, the browser may show 'No file chosen' again. That does not mean the document is gone. The document remains indexed and appears in the Indexed Document Library below.

5. Indexed document library

The Indexed Document Library shows documents that are already uploaded, converted, and searchable. Users do not need to upload the same document again after it appears in the library.

Library field	Meaning
Document	The original uploaded file name and internal document ID.
Indexed text	The amount of extracted text available for search.
Preview	A short preview of the converted document text.
View text	Opens the extracted text generated from the uploaded document.
Ask test question	Runs a sample retrieval question against the indexed content.

6. Search and ask/retrieve context

Search

Use Search Knowledge Base when you want to search across indexed documents. Search results include document names, similarity scores, and relevant text passages.

Ask / retrieve context

Use Ask / Retrieve Context when you want to ask a question about indexed documents. By default, the product runs in private retrieval mode. It returns the most relevant source-backed passages rather than generating a polished LLM-written answer.

Generated answers are optional

No external LLM key is required for private retrieval mode. If a model provider is configured later, the system can be extended to generate polished summaries and natural-language answers.

Good test questions after uploading the included validation document include:

- What is the project codename?
- What is the unique validation phrase?
- Which services power the product?
- Which public ports should be opened?
- What is the backup helper command?

7. Docling UI playground

The Docling UI playground is available for testing document conversion behavior. It is useful for advanced users who want to inspect Docling separately from the main application.

```
http://<public-ip>/docling/ui/
```

Important distinction

Uploading a document in the Docling UI playground may convert the file, but it does not automatically index the document into Qdrant. To build the searchable knowledge base, upload documents through the main application page.

8. Helper commands

The AMI includes helper commands to make administration easier.

Command	Purpose
codecreator-ai-kb-url	Show application URLs and login credentials.
codecreator-ai-kb-status	Show product status, Docker containers, local health, and Nginx status.
codecreator-ai-kb-test	Run local and public health checks for the app, Qdrant, and Docling.
codecreator-ai-kb-credentials	Show the generated admin username and password.
codecreator-ai-kb-logs	View recent application logs.
codecreator-ai-kb-restart	Restart the application stack.
codecreator-ai-kb-backup	Create a backup of application data and indexed documents.
codecreator-ai-kb-mcp-info	Show MCP-ready retrieval information.

9. Backup guidance

Use the included backup helper to create a backup before making major changes, applying upgrades, or stopping the instance for long periods.

```
sudo codecreator-ai-kb-backup
```

The backup helper is intended to preserve important application data, converted document text, uploaded files, and vector database data. Store backups in a safe location according to your internal retention policy.

10. Security and networking

The product is designed for a simple public IP based setup. The main web application is protected by Basic Auth, and internal services are not exposed publicly by default.

Port	Public?	Purpose
22 TCP	Yes	SSH administration using the ubuntu user and Amazon private key.
80 TCP	Yes	Main web application, health check, and protected web interface.
5001 TCP	No	Docling internal service bound to localhost/internal Docker networking.
6333 TCP	No	Qdrant internal HTTP API bound to localhost/internal Docker networking.
6334 TCP	No	Qdrant internal gRPC API bound to localhost/internal Docker networking.
8501 TCP	No	Private application backend bound to localhost.

Recommended security practices:

- Restrict SSH access to trusted IP addresses whenever possible.
- Change or rotate credentials according to your internal policy.
- Use snapshots or backups before major changes.
- Do not open Docling, Qdrant, or backend application ports to the public internet.
- For production use, consider adding a domain name and HTTPS reverse proxy configuration.

11. Troubleshooting

Issue	What to do
Web page does not load immediately	Wait 3 to 5 minutes after first boot and refresh the browser. Check codecreator-ai-kb-test.
Docling shows connection refused during first minutes	This usually means Docling is still warming up. Wait a few minutes and rerun codecreator-ai-kb-test.
Uploaded file seems gone after returning	The browser file picker resets after upload. Check the Indexed Document

Issue	What to do
home	Library. If the document appears there, it is still indexed.
Ask says no indexed passages found	Upload and index a document from the main application page. Do not use the Docling playground for indexing.
No polished summary is generated	Private retrieval mode is active. Configure an external or local model provider later for generated summaries.
Large PDF conversion is slow	Use a larger instance such as m5.xlarge or m6i.xlarge and allow several minutes for conversion.

```
codecreator-ai-kb-test
codecreator-ai-kb-status
codecreator-ai-kb-logs
```

12. Recommended instance sizes

Use case	Recommended instance	Notes
Small demo or light testing	t3.large	Use for light testing only. Large PDFs may be slow.
Standard use	m5.large or m6i.large	Recommended baseline for most customers.
Larger PDFs or heavier conversion	m5.xlarge or m6i.xlarge	More CPU and memory improve conversion experience.
Larger document libraries	m6i.xlarge with 100 GB+ EBS	Increase EBS capacity based on uploaded document volume.

13. Support notes

When requesting support, include the instance type, operating region, approximate time since launch, and output from the following commands:

```
cat /home/ubuntu/FIRST_LOGIN.txt
codecreator-ai-kb-status
codecreator-ai-kb-test
codecreator-ai-kb-logs
```

For security, do not share private documents, confidential business files, or sensitive passwords unless specifically required and approved by your organization.

Quick start checklist

1. Launch the AMI from AWS Marketplace.
2. Open ports 22 and 80 in the security group.
3. SSH as ubuntu and read /home/ubuntu/FIRST_LOGIN.txt.
4. Wait 3 to 5 minutes for first boot and Docling warm-up.
5. Open the main application URL and log in.
6. Upload a document through the main application page.
7. Confirm it appears in the Indexed Document Library.
8. Search or ask/retrieve context from the indexed document.