Custom Search Engine On AWS

Project Report

Aranya Singh Chauhan (CSE, SRM IST)
Eshaan Mathakari (ECE, SRM IST)

Problem Statement

Search as a capability is an important feature that is required by almost all medium and large enterprises as search helps filter relevant and required information in the world of big data. Search helps find relevant information quickly and saves time to go through vast information.

When given a large dataset, it becomes extremely difficult for an individual or an organization to fetch a piece of information that they seek. In addition to this, even with traditional searching methods (searching for keywords in a document), it can be tough to locate that particular piece of information if one isn't sure about what keywords to look for in the dataset.

Project Description

- The project aims to deploy a search engine on AWS, that not only eliminates the problems mentioned in the section above but also provides features that make the overall system keep on improving over time with the help of feedback provided by the users.

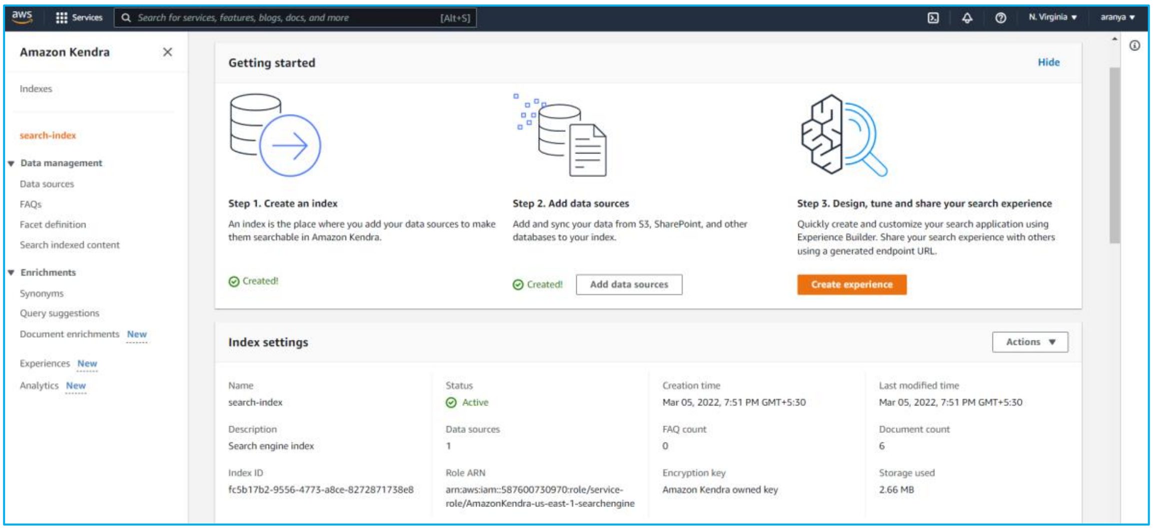
- The project makes use of **AWS Kendra**, which is an intelligent search service provided by machine learning. With AWS Kendra, we build a search engine that can use a dataset stored within an **AWS Simple Storage Service (Amazon S3) Bucket**.

- The search not only looks for keywords in the dataset, but it can also look for synonyms to the keywords that are being searched. This gives the user the ability to search for relevant content even if they are not aware of the exact keywords that they need to find. The users can even make use of natural language as their queries.

- Most commonly asked questions/queries related to the dataset can be added to the search engine as a configuration as well. This helps the search engine display the results extremely quickly if the query that the user is entering contains an FAQ, thus reducing the engine's time complexity.

Approach Used

1. **Index Creation on AWS Kendra:**
   An *index* serves the purpose of acting as a dataset source that is to be used by the search engine. Connectors are used to connect the dataset with AWS Kendra. Here, we have used an S3 bucket as the data source for the connector.



Getting started

Step 1. Create an index
An index is the place where you add your data sources to make them searchable in Amazon Kendra.
Created!

Step 2. Add data sources
Add and sync your data from S3, SharePoint, and other databases to your index.
Created!

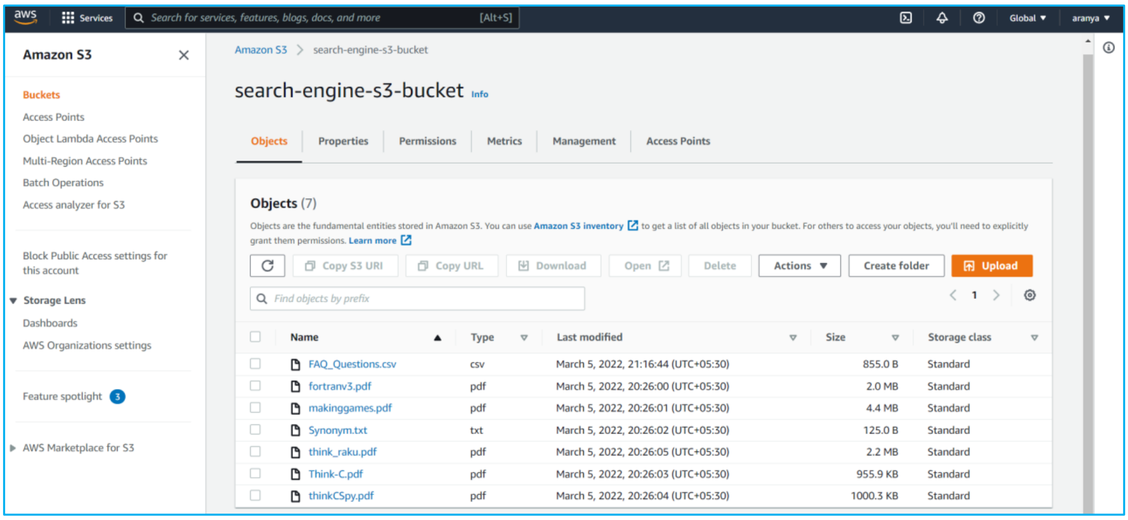
Step 3. Design, tune and share your search experience
Quickly create and customize your search application using Experience Builder. Share your search experience with others using a generated endpoint URL.

Index settings

Name: search-index
Status: Active
Creation time: Mar 05, 2022, 7:51 PM GMT+5:30
Last modified time: Mar 05, 2022, 7:51 PM GMT+5:30
Description: Search engine index
Data sources: 1
FAQ count: 0
Document count: 6
Index ID: fc5b17b2-9556-4773-a8ce-8272871738e8
Role ARN: arn:aws:iam::587600730970:role/service-role/AmazonKendra-us-east-1-searchengine
Encryption key: Amazon Kendra owned key
Storage used: 2.66 MB

AWS Kendra Index

2. **Uploading Database:** A *data source* is a location, such as an Amazon Simple Storage Service (Amazon S3) bucket, where you store the documents for indexing. One can automatically synchronize data sources with an Amazon Kendra index so that new, updated, or deleted documents in the data source are also added, updated, or deleted in the index for searching on.



search-engine-s3-bucket

Objects (7)

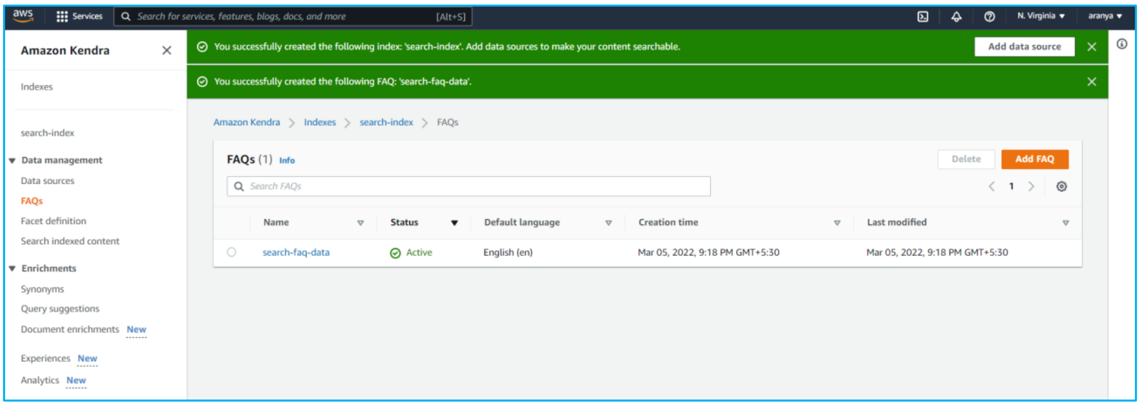
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

Name	Type	Last modified	Size	Storage class
FAQ_Questions.csv	csv	March 5, 2022, 21:16:44 (UTC+05:30)	855.0 B	Standard
fortranv3.pdf	pdf	March 5, 2022, 20:26:00 (UTC+05:30)	2.0 MB	Standard
makinggames.pdf	pdf	March 5, 2022, 20:26:01 (UTC+05:30)	4.4 MB	Standard
Synonym.txt	txt	March 5, 2022, 20:26:02 (UTC+05:30)	125.0 B	Standard
think_raku.pdf	pdf	March 5, 2022, 20:26:05 (UTC+05:30)	2.2 MB	Standard
Think-C.pdf	pdf	March 5, 2022, 20:26:03 (UTC+05:30)	955.9 KB	Standard
thinkCSpy.pdf	pdf	March 5, 2022, 20:26:04 (UTC+05:30)	1000.3 KB	Standard

Dataset for Search Engine using S3

3. **Adding Search Functionality to the Engine:** To get answers, we need to query an index. With Kendra, users can use natural language in their queries. The response contains information, such as the title, a text excerpt, and the location of documents in the index that provide the best answer.

4. **Adding FAQs and Synonyms to optimize search results:**



You successfully created the following index: 'search-index'. Add data sources to make your content searchable.

You successfully created the following FAQ: 'search-faq-data'.

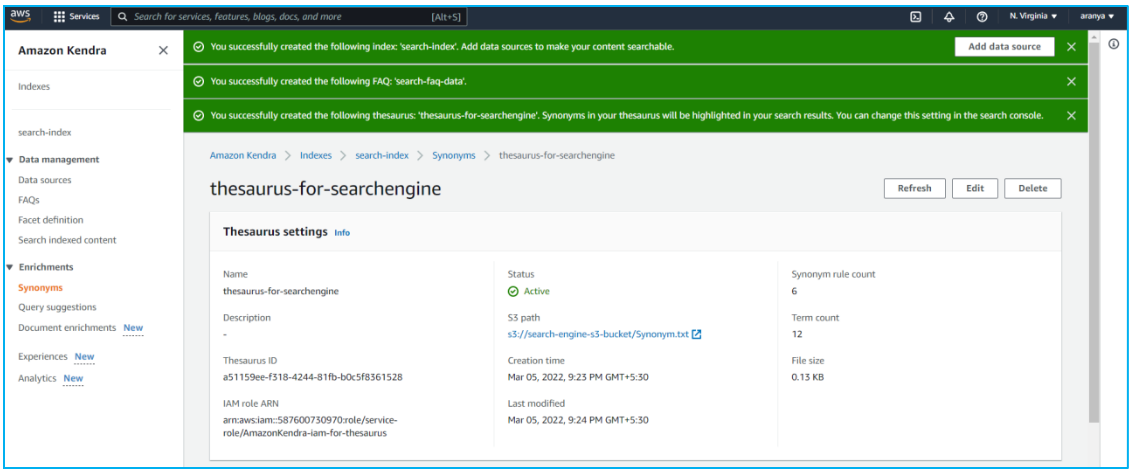
Amazon Kendra > Indexes > search-index > FAQs

FAQs (1)

	Name	Status	Default language	Creation time	Last modified
○	search-faq-data	Active	English (en)	Mar 05, 2022, 9:18 PM GMT+5:30	Mar 05, 2022, 9:18 PM GMT+5:30

Data for FAQ Search Function

In order for the search engine to be able to display the search results efficiently and quickly, we add documents containing FAQs related to the dataset along with synonyms for commonly searched terms. This not only gets the engine to display results faster, but also makes it easier for the user to search without having to worry about the keywords.



You successfully created the following index: 'search-index'. Add data sources to make your content searchable.

You successfully created the following FAQ: 'search-faq-data'.

You successfully created the following thesaurus: 'thesaurus-for-searchengine'. Synonyms in your thesaurus will be highlighted in your search results. You can change this setting in the search console.

Amazon Kendra > Indexes > search-index > Synonyms > thesaurus-for-searchengine

thesaurus-for-searchengine

Thesaurus settings

Name
thesaurus-for-searchengine

Description
-

Thesaurus ID
a51159ee-f318-4244-81fb-b0c5f8361528

IAM role ARN
arn:aws:iam::587600730970:role/service-role/AmazonKendra-iam-for-thesaurus

Status
Active

S3 path
s3://search-engine-s3-bucket/Synonym.txt

Creation time
Mar 05, 2022, 9:23 PM GMT+5:30

Last modified
Mar 05, 2022, 9:24 PM GMT+5:30

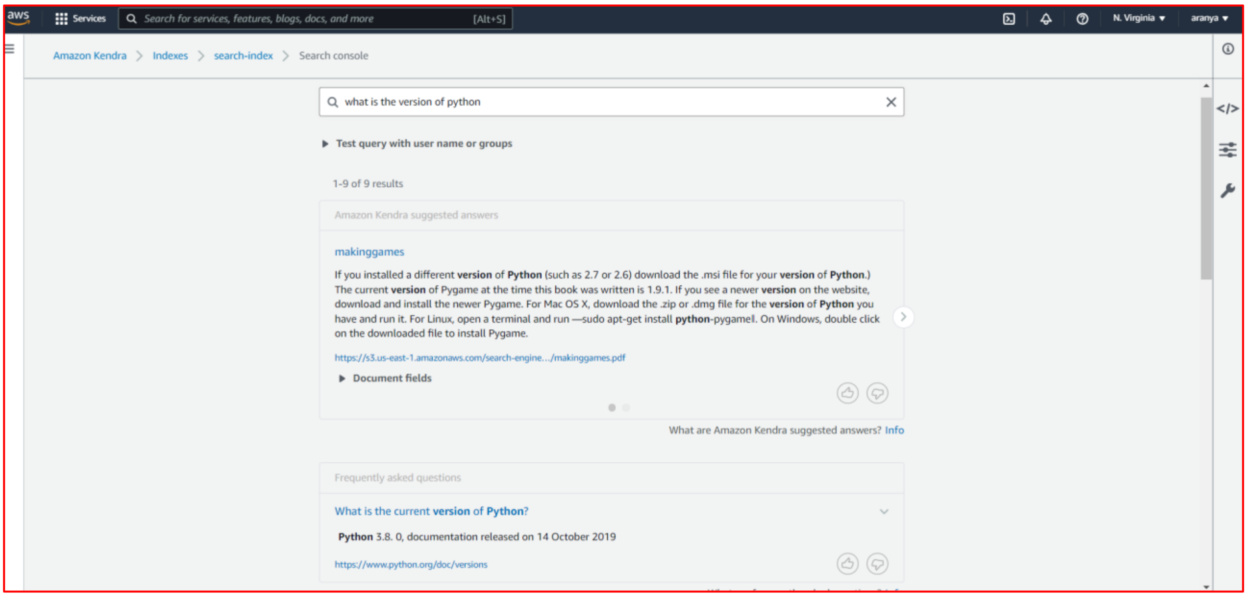
Synonym rule count
6

Term count
12

File size
0.13 KB

Thesaurus Data for Synonyms Function

Working of the Project:


aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
N. Virginia ▾
aranya ▾
Amazon Kendra > Indexes > search-index > Search console
Test query with user name or groups
What are Amazon Kendra suggested answers? Info

what is the version of python

1-9 of 9 results

Amazon Kendra suggested answers

makinggames

If you installed a different **version** of **Python** (such as 2.7 or 2.6) download the .msi file for your **version** of **Python**.) The current **version** of Pygame at the time this book was written is 1.9.1. If you see a newer **version** on the website, download and install the newer Pygame. For Mac OS X, download the .zip or .dmg file for the **version** of **Python** you have and run it. For Linux, open a terminal and run —sudo apt-get install **python**-pygamel. On Windows, double click on the downloaded file to install Pygame.

https://s3.us-east-1.amazonaws.com/search-engine.../makinggames.pdf

▶ Document fields

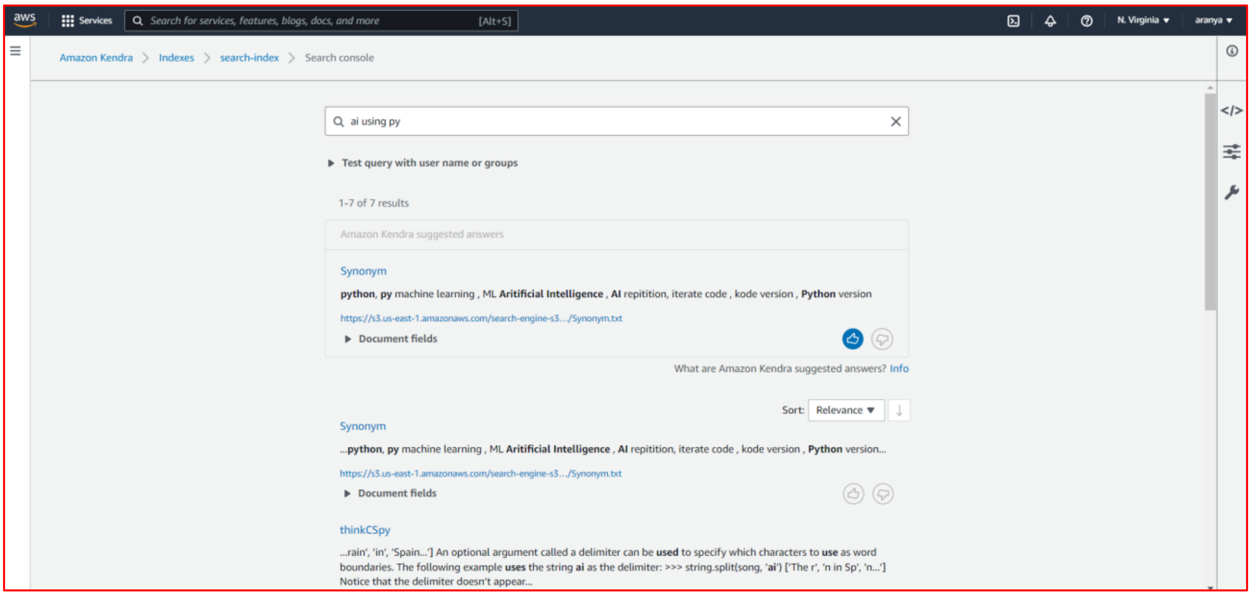
Frequently asked questions

What is the current **version** of **Python**?

Python 3.8. 0, documentation released on 14 October 2019

https://www.python.org/doc/versions

Figure 1FAQ Search Search Results for FAQs


aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
N. Virginia ▾
aranya ▾
Amazon Kendra > Indexes > search-index > Search console
Test query with user name or groups
What are Amazon Kendra suggested answers? Info
Sort: Relevance ▾

ai using py

1-7 of 7 results

Amazon Kendra suggested answers

Synonym

python, **py** machine learning , ML **Aritificial Intelligence** , **AI** repitition, iterate code , kode version , **Python** version

https://s3.us-east-1.amazonaws.com/search-engine-s3.../Synonym.txt

▶ Document fields

Synonym

...**python**, **py** machine learning , ML **Aritificial Intelligence** , **AI** repitition, iterate code , kode version , **Python** version...

https://s3.us-east-1.amazonaws.com/search-engine-s3.../Synonym.txt

▶ Document fields

thinkCSpy

...rain', 'in', 'Spain...'] An optional argument called a delimiter can be **used** to specify which characters to **use** as word boundaries. The following example **uses** the string **ai** as the delimiter: >>> string.split(song, '**ai**') ['The r', 'n in Sp', 'n...'] Notice that the delimiter doesn't appear...

Search with Synonyms, FAQ functions implemented.

Conclusion

The project was successfully deployed on AWS, using Kendra as the primary search tool and S3 for storage. Users can successfully search for content within the dataset without having to type-in the exact keywords. They can also provide feedback in regards to the results that are displayed, so that the tool can use its Machine Learning capabilities and improve when it comes to displaying accurate results.