

A background image showing two people shaking hands in a professional setting. The person on the left is wearing a light-colored sweater, and the person on the right is wearing a patterned sweater. They are standing in front of a window. In the foreground, there is a laptop, a small potted plant, and two coffee cups.

MSDS 697 | SPRING 2024 | GROUP 16

Automated Data Pipeline: Job Recommender Engine

Shagun Kala | Param Mehta | Belinda Ong
Sonal Shad | Laila Zaidi

Introduction



What we are solving for: Job seekers have to spend a lot of time searching multiple job titles and reading through individual job descriptions to understand if it is a good fit.


Solution: Personalized Job Recommender for Data Scientists.

- Automated data pipeline retrieves, processes, and stores daily job posts.
- User selects filters and submits resume.
- Resume is parsed, text is matched against job descriptions.
- Top three matches are displayed with LLM-based summaries of descriptions.


User Interface via Streamlit




×




UNIVERSITY OF
SAN FRANCISCO



[Github](#)


Job Statistics 

	Job Title	Total Number of Jobs	Average Salary	Total Number of Cities	Total Number of States
0	Data Scientist	175	164,946.41	211	39
1	Data Analyst	153	57,989.94	211	39
2	Machine Learning Engineer	166	174,939.26	211	39

Enter location 


New YorkSan FranciscoChicagoLos AngelesSeattle


No location selected

Enter Job Title 

Data ScientistData AnalystMachine Learning Engineer

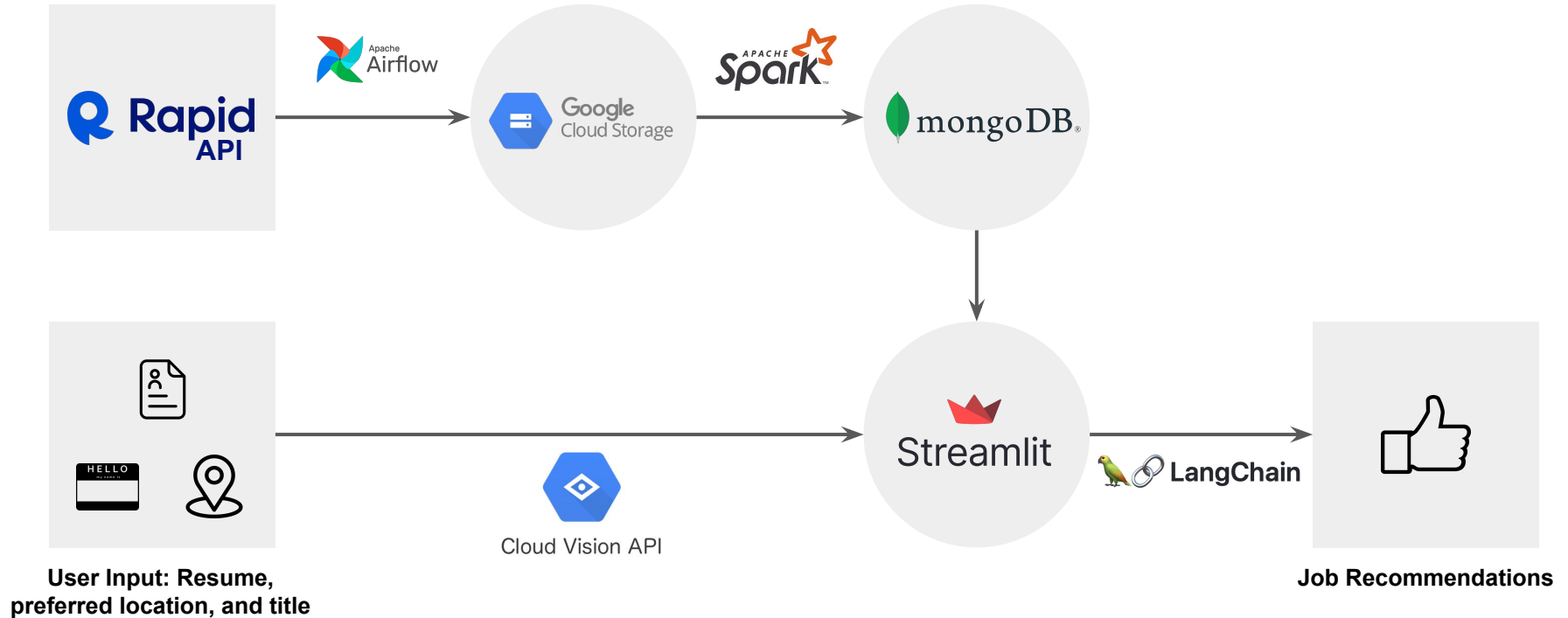
No job title selected

Upload your resume 

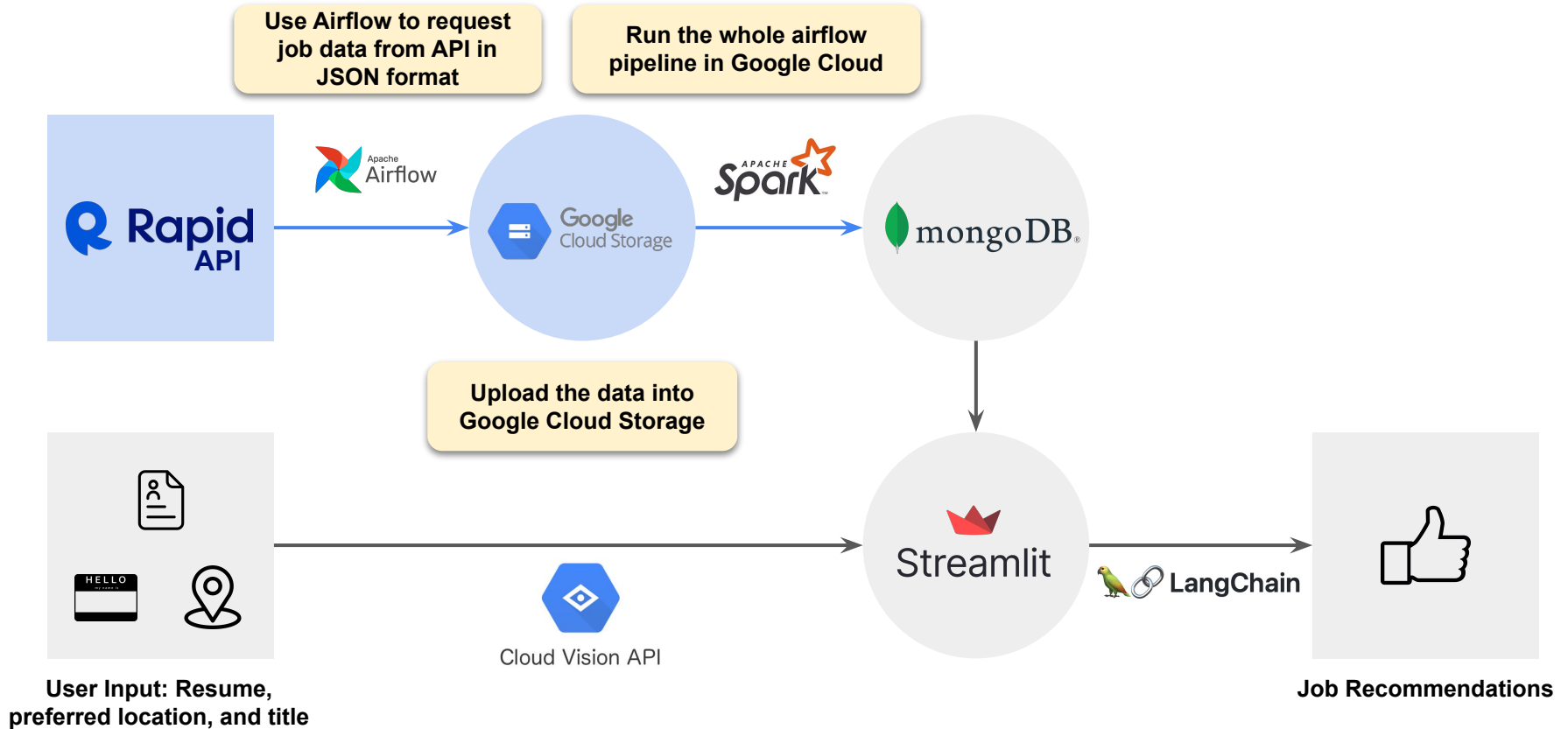
 Drag and drop file here
Limit 200MB per file • PDF

Browse files

Overview of Data Pipeline



Overview of Data Pipeline



Usage of Google Cloud Storage



Airflow will submit a request to the Rapid API every day and stores the JSON files in a new folder by date, then by job title

api_jobs_data_json

Location: us (multiple regions in United States) | Storage class: Standard | Public access: Not public | Protection: None

OBJECTS | CONFIGURATION | PERMISSIONS | PROTECTION | LIFECYCLE | OBSERVABILITY | INVENTORY REPORTS

Buckets > api_jobs_data_json

UPLOAD FILES | UPLOAD FOLDER | CREATE FOLDER | TRANSFER DATA

Filter by name prefix only | Filter | Filter objects and folders

<input type="checkbox"/>	Name	Size	Type	Created
<input type="checkbox"/>	2024-02-26/	—	Folder	—
<input type="checkbox"/>	2024-02-29/	—	Folder	—
<input type="checkbox"/>	2024-03-03/	—	Folder	—
<input type="checkbox"/>	parsed_resume.txt/	—	Folder	—
<input type="checkbox"/>	uploaded_resume_pdf/	—	Folder	—

api_jobs_data_json

Location: us (multiple regions in United States) | Storage class: Standard | Public access: Not public | Protection: None

OBJECTS | CONFIGURATION | PERMISSIONS | PROTECTION | LIFECYCLE | OBSERVABILITY | INVENTORY REPORTS

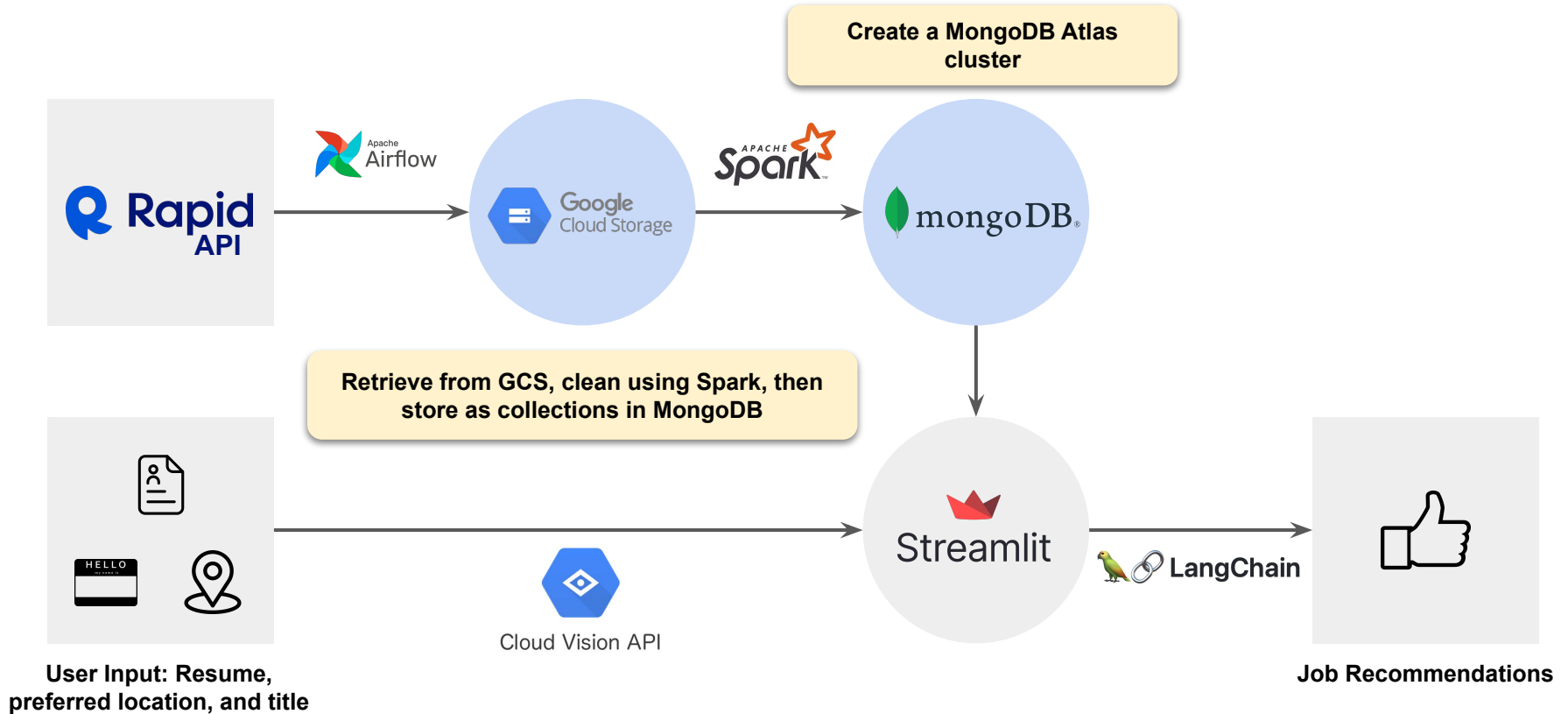
Buckets > api_jobs_data_json > 2024-03-03

UPLOAD FILES | UPLOAD FOLDER | CREATE FOLDER | TRANSFER DATA | MANAGE HOLDS | EDIT RETENTION | DOWNLOAD | DELETE

Filter by name prefix only | Filter | Filter objects and folders | Show deleted data

<input type="checkbox"/>	Name	Size	Type	Created	Storage class	Last modified	Public access	Version
<input type="checkbox"/>	DataAnalyst.json	1.4 MB	application/json	Mar 3, 2024, 12:58:51 PM	Standard	Mar 3, 2024, 12:58:51 PM	Not public	—
<input type="checkbox"/>	DataScientist.json	1.8 MB	application/json	Mar 3, 2024, 12:58:39 PM	Standard	Mar 3, 2024, 12:58:39 PM	Not public	—
<input type="checkbox"/>	MachineLearningEngineer.json	1.6 MB	application/json	Mar 3, 2024, 12:59:03 PM	Standard	Mar 3, 2024, 12:59:03 PM	Not public	—

Overview of Data Pipeline



Usage of MongoDB



Atlas | Li En Belinda... | Access Manager | Billing | All Clusters | Get Help | Li En Belinda

Project 0 | Data Services | App Services | Charts

Overview | DEPLOYMENT | Database | Data Lake | SERVICES | Device Sync | Triggers | Data API | Data Federation | Atlas Search | Stream Processing | Migration | SECURITY | Backup | Database Access | Network Access | Advanced | Goto

LI EN BELINDA'S ORG - 2024-02-20 > PROJECT 0 > DATABASES

MSDS697-Group16

VERSION 7.0.6 | REGION GCP Iowa (us-central1)

Overview | Real Time | Metrics

DATABASES: 1 | COLLECTIONS: 4

+ Create Database

Search Namespaces

- msds697-group16
 - demo
 - job_stats
 - jobs_data
 - new_jobs**

msds697-group16.new_jobs

STORAGE SIZE: 2.02MB | LOGICAL DATA SIZE: 3.64MB | TOTAL DOCUMENTS: 494 | INDEXES TOTAL SIZE: 36KB

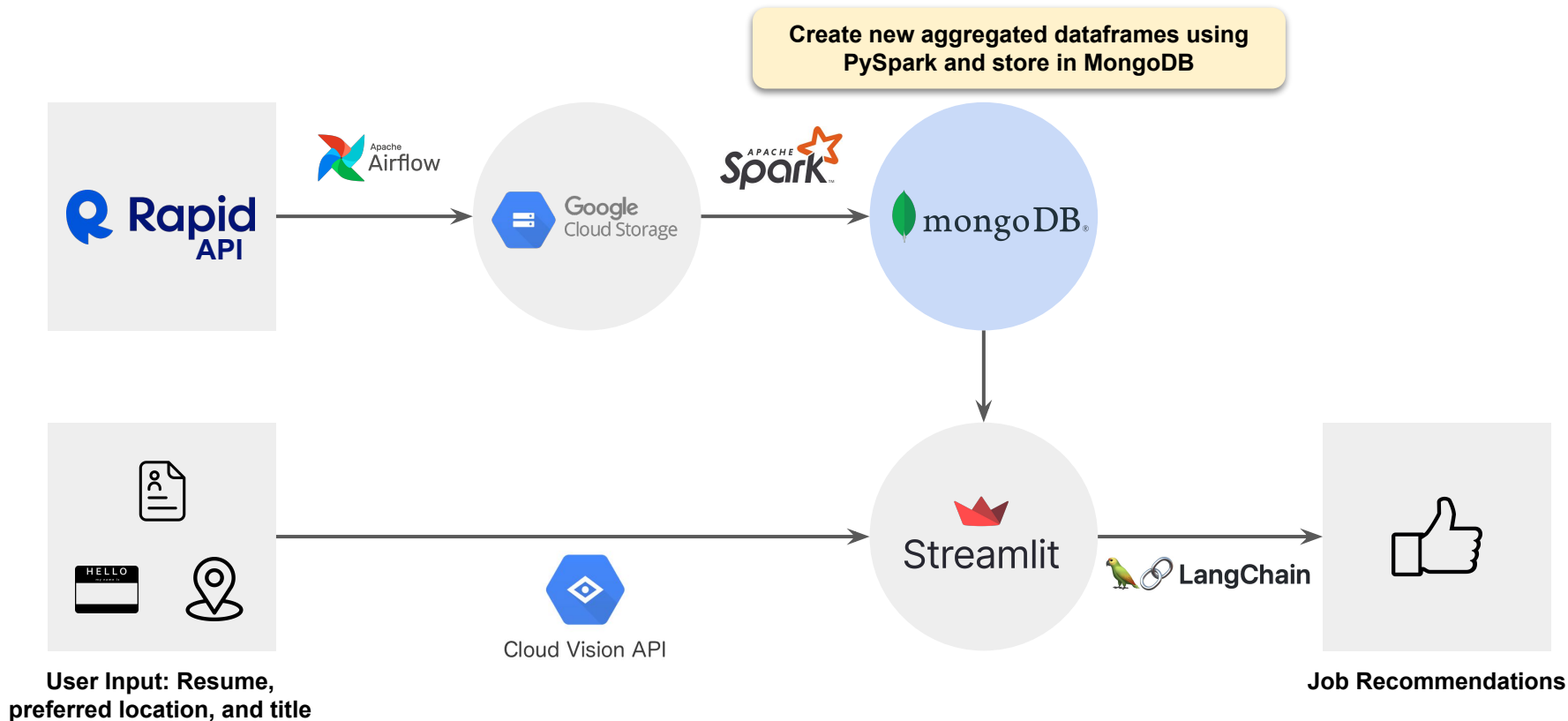
Find | Indexes | Schema Anti-Patterns 0 | Aggregation | Search Indexes

INSERT DOCUMENT

Filter | Type a query: { field: 'value' } | Reset | Apply | Options

```
_id: ObjectId('65e81d63d062144ada396d6b')
id: "2PQXj5Qn9PQCoLMQAAAAA=="
companyName: "Intuitive"
title: "Senior Data Scientist Product Analytics"
salary: "123740 - 189100"
jobUrl: "https://ai-jobs.net/job/130382-senior-data-scientist-product-analytics..."
location: "Sunnyvale - CA - US"
postedTime: "2024-02-08T20:43:18.000Z"
description: "Company DescriptionAt Intuitive, we are united behind our mission: we ..."
searchTitle: "Data Scientist"
clean_description: "compani descriptionat intuit unit behind mission believ minim invas ca..."
```


Overview of Data Pipeline



Usage of MongoDB



Atlas

Li En Belinda...

Access Manager

Billing

Project 0

Data Services

App Services

Charts

Overview

DEPLOYMENT

Database

Data Lake

SERVICES

Device Sync

Triggers

Data API

Data Federation

Atlas Search

Stream Processing

Migration

SECURITY

Backup

Database Access

Network Access

Advanced

Goto

LI EN BELINDA'S ORG - 2024-02-20 > PROJECT 0 > DATABASES

MSDS697-Group16

Overview

Real Time

Metrics

DATABASES: 1

COLLECTIONS: 4

+ Create Database

Search Namespaces

msds697-group16

demo

job_stats

jobs_data

new_jobs

msds697-group16.job_stats

STORAGE SIZE: 36KB

LOGICAL DATA SIZE: 354B

TOTAL DOCUMENTS: 3

INDEXES TOTAL SIZE: 36KB

Find

Indexes

Schema Anti-Patterns 0

Aggregation

Search Indexes

INSERT DOCUMENT

Filter

Type a query: { field: 'value' }

Reset

Apply

Options

_id: ObjectId('65e820f8689a8ba434c34683')

searchTitle: "Data Scientist"

total_jobs: 175

average_salary: 164946.41

city: 211

state: 39

_id: ObjectId('65e820f8689a8ba434c34684')

searchTitle: "Data Analyst"

total_jobs: 153

average_salary: 57989.94

city: 211

state: 39

_id: ObjectId('65e820f8689a8ba434c34685')

searchTitle: "Machine Learning Engineer"

total_jobs: 166

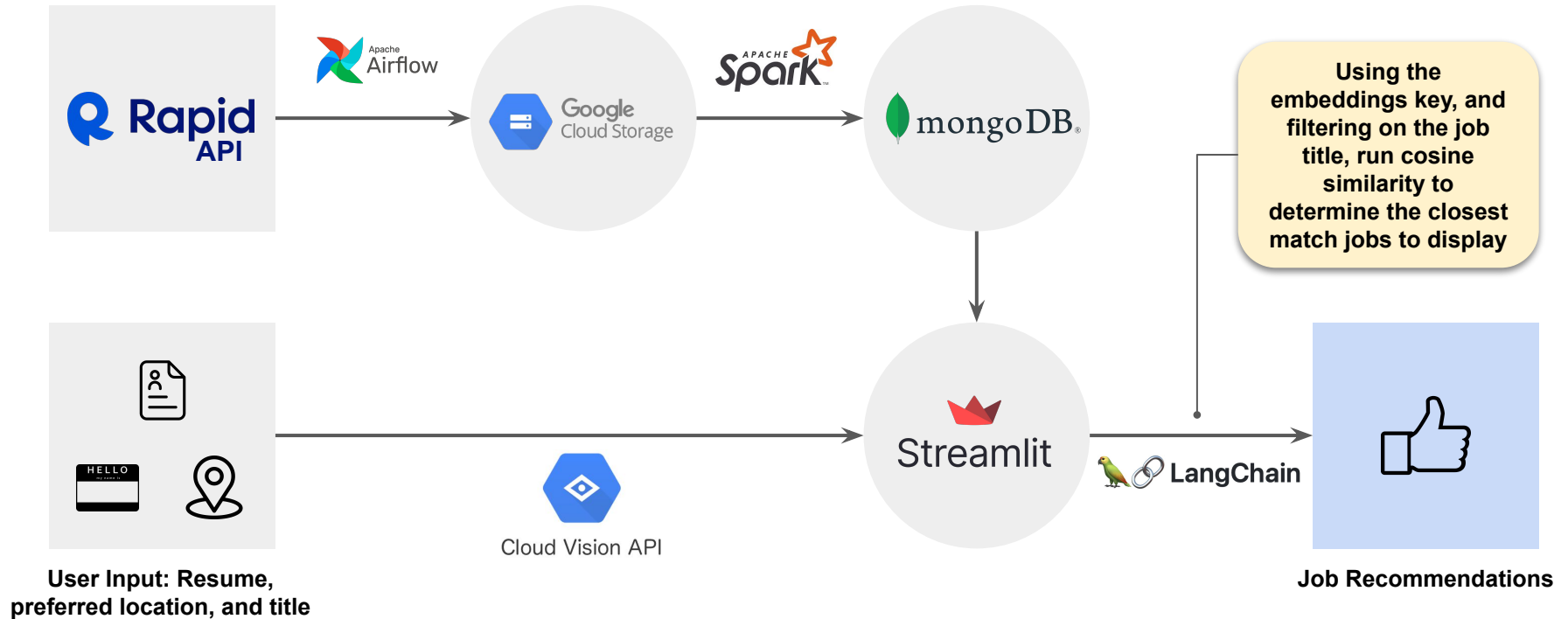
average_salary: 174939.26

Job Statistics

Streamlit


	Job Title	Total Number of Jobs	Average Salary	Total Number of Cities	Total Number of States
0	Data Scientist	175	164,946.41	211	39
1	Data Analyst	153	57,989.94	211	39
2	Machine Learning Engineer	166	174,939.26	211	39

Overview of Data Pipeline




Displaying the recommendations






UNIVERSITY OF
SAN FRANCISCO



[Github](#)

×

Deploy ⋮

 Param_Mehta_Resume.pdf 196.1KB

Find Jobs

Processing ...

Chewy

Machine Learning Engineer II

Similarity Score: 0.94

Location: Minneapolis - MN - US

Salary: None - None

Job URL: <https://careers.chewy.com/us/en/job/5772884/Machine-Learning-Engineer-II>

Matching Points:

- Both the resume and job description emphasize expertise in Machine Learning and Data Science, including data preparation, model building, and deployment.
- The resume showcases proficiency in natural language processing, time series analysis, and cloud computing (GCP), all of which are crucial skills mentioned in the job description.

Demo

[Github](#)

**Code base and
project report**

```
(truckx) parammehta@Params-MacBook-Air new %
```



A background image showing two people shaking hands in a professional setting. The person on the left is wearing a light-colored sweater, and the person on the right is wearing a patterned sweater. They are standing in front of a window. In the foreground, there is a laptop, a small potted plant, and two coffee cups.

MSDS 697 | SPRING 2024 | GROUP 16

Automated Data Pipeline: Job Recommender Engine

Shagun Kala | Param Mehta | Belinda Ong
Shnal Shad | Laila Zaidi

Generating the recommendations



Index Overview

This vector search index parses the data in **msds697-group16.jobs_data** and has the following configurations.

[Edit Index Definition](#)

```
1  {
2    "fields": [
3      {
4        "numDimensions": 768,
5        "path": "embedding",
6        "similarity": "cosine",
7        "type": "vector"
8      },
9      {
10       "path": "searchTitle",
11       "type": "filter"
12     }
13   ]
14 }
```

To determine the closest jobs to display

- Use the embeddings key
- Filter on job title
- Run cosine similarity

Usage of MongoDB



Li En Belinda... Access Manager Billing

All Clusters Get Help Li En Belinda

Project 0 Data Services App Services Charts

Overview DEPLOYMENT Database Data Lake SERVICES Device Sync Triggers Data API Data Federation Atlas Search Stream Processing Migration SECURITY Backup Database Access Network Access Advanced New On Atlas 4 Goto

LI EN BELINDA'S ORG - 2024-02-20 > PROJECT 0 > DATABASES

MSDS697-Group16

VERSION 7.0.6 REGION GCP IOWA (us-central1) CLUSTER TIER M0 Sandbox (General)

Overview Real Time Metrics Collections Atlas Search Profiler Performance Advisor Online Archive Cmd Line Tools

SANDBOX NODES REPLICA SET

Connect Configuration ...

TAGS

Use tags to efficiently label and categorize your clusters. Any tags you apply will display here.

[Learn more about tagging.](#)

ADD TAG

REGION IOWA (us-central1)

ac-xmr... shard-00-00.vnw6... SECONDARY

ac-xmrw... shard-00-01.vnw63... PRIMARY

ac-xmr... shard-00-02.vnw6... SECONDARY

This is a Shared Tier Cluster

If you need a database that's better for high-performance production applications, upgrade to a dedicated cluster.

Upgrade

Logical Size 12.1 MB

512.0 MB max

0.0 B

Last 30 Days

Operations R: 0 W: 0

0.2/s

0

Last 6 Hours

Connections 4

500 max

0

Last 6 Hours