Benchmarking suite

Steps:

Data Generation

Data Loading

Query generation (Kaplan-Meier and COX regression) and aggregation queries

Performance analysis and review

DataBase Setup

Used Docker Compose :

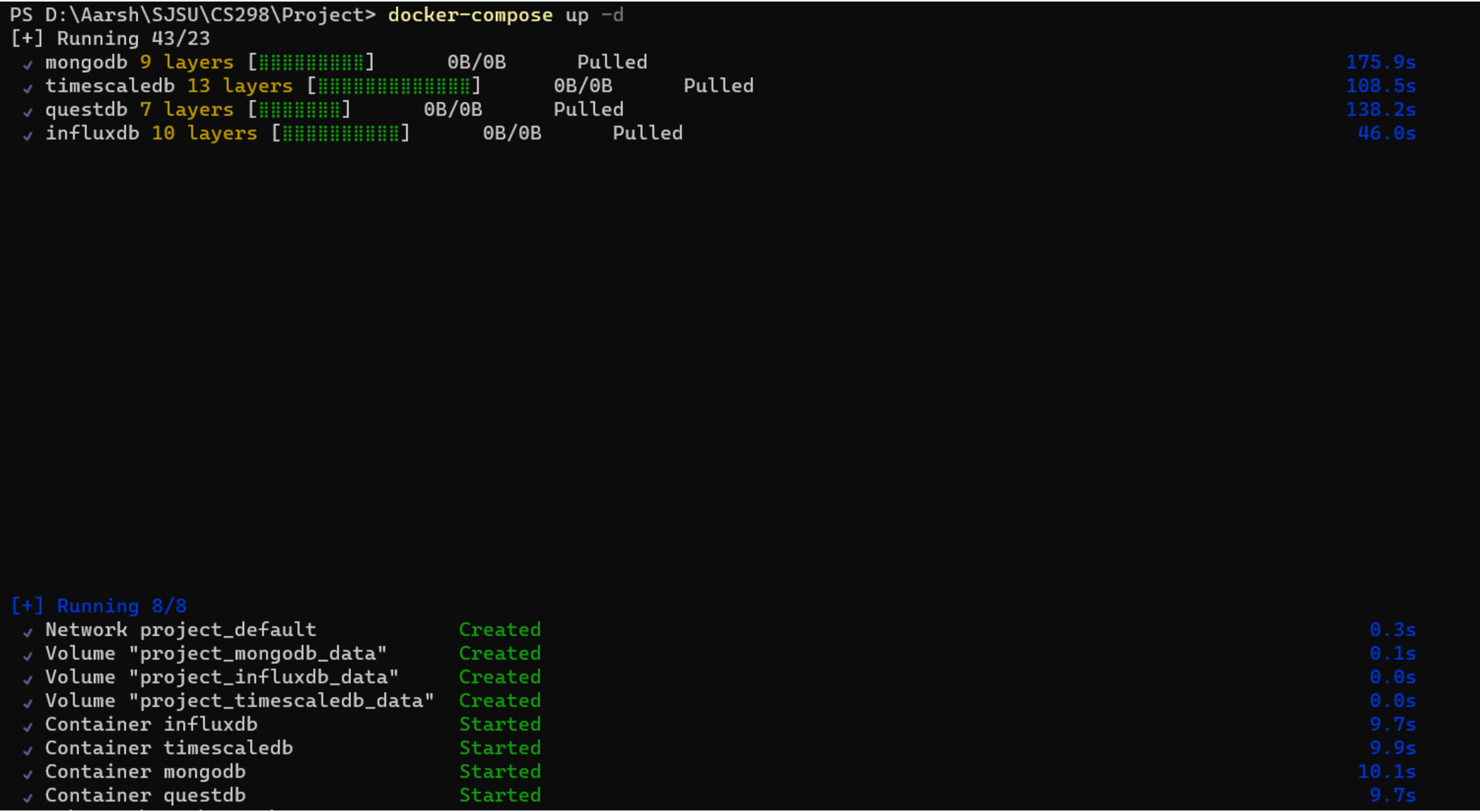
Docker-compose.yml file contains the necessary configurations like container,ports and environment information. This is to set up manually depending on needs . When this file is called with all or particular database , the docker container is created or started with the specified configuration.

Example :

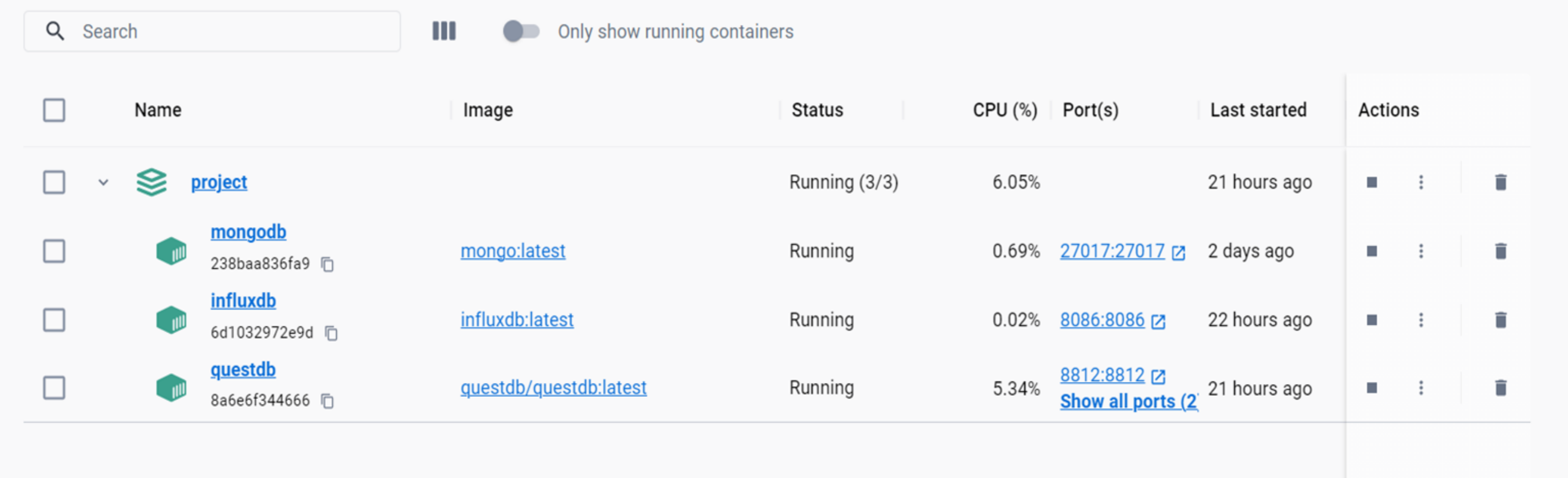
```
docker-compose up -d
```

or

```
docker-compose up -d mongodb
```



```
PS D:\Aarsh\SJSU\CS298\Project> docker-compose up -d
[+] Running 43/23
 ✔ mongodb 9 layers [⣿⣿⣿⣿⣿⣿⣿⣿⣿]      0B/0B      Pulled                                                 175.9s
 ✔ timescaledb 13 layers [⣿⣿⣿⣿⣿⣿⣿⣿⣿⣿⣿⣿⣿]      0B/0B      Pulled                                     108.5s
 ✔ questdb 7 layers [⣿⣿⣿⣿⣿⣿⣿]      0B/0B      Pulled                                                   138.2s
 ✔ influxdb 10 layers [⣿⣿⣿⣿⣿⣿⣿⣿⣿⣿]      0B/0B      Pulled                                               46.0s




















[+] Running 8/8
 ✔ Network project_default                Created                                                      0.3s
 ✔ Volume "project_mongodb_data"          Created                                                      0.1s
 ✔ Volume "project_influxdb_data"         Created                                                      0.0s
 ✔ Volume "project_timescaledb_data"      Created                                                      0.0s
 ✔ Container influxdb                     Started                                                      9.7s
 ✔ Container timescaledb                  Started                                                      9.9s
 ✔ Container mongodb                      Started                                                     10.1s
 ✔ Container questdb                      Started                                                      9.7s
```



□	Name	Image	Status	CPU (%)	Port(s)	Last started	Actions
□	project		Running (3/3)	6.05%		21 hours ago	■ ⋮ 🗑
□	mongodb 238baa836fa9	mongo:latest	Running	0.69%	27017:27017	2 days ago	■ ⋮ 🗑
□	influxdb 6d1032972e9d	influxdb:latest	Running	0.02%	8086:8086	22 hours ago	■ ⋮ 🗑
□	questdb 8a6e6f344666	questdb/questdb:latest	Running	5.34%	8812:8812 Show all ports (2	21 hours ago	■ ⋮ 🗑

Data Generation

Creating synthetic data with parameters to be tune for size and target database Modified the Kaggle synthetic dataset to take parameters.

https://www.kaggle.com/datasets/louise2001/survival-analysis-synthetic-data

Modified to be able to pass parameters

No of columns and target database

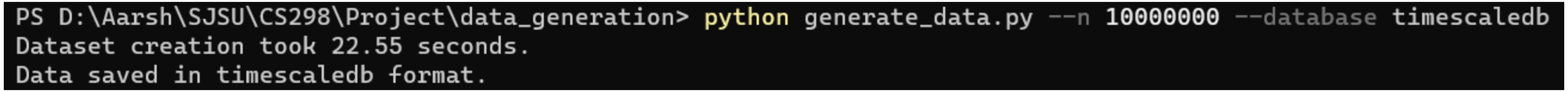
For this project,

If target_database = MongoDB ,then output is JSON for else CSV

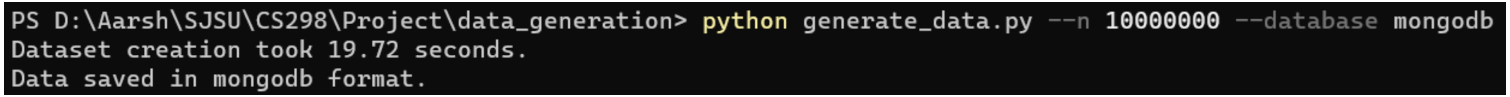
Example of data generation :

```
 python generate_data.py --n 1000 --database mongodb
```

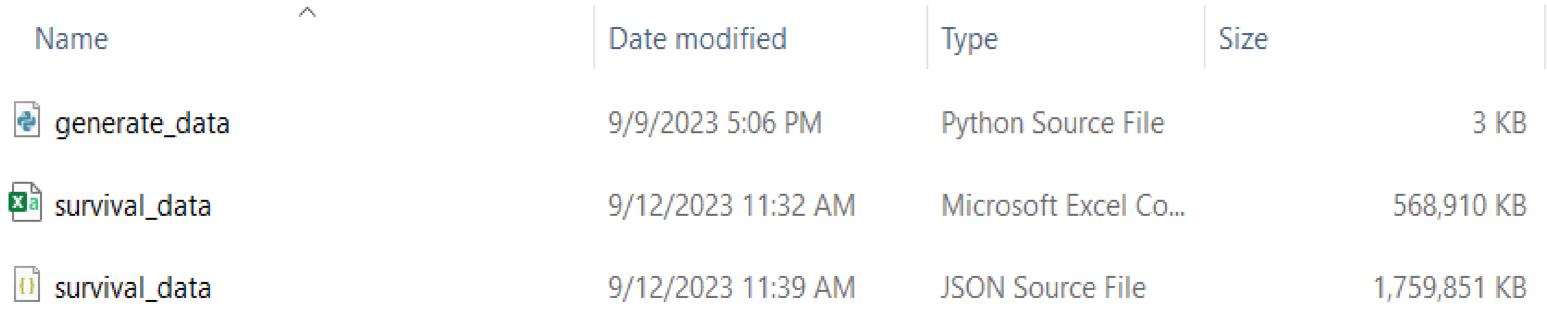
Example Output



```
PS D:\Aarsh\SJSU\CS298\Project\data_generation> python generate_data.py --n 10000000 --database timescaledb
Dataset creation took 22.55 seconds.
Data saved in timescaledb format.
```



```
PS D:\Aarsh\SJSU\CS298\Project\data_generation> python generate_data.py --n 10000000 --database mongodb
Dataset creation took 19.72 seconds.
Data saved in mongodb format.
```



Name	Date modified	Type	Size
generate_data	9/9/2023 5:06 PM	Python Source File	3 KB
survival_data	9/12/2023 11:32 AM	Microsoft Excel Co...	568,910 KB
survival_data	9/12/2023 11:39 AM	JSON Source File	1,759,851 KB

Kaplan-Meier Survival Analysis:

In Medical Research:

- The Kaplan-Meier method is used to measure the fraction of patients living for a certain amount of time after treatment.
- Calculate the Kaplan-Meier survival curves for data stored in each database. This involves calculating survival probabilities at different time points.
- Plot the Kaplan-Meier survival curves using a suitable visualization library (e.g., Matplotlib in Python). Each database performance can be evaluated by the quality of these plots.

Cox regression

In Medical Research:

- Cox regression is used to model the relationship between covariates (independent variables) and the hazard of an event occurring.This factors are basically affecting the survival like age, gender, treatment type.
- Construct Cox proportional hazards models using the data in each database.
- Compare the results of Cox regression models across databases.Estimated hazard ratios, confidence intervals can help in comparing the databases.