**FA23: DATA-225 Sec 11 - Db Systems for Analytics**

**Homework - - 4**

**Name :- Prayag Nikul Purani**

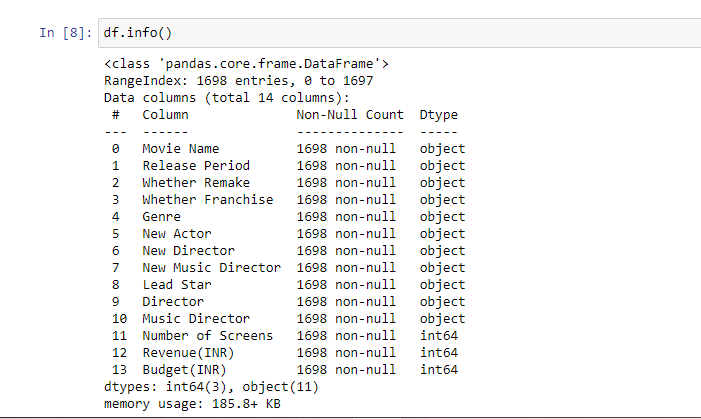
**SJSU Id :- 017416737**

**Dataset 1:**

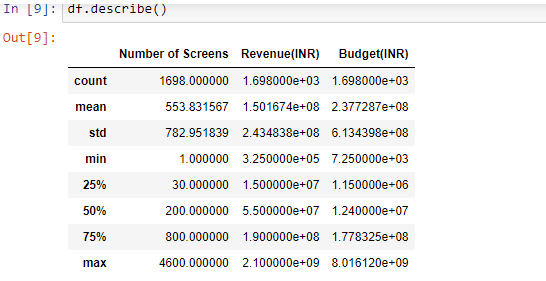
[**https://data.mendeley.com/datasets/3c57btcxy9/1**](https://data.mendeley.com/datasets/3c57btcxy9/1)

**Step 1: Cleaning the dataset**

So, first we will be importing the raw data to python and then see what could be done to it. After importing the data to python, we have divided the data into three csv files and renaming of some of the column names and then saving them to drive.

****

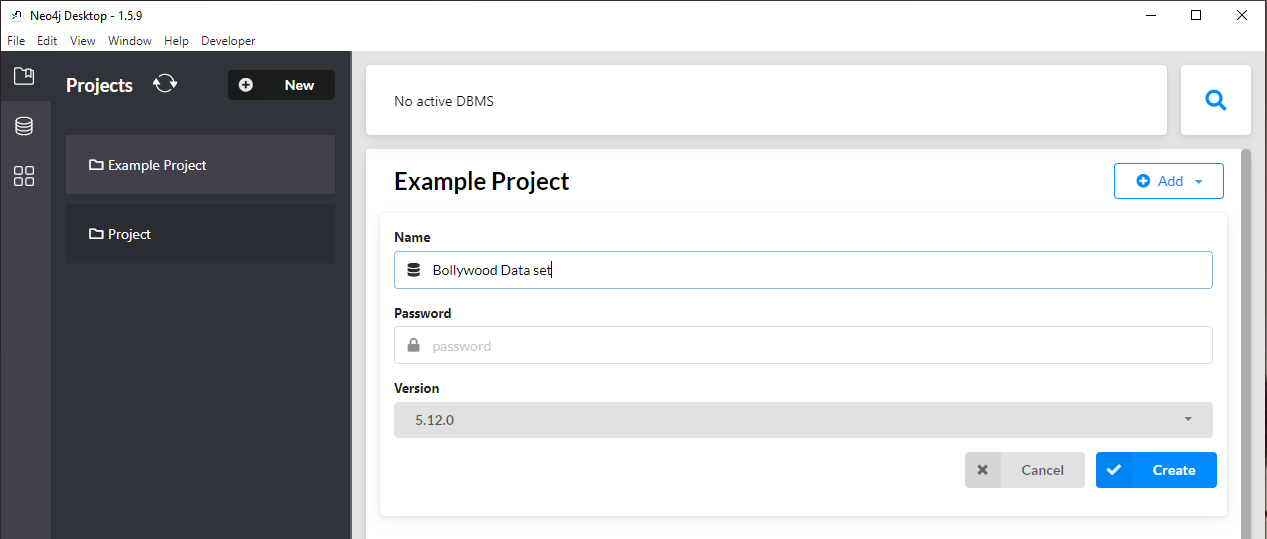
So we can see that the data is cleaned and has no null values and we can proceed with further steps and which is dividing the main data into 3 different csv files and renaming some of the columns



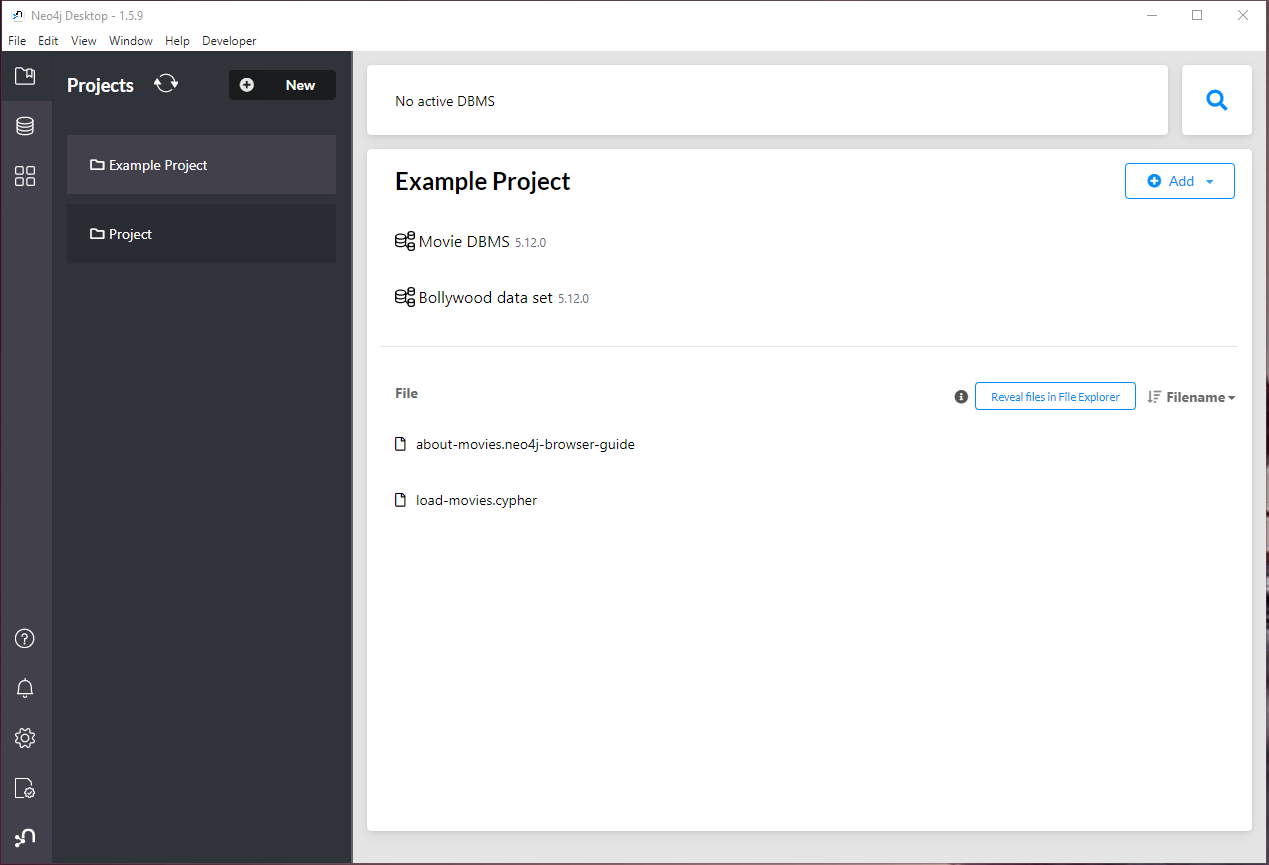
Now we will be performing some of the graph for knowing distribution



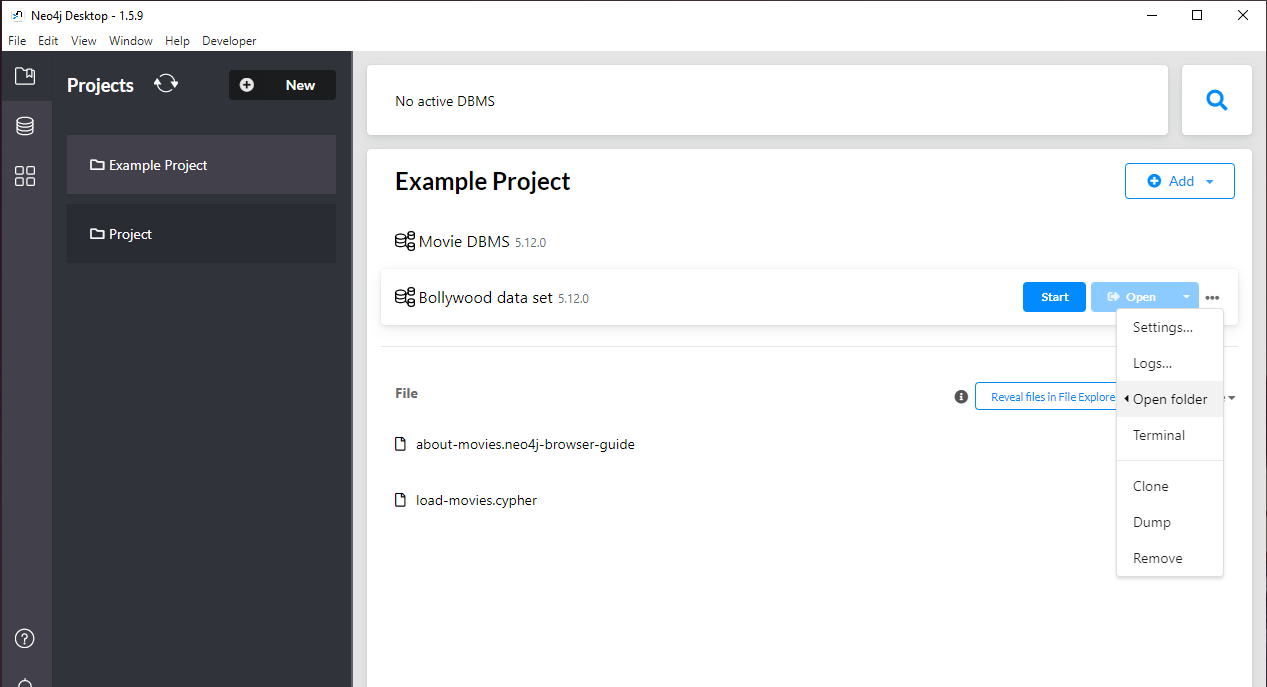
**Step 2: Importing the dataset to neo4j**

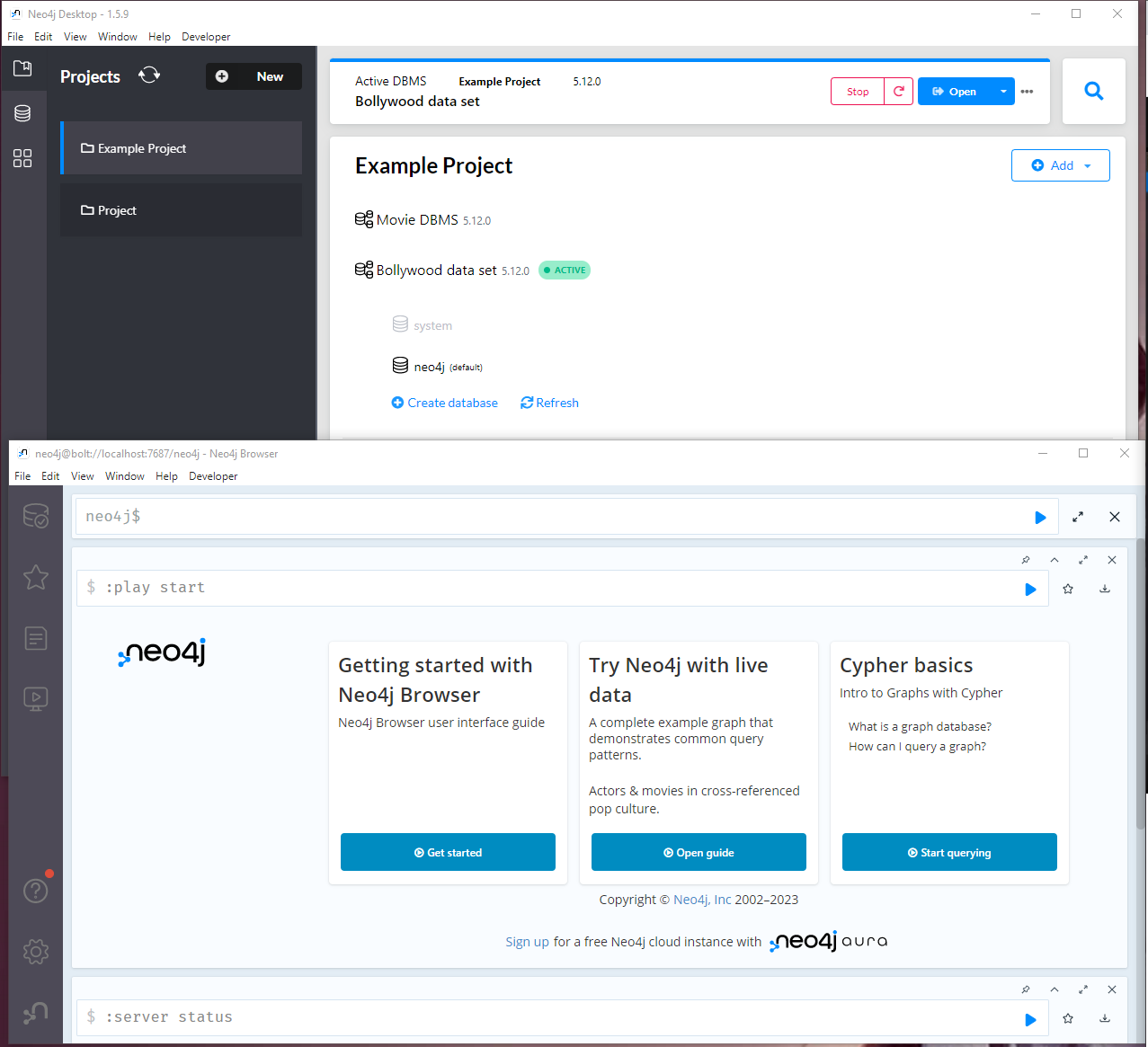
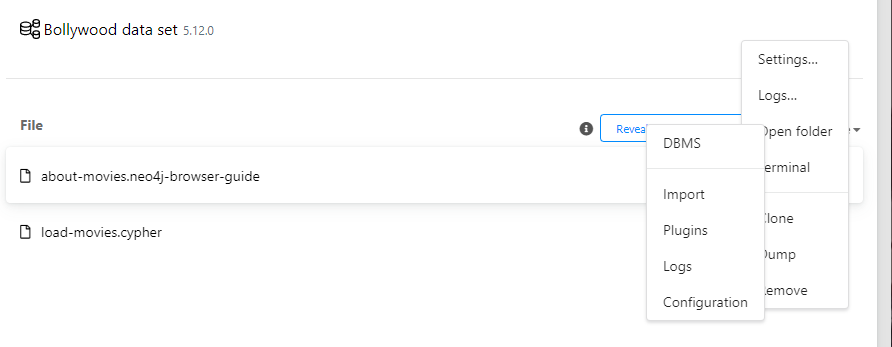


First we need to create the project and set the password in neo4j application after that we can go on with the other steps.



After creating the set we can see that it is displayed on the screen and after that we can proceed further by first opening the connection and then opening the neo4j browser in that the next step is to load the data from the csv file and applying the modelling for each file as we have 3 different csv file.

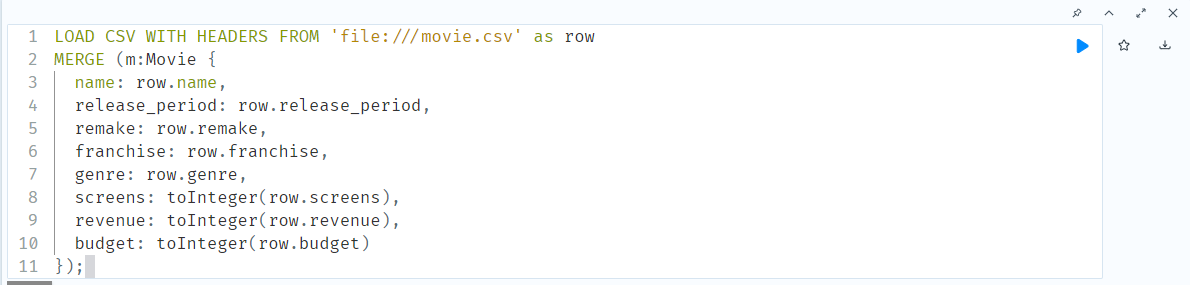




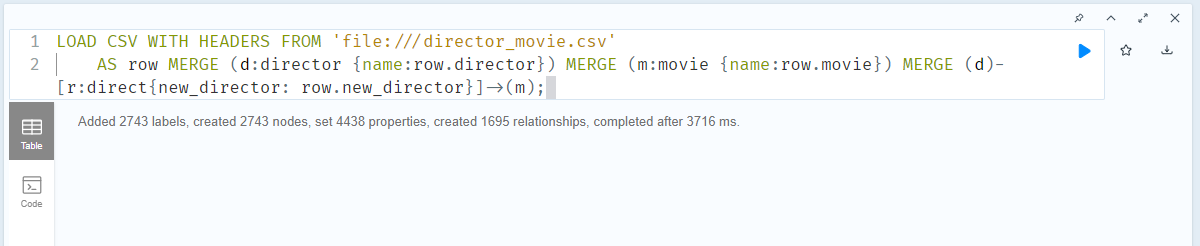
Now we have pasted the 3 files in to the neo4j folder now what we need to do is to model the data

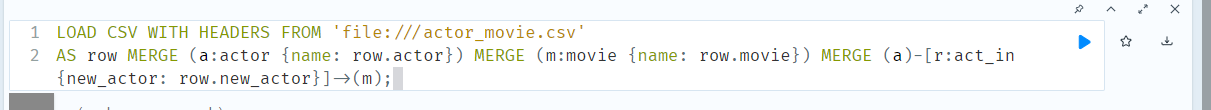
Launch the Neo4j Desktop. In the import folder of the project named "bollywood," place the three CSV files.

Next, launch Neo4j Browser and type the following instructions. The information will be imported into the database (for comprehensive details, refer to the section titled "2. Import data into Neo4j").

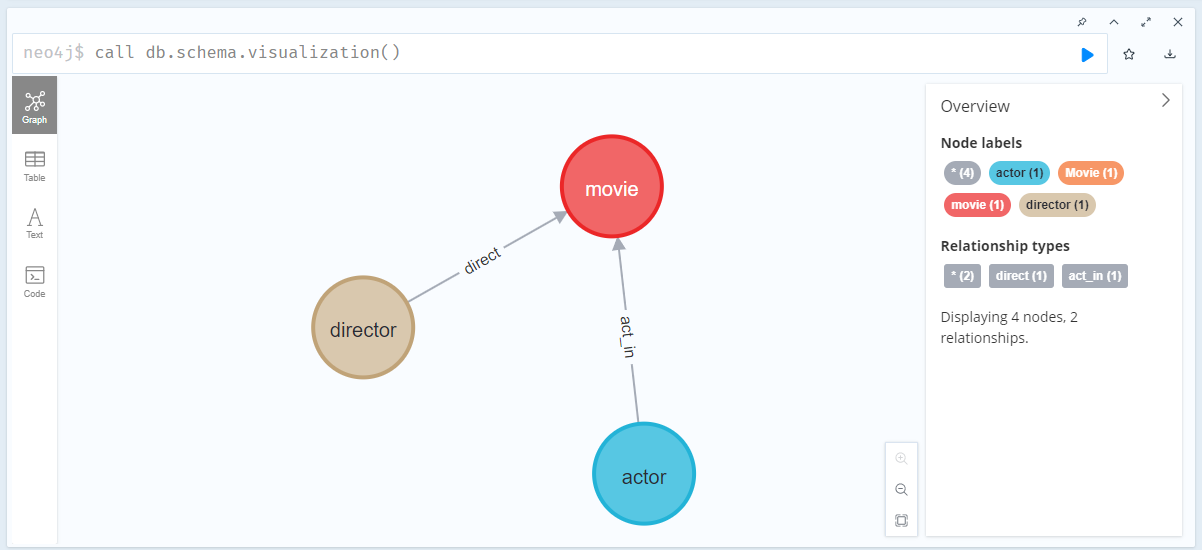


Now we are starting with the modeling of the data. So, first we will be importing the first csv file which is movies and the properties of this nodes are



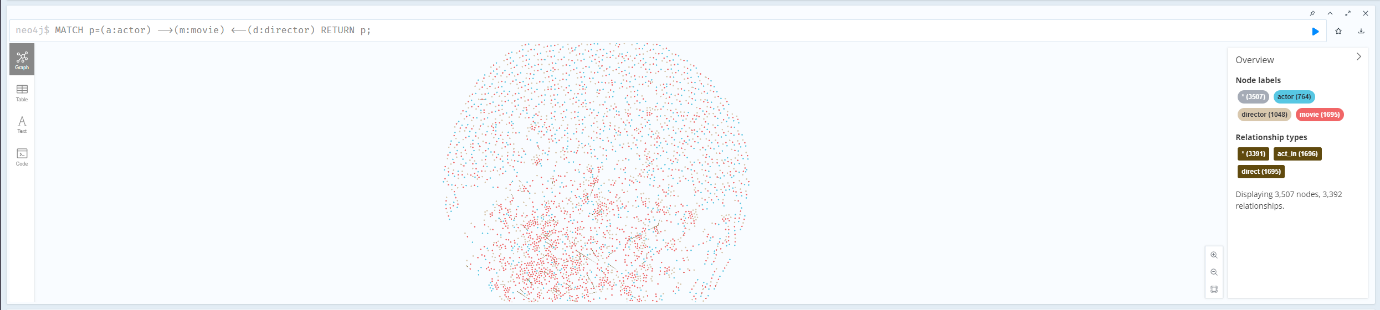


Now we have imported the rest two files and we have created the relationship between the nodes now the next step is to see whether what we have computed is correct or not so we will call db.schema.visualization is will plot the graph model for us and from it we can see whether what we have done is correct or to added something .



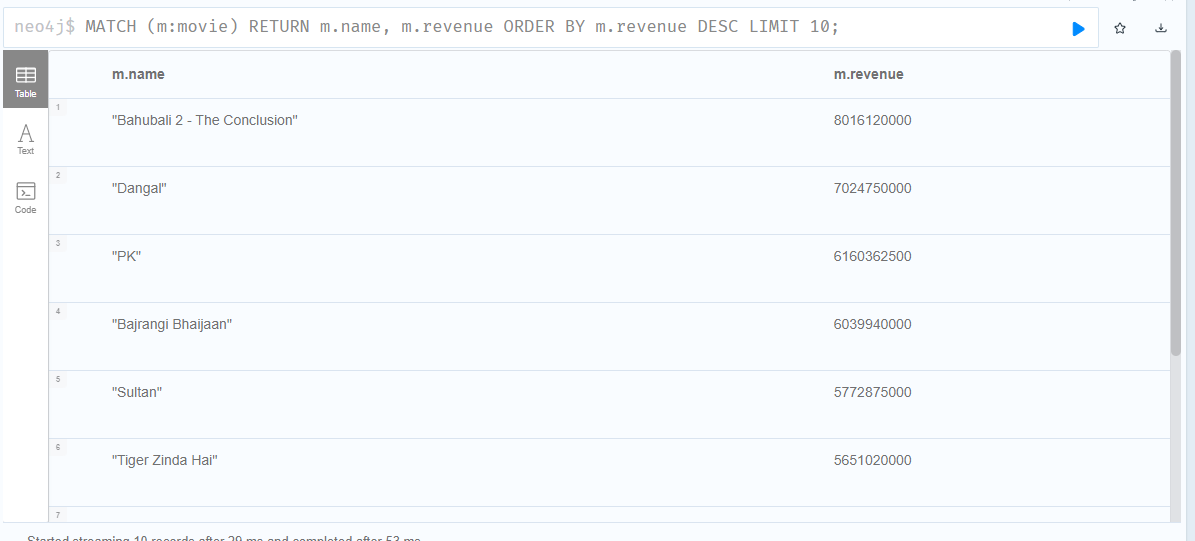
So, from the above graph we can see that the model is correct this what we needed to make and we got it.

Now we will see how our graph model with all the nodes will look. So, for this we need to update the setting of neo4j to display 3600 nodes instede of 500 which is the default setting so we will br able to see all the nodes present in the project.



Now we will do some queries

1. Finding top 10 movies with highest revenue.



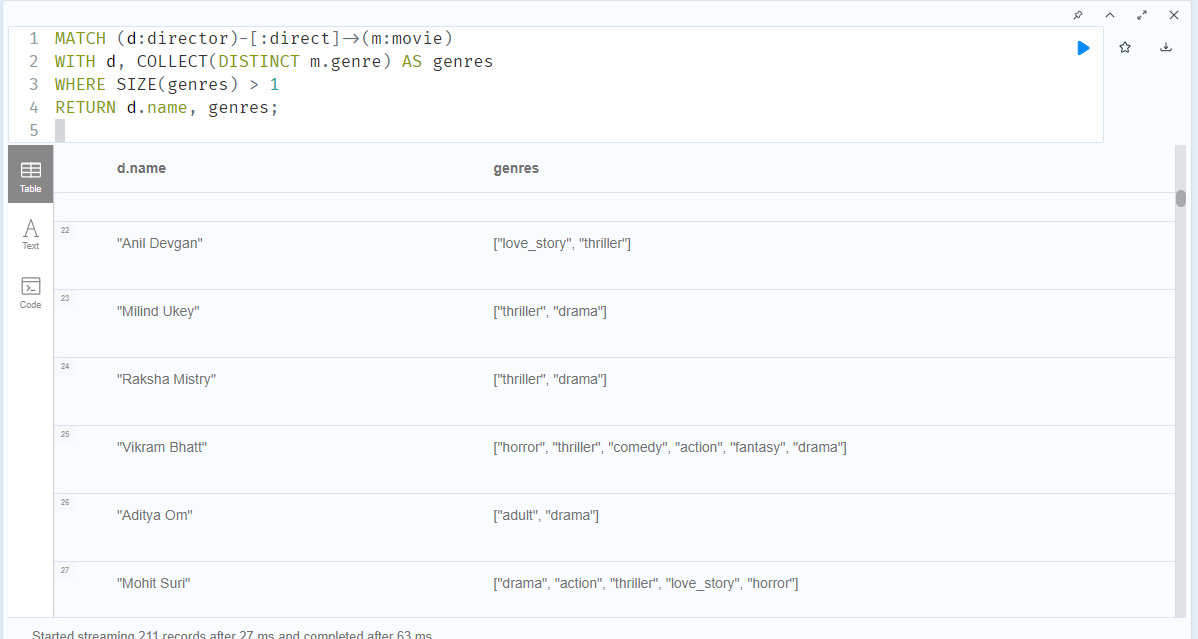
1. What if we need to find the different genre count present in the dataset.



1. Find movies that are part of the same franchise and list them in chronological order:



1. Find directors who have directed movies in multiple genres:



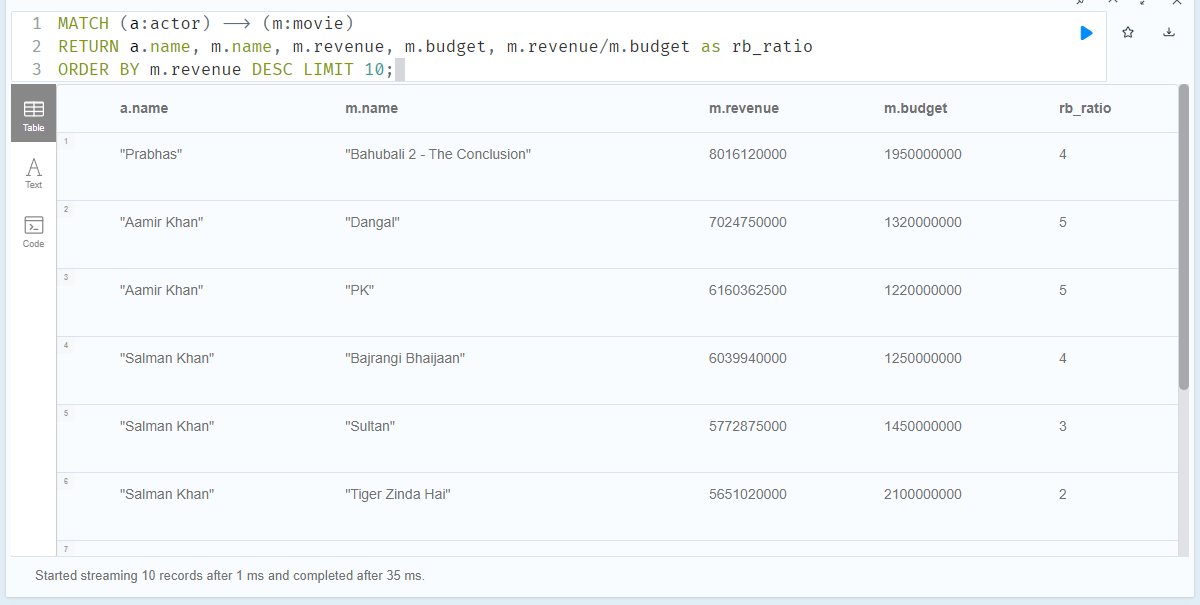
1. Find the average budget and revenue for movies in each genre:



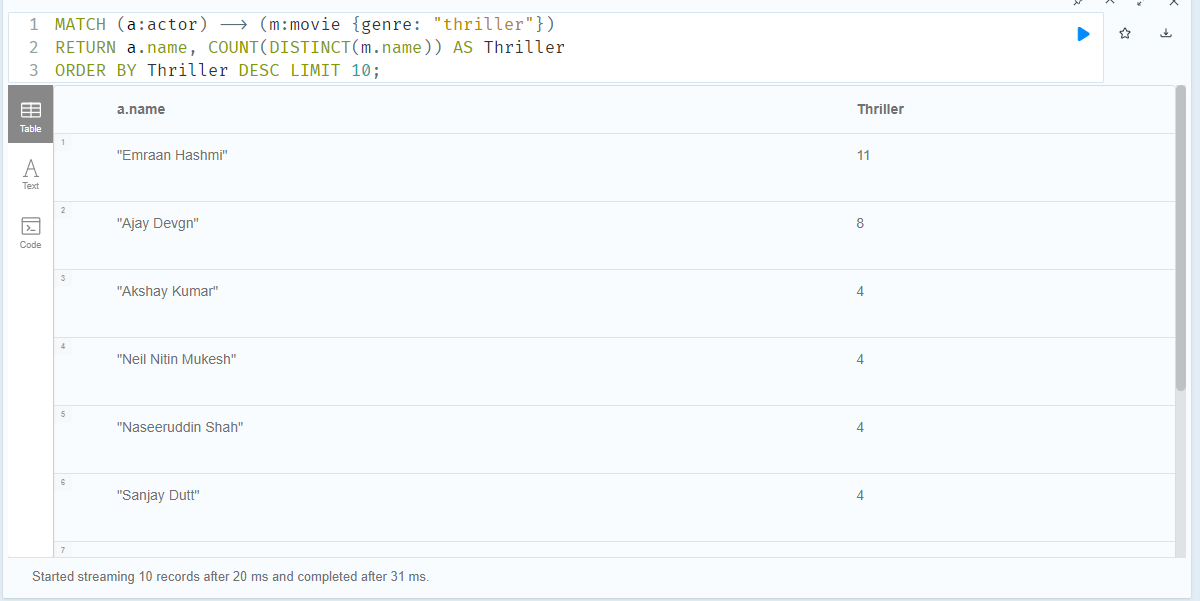
1. Find the most common actor pairs in movies directed by the same director:



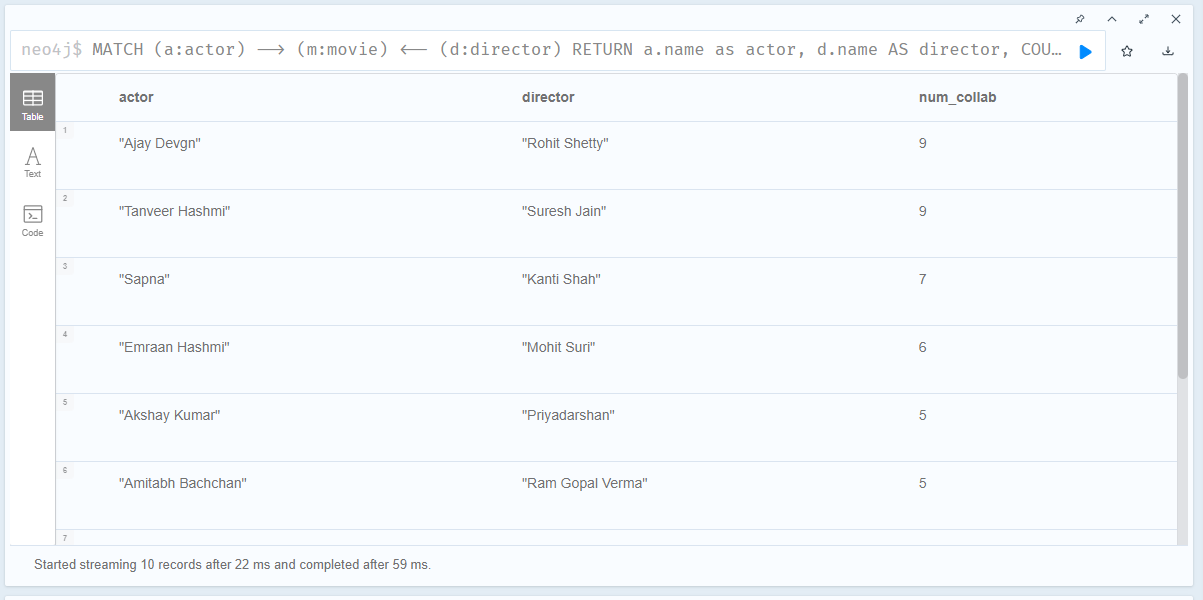
1. Top 10 movies with its details.



1. Which actors have the most thriller movies genre.



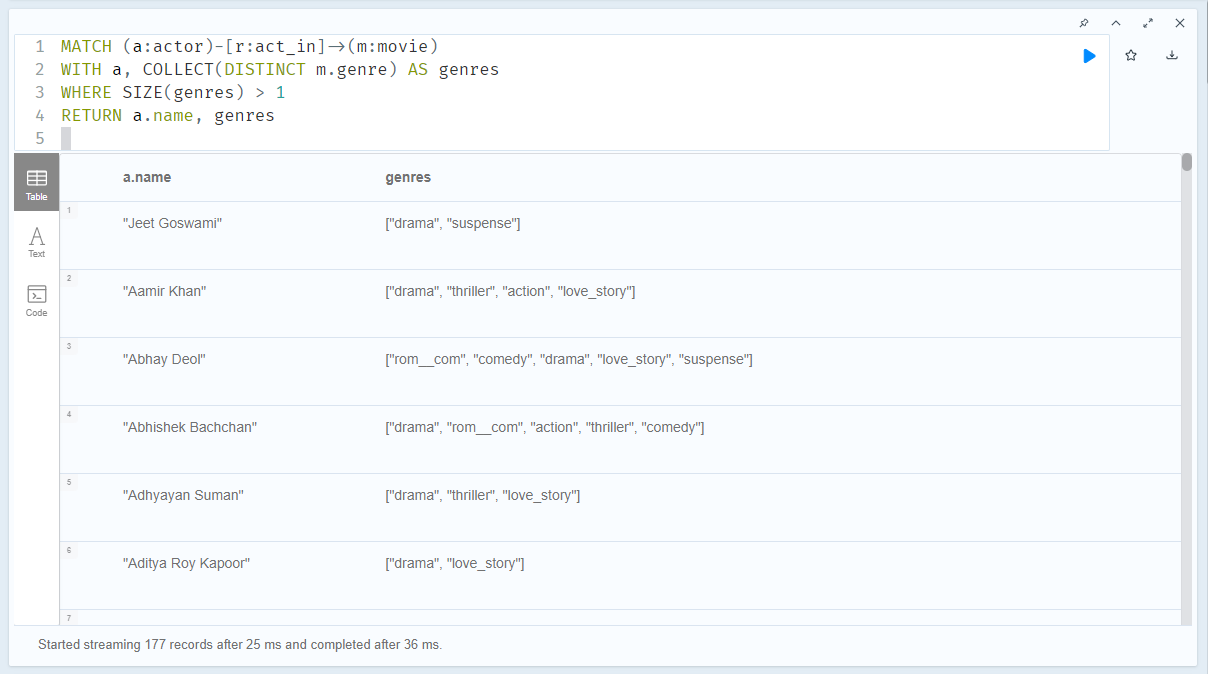
1. How often some actor-director duo worked together.



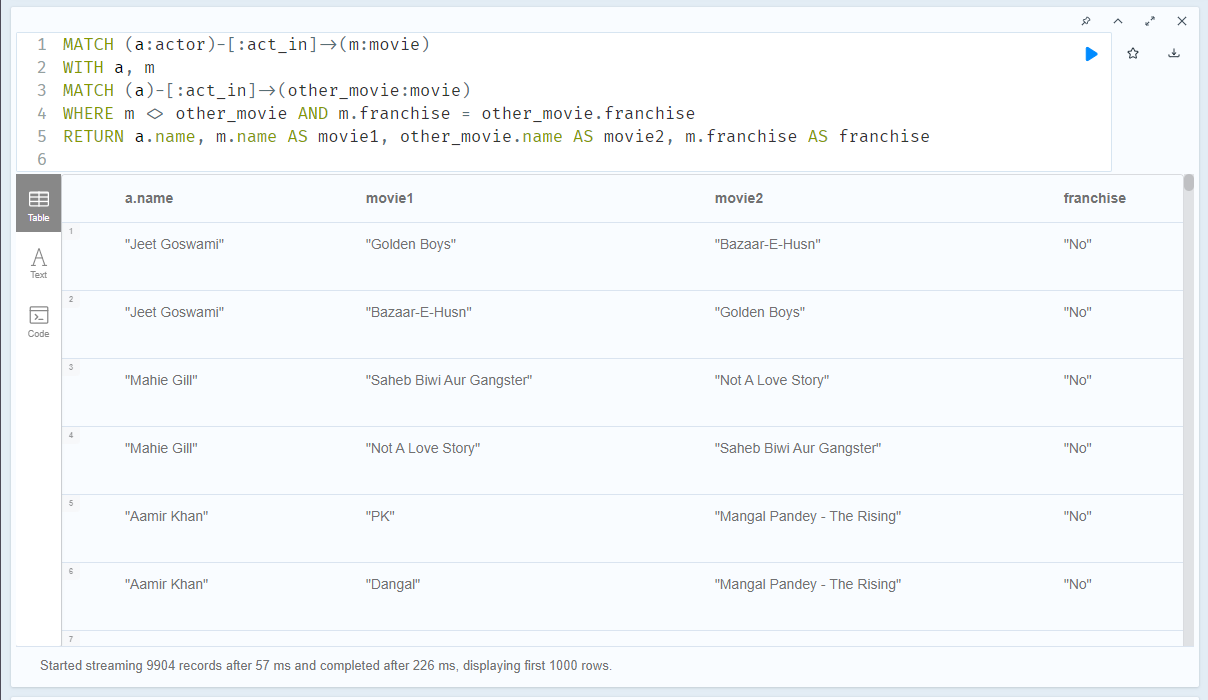
1. Find the most profitable actors: Identify actors who have appeared in movies with the highest revenue and calculate their total earnings from those movies.



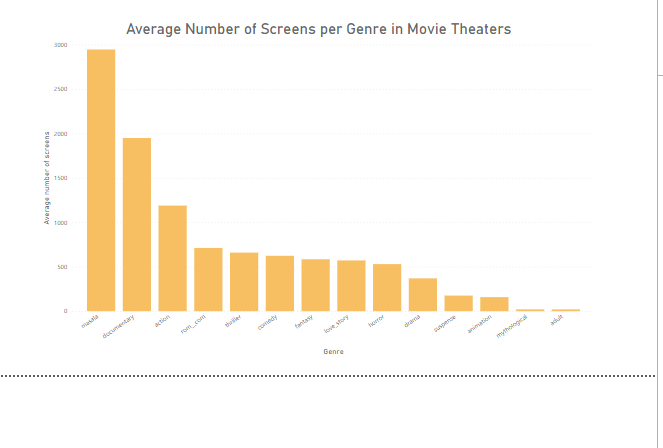
1. Find actors who have worked in different genres and their diversity: Identify actors who have appeared in movies from multiple genres and calculate the diversity of genres they've worked in.

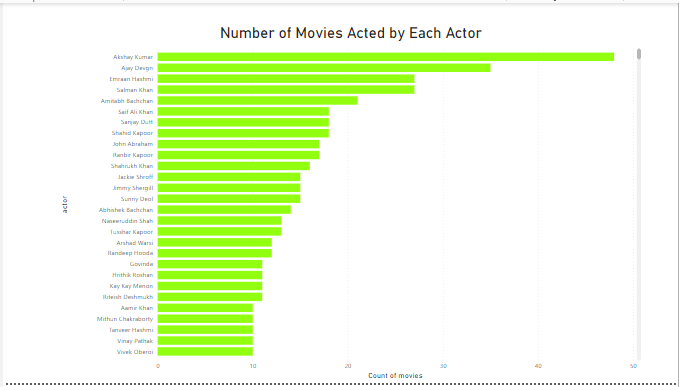


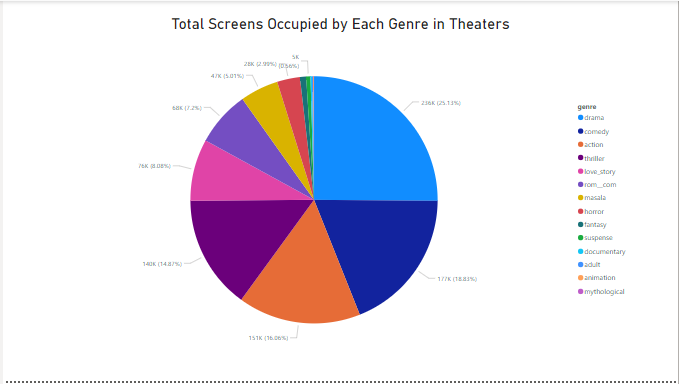
1. Find actors who have appeared in movies of the same franchise: Identify actors who have acted in multiple movies from the same franchise and list those movies.

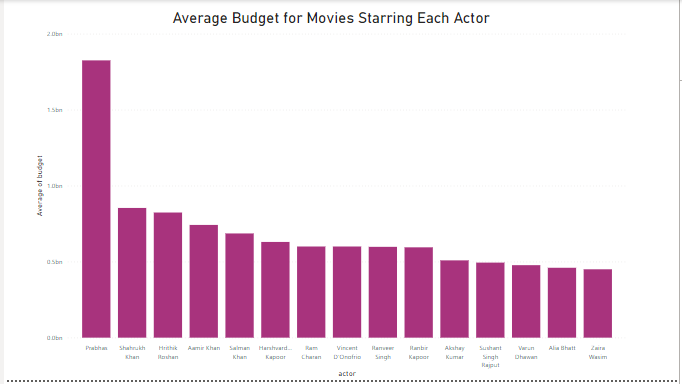


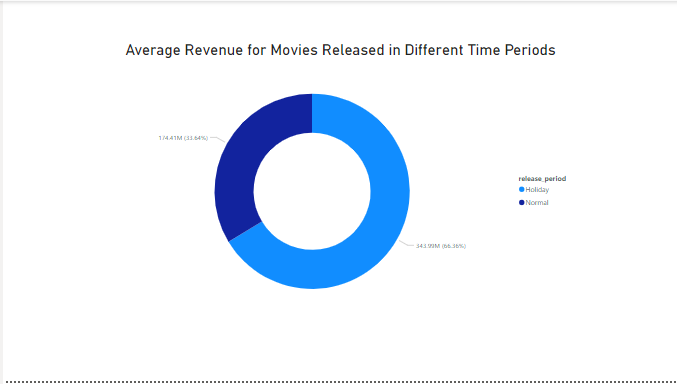
**Visualization:**

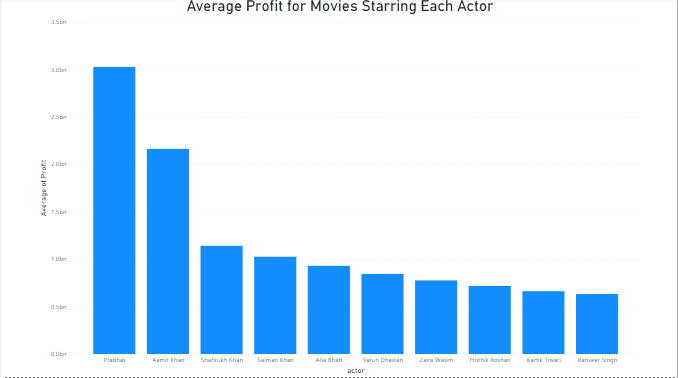
****

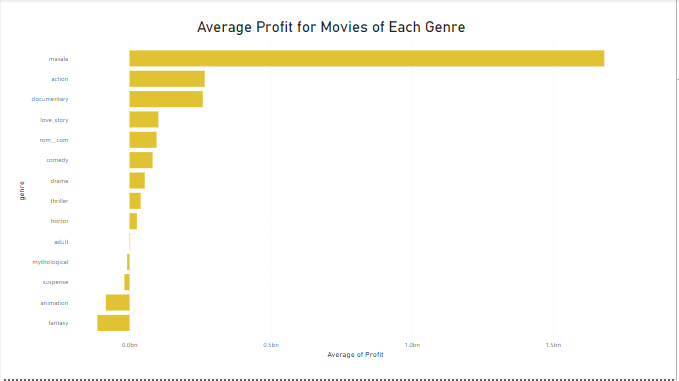
****

****

****

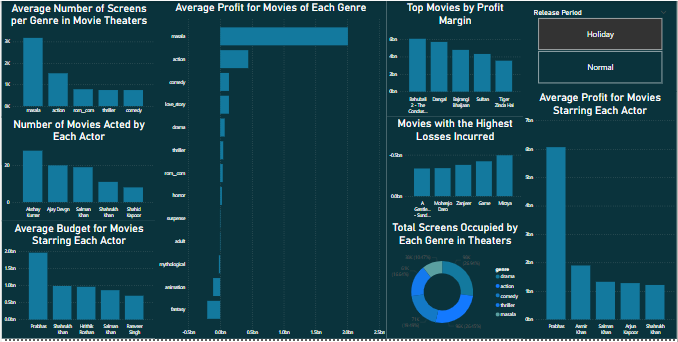
****

****

****

****

**The Final Dashboard:**

****

**References:**

<https://www.youtube.com/watch?v=J8vmqJrqd6w>

<https://medium.com/geekculture/analyzing-genomes-in-a-graph-database-27a45faa0ae8>

<https://enterprise-docs.anaconda.com/en/latest/data-science-workflows/data/neo4j.html>

<https://anaconda.org/conda-forge/neo4j-python-driver>

<https://towardsdatascience.com/neo4j-for-bollywood-5ceb371031f1>

<https://neo4j.com/docs/graph-data-science/current/installation/>

<https://community.neo4j.com/t/install-graph-data-science-in-neo4j-version-4/16539>

<https://www.youtube.com/watch?v=lmPS_UxF5Qw>

<https://www.youtube.com/watch?v=EDav6TfAOUs>

<https://www.youtube.com/watch?v=yluHRteVBNI&t=376s>