

GraPH THEORY AND ITS APPLICATIONS – ASSIGNMENT 4

Bollywood using neo4j



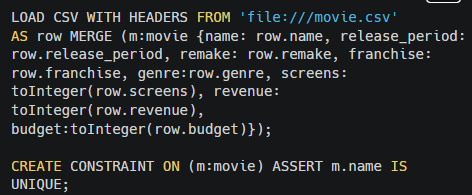
November 16, 2022

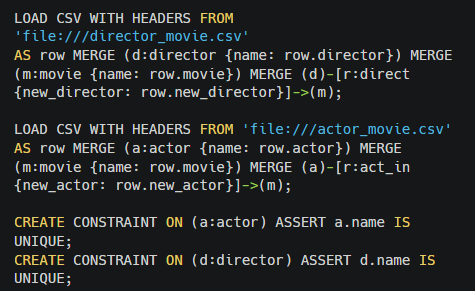
****

The dataset is from [https://boxofficeindia.com](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwie1Jq5y4fzAhXEg-AKHZYKC7MQFnoECAMQAQ&url=https%3A%2F%2Fboxofficeindia.com%2F&usg=AOvVaw33S6bzbkk-A0LfEboTANRv)

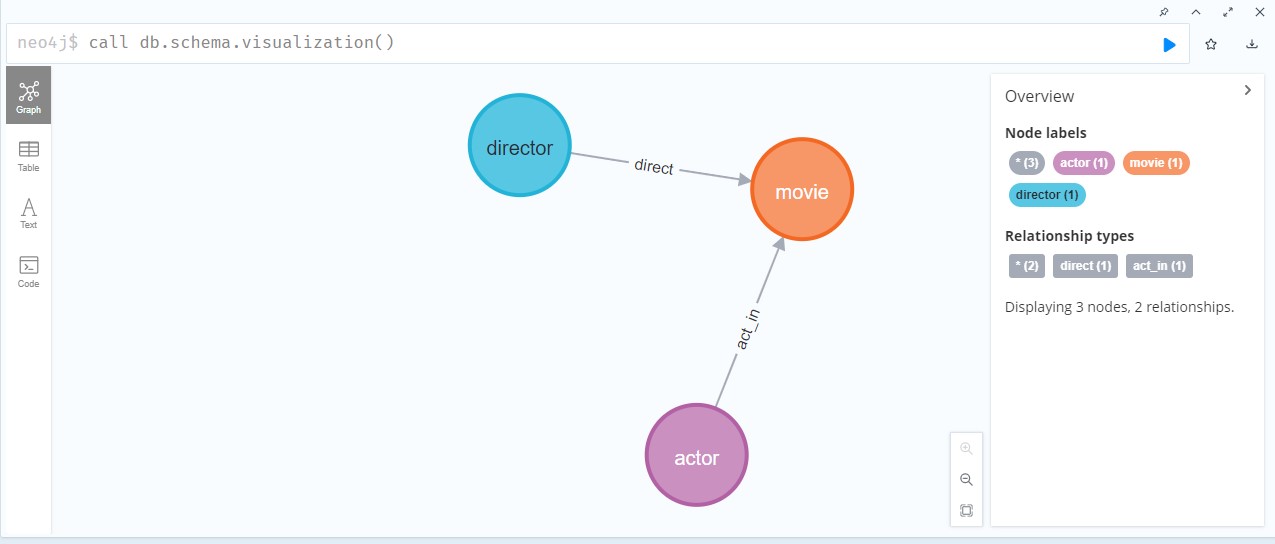
|  |  |  |
| --- | --- | --- |
| A.Prashanth | PES1UG20CS525 | Sec: I |
| Prateek Keshri | PES1UG20CS526 | Sec: I |

**1. Prepare and import the data**

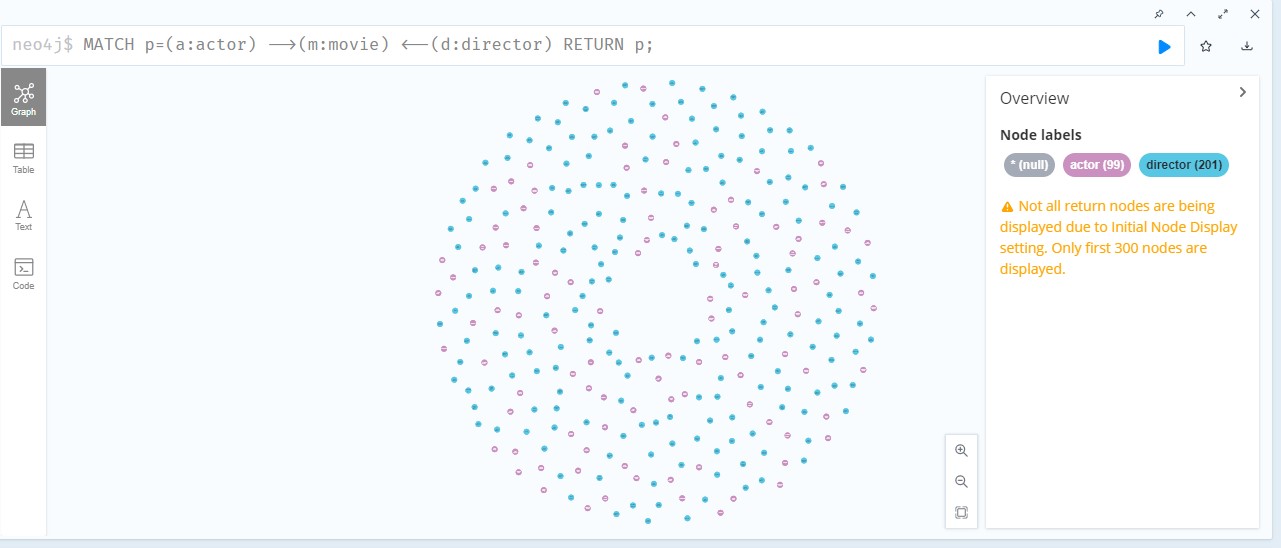
****

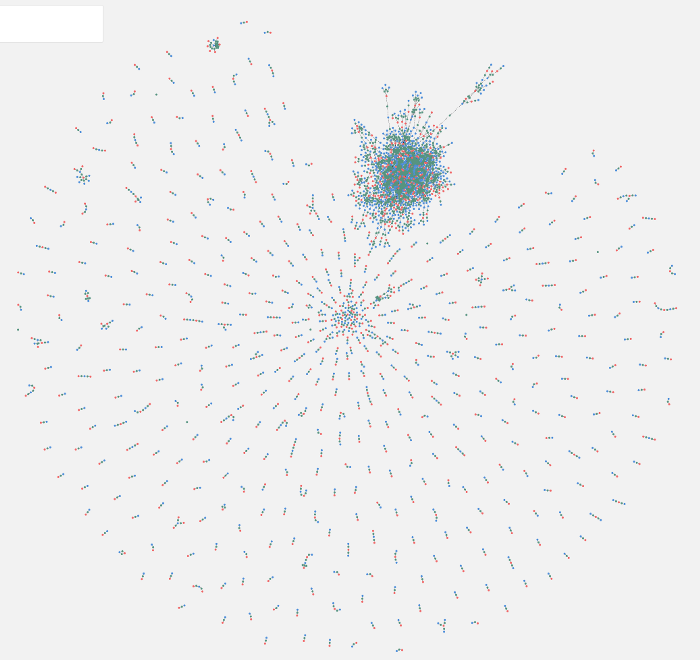
****

we can check how the three types of nodes (actor, director and movie) are connected:

 call db.schema.visualization()

# 2. Overview

The topological overview of the Bollywood dataset. Green: movie; Red: actor; Blue: director. Image by author.



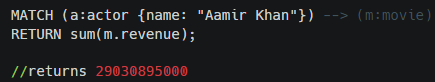
It is clear that this large cluster is held together by prolific Bollywood actors such as [the Three Khans](https://en.wikipedia.org/wiki/Khans_of_Bollywood), [Akshay Kumar](https://en.wikipedia.org/wiki/Akshay_Kumar" \t "_blank), [Emraan Hashmi](https://en.wikipedia.org/wiki/Emraan_Hashmi) and directors such as [Vikram Bhatt](https://en.wikipedia.org/wiki/Vikram_Bhatt) and [Mohit Suri](https://en.wikipedia.org/wiki/Mohit_Suri).

We can also have a look at the genre distribution in the dataset:



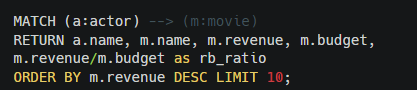
The documentation from Neo4j explains that [WITH is like the pipe operator](https://neo4j.com/docs/cypher-manual/current/clauses/with/) in Bash. In the query above, I calculated the total number of films with the WITH statement and then used it in the percentage calculation. The results suggested that dramas, comedies, and thrillers are the top three movie genres in Bollywood.

# 3. Get some statistics



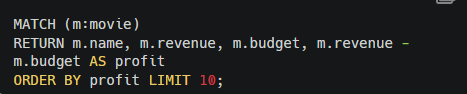


we can see the top earners:

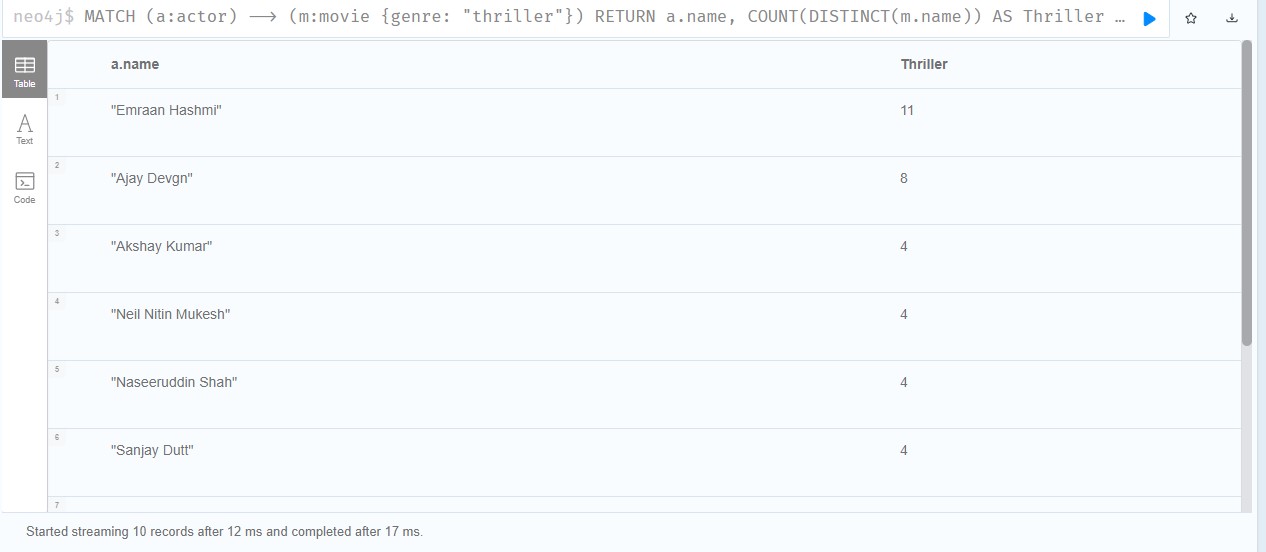




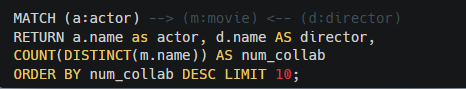
We can see the list of the biggest box-office bombs:

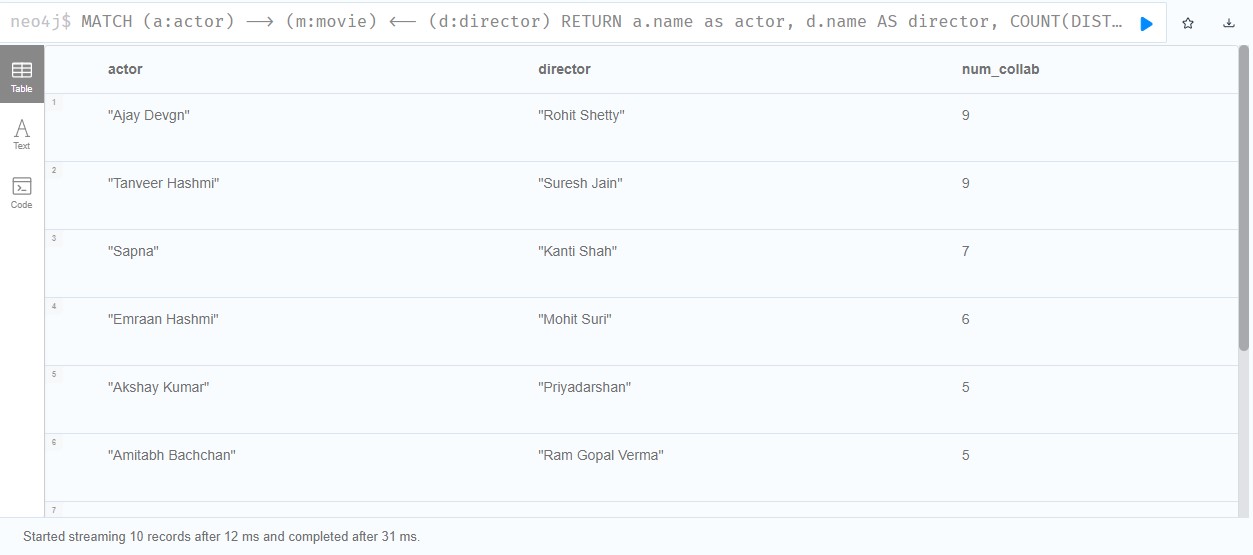




Afterwards, we can also see which actors have the most thrillers or horror movies under their belt.

Finally, let’s see how often some actor-director duo worked together.

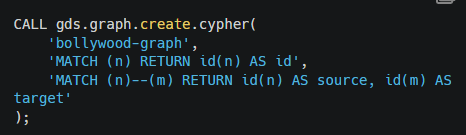




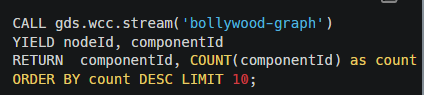
# 4. Community detection

the topological overview by Neo4j Bloom has shown us a large cluster centered around some of the biggest names in Bollywood. Now the question is whether it is possible to obtain that community through a query? With the help of Neo4j’s Graph Data Science Library (GDS), the answer is a resounding “Yes”.

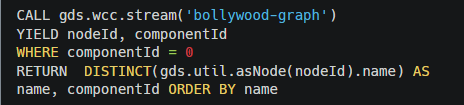
To use WCC, you need to enable the GDS plugin in our project

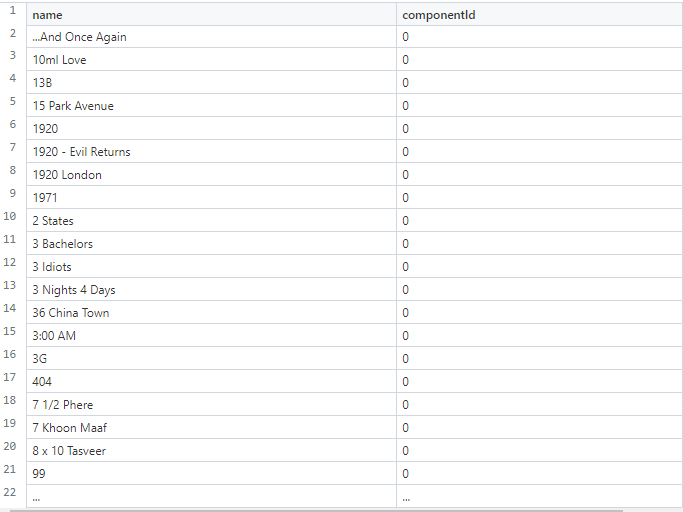


Afterwards, we run WCC. It returns the top 10 largest communities (called “components” in WCC):

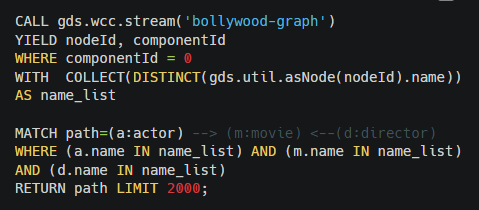


We can see that the Component 0 has as many as 1750 nodes. However, I have found that some nodes were double-counted. Let’s see what the nodes are with the DISTINCT function:





The query returns 1,734 nodes instead of the original 1,750. We can even show the networks in Neo4j Browser. But first, adjust the visualization parameters in Neo4j Browser



The COLLECT function transforms the names into a list. We then do a normal MATCH query and filter the results with that list.



**CONCLUSION:**

This project shows that Neo4j not only excels in relation-rich data, but it can also aggregate tabular data as easily as SQL. This tutorial showcases primarily the statistical functions in Neo4j. Readers can easily compare the SQL queries with my Cypher queries above and feel the differences in the syntax and the expressiveness of the two languages. According to Wikipedia, the syntax of Cypher is based on ASCII-art. So the queries look very visual and are easy to understand. Combined with its aggregating and Graph Data Science functions, Neo4j can do just anything that a relational database can do and then some.