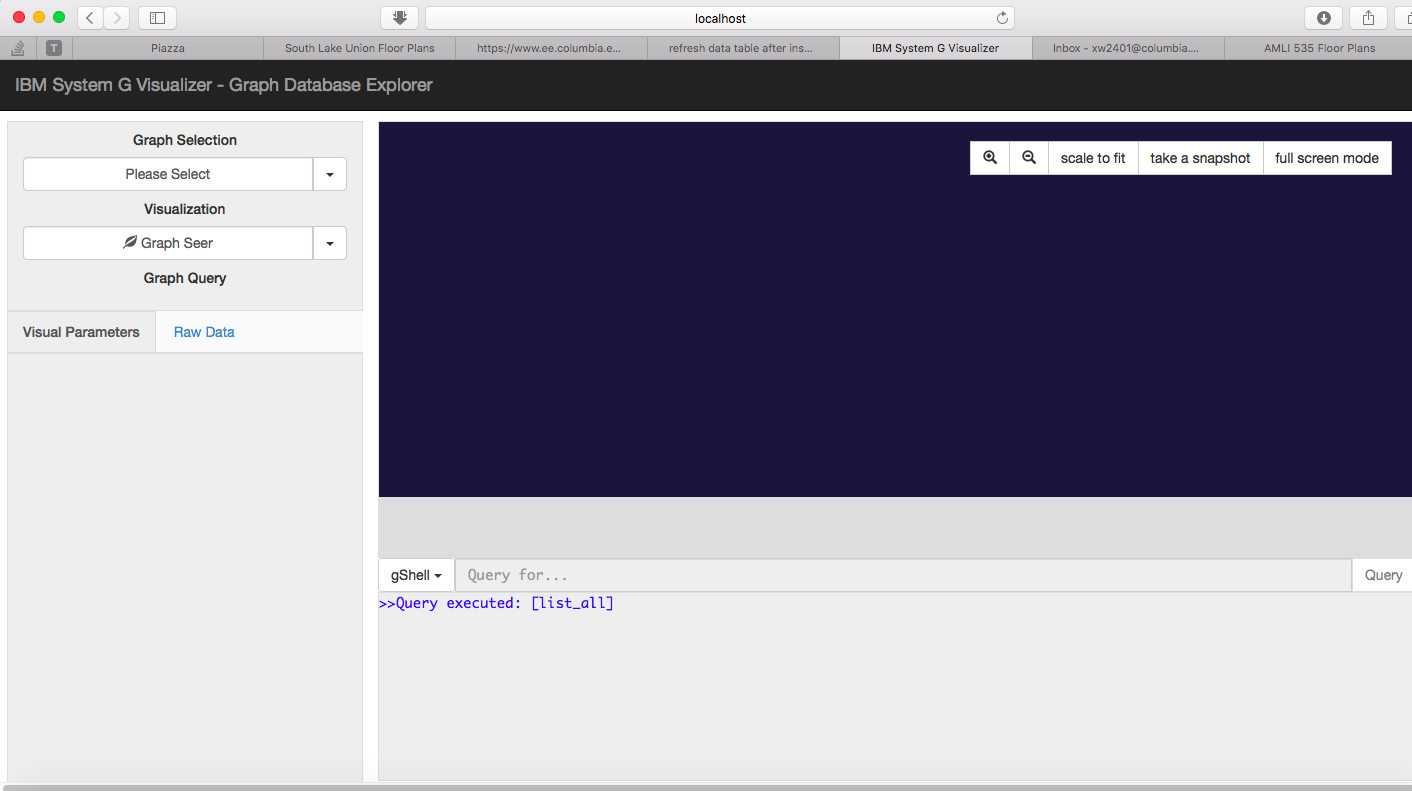
**Big Data Analytics Assignment 3**

**Name: Xucan Wang Uni: xw2401**

**1. Graph database**

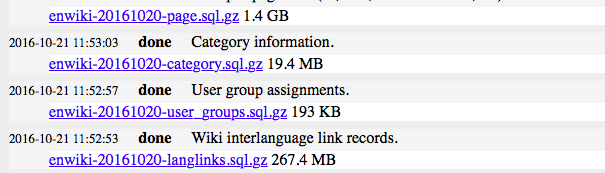
**1）Download IBM System G Graph Tools**

****

**2）Create a knowledge graph and try graph queries to find relevant items**

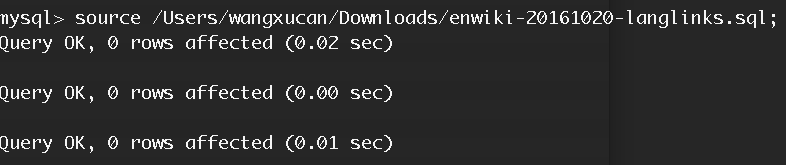
**A. dataset preparation part:**

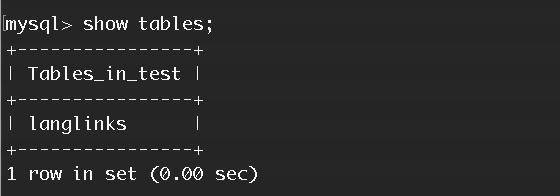
[**https://dumps.wikimedia.org/enwiki/20161020/**](https://dumps.wikimedia.org/enwiki/20161020/)

****

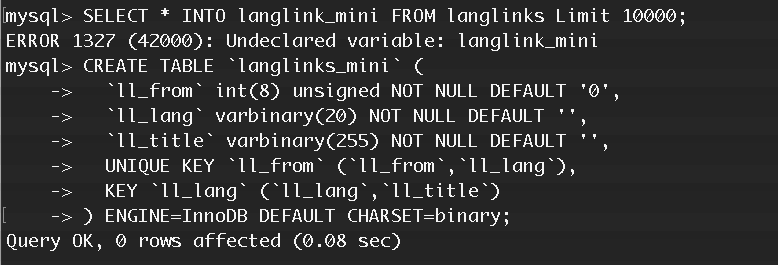
I used the page.sql and the langlinks.sql from the dataset above.

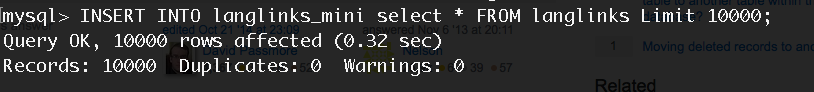
**B. Use the SQL file to create tables in MySQL database**





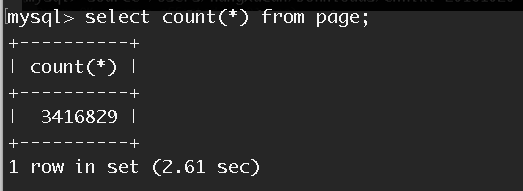
After the langlinks table is created, we only get a subset of the langlinks data and store the data into the langlink\_mini table (which has 10000 entries)



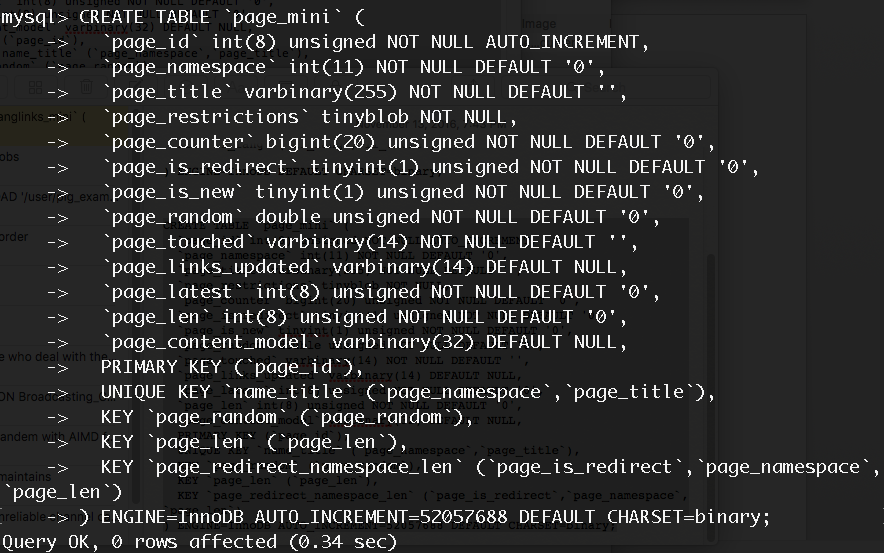






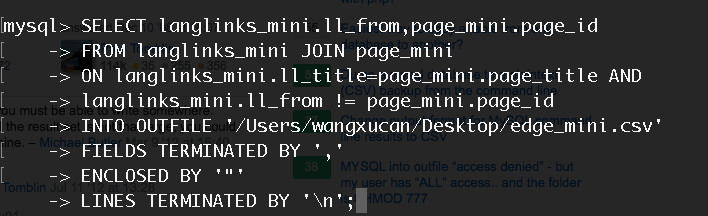


As the number of data in the page table is too oversized to do the join job, we still take a subset of the table and store it in the page\_mini table (we got 100000 records of page)



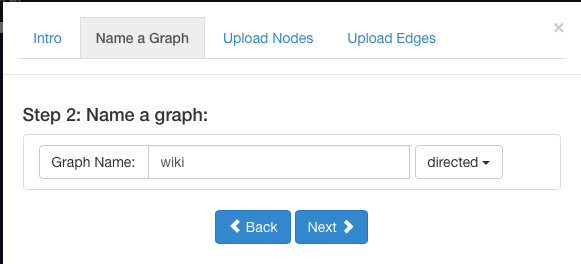


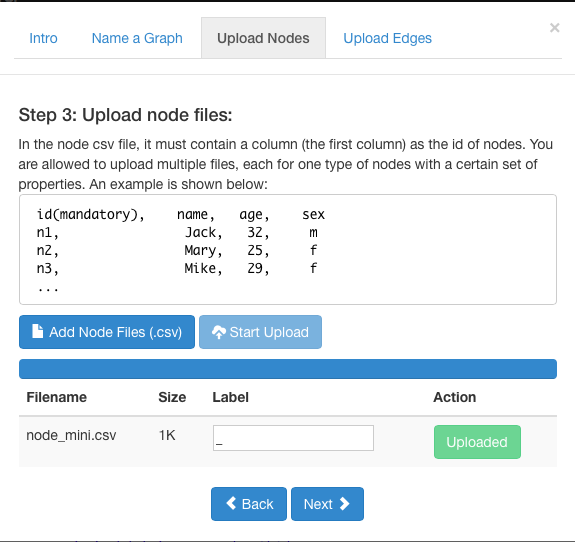
For now, we have prepared the datasets ready for join. According to the schema of the page table and langlinks table. We can get the edge.csv file with a JOIN action.

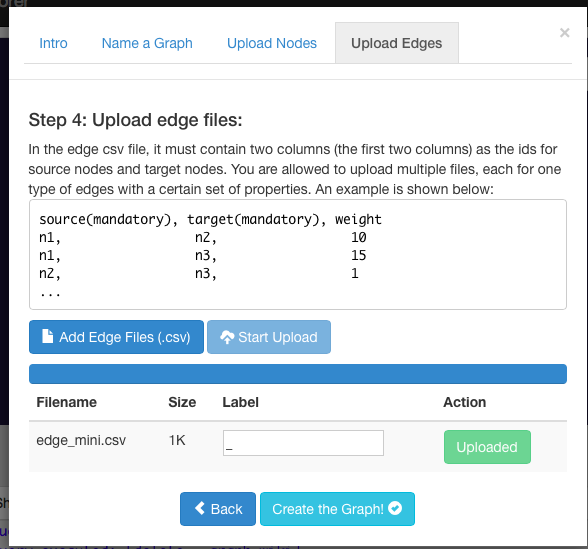


After this, we got the edge\_mini and the node\_mini csv file. Since we already got the edge\_mini.csv we can copy all the page\_id that exists in edge.csv to the node\_mini csv.

**C. Draw the graph using the IBM tool**

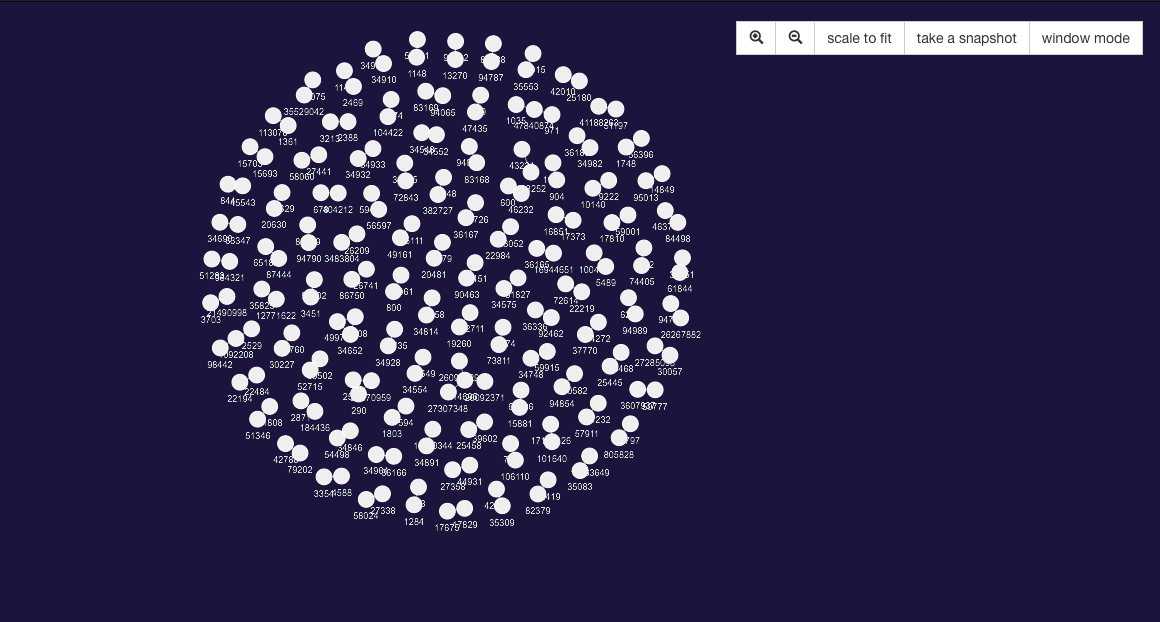
****

****

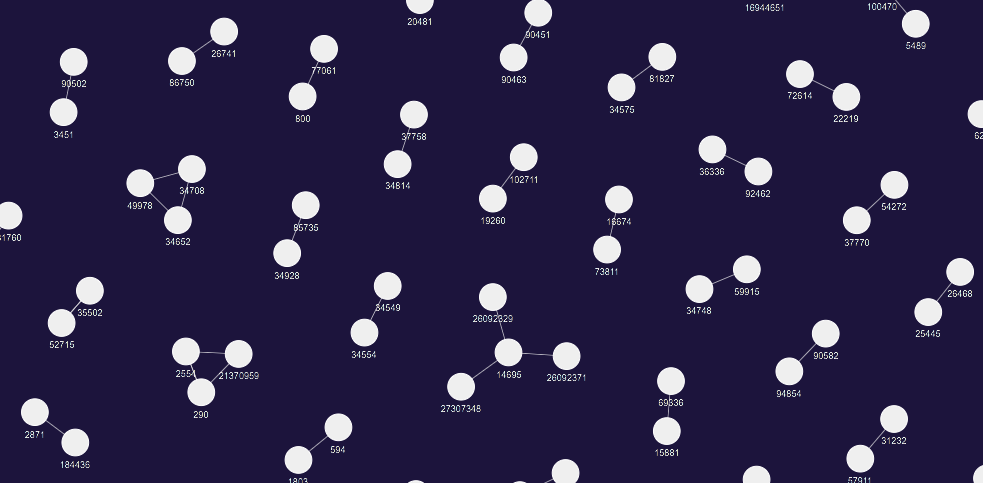
****

**After we draw the graph, it looks like this:**

****

****

**Part of the detailed graph looks like this:**

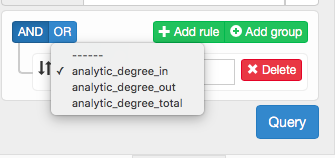
****

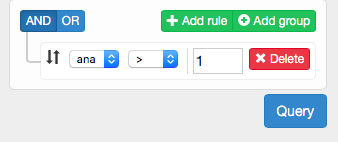
**The raw data on the side looks like this:**

****

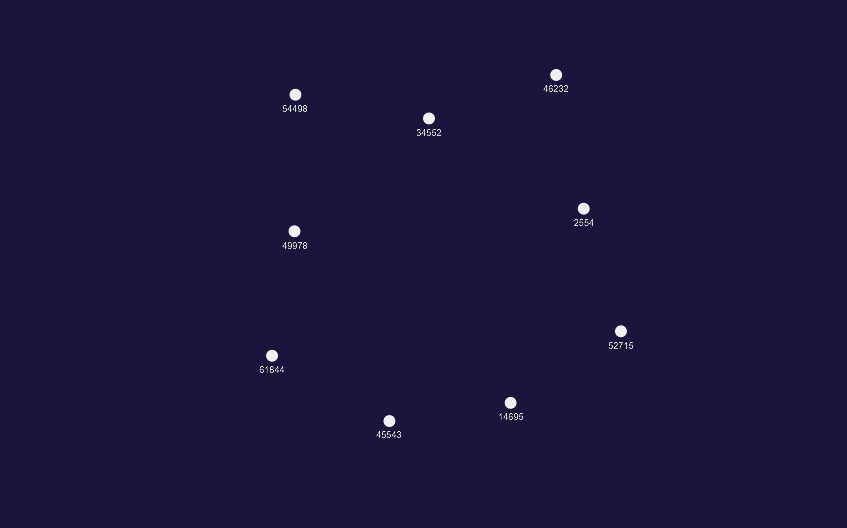
**D. Try graph queries on the nodes**

We can use the graph query to get some of the sub graph, for example:





In this query, we let the graph to show the nodes which has more than 1 in\_degree. The result looks like this:



When we go into one of the specific node 14695, we can find that, its in\_degree is 3.

