2020 IEEE 4th Information Technology,Networking,Electronic and Automation Control Conference (ITNEC 2020)

The Implementation Knowledge Graph of Air Crash Data based on Neo4j*

Yan Zou,Yan Liu
College of Mathematics and Computer Science, ChiFeng University
Inner Mongolia, China
329928728@qq.com

***Abstract*—The characteristics of neo4j graph database has been introduced, several methods of constructing graph database based on neo4j importing data have been compared and analyzed, and the "neo4j-admin" method has been finally selected to import the air crash data,to realize the visual display and query. Specific implementation steps as follows:(1) Data sets of air crash worldwide since 1908 [1] have been cleaned and preprocessed to remove incomplete lines and unimportant fields;(2) Crashed aircraft information table, accident details table and the above two table association relationship table have been designed and stored in the form of CSV files;(3)The CSV files have been imported into neo4j using the tool "neo4j-admin" , air crash knowledge graph has been generated; (4) This knowledge graph can be used to query the number of accidents since 1908, the place where the accidents occurred, the number of casualties and the time when the accidents occurred.**

Keywords—air crash data, knowledge graph, Neo4j, graph database, CSV files

I. INTRODUCTION

Relational databases such as SQL Server, MySQL, Oracle have been the mainstream of database management system for a long time. With the arrival of big data era, data are more diversified,RDBMS(Relational Database Management System) have exposed some shortcomings in data model, storage and query, the more prominent disadvantage is that: when dealing with various relational data information, it is accompanied by a large number of table JOIN operations. These join operations affect the efficiency of data queries. The graph database can solve this problem more efficiently. And it is more intuitive to use the form of graph to show the relationship between things in the real world[2,3]. With the application and development of knowledge graph in recent years, graph databases have also been widely used, such as Neo4j, gStore, FlockDB, AllegroGrap, GraphDB, InfiniteGraph, TITAN, OrientDb. Neo4j is a popular graph database management system with its fast reading and writing speed, flexible data design, strong adaptability, easy to use and convenient modeling. Neo4j is a high-performance non-relational graph database management system that stores structured data on the network (mathematically called graphs) instead of tables. It is an embedded, disk-based, fully transaction Java persistence engine with all the features of a mature database.

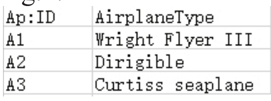
There are many ways of neo4j importing. For example, Cypher CREATE commands;The Cypher LOAD CSV statement is used to convert the data into CSV format and read the data through the LOAD CSV command;Batch Inserter tool can also be used; You can also use the neo4j official neo4j-import or neo4j-admin tool[4,5]. The advantages and disadvantages of these methods were shown as table 1.

In this paper,we have obtained a data set of air crash around the world since 1908 from Kaggle's website, which is stored in the form of CSV files. In view of the structural characteristics of the data file, we adopt the neo4j-admin commands to build the database of air crash graph based on neo4j, to present crash data in a clearer and more intuitive way.

II. DATA PROCESSING

The global air crash data sets since 1908 is a CSV table containing 5,269 lines, the fields included are: date of the accident, the specific time, the accident site, national army processing accident, flight number, type of aircraft, fatalities, and a brief description. We have cleaned and processed the data set, some fields with incomplete information have been removed, the information of four fields including the date of the accident, the place of the accident, the type of the aircraft and the fatalities have been retained, and an invalid information (neither the type of the aircraft nor the fatalities, but only contain the time of the accident) was deleted, and 5,268 pieces of data have been retained. We have designed three CSV tables to store the basic information of the aircraft, the time information of the accident and the relationship between them[6].

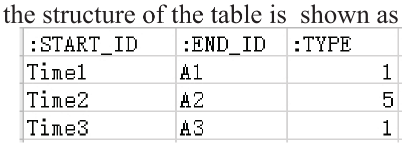
“Plates.csv”is used to store the basic information of airplanes, including two fields “Ap:ID”and “AirplaneType”, which has 2436 different aircraft types. Here, the values of “Ap:ID” field is required to be unique, and the structure of the table is shown as Fig.1.



Ap:ID	AirplaneType
A1	Wright Flyer III
A2	Dirigible
A3	Curtiss seaplane

Fig.1 The structure of “Plates. Csv” and its first three lines

“HappenTime.csv” is used to store the accident sites and time of the accidents,which including “Time: ID”, “HappenDate” and “IncidentSites” three fields, which has 5268 lines, the values of “Time: ID” field is also required to be unique, the structure of the table is shown as Fig.2.



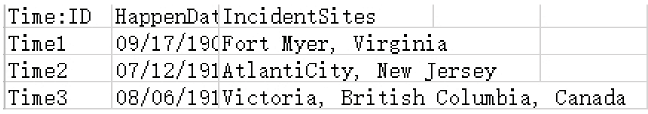
:START_ID	:END_ID	:TYPE
Time1	A1	1
Time2	A2	5
Time3	A3	1

Fig.2 The structure of “relationships.csv” and its first three lines

This work is supported by ChiFeng Science and Technology Bureau and ChiFeng University serve ChiFeng economic and social development application project(cfxyfc20184).

978-1-7281-4390-3/20/$31.00 ©2020 IEEE

Another table is “relationships.csv”, it is used to represent the relationship between the two above tables, it contains three fields: “:START_ID”, “:END_ID” and“: TYPE”, “:START_ID” is used to store the head nodes ID in the database, we have set the happened time of the accident as the head nodes, so“:START_ID”field storage is the value of “ Time: ID” of “HappenTime.csv”; “:END_ID”is used to store the ID of tail nodes in the graph database. Here, we use the name of aircraft type as tail node. “:TYPE”is used to store the relationships of head nodes and tail nodes. We regard the fatalities as the relationship between them, which is labeled on the edges between the two nodes. The structure of the table is shown as Fig.3.



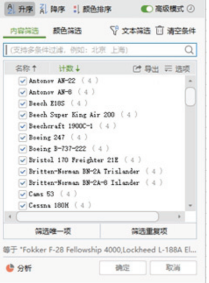
Time:ID	HappenDat	IncidentSites		
Time1	09/17/19(	Fort Myer, Virginia		
Time2	07/12/191	AtlantiCity, New Jersey		
Time3	08/06/191	Victoria, British Columbia, Canada		

Fig.3 The structure of “HappenTime.csv” and its first three lines

TABLE I. THE COMPARISON OF NEO4J DATA IMPORT METHODS

	Create commands	*load csv commands*	*Batch Inseter tool*	*Batch Import tool*	*neo4j-import or neo4j-admin tool*
Applicable scenario	*1~1w nodes*	*1w~10w nodes*	*Over 10 million nodes*	*Over 10 million nodes*	*Over 10 million nodes*
Speed	*slowly(1000 nodes/s)*	*generally(5000 nodes/s)*	*Very fast (tens of thousands nodes/s)*	*Very fast (tens of thousands nodes/s)*	*Very fast (tens of thousands nodes/s)*
Advantages	*Easy to use, can be inserted in real time*	*Easy to use, can load locally*	*Remote CSV; Real-time insertion*	*Batch Inserter can run the compiled jar package directly;can import data from an existing database*	*Official production, which takes up fewer resources than Batch Import*
Disadvantages	*Slow speed*	*need to convert the data into CSV format*	*need to convert it to CSV; Only available in JAVA; Neo4j must also be stopped when inserting*	*need to convert it to CSV; Neo4j must be stopped*	*need to convert the data to CSV; Neo4j must be stopped; You can only generate new databases, can not insert data into existing databases*

For data processing, the key operation lies in the processing of “Plates.csv” and “relationships.csv”For the“Plates.csv”table, we need to reprocess the rows with the same aircraft type name.Taking the behavior of aircraft type name repeating four times as an example, we use the "filter" function of Microsoft Excel to deduplicate one by one,The specific steps are shown as fig. 4.

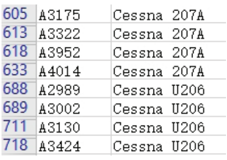


☑ Antonov AN-22 （4）
☑ Antonov AN-8 （4）
☑ Beech E18S （4）
☑ Beech Super King Air 200 （4）
☑ Beechcraft 1900C-1 （4）
☑ Boeing 247 （4）
☑ Boeing B-737-222 （4）
☑ Bristol 170 Freighter 21E （4）
☑ Britten-Norman BN-2A Trislander （4）
☑ Britten-Norman BN-2A-8 Islander （4）
☑ Cams 53 （4）
☑ Cessna 180H （4）

等于 "Fokker F-28 Fellowship 4000,Lockheed L-188A El...

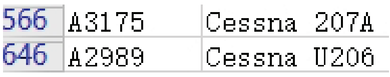
Fig.4 “Plates.csv” reprocess and filter

For each group of data screened out, we adopted the method of "keeping the minimum number", that is, the record of the large number was deleted, and the minimum number was taken as the unique number of the aircraft type.Before removing the duplicate lines is shown as fig. 5,and after removing the duplicate lines is shown as fig. 6.



605	A3175	Cessna 207A
613	A3322	Cessna 207A
618	A3952	Cessna 207A
633	A4014	Cessna 207A
688	A2989	Cessna U206
689	A3002	Cessna U206
711	A3130	Cessna U206
718	A3424	Cessna U206

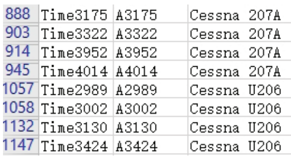
Fig.5 “Plates.csv” before reprocessing



566	A3175	Cessna 207A
646	A2989	Cessna U206

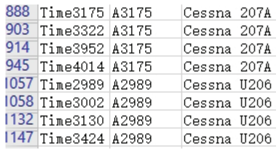
Fig. 6 “Plates.csv” after reprocessing

In this way, all duplicate rows in the “Plates.csv”were deprocessed manually, the original table which contains 5268 pieces of data,has been reduced to 2436. For “relationships.csv” ,the processing mainly refers to the aircraft will be the same name in the table using the same number, we adopt the method is "minimum number covering method", which select the smallest number as the serial number instead of others, before processing,the data in the table are shown as fig. 7, after processing the data is shown as fig. 8. In addition, for the missing rows of the aircraft TYPE name (“AirplaneType” field of the Plates.csv) and the missing rows of the fatalities (the “:TYPE” field of the relationships.csv), fill with "NULL" uniformly.



888	Time3175	A3175	Cessna 207A
903	Time3322	A3322	Cessna 207A
914	Time3952	A3952	Cessna 207A
945	Time4014	A4014	Cessna 207A
1057	Time2989	A2989	Cessna U206
1058	Time3002	A3002	Cessna U206
1132	Time3130	A3130	Cessna U206
1147	Time3424	A3424	Cessna U206

Fig.7 “relationships.csv” before processing



888	Time3175	A3175	Cessna 207A
903	Time3322	A3175	Cessna 207A
914	Time3952	A3175	Cessna 207A
945	Time4014	A3175	Cessna 207A
1057	Time2989	A2989	Cessna U206
1058	Time3002	A2989	Cessna U206
1132	Time3130	A2989	Cessna U206
1147	Time3424	A2989	Cessna U206

Fig. 8 “relationships.csv” after processing

III. VISUALIZATION OF THE GRAPH

We have used the "neo4j-admin" command tool to import the above CSV files, and the graph database is established as "plates. db", is shown as fig. 9. When we execute the command "START n=node(*) RETURN n" in the browser, it shows all the nodes and relationships in the plates.db, including 5267 “HappenTimes” (time of accident) nodes, 2435 “AirplaneType”(aircraft type name) nodes, and 5267 relationships. Fig. 10 shows a portion of overall graph. Let's take any one of these nodes to illustrate what the graph means. The Junkers f-13 has crashed 17 times since 1908, the values in the head node represent where the accident occurred,When we hover the cursor over the head node, the time of the accident is shown at the bottom of the graph, and the data on the edge represents the number of casualties.it is shown as fig. 11 (the "Initial Node Display" under the configuration is 300). Fig. 12 shows the plane with the highest number of accidents, known as the Douglas dc-3, with 334 crashes. The graph lists the location, time and casualties of each accident.

IV. CONCLUSIONS

Based on the open air crash data set, we use the “neo4j-admin” command to import csv files to build a plane crash graph database , and realize the visualization.The graph contains 7702 nodes and 5267 relationships,it has certain practical significance of the data statistics and convenient query.As a preliminary exploration of constructing knowledge graph, there are many aspects to be further research, in the future ,we will consider using crawler technology to crawl more data about a plane crash on the websites, from a multidimensional perspectives to interpret crash data, and using ontology tools [7,8]to design and build more complex plane crash data knowledge graph, make the graph construction have more practical application value.

REFERENCES

[1] plane crash data source: https://www.kaggle.com/saurograndi/airplane-crashes-since-1908.

[2] LI Ying,ZHANG,“Shu-guang.Application of knowledge mapping in analysis of the discipline development”.J.Journal of Medical Postgraduates,2013,vol. 26, pp. 875-877.

[3] KANG Jie-hua ，LUO Zhang-xuan,“Research on RDF data storage based on graph database Neo4j”.J.Information Technology,2015,vol.6,pp.115-117.

[4] HAO PeiHao, GAO Jie,“Visualization Analysis of Police Security Knowledge Map Based on Neo4j Graph Database”.J.Modern Computer,2018,vol.35,pp.8-11.

[5] Webber J,“A Programmatic Introduction to Neo4j”.C. Conference on Systems,Programming,and Applications:Software for Humanity,2012,pp.217-218.

[6] CSV table structure design: https://neo4j.com/.

[7] ZHANG Yong,Lv Jun-bai,“Research review on ontology modeling based on Protege”.J.Fujian Computer,2011,vol.1,pp.43-45.

[8] XIAO Qing-du,QU Liang-liang,HOU Xia,“Design and implementation of knowledge graph system of course system based on Neo4j graph database”.J.Computer Knowledge and Technology,2017,vol.13,pp.130-132.

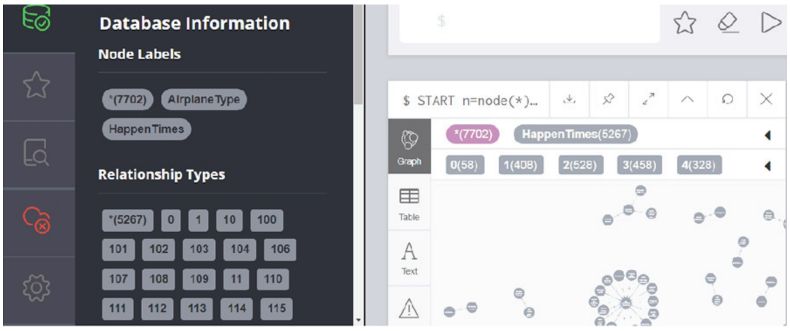
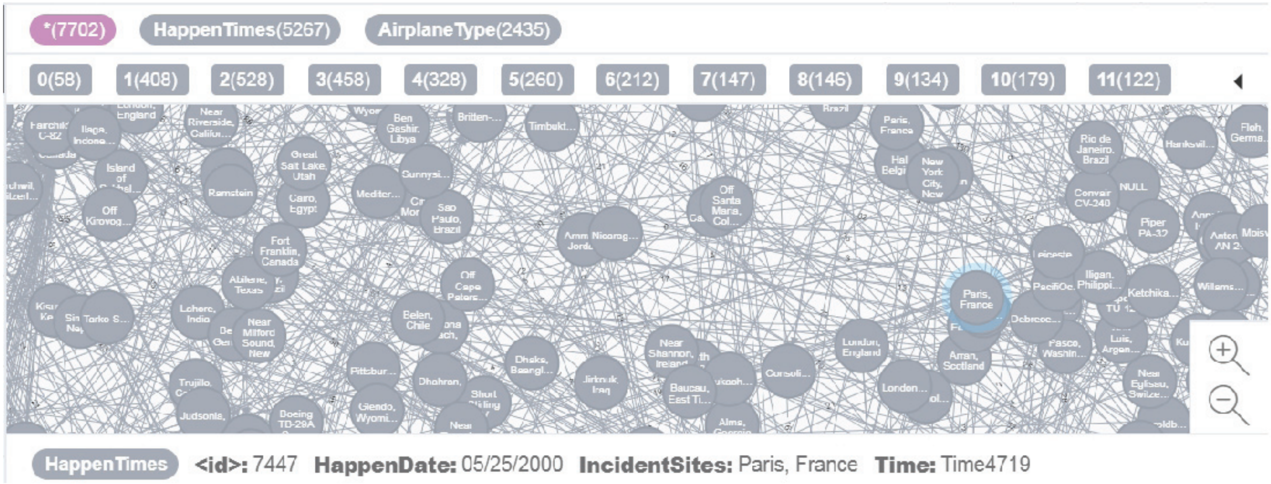
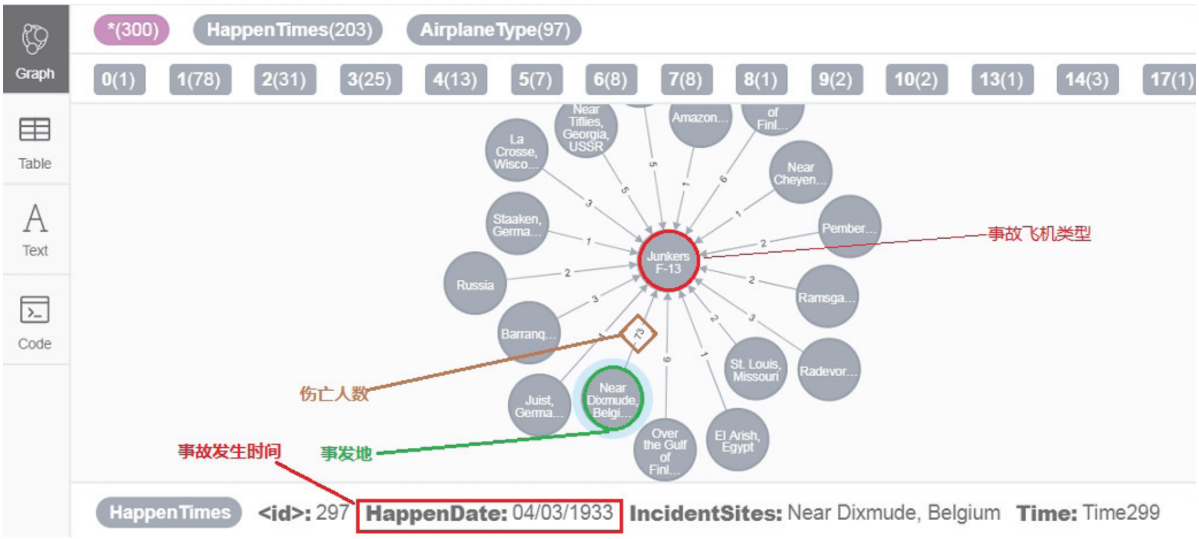


Fig.9 plates. db



HappenTimes <id>: 7447 HappenDate: 05/25/2000 IncidentSites: Paris, France Time: Time4719

Fig. 10 a portion of the overall graph



HappenTimes <id>: 297 HappenDate: 04/03/1933 IncidentSites: Near Dixmude, Belgium Time: Time299

Fig.11 the graph of "Junkers f-13"

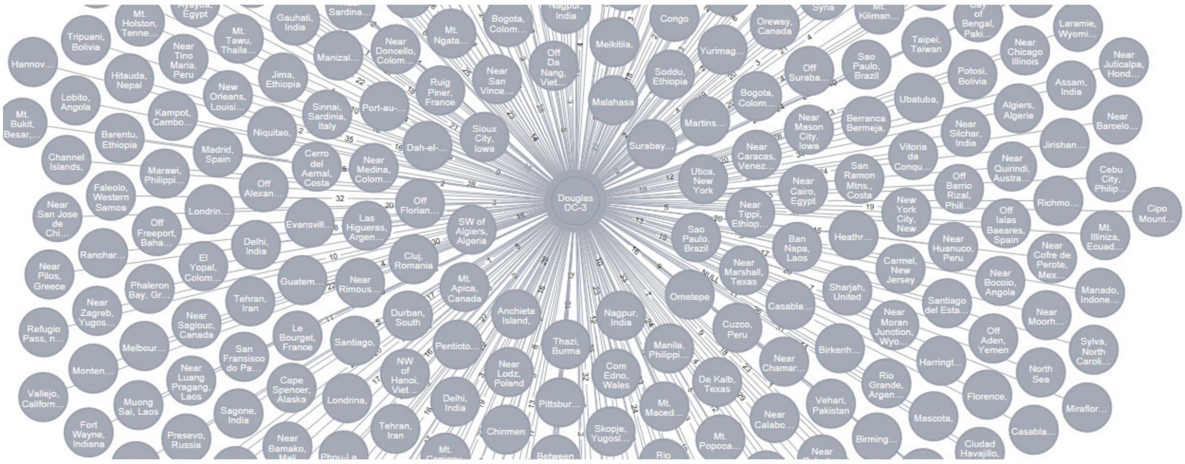


Fig. 12 the graph of Douglas dc-3(the aircraft with the highest number of accidents)