Using Fusion for solving heterogeneity and incompleteness of information in Big Data

Ehsan Azizi Khadem[1], Emad Fereshteh Nezhad[2]

[1.] MSc of Computer Engineering, Department of Computer Engineering, Lorestan University, Iran
[2.] MSc of Computer Engineering, Communications Regulatory Authority of I.R of Iran

Abstract: The Big Data term is used for describe large scale data that is stored and analyzed in an organization or company. The amount of data in this kind of information is often more than trabyte. Big Data is a new and very attractive field of database science that many engineers studying and mine about it in all of world. But this field has some challenges like heterogeneity and incompleteness. In this paper, we use fusion to preprocess data to solve two challenges mentioned above.

[Ehsan Azizi Khadem, Emad Fereshteh Nezhad. **Using Fusion for solving heterogeneity and incompleteness of information in Big Data.** *Nat Sci* 2014;12(12):100-103]. (ISSN: 1545-0740). http://www.sciencepub.net/nature. 15

Keywords: Big Data, Data Base, Fusion, Heterogeneity, Incompleteness

1. Introduction

During last 40 years data storage and analysis is a very important field of science that includes querying of science and data management to mine rules from large collection of data and to sophisticate prediction.[1][2]

Big Data is a defined term used to describe very large datasets like information of mobile users and their communication. Furthermore increasing of amount of data and hidden regularity in it guide users to more studies about new techniques tostore, retrieve and knowledge extraction from big data.[3] But some problems start right away during these techniques. For example posts and comments are weakly structured of piece of text, while images, sounds, and videos are strongly structured for storage and display, but not for mining, semantic search and so on. These challenges occurs underline many and sometime all of steps of Big Data storage and retrieve. We introduce some of these challenges and then present a manner to manage one of them.[4][5][6]

Problems And Challenges

As mentioned some challenges and problems are underline the Big Data analysis that must be manage for more speed and accuracy. The obvious thing people think of Big Data is its size. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades.[7][8]

Nowadays, data volume is increasing faster than compute resources and CPU sped. Furthermore in data distribution we meet many other problems like how to aggregate multiple disparate workloads with varying performance goals. [9]

Another challenge is traditional I/O systems. For example, hard disk drives were used to store persistent data but newer storage technologies do not have same large spread in performance between the sequential and random I/O performance. If you process the larger data set it takes longer time. There are many situations in that the result of analysis is required immediately. For example in ebanking transaction we need tremendous speed in request and response. But when we use large data set, it is so difficult. It is often necessary to find criteria to rapid access. In data analysis, this sort of search is likely to occur repeatedly. Scanning the entire data set to find suitable elements is obviously impractical. [10][11][12][13]

The privacy of data is another challenge in Big Data. Protecting the data that distribute in several resources is very difficult. The existing paradigm of differential privacy is a very important step in right direction, but it reduce information content too far in order to be useful in most practical cases. Furthermore we have challenge with human collaboration.[14] There are individual input data for several human experts that must be analyzed in Big Data. These multiple experts may be separated in space and time when it is too expensive to assemble an entire team together in one room. The data system has to accept this distributed expert input, and support collaboration. Now companies and organizations analyze business data for risk management, customer retention, and brand management and so on. Separate system handle varied tasks and using Big Data lead to have large size data sets. It needs more time and money to implement this structure because of workloads. The challenge is for underlying system architecture to be flexible enough that the components built on top of it for expressing the various kinds of processing tasks can tune it to efficiently run these different workloads. [15][16]

The very fact that Big Data analysis typically involves multiple phases high lights a challenge that arises routinely in practice: production system must run complex analytic pipelines, or workflows, at routine intervals, hourly or daily. But one of very important challenge is heterogeneity and incompleteness.[17][18][19]

Heterogeneity And Incompleteness

Machine analysis algorithm expects homogeneous data, and can not understand nuance. Data must be carefully structured as a first step in data analysis. For example, a student wants to select some courses in a term. We could create one record per selection a course by a student.[20][21] Another design is create a record for each course in a term and store all information about all students whom select that course in the record. The three design choices listed have successively less structure and conversely successively variety. However, the less structured design is likely to be more effective for many purposes for example questions relating to grades of a major will require an expensive join operation with the first two designs, but can be avoided with the latter. However, computer system works most efficiently if they can store multiple items that are all identical in size and structure. Efficient representation access and analysis of semi structured data required further work. Even after data cleaning and error correction, some incompleteness and errors in data are likely to remain. This incompleteness and errors must be managed during data analysis.[22][23][24] Doing this correctly is a challenge. Recent work on managing probabilistic data suggests one way to make progress. We present using fusion for solving heterogeneity and incompleteness of information in Big Data.

Fusion:

In MongoDB we have collections and documents. A collection is very similar to schema in relational databases. Collection is main structure of a NoSQL Data Base in MongoDB.[25]

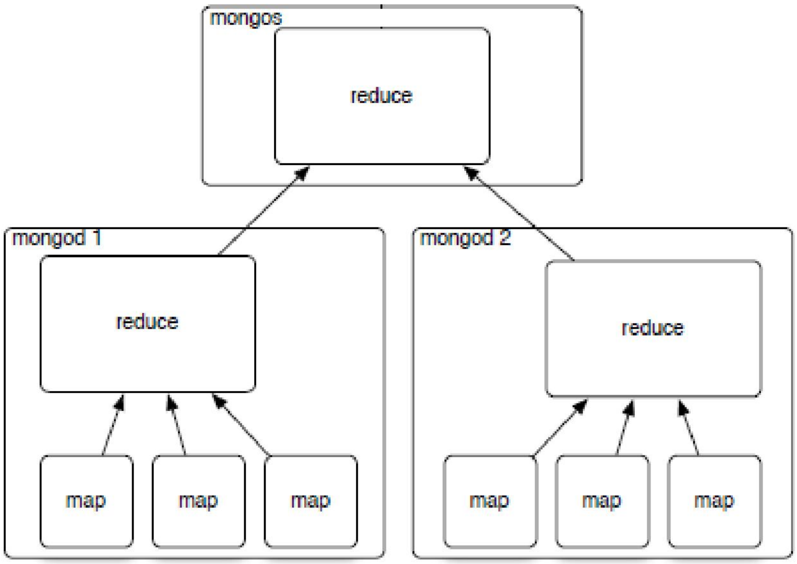
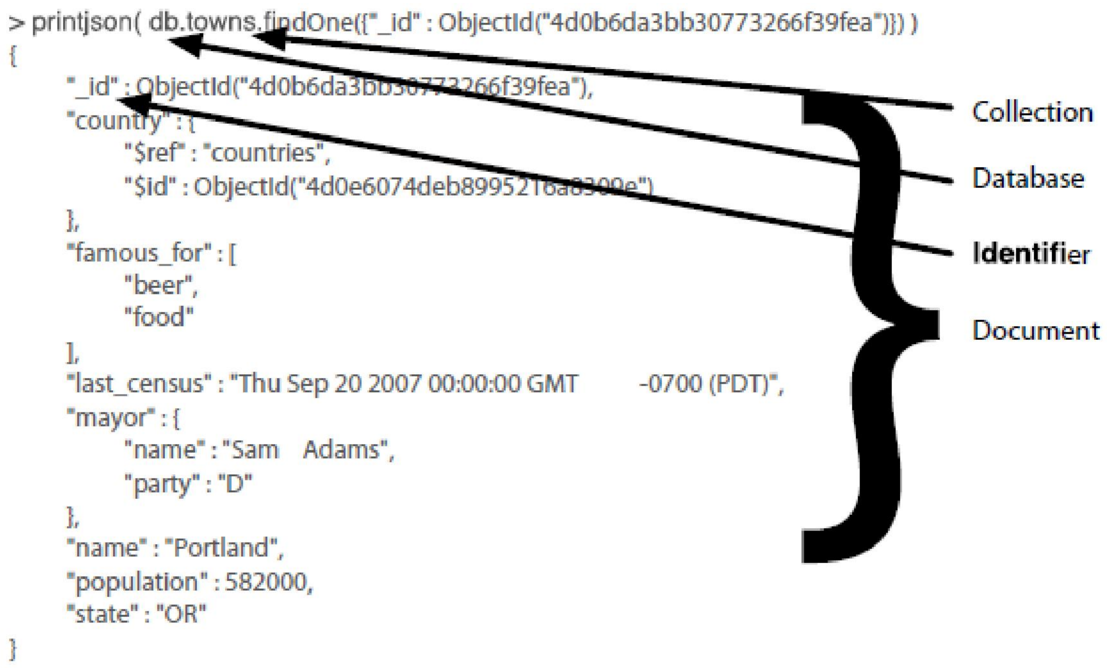


Figure 1: A mongoDB over two servers

If we have a few numbers of collections it seems that our database is simpler for design, search, retrieve data and etc. But in some projects perhaps there are many different collections for implementation. What can we do? For example, in an university database we must design a collection for students and another collection for courses and another for teachers and… But if we fuse these collections in one collection we find better performance. Although in relational databases we worry about fusing different tables, in Big Data we prefer it. When we design a collection for whole students, courses and teachers, it is not necessary to join between many collections and saving time and memory.

If we use explain() in find() method, we can compare times between two queries with two structures is mentioned above. It is clear that when field "millis" shows time to executing a query, we see it is very faster to executing a query in one collection against join two or three collection for find a document or documents.



```
> printjson( db.towns.findOne({"_id" : ObjectId("4d0b6da3bb30773266f39fea")}) )
{
      "_id" : ObjectId("4d0b6da3bb30773266f39fea"),
      "country" : {
            "$ref" : "countries",
            "$id" : ObjectId("4d0e6074deb8995216a8309e")
      },
      "famous_for" : [
            "beer",
            "food"
      ],
      "last_census" : "Thu Sep 20 2007 00:00:00 GMT        -0700 (PDT)",
      "mayor" : {
            "name" : "Sam   Adams",
            "party" : "D"
      },
      "name" : "Portland",
      "population" : 582000,
      "state" : "OR"
}
```

Figure 2 : A mongo document printed as a json

```
>db.university.find(num:1,co:2)
One collection:
{
"num" :  "1"
"co" : "2"
"milllis" : 25
}


Two collections :
{
"num" :  "1"
"co" : "2"
"milllis" : 263
}
Three collections :
{
"num" :  "1"
"co" : "2"
"milllis" : 3768
}
```

Conclusion:

Big Data is a very important solution for large scale data sets and it has a very tremendous ability to manage increasing data in volume and types. We have some challenges in this way that decreasing performance of database management systems. Some of these challenges are heterogeneity and incompleteness, scale, timeless, privacy, human collaboration and system architecture. We must solve problems mentioned above to increase speed and saving memory size. In heterogeneity and incompleteness if we fuse relative collections and construct whole collection, will have faster executing queries and effective saving of the data. Of course we must use vertical growing in data sets to distribute collections in several servers.

References:

1. Barabasi, A. (2003). *Linked: How everything is connected to everything else and what it means.* New York: Plume.
2. Barabasi, A., Albert, R., & Jeong H. (2000). Scale-free characteristics of random networks: The topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications, 281,* 69–77. doi:10.1016/S0378-4371 (00)00018-2
3. Bishop, S., Helbing, D., Lukowicz, P., & Conte, R. (2011). FuturICT: FET flagship pilot project. *Procedia Computer Science, 7,* 34–38.
4. Centola, D., Eguíluz, V. M., & Macy, M. W. (2007). Cascade dynamics of complex propagation. *Physica A: Statistical Mechanics and its Applications, 374*, 449–456.
5. Cointet, J. P., & Roth, C. (2009). Socio-semantic dynamics in a blog network. *International Conference on Computational Science and Engineering.* doi:10.1109/CSE.2009.105
6. Conte, R., Gilbert, N., Bonelli, G., & Helbing, D. (2011). FuturICT and social sciences: Big

Data, big thinking*Zeitschrift für Soziologie, 40,* 412–413.

7. Cortes, C., & Pregibon, D. (2001). Signature-based methods for data streams. *Journal of Knowledge Discoveryand Data Mining, 5*, 167–182.
8. Goetz, M., Leskovec, J., McGlohon, M., & Faloutsos, C. (2009). Modeling blog dynamics. *AAAI Conference onWeblogs and Social Media*, 2009. Retrieved from http://cs.stanford.edu/~jure/pubs/blogs-icwsm09.pdf
9. Helbing, D., & Balietti, S. (2011). From social data mining to forecasting socio-economic crises. *European Physical Journal-Special Topics, 195,* 3–68.
10. Hipp, J. R., & Perrin, A. J. (2009). The simultaneous effect of social distance and physical distance on the formation of neighborhood ties. *City & Community, 8,* 5–25.
11. Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. S. (1999). The Web as a graph: Measurements, models and methods. *Lecture Notes in Computer Science, 1627*, 1–17.
12. Kossinets, G., & Watts, D. J. (2009). Origins of homophily in an evolving social network. *American Journal ofSociology, 115*, 405–450.
13. Leskovic, J., Kleinberg, J., & Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1, Article 2.
14. Liljeros, F., Edling, C. R., Nunes Amaral, L. A., Stanley, H. E., & Aberg, Y. (2001). The web of human sexualcontacts. *Nature, 411*, 908–909.
15. Milne, D., & Witten, I. H. (2009). An open-source toolkit for mining Wikipedia. *Proceedings of the New Zealand Computer Science Research Student Conference, 9.*
16. Price, D. J. de S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal ofthe American Society for Information Science, 27,* 292–306. doi:10.1002/asi.4630270505
17. Raine, L., & Wellman, B. (2012). *Networked. The new social operating system.* Cambridge: MIT Press.
18. Reips, U.-D. (2011). Journal impact revisited. *International Journal of Internet Science, 6*(1), 1–7.
19. Reips, U.-D., & Garaizar, P. (2011). Mining Twitter: Microblogging as a source for psychological wisdom of thecrowds. *Behavior Research Methods, 43*, 635–642. doi:10.3758/s13428-011-0116-6
20. Schnegg, M. (2006). Reciprocity and the emergence of power laws in social networks. *International Journal ofModern Physics C, 17,* 1067–1076.
21. Snijders, C., & Weesie, J. (2009). Reputation in an online programmers' market. In K. S. Cook, C. Snijders, 23.V.Buskens, & C. Cheshire (Eds.), *Trust and reputation* (pp. 166–185). New York: Russel Sage Foundation.
22. Stephen, A. T., & Toubia, O. (2009). Explaining the power-law degree distribution in a social commerce network. *Social Networks, 31,* 262–270.
23. Prof. Walter Kriha,(2012).NoSQL Databases,(pp. 85-110)

12/13/2014