Explore the Challenges for Building Big Data Statistics and Visualization system

LIU,HONGYANG

March 29, 2020

1 Title

Explore the challenges for building big data statistics and visualization system

2 Articles synopsis

Article1: Big Data Visualization Tools: A Survey
Article2: Big data and visualization: methods, challenges and technology progress
Article3: Big Data Exploration, Visualization and Analytics
Article4: Exploration and visualization in the web of big linked data: A survey of the state of the art
Article5: Big Data: A Survey

2.1 Article 1:

The first article introduces the challenges and visualization tools for organizations, groups, and entrepreneurs while analyzing large data sets. One the one hand, the challenges increase the complexity of data representation. One the other hand, it reduces the effectiveness of the interactive website(Enrico Giacinto Caldarola, Picariello, and Castelluccia 2015). Large scale data sets would make the visual algorithms rather complicated and the hardware like storage or memory is far from enough to handle enormous data in a short time. The complexity of the system would dramatically increase with the

data sets expanding. As the responsive time becomes longer, customers would spend more time waiting and handling the data. In addition, the effectiveness of the system would be greatly impacted. There the authors proposed and survey the technical solutions for the visualization system.

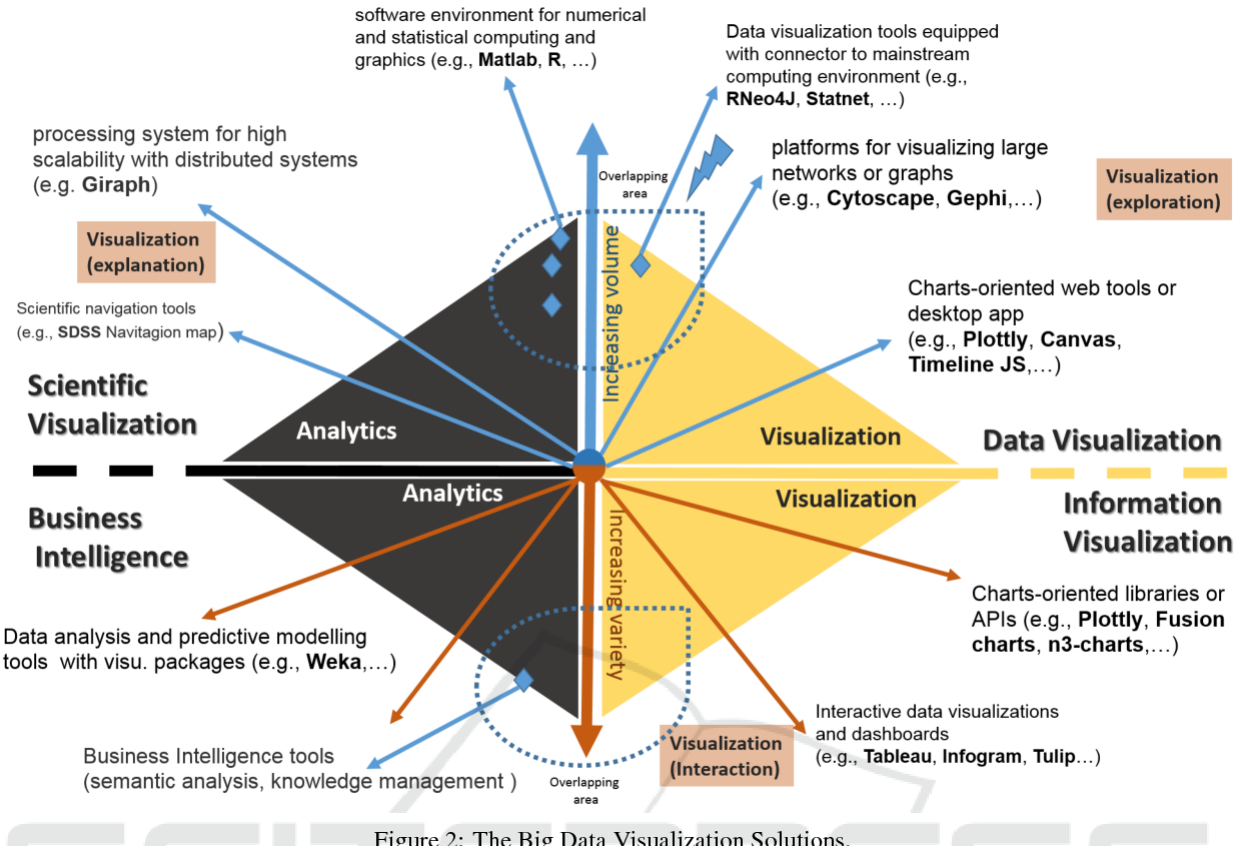

Figure 2: The Big Data Visualization Solutions

Figure 1: The Big Data Visualization Solutions

According to the author's theories, figure 1 introduces the procedures and four aspects of solutions proposed by previous articles(Enrico G Caldarola and Antonio M Rinaldi 2017). The four regions depicted in the figure show that the visualization tools could be classified into four categories and each category has its own analytic and visualization tools. Furthermore, the author proposes eight criteria to evaluate the four categories of current big data visualization tools. With the help of specific criteria and standards, the software and visual systems have been surveyed through the information

extracted from the products' websites.

However, this article mainly focuses on the classification of big data visualization tools and only considers several characteristics of the big data. The velocity is not including in this article's scope. And it only gives us limited methodologies on the construction of data visualization tools.

2.2 Article 2:

Name	Category	Data structure	Operating System	License	Query Languages	API availability	Latest Release (Date)
Apache Cassandra	NoSQL	Column-based	OS Independent	Open	CQL (Cassandra Query Language)	C++, C#, Python, Java	2.1.5 (2015-04-29)
Apache HBase	NoSQL	Table, Map	OS Independent	Open	HBase Query Predicates	Java	0.94.27 (2015-03-26)
Google BigTable	NoSQL	Table, Map	OS Independent	Open		Java, Python	
MongoDB	NoSQL	Document-based	Linux, Windows, OS X	Open	Mongo DB Command line language	Many third party client library exist	3.0 (2015-03-03)
Neo4j	NoSQL	Graph	Windows, Linux	Community, Commercial	Cypher Query Language	Rest API	2.2.2 (2015-05-21)
Apache CouchDB	NoSQL	Document	Ubuntu, Windows, Mac OS X	Open	CouchDB primitives	HTTP API	1.6.1 (2015)
OrientDB	NoSQL	Graph		Open	SQL	.NET. Php, Ruby, Python	2.1 (2015-05-05)
Terrastore	NoSQL	Document	OS Independent	Open	Primitives via HTTP	HTTP API	0.8.2 (2015-09)
FlockDB	NoSQL	Graph	OS Independent	Open	Native Language	Ruby and Apache Thrift API	1.8.5 (2012-03-09)
Hibari	NoSQL	Key-Value	OS Independent	Open	OS Independent	Native Erlang, Universal Binary Format (UBF/EBF), Apache Thrift, Amazon S3, JSON-RPC	source code
Riak	NoSQL	Key-Value	Unix (several distros)	Open	CRUD Operations via HTTP requests	Java, Ruby, Python, C#, Node.js, PHP, Erlang, HTTP Api, Python, Perl, Clojure, Scala, Smalltalk, and many others.	2.1.1 (-)
Hypertable	NewSQL	Table	Linux, Mac OS X	Open	SQL-like	C++, Java, Node.js, Perl, PHP, Python, Ruby	0.9.8.7 (-)
StarDog	NoSQL	Graph	Independent	Different licenses	SPARQL	Java	3.0.2 (2015-05-12)
Apache Hive	NewSQL	Table	OS Independent	Open	HiveQL (SQL-like)	Java	1.1.0 (2015-03-08)
InfoBright Community Edition	NoSQL	Column-oriented	Windows, Linux	Community, Commercial	SQL-like	ODBC, JDBC, C API, PHP, Visual Basic, Ruby, Perl and Python	4.0.7 (-)
Infinispan	NoSQL	key-value	OS Independent	Open	Own query DSL	Java, Ruby, Python, C++, .NET (via Hot Rod Protocol)	7.2.1 (2015-04)
Redis	NoSQL	key-value	Linux	BSD	Redis commands	Many	3.0.1 (2015-05-05)
Clustrix	NewSQL	Table	Linux	Commercial	SQL	Windows	6.0
VoltDB	NewSQL	Table	OS Independent	Both	SQL	Java, JDBC, ODBC	5.2

Figure 2: Evaluation synopsis of a set of technical characteristics

The second article has broad topics that aim at surveying the problems with big data. The purpose of the author is to choose the suitable tools for solving large scale data sets. Big data involves three characteristics like volume, velocity, and variety. The dimensions of big data increase the complexity whiling handling the data sets. In addition to challenges like heterogeneous data and data scaling, there are also some techniques problems like data storage. In this article, the author points a qualitative methodology to survey the databases from three categories: SQL, NoSQL, and NewSQL. He illustrates the strengths and weaknesses of the databases in each category.

Furthermore, the author compares the different databases from several categories and the results have been shown in figure 2. The figure shows that compared with the traditional SQL database, NoSQL database has much better performance while utilizing large volumes and heterogeneous data. Besides, it gives the big data software developers a data storage solution, which makes the individuals or companies reducing the cost and increasing revenue.

Overall, the contents of this article involving the survey of the prevalent big data storages tools and it helps us solve the storage of big data visualization. Therefore, It may reduce the response time when querying or retrieving data from the databases. However, we also found that some limitations of this article. For example, some criteria have not been taken into consideration; The scalability and has not been evaluated in the system. We still face some challenges while building a big data visualization system(Enrico Giacinto Caldarola and Antonio Maria Rinaldi 2015).

2.3 Article 3:

In this article, the author reviews several contemporary visualization solutions to solve the challenges of building a visualization system. The author's main point is that interactive systems should ensure that non-experts use interactive systems as easily as possible and help them gain insights or make business decisions without the help of professional information experts. To achieve this objective, the author introduces several works in multidisciplinary research papers through literature reviews. This work demonstrates several ideas for coping with visual issues. However, we could not find a specific implementation in this paper(Bikakis, Papastefanatos, and Papaemmanouil 2019).

2.4 Article 4:

The author explores recent works related to the data system in the *Web of Data* (WoD) and proposes their suggestions on the improvements of the WoD system. The first challenge illustrated in the article is the large scale data sizes, while the current interactive websites could only handle a small fragment of datasets. Another challenge mentioned in the article is dynamic generated data because the data is collected and increased with time. The

dynamic data sets limited the scalability of data processing. big data system developers should also consider the personal preferences. For example, the customer may offer some demands: (1) use the specific components or widgets(maps, charts, tables,etc.); (2) manage data according to user's preferences; (3) modify data parameter, etc. (4) define exact operational functionality to handle data.

Beyond these requirements above, the system should also offer some start-up tutorials or assistance to help users easily use the system and improve their efficiency while manipulating the interactive systems. The last challenge is concerning the visual presentation, as the performance of the WoD system, could is limited at a small set of data. When users have a higher requirement to do data summary or aggregation, they may need more efficient algorithms to handle the data sets and enough storage, memories and computing resources for improving the performance of visual systems.

To solve the challenges and problems, the article illustrates several visual tools in the WoD. The authors survey and compare the visual tools and classified them into five categories. After the comparison, the authors found that most of the tools in the WoD system do not fulfill the requirements which have been mentioned previously. To sum up, the development of the visualization system should consider the two requirements surveyed by the article to handle these challenges and improve performance in further work(Bikakis and Sellis 2016).

2.5 Article 5:

The last paper introduces conventional data conventional used for data visualization and the authors emphasize the extension of the traditional data visualization tools. However, these visual tools are far from enough in handling the big data sets. Figure 3 shows the difference between small and large size data with the challenges of scalability and dynamics. We could notice that with the data sets increasing, the challenges would become more severe, as the big sized data has high dimensions and requires faster data streaming speed to analyze and process for discovering patterns and insights.

To address the problems, the authors offer some potential solutions and present the new progress of methodologies to find the suitable big data visualization tools. The methodologies called(SWOT) have been shown in figure 4.

SWOT methods were summaries by the authors through analyzing the

Data type	Small, mid-sized	Big-sized
Static data	Well studied	Open issues type A
Dynamic data	Open issues type B	Highly challenging (A and B) >> A+B

Figure 3: The research status and challenge of visual analytics

factors of new approaches regarding big data visualization and this method is to help to distinguish the advantages and disadvantages when using new big data methodologies. They propose the method to facilitate developers to speed up the finding of new web-based tools(L. Wang, G. Wang, and Alexander 2015).

Strengths	Opportunities
• With the functions of visualization and interaction for visualizing data.	• Immersive visualization with virtual reality (VR) results in a better perception of data scape geometry and more intuitive data understanding.
• Able to visualize a variety of data types.	• The intrinsic human pattern recognition (or visual discovery) skills could be maximized.
Weaknesses	**Threats**
• There is room to improve in visualizing big data with high velocity or the combination of three high Vs (Volume + Velocity + Variety).	• Lack adequate visualization in a lot of Big Data applications.

Figure 4: The SWOT analysis of current big data visualization software tools

3 Research Gap:

From the literature reviews, we could find that the issues mentioned in the introduction section would influence the development of big data visualization systems. While several articles show that some challenges play a vital role in the system and some others not.

The paper *Big Data Visualization Tools: A Survey* by Caldarola, Enrico G and Rinaldi, Antonio M, proposes the solutions for the big data tools and survey current visualization systems from the eight criteria. However, the authors only consider the data dimensions including volume and variety, while the velocity is not in his scope(Enrico Giacinto Caldarola, Picariello, and Castelluccia 2015).

Int the paper *Big Data Exploration, Visualization and Analytics*, Bikakis, Nikos and Papastefanatos, George and Papaemmanouil, Olga reviewed the visualization solutions for the construction of big data system and illustrate that the system should ensure non-experts utilize it as easily as possible, while we could not find the specific implementations in this paper(Bikakis, Papastefanatos, and Papaemmanouil 2019).

4 methodologies

The problems and challenges has been illustrated in previous sections. Some solutions have also been discussed to address the challenges.

In *Big Data: A Survey*, Caldarola, Enrico Giacinto and Rinaldi, Antonio Maria illustrated the traditional SQL database has poor performance in handling variety and the large volume of data. And they give their criteria to survey the database tools. However, because of the limited number of criteria, some characteristics like scalability have not been evaluated in the system(Enrico Giacinto Caldarola and Antonio Maria Rinaldi 2015).

Because the big data visualization tools are a new era and the limited number of review paper. Readers could not find all solutions to all challenges introduced in previous parts. The paper *Big data and visualization: methods, challenges and technology progress* by Wang, Lidong and Wang, Guanghui and Alexander, Cheryl Ann propose *SWOT* methods to facilitate average developers to accelerate the speed of finding new web-based tools(L. Wang, G. Wang, and Alexander 2015). The methods could helm them easily distinguish the advantages and disadvantages while building the visualization

system.

References

Bikakis, Nikos, George Papastefanatos, and Olga Papaemmanouil (2019). *Big Data Exploration, Visualization and Analytics*.

Bikakis, Nikos and Timos Sellis (2016). "Exploration and visualization in the web of big linked data: A survey of the state of the art". In: *arXiv preprint arXiv:1601.08059*.

Caldarola, Enrico G and Antonio M Rinaldi (2017). "Big Data Visualization Tools: A Survey". In: *Proceedings of the 6th International Conference on Data Science, Technology and Applications*. SCITEPRESS-Science and Technology Publications, Lda, pp. 296–305.

Caldarola, Enrico Giacinto, Antonio Picariello, and Daniela Castelluccia (2015). "Modern enterprises in the bubble: Why big data matters". In: *ACM SIGSOFT Software Engineering Notes* 40.1, pp. 1–4.

Caldarola, Enrico Giacinto and Antonio Maria Rinaldi (2015). "Big Data: A Survey". In: *Proceedings of 4th International Conference on Data Management Technologies and Applications*. SCITEPRESS-Science and Technology Publications, Lda, pp. 362–370.

Wang, Lidong, Guanghui Wang, and Cheryl Ann Alexander (2015). "Big data and visualization: methods, challenges and technology progress". In: *Digital Technologies* 1.1, pp. 33–38.