ISBN 978-981-11-7861-0
Proceedings of 2018 the 8th International Workshop on Computer Science and Engineering (WCSE 2018)
Bangkok, 28-30 June, 2018, pp. 126-130
doi: 10.18178/wcse.2018.06.022

Competition or Cooperation? The Evolution of xSQL in Big Data Era

Tian Jing [1, 2 +]

[1] Postdoctoral Workstation of Credit Reference Center, the People's Bank of China
[2] Postdoctoral Research Station of Financial Research Institute, the People's Bank of China

Abstract. For decades, database technologies continue evolving to help people organize and manage data efficiently. Most really novel development of Hadoop/MapReduce has occurred to handle enormous volumes of unstructured data in last five years. Therefore, there are many discussions on which of the two techniques will finally survive during the competition. The paper comparatively provides several reflections concerning the evolution of database with the background of big data.

Keywords: database, hadoop, evolution, big data

1. Introduction

Every day, a great diversity of data explosively increases in different ways. It is estimated that the digital universe will be thirty-five zettabytes approximately in 2020. Over half a century, the landscape of database techniques change fast from traditional relational database management system (RDBMS, a.k.a. OldSQL) to NoSQL and NewSQL [1]. Nowadays, people turn to explore proper and high efficient methods to maximize mining the raw gritty data whose valuable content that is extracted by OldSQL at great cost. Big data analytics and Hadoop are rapidly emerging as the preferred solution to address business and technology trends that are disrupting traditional data management and processing. There is no doubt that database dominates the field of data management technologies as it evolves to various forms. Since new products of OldSQL, NoSQL and NewSQL, continue to emerge, it deserves to retrospect its development. In the paper, the author addresses his concerns in five aspects. The main contribution is that some crucial criteria are proposed to determine the most appropriate technical infrastructure of data governance and applications.

2. Commercial or Open-Source? The Variation of Openness

For a long time, cost is the principal factor when users plan to achieve databases for designing their systems particularly in the early stage of Internet. Therefore, most of enterprise users afford to purchase commercial products like Oracle, DB2 or SQLServer directly. Instead, open source database software like MySQL and PostgreSQL are popular in small business for they are free. The maturity of commercial offerings and reliability of service guarantee an enterprise to establish a new information system with high quality in a short time [2]. However, an increasing number of practitioners choose to utilize open-source database software as their core components of infrastructure within either intercontinental or small organization. This kind of trend could be explained that open source databases have been promoted significantly and applied successfully in a wide range. Actually they get competent on account of not only the price but also the performance compared with the commercial counterparts. Since users start to emphasize the capabilities of technological self-controlling on IT commodities, open-source software provides such an excellent way to master the essentiality from bottom to top. Moreover, developers are able to reconstruct a database by rewriting the source code freely in terms of business requirements.

+ Corresponding author. Tel.: +86-15010356525; fax: +86-10-57373157.
E-mail address: fellaini@vip.qq.com.

3. The Issue of Database Security Need to Be Highlighted

Most traditional RDBMS are deployed to run within Intranet isolated to the external network and Internet, for they are centralized and running in a single but powerful physical server. Whereas, more and more DBMS have supported to be adopted in cloud environment. A growing reliance on cloud services creates vulnerabilities for organizations. Vulnerabilities in cloud infrastructure provide the next frontier for cybercrime. Therefore, these databases are unable to be protected by classic security countermeasures.

Every month, a certain number of database threats is newly disclosed. For large vendors, prompt response and fixes released by the professional teams help to minimize users' loss. Unfortunately, open source software could not respond that timely as they are maintained by foundations, communities and independent programmers. Thus, the vulnerability attracts illegal attacks, establishing the black industrial chain. In 2016, tens of thousands of MongoDB (cloud) databases were hijacked and held for ransom after users left outdated versions exposed, without authentication turned on [3][4].

4. Survival or Extinction? What's the Future of RDBMS?

In the early 1970s, Ted Codd developed relational theory, based on the mathematics of set theory. Delivering rigor and accuracy to data access and manipulation, the mathematical basis and foundational theory of relational technology is unique within the world of database systems.

Ever since NoSQL began to prevail about ten years ago, some people have claimed that RDBMS would be shortly extinct. Nevertheless, that argument has never become true. The latest ranking provided by DB Engines (shown in Fig. 1) demonstrates that the top four databases resume RDBMS in February 2018. Moreover, six in ten of the list can be classified to RDBMS [5].

341 systems in ranking, February 2018

Rank			DBMS	Database Model	Score		
Feb 2018	Jan 2018	Feb 2017			Feb 2018	Jan 2018	Feb 2017
1.	1.	1.	Oracle	Relational DBMS	1303.28	-38.66	-100.55
2.	2.	2.	MySQL	Relational DBMS	1252.47	-47.24	-127.83
3.	3.	3.	Microsoft SQL Server	Relational DBMS	1122.04	-26.03	-81.42
4.	4.	4.	PostgreSQL	Relational DBMS	388.38	+2.19	+34.70
5.	5.	5.	MongoDB	Document store	336.42	+5.47	+0.92
6.	6.	6.	DB2	Relational DBMS	189.97	-0.30	+2.07
7.	7.	↑ 8.	Microsoft Access	Relational DBMS	130.07	+3.37	-3.32
8.	↑ 9.	↑ 10.	Redis	Key-value store	127.02	+3.88	+12.98
9.	↑ 10.	↑ 11.	Elasticsearch	Search engine	125.32	+2.76	+17.01
10.	↓ 8.	↓ 7.	Cassandra	Wide column store	122.78	-1.10	-11.60

Fig. 1: The ranking of database management systems in Feb. 2018.

Why does RDBMS survive in the fierce competition? It associates closely with these RDBMS vendors' endeavor in the sustaining promotion of compatibility with NoSQL. For example, Microsoft develops DocumentDB while Oracle releases Oracle NoSQL Database respectively. Multiple advantages of NoSQL databases have been integrated into the flagship products of the large enterprises via mergers and acquisitions. Therefore, both the functionalities and the performance of these alternatives, such as Oracle 12c, Microsoft SQL Server 2017 and IBM DB2 10.5, have been reinforced substantially.

5. Deep Inter-fusion Is the Prospective Trend

Many people presume that traditional RDBMS returns in the way of being NewSQL after the features of NewSQL have been unveiled and analyzed. The ACID-compatible NewSQL also supports processing massive information concurrently and distributed deployment in cloud, which combines the strengths of both OldSQL and NoSQL. Thereupon, it would provide a broad range for applications.

However, the classification of database is multi-dimensional without an authoritative standard. Fig. 2 illustrates a typical method of how to categorize the most common database products into a number of groups. Obviously, many products could be inclusively considered as one specific class such as either

relational or non-relational. The SPARIN principal provides a sequence of criteria for evaluating taxonomy, that is, “Scalability”, “Performance”, “Relaxed consistency”, “Agility”, “Intricacy” and “Necessity” [6].

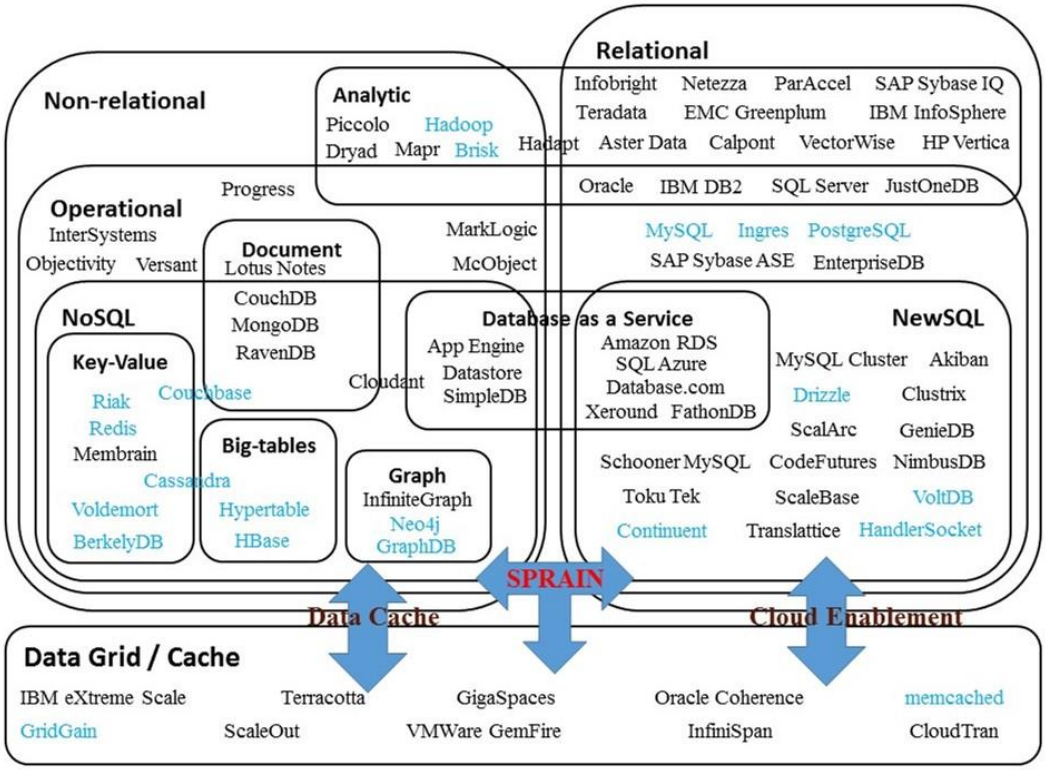

Fig. 2: A multi-dimensional classification of database products.

6. Big Data vs Data Warehouse, Which Is Better?

Big data technology brings a great challenge for traditional database methodologies such as data warehousing. In some circles today, there is a sort of “big data vs data warehouse” debate ongoing. Often the discussion casts Hadoop as the obvious heir apparent in the data processing world, with data warehouse cast as the traditional solution. There are pros and cons for both two alternatives.

6.1. Big Data Advocators

Advocators of big data assume that the disadvantages of data warehouse are hard to be conquered. It would be even tough for ETL and BI architecture to adjust accordingly when new sources are added [7].

On the contrary, big data techniques like Hadoop gains vast attractions and supports because of the low cost and high scalability. One hand, raw data could be easily loaded into a system without being discarded in terms of varieties or volumes. On the other hand, complicated variation could be readily processed. Many enterprises have migrated the data assets into Hadoop from a data warehouse especially in finance field.

6.2. Data warehouse advocators

People who prefer data warehouse argue that it is seemingly rigid and inflexible due to the IT policies rather than the architecture itself. In their opinions, data warehouse is the best scheme designed for analyzing data specially. At the same time, the unexpected indeterminacy would bring high risk once the analysis completely depends on a large volume of the targeting data that has not been pre-handled or modeled. So structured data should not always be a bad thing [8]. Although it is presumed that Hadoop would be mature enough to support all the capabilities owned by SQL and BI, it should be pointed out that the abundant achievements takes the top experts in computer science decades to establish databases and data warehouse.

History demonstrates that new technologies might be overestimated in a short time but be underestimated in a long time. Thus, more scientists and practitioners are prone to hold the opinion that Hadoop would be a vital component of the platform for analyzing data in the future rather than the exclusive one. Hadoop would be fused into data warehouse for tackling sharp increase of data amount, limitation of structural data processing, long elapse of operation, lack of support for rapid analysis and value discovery.

6.3. Hadoop against MPP or Hadoop plus MPP?

Both Hadoop and MPP have shined for it is purpose-built to handle complicated analysis and procession concurrently. However, they are working in distinct manners as Table 1 demonstrates [9][10].

TABLE I: A COMPARISON OF HADOOP AND MPP

Name	Hadoop	MPP
Data Features	Offline batch process, real-time query with simple logic, stream computing semi-structured and non-structured data	Offline batch process, real-time query with complex logic, real-time analysis, structured data (primarily) non-structured data (partly)
Data Scale	PB to several hundred-PB level	TB to PB level
Expansion	Prominently High Capability of being expanded to more than 5,000 nodes in a single cluster 100 PB data storage and processing	Moderately high A single cluster usually contains less than 100 nodes. The system will be instable once the nodes are increased to more than 100 and processing PB data.
Hardware	x86-based infrastructure	x86-based infrastructure
Complicated analysis across multi-table	Lack of high efficient indexes Lower performance of data storage and query optimization	Higher performance of complex analysis via indexing and partition
Real-time	Lower capability of real-time data processing due to lacking of optimization mechanism	Higher capability of real-time data processing
Difficulty of application development	Based on MapReduce Difficult	Base on SQL Easy
Future	The factual standard of big data Rapid development of the ecosystem Broad region of application	Technical Maturity Prone to be integrated to Hadoop
Cost of purchase	Low (about 5,000 RMB per TB) Open source (primarily)	High (30,000 – 50,000 RMB per TB) Commercial (such as Teradata, Netezza, GreenPlum, Vertica, ParAccel)
Cost of operation & maintenance	High Complicated in system maintenance, optimization and application development Lack of professionals	Low Mature in maintenance, optimization and application development of RDBMS Adequate professionals

It is clear that neither of them could be independent but high-efficient for an enterprise architecture, particularly applications demand the simultaneous capabilities of OLAP and OLTP. Thus, integrating Hadoop and MPP together becomes a vital theme of this research. The framework displayed in Fig. 3 facilitates enterprises to guarantee the performance of both OLAP and OLTP regardless of data formats.

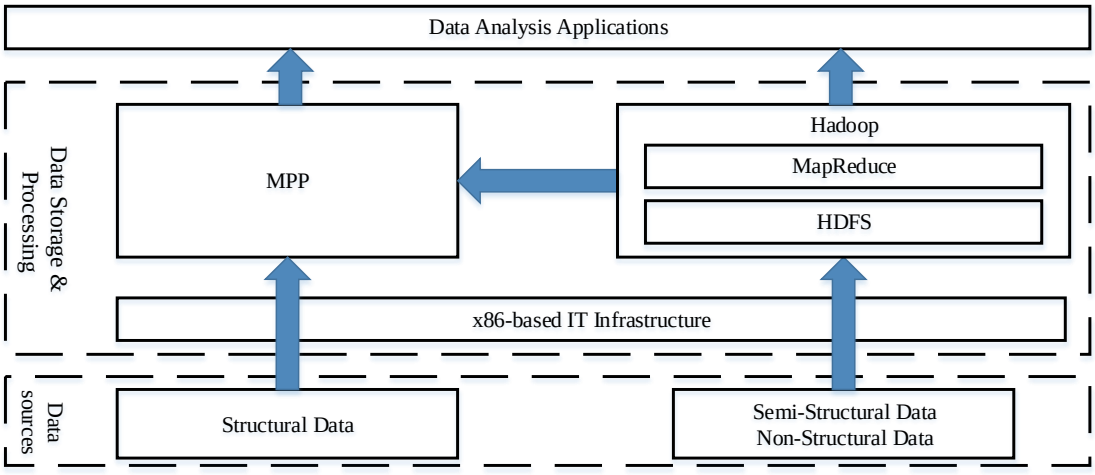

Fig. 3: A hybrid hierarchical framework for big data.

At the bottom layer, diverse patterns of data is introduced into the system from different sources. Structural data is directly brought to the MPP platform without complicated pre-process in batch. Semi-

structured data and non-structured data is loaded into HDFS and could be either provided to the upper service or persisted into the MPP database after standardization.

In the middle layer, Hadoop and MPP platforms are established based on the unified and x86-based infrastructure respectively. The two components are engaged to load and store big data collaboratively. The MPP database could receive the raw data either from original sources vertically or from Hadoop horizontally.

In the top layer, two primary types of applications could be developed. The OLTP business including query and data processing with high timeliness guarantee could be supported by MPP. The OLAP business including data pre-processing, data loading in batch. Hadoop would be an excellent supplement once the logic of OLTP is simple [11]. Generally, the scenarios involve reporting & analysis, interactive analysis, list-level data generation as well as data mining.

It indicates that the debate of simple replacement is somewhat misdirected and the discussion could lead companies away from the strategy they really should follow, namely a strategy of productive coexistence.

7. Summary

It is essential to evaluate data management and analysis technologies from time to time as daily data grows exponentially. Since OldSQL, NoSQL and NewSQL are developed successively, what will be the next xSQL? By comparing features, the MPP and Hadoop/MapReduce worlds are destined for unification. The integration of these two counterparts is just a temporary combination of parallel demanding on OLAP and OLTP. It is not hard to expect that they are more unified, rationalized and seamlessly integrated as a product in the future while enterprises can gain a competitive advantage by being adopters of the solution currently.

8. Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 61305054 & 61402016).

9. References

[1] Andrew Pavlo, Matthew Aslett, What's Really New with NewSQL? [J], *ACM Sigmod Record*, 2016, 45(2): 45-55

[2] https://blog.rdx.com/who-will-win-the-database-wars-open-source-vs-commercial-database-systems/

[3] Shivnandan Singh, Rakesh Kumar Rai, A Review Report on Security Threats on Database [J], *International Journal of Computer Science and Information Technologies*, 2014, 5(3): 3215-3219

[4] Michelle Malcher, Database Security. In: *DBA Transformations*. Apress, Berkeley, CA, 2018

[5] https://db-engines.com/en/ranking

[6] Matthew Aslett, NoSQL, NewSQL and Beyond: The answer to SPRAINed relational databases, https://blogs.the451group.com/information_management/2011/04/15/nosql-newsql-and-beyond/

[7] Lengdong Wu, Liyan Yuan, Jiahuai You, Survey of Large-Scale Data Management Systems for Big Data Applications [J], *Journal of Computer Science and Technology*, 2015, 30(1): 163-183

[8] Songting Chen, Cheetah: a high performance, custom data warehouse on top of MapReduce [C], *Proceedings of the VLDB Endowment*, 2010, 3(2):1459-1468

[9] Nikolay Golov, Lars Rönnbäck, Big Data normalization for massively parallel processing databases [J], *Computer Standards & Interfaces*, 2017, 54(2): 86-93

[10] Georgia Kougka, Anastasios Gounaris, The many faces of data-centric workflow optimization: a survey [J], *International Journal of Data Science and Analytics*, 2018, pp. 1-27

[11] Xin Lu, Fei Su, Haozhang Liu, Weiwei Chen, Xingzhou Cheng. A unified OLAP/OLTP big data processing framework in telecom industry [C], *Proceedings of 2016 16th International Symposium on Communications and Information Technologies*, pp. 290-295, 26-28 Sept. 2016, Qingdao, China.