SCIENCE CHINA
Information Sciences

• INSIGHT •

October 2016, Vol. 59 108101:1–108101:3
doi: 10.1007/s11432-015-0876-8

The golden age for popularizing big data

Lionel Ming-shuan NI[1], Jiang XIAO[2]* & Haoyu TAN[2]

[1]*Department of Computer and Information Science, the University of Macau, Macau, China;*
[2]*Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China*

Received March 4, 2016; accepted April 19, 2016; published online September 13, 2016

Citation Ni L M, Xiao J, Tan H Y, et al. The golden age for popularizing big data. Sci China Inf Sci, 2016, 59(10): 108101, doi: 10.1007/s11432-015-0876-8

The fast development of the Internet and Mobile Internet has opened up the era of big data. Both academia and industry have realized that the potential knowledge hidden in big data is huge and great business opportunities can be created by exploiting the data deeply and creatively. Back in 2000, as legacy RDBMS [1] could no longer meet the requirements of handling large volumes of data, new technologies that can provide efficient query and fast processing of big data were in urgent need. A big data technology pioneer, Google, has keenly realized that in a legacy RDBMS, either stand-alone or parallel, the conformity to a strict relational data model and the full supports of ACID requirements have become serious impediments to efficiently processing big data. Consequently, Google has started to deploy large clusters consisting of commodity hardware and developed a bunch of big data processing technologies since (at least) 2002. During the next five years, Google has successively developed a number of famous and extremely influential big data technologies such as a distributed file system called Google file system (GFS), a highly fault-tolerant parallel computing engine called MapReduce [2] and a NoSQL database called BigTable.

In 2008, Google has announced that MapReduce can sort 1 PB data in 68 s with 1000 servers. From then on, many companies have shown great interest in MapReduce and NoSQL databases. Under this circumstance, there were intensely inconclusive debates over RDBMS and NoSQL about which would be a better fit for big data marketplace [3]. There have been numerous benchmarking efforts in bridging the gap by the development of database system and data mining techniques. Promising to overcome major problems of traditional RDBMS and provide similar functionalities, NewSQL systems [4] such as Hive (on either MapReduce or Tez) and Spark SQL are attracting huge interest from both researchers and business users. The distinguished features of NewSQL are summerized by Turing Award 2014 winner Michael Stonebraker as "preserving SQL and ACID support, high per-node performance and scalability".

Nevertheless, a one-size-fits-all DBMS still seems impossible for the entire set of big data applications. For example, NoSQL will be a better fit when it comes to web applications of real-time query demand, such that it is vital to use tailor-made database techniques for different application scenarios. In addition to fundamental DBMS, data mining and machine learning for analytics and prediction have received increasing attention with innovative progress of big data technology during the recent 5 years. The breakthrough of deep learning technique came in 2006 which empowered the learning capability of conventional data min-

* Corresponding author (email: jiangxiao@ust.hk)

The authors declare that they have no conflict of interest.

ing models and soon expanded its usage to fields such as speech recognition, image recognition and natural language processing. The huge amount of training parameters of deep learning model pose challenges on the computing capacity. The advent of large-scale heterogeneous and parallel systems such as Tencent Mariana [5] that assembles heterogeneous architecture, GPU and distributed algorithm, makes the computing power never been greater.

Today, big data technology has made tremendous progress. For example, Baidu enlarged the storage size of single system from TB magnitude (Google's BigTable) to more than 1000 PB by Redis, and broke the 33 min TeraSort record created by Microsoft in 2004 with only 7.16 s in 2014. Tencent Mariana system took only five days to train the 1.5 million pictures and built up image classification model. In fact, now the efficiency issue of most applications has been well addressed by the set of big data toolkits, which consist of data extraction, data management and processing, data analytics and mining, data visualization and so on.

However, until recently, big data technologies still lack broader adoption by numerous traditional enterprises and small IT companies. Exploiting value from big data remains extremely challenging for them due to the technological barriers of implementation. Indeed, most successful big data practices are conducted by IT giants such as Google, Facebook, Baidu, Alibaba and Tencent who already have competitive advantages in ample technology strength, advanced equipment and abundant reserve of talents. The latest Gartner Hype Cycle for Emerging Technologies shows that big data technology has just passed from "Peak of Inflated Expectations" and slipped down into "Trough of Disillusionment" in 2015. We believe that the reason behind this turning point is that the challenges of efficiently processing big data have already been successfully addressed. The next goal of the development of big data technologies should be making them much easier to use or implement so that they can be adopted wildly by most institutions and companies. We therefore refer to the first 15 years of this century as "Big Data 1.0" which solves the issues of efficiently processing of big data and refer to 2015 onwards as "Big Data 2.0" which will solve the issues of realizing true values of data in boarder domains by popularizing big data technologies as much as possible.

In the following, we propose three key features that new big data technologies should provide in "Big Data 2.0".

End-to-end. End-to-end service refers to treating the whole process of turning data into value as a black box. In other words, an End-to-End big data system should have only two endpoints: one for the input of raw data and the other one for the output of final results. It is a key factor in successful implementation of a real-world big data project in which users may only have minimum IT skills. This will also help eliminate the communication cost between data scientists, engineers and domain experts. Guided by end-to-end principle, domain specialists only need to concentrate on the customers' requirements and explain the augments to scientist in a conceptual level, without penetration of sophisticated big data technologies. Take a telecom churn prediction system for example. Suppose domain experts define a potential "churner" as "a customer who has consumption reduction by $y\%$ in recent x months compared to preceding x month". In end-to-end manner, data scientists consider both x and y as optional augments for model with the ease of iterative discussion with experts. As a result, the cooperation between data scientists and domain experts will become less coupled, more agile and thereby more efficient.

All-in-cloud. All-in-cloud service refers to integrating all of the isolate services such as Hadoop as a Service, Machine Learning as a Service into Value as a Service lying on "end-to-end" foundation. It has the power to minimize the customization cost that keeps traditional enterprises from big data technology. Previous cloud-based efforts like Amazon elastic MapReduce Machine Learning as a Service have reduced deployment and maintenance investment. However, those services are only part-in-cloud as they just support a small parts of the whole system, far from the entire process of delivering data into value. By contrast, the novel all-in-cloud service enables the periodic ELT scheme to evolve into real-time input, which cuts off pre-loading stage and simplifies the implementation. In E-commence community, a recommendation system completely hosted in clouds can use an embedded code for data collection at its application front end and then put these data on the cloud via web API, and the recommendation results will be updated in real-time and provided to the vendor. It is clear that all-in-cloud service offers a more effective solution for data-value transition.

Interactive. Interactive visualization involves two aspects: (1) it creates visual representation (i.e., image or video) to make sense of raw data, and conveys the information in a clear yet fast way for users as well, and (2) it allows users to directly adjust analysis augments, receive instant feedback and predictive insights. It gains growing impor-

tance for making big data technologies accessible to a wide range of users. For example, Google Earth helps users to view the world from multiple perspectives and levels. As long as computers cannot completely replace humans for fast decision making, it is necessary to interactively integrate human intelligence in big data analysis.

Over the past 15 years, technologies developed in "Big Data 1.0" have been capable of providing efficient processing, management, and analysis of large amounts of data in most scenarios. Today, the age of "Big Data 2.0" which targets popularizing big data technologies has begun. Researchers and developers should embrace three features, namely "end-to-end", "all-in-cloud", and "interactive", to advance existing big data technologies. We believe that the only way to realize the true and great values of big data for both research and commercial purposes is to make the related technologies as friendly as possible to end users, most of which are domain exports other than data scientists or engineers.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 61300030).

References

1 Gary J, Liu D T, Nieto-Santisteban M, et al. Scientific data management in the coming decade. ACM SIGMOD Record, 2005, 34: 34–41
2 Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Commun ACM, 2008, 51: 107–113
3 Stonebraker M, Abadi D, Dewitt D J, et al. MapReduce and parallel DBMSs: friends or foes? Commun ACM, 2010, 53: 64–71
4 Xin R S, Rosen J, Zaharia M, et al. Shark: SQL and rich analytics at scale. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: ACM, 2013. 13–24
5 Zou Y Q, Jin X, Li Y, et al. Mariana: tencent deep learning platform and its applications. In: Proceedings of the VLDB Endowment, Hangzhou, 2014. 7: 1772–1777