Journal of Computer Sciences and Applications, 2015, Vol. 3, No. 6, 137-142
Available online at http://pubs.sciepub.com/jcsa/3/6/6

DOI:10.12691/jcsa-3-6-6

Future Trends in Cloud Computing and Big Data

Smaranika Mohapatra*, Jharana Paikaray, Neelamani Samal

Department of Computer Science & Engg, Gandhi Institute for Education & Technology, Bhubaneswar-752060, Odisha, India
*Corresponding author: smaranikka.88@gmail.com

Abstract In recent years, accompanied by lower prices of information and communications technology (ICT) equipment and networks, various items of data gleaned from the real world have come to be accumulated in cloud data centers. There are increasing hopes that analysis of this massive amount of data will provide insight that is valuable to both businesses and society. Since tens of terabytes (TBs) or tens of petabytes (PBs) of data, big data, should be handled to make full use of it, there needs to be a new type of technology different from ordinary ICT. It revolves around different areas of analytics and Big Data. Through a detailed survey, we identify possible gaps in technology and provide recommendations for the research community on future directions on Cloud-supported Big Data computing and analytics solutions.

Keywords: *big data, cloud computing analytics, data management*

Cite This Article: Smaranika Mohapatra, Jharana Paikaray, and Neelamani Samal, "Future Trends in Cloud Computing and Big Data." *Journal of Computer Sciences and Applications*, vol. 3, no. 6 (2015): 137-142. doi: 10.12691/jcsa-3-6-6.

1. Introduction

Society is becoming increasingly more instrumented and as a result, organizations are producing and storing vast amounts of data. The ability to cross-relate private information on consumer preferences and products with information from tweets, blogs, product evaluations, and data from social networks opens a wide range of possibilities for organizations to understand the needs of their customers, predict their wants and demands, and optimize the use of resources. This paradigm is being popularly termed as Big Data. Big data offers substantial value to organizations willing to adopt it, but at the same time poses a considerable number of challenges for the realization of such added value. An organization willing to use analytics technology frequently acquires expensive software licenses; employs large computing infrastructure; and pays for consulting hours of analysts who work with the organization to better understand its business, organize its data, and integrate it for analytics [1]. This joint effort of organization and analysts often aims to help the organization understand its customers' needs, behaviors, and future demands for new products or marketing strategies. Such effort, however, is generally costly and often lacks flexibility.

Cloud computing has been revolutionizing the IT industry by adding flexibility to the way IT is consumed, enabling organizations to pay only for the resources and services they use. The most often claimed benefits of Clouds include offering resources in a pay-as-you-go fashion, improved availability and elasticity, and cost reduction. Clouds can prevent organizations from spending money for maintaining peak-provisioned IT infrastructure that they are unlikely to use most of the time.

Figure 1 depicts the common phases of a traditional analytics workflow for Big Data. This paper discusses approaches and environments for carrying out analytics on Clouds for Big Data applications, survey approaches, environments, and technologies on areas that are key to Big Data analytics capabilities and discuss how they help building analytics solutions for Clouds. The paper is organized as follows:

in Section 2 we represent the Data Management and supporting architectures, Section 3 gives the details of model development. Section 4 is related to visualization and user interaction and Section 5 consists of business models, Section 6 consists of different challenges faced by analytics.

We focus on the most important technical issues on enabling Cloud analytics, but also highlight some of the non-technical challenges faced by organizations that want to provide analytics as a service in the Cloud.

2. Data Management

Data analysis is the most time-consuming and labour intensive jobs of analytics. Performing analytics on large volumes of data requires efficient methods to store, filter, transform, and retrieve the data. Cloud deployment models adopted by different enterprises, where Clouds can be for instance:

- Private: deployed on a private network, managed by the organization itself or by a third party. A private Cloud is suitable for businesses that require the highest level of control of security and data privacy
- Public: deployed off-site over the Internet and available to the general public. Public Cloud offers high efficiency and shared resources with low cost. The analytics services and data management are handled by the provider and the quality of service (e.g. privacy, security, and availability) is specified

in a contract. Organizations can leverage these Clouds to carry out analytics with a reduced cost or share insights of public analytics results.

- Hybrid: combines both Clouds where additional resources from a public Cloud can be provided as needed to a private Cloud. Customers can develop and deploy analytics applications using a private environment, thus reaping benefits from elasticity and higher degree of security than using only a public Cloud.

Regarding the availability of data and analytics models[2] some scenarios are envisaged as :(i) data and models are private; (ii) data is public, models are private; (iii) data and models are public; and (iv) data is private, models are public.

Cloud-enabled Big Data analytics needs to explore means to allocate and utilize these specialized resources in a proper manner. The rest of this section discusses existing solutions on data management irrespective of where data experts are physically located, focusing on storage and retrieval of data for analytics; data diversity, velocity and integration; and resource scheduling for data processing task.

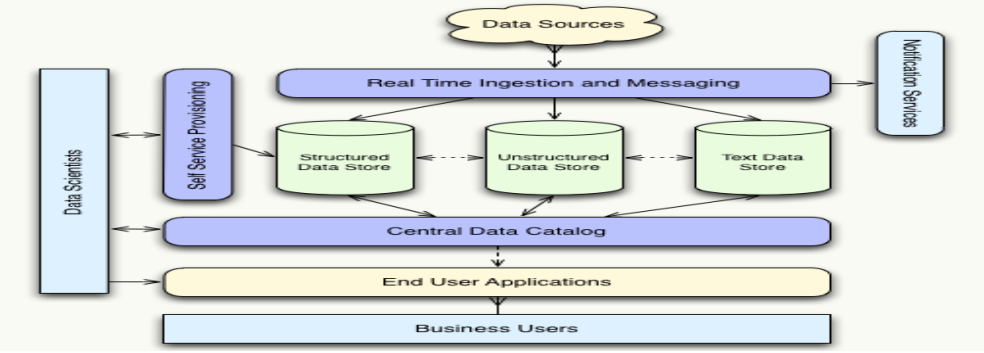

Figure 1. Big Data workflow

2.1. Data Variety and Velocity

Big Data is characterized by what is often referred to as a multi-V model, as depicted in Figure 2. Variety represents the data types, velocity refers to the rate at which the data is produced and processed, and volume defines the amount of data. Veracity refers to how much the data can be trusted given the reliability of its source[3], whereas value corresponds the monetary worth that a company can derive from employing Big Data computing. Although the choice of Vs used to explain Big Data is often arbitrary and varies across reports and articles on the Web – e.g. as of writing Viability is becoming a new V – variety, velocity, and volume[4,5] are the items most commonly mentioned. Regarding Variety, it can be observed that over the years, substantial amount of data has been made publicly available for scientific and business uses. Examples include repositories with government statistics1; historical weather information and forecasts; DNA sequencing; information on traffic conditions in large metropolitan areas; product reviews and comments; demographics comments, pictures, and videos posted on social network Web sites; information gathered using citizen-science platforms[6] and data collected by a multitude of sensors measuring various environmental conditions such as temperature, air humidity, air quality, and precipitation. An example illustrating the need for such a variety within a single analytics application is the Eco-Intelligence [7].

2.2. Data Storage

Several solutions were proposed to store and retrieve large amounts of data demanded by Big Data, some of which are currently used in Clouds. Internet-scale file systems such as the Google File System (GFS) [8] attempt to provide the robustness, scalability, and reliability that certain Internet services need. Other solutions provide object-store capabilities where files can be replicated across multiple geographical sites to improve redundancy, scalability, and data availability. Examples include Amazon Simple Storage Service (S3)[3],Nirvanix Cloud Storage,4 OpenStack Swift5.Nevertheless, in the context of Big Data, this approach of moving data to computation nodes would generate large ratio of data transfer time to processing time. Thus, a different approach is preferred, where computation is moved to where the data is. The same approach of exploring data locality was explored previously in scientific workflows and in Data Grids.

In the context of Big Data analytics, MapReduce presents an interesting model where data locality is explored to improve the performance of applications. Hadoop, an open source MapReduce implementation, allows for the creation of clusters that use the Hadoop Distributed File System (HDFS) to partition and replicate data sets to nodes where they are more likely to be consumed by mappers. In addition to exploiting concurrency of large numbers of nodes, HDFS minimizes the impact of failures by replicating data sets to a configurable number of nodes to develop an analytics platform to process Facebook's large data sets. The platform uses Scribe to aggregate logs from Web servers and then exports them to HDFS files and uses a Hive–Hadoop cluster to execute analytics jobs. The platform includes replication and compression techniques and columnar compression of Hive7 to store large amounts of data. Among the drawbacks of Cloud storage techniques and MapReduce implementations, there is the fact that they require the customer to learn a new set of APIs to build analytics solutions for the Cloud. Attempts have been made to provide hybrid solutions that incorporate MapReduce to perform some of the queries and data processing required by DBMS's. Cohen et al. [9] provide a parallel database design for analytics that supports SQL

and MapReduce scripting on top of a DBMS to integrate multiple data sources. Data processing and analytics capabilities are moving towards Enterprise Data Warehouses (EDWs), or are being deployed in data hubs to facilitate reuse across various data sets. With respect to EDW, some Cloud providers offer solutions that promise to scale to one petabyte of data or more. Amazon Redshift [10], for instance, offers columnar storage and data compression and aims to deliver high query performance by exploring a series of features, including a massively parallel processing architecture using high performance hardware, mesh networks, locally attached storage, and zone maps to reduce the I/O required by queries. Another distinctive trend in Cloud computing is the increasing use of NoSQL databases as the preferred method for storing and retrieving information. NoSQL adopts a non-relational model for data storage. Han et al. [11] presented a survey of NoSQL databases with emphasis on their advantages and limitations for Cloud computing. The survey classifies NoSQL systems according to their capacity in addressing different pairs of CAP (consistency, availability, partitioning). The survey also explores the data model that the studied NoSQL systems support.

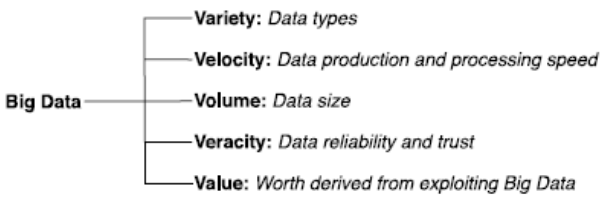

Figure 2. Some 'Vs' of Big Data

2.3. Data Integration Solutions

EDWs or Cloud based data warehouses, however, create certain issues with respect to data integration and the addition of new data sources. Standard formats and interfaces can be essential to achieve economies of scale and meet the needs of a large number of customers [12]. A Software as a Service (SaaS) solution that offers analytics functionalities on a subscription model; and appliances with the business analytics infrastructure, hence providing a model that allows a customer to migrate gradually from an on-premise analytics to a scenario with Cloud-provided analytics infrastructure, multi-flow solution for analytics that can be deployed on the Cloud. The multi-flow approach provides a range of possible analytics operators and flows to compose analytics solutions; viewed as workflows or instantiations of a multi-flow solution. The Business Process Execution Language (BPEL) data transition approach is used for data exchange by passing references to data between services to reduce the execution time and guarantee the correct data processing of an analytics process. A generic data Cloud layer is introduced to handle heterogeneous data Clouds, and is responsible for mapping generic operations to each Cloud implementation.

2.4. Data Processing and Resource Management

MapReduce [12] is one of the most popular programming models to process large amounts of data on clusters of computers. Hadoop [13] is the most used open source MapReduce implementation, also made available by several Cloud providers.Hadoop uses the HDFS file system to partition and replicate data sets across multiple nodes, such that when running a MapReduce application, a mapper is likely to access data that is locally stored on the cluster node where it is executing. Hadoop provides data parallelism and its data and task replication schemes enable fault tolerance, but what is often criticized about it is the time required to load data into HDFS and the lack of reuse of data produced by mappers. MapReduce is a model created to exploit commodity hardware, but when executed on reliable infrastructure, the mechanisms it provides to deal with failures may not be entirely essential.

Starfish [14], a data analytics system built atop Hadoop, focuses on improving the performance of clusters throughout the data lifecycle in analytics, without requiring users to understand the available configuration options. Starfish employs techniques at several levels to optimise the execution of MapReduce jobs. It uses dynamic instrumentation to profile jobs and optimises workflows by minimising the impact of data unbalance and by balancing the load of executions. The eXtreme Analytics Platform (XAP) [15] enables analytics supporting multiple data sources, data types (structured and unstructured), and multiple types of analyses. The target infrastructure of the architecture is a cluster running a distributed file system. A modified version of Hadoop, deployed in the cluster, contains an application scheduler (FLEX) able to better utilise the available resources than the default Hadoop scheduler. The analytics jobs are created via a high-level language script, called Jaql, that converts the high-level descriptive input into an analytics MapReduce workflow that is executed in the target infrastructure.

Realtime analysis of Big Data is a hot topic, with Cloud providers increasingly offering solutions that can be used as building blocks of stream and complex event processing systems. AWS Kinesis is an elastic system for real-time processing of streaming data that can handle multiple sources, be used to build dashboards, handle events, and generate alerts. It allows for integration with other AWS services.

3. Model Building

The data storage and Data as a Service (DaaS) capabilities provided by Clouds are important, but for analytics, it is equally relevant to use the data to build models that can be utilised for forecasts and prescriptions. Amazon EC2 as a hosting platform for the Zementis' ADAPA model [16] scoring engine. Predictive models, expressed in Predictive Model Markup Language (PMML),are deployed in the Cloud and exposed via Web Services interfaces. Users can access the models with Web browser technologies to compose their data mining solutions. Existing work also advocates the use of PMML as a language to exchange information about predictive models. Zementis [16] also provides technologies for data analysis and model building that can run either on a customer's premises or be allocated as SaaS using Infrastructure as a Service (IaaS) provided by solutions such as Amazon EC2 and IBM SmartCloud Enterprise [17]. Google Prediction API [18] allows users to create machine learning models to predict numeric values for a

new item based on values of previously submitted training data or predict a category that best describes an item. The prediction API allows users to submit training data as comma separated files following certain conventions, create models, share their models or use models that others shared. With the Google Prediction API, users can develop applications to perform analytics tasks such as sentiment analysis, purchase prediction, provide recommendations, analyze churn, and detect spam.

4. Visualisation and User Interaction

With the increasing amounts of data with which analyses need to cope, good visualisation tools are crucial. These tools should consider the quality of data and presentation to facilitate navigation [19]. The quality of data and presentation to facilitate navigation. The type of visualisation may need to be selected according to the amount of data to be displayed, to improve both displaying and performance. Visualisation can assist in the three major types of analytics: descriptive, predictive, and prescriptive. Users typically submit their jobs and wait until the execution is complete to download and analyse sample results to validate full runs. As this back and forth of data is not well supported by the Cloud, the authors issue a call to arms for both research and development of better interactive interfaces for Big Data analytics where users iteratively pose queries and see rapid responses. Several projects attempt to provide a range of visualisation methods from which users can select a set that suits their requirements. ManyEyes [20] from IBM allows users to upload their data, select a visualisation method – varying from basic to advanced – and publish their results. Users may also navigate through existing visualisations and discuss their findings and experience with peers. Selecting data sources automatically or semi-automatically is also an important feature to help users perform analytics. PanXpan [22] is an example of a tool that automatically identifies the fields in structured data sets based on user analytics module selection. FusionCharts [23] is another tool to allow users to visually select a subset of data from the plotted data points to be submitted back to the server for further processing. Besides visualisation of raw data, summarised content in form of reports are essential to perform predictive and prescriptive analytics. Several solutions have explored report generation and visualisation. For instance, SAP Crystal Solutions [21] provides BI functionalities via which customers can explore available data to build reports with interactive charts, what-if scenarios, and dashboards. The produced reports can be visualised on the Web, e-mail, Microsoft Office, or be embedded into enterprise applications. Another example on report visualisation is Cloud9 Analytics [24], which aims to automate reports and dashboards, based on data from CRM and other systems. It provides features for sales reports, sales analytics, and sales forecasts and pipeline management. By exploring history data and using the notion of risk, it offers customers clues on which projects they should invest their resources and what projects or products require immediate action.

5. Business Models and Non-technical Challenges

In addition to providing tools that customers can use to build their Big Data analytics solutions on the Cloud, models for delivering analytics capabilities as services on a Cloud have been discussed in previous work. Some of the potential business models proposed in their work include:

- **Hosting customer analytics jobs in a shared platform**: suitable for an enterprise or organisation that has multiple analytics departments. Traditionally, these departments have to develop their own analytics solutions and maintain their own clusters. With a shared platform they can upload their solutions to execute on a shared infrastructure, therefore reducing operation and maintenance costs.
- **A full stack designed to provide customers with end-to-end solutions**: appropriate for companies that do not have expertise on analysis. In this model, analytical service providers publish domain-specific analytical stream templates as services. The provider is responsible for hosting the software stack and managing the resources necessary to perform the analyses.
- **Expose analytics models as hosted services**: analytics capabilities are hosted on the Cloud and exposed to customers as services. This model is proposed to companies that do not have enough data to make good predictions.

Cloud-enabled Big Data analytics poses several challenges with respect to replicability of analyses. When not delivered by a Cloud, analytics solutions are customer-specific and models often have to be updated to consider new data. Cloud solutions for analytics need to balance generality and usefulness.

Other challenges

In business models where high-level analytics services may be delivered by the Cloud, human expertise cannot be easily replaced by machine learning and Big Data analysis [27]; in certain scenarios, there may be a need for human analysts to remain in the loop [25]. Management should adapt to Big Data scenarios and deal with challenges such as how to assist human analysts in gaining insights and how to explore methods that can help managers in making quicker decisions. Application profiling is often necessary to estimate the costs of running analytics on a Cloud platform. Users need to develop their applications to target Cloud platforms; an effort that should be carried out only after estimating the costs of transferring data to the Cloud, allocating virtual machines, and running the analysis. This cost estimation is not a trivial task to perform in current Cloud offerings. Although best practices for using some data processing services are available [26], there should be tools that assist customers to estimate the costs and risks of performing analytics on the Cloud. Data ingestion by Cloud solutions is often a weak point, whereas debugging and validation of developed solutions is a challenging and tedious process. the manner analytics is executed on Cloud platforms resembles the batch job scenario: users submit a job and wait until tasks are executed and then download the results. Once an analysis is complete, they download sample results that are enough to validate the

analysis task and after that perform further analysis. Current Cloud environments lack this interactive process, and techniques should be developed to facilitate interactivity and to include analysts in the loop by providing means to reduce their time to insight. Systems and techniques that iteratively refine answers to queries and give users more control of processing are desired. market research shows that inadequate staffing and skills, lack of business support, and problems with analytics software are some of the barriers faced by corporations when performing analytics [4]. These issues can be exacerbated by the Cloud as the resources and analysts involved in certain analytics tasks may be offered by a Cloud provider and may move from one customer engagement to another. terms such as Analytics as a Service (AaaS) and Big Data as a Service (BDaaS) are becoming popular. They comprise services for data analysis similarly as IaaS offers computing resources. However, these analytics services still lack well defined contracts since it may be difficult to measure quality and reliability of results and input data, provide promises on execution times, and guarantees on methods and experts responsible for analysing the data. Therefore, there are fundamental gaps on tools to assist service providers and clients to perform these tasks and facilitate the definition of contracts for both parties.

6. Summary and Conclusions

The amount of data currently generated by the various activities of the society has never been so big, and is being generated in an ever increasing speed. This Big Data trend is being seen by industries as a way of obtaining advantage over their competitors: if one business is able to make sense of the information contained in the data reasonably quicker, it will be able to get more costumers, increase the revenue per customer, optimise its operation, and reduce its costs. Nevertheless, Big Data analytics is still a challenging and time demanding task that requires expensive software, large computational infrastructure, and effort. Cloud computing helps in alleviating these problems by providing resources on-demand with costs proportional to the actual usage. Furthermore, it enables infrastructures to be scaled up and down rapidly, adapting the system to the actual demand. Although Cloud infrastructure offers such elastic capacity to supply computational resources on demand, the area of Cloud supported analytics is still in its early days. In this paper, we discussed the key stages of analytics workflows, and surveyed the state-of-the-art of each stage in the context of Cloud-supported analytics. Surveyed work was classified in three key groups: Data Management (which encompasses data variety, data storage, data integration solutions, and data processing and resource management), Model Building and Scoring, and Visualisation and User Interactions. For each of these areas, ongoing work was analysed and key open challenges were discussed. This survey concluded with an analysis of business models for Cloud-assisted data analytics and other non-technical challenges.

The area of Big Data Computing using Cloud resources is moving fast, and after surveying the current solutions we identified some key lessons:

- There are plenty of solutions for Big Data related to Cloud computing. Such a large number of solutions have been created because of the wide range of analytics requirements, but they may, sometimes, overwhelm non-experienced users. Analytics can be descriptive, predictive, prescriptive; Big Data can have various levels of variety, velocity, volume, and veracity. Therefore, it is important to understand the requirements in order to choose appropriate Big Data tools;
- It is also clear that analytics is a complex process that demands people with expertise in cleaning up data, understanding and selecting proper methods, and analysing results. Tools are fundamental to help people perform these tasks. In addition, depending on the complexity and costs involved in carrying out these tasks, providers who offer Analytics as a Service or Big Data as a Service can be a promising alternative compared to performing these tasks in-house;
- Cloud computing plays a key role for Big Data; not only because it provides infrastructure and tools, but also because it is a business model that Big Data analytics can follow (e.g. Analytics as a Service (AaaS) or Big Data as a Service (BDaaS)). However, AaaS/BDaaS brings several challenges because the customer and provider's staff are much more involved in the loop than in traditional Cloud providers offering infrastructure/ platform/software as a service.

Recurrent themes among the observed future work include

(i) The development of standards and APIs enabling users to easily switch among solutions and

(ii) The ability of getting the most of the elasticity capacity of the Cloud infrastructure. The latter includes expressive languages that enable users to describe the problem in simple terms whilst decomposing such high-level description in highly concurrent subtasks and keeping good performance efficiency even for large numbers of computing resources. If this can be achieved, the only limitations for an arbitrary short processing time would be market issues, namely the relation between the cost for running the analytics and the financial return brought for the obtained knowledge.

References

[1] X. Sun, B. Gao, Y. Zhang, W. An, H. Cao, C. Guo, W. Sun, Towards delivering analytical solutions in cloud: Business models and technical challenges, in: Proceedings of the IEEE 8th International Conference on e-Business Engineering (ICEBE 2011), IEEE Computer Society, Washington, USA, 2011, pp. 347-351.

[2] P.R. Krishna, K.I. Varma, Cloud Analytics: A Path Towards Next Generation Affordable BI, White paper, Infosys (2012).

[3] P.S. Yu, On mining big data, in: J. Wang, H. Xiong, Y. Ishikawa, J. Xu, J. Zhou (Eds.), Web-AgeInformation Management, in: Lecture Notes in Computer Science, vol. 7923, Springer-Verlag, Berlin, Heidelberg, 2013, p. XIV.

[4] P. Russom, Big Data Analytics, TDWI best practices report, The Data Warehousing Institute (TDWI) Research (2011).

[5] X. Sun, B. Gao, Y. Zhang, W. An, H. Cao, C. Guo, W. Sun, Towards delivering analytical solutions in cloud: Business models

and technical challenges, in: Proceedings of the IEEE 8th International Conference on e-Business Engineering (ICEBE 2011), IEEE Computer Society, Washington, USA, 2011, pp. 347-351.

[6] R. Bonney, J.L. Shirk, T.B. Phillips, A. Wiggins, H.L. Ballard, A.J. Miller-Rushing, J.K. Parrish, Next steps for citizen science, Science 343 (2014) 1436-1437.

[7] X. Zhang, E. Zhang, B. Song, F. Wei, Towards Building an Integrated Information Platform for Eco-city, in: Proceedings of the 7th International Conference on e Business Engineering (ICEBE 2010), 2010, pp. 393-398.

[8] S. Ghemawat, H. Gobioff, S.-T. Leung, The google file system, in: Proceedings of the 9th ACM Symposium on Operating Systems Principles (SOSP 2003), ACM, New York, USA, 2003, pp. 29-43.

[9] J. Cohen, B. Dolan, M. Dunlap, J.M. Hellerstein, C. Welton, MAD skills: new analysis practices for big data, Proceedings of the VLDB Endow 2 (2) (2009) 1481-1492.

[10] Amazon redshift, http://aws.amazon.com/redshift/.

[11] J. Han, H. E, G. Le, J. Du, Survey on NoSQL database, in: 6th International Conference on Pervasive http://media.amazonwebservices.com/AWS_Amazon_ER_Best_Practices. pdf.

[12] J. Dean, S. Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, Communications of the ACM 51(1).

[13] Apache Hadoop, http://hadoop.apache.org.

[14] H. Herodotou, H. Lim, G. Luo, N. Borisov, L. Dong, F.B. Cetin, S. Babu, Starfish: A Self-tuning System for Big Data Analytics, in: Proceedings of the 5th Biennial Conference on Innovative Data Systems Research (CIDR 2011), 2011, pp. 261-272.

[15] A. Balmin, K. Beyer, V. Ercegovac, J.M.F. Ozcan, H. Pirahesh, E. Shekita, Y. Sismanis, S. Tata, Y. Tian, A platform for eXtreme Analytics, IBM J. Res. Dev. 57 (3-4) (2013) 4:1-4:11.

[16] Zementis – adaptive decision technology, http://www.zementis.com (2012).

[17] IBM SmartCloud Enterprise, http://www-935.ibm.com/services/us/en/ cloud-enterprise/ (2012).

[18] Google Prediction API, https://developers.google.com/prediction/.

[19] P. Deyhim, Best practices for Amazon EMR, White paper, Amazon (2013). URL http://media.amazonwebservices.com/AWS_Amazon_EMR_Best_Practices. pdf.

[20] F.B. Viegas, M. Wattenberg, F. van Ham, J. Kriss, M. McKeon, ManyEyes: a Site for Visualization at Internet Scale, IEEE Trans. Vis. Comput. Graphics 13 (6) (2007) 1121-1128.

[21] SAP Crystal Solutions,http://www.crystalreports.com/.

[22] panXpan, https://www.panxpan.com.

[23] FusionChars, http://www.fusioncharts.com/.

[24] Cloud9 Analytics, http://www.cloud9analytics.com.

[25] D. Lazer, R. Kennedy, G. King, A. Vespignani, The Parable of google flu: Traps in big data analysis, Science 343 (2014) 1203-1205.

[26] P. Deyhim, Best practices for Amazon EMR, Whitepaper , Amazon (2013). URL。

[27] A. McAfee, E. Brynjolfsson, Big data: The management revolution, Harv. Bus. Rev. (2012) 60-68.