Bigdata Computing and Clouds

Shreelakshmi C.M
Department of Computer Science and Engineering
GSSS Institute of Engineering and Technology for Women Mysuru, Karnataka, India

Abstract- *Big data computing demands a large storage and computing for data processing and creation that could be delivered from on-premise or clouds infrastructures. This paper provides the evolution of big data computing, differences between traditional data warehousing and big data, taxonomy of big data computing and underpinning technologies, integrated platform of big data and clouds known as big data clouds, layered architecture and components of big data cloud, and finally open-technical challenges and future directions. Approaches and environments for carrying out analytics on Clouds for Big Data applications are discussed. It revolves around four important areas of analytics and Big Data, namely (i) data management and supporting architectures; (ii) model development and scoring; (iii) visualisation and user interaction;and (iv) business models. Through a detailed survey, we identify possible gaps in technology andprovide recommendations for the research community on future directions on Cloud-supported Big Datacomputing and analytics solutions.*

Keywords- Bigdata,Cloud computing,Data managent,Bigdata computing

I. INTRODUCTION

Big data computing is an emerging data science paradigm of multidimensional information mining for scientific discovery and business analytics over large-scale infrastructure. The data collected/produced from several scientific explorations and business transactions often require tools to facilitate efficient data management, analysis, validation, visualization, and dissemination while preserving the intrinsic value of the data [1][5]. New advancements in semiconductor technologies are eventually leading to faster computing, large-scale storage, and faster and powerful networks at lower prices, enabling large volumes of data preservation and utilization at faster rate. Recent advancements in cloud computing technologies are enabling to preserve every bit of the gathered and processed data, based on subscription models, providing high availability of storage and computation at affordable price. Conventional data warehousing systems are based on predetermined analytics over the abstracted data and employ cleansing and transforming into another database known as data marts – which are periodically updated with the similar type of rolled-up data. However, big data systems work on non predetermined analytics; hence, no need of data cleansing and transformations procedures.

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 organisations to understand the needs of their customers, predict their wants and demands, and optimise the use of resources. This paradigm is being popularly termed as Big Data. Despite the popularity on analytics and Big Data, putting them into practice is still a complex and time consuming endeavour. Big Data offers substantial value to organisations willing to adopt it, but at the same time poses a considerable number of challenges for the realisation of such added value.

An organisation willing to use analytics technology frequently acquires expensive software licences; employs large computing infrastructure; and pays for consulting hours of analysts who work with the organisation to better understand its business, organise its data, and integrate it for analytics [3]. This joint effort of organization and analysts often aims to help the organisation understand its customers' needs, behaviours, and future demands for new products or marketing strategies.

A. Data management

One of the most time-consuming and labour-intensive tasks of analytics is preparation of data for analysis; a problem often exacerbated by Big Data as it stretches existing infrastructure to its limits.Performing analytics on large volumes of data requires efficient methods to store, filter, transform, and retrieve the data. Some of the challenges of deploying data management solutions on Cloud environments have been known for some time [4][5][6], and solutions to perform analytics on the Cloud face similar challenges. Cloud analytics solutions need to consider the multiple Cloud deployment models adopted by enterprises, where Clouds can be for instance:

- Private: deployed on a private network, managed by the organisation 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. In such conditions, this type of Cloud infrastructure can be used to share the

services and data more efficiently across the different epartments of a large enterprise.

- Public: deployed off-site over the Internet and available to the general public. Public Cloud offers high efficiency and sharedresources 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. Organisations 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.

B. Big database

Big data addresses the data management and analysis issues in several areas of business intelligence, engineering, and scientific explorations. Traditional databases segregate the operational and historical data for operational and analysis reasoning, which are mostly structured. However, big data bases address the data analytics over an integrated scale out compute and data platform for unstructured data in near real time. Figure 1 depicts several issues in traditional data (data warehousing online transaction processing/online analytical processing) and big data technologies that are classified into major areas like infrastructure, data handling, and decision support software as described in the succeeding texts.

- Decision support/intelligent software tools: Big data technologies address various decision supporting tools for searching the large data volumes and construct the relations and extract the information based on several analytical methods. These tools would address several machine-learning techniques, decision support systems, and statistical modeling tools.

- Large-scale data handling: rapidly growing data distributed over several storages and compute nodes with multidimensional data formats.

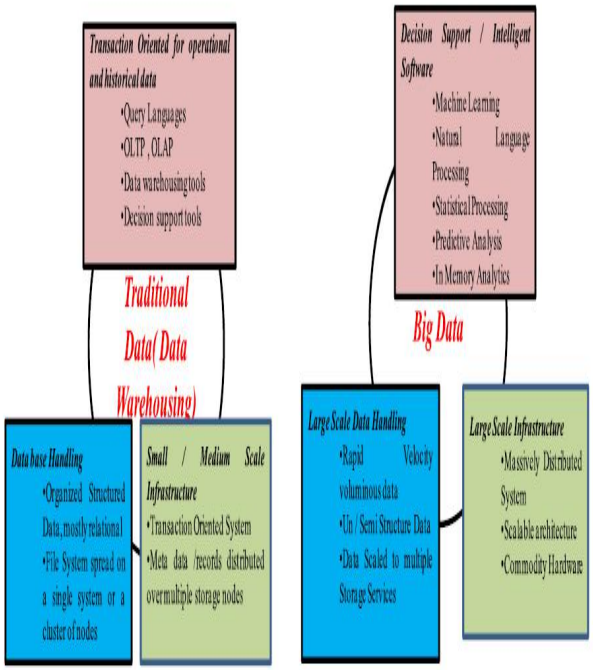

Figure 1. Big data versus traditional data (data warehousing) models. OLTP, online transactional processing; OLAP, online analytical processing.

Figure 1 depicts several issues in traditional data (data warehousing online transaction processing/online analytical processing) and big data technologies that are classified into major areas like infrastructure, data handling, and decision support software as described in the succeeding texts.

Decision support/intelligent software tools: Big data technologies address various decision supporting tools for searching the large data volumes and construct the relations and extract the information based on several analytical methods. These tools would address several machine-learning techniques, decision support systems, and statistical modeling tools.

- Large-scale data handling: rapidly growing data distributed over several storages and compute nodes with multidimensional data formats.

Table I. Traditional data warehousing versus big data issues

Serial nos.	Property	Traditional data warehousing	Big data-specific issues
1	Data volume	Data are segregated into operational and historical data. Applies extract, transformation, and load mechanisms for processing. As the data volumes are increased, the historical data are filtered from warehouse system for faster database queries.	High volume of data from several sources like web, sensor networks, social networks, and scientific experiments. Capable of handling operational and historical data together, which could be replicated on multiple storage devices for high availability and throughput.
2	Speed	Transaction-oriented and the data in turn generated from the transactions are low.	High data growth due to several sources like web and scientific sensors streaming experiments.
3	Data formats	Semi/structured data like XML and relational.	Multi-structured data handling such as relational, and un/semi-structured such as text, XML, video streaming, and so on.
4	Applicable platforms	Online transactional processing, relational database management system.	Big data analytics, text mining, video analytics, web log mining, scientific data exploration, intrinsic information extractions, graph analytics, social networking, in-memory analytics, and statistical and predictive analytics.
5	Programming methodologies/ languages	Query language like SQL.	Data-intensive computing languages for batch processing and stream computing like Map/Reduce and NoSQL programming.
6	Data backup/archival	Files/relational data need to have data backup procedures or mechanisms. Traditional data works on regular, incremental, and full backup mechanisms that are already established.	Due to large and high speeds of the data growth rates, the conventional methods are not adequate; hence, techniques such as differential backup mechanisms need to be explored.
7	DR	Data are replicated at several places to address the disaster.	DR techniques could be separated from mission critical and non critical data.
8	Relationship with clouds	Relational data bases/data warehousing tools as services over cloud infrastructures.	On-demand big data infrastructure setup, analytic services by several cloud, and big data providers.
9	Data deduplication	Applicable to transactional record deduplication while merging database records.	File and block level deduplication mechanisms need to be explored for continuous growing and stream-oriented data.
10	System users	Administrators, developers, and end users.	Data scientists and analytics end users.
11	Theorem applicable	Follows CAP theorem [20] with ACID [21] properties.	Follows CAP theorem with BASE properties [22].

DR, disaster recovery; SQL, structured query language; BASE, basically available, soft state, and eventually consistent; CAP, consistency, availability, and partition tolerance.

II. BACKGROUND AND METHODOLOGY

Organisations are increasingly generating large volumes of data as result of instrumented business processes, monitoring of user activity [7],[8], web site tracking, sensors, finance, accounting, among other reasons. With the advent of social network Web sites, users create records of their lives by daily posting details of activities they perform, events they attend, places they visit, pictures they take, and things they enjoy and want. This data deluge is often referred to as Big Data[9][10][11]; a term that conveys the challenges it poses on existing infrastructure with respect to storage, management, interoperability, governance, and analysis of the data. In today's competitive market, being able to explore data to understand customer behaviour, segment customer base, offer customised services, and gain insights from data provided by multiple sources is key to competitive advantage. Although decision makers would like to base their decisions and actions on insights gained from this data , making sense of data, extracting non obvious patterns, and using these patterns to predict future behaviour are not new topics. Knowledge Discovery in Data (KDD) [12] aims to extract non obvious information using careful and detailed analysis and interpretation. Data mining [13][14], more specifically, aims to discover previously unknown interrelations among apparently unrelated attributes of data sets by applying methods from several areas including machine learning, database systems, and statistics. Analytics comprises techniques of KDD, data mining, text mining, statistical and quantitative analysis, explanatory and predictive models, and advanced and interactive visualisation to drive decisions and actions [15].

Figure 2 depicts the common phases of a traditional analytics workflow for Big Data. Data from various sources, including databases, streams, marts, and data warehouses, are used to build models. The large volume and different types of the data can demand pre-processing tasks for integrating the data, cleaning it, and filtering it. The prepared data is used to train a model and to estimate its parameters. Once the model is estimated, it should be validated before its consumption. Normally this phase requires the use of the original input data and specific methods to validate the created model. Finally, the model is consumed and applied to data as it arrives. This phase, called model scoring, is used to generate predictions, prescriptions, and recommendations. The results are interpreted and evaluated, used to generate new models or calibrate existing ones, or are integrated to pre-processed data. Analytics solutions can be classified as descriptive, predictive, or prescriptive as illustrated in Fig. 3. Descriptive analytics uses historical data to identify patterns and create management reports; it is concerned with modelling past behaviour. Predictive analytics attempts to predict the future by analysing current and historical data. Prescriptive solutions assist analysts in decisions by determining actions and assessing their impact regarding business objectives, requirements, and constraints. Despite the hype about it, using analytics is still a labour intensive endeavour. This is because current solutions for analytics are often based on proprietary appliances or software systems built for general purposes. Thus, significant effort is needed to tailor such solutions to the specific needs of the organisation, which includes integrating different data sources and deploying the software on the company's hardware (or, in the case of appliances, integrating the appliance hardware with the rest of the company's systems) [18]. Such solutions are usually developed and hosted on the customer's premises, are generally complex, and their operations can take hours to execute. Cloud computing provides an interesting model for analytics, where solutions can be hosted on the Cloud and consumed by customers in a pay-as-you-go fashion. For this delivery model to become reality, however, several technical issues must be addressed, such as data management, tuning of models, privacy, data quality, and data currency. This work highlights technical issues and surveys existing work on solutions to provide analytics capabilities for Big Data on the Cloud. Considering the traditional analytics workflow presented in Fig.1, we focus on key issues in the phases of an analytics solution. With Big Data it is evident that many of the challenges of Cloud analytics concern data management, integration, and processing. Previous work has focused on issues such as data formats, data representation, storage, access, privacy, and data

quality. Security is certainly a key challenge for hosting analytics solutions on public Clouds. We consider, however, that security is an extensive topic and would hence deserve a study of its own.

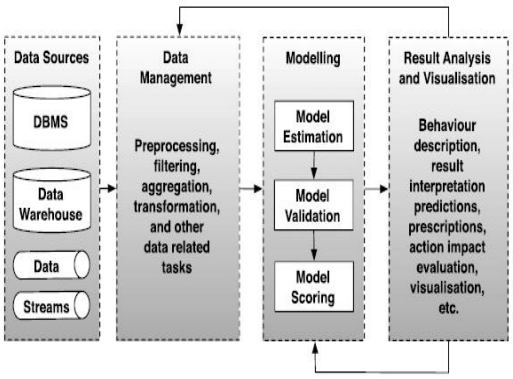

Figure 2: Overview of Analytics Workflow for Bigdata

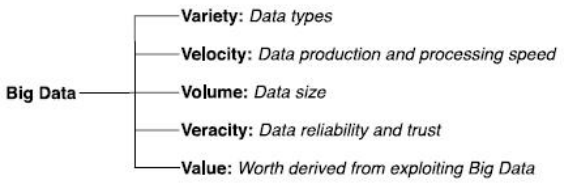

Fig. 3. Some 'Vs' of Big Data.

Figure 3: Categories Of Analytics

III. BIG DATA IN CLOUDS: AN INTEGRATED BIG DATA AND CLOUD PLATFORM

Big data in clouds is a new generation data-intensive platform for quickly building the analytics and deploying over an elastically scalable infrastructure. Based on the services rendered to the end users, these are broadly classified into four types as described in the succeeding texts.

- Public big data clouds: large-scale data organization and processing over the elastically scalable clouds infrastructure. The resources are served over Internet as pay-as-go computing models. The examples include Amazon big data computing in clouds [17], Windows Azure HDInsight [20], RackSpace Cloudera Hadoop [19], and Google cloud platform of big data computing [20].
- Private big data clouds: deployment of big data platform within the enterprise over a virtualized infrastructure, with a greater control and privacy to the single organization.
- Hybrid big data clouds: federation of public and private big data clouds for scalability, disaster recovery, and high availability. In this deployment, the private tasks can be migrated to the public infrastructure during peak workloads.
- Big data access networks and computing platform: Integrated platform of data, computing, and analytics delivered as services by multiple distinct providers.Big data computing in clouds also known as 'big data clouds' is data-intensive analytics platform of large-scale, distributed compute, and storage infrastructures.

The features of big data clouds are as follows:

(i) large-scale distributed compute and data storages: wide range of computing facilities with seamless access to scalable storage repositories and data services;
(ii) information-defined data storage: metadata-based data access instead of path and filenames;
(iii) distributed virtual file system: File system could be dynamically created and mapped to the computing cluster;
(iv) seamless access of computing and data: transparent access to large-scale data and compute resources; dynamic selection of data containers and compute resources: able to handle dynamic creation of virtual machines and able to access large-scale distributed data sources increasing the data location proximity;
(v) high performance data and computation: Compute and data should be high performance driven;
(vi) multidimension data handling: support for several forms of data with necessary tools for processing;
(vii) analytics platform services: able to develop, deploy, and use analytics over the environment;
(viii) high availability of computing and data: replication mechanisms for both computing and data; and
(ix) platform for data-intensive computing: support for both traditional and emerging data-intensive computing models and scalable deployment and execution of applications.

The content from several sources like social media, web logs, scientific studies,sensor networks, business transactions, and so on are growing rapidly. Deriving useful information for decision-making from such large data, fusing the information from several sources would be a challenging task. The elements of big data access network are as follows: data services, big data computing platform, data scientist, and computing cloud described in the succeeding texts.

- Data and platform services: Several providers, those who provide services for accessing both data and platform services for computing on the data, for example, Google data APIs (GData) , provide protocols for reading and writing data on the web for several services like content

API for shopping, Google analytics, spreadsheets, and YouTube.

- Big data computing platform: platform for managing the various data sources including data management, access, programming models, schedulers, security, and so on. The platform includes various tools for accessing other data platforms using streaming, web services, and APIs. Other data platforms include data services from relational data stores, Google data, social networking, and so on.
- Data scientist: analytics developers having access to the computing platform.
- Computing cloud: computing infrastructure from private/public/hybrid clouds.

Figure 4 describes Integrated cloud and big data compute network. USGS, United States Geological Survey; HTTP, hypertext transfer protocol; REST, representational state transfer.

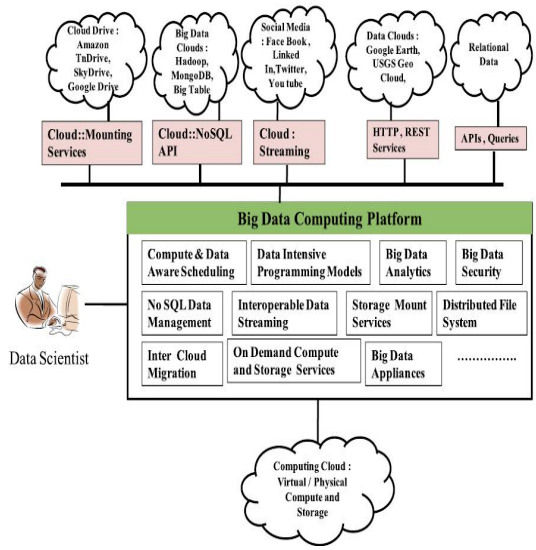

Figure 4 Integrated cloud and big data compute network. USGS, United States Geological Survey; HTTP, hypertext transfer protocol; REST, representational state transfer.

A. Big data clouds for the enterprise

Big data clouds enable enterprises to save money, grow revenue, and achieve many other business objectives in any vertical by quickly building their big data databases and writing analytics for mining the information.

The benefits of big data clouds for the enterprises are mentioned in the succeeding texts.

- Build new applications: Big data clouds would allow enterprises to collect billions of real-time data points on its products, resources, or customers and then repackage that instantaneously to optimize customer experience or resource utilization.
- Improve the effectiveness and minimize the cost: Big data clouds offer services and pay-as-go consumption model similar to cloud services. This pricing model would effectively reduce both the cost of the applications development by minimizing the cost of development tools.
- Realize new sources of information and build applications to gain competitive advantage: The information could be quickly fused from several big data databases and rapidly build applications for several platforms like hand-held and mobile devices.
- Increase in customer loyalty: Increase in the amount of data sharing within the organization and the speed with which it is updated allows businesses and other organizations to more rapidly and accurately respond to customer demand.

B. Elements of big data and cloud

Big data and traditional data warehousing mechanisms differ with each other in several ways like large-scale data organization, and querying followed by platforms and tools to the data scientists for analytics development.

(i) Big data infrastructure services: This layer offers core services such as compute, storage, and data services for big data computing as described in the succeeding texts.

(a) Basic storage service: provides basis services for data delivery that is organized either on physical or virtual infrastructure and supports various operations like create, delete, modify, and update with a unified data model supporting various types of data.

(b) Data organization and access service: Data organization provides management and location of data resources for all kinds of data, and selection, query transformation, aggregation and representation of query results, and semantic querying for selecting the data of interest.

(c) Processing service: mechanism to access the data of interest, transferring to the compute node, efficient scheduling mechanism to process the data, programming methodologies, and various tools and techniques to handle the variety of data formats.

The elements of big data infrastructure services are described in the succeeding texts.

- Computing clouds: on-demand provisioning of compute resources, which could expand or shrink based on the analytics requirements.

- Storage cloud: large volume of storage offered over the network. The storages offered include file system, block storages, and object-based storage. Storage clouds offer to create file system of choice and also elastically scalable. Storage clouds can be accessed based on the pricing models that are usually based on data volumes and transactions/data transfer. The several services offered by storage clouds are raw, block, and object-based storages
- Data clouds: Data clouds are similar to storage clouds; however, unlike storage space delivery, they offer data as a service. Data clouds offer tools and techniques to publish the data, tag the data, discovery the data, and process the data of interest. Data clouds operate on domainspecific data leveraging the storage clouds to serve data as a service based on the four steps of 'standard scientific model' [40] such as data collection, analysis, analyzed reports, and longterm preservation of the data. (ii) Big data platform services: This layer offers schedulers, query mechanisms for data retrieval, and data-intensive programming models to address several big data analytic problems. (iii) Big data analytics services: big data analytics as services over big data cloud infrastructure. The services would be offered to enterprises based on service-level agreements (SLAs) meeting QoS parameters.

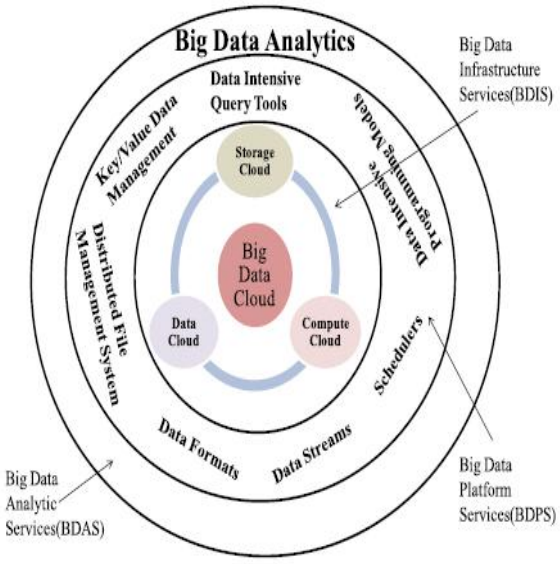

Figure 5 Big data Cloud Components

IV. CONCLUSION

Big data computing is an emerging platform for data analytics to address large-scale multidimensional data for knowledge discovery and decision-making. In this paper, we have studied, characterized, and categorized several aspects of big data computing systems. Big data technology is evolving and changing the present traditional data bases with effective data organization, large computing, and data workloads processing with new innovative analytics tools bundled with statistical and machine-learning techniques. With the maturity of cloud computing technologies, big data technologies are accelerating in several areas of business, science, and engineering to solve data intensive problems.

REFERENCES

[1] Dean J, Ghemawat S. MapReduce: simplified data processing on large cluster. Communications of the ACM 2008; 51(1): 107–113.

[2] Advancing discovery in science and engineering, the role of basic computing research, http://www.cra.org/ccc/files/ docs/Natl_Priorities/web_data_spring.pdf [last accessed 10 August 2014].

[3] 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.

[4] D.S. Katz, S. Jha, M. Parashar, O. Rana, J.B. Weissman, Survey and Analysis of Production Distributed Computing Infrastructures, CoRR abs/1208.2649.

[5] S. Sakr, A. Liu, D. Batista, M. Alomari, A survey of large scale data management approaches in cloud environments, IEEE Communications Surveys Tutorials 13 (3) (2011) 311–336.

[6] Storm: distributed and fault-tolerant realtime computation, http://storm. incubator.apache.org.

[7] Attention, shoppers: Store is tracking your cell, New York Times. URL http://www.nytimes.com/2013/07/15/business/attention-shopper-storesare-tracking-your-cell.html.

[8] Unlocking Game-Changing Wireless Capabilities: Cisco and SITA help Copenhagen Airport Develop New Services for Transforming the Passenger Experience, Customer case study, CISCO (2012). URL http://www.cisco.com/ en/US/prod/collateral/wireless/c36_696714_00_copenhagen_airport_cs.pdf.

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

[10] B. Franks, Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics, first ed., in: Wiley and SAS Business Series, Wiley, 2012.

[11] Buyya R, Shin Yeo C, Venugopal S, Brobergand J, Brandic I. Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility. Future Generation Computer Systems 2009; 25(6): 599–616.

[12] Fair scheduler, http://hadoop.apache.org/docs/r1.2.1/fair_scheduler.pdf [last accessed 20 February 2015].

[13] InfiniteGraph: the distributed graph database, a performance and distributed performance benchmark of InfiniteGraph and a leading open source graph database using synthetic data, infinite graph, white paper from objectivity, http://www.objectivity.com/wp-content/uploads/Objectivity_WP_IG_Distr_Benchmark.pdf, 2012 [last accessed 20 December 2014].

[14] Aggarwal CC, Zhai C. Probabilistic Models for Text Mining: In Mining Text Data. Kluwer Academic Publishers: Netherlands, 2012: 257–294.

[15] Karlof H, Suri S, Vassilvitskii S. A model of computation for MapReduce, in: Proceedings of the 21st ACM-SIAM Symposium on Discrete Algorithms (SODA 2010), Austin, Texas, January 2010.

[16] InfiniteGraph: the distributed graph database, a performance and distributed performance benchmark of InfiniteGraph and a leading open source graph database using synthetic data, infinite graph, white paper from objectivity, http://www.objectivity.com/wp-content/uploads/Objectivity_WP_IG_Distr_Benchmark.pdf, 2012 [last accessed 20 December 2014].

[17] Chang F, Dean J, Ghemawat S, Heish W. C., Wallach D. A, Burrows M, Chandra T, Fikes A, Gruber R. E. Big table: a distributed storage system for structured data, in: Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI 2006), Seattle, WA, Nov 2006.

[18] . Amazon elastic MapReduce, developer guide, 2015, http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-dg.pdf [last accessed 1 November 2014].

[19] . Buyya R, Vecchiola C, Selvi T. Mastering in Cloud Computing – Foundations and Applications Programming. Morgan Kaufman: USA, 2013.

[20] 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.