UDC 004=111 DOI 10.5862/JCSTCS.212.10

S.V. Mescheryakov, A.O. Rudenko, D.A. Shchemelinin

ASE INTERNATIONAL CONFERENCES ON BIG DATA SCIENCE AND COMPUTING

С.В. Мещеряков, А.О. Руденко, Д.А. Щемелинин

МЕЖДУНАРОДНЫЕ КОНФЕРЕНЦИИ ASE ПО НАУКЕ И КОМПЬЮТЕРНОЙ ОБРАБОТКЕ БОЛЬШИХ ДАННЫХ

This paper is the analytical overview of the series of the International conferences on big data, which are being organized annually by the Academy of Science and Engineering (ASE) in the USA at Stanford, Cambridge, Harvard, and other famous universities [1]. The conference proceedings are published with open Internet access in the ASE scientific digital library [2] while the paper abstracts are indexed in the world leading citation database Scopus [3]. This article briefly describes the most interesting, to authors' opinion, papers and poster presentations having innovative ideas in big data computing area.

BIG DATA; DATABASE; PERFORMANCE; CAPACITY; CLOUD COMPUTING.

Статья представляет собой аналитический обзор серии международных конференций по компьютерной обработке больших данных, которые организуются ежегодно Академией науки и техники США (ASE) в Стэнфорде, Кембридже, Гарварде и других известных университетах [1]. Тезисы конференций опубликованы с открытым Интернет-доступом в научной электронной библиотеке ASE [2], аннотации работ индексированы в ведущей мировой базе данных цитирования Scopus [3]. Кратко описаны наиболее интересные, по мнению авторов, презентации и стендовые доклады, содержащие инновационные идеи в области компьютерной обработки больших данных.

БОЛЬШИЕ ДАННЫЕ; БАЗА ДАННЫХ; ПРОИЗВОДИТЕЛЬНОСТЬ; ЗАГРУЖЕННОСТЬ; ОБЛАЧНЫЕ ВЫЧИСЛЕНИЯ.

What is Big Data?

IT world is a dynamic and data intensive area, moving towards big data and increasing Internet network traffic. Big data is now a growing challenge for many IT companies, especially based on cloud computing services [4–9].

There is still no consensus on how to specify the data volume and define the term "Big Data" – terabytes or petabytes or exabytes or zettabytes [9]. Every IT organization refers to its own data growth to such a big volume and complex infrastructure, which is hard for transferring, storing, processing, analysis and visualization within existing computing architecture. The customer demands are increasing along with development of new IT devices and services, including mobile devices, cloud-based applications, social media and networking, audio and video conferencing, etc., showing accelerated growth over the past 5 years (Fig. 1).

In 2014, the big data problems were focused on ASE conferences 3 times at different places – Stanford University, CA, USA, May 27-31; Tsinghua University, China, August 4-7; and Harvard University, MA, USA, December 13-16. Each conference is a big international forum, having average paper acceptance rate of 8.5 %, bringing together industry companies, academic scientists and other IT specialists from all over the world to share their experience and exchange the advanced results in big data computing.

All the conference sessions and topics of interest in big data are divided into the following independent parts [11–13]:

1. Big Data Science and Engineering;
2. Economic Computing;
3. Social and Biomedical Informatics;

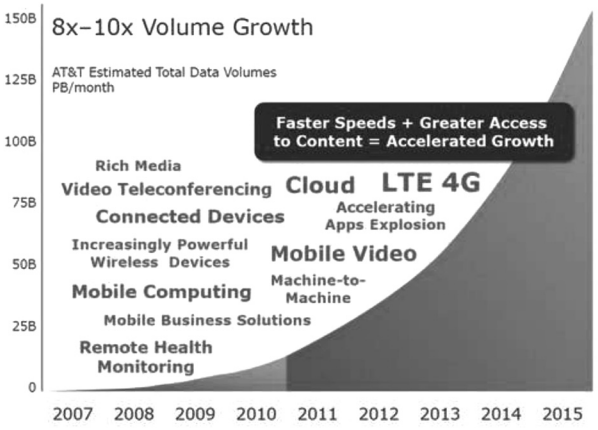

Fig. 1. Big Data Growth over the Past Years, PB/month [10]

4. Cyber Security;

In addition to industry speeches, the poster presentations, workshops, exhibitions and other initiatives are organized by ASE that allow attracting much more participants with or without a thesis published in conference proceedings. The most relevant presentations, innovative ideas and hot topics of big data computing are briefly described below.

Big Data Storing and Performance Testing

Most of enterprise-scale environments are now cloud-based and distributed across datacenters. In a big cloud infrastructure, the performance of data storage is the main bottleneck due to big data flow between multiple remote hosts, especially database (DB) servers. Traditional object-relational approach to data structure developed in early 1980th by Michael Stonebraker [4] is no longer effective for storing big data due to a lot of DB rules, dependencies and constraints. In solving performance, capacity, scalability and other big data challenges, new noSQL solutions, such as Cassandra, MongoDB, HBase, Hadoop, CouchDB, GraphDB, Redis, etc., offer the following benefits over traditional relational DBs:

1. Higher performance of noSQL DB that is important for web applications.

2. Flexible modification of unstructured data on the fly without downtime.

3. Better scalability of noSQL DB due to multi-node architecture.

4. Replication and automatic sharing between primary and secondary nodes.

5. Failure tolerance for a single node degradation, outage or data loss.

6. Most of noSQL DBs are open source and free to install.

Failure tolerance is the key idea of noSQL solutions, meaning that the entire production system will not fail in case a single node is down due to automatic replication and physical sharing between DB nodes, thus increasing overall system performance. On the other hand, the architecture of noSQL DB is application specific and DB tools for benchmarking and analysis are not yet well developed.

The focus of performance analysis is testing the stability of both DB and application against specific workload. Below is the list of products of leading IT companies, which are the most popular in noSQL testing.

1. Yahoo Cloud Serving Benchmark (YCSB) [14]. YCSB is the universal cloud service client for performance benchmarking on the noSQL DBs, supporting Cassandra, Mongo, Redis, etc. Reads, writes and updates can be tested against a sample DB to evaluate the performance of various solutions under specified workload. YCSB can run with an arbitrary number of query threads and multiple hosts in parallel, measuring the throughput in operations per second (IOPS) and the latency of DB operations.

2. SandStorm from Impetus Technologies [15]. SandStorm is the automated tool, providing a load testing and monitoring resources of Cassandra, Mongo and Kafka stack. Real world scenarios can be created to simulate various network conditions and evaluate the performance and scalability of entire applications, including web, mobile, cloud and big data, for the purpose of reducing the cost of cloud environment.

3. Cassandra and Python Stress Tests [16, 17]. The popular py_stress code written in Python and the Java-based utility are used to carry out the stress tests specifically against Cassandra cluster.

4. JMeter Cassandra Plugin by Netflix [18]. JMeter plugin is the fully configurable client of Cassandra provided by Netflix Company to execute different loads on Cassandra clusters.

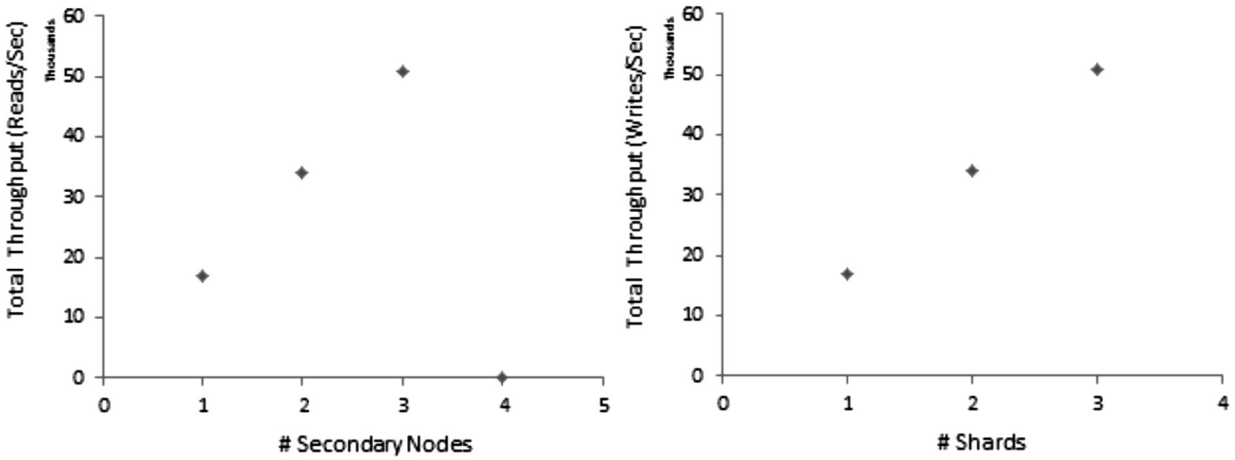

Fig. 2. Scalability for Read and Write Traffic [21]

The performance of specific noSQL DBs like Hadoop, HBase in a distributed infrastructure is evaluated in papers [19, 20]. Obviously, replication of big data causes a latency lag depending on replica set. To reach higher write performance, the DB transaction is committed right after the data is written to a primary node without waiting for secondary replicas. For better read performance, the DB requests are routed to different nodes in parallel. With this read-write solution, the overall system throughput should scale linearly upon adding new secondary nodes as shown in Fig. 2, or otherwise there is a bottleneck somewhere in network traffic.

More valuable conclusions and results of performance testing are presented in [21, 22], when comparing Cassandra with MongoDB (Fig. 3). At the left graph, the correlation with disk usage is observed. The more disk free space the higher throughput. At the right graph, the dependence on the amount of documents returning from a single query is found. DB query rate is acceptable as long as the returning data fits in memory without access to disk, while the performance of heavy queries can degrade dramatically to zero. The solution is to increase DB cache, or install more memory, or re-implement potentially heavy DB queries.

The results of big data performance testing specifically for Zabbix monitoring system are presented in [7, 9]. Fig. 4 shows that Zabbix DB server is much more intensive in write rather than read operations. That is reasonable

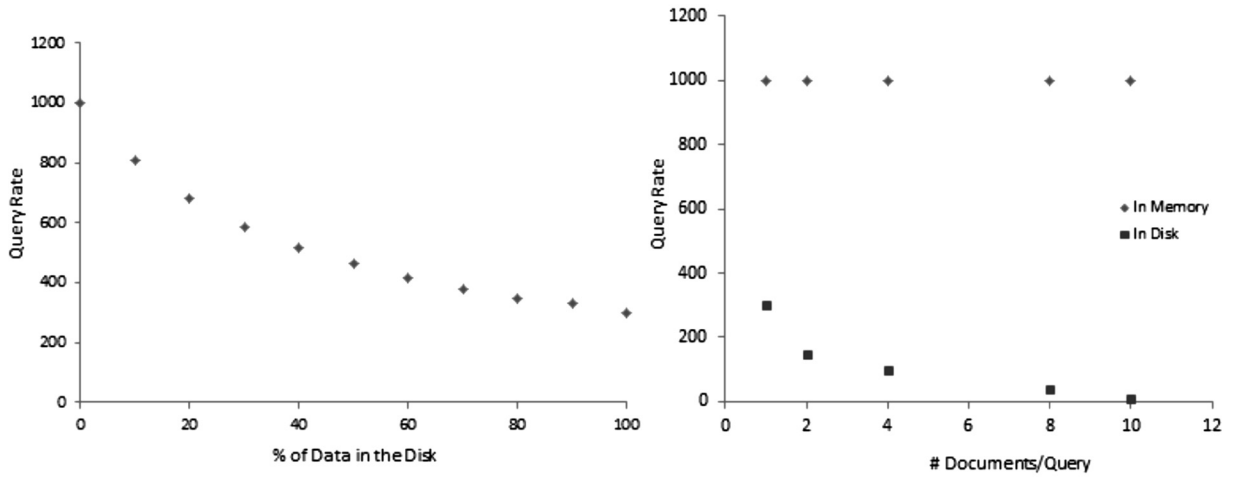

Fig. 3. Performance Testing with MongoDB [21]

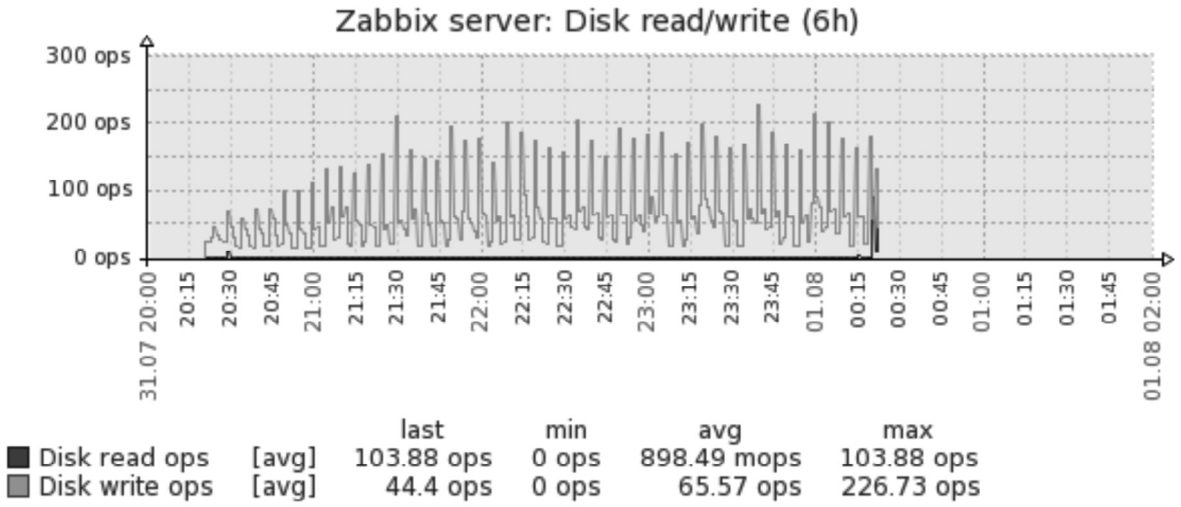

Fig. 4. Read-write Performance of Zabbix Monitoring DB Server [9]

because in the monitoring system the bigger data traffic is sent from multiple hosts to central DB with specified polling time interval, while read from DB is issued for analysis and troubleshooting purposes not so often. Thus, when the monitoring system comes to scalability, Cassandra is more applicable and will scale up better in terms of write performance, while MongoDB is better designed for read intensive traffic.

Big Software Deployments in Big Cloud Environments

This hot topic for big IT companies is addressed by the authors of this paper [23] in their poster presentation to the 2nd ASE International Conference on Big Data Science and Computing. Automatic Deployment System (ADS) is designed and implemented at RingCentral Internet Telecommunication Company for the purpose of automation and improving the time-consuming process of frequently repeatable building over 10K virtual machines (VMs) in big data centers.

ADS architecture is based on Puppet, Foreman, Opscode software and RingCentral in-house applications. Puppet [24] is IT product that helps system administrators proactively manage the entire infrastructure throughout the whole lifecycle (Fig. 5), automating repetitive deployment tasks, scaling from 10s to 1000s of VMs, both on-premise and in

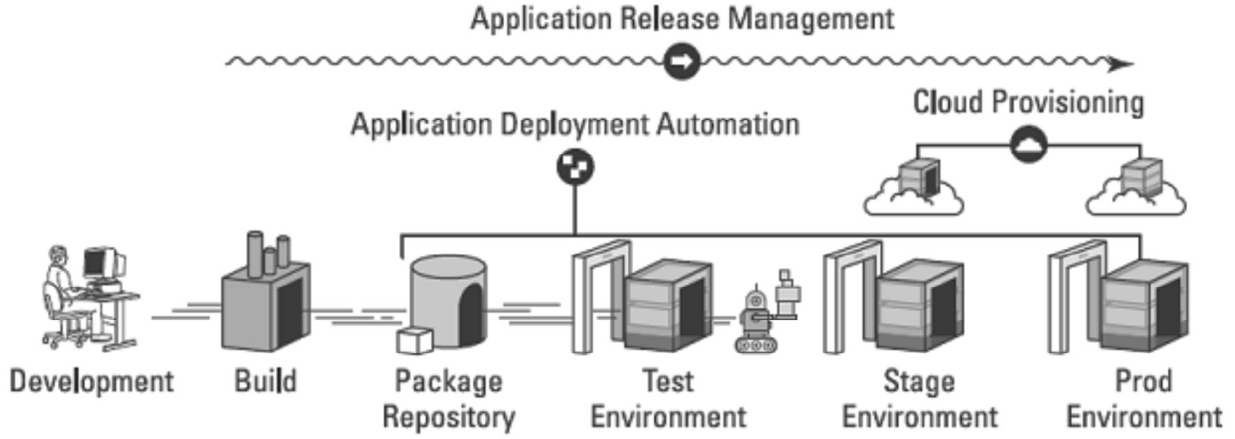

Fig. 5. Typical Deployment Delivery Pipeline [25]

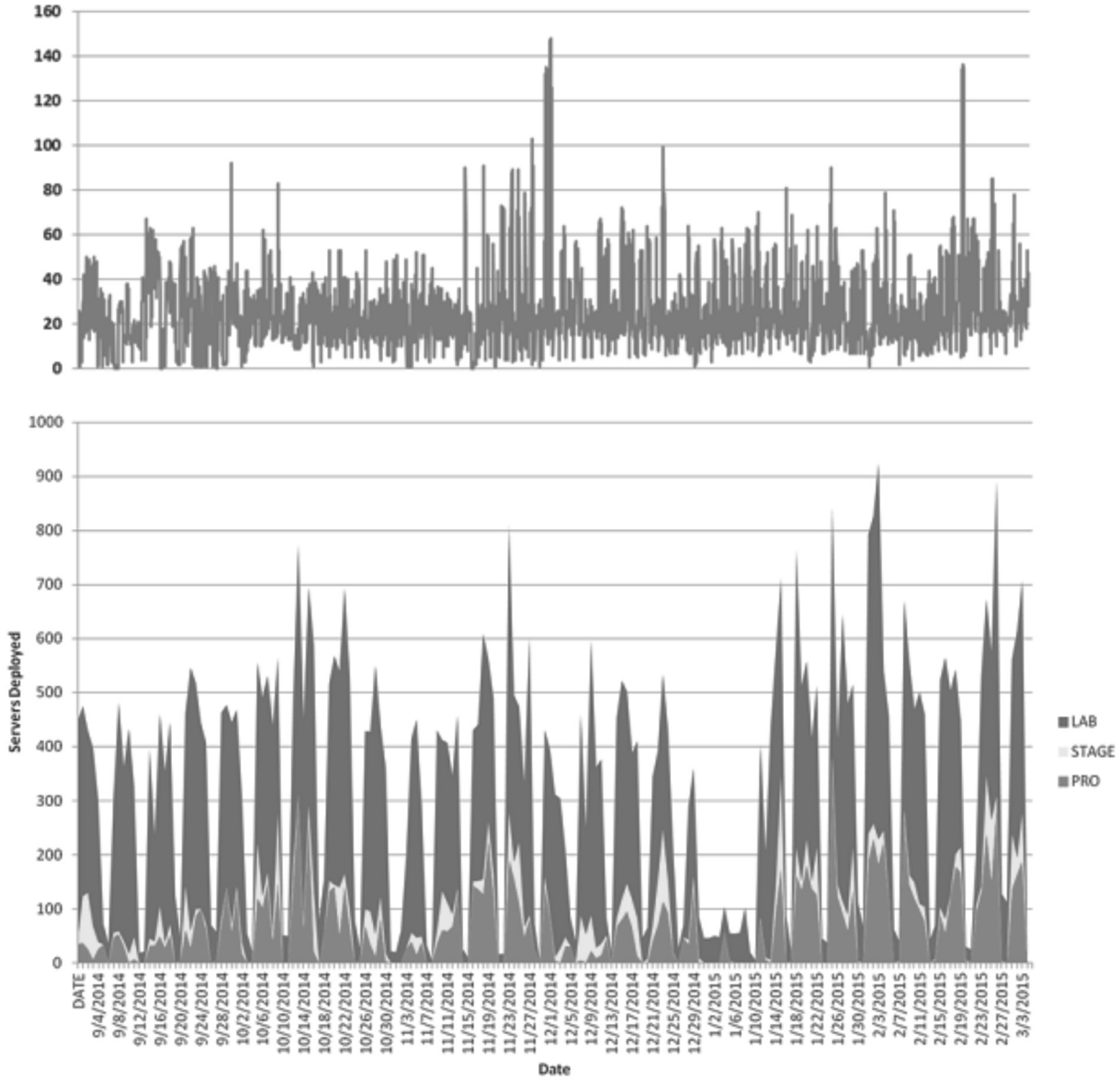

Fig. 6. Amount and Durations of ADS Deployments, min/host [23]

the cloud. Foreman server [26] is the core of ADS architecture, also providing web user application. All of the Puppet manifests, VM images and snapshots are created and stored in the ADS central DB repository for future VM updates and possible rollbacks. Opscode Chef [27] is automation solution that transforms IT infrastructure into the Ruby code, describing in terms of cookbooks and recipes the ingredients and step-by-step instructions for assembling them together into running system. Chef definitions are also chosen by leading IT companies like Facebook and Amazon.

Big data scalability leads to proportional increase of software deployment amount and IT maintenance expenses. With ADS, the duration of a single VM deployment and customization is reduced from hours to minutes depending on the VM image size (Fig. 6). ADS performance now allows building up to 6 big data locations, each having about 150 VMs, within 4-hour IT maintenance period. The benefits of ADS are obvious – provide high availability and continuous delivery of IT services without

outage to business customers, improve the quality of service (QoS) along with excluding manual steps and human errors, save operations efforts and resources in big IT companies.

Big Data Flow and Data Intensive Applications

Information exchange among national economies, companies and people have reached previously unimagined levels and are playing an ever-larger role in modern digital age. Being unconnected to the global Internet network means stay behind the high technologies. According to recent researches and analytical reports of McKinsey Global Institute (MGI) [28], the countries with global network communications and big data flows increase their Gross Domestic Product (GDP) up to 40 % faster than unconnected countries do. Germany, Hong Kong and USA are in the top of MGI Connectedness Index while some developing economies, such as Brazil, India, Saudi Arabia, are climbing up the ranks rapidly due to expanding their products and services into global flows.

Mobile Internet and cloud computing services, which became very popular and therefore data intensive, are in the top 12 list of high technologies that had radically changed IT world. Global online data traffic across borders grew 8 times since 2012, including 40 % increase of Skype international communications. Internet telephony, health care, retail, financial services, online education, transportation and navigation, production monitoring, social media and networking are the leading sectors of big data application.

Big data stream is often caused by high workload, and vice versa. Keynote speech about massive online data flow is provided by Dr. Kalyan Veeramachaneni from Massachusetts Institute of Technology (MIT), Cambridge, MA, USA [11]. Many universities including MIT offer online courses like [4], which are commonly known as Massive Open Online Courses (MOOCs), for millions of students and postgraduates around the IT world. For a single MOOC, the traffic is around 200M click stream events, 10M assignment submissions and 90K forum posts [11]. On the way to solve the problem of high load and scalability, the custom platforms and tools are built at MIT, which allow MOOC instructors teach more effectively, see how the students study the topics, improve student's engagement, and prevent possible web service outage.

Big data flow can be caused by network storm, denial of service (DoS) attack, email bomb, fax or SMS or media broadcast, and other undesirable user activity, which is often called flood. Cyber security questions related to big data are addressed in the keynote lectures from the University of Southern California [10]. To their opinion, current security implications are intensified by big data, especially on cloud-based platforms where a victim is exploited remotely. Cross domain solution (CDS) is proposed as a firewall, protecting network domain against data leakage and intrusion from the Internet.

One more keynote presentation of Dr. Amr Awadallah, CTO and founder of Cloudera Inc., leading company in big data management, refers to Apache Hadoop as noSQL solution, supporting 700+ hardware and software systems, 15K+ trained big data specialists since 2008 [10]. Hadoop uses prescriptive relational DB schema for write and descriptive data modeling for read. New columns are added explicitly to DB table before it is populated with new data. Read data flow may start on the fly and will appear retroactively once DB schema properly describes it. That gives the benefits of high performance, providing together a data storing and real time access for reporting and analysis, auto-scaling with proven growth to 1PB of data per 1K nodes, without requiring developers to redesign data warehouse architecture and algorithms in data intensive applications.

The future of data intensive applications in common and Hadoop framework in particular is outlined by Dr. Milind Bhandarkar, the founding team member at Yahoo!, contributing and working with Hadoop since the earliest version 0.1 [29].

Big Data Analytics and Tools

A promising technique for big data analytics is online learning. Massive Online Analysis (MOA) is the popular open source framework for data stream mining with an actively growing community [30]. It uses different verification tools and machine learning algorithms, such as classification, regression, clustering, outlier and

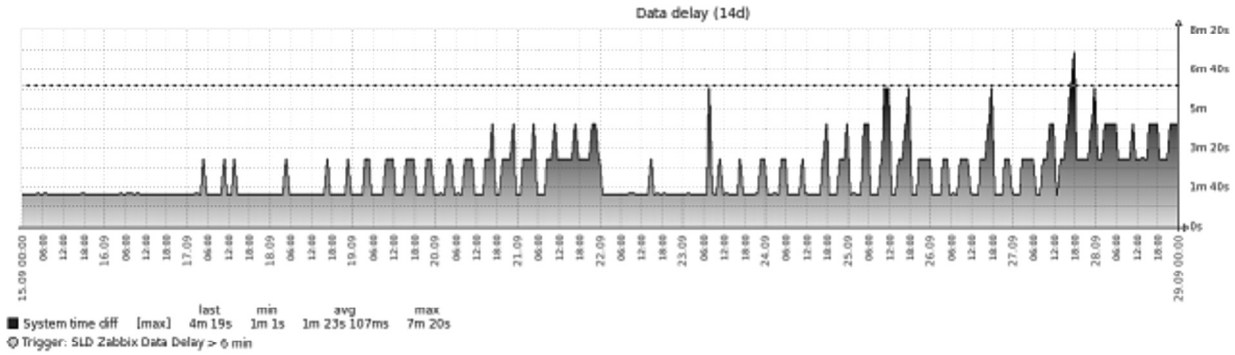

Fig. 7. Zabbix Data Delay Statistics for Two Weeks [9]

concept drift detection, and more demanding issues.

Zabbix [7, 9] can also be used for data analysis, not only as monitoring system, by means of SQL language or specially designed graphs and dashboards in Zabbix application. Zabbix is the open source enterprise-class solution and all the monitoring data is stored in a centralized DB. One more benefit is that all the data is separated into real time (from one week to a month) and the history, which can be stored in separate DBs of different type, either relational or noSQL. Historical data is represented in trends as max, min and average values, allowing significantly reduce DB volume. The key disadvantage of Zabbix architecture is data delay up to 6 min between the local host time and the moment of data availability in Zabbix DB, depending on the proxy servers' amount and performance (Fig. 7).

The disadvantage of proxy data delay is resolved in Sumo Logic [31], where the data is transferred directly to the cloud storage and is available for analysis immediately in real time. Sumo Logic is specially designed for and is efficient as log analyzer having predefined dashboards and its own query language like SQL with regular expressions. In a big cloud distributed environment, system and application logs are also big data, producing terabytes of daily rate and slowing down the performance of production applications when parsing the logs locally. The main disadvantage of Sumo Logic, to opinion of the most IT players [5], is that it's rather expensive solution for small and even big companies like Netflix [18]. Regular payment depends on the volume of data retention on the external storage of the 3rd party vendor. Other limitations are 500 lines per data collector and 15K events per second as maximum. One more weak point of any cloud in opposite to in-house solution is cyber security of collected logs data, which may contain sensitive information, such as user accounts and passwords.

In big data analytics, a crucial part is visualization. With millions of statistical data values, typical graphics become cluttered, hard to read and analyze. Reducing and subsampling the data for better visibility is not good workaround. At North Carolina State University [32], new methods of statistical modeling and scalable interactive representation of a big data network using (un)directed (un)weighted graphs are proposed. Good visualization helps the analysts explore valuable information about data distribution, clustering, trends, scalability, performance, and other big data properties.

Using novel methodology of crowdsourcing (versus well-known outsourcing and insourcing) for big data analytics and management is presented by the laboratory of Stanford University [33]. Traditional entity resolution (ER) algorithm in relational databases is challenging for huge data sets because of many to many checks like "what matches what", consuming time and computing resources. To evaluate efficiently and reduce expenses, the crowd ER strategy allows obtaining information for a particular project by enlisting the services of a lot of people, either paid or unpaid, typically via the Internet (Fig. 8). The idea is to cut the input data set and verify only critical pairwise similarities close to a specified threshold. The key point of crowdsourcing is to use humans

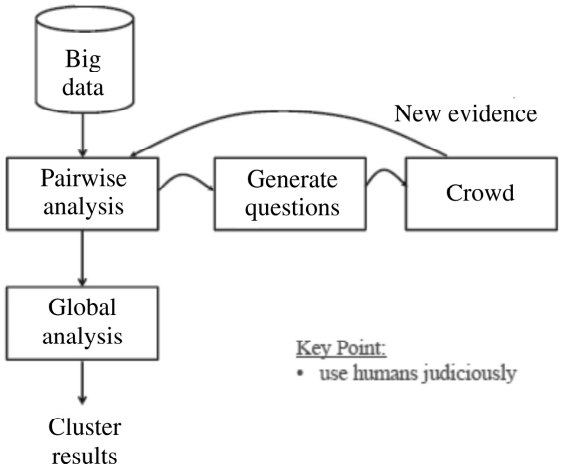

Fig. 8. Using Crowdsourcing for Big Data Analytics

judiciously, generate questions and consider all possible combinations.

In [34], the crowdsourcing technique is investigated in depth, further enhanced using the power of social networks and is applied to create a self-sustainable crowd-engagement chain for big data harvesting and better decision making.

The Future of Big Data

There are five key dimensions of big data, so-called 5V's:

1. Volume of big data stored in the IT world tends to grow rapidly, according to IDC analysis and forecast [10], from 8 million PB in 2015 to 40 ZB in 2020 that is over 5 TB per person.

2. Velocity of big data streaming in real time will also accelerate upon network bandwidth extension and people's demand. For example, sharing new data at Facebook is over 2.5 billion posts per day that is 500 TB daily rate.

3. Variety of big data implementations, including text, image, audio and video processing, moves from traditional structured databases towards semi-structured and unstructured organization that is of higher performance, better scalable and flexible for future updates.

4. Veracity, consistency and integrity of big data will strongly depend on failure tolerance idea based on multi-node architecture, distributed parallel computing, replication and sharing between nodes.

5. Value of global data exchange is hard to overestimate and, in recent MGI analytical reports [28], is predicted as a benefit of up to 40 % growth to national GDP, especially for developing economies.

According to Cisco and other research reports [35], big data and cloud computing technologies are now being extended to the fog computing paradigm, meaning that IT services are hosted at the network edge close to end user location, spanning multiple domains and distributed over heterogeneous platforms. Even the geographical position of mobile smartphones and other wireless end devices are taken into account in real time analytics. The goal is to reduce network traffic, facilitate service mobility across platforms, speed up data delivery to a client and improve QoS. In the Internet of Everything (IoE) applications, various computers, vehicles, mobile and other devices are interconnected around the Internet backbone. Fog networking supports densely distributed data collections, hence adding one more big data dimension to the 5V list.

A long-term strategy of big data programs and projects in various engineering areas is to escalate big data concerns to in-depth core scientific research, promote all-around IT resources available today and in the future. A good experience is comprehensive integration with different fields of inquiry, collaboration of multi-disciplinary teams and communities. That gives more opportunities and accelerates the progress of scientific discovery and development.

ASE International conferences continue to address a wide range of big data problems. Such conferences are connecting the scientists and industry practitioners from leading IT organizations, who are always focused on new

solutions to predict, analyze, improve and resolve big data challenges in practice.

The next International Conference on Big Data is scheduled to September 17, 2015, Xian, China [36]. We are looking forward for innovative ideas and good examples of successful implementation in big data and high performance computing.

REFERENCES / СПИСОК ЛИТЕРАТУРЫ

1. *ASE International Conferences on Big Data Science and Computing*. Available: http://www.scienceengineeringacademy.org/asesite/conferences/

2. ASE Open Access Scientific Digital Library. *Big Data 2014 Conference Proceedings*. Available: http://www.ase360.org/handle/123456789/24

3. *Elsevier's Scopus Citation Index Database. Free search for author on official web site*. Available: http://www.scopus.com/search/form/authorFreeLookup.url

4. **Kucherova K.N., Mescheryakov S.V.** Tackling the Challenges of Big Data On-Line, *Modern Infocommunication and Remote Technologies in the Educational Space of School and Higher Education Institution: Materials of the 2nd International Scientific Conference*. Prague, Czech Republic, Vedecko Vydavatelske Centrum "Sociosfera-CZ", 2015. Available: http://sociosphera.com/conference/2015/sovremennye_infokommunikacionnye_i_distancionnye_tehnologii_v_obrazovatelnom_prostranstve/

5. **Mescheryakov S.V., Shchemelinin D.A.** International Conference for the Performance Evaluation and Capacity Analysis by CMG, *St. Petersburg State Polytechnical University Journal. Computer Science. Telecommunications and Control System*. St. Petersburg: SPbGPU Publ., 2014, No. 1(188), Pp. 99–104.

6. **Mescheryakov S., Shchemelinin D., Efimov V.** Adaptive Control of Cloud Computing Resources in the Internet Telecommunication Multiservice System, *The 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops*. St. Petersburg, Russia, 2014. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7002117

7. **Mescheryakov S.V., Shchemelinin D.A.** Analytical Overview of Zabbix International Conference 2013, *St. Petersburg State Polytechnical University Journal. Computer Science. Telecommunications and Control System,* St. Petersburg: SPbGPU Publ., 2014, No. 1(188), Pp. 91–98.

8. **Mescheryakov S., Shchemelinin D.** Capacity Management of Java-based Business Applications Running on Virtualized Environment, *Proc. of the 39th International Conference for the Performance and Capacity by CMG*. La Jolla, USA, 2013. Available: http://www.cmg.org/conference/

9. **Mescheryakov S., Shchemelinin D.** Proactive and Reactive Monitoring, *The 4th International Zabbix Conference on Scaling and High Performance Monitoring*. Riga, Latvia, 2014. Available: http://www.zabbix.com/conf2014_agenda.php

10. *IEEE Xplore Digital Library. Conference Abstracts and Proceedings*. Available: http://ieeexplore.ieee.org/search/searchresult.jsp?punumber=6784145

11. *The 2nd ASE International Conference on Big Data Science, Social Computing and Cyber Security*. Stanford University, Stanford, CA, USA, 2014. Available: http://www.scienceengineering.org/ase/conference/2014/bigdata/sanjose/website/

12. *The 3rd ASE International Conference on Big Data Science, Social and Economic Computing*. Tsinghua University, Beijing, China, 2014. Available: http://www.scienceengineering.org/ase/conference/2014/bigdata/beijing/website/

13. *The 4th ASE International Conference on Big Data, Social and Biomedical Computing*. Harvard University, Cambridge, MA, USA, 2014. Available: http://www.scienceengineering.org/ase/conference/2014/bigdata/boston/website/

14. **Cooper B.F., Silberstein A., Tam E., Ramakrishnan R., Sears R.** Benchmarking Cloud Serving Systems with YCSB. *ACM Symposium on Cloud Computing*, Indianapolis, IN, USA, 2010. Available: http://www.brianfrankcooper.net/pubs/ycsb.pdf

15. *Impetus Technologies official site*. Available: http://sandstorm.impetus.com/

16. *Apache Cassandra Documentation*. Available: http://www.datastax.com/docs/1.1/references/stress_java

17. *Cassandra Configuration and Tuning*. Available: https://docs.jboss.org/author/display/RHQ/Cassandra+Configuration+and+Tuning

18. *Apache JMeter Cassandra Plugin by Netflix Company*. Available: https://github.com/Netflix/CassJMeter/wiki

19. **Wlodarczyk T.W., Han Y., Rong C.** Performance Analysis of Hadoop for Query Processing, *IEEE Workshops of International Conference on Advanced Information Networking and Applications*. 2011, Pp. 507–513.

20. **Vora M.N.** Hadoop-HBase for Large-scale Data, *International Conference on Computer Science and Network Technology*. 2011, Pp. 601–605.

21. **Gurijala A.** Methodical Benchmarking of NoSQL Database Systems, *Proc. of the 39th International Conference for the Performance and Capacity by CMG*. La Jolla, USA, 2013. Available: http://www.cmg.org/conference/

22. **Batterywala M.** Performance Testing of NoSQL Applications, *Proc. of the 39th International Conference for the Performance and Capacity by CMG*. La Jolla, USA, 2013. Available: http://www.

cmg.org/conference/

23. **Mescheryakov S., Shchemelinin D.** Big Software Deployments in a Big Enterprise Environment, *Proc. of the 2nd ASE International Conference on Big Data Science and Computing*. Stanford, CA, USA, 2014. Available: http://www.ase360.org/handle/123456789/135/

24. *What is Puppet?* Puppet Labs, 2013. Available: http://puppetlabs.com/puppet/what-is-puppet/

25. **Sanjeev S.** *DevOps for Dummies*, The 2nd IBM Limited Edition. John Wiley & Sons Inc., Hoboken, NJ, USA, 2014. Available: http://ibm.co/1dSqfyh/, https://sdarchitect.wordpress.com/2013/10/09/devops-for-dummies-book-is-available-for-download/

26. *Foreman 1.3 Manual*. Available: http://www.theforeman.org/manuals/1.3/index.html

27. *Opscode Chef*. Available: http://www.opscode.com/chef/

28. *Global Flows in a Digital Age*. McKinsey Global Institute, 2014. Available: http://www.mckinsey.com/insights/globalization/global_flows_in_a_digital_age/

29. **Bhandarkar M.** Future of Data Intensive Applications, *Proc. of the 2nd ASE International Conference on Big Data Science and Computing*. Stanford, CA, USA, 2014. Available: http://www.ase360.org/handle/123456789/24

30. *Massive Online Analysis*. Available: http://moa.cms.waikato.ac.nz/

31. *Sumo Logic Analytic Services*. Available: http://www.sumologic.com/

32. **Selim H., Chopade P., Zhan J.** Statistical Modeling and Scalable, Interactive Visualization of Large Scale Big Data Networks, *Proc. of the 2nd ASE International Conference on Big Data Science and Computing*. Stanford, CA, USA, 2014. Available: http://www.ase360.org/handle/123456789/139/

33. **Whang S.E., Lofgren P., Garcia-Molina H.** Using Crowdsourcing for Data Analytics (Stanford University), *Proc. of the 39th IEEE International Conference on Very Large Data Bases*. Trento, Italy, 2013. Available: http://ieeexplore.ieee.org/search/searchresult.jsp?punumber=6784145

34. **Xydopoulos G., Basnayake H., Louvieris P., Stergioulas L.** A Novel Crowd-sourcing Technique for Big Data Harvesting on Social Media, *Proc. of the 2nd ASE International Conference on Big Data Science and Computing*. Stanford, CA, USA, 2014. Available: http://www.ase360.org/handle/123456789/153/

35. *Fog Computing, Ecosystem, Architecture and Applications*, Cisco Research. Available: http://www.cisco.com/web/about/ac50/ac207/crc_new/university/RFP/rfp13078.html

36. *The 7th International Big Data Conference*. Xian, China, 2015. Available: http://www.cyberc.org/

MESCHERYAKOV, Sergey V. *St. Petersburg Polytechnic University*.
195251, Politekhnicheskaya Str. 29, St. Petersburg, Russia.
E-mail: serg-phd@mail.ru

МЕЩЕРЯКОВ Сергей Владимирович – *профессор кафедры инженерной графики и дизайна Санкт-Петербургского государственного политехнического университета, доктор технических наук.*
195251, Россия, Санкт-Петербург, ул. Политехническая, д. 29.
E-mail: serg-phd@mail.ru

RUDENKO, Alexander O. *St. Petersburg Polytechnic University*.
195251, Politekhnicheskaya Str. 29, St. Petersburg, Russia.
E-mail: rudenko.ao@gmail.com

РУДЕНКО Александр Олегович – *аспирант кафедры инженерной графики и дизайна Санкт-Петербургского государственного политехнического университета.*
195251, Россия, Санкт-Петербург, ул. Политехническая, д. 29.
E-mail: rudenko.ao@gmail.com

SHCHEMELININ, Dmitry A. *RingCentral Inc.*.
1400 Fashion Island Blvd., San Mateo, CA, USA 94404.
E-mail: dshchmel@gmail.com

ЩЕМЕЛИНИН Дмитрий Александрович – *руководитель департамента эксплуатации и развития RingCentral, кандидат технических наук.*
1400 Fashion Island Blvd., San Mateo, CA, USA 94404.
E-mail: dshchmel@gmail.com