Cloud Search Based Applications for Big Data - Challenges and Methodologies for Acceleration

George Suciu[1,3], Ana Maria Sticlan[3], Cristina Butca[2,3], Alexandru Vulpe[1(✉)], Alexandru Stancu[1], and Simona Halunga[1]

[1] Faculty of Electronics, Telecommunications and IT, Telecommunication Department, University Politehnica of Bucharest, Sector 6, 060071 Bucharest, Romania
george@beia.ro, {alex.vulpe, alex.stancu}@radio.pub.ro, shalunga@elcom.pub.ro

[2] Faculty of Automatic Control and Computers, Computer Science and Engineering Department, University Politehnica of Bucharest, Sector 6, 060071 Bucharest, Romania
{anamaria.jalba, cristina.butca}@beia.ro

[3] R&D Department, Beia Consult International, Sector 4, 041386 Bucharest, Romania

Abstract. Innovation in Search Based Applications (SBAs) requires more than just creation of technology and use of Cloud Computing or Big Data (BD) platforms. Furthermore, the problem of acceleration in the aggregation and analysis of heterogeneous cloud-based data needs to be addressed. This paper fills a gap in the Cloud Computing literature by providing a general overview of the challenges and methodologies for acceleration of search applications for BD. The main contribution of this paper consists in analyzing cloud techniques that can be used for faster search of large volumes of data. Finally, the components and interfaces of the proposed SBA based on EXALEAD CloudView are presented and discussed.

Keywords: Acceleration · Big Data · Cloud Computing · EXALEAD CloudView · Search Based Applications

1 Introduction

The innovation flow is definitely becoming faster and faster, not only for technology but also for services and business models. Speed is becoming the key to market access. Globalisation of the economy is the rule in all the intensive software systems markets [1]. In today's hypercompetitive business environment, companies not only have to find and analyze the relevant data they need, they must do it quickly. Visualization helps organizations perform analyses and make decisions much more rapidly, but the challenge consists of going through the sheer volumes of data and accessing the level

F. Pop and M. Potop-Butucaru (Eds.): ARMS-CC 2015, LNCS 9438, pp. 177–185, 2015.
DOI: 10.1007/978-3-319-28448-4_13

of detail needed, all at a high speed [2]. In this paper we describe the challenges and methodologies for acceleration of a search based application (SBA) for Big Data. Furthermore, we provide a general survey on search based solutions for Big Data, which will be presented below.

BD is typically considered to be a data collection that has grown so large it can't be effectively or affordably managed (or exploited) using conventional data management tools: for example, classic relational database management systems (RDBMS) or conventional search engines, depending on the task at hand. Cloudera Enterprise [3] is designed specifically for mission-critical environments and includes CDH [4], one of the most popular open source Hadoop-based platform, as well as advanced system management and data management tools.

Cloudera Enterprise, with Apache Hadoop at the core, is unified into one integrated system, bringing different users and application workloads to one pool of data on a common infrastructure. Also, no data movement is required, offering perimeter security, authentication, granular authorization and data protection, enterprise-grade data auditing, data lineage, and data discovery. Moreover, the solution provides managed native high-availability, fault-tolerance and self-healing storage, automated backup and disaster recovery, and advanced system and data management, and ensures that data and applications remain private, offering an open platform to connect with all existing investments in technology and skills. Moreover, the cloud computing SBAs can be used for accelerating business decisions using BD, for example taking real-time meaningful information from sensors about environmental conditions which can be utilized by farmers for precision agriculture or environmental telemetry [5, 6].

MarkLogic [7] is a solution whose search and query capability makes it easier to find better answers in BD. As an Enterprise NoSQL database, MarkLogic gives organizations the ability to accelerate virtually any query over today's BD, thanks to sophisticated, best-in-class indexes. These same indexes also power full-text search, and MarkLogic is consistently chosen to power enterprise search applications over other offerings from the world's largest search engine companies. MarkLogic has enterprise search built-in, enabling organizations to turn BD into useful results, without the need to shred the data. MarkLogic indexes data on load and makes it immediately searchable.

Fusion [8] is a solution that is aiming to make BD searches as simple as googling. This solution implements and extends the open source Apache Solr [9] search framework. Fusion also provides analytics features, BD discovery and importation tools that can connect to a variety of database types, machine learning and natural language search.

The rest of the paper is organized as follows: Sect. 2 presents related work, Sect. 3 analyses challenges identified for the acceleration process, Sect. 4 details the methodologies and techniques for Big Data SBAs, followed by Sect. 5 that concludes the paper and envisions future work.

2 Related Work

The acceleration of search based applications for BD faces multiple challenges, which emerge from several industrial example scenarios, for instance by using business accelerators or virtual accelerators. This section analyses the main approaches and reviews the state of the art.

2.1 Business Accelerators

A business accelerators' main goal is to produce successful firms that will leave the program financially viable and freestanding in the go-to-market process. Critical to the definition of an accelerator is the provision of management guidance, technical assistance and consulting tailored to young growing companies. National Business Incubation Association [10] estimates that there are about 7,000 business accelerators or incubators worldwide, out of which about 1400 are found in North America and approx. 500 in China. In 2012, there were around 1200 incubators in Europe, generating 30.000 gross new jobs/year and the ICT sector incubation is among the most common industry specializations of incubation service providers in Europe, as presented in Fig. 1.

Business activities in which European business incubators specialise in

Business Activities	Number	Percentage
(1) Sales, marketing and distribution	5	0.4
(2) Business and financial services	8	0.6
(3) Advanced/ high-tech manufacturing	263	18.6
(4) Information & communication technologies	258	18.2
(5) Research & development	173	12.2
(6) Biotechnology/ pharmaceuticals	201	14.2
(7) Knowledge-based industries/ new economy companies	162	11.5
(8) Other manufacturing activities	86	6.1
(9) Other service activities	124	8.8
(10) A combination of some/ all of these activities	134	9.5
Total *(multiple responses possible)*	**1,414**	**100.0**

(Source: CSES analysis of DG Enterprise, Incubator database)

Fig. 1. Business activities in which European business incubators specialise

Accelerators vary widely in their size and service offering and may provide clients, for example, access to appropriate rental space and flexible leases, shared basic business services and equipment, technology support services and assistance in obtaining the financing necessary for company growth. Incubators will play the role of a partner in this project as many innovations are tested by start-ups inside this kind of structure.

2.2 Virtual Accelerators

Virtual acceleration is defined as the delivery of services solely through electronic means, eliminating requirement for geographical proximity of the clients. Examples of virtual accelerators operating in Europe can be found in the New Hampshire Virtual Business Incubator [11] or the FIWARE platform [12].

Other virtual accelerators can be built on communities, networks and events, where mostly organisations have as goal to accelerate the go-to-market using BD [13]. Furthermore, infrastructure and facility providers offer facilities to start-ups in the domain, that receive also support from public acceleration initiatives such as government stimulations.

Further acceleration for BD search based applications can be obtained by using technical consultancy services and internal venturing programs of large companies.

Several large ICT companies have installed proprietary venturing programs. They are often custom made for the company and therefore public access to details about how these programs deal with Validated Learning is not always available.

Except for the last category, these initiatives are mainly focused on start-ups in a very early stage. They are not adapted to offering support to a large industry. Often these initiatives are venture driven and require equity; and are now emerging in Europe.

3 Challenges for Acceleration

Over the past decades, ICT companies have faced many kinds of challenges with respect to introducing their products or services in the market. We identify in the following three major development waves, each bringing a very different breed of problems to ICT companies:

- **Solving the engineering headache:** Delivering software products on time, within budget and with an acceptable quality has always been a challenge for software companies. In this historical context, software companies have specialized in what can be called "The art of software engineering".
- **Solving the innovation headache:** As software and ICT became more and more an instrument for innovation, ICT companies were faced with a new generation of challenges that can be classified as "The art of software innovation".
- **Accelerating and monetizing innovations:** ICT companies have realized that innovation is much more than technology itself and they struggle to translate the scientific advances they make into marketable innovations. It is not sufficient to install the appropriate innovation processes or to generate the ideas, but it also requires the infrastructure and instruments to commercialize these inventions and innovations.

The current state of the art can be classified into three categories:

- **Tool support for the progress follow up of validated learning:** An example of this is Lean Launch Lab, an approach developed by Steve Blank and applied both in his class at the Stanford University, but also with spin-offs from the National Science Foundation [14].

- **Technological support/methodologies for defining and validating features of Minimal Viable Products:** In addition to face-to-face interviews, some companies use prototypes to test specific assumptions, for example in an early stage, creating a dynamic mockup of a product feature, putting it online and observing user response through data analysis.
- **Technological support for observing users and collecting metrics:** Validated Learning can be achieved by studying the usage data collected from customers that use the applications. Solutions for easy implementation of data collection exist; some even free of charge (for example Google Analytics [15] is a "freemium" example, Piwik [16] an open source one).

4 Methodologies and Techniques for Acceleration

Multimedia content is the fastest growing type of user-generated content, with millions of photos, audio files and videos uploaded to the Web and enterprise servers daily. Recently, technologies like automatic speech-to-text transcription and object recognition processing (called Content-Based Image Retrieval, or CBIR [17]) are enabling us to structure this content from the inside out, and paving the way toward new accessibility for large-volume multimedia collections. We expect for this trend to have a significant impact in fields like medicine, media, publishing, environmental science, forensics and digital asset management.

A search system is therefore a Data Management System like its NoSQL and NewSQL counterparts, and it achieves massive scalability in much the same way, i.e. through distributed architectures, parallel processing, column-oriented data models, etc. However, it is the semantic capabilities and high usability of search-based DMS that make them ideal complements to (and in some cases, alternatives to) NoSQL and NewSQL systems [18].

Search platforms are responsive because they are optimized for fast query processing against large volumes of data (read operations), and because most of the calculations they use to produce dashboard analytics and ad hoc drilling are automatically executed, as part of routine indexing processes, the results are there waiting to be exploited with no processing overhead. For example, Exalaed CloudView [19] extends analytic possibilities with high-performance query-time computations.

What's more, all of these out-of-the-box search, access and analysis capabilities can be rapidly packaged into secure, task-oriented business applications to help you extract real bottom-line value out of BD investments in a matter of days or weeks. For all these reasons, search platforms serve as perfect complements to NoSQL and NewSQL systems, and, in some contexts, provide a pragmatic alternative to them.

We propose a SBA based on EXALEAD CloudView that combines web-scale semantic technologies, rapid drag-and-drop application development and hybrid quantitative/qualitative analytics to deliver a consumer-style information experience to mission-critical business processes. Furthermore, acceleration techniques can be applied to meet demands for real-time, in-context, accurately-delivered information,

accessible from diverse web and enterprise BD sources, yet delivered faster and with less cost than with traditional application architectures [19].

CloudView offers an extensive connector suite, a drag-and-drop development framework and a library of 100+ application widgets to support mobile and fixed applications that incorporate a broad range of functionality, including search and faceted navigation, quantitative and qualitative analytics, rich content mashups, and sentiment analysis. The SBA platform is composed of four core components administered via a Management and Monitoring console, as presented in Fig. 2.

Fig. 2. Components of the Big Data search application

In a BD world, meaningful context begins with the right connections and fast contextual information delivery. SBAs helps organization cut costs and increase revenue by solving one of today is most vexing information systems challenges: achieving a unified view of information across data silos to support global search and discovery, enabling innovative new business applications.

This includes an intelligent extraction of complex structured data and associated rich metadata (attributes, rules, relationships, etc.) from the world's most sophisticated enterprise applications and data warehouse systems.

For some, BD simply means Big Headaches [20], raising difficult issues of information system cost, scaling and performance, as well as data security, privacy and ownership. Furthermore, advantages and challenges of public cloud versus private cloud need to be considered [21]. Also, it carries the potential for breakthrough insights and innovation in business, science, medicine and government, machines and data together to reveal the natural information intelligence locked inside mountains of BD, as depicted in Fig. 3.

The classic data management mission represents transforming raw data into action-guiding wisdom [22]. In the era of BD, the challenge is to find fast, automated, industrial-grade methods for accomplishing this transformation.

We analyzed CloudViews' usability, agility and performance for search and search-based applications (SBAs) for BD environments using the following criteria:

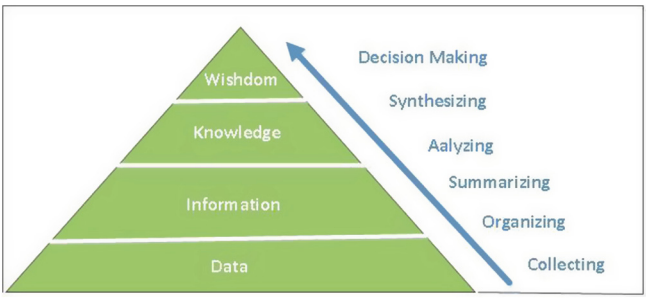

Fig. 3. Mountain methodology for Big Data search application

- **Performance:** is a search platform which offers wide access to information on infrastructure level and is used for both SBA (Search Business Application) for online and enterprise level. This solution can provide secure sub-second query processing against billions of records for thousands of simultaneous users.
- **Connectivity:** CloudView uses most advanced Web crawler to gather multi-format data from virtually any source, including an intelligent extraction of complex structured data and associated rich metadata (attributes, rules, relationships, etc.) from the enterprise applications and data warehouse systems, and processing of the "grey" data that constitutes the bulk of Big Data collection.
- **Analytics:** The platform supports query-time computation of complex numerical, geophysical and virtual aggregates and clusters, and supports dynamic 2D faceting for creating advanced pivot-style tables.
- **Business application development:** The SBA platform is unique, because it provides a drag-and-drop development framework, the Mash-up Builder, for rapidly constructing high value business applications on top of Big Data sources, including applications optimized for mobile delivery.

Consequently, CloudView can be used for the acceleration of large volume SBAs, using heterogeneous big data sources for processes that reveal accurate information assets in time critical applications.

Big data technology must support search, development, governance and analytics services for all data types—from transaction and application data to machine and sensor data, to social, image and geospatial data, and more.

For operational Big Data workloads, NoSQL Big Data systems such as document databases have emerged to address a broad set of applications, and other architectures, such as key-value stores, column family stores, and graph databases are optimized for more specific applications. NoSQL technologies, which were developed to address the shortcomings of relational databases in the modern computing environment, are faster and scale much more quickly and inexpensively than relational databases.

Critically, NoSQL Big Data systems are designed to take advantage of new cloud computing architectures that have emerged over the past decade to allow massive computations to be run inexpensively and efficiently. This makes operational Big Data workloads much easier to manage, cheaper and faster to implement.

Analytical Big Data workloads, on the other hand, tend to be addressed by MPP database systems and MapReduce. These technologies are also a reaction to the limitations of traditional relational databases and their lack of ability to scale beyond the resources of a single server. Furthermore, MapReduce provides a new method of analyzing data that is complementary to the capabilities provided by SQL.

5 Conclusions

This paper analysed the challenges and methodologies for acceleration of Search Based Applications for Big Data and proposed a practical implementation by using EXALEAD CloudView platform.

Furthermore, the paper presented a general overview on the techniques and interfaces of SBAs for Big Data and described how EXALEAD CloudView can be applied for the acceleration of SBAs using heterogeneous BD sources, thus revealing accurate information assets in time critical applications.

As future work we envision to develop the proposed solution for analyzing environmental conditions from sensor data which can be utilized in agriculture for precision farming.

Acknowledgments. The work has been funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreement POSDRU/159/1.5/S/134398 and supported in part by UEFISCDI Romania under grants no. 20/2012 "Scalable Radio Transceiver for Instrumental Wireless Sensor Networks - SaRaT-IWSN", TELE-GREEN, NMSDMON, CarbaDetect, MobiWay, EV-BAT and CommCenter projects, grant no. 262EU/2013 "eWALL" support project, grant no. 337E/2014 "Accelerate" project, by European Commission by FP7 IP project no. 610658/2013 "eWALL for Active Long Living - eWALL" and European Union's Horizon 2020 research and innovation program under grant agreement No. 643963 (SWITCH project).

References

1. Zhao, Z.Q., Zou, X.R., Li, C.P.: Design of ERP management information system for SME. Appl. Mech. Mater. **608**, 440–444 (2014)
2. Jagadish, H.V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan, R., Shahabi, C.: Big data and its technical challenges. Commun. ACM **57**(7), 86–94 (2014)
3. Dhar, S., Mazumdar, S.: Challenges and best practices for enterprise adoption of big data technologies. In: 2014 IEEE International Technology Management Conference (ITMC), pp. 1–4 (2014)
4. Kashyap, K., Deka, C., Rakshit, S.: A review on big data, hadoop and its impact on business. Int. J. Innovative Res. Dev. **3**(12), 1–4 (2014)
5. Waga, D., Rabah, K.: Environmental conditions' big data management and cloud computing analytics for sustainable agriculture. World J. Comput. Appl. Technol. **2**, 73–81 (2014)
6. Ochian, A., Suciu, G., Fratu, O., Suciu, V.: Big data search for environmental telemetry. In: 2014 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), pp. 182–184 (2014)

7. Hunter, J., Grimm, R.: A JSON facade on MarkLogic server. XML Prague, pp. 25–29 (2011)
8. Feinleib, D.: The big data landscape. In: Big Data Bootcamp, pp. 15–34. Apress (2014)
9. Smiley, D., Pugh, D.E.: Apache Solr 3 Enterprise Search Server. Packt Publishing Ltd., Birmingham (2011)
10. Isabelle, D.A.: Key factors affecting a technology entrepreneur's choice of incubator or accelerator. Technol. Innov. Manag. Rev. **3**(2), 16–22 (2013)
11. Kuesten, C.: Knowledge matters: technology, innovation, and entrepreneurship in innovation networks and knowledge. J. Prod. Innov. Manag. **29**(2), 332–334 (2012)
12. Villaseñor, E., Estrada, H.: Informetric mapping of big data in FI-WARE. In: Proceedings of the 15th Annual International Conference on Digital Government Research, pp. 348–349. ACM (2014)
13. Sand, G., Tsitouras, L., Dimitrakopoulos, G., Chatzigiannakis, V.: A big data aggregation, analysis and exploitation integrated platform for increasing social management intelligence. In: 2014 IEEE International Conference on Big Data, pp. 40–47. IEEE (2014)
14. Neumeyer, X.: Examining the role of inquiry-based learning in entrepreneurship education. In: NCIAA Conference, Washington, DC (2013)
15. Plaza, B.: Google analytics for measuring website performance. Tourism Manag. **32**, 477–481 (2011)
16. Miller, S.A.: Piwik Web Analytics Essentials. Packt Publishing Ltd., Birmingham (2012)
17. Hole, A.W., Prabhakar, L.R.: Design and implementation of content based image retrieval using data mining and image processing techniques. Database **3**(3), 219–224 (2015)
18. Grolinger, K.: Data management in cloud environments: NoSQL and NewSQL data stores. J. Cloud Comput. Adv. Syst. Appl. **2**(22), 1–24 (2013)
19. Eckstein, R.: Interactive Search Processes in Complex Work Situations: A Retrieval Framework, vol. 10, pp. 62–67. University of Bamberg Press, Bamberg (2011)
20. Chen, H., Chiang, R., Storei, V.C.: Business intelligence and analytics: from big data to big impact. MIS Q. **36**(4), 1165–1188 (2012)
21. Suciu, G., Ularu, E.G., Craciunescu, R.: Public versus private cloud adoption—a case study based on open source cloud platforms. In: 20th IEEE Telecommunications Forum (TELFOR), pp. 494-497 (2012)
22. Minelli, M., Chambers, M., Dhiraj, A.: Big data technology, Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses. Wiley, Hoboken (2013)