TECHNICAL REPORT

A Roadmap for Big-Data Research and Education

Olov Schelén
Ahmed Elragal
Moutaz Haddara

LULEÅ UNIVERSITY OF TECHNOLOGY

LULEÅ
UNIVERSITY
OF TECHNOLOGY

A Roadmap for Big-Data Research and Education

Olov Schelén, Ahmed Elragal, Moutaz Haddara

Luleå University of Technology
Department of Computer Science, Electrical and Space Engineering
Division of Computer Science

ISSN 1402-1536
ISBN 978-91-7583-275-3 (pdf)

Luleå 2015

www.ltu.se

A roadmap for big-data research and education

Olov Schelén, Ahmed Elragal, Moutaz Haddara
Luleå University of Technology
{olov.schelen, ahmed.elragal, moutaz.haddara}@ltu.se

March 24, 2015

Abstract

The research area known as *big data* is characterized by the 3 V's, which are *volume*; *variety*; and *velocity*. Recently, also *veracity* and *value* have been associated with big data and that adds up to the 5 V's. Big data related information systems (IS) are typically highly distributed and scalable in order to handle the huge datasets in organizations. Data processing in such systems includes creation, retrieval, storage, analysis, presentation, visualization, and any other activity that is typical for IS systems. Big data is often associated with business analytics, cloud services, or industrial systems.

This document presents a brief overview of the state of the art in selected topics of big data research, with the purpose of providing input to a roadmap for research and education at Luleå University of Technology (LTU). The selection of topics is based on assessments of where LTU can make an impact based on current and anticipated research strengths and position with industry (e.g., process industry, data centers and cloud application providers). Topics include distributed systems, mobility, Internet of Things, and advanced analytics.

Contents

1 Background and outline

At LTU a total of seven areas of excellence were defined to promote interdisciplinary research and innovation in strategically important areas. This report is to provide input to the roadmap of two of these areas: Enabling Information Communication Technology (EICT), and Intelligent Industrial Processes (IIP). The document is will be enhanced incrementally and provide part of a work program that may be carried out at LTU together with partners in academia and industry.

The structure of the main parts of the document is as follows. First, the current local arenas for research and innovation on big-data, cloud and data-centers are presented. Second, there are brief overviews of the global industry perspective and the academic perspective for big data respectively. Third, the core of the report contains a number of challenges and opportunities that are potential candidates for a roadmap. This part is structured like a cook book. New "recipes" can be added and improved independently and in parallel.

Reflections are provided in each subsection. Also early conclusions are drawn for each recipe presented. This structure is to localize information in order to make it easy to add new sections and recipes.

2 Big-data arenas at LTU and vicinity

There are several efforts at LTU, or in the vicinity of LTU, that are (or will become) arenas for industry collaboration, funding and visibility of results in the areas of big data, clouds, datacenters, industry automation, etc. The following four arenas are especially important.

Process IT [1] has been around for 10 years and demonstrated successful research in cooperation with process industry. This is one of the major success stories of LTU and it will continue to provide a good platform for research and development.

Cloudberry Datacenters [2] provides a business arena in Research, Innovation, Design and Education, and also attracts and distributes funding for education and pilot projects in cooperation between industry and academia. This arena has received a lot of attention lately. It has established initial funding and a quite impressive partner list. There is support for pilot projects in cooperation between industry and academia (at LTU and elsewhere) that shoud be interesting to researchers. Organizationally, Cloudberry is based at LTU.

LTU Business AB and the county municipality of Norrbotten (swe länstyrelsen i norrbotten) have ongoing work to create a strong datacenter region in the vicinity of LTU that includes efforts on research and innovation. This would establish Sweden as a major place for datacenter and cloud research [3].

Big Data Analytics research group is being established since beginning of 2014 with publications in the area of analytics, and also actively engaged with leading industry players.

3 Industry surveys on big-data

There are a number of surveys performed to assess the priorities and efforts on big data in the industry. Jaspersoft made a survey on enterprise use of big data in corporate decision making [4]. About 60% of the respondents where application developers. Some takeaways are that the most polpular big-data sources are customer relationship management (CRM), Financials, e-commerce, retail, and supply chain information. The most popular data stores are relational databases (56%), MongoDB (23%), analytic databases (14%), Hadoop HDFS (12%), and Hive (4%). The usage of big data is supported by an other survey by BARC [5], which shows the response for 508 corporations (Figure 1).

A survey by Gigaspaces shows that the interest for Real-Time Stream Processing and Cloud-Based Big Data is increasing in today's enterprises. The survey indicated that there is increasing readiness to use streaming solutions to deal with the challenges of

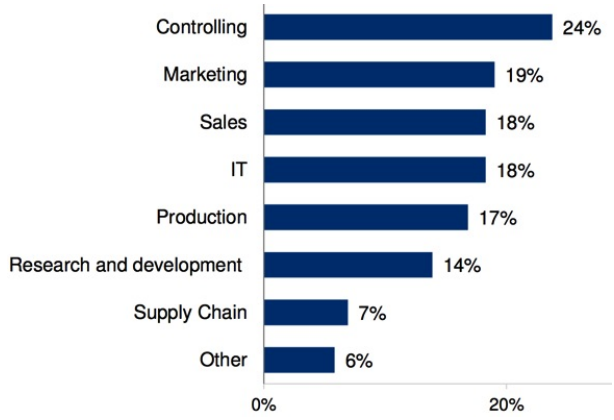

Figure 1: In which areas does your company use big data analysis? (source: [5])

Big Data and speed up big data processing. There is a trend towards being more real-time and handling data on the fly as it is collected.

Some reflections on this follows. The phenomenon of big data probably originates from the areas of business analytics and corporate decision making realizing that market advantages can be obtained by analyzing so far unstructured data on customer and market behavior. Historically, such business has been based on traditional technologies of structured data and therefore a new approach for handling data has been needed to effectively obtain the desired results.

In parallel, or even before the above mentioned, the latest development in cloud services (e.g, software as a service) provided by large companies such as facebook, google, amazon and yahoo has included new technologies in big data that are already in operation. That is a pragmatic example of doing real implementations on big data that were in place before the real phenomenon emerged. It is important to note that that the technologies used for decision support systems (as mentioned in the surveys) are well aligned with what is seen in cloud and data center industry in general. LTU has reasonably good coverage of the mentioned technologies in undergraduate courses, besides that coverage of analytic databases should be strengthened. Keeping well acquainted with the typical frameworks that are used in industry and in experimental research is a key success factor to understand weaknesses and build a long term road map. A couple of these frameworks are covered in the section 5.

The McKinsey Report [13] states that big data could provide significant growth and financial opportunities to the industry. For example, the report suggests that the full potential utilization of big data could increase the net margin by 60% in the US retail sector. In addition, the manufacturing sector could enjoy up to 50% decrease in product development and assembly costs [13]. On the other hand, the public sector could also benefit from big data utilization and analytics. It is expected that the European public sector could experience an annual 0.5% productivity growth, and 250 Billion Euros value.

4 Academic research directions

The investigation of the state-of-the-art in big data literature reveals various dimensions which will be illustrated in the below paragraphs. The aim of this report is to show just a few examples of papers that discuss big data research agendas. We do not cover the plethora of papers providing more specific research results.

Cuzzocrea et al [6] state that big data research is driven by real-life applications and systems, such as representing, modeling, processing, querying and mining massive, distributed, large-scale repositories (mostly being of unstructured nature). They discuss three important aspects of Big Data research: On-Line Analytic Processing (OLAP) over Big Data, Big Data Posting, and Privacy of Big Data. They also depict future research directions. Therein, they state that one of the most significant application scenarios where Big Data arise is in scientific computing. There, scientists and researchers produce huge amounts of data per-day via experiments (e.g., think of disciplines like high-energy physics, astronomy, biology, bio-medicine, and so forth) but extract-

ing useful knowledge for decision making purposes from these massive, large-scale data repositories is almost impossible for actual DBMS-inspired analysis tools. From a methodological point of view, two main research challenges arise. The first one is represented by the issue of conveying Big Data stored in heterogeneous and different-in-nature data sources (e.g., legacy systems, Web, scientific data repositories, sensor and stream databases, social net- works) into a structured, hence well-interpretable, format, which is then ready to populating OLAP data cubes modeling the target analytics. The second one is represented by the issue of managing, processing and transforming the extracted structured data repositories in order to derive Business Intelligence (BI) components like diagrams, plots, dashboards, and so forth, for decision making purposes, hence effectively realizing the complex analytics view. Besides that general statement there is, however, a lack of concrete examples on big data in scientific computing.

Lindman et al [7] propose a research agenda for open data service research. They define the term open data and structure the agenda along the following topics: 1. Technologies, 2. Information, 3. Processes and activities, 4. Products and Services, 5. Participants (including developers, data owners, and service developers), 6. Customers, 7. Environment.

5 Basic technology and toolsets

Undertaking a research on big data requires the use of various tools which belong to the three main domains of big data: storage, processing, and analytics. Hence, decisions need to be made by the data scientist on which toolsets to use while undertaking a research. The toolset ranges from programming languages, software packages (primarily open source), simulators, and theoretical frameworks. There are commonly used toolsets for building proof-of-concepts and for experimental verification, and for theoretical (analytical) verification

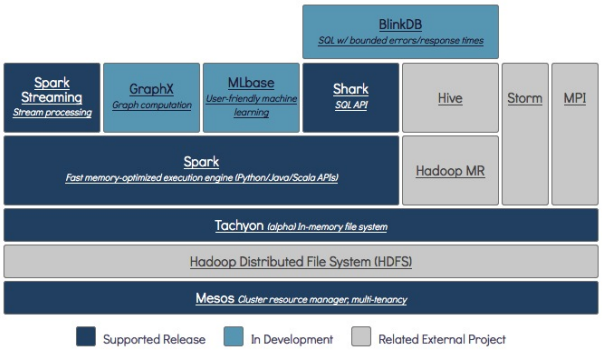

Figure 2: The Berkeley Data Analytics Stack (BDAS)

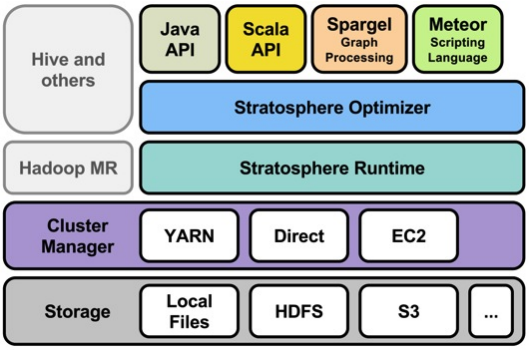

Figure 3: The Stratosphere stack

(qualitatively and quantitatively). The options are however a bit diverse. Agreeing on, and having good understanding of, toolsets is a prerequisite for successful education and research. The traditional toolsets used in CSEE education and research are applicable and necessary in addressing big data. In addition there are some specific big data management and analysis stacks that should be considered. Two of them are mentioned here: Berkeley Data Analytics Stack (BDAS) [8] (Figure 2) and Stratosphere [9] (Figure 3)

Both BDAS and Stratosphere seamlessly support common frameworks such as Hadoop and Hive. They also support programming in Java and Scala, the latter which originates from declarative (funcional) programming. There are scalable data storages with different paradigms of access methods.

The purpose of this report is not to cover these stacks in detail. It can be noted, however, the strong trend in massive scaling and distribution as provided by Yarn or Mesos. Light weight virtualization in the form of process containers is another trend that should be incorporated. Although it is important to have a preferred toolset, it should be noted that the options are many and as new alternatives are frequently occurring, the toolset must be discussed and adapted quite often.

6 MSc in Data Science- suggested postgraduate program

To become successful in industry and in pursuing post graduate studies (especially if an experimental approach is taken) students should become skilled at using common big-data tools and have some experience of building real proof of concepts. There are some efforts on defining big-data curricula for undergraduate and masters level education. Silva et. al. [10] state that the extensive and effective use of systems incorporating big data in many application scenarios, has lead to that these systems have become a key component in the broad landscape of database systems, which creates the need to integrate the study of Big Data Management Systems (BDMS) as part of the computing curricula. They also describe an array of course resources (e.g., virtual machines, sample projects, and in-class exercises) and how these resources support the learning outcomes and enable a hands-on experience with Big Data technologies. They also provides a categorization of the tools and concrete examples of open-source tools in each category.

LTU has already taken a grip on big-data technologies in various undergraduate courses, but further efforts on synchronizing the courses in the education plan is a key to success. Note that the curricula must be broader than the BDMS base described above, and is should also include some specifics that have bearing on current LTU research efforts. There is a great opportunity to mix in current strength such as mobility, networking, process control with big data. This requires further pedagogical study on such newly suggested program.

7 National and international strategic agendas

A current Swedish initiative is *The big data analytics network*, which is coordinated by SICS. This initiative has gathered over 40 partners in industry and academia. LTU is one of them. There is a document *Big data analytics - a research and innovation agenda for Sweden* [11] [12] that describes the directions. Another relevant initiative is the Vinnova funded strategic innovation agenda called the *Internet-of-Things*.

8 Top research groups in the field

In Sweden there are Umeå University (UMU), KTH, SICS, and Chalmers that have especially strong research agendas in cloud and big-data. UMU are specialists in grid computing, virtualization, schelduling and live migration of virtual machines. They also run a large VR funded project in cloud control in cooperation with the control theory group at Lund University. KTH and SICS are drivers of the national big-data agenda, and they have a strong background in big data systems. Futhermore, they are contributors to the Stratosphere project (see below). Chalmers run an SSF funded projekt called *Data Intensive Computing*.

Internationally UC Berkeley AMPLab - AMPLab (Algorithms, Machines, and People Lab) is a collaborative effort at UC Berkeley addressing Big Data analytics problems. Software components built by AMPLab is integrated in the open source Berkeley Data Analytics Stack (BDAS). TU-Berlin DIMA - The Database Systems and Information Management Research group (DIMA) conducts research in the field of information management on cloud through the Stratosphere project. Stratosphere is

Figure 4: SWOT Analysis

the European counterpart of Spark/BDAS and exploits the power of parallel computing for complex information management applications. In 2012, eight Universities and research institutes started a consortium to productize Stratosphere through the Europa-EIT project. MIT Big Data - The MIT Big Data Initiative launched in May 2012, aims to develop scalable systems and platforms across multiple application domains.

In Germany, Fraunhofers IAIS provides solutions to aid enterprises in optimizing their products and services via implementing intelligent knowledge management practices and applications. The IAIS focuses on the analytical technics such as data mining. Their research effort in big data has resulted in the development of a big data architecture called living lab. The lab provides a scalable framework that supports learning and experimentations, with the capability for streaming data.

9 LTU SWOT Analysis

This is a preliminary SWOT analysis:

A reflection from this is that LTU would probably benefit from focusing on experimental research which includes building proof-of concepts, strong experimental evaluations (applying theoretical evaluations where possible), and cross disciplinary research involving active industry cooperation where applicable. This approach would not maximize the number of publications over time ratio, but good experimental work can obtain a high number of references and be suitable for technology transfer to industry and start-ups.

10 Roadmap opportunities

This section contains some opportunities for LTU, selected in consideration of state of the art, arenas and SWOT. The intention is that each of the opportunities should contain both near term and long term challenges, providing opportunities for sustained research until the year 2020.

10.1 Cloud computing, distributed systems and mobility

We foresee that the future cloud model may be much more distributed to support mobility and dynamic private/group clouds that are not necessarily under uniform management. This poses a number of challenges including light weight virtualization, addressing, monitoring, profiling, control and actuation, security, etc. The big data and cloud development is partly based on fundamental research in distributed systems which has been performed since the 1980's. Large scale systems performance can be improved a lot by understanding the minimal requirements and trade-offs. Although there is a big base established knowledge on algoritmns and data structures for distributed systems, there are very good opportunities to provide enhancements that meet particular application demands. Also advancing on light weight virtualization and monitoring/measurement of such systems will cater for sustainable and resource efficient cloud implementations.

10.2 Big Data Analytics

A recent study examined the state-of-the-art in big data analytics [14]. The paper aimed at elucidating knowledge on the characteristics of big data analytics literature as well as explores the areas that lack sufficient research within the big data analytics domain. Towards that end, their research has reviewed 24 publications between 2010 and 2014 [14, 15]. Results of text mining the papers revealed that they belong to three clusters with both common as well as distinct characteristics. The reviewed papers were clustered into three main themes, 1) technical algorithms; 2) processing, cloud computing, opportunities & challenges; and 3) performance, prediction, and distributed systems. The research suggested the below areas of future research in big data analytics: Access to source data set: BDA assumes the availability & access to original data. Such primary data may not always be available for analytics purposes. Accordingly, we believe that big data analytics should be able to be implemented and run without the luxury of primary data. Understandability of discovered patterns: while advances in data mining encompass very powerful algorithms, there are fewer advances on driving the knowledge discovery process towards results appropriate for human consumption. Privacy preserving: The demand for privacy-preservation in data mining emerges in two different related contexts: 1. Personal data must be protected from disclosure towards everyone; & 2. Confidential data must be protected from disclosure towards partners. Algorithm tractability: Mining techniques are beginning to encounter problems as the volume of data requiring analysis grows disproportionately with the comparatively slower improvements in I/O channel speeds. That is, many mining techniques are becoming heavily I/O bound and this is limiting their benefits. Methods to reduce the amount of data have been presented in the literature including statistical methods e.g., dimension reduction.

10.3 IoT, smart Cities, and Big Data

The Internet of Things (IoT) is an ecosystem of (physical) objects connected to the Internet, qualified of identifying themselves and communicating with other objects on the network. IoT enabled sensors and objects to bring unprecedented analysis potentials, together with big opportunities for business benefits. The IoT has significant potential to transform business as well as human life. IoT is promising unprecedented connectivity among objects and the gathering of massive amounts of data, specially in a city that is full of smart objects i.e., smart cities. This may seem straightforward, but it is not! Even if at the organizational level there are numerous challenges. Needless to say that the magnitude of the challenge at the (smart) city level, is rather gigantic. Inevitably, neither organizations nor (smart) cities could reap the benefits of IoT without addressing their core challenges. Basically, IoT requires objects to be connected online and start communicating data and that creates sheer volume of big data. What is required next is the ability to capture, store, process, and analyze this big data created by the IoT objects. Gartners 2012 as well as 2013 repots of Hype Technology Cycles have put big data and IoT as top technologies (first two positions)! IDC, Intel, and UN predicted that by 2020, there would be 200 billion objects connected to the Internet. The big question here is the ability to process big data generated by IoT, in terms of computational power and throughputs! As an example, a simple sensing and monitoring application for a group of 100 sensors is capable of producing 4PB of raw data in a year. In another occasion, we should be able to handle around 500K records (regardless of record size), generated in less than a minute!

10.4 Cyber-physical systems and Data Center Infrastructure Management (DCIM)

Cyber-physical systems (CPS) are systems of physical entities whose operations are monitored, co-

ordinated, controlled and integrated by a computing and communication core. Early examples, that are known as embedded systems, can be found in several areas (e.g., aerospace, automotive, chemical processes, civil infrastructure, energy, healthcare, manufacturing, transportation, entertainment, consumer appliances). It is common to use CPS as an umbrella that covers all kinds of systems with any physical devices in. That makes it very broad. However, for our research agenda a smaller scope should be attainable, like next generation embedded systems, industrial automation, vehicle automation, etc. Datacenters require buildings, equipment and a lot of energy. This effort is focused on building automation and physical equipment automation. Especially there is a challenge in lowering the energy consumption. This effort involves modeling and simulation as well as deploying sensors. It is expected that daily operations of datacenters will be more automated, including autonomous vehicles and robots to replace failing hardware.

11 Conclusions

Within the vast area of *big data*, some initial proposals for research topics until the year 2020 have been made. Also, the need for identifying suitable tool sets (theories, software platforms, simulators, etc) that can serve as a common ground for researchers was discussed. Furthermore, the importance of introducing such tools already in undergraduate teaching was argued. Strengths, weaknesses, opportunities and threats were identified, and a generic approach towards experimental and cross disciplinary research were made. These are topics that should be incrementally re-evaluated to ensure a successful roadmap evolution.

12 Acknowledgements

This work has been funded by the strategic areas of Enabling Information Communication Technology (EICT) and Intelligent Industrial Processes (IIP).

References

[1] "Process it innovations." [Online]. Available: http://www.processitinnovations.se

[2] "Cloudberry dataenters." [Online]. Available: http://www.cloudberry-datacenters.com

[3] "Strategi fr att skapa en vrldsledande teknikregion i norrbotten fr klimatsmarta effektiva datacenter." [Online]. Available: http://www.lansstyrelsen.se/norrbotten/SiteCollectionDocuments/Sv/nyheter/rapport-strategi-skapa-varldsledande-teknikregion-NB-I pdf

[4] "Jaspersoft big data survey," 2014. [Online]. Available: http://www.jaspersoft.com

[5] G. BARC Institute, Wuerzburg, "Big data survey europe - usage, technology and budgets in european best-practice companies," 2013. [Online]. Available: http://www.pmone.com/fileadmin/user_upload/doc/study/BARC_BIG_DATA_SURVEY_EN_final.pdf

[6] A. Cuzzocrea, D. Saccà, and J. D. Ullman, "Big data: A research agenda," in *Proceedings of the 17th International Database Engineering & Applications Symposium*, ser. IDEAS '13. New York, NY, USA: ACM, 2013, pp. 198–203. [Online]. Available: http://doi.acm.org/10.1145/2513591.2527071

[7] J. Lindman, M. Rossi, and V. Tuunainen, "Open data services: Research agenda," in *System Sciences (HICSS), 2013 46th Hawaii International Conference on*, Jan 2013, pp. 1239–1246.

[8] "Berkeley data analytics stack." [Online]. Available: https://amplab.cs.berkeley.edu/software/

[9] "Stratosphere big data analytics platform." [Online]. Available: http://stratosphere.eu

[10] Y. N. Silva, S. W. Dietrich, J. M. Reed, and L. M. Tsosie, "Integrating big data into the computing curricula," in *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '14. New York, NY, USA: ACM, 2014, pp. 139–144. [Online]. Available: http://doi.acm.org.proxy.lib.ltu.se/10.1145/2538862.2538877

[11] O. Görnerup, D. Gillblad, A. Holst, and B. Bjurling, "Big data analytics - a research and innovation agenda for sweden." [Online]. Available: http://www.vinnova.se/PageFiles/0/BigDataAnalytics.pdf

[12] "Sics big data analytics." [Online]. Available: https://www.sics.se/projects/big-data-analytics

[13] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. Hung Byers "Big data: The next frontier for innovation, competition, and productivity" [Online]. Available: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[14] A. Elragal, M. Haddara "Big data abalytics: a test mining-based litterature analysis", NOKOBIT, Fredrikstad, Norway, 2014.

[15] N. Elgeny, A. Elragal "Big data abalytics:: A Literature Review Paper", Lecture Notes in Computer Science Volume 8557, 2014, pp 214-227