BIG DATA ANALYTICS IN INFORMATION RETRIEVAL: PROMISE AND POTENTIAL

[1]DINESH MAVALURU, [2]R. SHRIRAM, [3]VIJAYAN SUGUMARAN

[1]Research Scholar, Department of Computer Applications, B.S. Abdur Rahman University
[2]Professor, Department of Computer Science and Engineering, B.S. Abdur Rahman University
[3]Professor, Department of Decision and Information Sciences, Oakland University

Abstract- Objective: To describe the promise and potential of big data analytics in Information Retrieval. Methods: The paper describes the emerging field of big data analytics in information retrieval, discusses the benefits, outlines an architectural framework and methodology, describes examples reported in the literature, briefly discusses the challenges, and offers conclusions. Results: The paper provides a broad overview of big data analytics for information retrieval researchers. Conclusions: Big data analytics in information retrieval is evolving into a promising field for providing insight from very large data sets and improving outcomes while reducing costs. Its potential is great; however there remain challenges to overcome.

Index Terms- Analytics, Big Data, Framework, Hadoop, Information Retrieval and Methodology.

I. INTRODUCTION

The World Wide Web historically has generated large amounts of data and the techniques for information retrieval from large data sets play a very important role as the size of the world-wide web exceeded 800 million pages in 1999 [1] to 11.5 billion in 2005 [2], and possibly more than 30 billion nowadays. While most data is stored in databases, the current trend is toward big data analytics. Driven by mandatory requirements and the potential to improve the quality of Information retrieval and delivery, reducing the time to retrieve the results from these massive quantities of data (known as 'big data') hold the promise of supporting a wide range of users and information retrieval functions, including among others decision support, surveillance, and knowledge management [3]. Reports say data from the U.S. Google databases alone reached, in 2011, 350 exabytes. At this rate of growth, big data for U.S. World Wide Web will soon reach the zettabyte (1024 gigabytes) scale and, not long after, the yottabyte (1024 zettabytes) [4]. By definition, big data in information retrieval refers to electronic data sets so large and complex that they are difficult (or impossible) to manage with traditional software and/or hardware; nor can they be easily managed with traditional or common data management tools and methods [5]. Big data in information retrieval is overwhelming not only because of its volume but also because of the diversity of data types and the speed at which it must be managed [6]. The totality of data related to user query for information and wellbeing may make up "big data" in the information retrieval systems. It includes all the data which structured or un structured data from different data sources like databases, social media posts, including Twitter feeds (so-called tweets) [7], blogs [8], status updates on Facebook and other platforms, and web pages; news feeds, and articles in journals. For the big data scientist, there is, amongst this vast amount and array of data, opportunity. By discovering associations and understanding patterns and trends within the data, big data analytics has the potential to improve retrieval efficiency, to improve the efficiency and lower costs. Thus, big data analytics applications in information retrieval take advantage of the explosion in data to extract insights for making better informed decisions [9], and as a research category are referred to as big data analytics in information retrieval [10]. When big data is synthesized and analyzed, the above-mentioned associations, patterns and trends are revealed. Healthcare providers and other stakeholders in the information retrieval system can develop systematic understanding of the information, resulting in higher quality information at lesser time and in better outcomes overall [11]. The potential for big data analytics in information retrieval to lead to better outcomes exists across many scenarios, for example: by analyzing user characteristics and the cost and outcomes of system to identify the most efficient and cost effective retrieval systems and offer analysis and tools, thereby influencing provider behavior; applying advanced analytics to user profiles (e.g., segmentation and predictive modeling) to proactively identify individuals who would benefit from using these types of systems; identifying, predicting and minimizing spam by implementing advanced analytic systems for spam detection and checking the accuracy and consistency of claims; and, implementing much nearer to real-time, claim authorization; creating new revenue streams by aggregating and synthesizing user information to provide data and services to third parties, for example, licensing data to assist social

networking companies in identifying users for attachment with sports. Many payers are developing and deploying mobile apps that help user to identify the information and save the time to retrieve that information, locate providers and improve their efficiency of results retrieved by the system. Via analytics, payers are able to monitor adherence to usage and reliability trends that lead to users benefits [12]. This paper provides an overview of big data analytics in information retrieval as it is emerging as a discipline. First, we define and discuss the various advantages and characteristics of big data analytics in information retrieval. Then we describe the architectural framework of big data analytics in information retrieval systems. Third, the big data analytics application development methodology is described. Fourth, we provide examples of big data analytics in information retrieval reported in the literature. Fifth, the challenges are identified. Lastly, we offer conclusions and future directions.

II. BIG DATA ANALYTICS IN INFORMATION RETRIEVAL

Data volume is expected to grow dramatically in the years ahead over the World Wide Web. In addition, information retrieval models are changing; meaningful use and pay for performance are emerging as critical new factors in today's information retrieval systems. Although profit is not and should not be a primary motivator, it is vitally important for organizations to acquire the available tools, infrastructure, and techniques to leverage big data effectively or else risk losing potentially millions of dollars in revenue and profits [19]. What exactly is big data? A report delivered to the U.S. Congress in August 2012 defines big data as "large volumes of high velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of the information." Big data encompasses such characteristics as variety, velocity and, specifically with respect to information retrieval systems, veracity. Existing analytical techniques can be applied to the vast amount of existing information to reach a deeper understanding of outcomes, which then can be applied at the point of retrieval.

III. ADVANTAGES TO INFORMATION RETRIEVAL

By combining different retrieval techniques and effectively using big data, information retrieval systems ranging from small and multi-provider groups to large networks and organizations stand to realize significant benefits. Potential benefits include increasing the efficiency at earlier stages when they can be retrieved more easily and effectively; managing specific data bases and detecting results more quickly and efficiently related to the user query. Numerous questions can be addressed with big data analytics. Certain developments or outcomes may be predicted and/or estimated based on vast amounts of historical data, such as: length of stay (LOS); users who will choose the results; users who likely will not benefit from results retrieved; complications, etc. McKinsey estimates that big data analytics can enable more than $300 billion in savings per year in U.S. alone, two thirds of that through reductions of approximately 8% in national expenditures. Big data could help reduce waste and inefficiency in the following areas: Comparative effectiveness research to determine more relevant and cost-effective ways to retrieve results for the given user query.

Research & development in statistical tools and algorithms to improve information retrieval system design and user access to the system, thus reducing trial failures and speeding new frameworks to users; and analyzing data to identify follow-on indications and discover adverse effects before it reaches the user.

Analyzing data patterns and tracking data outbreaks and transmission to improve data surveillance and speed response to the given user query; faster development of more accurately targeted users, e.g., choosing the annual data access; and, turning large amounts of data into actionable information that can be used to identify needs, provide services, and predict and prevent crises, especially for the benefit of users. In addition, big data analytics in information retrieval can contribute to suggestion based retrieval systems: Combine and analyze a variety of structured and unstructured data, financial and operational data, clinical data, and genomic data to match queries with outcomes. Apply advanced analytics to user profiles (e.g., segmentation and predictive modeling) to identify individuals who would benefit from proactive or minor changes in the system.

IV. THE 4 "VS" OF BIG DATA ANALYTICS IN INFORMATION RETRIEVAL

Like big data in information retrieval, the analytics associated with big data is described by three primary characteristics: volume, velocity and variety. Over time, related data will be created and accumulated continuously, resulting in an incredible volume of data. The already daunting volume of existing data includes different formats and different sources of data. Newer forms of big data, such as 3D imaging, animation and biometric sensor readings, are also fueling this exponential growth. Fortunately, advances in data management, particularly virtualization and cloud computing, are facilitating the development of platforms for more effective capture, storage and

manipulation of large volumes of data. Data is accumulated in real-time and at a rapid pace, or high velocity. The constant flow of new data accumulating at unprecedented rates presents new challenges. Just as the volume and variety of data that is collected and stored has changed, so too has the velocity at which it is generated and that is necessary for retrieving, analyzing, comparing and making decisions based on the output. Velocity of mounting data increases with data that represents regular monitoring, such as multiple daily updating, readings of different sources, and static data. Meanwhile, in many situations, constant real-time data can mean the difference between structured and unstructured data. Future applications of real-time data, such as detecting infections as early as possible, identifying them swiftly and applying the right procedure could reduce the risk. The ability to perform real-time analytics against such high-volume data in motion and across all specialties would revolutionize the information retrieval systems. Therein lies the variety. As the nature of data has evolved, so too have analytics techniques scaled up to the complex and sophisticated analytics necessary to accommodate volume, velocity and variety. Gone are the days of data collected exclusively in electronic records and other structured formats. Increasingly, the data is in multimedia format and unstructured. The enormous variety of data— structured, unstructured and semi-structured—is a dimension that makes data both interesting and challenging.

Structured data is data that can be easily stored, queried, recalled, analyzed and manipulated by machine. Historically, in some databases, structured and semi-structured data includes instrument readings and data generated by the ongoing conversion of paper records to electronic data.

Already, new data streams—structured and unstructured are cascading into the fitness or effectives of retrieval systems, social media research and other sources. But relatively little of this data can presently be captured, stored and organized so that it can be manipulated by computers and analyzed for useful information. Healthcare applications in particular need more efficient ways to combine and convert varieties of data including automating conversion from unstructured to structured data.

The structured data include familiar input record fields such as user name, data of birth, address, codes, and other information easily coded into and handled by automated databases. The need to field-code data at the point of care for electronic handling is a major barrier to acceptance of the user. On the other hand, most providers agree that an easy way to reduce prescription errors is to use digital entries rather than unstructured data. The potential of big data in information retrieval lies in combining traditional data with new forms of data. We are already seeing data sets from a multitude of sources support faster and more reliable research and discovery.

Researchers have introduced a fourth characteristic, veracity, or ‘data assurance’. That is, the big data, analytics and outcomes are error-free and credible. Of course, veracity is the goal, not (yet) the reality. Data quality issues are of acute concern in information retrieval for two reasons: decisions depend on having the accurate information, and the quality of data, especially unstructured data, is highly variable and all too often incorrect.

Veracity assumes the simultaneous scaling up in granularity and performance of the architectures and platforms, algorithms, methodologies and tools to match the demands of big data. The analytics architectures and tools for structured and unstructured big data are very different from traditional business intelligence (BI) tools. They are necessarily of industrial strength. For example, big data analytics in information retrieval would be executed in distributed processing across several servers (“nodes”), utilizing the paradigm of parallel computing and ‘divide and process’ approach. Likewise, models and techniques—such as data mining and statistical approaches, algorithms, visualization techniques—need to take into account the characteristics of big data analytics. Traditional data management assumes that the warehoused data is certain, clean, and precise.

Veracity in web data faces many of the same issues as in traditional data. Improving efficiency, avoiding errors and reducing costs depend on high-quality data and efficiency, accuracy and more precise targeting of processes by retrieving. But increased variety and high velocity hinder the ability to cleanse data before analyzing it and making decisions, magnifying the issue of data “trust”. The ‘4Vs’ are an appropriate starting point for a discussion about big data analytics in information retrieval. But there are other issues to consider, such as the number of architectures and platforms, and the dominance of the open source paradigm in the availability of tools. Consider, too, the challenge of developing methodologies and the need for user-friendly interfaces. While the overall cost of hardware and software is declining, these issues have to be addressed to harness and maximize the potential of big data analytics in information retrieval.

V. ARCHITECTURAL FRAMEWORK

Architectural framework is the conceptual framework for a big data analytics project in information retrieval, which is similar to that of a traditional informatics or

analytics project. The key difference lies in how processing is executed. In a regular analytics project, the analysis can be performed with a business intelligence tool installed on a stand-alone system, such as a desktop or laptop. Because big data is by definition large, processing is broken down and executed across multiple nodes. The concept of distributed processing has existed for decades. What is relatively new is its use in analyzing very large data sets as retrieval system providers start to tap into their large data repositories to gain insight for making better-informed decisions.

Furthermore, open source platforms such as Hadoop/ MapReduce, available on the cloud, have encouraged the application of big data analytics in information retrieval. While the algorithms and models are similar, the user interfaces of traditional analytics tools and those used for big data are entirely different; traditional analytics tools have become very user friendly and transparent. Big data analytics tools, on the other hand, are extremely complex, programming intensive, and require the application of a variety of skills. They have emerged in an ad hoc fashion mostly as open-source development tools and platforms, and therefore they lack the support and user-friendliness that vendor-driven proprietary tools possess. As Figure 1 indicates, the complexity begins with the data itself. Big data in information retrieval can come from internal and external sources often in multiple formats (flat files, .csv, relational tables, ASCII/text, etc.) and residing at multiple locations (geographic as well as in different providers' sites) in numerous legacy and other applications (transaction processing applications, databases, etc.). Sources and data types include:

- Web and social media data: Clickstream and interaction data from Facebook, Twitter, LinkedIn, blogs, and the like. It can also include health plan websites, smartphone apps, etc.
- Machine to machine data: readings from remote sensors, meters, and other vital sign devices.
- Big transaction data: health care claims and other billing records increasingly available in semi- structured and unstructured formats.
- Biometric data: finger prints, genetics, handwriting, retinal scans, x-ray and other medical images, blood pressure, pulse and pulse-oximetry readings, and other similar types of data.
- Human-generated data: unstructured and semi- structured data such as EMRs, physicians notes, email, and paper documents.

For the purpose of big data analytics, this data has to be pooled. In the second component the data is in a 'raw' state and needs to be processed or transformed, at which point several options are available. A service oriented architecture based approach combined with web services (middleware) is one possibility. The data stays raw and services are used to call, retrieve and process the data. Another approach is data warehousing wherein data from various sources is aggregated and made ready for processing, although the data is not available in real time. Via the steps of extract, transform, and load (ETL), data from diverse sources is cleansed and readied. Depending on whether the data is structured or unstructured, several data formats can be input to the big data analytics platform.

In the next component of the conceptual framework, several decisions are made regarding the data input approach, distributed design, tool selection and analytics models. Finally, on the far right, the four typical applications of big data analytics in information retrieval are shown. These include queries, reports, OLAP, and data mining. Visualization is an overarching theme across the four applications. Drawing from such fields as statistics, computer science, applied mathematics and economics, a wide variety of techniques and technologies has been developed and adapted to aggregate, manipulate, analyze, and visualize big data in information retrieval.

The most significant platform for big data analytics is the open-source distributed data processing platform Hadoop (Apache platform), initially developed for such routine functions as aggregating web search indexes. It belongs to the class "NoSQL" technologies—others include CouchDB and MongoDB—that evolved to aggregate data in unique ways. Hadoop has the potential to process extremely large amounts of data mainly by allocating partitioned data sets to numerous servers (nodes), each of which solves different parts of the larger problem and then integrates them for the final result. Hadoop can serve the twin roles of data organizer and analytics tool.

It offers a great deal of potential in enabling enterprises to harness the data that has been, until now, difficult to manage and analyze. Specifically, Hadoop makes it possible to process extremely large volumes of data with various structures or no structure at all.

But Hadoop can be challenging to install, configure and administer, and individuals with Hadoop skills are not easily found. Furthermore, for these reasons, it appears organizations are not quite ready to embrace Hadoop completely.

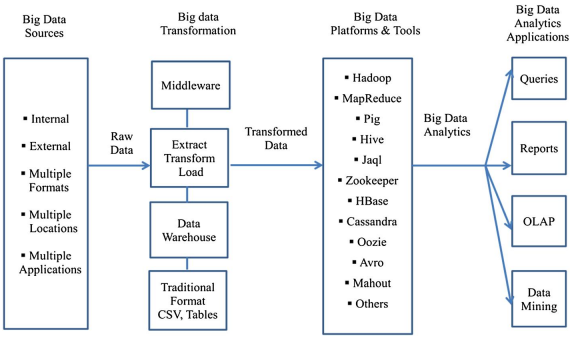

Figure 1. Conceptual Architecture of big data analytics

The surrounding ecosystem of additional platforms and tools supports the Hadoop distributed platform. Numerous vendors—including AWS, Cloudera, Hortonworks, and MapR Technologies—distribute open source Hadoop platforms. Many proprietary options are also available, such as IBM's BigInsights. Further, many of these platforms are cloud versions, making them widely available. Cassandra, HBase, and MongoDB, described above, are used widely for the database component. While the available frameworks and tools are mostly open source and wrapped around Hadoop and related platforms, there are numerous trade-offs that developers and users of big data analytics in information retrieval must consider. While the development costs may be lower since these tools are open source and free of charge, the downsides are the lack of technical support and minimal security. These are, of course, significant drawbacks, and therefore the trade-offs must be addressed. Additionally, these platforms/tools require a great deal of programming, skills the typical end-users of information retrieval systems may not possess. Furthermore, considering the only recent emergence of big data analytics in information retrieval, governance issues including ownership, privacy, security, and standards have yet to be addressed. In the next section we offer an applied big data analytics methodology in information retrieval system to develop and implement a big data project for retrieval system providers.

VI. METHODOLOGY

While several different methodologies are being developed in this rapidly emerging discipline, here we outline one that is practical and hands-on. In Step 1, the interdisciplinary big data analytics for information retrieval team develops a 'concept statement'. This is a first cut at establishing the need for such a project. The concept statement is followed by a description of the project's significance. The information retrieval system will note that there are trade-offs in terms of alternative options, cost, scalability, etc. Once the concept statement is approved, the team can proceed to Step 2, the proposal development stage. Here, more details are filled in. Based on the concept statement, several questions are addressed: What problem is being addressed? Why is it important and interesting to the information retrieval provider? What is the case for a 'big data' analytics approach? The project team also should provide background information on the problem domain as well as prior projects and research done in this domain. Next, in Step 3, the steps in the methodology are fleshed out and implemented. The concept statement is broken down into a series of propositions. Simultaneously, the independent and dependent variables or indicators are identified. The data sources, as outlined in Figure 1, are also identified; the data is collected, described, and transformed in preparation for analytics. A very important step at this point is platform/tool evaluation and selection. There are several options available, as indicated previously, including AWS Hadoop, Cloudera, and IBM Big Insights. The next step is to apply the various big data analytics techniques to the data. This process differs from routine analytics only in that the techniques are scaled up to large data sets.

Through a series of iterations and what-if analyses, insight is gained from the big data analytics. From the insight, informed decisions can be made. In Step 4, the models and their findings are tested and validated and presented to stakeholders for action. Implementation is a staged approach with feedback loops built in at each stage to minimize risk of failure.

We draw on publicly available material from numerous sources, including vendor sites. In this emerging discipline, there is little independent research to cite. These examples are from secondary sources. Nevertheless, they are illustrative of the potential of big data analytics in information retrieval.

VII. CHALLENGES

At minimum, a big data analytics platform in information retrieval must support the key functions necessary for processing the data. The criteria for platform evaluation may include availability, continuity, ease of use, scalability, ability to manipulate at different levels of granularity, privacy and security enablement, and quality assurance. In addition, while most platforms currently available are open source, the typical advantages and limitations of open source platforms apply. To succeed, big data analytics in information retrieval systems needs to be packaged so it is menu driven, user-friendly and transparent. Real-time big data analytics is a key requirement for information retrieval systems. The lag between data collection and processing has to be addressed. The dynamic availability of numerous analytics algorithms, models and methods in a pull-down type of menu is also necessary for large-scale adoption. The important managerial issues

of ownership, governance and standards have to be considered. And woven through these issues are those of continuous data acquisition and data cleansing. Data is rarely standardized, often fragmented, or generated in legacy IT systems with incompatible formats. This great challenge needs to be addressed as well.

CONCLUSIONS

Big data analytics has the potential to transform the way information retrieval systems use sophisticated technologies to gain insight from their data and other data repositories and make informed decisions. In the future we'll see the rapid, widespread implementation and use of big data analytics across the IT industry. To that end, the several challenges highlighted above, must be addressed. As big data analytics becomes more mainstream, issues such as guaranteeing privacy, safeguarding security, establishing standards and governance, and continually improving the tools and technologies will garner attention. Big data analytics and applications in information retrieval are at a nascent stage of development, but rapid advances in platforms and tools can accelerate their growing process.

REFERENCES

[1] Bian J, Topaloglu U, Yu F, Yu F, "Towards Large-scale Twitter Mining for Drugrelated Adverse Events," Maui, Hawaii: SHB; 2012.

[2] S. Lawrence and C. L. Giles, "Accessibility of information on the web," Nature, vol. 400, pp. 107-109, 1999.

[3] A. Gulli and A. Signorini, "The indexable web is more than 11.5 billion pages," Proceedings of 14th International Conference on World Wide Web, Special interest tracks and posters, 2005.

[4] Q. Liu and J. Wang, "Two k-winners-take-all networks with discontinuous activation functions," Neural Networks, vol. 21, no. 2-3, pp. 406-413, 2008.

[5] Put big data to work for your business – with SAP solutions and technology. Company website, SAP, 2013.

[6] Daniel J. Abadi. Tradeoffs between Parallel Database Systems, Hadoop, and HadoopDB as Platforms for Petabyte-Scale Analysis. In Proceedings of the 22nd International Conference on Scientific and Statistical Database Management, SSDBM '10, pages 1–3, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3-642-13817-9, 978-3-642-13817-1.

[7] Daniel J. Abadi. Consistency Tradeoffs in Modern Distributed Database System Design: CAP is Only Part of the Story. Computer, 45(2):37–42, February 2012.

[8] Andreas Bauer and Holger Günzel, editors. Data Warehouse Systeme: Architektur, Entwicklung, Anwendung. dpunkt.verlag GmbH, 4th edition edition, 2013.

[9] Edmon Begoli. A Short Survey on the State of the Art in Architectures and Platforms for Large Scale Data Analysis and Knowledge Discovery from Data. In Joint Working IEEE/IFIP Conference on Software Architecture and European Conference on Software Architecture, WICSA/ECSA '12, pages 177–183, New York, NY, USA, 2012.

[10] Edmon Begoli and James Horey. Design Principles for Effective Knowledge Discovery from Big Data. In Joint Working IEEE/IFIP Conference on Software Architecture and European Conference on Software Architecture, WICSA/ECSA '12, pages 215–218, New York, NY, USA, 2012.

[11] Kevin S. Beyer, Vuk Ercegovac, Rainer Gemulla, Andrey Balmin, Mohamed Y. Eltabakh, Carl-Christian Kanne, Fatma Özcan, and Eugene J. Shekita. Jaql: A Scripting Language for Large Scale Semistructured Data Analysis. In Proceedings of the VLDB Endowment, volume 4 of PVLDB, pages 1272–1283, 2011.

[12] Raghupathi W: Data Mining in Health Care. In Healthcare Informatics: Improving Efficiency and Productivity. Edited by Kudyba S. Taylor & Francis; 2010:211–223.

★★★