Big data Analytic – A Review

Anuradha Bhatia,
Sr. Lecturer,
VES Polytechnic, Chembur , Mumbai-400094,
anubhatia31@rediffmail.com

Abstract - Today, software for every layer of the enterprise stack is available under a permissive open source license. In fact, the world's most popular OS (Linux), Web server (Apache HTTP Server), relational database (MySQL), and Apache Hadoop distribution (CDH from Cloudera – downloaded more than all alternatives combined) are all open source software. Many people intuitively recognize the surface benefits of source code being available for inspection and modification. However, for enterprise software buyers, it's also important to understand how deep, direct involvement by your support vendor in the open source development process, properly conceived for customer benefit, has a tangible impact beyond the simple availability of source code. Otherwise, the "open source" label has limited practical meaning beyond its first-glance appeal.

Keyword: Hadoop, MapReduce, distributed systems, Clustering, HDFS, cloudera, open source,Big data, Apache Hadoop, MapReduce, data refinery, data warehouse.

I. INTRODUCTION

Big datais a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, cu ration, storage, search, sharing, transfer, analysis, and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."

We are awash in a flood of data today. In a broad range of application areas, data is being collected at unprecedented scale. Decisions that previously were based on guesswork, or on pain staking constructed models of reality, can now be made based on the data itself. Such Big Data analysis now drives nearly every aspect of our modern society, including mobile services, Retail, manufacturing, financial services, life sciences, and physical sciences. Scientific research has been revolutionized by Big Data.

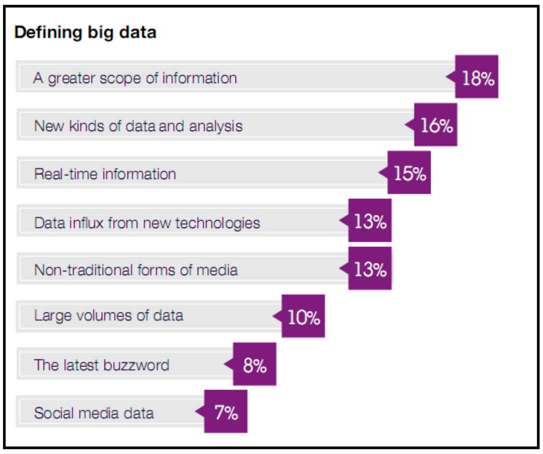

Fig 1: Defining Big Data

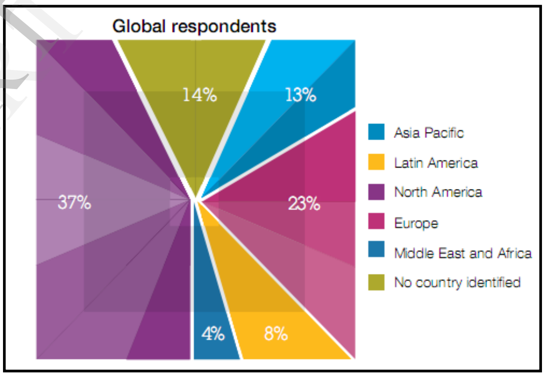

Fig 2: Global Respondents of Big Data

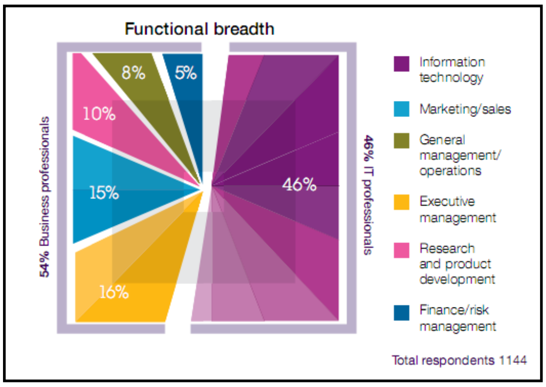

Fig 3: Functional Breadth of Big Data

Heterogeneity, scale, timeliness, complexity, and privacy problems with Big Data impede progress at all phases of the pipeline that can create value from data. The problems start right away during data acquisition, when the data tsunami requires us to make decisions, currently in an ad hoc manner, about what data to keep and what to discard, and how to store what we keep reliably with the right metadata. Much data today is not natively in structured format; for example, tweets and blogs are weakly structured pieces of text, while images and video are structured for storage and display, but not for semantic content and search: transforming such content into a structured format for later analysis is a major challenge. The value of data explodes when it can be linked with other data, thus data integration is a major creator of value. Since most data is directly generated in digital format today, we have the opportunity and the challenge both to influence the creation to facilitate later linkage and to automatically link previously created data. Data analysis, organization, retrieval, and modelling are other foundational challenges. Data analysis is a clear bottleneck in many applications, both due to lack of scalability of the underlying algorithms and due to the complexity of the data that needs to be analyzed. Finally, presentation of the results and its interpretation by non-technical domain experts is crucial to extracting actionable knowledge.

Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers". What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyse the data set in its domain. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options.

For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."Enterprise users have been building analytic applications for years. Today, nearly every enterprise already has analytic applications that they use every day, and these are relatively well understood and captured in the graphic below.

Data comes from a set of data sources – most typically from the enterprise applications: ERP, CRM, custom applications that power the business

- That data is extracted, transformed, and loaded into a data system: a relational Database Management System (RDBMS), an Enterprise Data Warehouse (EDW), or even a Massively Parallel Processing system (MPP)
- A set of analytical applications – either packaged (e.g. SAS) or custom, then point at the data in those systems to enable users to garner insights from that data.

The general flow looks something like this and is depicted in Fig4 below:

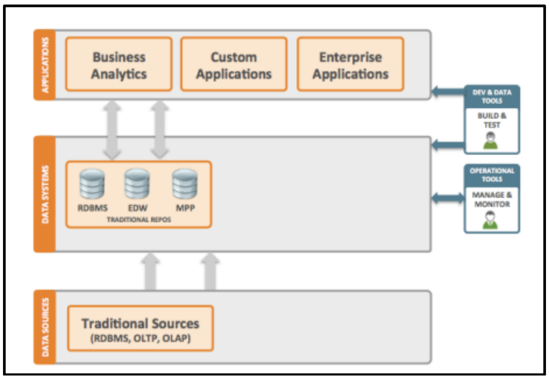

Fig 4: Architecture

Times have changed however. Recently, there has been an explosion of data entering into this
Landscape. And it isn't just more records coming from the ERP or CRM systems: it is an entirely new class of data that was never envisioned when those data systems first came into being. It is machine generated data, sensor data, social data, web logs and other such types that are both growing exponentially, but also often (but not always) unstructured in nature.

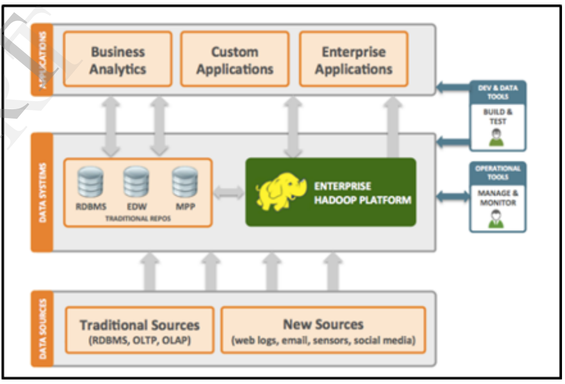

Fig 5: Hadoop Structure

It also includes data that was once thought of as low to medium value or even exhaust data, too expensive to store and analyse. And it is this type of data that is turning the conversation from "data analytics" to "big data analytics": because so much insight can be gleaned for business advantage.

II. DATA AVAILABILITY

As depicted in Figure 10, we see how data availability requirements change dramatically as companies mature their big data efforts. Analysis of responses revealed that no matter the stage of big data adoption, organizations face increasing demands to reduce the latency from data capture to action. Executives, it seems, are increasingly considering the

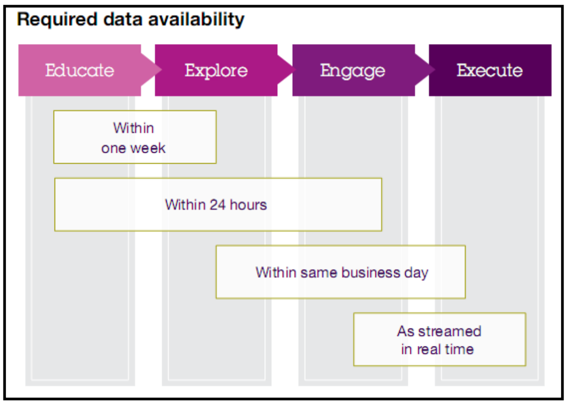

Fig 6: Big Data required Data Availability

value of timely data in making strategic and day-to-day business decisions. Data is no longer just something that supports a decision; it is a mission-critical component in making that decision.

III. BIG DATA OBSTACLES

Challenges that inhibit big data adoption differ as organizations move through each of the big data adoption stages. But our findings show one consistent challenge – regardless of stage – and that is the ability to articulate a compelling business case (see Figure 11). At every stage, big data efforts come under fiscal scrutiny. The current global economic landscape has left businesses with little appetite for new technology investments without measurable benefits – a requirement that, of course, is not exclusive to big data initiatives. After organizations successfully implement POCs, the biggest challenge becomes finding the skills to operationalize big data, including: technical, analytical and governance skills.

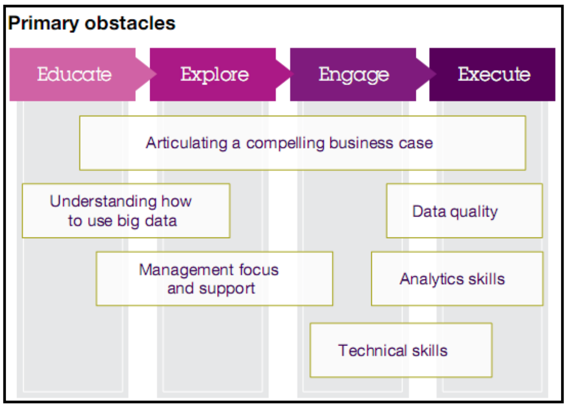
Primary obstacles
Educate
Explore
Engage
Execute
Articulating a compelling business case
Understanding how to use big data
Data quality
Management focus and support
Analytics skills
Technical skills

Fig 7: Big Data Obstacles

IV. RECOMMENDATIONS: ULTIVATING BIG DATA ADOPTION

Work Study findings provided new insights into how organizations at each stage are advancing their big data efforts. Driven by the need to solve business challenges, in light of both advancing technologies and the changing nature of data, organizations are starting to look closer at big data's potential benefits. To extract more value from big data, we offer a broad set of recommendations to organizations as they proceed down the path of big data.

Mass digitization, one of the forces that helped to create the surge in big data, has also changed the balance of power between the individual and the institution. If organizations are to understand and provide value to empowered customers and citizens, they have to concentrate on getting to know their customers as individuals. They will also need to invest in new technologies and advanced analytics to gain better insights into individual customer interactions and preferences.

V. HADOOP - THE ACTIVE ARCHIVE

In a 2003 interview with ACM, Jim Gray claimed that hard disks can be treated as tape. While it may take many more years for magnetic tape archives to be retired, today some portions of tape workloads are already being redirected to Hadoop clusters. This shift is occurring for two fundamental reasons. First, while it may appear inexpensive to storedata on tape, the true cost comes with the difficulty of retrieval. Not only is the data stored offline, requiring hours if not days to restore, but tape cartridges themselves are prone to degradation over time making data loss a reality and forcing companies to factor in those costs. To make matters worse, tape formats change every couple of years requiring

Organizations to either perform massive data migrations to the newest tape format or risk the inability to restore data from obsolete tapes.

VI. CONCLUSION

We have entered an era of Big Data. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

Hadoop and the data warehouse will often work together in a single information supply chain.When it comes to Big Data, Hadoop excels in handling raw , unstructured and complex data withvastprogrammingflexibility. Data warehouses also manage big structured data, integrating subject areas and providing interactive performance through BI tools. It is rapidly becoming a symbiotic relationship. Some differences are clear and identifying workloads or data that runs best on one or the other will be dependent on your organization and use cases. As with all platform selections, careful analysis of the business and technical requirements should be done before platform selection to ensure the best outcome. Having both Hadoop and a data warehouse onsite greatly helps everyone learn when to use which.

REFERENCES

[1]Dean, J. and Ghemawat, S., “MapReduce: Simplified Data Processing on Large Clusters.” Appeared in Proceedings of the Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, 2004.Available online at http://labs.google.com/papers/mapreduce.html, March 2010.

[2]Apache Nutch project, http://lucene. Apache .org/nutch/, March 2010.

[3]Apache Hadoop project, http://hadoop. Apache .org/, March 2010.

[4]Patterson, J., “The Smart Grid: Hadoop at the Tennessee Valley Authority (TVA),” Cloudera Blog, http://www.cloudera.com/ blog/2009/06/smart-grid-hadoop-tennessee-valley -authority-tva/, June 2, 2009.

[5] White, T., Hadoop: The Definitive Guide. O’Reilly Media, May 2009.

[6] Lin, J. and Dyer, C., Data-Intensive Text Processing with MapReduce. To be published by Morgan and Claypool. Pre-press edition available at http://www.umiacs.umd.edu/~ jimmylin/MapReduce-book-20100219.pdf, February 2010.

[7]Segaran, T. and Hammerbacher, J., Beautiful Data. O’Reilly Media, July 2009.

[8] Hadoop and the Data Warehouse:When to Use Which,Dr. AmrAwadallah,Founder and CTO, Cloudera, Inc.DanGraham,General Manager, Enterprise Systems, Teradata Corporation.

[9] Anuradha Bhatia and GauravVaswani, “Big Data– An Overview”, International Journal of Engineering Sciences & Research Technology, Vol 2, No8,August (2013). (www.ijesrt.org) .

[10] Hey, T., Tansley, S. & Tolle, K. (2009) The Fourth Paradigm. “Data-intensive scientifc discovery”, Microsoft.

[11] Hilbert, M. & Lopez, P. (2011) “The world’s technological capacity to store, communicate and compute information”, Science 332, 1 April 2011, 60-65.

[12] IDC (2010) “IDC Digital Universe Study, sponsored by EMC”, May 2010, available at http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm

[13] ICSU Strategic Plan 2006-2011, International Council for Science, Paris, 64pp.

[14] All the reports are available at the ICSU website, www.icsu.org

[15] Big Data – A review by Anuradha Bhatia and Gaurav Vaswani in IJESRT , [Bhatia,2(8): August, 2013].

[16] http://www.tei-c.org/index.xml

[17] ttp://history.state.gov/historicaldocuments

[18] Leetaru, K. (forthcoming). “Fulltext Geocoding Versus Spatial Metadata For Large Text Archives: Towards a Geographically Enriched Wikipedia”, D-Lib Magazine