Volume 6, Issue 7, July 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Enhancing Data Analytics Using Big Data

Dr. Maulik N. Pandya
H.O.D. – Assistant Professor, M.Sc. IT Department
Shri A. N. Patel P. G. Institute, Anand, Gujarat, India

Kalpit G. Soni
Assistant Professor, M.Sc. IT Department
Shri A. N. Patel P. G. Institute, Anand, Gujarat, India

Abstract - Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. Big data is more than simply a matter of size; it is an opportunity to find insights in new and emerging types of data and content, to make your business more agile, and to answer questions that were previously considered beyond your reach. Until now, there was no practical way to harvest this opportunity. Big Data bring new opportunities to modern society and challenges to data scientists. On the one hand, Big Data hold great promises for discovering subtle population patterns and heterogeneities that are not possible with small-scale data. On the other hand, the massive sample size and high dimensionality of Big Data introduce unique computational and statistical challenges, including scalability and storage bottleneck, noise accumulation, spurious correlation, incidental and measurement errors. These challenges are distinguished and require new computational and statistical paradigm. This paper gives overviews on the salient features of Big Data and how these features impact on paradigm change on statistical and computational methods as well as computing architectures. We also provide various new perspectives on the Big Data analysis and computation.

Keywords: Big data, Operational Big Data, Analytical Big Data, Hadoop,

I. INTRODUCTION

Big data is the term for 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, curation, 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."

As of 2012, limits on the size of data sets that are feasible to process in a reasonable amount of time were on the order of exabytes of data, that is, millions of terabytes. Scientists regularly encounter limitations due to large data sets in many areas, including meteorology, genomics, connectomics, complex physics simulations, and biological and environmental research. The limitations also affect Internet search, finance and business informatics. Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency identification readers, and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 exabytes (2.5×10^{18}) of data were created. The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization.

II. HISTORY

In a 2001 research report and related lectures, META Group (now Gartner) analyst Doug Laney defined data growth challenges and opportunities as being three-dimensional,

1. increasing **V**olume (amount of data)
2. **V**elocity (speed of data in and out)
3. **V**ariety (range of data types and sources)

Now much of the industry, continue to use this "3Vs" model for describing big data. In 2012, Gartner updated its definition as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization." Additionally, a new **V** "Veracity" is added by some organizations to describe it.

If Gartner's definition (the 3Vs) is still widely used, the growing maturity of the concept fosters a more sound difference between big data and Business Intelligence, regarding data and their use:

- Business Intelligence uses descriptive statistics with data with high information density to measure things, detect trends etc.;
- Big data uses inductive statistics and concepts from nonlinear system identification [19] to infer laws (regressions, nonlinear relationships, and causal effects) from large data sets [20] to reveal relationships, dependencies and perform predictions of outcomes and behaviors.

III. SELECTING A BIG DATA TECHNOLOGY: OPERATIONAL VS. ANALYTICAL

The Big Data landscape is dominated by two classes of technology: systems that provide operational capabilities for real-time, interactive workloads where data is primarily captured and stored; and systems that provide analytical capabilities for retrospective, complex analysis that may touch most or all of the data. These classes of technology are complementary and frequently deployed together.

Operational and analytical workloads for Big Data present opposing requirements and systems have evolved to address their particular demands separately and in very different ways. Each has driven the creation of new technology architectures. Operational systems, such as the NoSQL databases, focus on servicing highly concurrent requests while exhibiting low latency for responses operating on highly selective access criteria. Analytical systems, on the other hand, tend to focus on high throughput; queries can be very complex and touch most if not all of the data in the system at any time. Both systems tend to operate over many servers operating in a cluster, managing tens or hundreds of terabytes of data across billions of records.

3.1 Operational Big Data

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, and cheaper and faster to implement.

In addition to user interactions with data, most operational systems need to provide some degree of real-time intelligence about the active data in the system. For example in a multi-user game or financial application, aggregates for user activities or instrument performance are displayed to users to inform their next actions. Some NoSQL systems can provide insights into patterns and trends based on real-time data with minimal coding and without the need for data scientists and additional infrastructure.

3.2 Analytical Big Data

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.

As applications gain traction and their users generate increasing volumes of data, there are a number of retrospective analytical workloads that provide real value to the business. Where these workloads involve algorithms that are more sophisticated than simple aggregation, MapReduce has emerged as the first choice for Big Data analytics. Some NoSQL systems provide native MapReduce functionality that allows for analytics to be performed on operational data in place. Alternately, data can be copied from NoSQL systems into analytical systems such as Hadoop for MapReduce.

Table 1. Overview of Operational vs. Analytical Systems

	Operational	**Analytical**
Latency	1ms - 100ms	1min - 100min
Concurrency	1000 - 100,000	1 - 10
Access Pattern	Writes and Reads	Reads
Queries	Selective	Unselective
Data Scope	Operational	Retrospective
End User	Customer	Data Scientist
Technology	NoSQL	Hadoop MapReduce, MPP Database

IV. CONSIDERATIONS FOR DECISION MAKERS

While many Big Data technologies are mature enough to be used for mission-critical, production use cases, it is still nascent in some regards. Accordingly, the way forward is not always clear. As organizations develop Big Data strategies, there are a number of dimensions to consider when selecting technology partners, including:

1. Online vs. Offline Big Data
2. Software Licensing Models
3. Community
4. Developer Appeal

5. Agility
6. General Purpose vs. Niche Solutions

4.1 Online vs. Offline Big Data

Big Data can take both online and offline forms. Online Big Data refers to data that is created, ingested, trans- formed, managed and/or analyzed in real-time to support operational applications and their users. Big Data is born online. Latency for these applications must be very low and availability must be high in order to meet SLAs and user expectations for modern application performance. This includes a vast array of applications, from social networking news feeds, to analytics to real-time ad servers to complex CRM applications. Examples of online Big Data databases include MongoDB and other NoSQL databases.

Offline Big Data encompasses applications that ingest, transform, manage and/or analyze Big Data in a batch context. They typically do not create new data. For these applications, response time can be slow (up to hours or days), which is often acceptable for this type of use case. Since they usually produce a static (vs. operational) output, such as a report or dashboard, they can even go offline temporarily without impacting the overall goal or end product. Examples of offline Big Data applications include Hadoop-based workloads; modern data warehouses; extract, transform, load (ETL) applications; and business intelligence tools.

Organizations evaluating which Big Data technologies to adopt should consider how they intend to use their data. For those looking to build applications that support real-time, operational use cases, they will need an operational data store like MongoDB. For those that need a place to conduct long-running analysis offline, perhaps to inform decision-making processes, offline solutions like Hadoop can be an effective tool. Organizations pursuing both use cases can do so in tandem, and they will sometimes find integrations between online and offline Big Data technologies. For instance, MongoDB provides integration with Hadoop.

4.2 Software License Model

There are three general types of licenses for Big Data software technologies:

Proprietary. The software product is owned and controlled by a software company. The source code is not available to licensees. Customers typically license the product through a perpetual license that entitles them to indefinite use, with annual maintenance fees for support and software upgrades. Examples of this model include databases from Oracle, IBM and Terradata.

Open-Source. The software product and source code are freely available to use. Companies monetize the software product by selling subscriptions and adjacent products with value-added components, such as management tools and support services. Examples of this model include MongoDB (by MongoDB, Inc.) and Hadoop (by Cloudera and others).

Cloud Service. The service is hosted in a cloud- based environment outside of customers' data centers and delivered over the public Internet. The predominant business model is metered (i.e., pay-per-use) or subscription-based. Examples of this model include Google App Engine and Amazon Elastic MapReduce.

For many Fortune 1000 companies, regulations and internal policies around data privacy limit their ability to leverage cloud-based solutions. As a result, most Big Data initiatives are driven with technologies deployed on-premise. Most of the Big Data pioneers are web companies that developed powerful software and hardware, which they open-sourced to the larger community. Accordingly, most of the software used for Big Data projects is open-source.

4.3 Community

In these early days of Big Data, there is an opportunity to learn from others. Organizations should consider how many other initiatives are being pursued using the same technologies and with similar objectives. To understand a given technology's adoption, organiza- tions should consider the following:

- The number of users
- The prevalence of local, community-organized events
- The health and activity of online forums such as Google Groups and StackOverflow
- The availability of conferences, how frequently they occur and whether they are well-attended

4.4 Developer Appeal

The market for Big Data talent is tight. The nation's top engineers and data scientists often flock to companies like Google and Facebook, which are known havens for the brightest minds and places where one will be exposed to leading edge technology. If enterprises want to compete for this talent, they have to offer more than money.

By offering developers the opportunity to work on tough problems, and by using a technology that has strong developer interest, a vibrant community, and an auspicious long-term future, organizations can attract the brightest minds. They can also increase the pool of candidates by choosing technologies that are easy to learn and use — which are often the ones that appeal most to developers. Furthermore, technologies that have strong developer appeal tend to make for more productive teams who feel they are empowered by their tools rather than encumbered by poorly-designed, legacy technology. Productive developer teams reduce time to market for new initiatives and reduce development costs, as well.

4.5 Agility

Organizations should use Big Data products that enable them to be agile. They will benefit from technologies that get out of the way and allow teams to focus on what they can do with their data, rather than how to deploy new applications and

infrastructure. This will make it easy to explore a variety of paths and hypotheses for extracting value from the data and to iterate quickly in response to changing business needs.

In this context, agility comprises three primary components:

Ease of Use. A technology that is easy for developers to learn and understand -- either because of the way it's architected, the availability of tools and information, or both -- will enable teams to get Big Data projects started and to realize value quickly. Technologies with steep learning curves and fewer resources to support education will make for a longer road to project execution.

Technological Flexibility. The product should make it relatively easy to change requirements on the fly—such as how data is modeled, which data is used, where data is pulled from and how it gets processed as teams develop new findings and adapt to internal and external needs. Dynamic data models (also known as schemas) and scalability are capabilities to seek out.

Licensing Freedom. Open-source products are typically easier to adopt, as teams can get started quickly with free community versions of the software. They are also usually easier to scale from a licensing standpoint, as teams can buy more licenses as requirements increase. By contrast, in many cases proprietary software vendors require large, upfront license purchases, which make it harder for teams to get moving quickly and to scale in the future.

MongoDB's ease of use, dynamic data model and open- source licensing model make it the most agile online Big Data solution available.

4.6 General Purpose vs. Niche Solutions

Organizations are constantly trying to standardize on fewer technologies to reduce complexity, to improve their competency in the selected tools and to make their vendor relationships more productive. Organizations should consider whether adopting a Big Data technology helps them address a single initiative or many initiatives. If the technology is general purpose, the expertise, infrastructure, skills, integrations and other investments of the initial project can be amortized across many projects. Organizations may find that a niche technology may be a better fit for a single project, but that a more general purpose tool is the better option for the organization as a whole.

V. APPLICATIONS OF BIG DATA

5.1 Science and research

- When the Sloan Digital Sky Survey (SDSS) began collecting astronomical data in 2000, it amassed more in its first few weeks than all data collected in the history of astronomy. Continuing at a rate of about 200 GB per night, SDSS has amassed more than 140 terabytes of information. When the Large Synoptic Survey Telescope, successor to SDSS, comes online in 2016 it is anticipated to acquire that amount of data every five days.
- Decoding the human genome originally took 10 years to process, now it can be achieved in less than a week: the DNA sequencers have divided the sequencing cost by 10,000 in the last ten years, which is 100 times faster than the reduction in cost predicted by Moore's Law.
- The NASA Center for Climate Simulation (NCCS) stores 32 petabytes of climate observations and simulations on the Discover supercomputing cluster.[26]

5.2 Government

- In 2012, the Obama administration announced the Big Data Research and Development Initiative, which explored how big data could be used to address important problems faced by the government. The initiative was composed of 84 different big data programs spread across six departments.
- Big data analysis played a large role in Barack Obama's successful 2012 re-election campaign.
- The United States Federal Government owns six of the ten most powerful supercomputers in the world.
- The Utah Data Center is a data center currently being constructed by the United States National Security Agency. When finished, the facility will be able to handle a large amount of information collected by the NSA over the Internet. The exact amount of storage space is unknown, but more recent sources claim it will be on the order of a few Exabytes.

5.3 Private sector

- eBay.com uses two data warehouses at 7.5 petabytes and 40PB as well as a 40PB Hadoop cluster for search, consumer recommendations, and merchandising. Inside eBay's 90PB data warehouse
- Amazon.com handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. The core technology that keeps Amazon running is Linux-based and as of 2005 they had the world's three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB.
- Walmart handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes (2560 terabytes) of data – the equivalent of 167 times the information contained in all the books in the US Library of Congress.
- Facebook handles 50 billion photos from its user base.
- FICO Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide.
- The volume of business data worldwide, across all companies, doubles every 1.2 years, according to estimates.
- Windermere Real Estate uses anonymous GPS signals from nearly 100 million drivers to help new home buyers determine their typical drive times to and from work throughout various times of the day.

5.4 International development

Research on the effective usage of information and communication technologies for development (also known as ICT4D) suggests that big data technology can make important contributions but also present unique challenges toInternational development. Advancements in big data analysis offer cost-effective opportunities to improve decision-making in critical development areas such as health care, employment, economic productivity, crime, security, and natural disaster and resource management. However, longstanding challenges for developing regions such as inadequate technological infrastructure and economic and human resource scarcity exacerbate existing concerns with big data such as privacy, imperfect methodology, and interoperability issues.

5.5 Market

"Big Data" has increased the demand of information management specialists in that Software AG, Oracle Corporation, IBM, Microsoft, SAP, EMC, HP and Dell have spent more than $15 billion on software firms only specializing in data management and analytics. In 2010, this industry on its own was worth more than $100 billion and was growing at almost 10 percent a year: about twice as fast as the software business as a whole.

Developed economies make increasing use of data-intensive technologies. There are 4.6 billion mobile-phone subscriptions worldwide and there are between 1 billion and 2 billion people accessing the internet. Between 1990 and 2005, more than 1 billion people worldwide entered the middle class which means more and more people who gain money will become more literate which in turn leads to information growth. The world's effective capacity to exchange information through telecommunication networks was 281 petabytes in 1986, 471 petabytes in 1993, 2.2 exabytes in 2000, 65 exabytes in 2007 and it is predicted that the amount of traffic flowing over the internet will reach 667 exabytes annually by 2014. It is estimated that one third of the globally stored information is in the form of alphanumeric text and still image data, which is the format most useful for most big data applications. This also shows the potential of yet unused data (i.e. in the form of video and audio content).

VI. BIG DATA SOFTWARE

- Hadoop - Apache Foundation
- MongoDB - MongoDB, Inc
- Splunk - SplunkInc
- HP Vertica - HP

VII. CONCLUSION

Big data isn't just hype – and it's much more than a buzz phrase. Today, companies across industries are finding they not only need to manage increasingly large data volumes in their real-time systems, but also analyze that information so they can make the right decisions fast to compete effectively in the market. The Big Data is the future of Database handling.

REFERENCES

[1] "Data, data everywhere". The Economist. 25 February 2010. Retrieved 9 December 2012

[2] Francis, Matthew (2012-04-02). "Future telescope array drives development of exabyte processing". Retrieved 2012-10-24.

[3] "Community cleverness required". Nature 455 (7209): 1. 4 September 2008. .

[4] "Sandia sees data management challenges spiral". HPC Projects. 4 August 2009.

[5] Reichman, O.J.; Jones, M.B.; Schildhauer, M.P. (2011). "Challenges and Opportunities of Open Data in Ecology".Science 331 (6018): 703–5. doi:10.1126/science.1197962.

[6] "Data Crush by Christopher Surdak". Retrieved 14 February 2014.

[7] Hellerstein, Joe (9 November 2008). "Parallel Programming in the Age of Big Data". Gigaom Blog.

[8] Segaran, Toby; Hammerbacher, Jeff (2009). Beautiful Data: The Stories Behind Elegant Data Solutions. O'Reilly Media. p. 257. ISBN 978-0-596-15711-1.

[9] "IBM What is big data? — Bringing big data to the enterprise". www.ibm.com. Retrieved 2013-08-26.

[10] Oracle and FSN, "Mastering Big Data: CFO Strategies to Transform Insight into Opportunity", December 2012

[11] Jacobs, A. (6 July 2009). "The Pathologies of Big Data". ACMQueue.

[12] Magoulas, Roger; Lorica, Ben (February 2009). "Introduction to Big Data". Release 2.0 (Sebastopol CA: O'Reilly Media) (11).

[13] Snijders, C., Matzat, U., &Reips, U.-D. (2012). 'Big Data': Big gaps of knowledge in the field of Internet.International Journal of Internet Science, 7, 1-5. http://www.ijis.net/ijis7_1/ijis7_1_editorial.html

[14] Laney, Douglas. "3D Data Management: Controlling Data Volume, Velocity and Variety". Gartner. Retrieved 6 February 2001.

[15] Beyer, Mark. "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data". Gartner. Archived from the original on 10 July 2011. Retrieved 13 July 2011.

[16] Laney, Douglas. "The Importance of 'Big Data': A Definition". Gartner. Retrieved 21 June 2012.

[17] "What is Big Data?". Villanova University.
[18] Billings S.A. "Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains". Wiley, 2013
[19] Delort P., Big data Paris 2013 http://www.andsi.fr/tag/dsi-big-data/
[20] Big Data Solution Offering". MIKE2.0. Retrieved 8 Dec 2013
[21] Bertolucci, Jeff "Hadoop: From Experiment To Leading Big Data Platform", "Information Week", 2013. Retrieved on 14 November 2013

ABOUT AUTHOR

Dr. Maulik N. Pandya
He is Head of Department and Assistant Professor in M.Sc. IT Department at Shri. A. N. Patel P. G. Institute, Anand. He completed his Ph.D. in mobile communication technology. He has registered a patent for innovative technology of money transaction over small computing devices. His research interest includes JAVA Technologies, Database Management.

Mr. Kalpit G. Soni
He is an Assistant Professor in M.Sc. IT Department at Shri. A. N. Patel P. G. Institute, Anand. He is pursuing his Ph.D. on Cloud computing based innovative research. His research interest includes Cloud Computing, Networking and Data Warehousing.

