Design Challenges in Big Data

K . Swathi
Department of Master Of Computer Applications
Prasad. V. Potluri Siddhartha Institute of Technology
Vijayawada, India

V. V. Subrahmanyam
Indira Gandhi National Open University
(IGNOU)
New Delhi, India

Abstract - Big Data is a broad term for large and complex datasets where traditional data processing applications are inadequate. The integration of this huge data sets is quite complex. There are several issues and challenges one can face during this integration such as analysis, data curation, capture, processing, sharing, search, visualization, information privacy, management and storage. The core elements of the big data platform are to handle the data in new ways as compared to the traditional relational database. Accuracy in managing big data will lead to more confident decision making. In this paper we analyze the issues and design challenges in Big Data.

Keywords—Big data, Issues, Design Challenges

I. INTRODUCTION

Big data was been in the formal literature of Database Management Systems since 1990s. It refers not only to specific, large datasets, but also to data collections that consolidate many datasets from multiple sources, and even to the techniques used to manage and analyze the data.

Big Data has the potential to revolutionize much more than just research. Google’s work on Google File System and MapReduce, and subsequent open source work on systems like Hadoop, have led to arguably the most extensive development and adoption of Big Data technologies, led by companies focused on the Web, such as Facebook, LinkedIn, Microsoft, Quantcast, and Twitter. They have become the indispensable foundation for applications ranging from Web search to content recommendation and compu-tational advertising. There have been persuasive cases made for the value of Big Data for healthcare (through home-based continuous monitoring and through integration across providers), urban planning (through fusion of high-fidelity geographical data), intelligent transportation (through analysis and visualization of live and detailed road network data), environmental modeling (through sensor networks ubiquitously collecting data), energy saving (through unveiling patterns of use), smart materials (through the new materials genome initiative), machine translation between natural languages (through analysis of large corpora), education (particularly with online courses), computational social sciences (a new methodology growing fast in popularity because of the dramatically lowered cost of obtaining data), systemic risk analysis in finance (through integrated analysis of a web of contracts to find dependencies between financial entities), homeland security (through analysis of social networks and financial transactions of possible terrorists), computer security (through analysis of logged events, known as Security In-formation and Event Management, or SIEM), and so on.

In the scientific domain, by revealing the genetic origin of illnesses, such as mutations related to cancer, the Human Genome Project, completed in 2003, is one project that’s a testament to the promises of big data. Consequently, researchers are embarking on two major efforts, the Human Brain Project and the US BRAIN Initiative, in a quest to construct a supercomputer simulation of the brain’s inner workings, in addition to mapping the activity of about 100 billion neurons in the hope of unlocking answers to Alzheimer’s and Parkinson’s. In August 2010, the White House, OMB, and OSTP proclaimed that Big Data is a national challenge and priority along with healthcare and national security [1]. The National Science Foundation, the National Institutes of Health, the U.S. Geological Survey, the Departments of Defense and Energy, and the Defense Advanced Research Projects Agency announced a joint R&D initiative in March 2012 that will invest more than $200 million to develop new big data tools and techniques [14].

A. *Big Data Market Size*

The Big Data technology and services market represents a fast-growing multibillion-dollar worldwide opportunity. In fact, a recent Wikibon forecast shows that the Big Data technology and services market grown from $7.3 billion in 2011 to $38.4 billion in 2015 and will grow to $50.1 billion in 2017 as shown in Fig.1. This is about six times the growth rate of the overall information technology market.

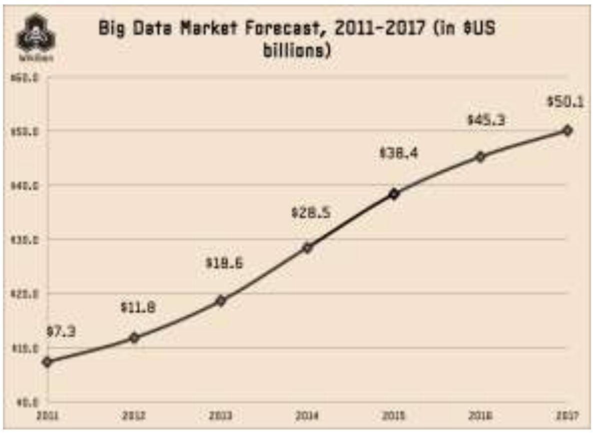

Fig. 1. Big Data Market Forecast (Source: Wikibon)

In this paper we had identified critical issues associated with data storage, management, and processing and also the Big Data design challenges.

II. ISSUES

There are three fundamental issue areas that need to be addressed in dealing with big data: storage issues, management issues, and processing issues. Each of these represents a large set of technical research problems in its own right.

A. *Storage*

The quantity of data has exploded each time we have invented a new storage medium. What is different about the most recent explosion – due largely to social media – is that there has been no new storage medium. Moreover, data is being created by everyone and everything (e.g., devices, etc) – not just, as heretofore, by professionals such as scientist, journalists, writers, etc.

Current disk technology limits are about 4 terabytes per disk. So, 1 exabyte would require 25,000 disks. Even if an exabyte of data could be processed on a single computer system, it would be unable to directly attach the requisite number of disks. Access to that data would overwhelm current communication networks. Assuming that a 1 gigabyte per second network has an effective sustainable transfer rate of 80%, the sustainable bandwidth is about 100 megabytes. Thus, transferring an exabyte would take about 2800 hours, if we assume that a sustained transfer could be maintained. It would take longer to transmit the data from a collection or storage point to a processing point than it would to actually process it!

Two solutions manifest themselves. First, process the data “in place” and transmit only the resulting information. In other words, “bring the code to the data”, Vs. the traditional method of “bring the data to the code.” Second, perform triage on the data and transmit only that data which is critical to downstream analysis. In either case, integrity and provenance metadata should be transmitted along with the actual data.

B. *Management*

Management will, perhaps, be the most difficult problem to address with big data. Unlike the collection of data by manual methods, where rigorous protocols are often followed in order to ensure accuracy and validity, digital data collection is much more relaxed. The richness of digital data representation prohibits a bespoke methodology for data collection. Data qualification often focuses more on missing data or outliers than trying to validate every item. Data is often very fine-grained such as clickstream or metering data. Given the volume, it is impractical to validate every data item: new approaches to data qualification and validation are needed.

The sources of this data are varied - both temporally and spatially, by format, and by method of collection. Individuals contribute digital data in mediums comfortable to them: documents, drawings, pictures, sound and video recordings, models, software behaviors, user interface designs, etc – with or without adequate metadata describing what, when, where, who, why and how it was collected and its provenance. Yet, all this data is readily available for inspection and analysis.

Going forward, data and information provenance will become a critical issue. JASON has noted [10] that “there is no universally accepted way to store raw data, … reduced data, and … the code and parameter choices that produced the data.” Further, they note:

“We are unaware of any robust, open source, platform-independent solution to this problem.” As far as we know, this remains true today. To summarize, there is no perfect big data management solution yet. This represents an important gap in the research literature on big data that needs to be filled.

C. *Processing*

Assume that an exabyte of data needs to be processed in its entirety. For simplicity, assume the data is chunked into blocks of 8 words, so 1 exabyte = 1K petabytes. Assuming a processor expends 100 instructions on one block at 5 gigahertz, the time required for end-to-end processing would be 20 nanoseconds. To process 1K petabytes would require a total end-to-end processing time of roughly 635 years. Thus, effective processing of exabytes of data will require extensive parallel processing and new analytics algorithms in order to provide timely and actionable information.

III. Big Data Design Challenges

There are numerous challenges requiring long-term research to working with big data. Stonebreaker and Hong [18] argue that the design for the systems and components that work with big data will require an understanding of both the needs of the users and the technologies that can be used to solve the problem being investigated – i.e., not all big data and its requirements are the same. In this instance, since the data that is newly created (envisioned and collected), is neither truly known nor well understood, designers will need to consider interfaces, graphics, and icons; application organization; and conceptual models, metaphors, and functionality. Because the end users will not often be the system designers, this presents an additional design challenge.

There are unknown challenges that will arise with each increase in scale and development of new analytics. Some of these challenges will be intractactable with the tools and techniques at hand. We believe these challenges to be just “over the horizon” with the next jump to zettabyte-size data sets.

A. *Data Input and Output Processes*

A major issue raised in big data design is the output process. Jacobs [9] summarized the issue very succinctly – “…its easier to get the data in than out.” His work shows that data entry and storage can be handled with processes currently used for relational databases. But, the tools designed for transaction processing that add, update, search for, and retrieve small to large amounts of data are not capable of extracting the huge volumes and cannot be executed in seconds or a few minutes.

How to access very large quantities of semi- or unstructured data, and how to utilize as yet unknown tool designs is not known. It is clear the problem may neither be solved by dimensional modeling and online analytical processing (OLAP), which may be slow or have limited functionality, nor by simply reading all the data into memory. Technical considerations that must be factored into the design include the ratio of the speed of sequential disk reads to the speed of random memory access. The current technology shows that random access to memory is 150,000 times slower than sequential access. Joined tables, an assumed requirement of associating large volumes of disparate but somehow related data, perhaps by observations over time alone, will come at further huge performance costs.

B. *Quality Vs Quantity*

An emerging challenge for big data users is “quantity vs. quality”. As users acquire and have access to more data (quantity), they often want even more. For some users, the acquisition of data has become an addiction. Perhaps, because they believe that with enough data, they will be able to perfectly explain whatever phenomenon they are interested in.

Conversely, a big data user may focus on quality which means not having all the data available, but having a (very) large quantity of high quality data that can be used to draw precise and high-valued conclusions.

- How do we decide which data is irrelevant versus selecting the most relevant data?
- How do we ensure that all data of a given type is reliable and accurate? Or, maybe just approximately accurate?
- How much data is enough to make an estimate or prediction of the specific probability and accuracy of a given event?
- How do we assess the “value” of data in decision making? Is more necessarily better?

Another way of looking at this problem is, what is the level of precision that the user requires? For example, trend analysis may not require the precision that traditional DB systems provide, but which requires massive processing in a Big Data environment. This problem also manifests itself in the “speed versus scale” challenge discussed below.

C. *Data Growth versus Data Expansion*

Most organizations expect their data to grow over their lifetime as the organization increases its services, its business and business partners and clients, its projects and facilities, and its employees. Few businesses adequately consider data expansion, which occurs when the data records grow in richness, when they evolve over time with additional information as new techniques, processes and information demands evolve. Most data is time-varying – the same data items can be collected over and over with different values based on a timestamp. Much of this data is required for retrospective analysis – particularly that which is used in estimative and predictive analytics.

D. *Speed Vs Scale*

As the volume of data grows, the “big” may morph from the scale of the data warehouse to the amount of data that can be processed in a given interval, say 24 hours. Gaining insight into the problem being analyzed is often more important than processing all of the data. Time-to-information is critical when one considers (near) real-time processes that generate near-continuous data, such as radio frequency identifiers (RFIDs – used to read electronic data wirelessly, such as with EZPass tags) and other types of sensors. An organization must determine how much data is enough in setting its processing interval because this will drive the processing system architecture, the characteristics of the computational engines, and the algorithm structure and implementation.

That said, another major challenge is *data dissemination*. The bottleneck is the communications middleware. While communication hardware speeds are increasing with new technologies, message handling speeds are decreasing only slowly. The computation versus communication dichotomy has not been fully resolved by large data store systems such as HDFS or Accumulo for exabyte-sized data sets.

E. *Structured Vs Unstructured Data*

Translation between structured data with well-defined data definitions (often in tables) as stored in relational databases, and unstructured data (e.g., free text, graphics, multi-media, etc.) suitable for analytics can impede end-to-end processing performance. The emergence of non-relational, distributed, analytics-oriented databases such as NoSQL, MongoDB, SciDB and linked data DBs provides dynamic flexibility in representing and organizing information.

Unlike a data set, a *data source* has no beginning and no end. One begins collecting and continues to do so until one has enough data or runs out of patience or money or both. The data streams in with varied speed, frequency, volume, and complexity. The data stream may dynamically change in two ways: (1) the data formats change, necessitating changes in the way the analytics process the data, or (2) the data itself changes necessitating different analytics to process it. A complicating factor is the implicit assumption that the data streams are well-behaved and that the data arrive more or less in order. In reality, data streams are not so well-behaved and often experience disruptions and mixed-in data, possibly unrelated, to the primary data of interest. There is a need to rethink data stream processing to, perhaps, emphasize continuous analytics over discontinuous and distributed data streams.

F. *Data Ownership*

Data ownership presents a critical and ongoing challenge, particularly in the social media arena. While petabytes of social media data reside on the servers of Facebook, MySpace, and Twitter, it is not really owned by them (although they may contend so because of residency). Certainly, the "owners" of the pages or accounts believe they own the data. This dichotomy will have to be resolved in court. Kaisler, Money and Cohen [12] addressed this issue with respect to cloud computing as well as other legal aspects that we will not delve into here.

With ownership comes a modicum of responsibility for ensuring its accuracy. This may not be required of individuals, but almost certainly is so of businesses and public organizations. However, enforcement of such an assumption (much less a policy) is extremely difficult. Simple user agreements will not suffice since no social media purveyor has the resources to check every data item on its servers.

With the advent of numerous social media sites, there is a trend in big data analytics towards mixing of first-party, reasonably verified data, with public and third-part external data, which has largely not been validated and verified by any formal methodology. The addition of unverified data: compromises the fidelity of the dataset; may introduce non-relevant entities; and may lead to erroneous linkages among entities. As a result, the accuracy of conclusions drawn from processing this mixed data varies widely.

- When does the validity of (publicly available) data expire?
- If data validity is expired, should the data be removed from public-facing websites or data sets?
- Where and how do we archive expired data? Should we archive it?
- Who has responsibility for the fidelity and accuracy of the data? Or, it a case of user beware?

G. *Compliance and Security*

In certain domains, such as social media and health information, as more data is accumulated about individuals, there is a fear that certain organizations will know too much about individuals. For example, data collected in electronic health record systems in accordance with HIPAA/HITECH provisions is already raising concerns about violations of one's privacy. Developing algorithms that randomize personal data among a large data set enough to ensure privacy is a key research problem.

Perhaps the biggest threat to personal security is the unregulated accumulation of data by numerous social media companies. This data represents a severe security concern, especially when many individuals so willingly surrender such information. Questions of accuracy, dissemination, expiration, and access abound. For example, the State of Maryland became the first state to prohibit by law employers asking for Facebook and other social media passwords during employment interviews and afterwards.

International Data Corporation (IDC) coined the term "digital shadow" to reflect the amount of data concerning an individual which has been collected, organized, and perhaps analyzed, to form an aggregate "picture" of the individual. It is the information about you that is much greater than the information you create and/or release about yourself. A key problem is how much of this information – either original or derived – do we want to remain private?

Clearly, some big data must be secured with respect to privacy and security laws and regulations. IDC suggested five levels of increasing security [8]: privacy, compliance-driven, custodial, confidential, and lockdown. Further research is required to clearly define these security levels and map them against both current law and current analytics. For example, in Facebook, one can restrict pages to 'friends'. But, if Facebook runs an analytic over its databases to extract all the friend's linkages in an expanding graph, at what security level should that analytic operate? e.g., how many of an individual's friends should be revealed by such an analytic at a given level if the individual (has the ability to and) has marked those friends at certain security levels?

H. *The Value of "Some Data" versus "All Data"*

Not all data is created equal; some data is more valuable than other data – temporally, spatially, contextually, etc. Previously, storage limitations required data filtering and deciding what data to keep. Historically, we converted what we could and threw the rest away (figuratively, and often, literally).

The concept of "quantitative qualitative computation" suggests that we need new mechanisms for converting latent, unstructured text, image or audio information into numerical indicators to make them computationally tractable. With big data and our enhanced analytical capabilities, the trend is towards keeping everything with the assumption that analytical significance will emerge over time. However, at any point in time the amount of data we need to analyze for specific decisions represents only a very small fraction of all the data available in a data source and most data will go un-analyzed.

- For a given problem domain, what is the minimum data volume required for descriptive, estimative, predictive and prescriptive analytics and decision modeling with a specified accuracy?
- For a given data velocity, how do we update our data volume to ensure continued accuracy and support (near) real-time processing?
- For a given problem domain, what constitutes an analytic science for non-numerical data?
- "What if we know everything?" – What do we do next?

I. Distributed Data and Distributed Processing

The allure of hardware replication and system expandability as represented by cloud computing along with the MapReduce and Message Passing Interface (MPI) parallel programming systems offers one solution to these challenges by utilizing a distributed approach. Even with this approach, significant performance degradation can still occur because of the need for communication between the nodes. An Open Research question is which big data problems are “Map Reducible”?

J. Troubles of Upscaling

The most typical feature of big data is its dramatic ability to grow. And one of the most serious challenges of big data is associated exactly with this. The design of the solution may be thought through and adjusted to upscaling with no extra efforts. The real problem isn’t the actual process of introducing new processing and storing capacities, however, it lies in the complexity of scaling up so, that the system’s performance doesn’t decline and would stay within the budget.

K. Real-time can be Complex

Lot of data keeps updating every second, and organizations need to be aware of that. This comes under the Volume and Velocity characteristics of Big Data. For instance, if a retail company wants to analyze customer behavior, real-time data from their current purchases can help. There are Data Analysis tools available for the same – Veracity and Velocity. They come with ETL engines, visualization, computation engines, frameworks and other necessary inputs.

It is important for businesses to keep themselves updated with this data, along with the “stagnant” and always available data. This will help build better insights and enhance decision-making capabilities.

However, not all organizations are able to keep up with real-time data, as they are not updated with the evolving nature of the tools and technologies needed. Currently, there are a few reliable tools, though many still lack the necessary sophistication.

IV. 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 issues and 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.

REFERENCES

[1] American Institute of Physics (AIP). 2010. College Park, MD, (http://www.aip.org/fyi/2010/)

[2] Ayres, I. 2007. *Supercrunchers*, Bantam Books, New York, NY

[3] Boyd, D. and K. Craford. 2011. “Six Provocations for Big Data”, Oxford Internet Institute’s “A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society”

[4] The Economist. 2010. “Data, Data Everywhere”, (online edition, February 28) http://www.economist.com/node/15557443

[5] Felten, E. 2010. “Needle in a Haystack Problems”, https://freedom-to-tinker.com/blog/felten/needle-haystack-problems/

[6] Fox, B. 2011. “Leveraging Big Data for Big Impact”, Health Management Technology, http://www.healthmgttech.com/

[7] Freeman, K. 2011. http://en.wikipedia.org/wiki/File:Kencf0618Facebook Network.jpg

[8] Gantz, J. and E. Reinsel. 2011. “Extracting Value from Chaos”, IDC’s Digital Universe Study, sponsored by EMC

[9] Jacobs, A. 2009. “Pathologies of Big Data”, *Communications of the ACM*, 52(8):36-44

[10] JASON. 2008. “Data Analysis Challenges”, The Mitre Corporation, McLean, VA, JSR-08-142

[11] Kaisler, S. 2012. “Advanced Analytics”, CATALYST Technical Report, i_SW Corporation, Arlington, VA

[12] Kaisler, S., W. Money, and S. J. Cohen. 2012. “A Decision Framework for Cloud Computing”, *45th Hawaii International Conference on System Sciences*, Grand Wailea, Maui, HI, Jan 4-7, 2012

[13] Kang, U. 2012. “Mining Tera-scale Graphs with MapReduce: Theory, Engineering, and Discoveries”, PhD. Thesis, Computer Science, Carnegie-Mellon University, Pittsburgh, PA

[15] Mervis, J. 2012. “Agencies Rally to Tackle Big Data”, Science, 336(4):22, June 6, 2012 Popp, R., S. Kaisler, et al. 2006. “Assessing Nation-State Fragility and Instability”, *IEEE Aerospace Conference*, 2006, Big Sky, MT

[16] Ritchey, T. 2005. "Wicked Problems: Structuring Social Messes with Morphological Analysis", Swedish Morphological Society, http://www.swemorph.com/wp.html

[17] Rittel, H. and M. Webber. *1973. “*Dilemmas in a General theory of Planning*”, in Policy Sciences*, Vol. 4, Elsevier Scientific, Amsterdam, the Netherlands, pp. 155-169

[18] Stonebraker, M. and J. Hong. 2012. “Researchers' Big Data Crisis; Understanding Design and Functionality”, Communications of the ACM, 55(2):10-11

[19] Taleb, N. 2010. *The Black Swan: The Impact of the Highly Improbable*, Random House, New York, NY

[20] Big. James Manyika,Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and AngelaHung Byers. Big Data: The next frontier for innovation, competition, and productivity McKinsey Global Institute. May 2011.

[21] Daniel Fasel and Andreas Meier, editors. Big Data. Springer Publisher, Heidelberg, Germany, 2013.

[22]Lisa Kart. Market Trends: Big Data Opportunities in Vertical Industries. Technical report, Gartner, 2012.

[23]IBM. IBM Research Dublin. http://www.research.ibm.com/labs/ireland/, March 2014.

[24]Wikibon Report, 2011.