Big Data: Challenges and Opportunities

Anuranjan Misra, Anshul Sharma, Preeti Gulia, Akanksha Bana

Abstract- Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This paper presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.

Index Terms—BigData, definition of big data, meseure of big data, Challenges in big data

I. INTRODUCTION

The definition of big data refers to groups of data that are so large and unwieldy that regular database management tools have difficulty capturing, storing, sharing and managing the information. Big Data refers to the massive amounts of data that collect over time that are difficult to analyze and handle using common database management tools. Big Data includes business transactions, e-mail messages, photos, surveillance videos and activity logs (see machine-generated data). Scientific data from sensors can reach mammoth proportions over time, and Big Data also includes unstructured text posted on the Web, such as blogs and social media. Big data refers to a process that is used when traditional data mining and handling techniques cannot uncover the insights and meaning of the underlying data. Data that is unstructured or time sensitive or simply very large cannot be processed by relational database engines. This type of data requires a different processing approach called big data, which uses massive parallelism on readily-available hardware.

II. THE SEVEN STEPS OF BIG DATA DELIVERY

- Collect: Data is collected from the data sources and distributed across multiple nodes – often a grid – each of which processes a subset of data in parallel.
  Process: The system then uses that same high-powered parallelism to perform fast computations against the data on each node. Next, the nodes reduce the resulting data findings into more consumable data sets to be used by either a human being (in the case of analytics) or machine (in the case of large-scale interpretation of results).
- Manage: Often the big data being processed is heterogeneous, originating from different transactional systems. Nearly all of that data needs to be understood, defined, annotated, cleansed and audited for security purposes.
- Measure: Companies will often measure the rate at which data can be integrated with other customer behaviors or records, and whether the rate of integration or correction is increasing over time. Business requirements should determine the type of measurement and the ongoing tracking.
- Consume: The resulting use of the data should fit in with the original requirement for the processing. For instance, if bringing in a few hundred terabytes of social media interactions demonstrates whether and how social media data delivers additional product purchases, then there should be rules for how social media data is accessed and updated. This is equally important for machine-to-machine data access.
- Store: As the "data-as-a-service" trend takes shape, increasingly the data stays in a single location, while the programs that access it move around. Whether the data is stored for short-term batch processing or longer-term retention, storage solutions should be deliberately addressed.
- Govern: Data governance encompasses the policies and oversight of data from a business perspective. As defined, data governance applies to each of the six preceding stages of big data delivery.

III. CHALLENGES

Big data presents a number of challenges relating to its complexity. One challenge is how we can understand and use big data when it comes in an unstructured format, such as text or video. Another challenge is how we can capture the most important data as it happens and deliver that to the right people in real-time. A third challenge is how we can store the data, and how we can analyze and understand it given its size and our computational capacity. And there are numerous other challenges, from privacy and security to access and deployment.

IV. CHALLENGES IN IT MANAGEMENT

1. Determining what data (both structured and unstructured , and internal & external) to use for different business decision
2. Being able to handle large volume, velocity, variety of big data.
3. Getting business units to share information across organization.
4. Building high levels of trust between the data scientist who present insight BIS date.

Manuscript Received on July 2014.

Dr. Anuranjan Misra, Prof. & Dean at Bhagwant Institute of Technology, Ghaziabad, India.

Ms. Anshul Sharma, M.Tech Scholar of Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India.

Dr. Preeti Gulia, Asst. Prof., Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India.

Ms. Akanksha Bana, Asst. Prof. in Computer Science & Engineering Department at Bhagwant Institute of Technology, Ghaziabad, India.

5. Getting top management in the company to approve investment in Big Data and its related investment.

V. CONCLUSION & OPPURTUNITIES

Since the Internet's introduction, we've been steadily moving from text-based communications to richer data that include images, videos, and interactive maps as well as associated metadata such as geolocation information and time and date stamps. Twenty years ago, ISDN lines couldn't handle much more than basic graphics, but today's high-speed communication networks enable the transmission of storage-intensive data types. For instance, Smartphone users can take high-quality photographs and videos and upload them directly to social networking sites via Wi-Fi and 3G or 4G cellular networks. We've also been steadily increasing the amount of data captured in bidirectional interactions, both people-to-machine and machine-to-machine, by using telematics and telemetry devices in systems of systems. Of even greater importance are e-health networks that allow for data merging and sharing of high-resolution images in the form of patient x-rays, CT scans, and MRIs between stakeholders. Advances in data storage and mining technologies make it possible to preserve increasing amounts of data generated directly or indirectly by users and analyze it to yield valuable new insights. For example, companies can study consumer purchasing trends to better target marketing. In addition, near-real-time data from mobile phones could provide detailed characteristics about shoppers that help reveal their complex decision-making processes as they walk through malls. Big data can expose people's hidden behavioral patterns and even shed light on their intentions. More precisely, it can bridge the gap between what people want to do and what they actually do as well as how they interact with others and their environment. This information is useful to government agencies as well as private companies to support decision making in areas ranging from law enforcement to social services to homeland security. It's particularly of interest to applied areas of situational awareness and the anticipatory approaches required for near-real-time discovery.

REFERENCES

1. A. Jacobs, The pathologies of big data, Commun. ACM Vol. 52 (8) (2009) pp. 36{44.
2. R. B. Miller, Response time in man-computer conversational transactions, in: Proceedings of the December 9-11, 1968, fall joint computer conference, part I, AFIPS '68 (Fall, part I), New York, NY, USA, 1968, pp. 267{277.
3. S. C. Seow, User and system response times, in: Designing and Engineering Time: The Psychology of Time Perception in Software, Addison-Wesley Professional, 2008, pp. 33{48.
4. M. de Berg, O. Cheong, M. van Kreveld, M. Overmars, 1-dimensional range searching, in: Computational Geometry: Algorithms and Application 2ed, Springer Berlin Heidelberg, 2008, pp. 96{99.

AUTHORS PROFILE

Dr. Anuranjan Misra, is Professor & Dean at Bhagwant Institute of Technology, Ghaziabad, India. He had authored 30 books, 100 research papers. His books are in many Indian & Foreign Universities Syllabus.

Ms. Anshul Sharma, is M.Tech Scholar of Department of Computer Science & Applications Maharshi Dayanand University, Rohtak, India.

Dr. Preeti Gulia, is an Assistant Professor, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India.

Ms. Akanksha Bana, is an Assistant Professor in Computer Science & Engineering Department at Bhagwant Institute of Technology, Ghaziabad, India.