ISSN: 2393-8528

IJICSE
Open Access

Contents lists available at www.ijicse.in

International Journal of Innovative Computer Science & Engineering

Volume 2 Issue 4; September-October-2015; Page No.16-20

IDENTIFYING THE CORE PHASES IN THE BIG DATA AND THEIR RESEARCH OPPORTUNITIES

Dr. M. Balamurugan[1], P. Mathiazhagan[2]

[1]Associate Professor, Department of Computer Science, Engineering and Applications, Bharathidasan University, Trichy, Tamilnadu, India

[2]Research Scholar, Department of Computer Science, Engineering and Applications, Bharathidasan University, Trichy, Tamilnadu, India

ARTICLE INFO

Received 01Sep. 2015
Accepted 30 Sep. 2015

Corresponding Author:

Dr. M. Balamurugan
Associate Professor, Department of Computer Science, Engineering and Applications, Bharathidasan University, Trichy, Tamilnadu, India

Key words: ***Big Data, Data Process, Data Storage, Visualization.***

ABSTRACT

One perspective is that big data is more and different kinds of data than is easily handled by traditional relational database management systems (RDBMSs). For example, 10 terabytes to be big data, but any numerical definition is likely to change over time as organizations collect, store, and analyze more data. Purpose of this analysis to identify the right challenges research opportunities. There are multiple research work happens currently primitive research work is to explore the growth of big data on the different dimension beyond to identify the core research phase such as storage, processing, and visualizing. For that, this study delivers various research opportunities for big data analytics.

I. INTRODUCTION

Big data is a term that is used to describe data thatis high volume, high velocity, and/or high variety; requires new technologies and techniques to capture, store and analyze it; and is used to enhance decision making, provide insight and discovery, and support and optimizeprocesses.Another useful perspective is to characterize big data as having high volume, high velocity, and high varietythe three Vs [Russom, 2011]:

- High volume: the amount or quantity of data
- High velocity: the rate at which data is created
- High variety: the different types of data

It is important to understand that what is thought to be big data today won't seem so big in the future [Franks, 2012]. Many data sources are currently untappedor at least underutilized. For example, every customer e-mail, customer-service chat, and social media comment may be captured, stored, and analyzed to better understand customers' sentiments. Web browsing data may capture every mouse movement in order to better understand customers' shopping behaviors. Radio frequency identification (RFID) tags may be placed on every single piece of merchandise in order to assess the condition and location of every item. It will show the projected growth of big data.

Big datahas many sources. For example, every mouse click on a web sitecan be captured in Web log files and analyzed in order to better understand shoppers' buying behaviors and to influence their shopping by dynamically recommending products. Social media sources such as Facebook and Twitter generate tremendous amounts of comments and tweets. This data can be captured and analyzed to understand, for example, what people think about new product introductions. Machines, such as smart meters, generate data. These meters continuously stream data about electricity, water, or gas consumption that can be shared with customers and combined with pricing plans to motivate customers to move some of their energy consumption such as for washing clothes to non-peak hours.

There is a tremendous amount of geospatial (e.g., GPS)data, such as that created by cell phones that can be used by applications like Four Square to help you know the locations of friends and to receive offers from nearby stores and restaurants. Image, voice, and audio data can be analyzed for applications such as facial recognition systems in security systems.

II. RELATED WORKS

A. Data Storage

Cloud-based Services: The cloud is now in the mainstream of computing. With the cloud, computing resources are virtualized and offered as a service over the Internet. The potential benefits of the cloud include access to specialized resources, quick deployment, easily expanded capacity, the ability to discontinue a cloud service when it is no longer needed, cost savings and good backup and recovery. These same benefits make the cloud attractive for big data and analytics.Public clouds are offered by thirdparty providers;private clouds are implemented within a company's firewall. A concern about data security is a primary reason that private clouds are sometimes preferred over public clouds. We will discuss public cloudsalthough the same approaches and technologies are used with private clouds.

Cloud services are available as softwareasaservice (SaaS), platformasaservice (PaaS), or infrastructureasaservice (IaaS), depending on what software is provided. Cloud services are all similar in that a company's data is loaded to the cloud, stored, and analyzed, and the results are downloaded to users and applications.With SaaS, the vendor provides the hardware, application software, operating system, and storage. The user uploads data and uses the application software to either develop an application (e.g., reports) or simply process the data using the software (e.g., credit scoring). Many BI and analytics vendors offer cloud services versions of their software, including Cognos, Business Objects, MicroStrategy, and SAS. SaaS is a particularly attractive option for firms that lack the financial or human resources to implement and maintain the software and applications inhouse.

PaaS differs from SaaS in that the vendor does not provide the software for building or running specific applications; this is up to the company. Only the basic platform is provided. The benefits of this approach include not having to maintain the computing infrastructure for applications that are developed; access to a dependable, highly scalable infrastructure; greater agility in developing new applications; and possible cost savings. Examples of PaaS include Oracle Cloud Computing, Microsoft Windows Azure, and Google App Engine.With IaaS, the vendor provides raw computing power and storage; neither operating system nor application software are included. Customers upload an image that includes the application and operating system. Becausethe customer provides the operating system, different ones can be used with different applications. IaaS vendors' offerings include Amazon EC2 (part of the Amazon Web Services offerings), Rackspace, and Google Compute Engine.

Non-Relational (NoSQL) Databases: Relational databases have been a computing mainstay since the 1970s. Data is stored in rows and columns and can be accessed through SQL queries. By way of contrast, non-relationalNoSQLdatabases are relatively new (1998), can store data of any structure, and do not rely on SQL to retrieve data (though some do support SQL and are perhaps better called "not only SQL databases"). Data such as XML, text, audio, video, image, and application specific document files are often stored and retrieved "as is" through keyvalue pairs that use keys to provide links to where files are stored on disk. There are specialized non-SQL databases that are designed for specific kinds of data such as documents and graphsand use their own storage and retrieval methods. Non-relational databases such as Apache Cassandra, MongoDB, and Apache Couchbasetend to be open-source; store data in a scaleout, distributed architecture; and run on lowcost, commodity servers. Hadoop/MapReduce, which is discussed next, is one example of a non-relational database. Because non-relational databases are often relatively new and open source, they are not as well supported as established RDBMS. They also are weaker on security, which can limittheir usefulness for some applicationssuch as financialones.

Hadoop/MapReduceOf all the platforms and approaches to storing and analyzing big data, none is receiving more attention than Hadoop/MapReduce. Its origins trace back to the early 2000s, whencompanies such as Google, Yahoo!, and Facebook needed the ability to store and analyze massive amounts of data from the Internet. Because no commercial solutions were available, these and other companies had to develop their own. Important to the development of Hadoop/MapReduce were Doug Cutting and Mike Cafarella,who were working on an open-source Web search engine project called Nutch when Google published papers on the Google File System (2003) and MapReduce (2004). Impressed with Google's work, Cutting and Cafarella incorporated the concepts intoNutch. Wanting greater opportunities to further his work, Cutting went to work for Yahoo!, which had its own big data projects under way [Harris, 2013].

Apache Hadoopis a software framework for processing large amounts of data across potentially massively parallel clusters of servers. To illustrate, Yahoo has over 42,000 servers in its Hadoop installation. Hadoop is open source and can be downloaded at related official website. The key component of Hadoop is the Hadoop Distributed File System (HDFS), which manages the data spread across the various servers. It is because of HDFS that so many servers can be managed in parallel. HDFS is file based and does not need a data model to store and process data. It can store data of any structure, but is not a RDBMS. HDFS can manage the storage andaccess of any type of data (e.g., Web logs, XML files) as long as the data can be put in a file and copied into HDFS.

The Hadoop infrastructure typically runs MapReduce programs (using a programming or scripting language such as Java, Python, C, R, or Perl) in parallel. MapReduce takes large datasets, extracts and transforms useful data, distributes the data to the various servers where processing occurs, and assembles the results into a smaller, easiertoanalyze file. It does not perform analytics rather, it provides the framework that controls the programs (often written in Java) that perform the analytics. Currently, jobs can be run only in batch, which limits the use of Hadoop/MapReduce for near realtime applications. Although Hadoop and MapReduce are discussed and typically used together, they can be used separately. That is, Hadoop can be used without MapReduce and vice versa.This is a simple processing task that could also be done with SQL and a RDBMS, but provides a good example of Hadoop/MapReduce processing. At the left is a data file with records containing Deer, Bear, River, and Car. The objective is to count the number of times each word occurs. Thefirst step is to split the records and distribute them across the clusters of servers (there are only three in this simple example). These splits are then processed by multiple map programssuch as JavaandR running on the servers. The objective in this example is to group the data by a split based on the words. The MapReduce system then merges the shuffle/sort results for input to the reduce program, which then summarizes the number of times each word occurs. This output can then be input to a data warehouse where it may be combined with other data for analysis or accessed directly by various BI tools (e.g., Tableau, MicroStrategy).

B. Data Processing

Data processing is the act of handling or manipulating data in some fashion.Regardless of the activities involved in it, processing tries to assign meaning todata. Thus, the ultimate goal of processing is to transform data into information.Data processing is the process through which facts and figures are collected,assigned meaning, communicated to others and retained for future use. Hencewe can define data processing as a series of actions or operations that convertsdata into useful information. We use the term 'data processing system' to includethe resources that are used to accomplish the processing of data.

These tools shown below have been designed to assist inthedebugging and the transformation of data.They are useful to clean and refine messy data, and convert it into appropriate formats. Often, large data sets represented in tabular formats contain typos, inaccuracies e.g., dates expressed in different formats, cells with abbreviated/expanded names, encoding errors, blank cells, etc., whose manual correction is unfeasible.These tools accelerate the process that enhances the quality of the information, and makes the data complete and easy to re-use.

DataWrangular

An interactiveweb application for data cleaning and transformation, Wrangler combines directmanipulation of visualized data with automatic inference of relevant data transformation. It enables analysts to repeatedly scan the space of applicable operations and anticipate its effects. It leverages semantic data types (geographical locations, dates, classification codes) to aid validation and type conversion.

Google Refine

A free tool designed with the objective to assist in understanding the structure and quality ofthe data, allowing the correction of certain common errors in data.It supports a wide range of formats: TSV, CSV, *SV, Excel (. xls and xlsx), JSON, XML, RDF, XML, and Google Data documents.The data source can be provided in 4 ways: upload a local file, from a URL importing data from tables in web pages, in XML documents, etc.), paste data from the clipboard, and link a Google Docs document.After treatment of the information, data can beexported in TSV (Tab Separated Values), CSV (comma separated values), and Excel formats, and in HTML table.Google Refine has three key features:

- Data Cleansing: It enables changing cell content and field unification. This action may beperformed manually or assisted by the program (the system cansuggest optimizations).It offers predefined operations such as collapsing consecutive whitespaces in texts, scape/unscape HTML entities, changing letter case, converting text to dates, blanking out cells, among others.

- Data transformation: Transformations through GREL (Google Refine Expression Language) instructions.It enablesthesplittingofcolumns, creating new columns based on values of other columns and combining cells to create new columns among other features.

- Creation of new data fields: New data fields may be created byexternal services to obtain new data from existing data, or using Freebase (free collaborative database) to complement the data.

3. Visualization Tools

Tableau

Tableau is often regarded as the grand master of data visualization software and for good reason. Tableau has a very large customer base of 57000+ accounts across many industries due to its simplicity of use and ability to produce interactive visualizations far beyond those provided by general BI solutions. It is particularly well suited to handling the huge and very fast-changing datasets which are used in Big Data operations, including artificial intelligence and machine learning applications, thanks to integration with a large number of advanced database solutions including Hadoop, Amazon AWS, My SQL, SAP and Teradata. Extensive research and testing has gone into enabling Tableau to create graphics and visualizations as efficiently as possible, and to make them easy for humans to understand.

Qlikview

Qlik with their Qlikview tool is the other major player in this space and Tableau's biggest competitor. The vendor has over 40,000 customer accounts across over 100 countries, and those that use it frequently cite its highly customizable setup and wide feature range as a key advantage. This however can mean that it takes more time to get to grips with and use it to its full potential. In addition to its data visualization capabilities Qlikview offers powerful business intelligence, analytics and enterprise reporting capabilities and I particularly like the clean and clutter-free user interface. Qlikview is commonly used alongside its sister package, Qliksense, which handles data exploration and discovery. There is also a strong community and there are plenty of third-party resources available online to help new users understand how to integrate it in their projects.

Google Chart

This Google Developers tool enables the creation of graphic images as PNG.It is free to use but with some limitations. Initially, its use was limited to 50,000 requests per URL and day, but now this limit stands at 250,000.In order to avoid this limitation, generated images may be stored on an external server running as a cache of images.There isa variety of graph types, offered as JavaScript classes. One advantage with this graphic generation system is that users do not need to install any component in environment or server.

Protovis

A JavaScriptoriented graphics library performing visualizations.It provides developers a large set of components and tools, enabling customization of the displays with direct control.Some of the most relevant features of this library are:

• Unlimited flexibility: It is based on a declarative grammar and datadriven framework.

•Simple graphics settings, based onchaining method.

•Focused on statistical graphics, its development method also enables structured, datadriven visualizations.

•It incorporates some statistical functions for data preparation.The main Protovis' disadvantage is that it is a heavy library (weighing more than 700 Kb), designed for either Intranets or fast connections.

III. CONCLUSION

The main focus of this paper is to give a brief survey of some of the multi-dimensional visualizationtechniques that are used in big data, knowing full well that the techniques are not limited to the ones that have been discussed in this paper as there are much more to this. Big data is a field of data that is fast emerging most especially in the business sector of our economy and lots of research is being done every day.The exact result they want when trying to visualize their data using any of the data visualization techniques. One main reason why this challenge is still persistent is because these techniques are being put to use wrongly, many of our business people still don't know what technique is best to use when they want to carry out a particular task and so they end up choosing the wrong visualization technique for the right data and they eventually end up getting wrong results. The use of the data visualization techniques used in big data could be interesting and at times challenging as well, it all depends on how effective you put it to use but for you to be able to choose the best underlying visualization technique to display your data effectively, you must first of all understand the data want to visualize with its size and cardinality, and their research development.

Reference

1. Hugh J. Watson, Big data Analytics: Concepts, Technologies, and Applications, Communications of the Association for Information Systems, Vol.34, pp. 1247-1268, April 2014.
2. Antoni Artigues, et.al, "Scientific Big Data Visualization: a Coupled Tools Approach", Supercomputing Frontiers and Innovations, Doi: 10.14529 / jsfi140301, 2014.
3. Samuel Ajibade, "An Overview of Big Data Visualization Techniques in Data Mining", International Journal of Computer Science and Information Technology Research, Vol.4, Issue 3, pp. 105-113, ISSN: 2348-1196, 2348-120X, 2014.
4. Nselberg, B. Dimsdale, "Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry; Visualization", 90, San Francisco, CA, 1990, pp 361-370.
5. D. A. Keim, H. Kriegel. "Visualization Techniques for Mining Large Databases: A Comparison;", IEEE Transactions on Knowledge and Data Engineering, Special Issue on Data Mining; Vol. 8, No. 6, December 1996, pp 923-938.
6. M. C. Oliveira, H. Levkowitz. "Visual Data Exploration to Visual Data Mining: A Survey", IEEE

Transaction on Visualization and Computer Graphs 9(3), 2003, 378–394.

7. M. Khan, S. S. Khan. Data and Information Visualization Methods and Interactive Mechanisms: A Survey, International Journal of Computer Applications, 34(1), 2011, pp. 1-14.
8. Robert Redpath, a Comparative Study of Visualization Techniques for Data Mining- a Thesis Submitted to the School of Computer Science and Software Engineering Monash University, 2000.
9. SAS, Data Visualization Techniques: From Basics to Big Data with SAS® Visual Analytics, 2014.
10. S. Vijaykumar, S. G. Saravanakumar, "Future Robotic Memory management", Advances in Digital Image Processing and Information Technology Communications in Computer and Information science, Volume 205, 2011, pp 315-325. ISSN: 1865-0929. DOI: 10.1007/978-3-642-24055-3_32.
11. A. Nancy, "Quality of Service Enhancement Analysis Process in Mobile Ad-Hoc Networks", International Journal of Computer Science and Mobile Applications (IJCSMA), Vol.3, Issue.4, Apr 2015, pg.1-7, ISSN: 2321-8363.