Survey on Big Data by Coordinating Mapreduce to Integrate Variety of Data

N. Monica[1], K. Ramesh Kumar[2]

[1]M. Tech Scholar, Department of Information and Technology, Hindustan University, Chennai, Tamilnadu, India

[2]Associate Professor, Department of Information and Technology, Hindustan University, Chennai, Tamilnadu, India

Abstract: ***The growing impact of Big Data deduces the importance of analyzing vast amount of data with a frequent and rapid rate of growth and change in databases and data warehouses. Due to the different types of data, it requires a specialized merging system to combine variety of data to provide better performance, robustness, flexibility and scalability. Hence, it is important to identify a sophisticated strategy for merging different types of data in a way they provide the best result. This survey study gives an overview of variety of data and analyzed the problem of integrating unstructured data with the traditional structured data. Moreover, tools currently available either analyze the structured data or the unstructured data but not the both. As a consequence, few technologies of Big Data are proposed in survey for integrating variety of data. Present study reviews Big Data technology and other related issue available in the survey.***

Keywords: Big Data, structured data, semi structured data, unstructured data, integration, technology.

1. Introduction

Big data is the process of examining large amounts of data of a variety of types to uncover hidden patterns, unknown correlations and other useful information. Such information can provide competitive advantages over rival organizations and result in business benefits, such as more effective marketing and increased revenue [1]. The primary goal of big data is to help companies make better business decisions by enabling data scientists and other users to analyze huge volumes of transaction data as well as other data sources that may be left untapped by conventional business intelligence programs. These other data sources may include Web server logs and Internet click stream data, social media activity reports, mobile-phone call detail records and information captured by sensors. Some people exclusively associate big data and big data analytics with unstructured data of that sort, but consulting firms like Gartner Inc. and Forrester Research Inc. also consider transactions and other structured data to be valid forms of big data.

Big data can be done with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics and knowledge mining. But the unstructured data sources used for big data analytics may not fit in traditional data warehouses. Furthermore, traditional data warehouses may not be able to handle the processing demands posed by big data.

Potential pitfalls that can trip up organizations on big data analytics initiatives include a lack of internal analytics skills and the high cost of hiring experienced analytics professionals, plus challenges in integrating Hadoop systems and data warehouses, although vendors are starting to offer software connectors between those technologies.
Mainstream definition of big data as the three Vs: volume, velocity and variety.

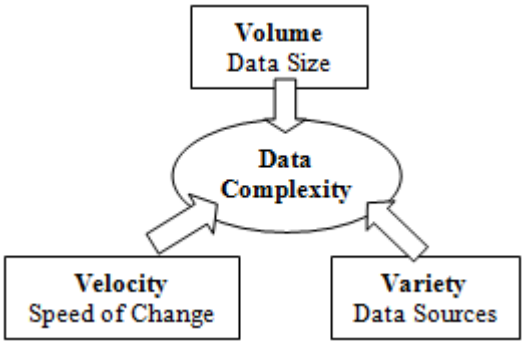

Figure 1: Three V's of big data

1.1 Volume

Big data is always large in volume. It actually doesn't have to be a certain number of petabytes to qualify. If your store of old data and new incoming data has gotten so large that you are having difficulty handling it, that's big data. Remember that it's going to keep getting bigger. Your consultant needs to recommend a scalable solution that can grow with your data. Many factors contribute to the increase in data volume. Transaction-based data stored through the years. Unstructured data streaming in from social media. Increasing amounts of sensor and machine-to-machine data being collected. In the past, excessive data volume was a storage issue. But with decreasing storage costs, other issues emerge, including how to determine relevance within large data volumes and how to use analytics to create value from relevant data.

1.2 Velocity

Velocity or speed refers to how fast the data is coming in, but also to how fast you need to be able to analyze and utilize it [2]. If you have one or more business processes that require real-time data analysis, you have a velocity challenge. Solving this issue might mean expanding your private cloud using a hybrid model that allows bursting for additional compute power as-needed for data analysis. Your consultant may need to offer suggestions for hardware,

software, and business process changes to handle today's high-speed data. Data is streaming in at unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. Reacting quickly enough to deal with data velocity is a challenge for most organizations.

1.3 Variety

Variety points to the number of sources or incoming vectors leading to your databases. That might be embedded sensor data, phone conversations, documents, video uploads or feeds, social media, and much more [3] .Variety in data means variety in databases .Data today comes in all types of formats. Structured, numeric data in traditional databases. Information created from line-of-business applications. Unstructured text documents, email, video, audio, stock ticker data and financial transactions. Managing, merging and governing different varieties of data is something many organizations still grapple with. There are also two additional dimensions when thinking about big data:

1.4 Variability

In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Even more so with unstructured data involved.

1.5 Complexity

Today's data comes from multiple sources. And it is still an undertaking to link, match, cleanse and transform data across systems [4]. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control.

2. Variety of Data

Big data incorporates varieties of data. Data today comes in all types of form traditional databases to hierarchical data stores created by end users and OLAP system, to text documents, email, meter-collected data, video, audio and financial transactions.

2.1 Structured Data

Any data stored in a well defined non propriety system. Data is primarily text based. Typically conforms to ACID property. Data can be designated as structured or unstructured data for classification within an organization. The term structured data refers to *data that is identifiable* because it is organized in a structure [5]. The most common form of structured data records is a database where specific information is stored based on a methodology of columns and rows. Structured data is also searchable by data type within content. Structured data is understood by computers and is also efficiently organized for human readers. In contrast, unstructured data has no identifiable structure. **Example:** Database, Data Warehouses, Electronic Spreadsheets

2.2 Semi structured Data

Semi-structured data is data that is neither raw data, nor typed data in a conventional database system. It is structured data, but it is not organized in a rational model, like a table or an object-based graph. A lot of data found on the Web can be described as semi-structured. Data integration especially makes use of semi-structured data. Any data stored in a system that conforms to some rules and can be propriety. Data is primarily text based which does not have to conform to ACID property. Semi-structured data is a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. Therefore, it is also known as schema less or self-describing structure. In the semi-structured data, the entities belonging to the same class may have different attributes even though they are grouped together, and the attributes' order is not important. Semi-structured data is increasingly occurring since the advent of the Internet where full-text documents and databases are not the only forms of data anymore and different applications need a medium for exchanging information. In object-oriented databases, one often finds semi-structured data. **Example:** Web Posts, Blogs, Wiki pages, Forums, Tweets, Instant Messages

2.3 Unstructured Data

Any data stored in a well defined propriety system. Binary data that conforms mostly to an agreed standard. Techniques such as knowledge mining and text analytics and noisy-text analytics provide different methods to find patterns in, or otherwise interpret, this information. Common techniques for structuring text usually involve manual tagging with metadata or part-of-speech tagging for further text mining-based structuring. Unstructured Information Management Architecture provides a common framework for processing this information to extract meaning and create structured data about the information. Software that creates machine-process able structure exploits the linguistic, auditory, and visual structure inherent in all forms of human communication [6] . Algorithms can infer this inherent structure from text, for instance, by examining word morphology, sentence syntax, and other small- and large-scale patterns. Unstructured information can then be enriched and tagged to address ambiguities and relevancy-based techniques then used to facilitate search and discovery. Examples of "unstructured data" may include books, journals, documents, metadata, health records, audio, video, analog data, files, and unstructured text such as the body of an e-mail message, Web page, or word-processor document. While the main content being conveyed does not have a defined structure, it generally comes packaged in objects that themselves have structure and are thus a mix of structured and unstructured data, but collectively this is still referred to as unstructured data [7]. Since unstructured data commonly occurs in electronic documents, the use of a content or document management system which can categorize entire documents is often preferred over data transfer and manipulation from within the documents. Document management thus provides the means to convey structure onto document collections. Search engines have become popular tools for indexing and

searching through such data, especially text. Example: PowerPoint, Word Documents, Email, PDF, Audio files, Video, Graphics and Multimedia

3. Comparison Between Structured, Semi Structured and Unstructured Data

Table 1: Comparison between structured, semi structured and unstructured data

	Structured Data	Semi Structured Data	Unstructured Data
Technology	Relational database table	XML / RDF	Character and binary data
Transaction Management	Matured transaction, various concurrency techniques	Transaction management adapted from RDBMS not matured	No transaction management, no concurrency
Version Management	Versioning over tuples, rows, tables etc.	Versioning over tuples or graphs is possible	Versioned as a whole
Flexibility	Schema dependent rigorous schema	Flexible, tolerant schema	Very flexible, absence of schema
Scalability	Scaling DB schema is difficult	Schema scaling is simple	Very scalable
Robustness	Very robust	New technology not widely spread	-
Query Performance	Structured query allows complex joins	Queries over anonymous nodes are possible	Only textual queries possible

4. Integrating Structured and Unstructured Data

Big data has evolved from solutions that focus on integrating and processing large volumes of structured and unstructured data to a broad and growing technology that enables many new and innovative data management and analytic solutions. Despite the inevitable confusion caused by this rapid change in data warehouse, it is becoming quite clear that many users are gaining considerable value from big data and that this value is best gained from a hybrid of new and existing data systems. Data processing needs are changing with the ever increasing amounts of both structured and unstructured data. While the processing of structured data typically relies on the well developed field of relational database management systems, MapReduce is a programming model developed to cope with processing immense amounts of unstructured data. MapReduce, however, offers features and advantages that can be exploited to integrate both structured and unstructured data. There are many different techniques for integrating disparate data systems. Three requirements, however, need to be addressed, regardless of the technique used:

1. Copying data between systems
Copying Hadoop MapReduce processing results into a relational database for further analysis.

2. Capturing source data, transforming it, and loading into a target system
Extracting useful business information from Hadoop web log data and loading it into a relational system for analysis.

3. Querying and analyzing data managed by multiple data systems

Running an SQL query that accesses data managed by both a relational database system and Hadoop HDFS. The first two of these tasks can be considered to be data integration, whereas the third task falls more into the category of data manipulation. Data integration can be done in batch or by continuously replicating data between systems. The integration may be done using a data transformation engine, or by generating batch code to do the required work. Data manipulation can be done in batch or interactively. The actual approach used for each of these tasks will depend both on the volume of data involved and performance requirements. From a performance perspective, it is important that the software enabling data integration or manipulation exploit the capabilities provided by the underlying data systems. An example here would be for the software to take advantage of the parallel computing processing capabilities of both a relational database system and Hadoop. Another example would be to use optimized MR code to access Hadoop HDFS data, rather than generating HIVEQL or Pig statements to do the job, which may lead to less efficient MR processing. For certain types of Hadoop data retrieval, however, direct access to the native HDFS file system will provide better performance.

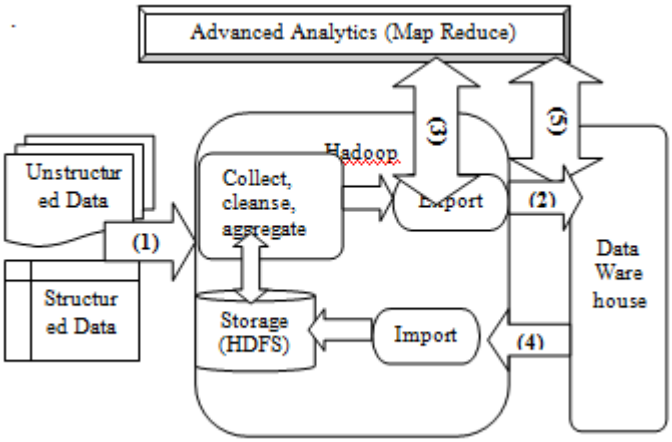

Figure 2: Integrated Structured and Unstructured Data

5. Challenges of Big Data

Many organizations are concerned that the amount of amassed data is becoming so large that it is difficult to find the most valuable pieces of information. Until recently, organizations have been limited to using subsets of their data, or they were constrained to simplistic analyses because the sheer volumes of data overwhelmed their processing platforms. The challenges include capture, 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 [8].

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. 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.

6. Issues of Big Data

6.1 Issue 1: Skills

There's definitely a skills issue. It's probably part of why the first wave of Big Data initiatives hasn't been terribly successful. Essentially, there are the same fundamental challenges as with small-data initiatives in the sense that getting real knowledge out of data is not really an IT capability. It's an analytics and data-science capability and that skill set is not there. It is compounded when you add big data to it because the Big Data technologies in play require much more of a software development bent, rather than an IT systems management skill set [9].

6.2 Issue 2: Data Structure

The challenge is that it all goes back to the beginning and how you structure data to make it accessible for ad-hoc analysis and make it flexible enough that you can get some things out. Companies like Tableau Software and Qlik Tech have shown the world and advanced users of Microsoft Excel is that you don't have to be a database expert to start manipulating data in an ad-hoc way and coming up with interesting views and insights provided when you started, the data warehouse is appropriately structured. It's very hard to fix after the fact. The challenge today is that most enterprise data warehouses view a customer or an entity that the business works with as a row of data rather than a column [10]. That row is populated and updated perhaps on a daily basis with snapshot or aggregate views of the current state of the customer. But you've collapsed away all the data that tells you about what the individual entity has actually done and the things that have accumulated about them over the course of their relationship. That makes it much harder to go back as a BI and analytics insights team and recover and start building models that are predictive or actionable in shaping behavior or changing the relationship you have as an enterprise with your customers.

6.3 Issue 3: Collection of data

The challenge often see is the data is collected and persistently stored, frequently for the purpose of disaster recovery. But that kind of long-term or expanded storage perpetuates the same schema that exists live, rather than perpetuating data in a more native form that you could go back to and then change how you subsequently process it and bring it live. It burns in whatever the first thinking was about how that data should be used. So the most successful enterprises I've seen have been ones where they have an archival process that stores in very cold or slow storage the most basic data and doesn't view that as a disaster recovery system and then has a disaster-recovery system for the current live system.

7. Emerging Technologies of Big Data

Big Data is broad and encompasses many trends and new technology developments, to give a very good overview of top ten emerging technologies that are helping users cope with and handle Big Data in a cost-effective manner.

1. Column-oriented databases
2. NoSQL databases
3. MapReduce
4. Hadoop
5. HIVE

7.1 Column-oriented databases

Traditional, row-oriented databases are excellent for online transaction processing with high update speeds, but they fall short on query performance as the data volumes grow and as data become more unstructured. Column-oriented databases store data with a focus on columns, instead of rows, allowing for huge data compression and very fast query times. The downside to these databases is that they will generally only allow batch updates, having a much slower update time than traditional models.

7.2 NoSQL databases

There are several database types that fit into this category, such as key value stores and document stores, which focus on the storage and retrieval of large volumes of unstructured, semi-structured, or even structured data. They achieve performance gains by doing away with some of the restrictions traditionally associated with conventional databases, such as read-write consistency, in exchange for scalability and distributed processing.

7.3 Map Reduce

The most widely known technology that helps to handle large data would be a distribution data process framework of the Map-Reduce method, such as Apache Hadoop. Data processing via the Mapreduce method has the following characteristics:

- It operates via regular computer that uses built-in hard disk, not a special storage. Each computer has extremely weak correlation where expansion can be hundreds and thousands of computers.
- Since many computers are participating in processing, system errors and hardware errors are assumed as general circumstances, rather than exceptional.
- With a simplified and abstracted basic operation of Map and Reduce, you can solve many complicated problems. Programmers who are not familiar with parallel programs can easily perform parallel processing for data.
- It supports high throughput by using many computers.

The following figure displays the implementation flow of the map-reduce method. Data stored in the HDFS storage is divided to available worker and expressed a value type, and

results are stored in a local disk. The data is complied by reducing worker and generate a result file. Depending on the characteristics of data storage, make the best use of locality by reducing the gap between a node which is processing data and source data location by placing worker in the location where data is stored [13]. Each worker can be implemented in various languages through streaming interface.

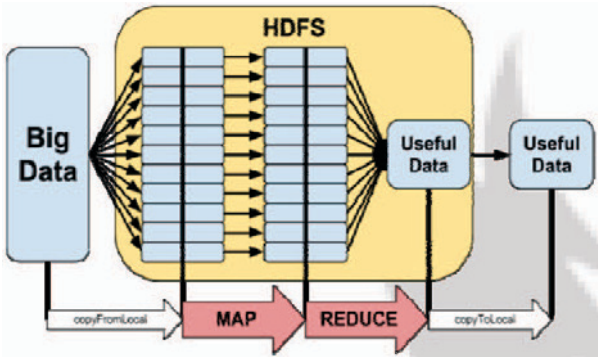

Figure 3: Map Reduce

This is a programming paradigm that allows for massive job execution scalability against thousands of servers or clusters of servers. Any MapReduce implementation consists of two tasks:

- The Map task, where an input dataset is converted into a different set of key pairs, or tuples.
- The Reduce task, where several of the outputs of the Map task are combined to form a reduced set of tuples.

7.4 Hadoop

Hadoop is by far the most popular implementation of MapReduce, being an entirely open source platform for handling Big Data. It is flexible enough to be able to work with multiple data sources, either aggregating multiple sources of data in order to do large scale processing, or even reading data from a database in order to run processor-intensive machine learning jobs. It has several different applications, but one of the top use cases is for large volumes of constantly changing data, such as location-based data from weather or traffic sensors, web-based or social media data, or machine-to-machine transactional data.

7.5 HIVE

HIVE is a SQL-like bridge that allows conventional BI applications to run queries against a Hadoop cluster. It was developed originally by Facebook, but has been made open source for some time now, and it's a higher-level abstraction of the Hadoop framework that allows anyone to make queries against data stored in a Hadoop cluster just as if they were manipulating a conventional data store [14]. It amplifies the reach of Hadoop, making it more familiar for BI users. Apache HIVE helps to analyze large data by using the query language called HIVEQL for data source, such as HDFS or HBase. Architecture is divided into Map Reduce oriented execution, Meta data information for data storage, and an execution part that receives a query from user or applications for execution. To support expansion by user, it allows user specified function at the scalar value, aggregation, and table level.

8. Uses of Big Data

Technologies today not only support the collection and storage of large amounts of data, they provide the ability to understand and take advantage of its full value, which helps organizations run more efficiently and profitably. For instance, with big data and big data analytics, it is possible to [15].

- Analyze millions of SKUs to determine optimal prices that maximize profit and clear inventory.
- Recalculate entire risk portfolios in minutes and understand future possibilities to mitigate risk.
- Mine customer data for insights that drive new strategies for customer acquisition, retention, campaign optimization and next best offers.
- Quickly identify customers who matter the most.
- Generate retail coupons at the point of sale based on the customer's current and past purchases, ensuring a higher redemption rate.
- Send tailored recommendations to mobile devices at just the right time, while customers are in the right location to take advantage of offers.
- Analyze data from social media to detect new market trends and changes in demand.
- Use clickstream analysis and knowledge mining to detect fraudulent behavior.
- Determine root causes of failures, issues and defects by investigating user sessions, network logs and machine sensors.

9. Examples of Big Data

Here are some real-world examples of Big Data in action;

- Consumer product companies and retail organizations are monitoring social media like Facebook and Twitter to get an unprecedented view into customer behavior, preferences, and product perception.
- Manufacturers are monitoring minute vibration data from their equipment, which changes slightly as it wears down, to predict the optimal time to replace or maintain. Replacing it too soon wastes money; replacing it too late triggers an expensive work stoppage
- Manufacturers are also monitoring social networks, but with a different goal than marketers: They are using it to detect aftermarket support issues before a warranty failure becomes publicly detrimental.
- Financial Services organizations are using data mined from customer interactions to slice and dice their users into finely tuned segments. This enables these financial institutions to create increasingly relevant and sophisticated offers.
- Advertising and marketing agencies are tracking social media to understand responsiveness to campaigns, promotions, and other advertising mediums.
- Insurance companies are using Big Data analysis to see which home insurance applications can be immediately processed, and which ones need a validating in-person visit from an agent.

As an example, one task manager process maintains the list of tasks to be performed, and doles out those tasks based on data locality to minimize the time associated with data latency, which slows down the computing process. The last part of making a developed application executable is the system configuration.

10. Conclusion

The main advantage of integrating structured and unstructured data is to provide a higher degree of flexibility. In this paper we compared structured, unstructured and semi-structured data and discussed about integration of structured and unstructured data using mapreduce and hadoop framework. We had also discussed about the issues and uses of Big Data with various examples.

11. Future Scope of the Study

Moreover, tools currently available either analyze the structured data or the unstructured data but not the both. Hence, future implementation could be towards integrating varieties of data with best efficient direction which is highly recommended.

References

[1] Blumberg, R., Atre, S.: The Problem with Unstructured Data. http://www.dmreview.com/issues/20030201/6287-1.html (19.02.2009) (2003)
[2] Abiteboul, S., Buneman, P., Suciu, D.: Data on the Web: from relations to semistructured data and XML. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA (1999)
[3] Manola, F., Miller, E.: Resource Description Framework (RDF): Concepts and Abstract Syntax. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/ (19.02.2009) (2004)
[4] Brickley, D., Guha, R.: RDF Vocabulary Description Language 1.0: RDF Schema. http://www.w3.org/TR/2004/REC-rdf-schema-20040210/ (19.02.2009) (2004)
[5] Aramburu, Juan Manuel Perez, Maria Jose, Rafael Berlanga and Torben Bach Pedersen, Integrating Data warehouse with Web Data: A Survey, IEEE transaction on knowledge and Data Engineering, Volume 20, July 2008.
[6] Allen, D.: Seam in Action. Manning Publications Co. Greenwich, CT, USA (2008)
[7] Bauer, C.: Java Persistence with Hibernate. Manning Publications Co. Greenwich, CT, USA (2006)
[8] DeMichiel, L., Keith, M.: JSR 220: Enterprise JavaBeansTM, Version 3.0. http://java.sun.com/products/ejb/docs.html (20.02.2009) (2006)
[9] Cheung, S., Matena, V.: Java Transaction API (JTA). http://java.sun.com/javaee/technologies/ jta/index. jsphttp:// java.sun.com/javaee/technologies/jta/index.jsp (19.02.2009) (2002)
[10] Basel Kayyali, David Knotl, Peter Crroves and Steve Van Kuiken. The Big Data revolution in healthcare. Accelerating value and innovation, January 2013.
[11] Connolly, T., Begg, C.: Database Systems: A Practical Approach to Design, Implementation, and Management. Addison Wesley Publishing Company (2005)
[12] Stroka, S.: Transaction Management in Federated, Heterogeneous Database Systems for Semantic Social Software Applications. (2009)
[13] Francois Bry, Michael Eckert, J.K., Weiand, K.: What the User interacts with: Reections On Conceptual Models For Semantic Wikis. (2009)
[14] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for projected clustering of high dimensional data streams. In Proc. 2004 Int. Conf. Very Large Data Bases (VLDB'04), pages 852{863, Toronto, Canada, Aug. 2004.
[15] Chamikara Jayalath, Julian Stephen and Patrick Eugster. From the Cloud to the Atmosphere. Running Map Reduce across Datacenters, IEEE Transaction on Computers special issues on cloud of clouds, 2013.

Author Profile

N. Monica pursuing M .Tech in Information Technology from Hindustan Institute of Technology and Science at Chennai. Her research focus lies on integrating structured and unstructured data to provide higher degree of flexibility.

Dr. K. Ramesh Kumar he received his PhD degree in Computer Science and Engineering from Alagappa University, Karaikudi, India in 2011. His PhD involved the development of algorithms to frequent pattern and association rule mining, and constructs new dataset for AIDS/HIV infected patients' case history. He had published 32 research papers in various journals and conferences.