International Journal of Information and Computation Technology.
ISSN 0974-2239 Volume 3, Number 7 (2013), pp. 655-662
© International Research Publications House
http://www. irphouse.com /ijict.htm

Data Refinery with Big Data Aspects

Rajan Kumar[1] and Sarika Gupta[2]

[1]Department of Information Technology, Dronacharya College of Engineering, Greater Noida U.P, India
[2]Department of Information Technology, Dronacharya College of Engineering, Greater Noida U.P, India.

Abstract

Big data has become an extremely popular term, due to the well documentary explosion in the amount of data being stored and processed by today businesses. According to IDC [1] the amount of digital information produced in 2011 will be ten times that produced in 2006: 1,800 Exabyte's. The majority of this data will be "unstructured" – complex data poorly-suited to management by structured storage systems like relational databases. Unstructured data comes from many sources and takes many forms –web logs, text files, sensor readings, user-generated Content like product reviews or text messages, audio, video and still imagery and more. 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. Through our work, we examine the data refinery objects and application that mostly used to handle the big data aspects like Hadoop, MapReduce, cloud database, TeraData[2] and many other platform that try to analysis the best way to handle this problem.

Keywords: Big Data, cloud architecture, Hadoop, MapReduce, TeraData.

1. Introduction

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process the data within time. Another reason that big data is a hot topic in the market today and new technology that enables an organization to take advantage of the natural resources of big data. Big data itself is not new – it has been here for a while and growing exponentially. What is new is the technology to process and analyze it. The purpose of big data technology is to cost effectively manage and analyze all of the available data. We can see the wide variety of sources of big data. It comes from our traditional systems – Billing systems, ERP systems[3], CRM systems[4].It also comes from machine data – from sensors, network switches. Moreover, it comes from humans – website data, social media, etc. In business publications and IT trade journals, the buzz about “big data” challenges is nearly deafening. Rapidly growing volumes of data – from transactional systems like enterprise resource planning (ERP) software and non-transactional sources such as web logs, customer call center records, and video images – are everywhere. A tsunami of data, some experts call it. In a 2001 research report and related lectures, META Group defined data growth challenges and opportunities as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). Now much of the industry, continue to use this "3Vs" model for describing big data.

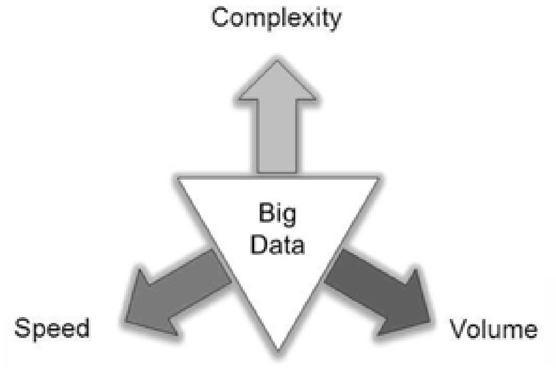

Figure 1: 3Vs model of Big Data integrity.

1.1Volume-The amount of data generated by companies – and their customers, competitors, and partners – continues to grow exponentially. According to industry analyst IDC, the digital universe created and replicated 1.8 trillion gigabytes in 2011.2 that is the equivalent of 57.5 billion 32GB Apple iPods.

1.2 Velocity-Data continues changing at an increasing rate of speed, making it difficult for companies to capture and analyze. For example, machine-generated data from sensors and web log data is being ingested in real-time by many applications. Without real-time analytics to decipher these dynamic data streams, companies cannot make sense of the information in time to take meaningful action.

1.3 Variety-It is no longer enough to collect just transactional data – such as sales, inventory details, or procurement information. Analysts are increasingly interested in new data types, such as sentiments expressed in product reviews, unstructured text from call records and service reports, online behavior such as click streams, images and videos, and geospatial and temporal details. These data types add richness that supports more analyses that are detailed.

Moreover they also analysis about the complexity

1.4 Complexity-With more details and sources, the data is more complex and difficult to analyze. In the past, banks used just transactional data to predict the probability of a customer closing an account. Now, these companies want to understand the “last mile” of the customer’s decision process. By gaining visibility into common consumer behavior patterns across the web site, social networks, call centers, and branches, banks can address issues affecting customer loyalty before consumers decide to defect. Analyzing and detecting patterns – on the fly across and all customer records – is time-consuming and costly. Replicating that effort over time can be even more challenging.

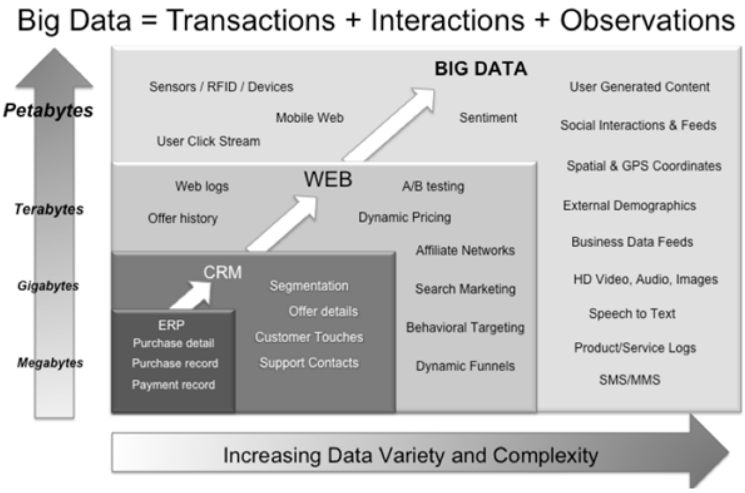

Figure 1.2: Increasing size of data with variety velocity and complexity.

2. Management of Big Data Aspects

Every major sector of industry and organization are hit by their own data that need to manage it , some of them developed their own technique hand and other takes help from data handler tools. These reliable technique and tools may be used different kind of handler but they have to pass from common problems.

Dealing with big data required two major things:

2.1 Inexpensive, reliable storage

2.2 New tool for analyzing the structure the unstructured data

2.1 Reliable Storage: HDFS Hadoop includes a fault-tolerant storage system called the Hadoop Distributed File System, or HDFS. HDFS is able to store huge amounts of information, scale up incrementally and survive the failure of significant parts of the storage Infrastructure without losing data. Hadoop creates clusters of machines and coordinates work among them. Clusters can be built with inexpensive computers. If one fails, Hadoop continues to operate the cluster without losing data or interrupting work, by shifting work to the remaining machines in the cluster. HDFS manages storage on the cluster by breaking incoming files into pieces, called "blocks," and storing each of the blocks redundantly across the pool of servers. In the common case, HDFS stores three complete copies of each file by copying each piece to three different servers.

HDFS has several useful features. In the very simple example shown, any two servers can fail, and the entire file will still be available. HDFS notices when a block or a node is lost, and creates a new copy of missing data from the replicas it.

2.2 New tool for analyzing the structure the unstructured data is most usable in today's business and IT operations because of some important and reliable factor that allowing users to:

- Capture and refine data from a wide variety of sources
- Perform necessary multi-structured data preprocessing
- Develop rapid analytics Process embedded analytics, analyzing both relational and non-relational data
- Produce semi-structured data as output, often with metadata and heuristic analysis
- Solve new analytical workloads with reduced time to insight
- Use massively parallel storage in Hadoop to efficiently store and retain data

2.3 HADOOP Big Idea: Many popular tools for enterprise data management relational database systems are designed to make simple queries run quickly. They use techniques like indexing to examine just a small portion of all the available data in order to answer a question. Hadoop is a different sort of tool. Hadoop is aimed at problems that require examination of all the available data. For example, text analysis and image processing generally require that every single record be read, and often interpreted in the context of similar records. Hadoop uses a technique called MapReduce to carry out this exhaustive analysis quickly. In the previous section, we saw that HDFS distributes blocks from a single file among a large number of servers for reliability. Hadoop takes advantage of this data distribution by pushing the work involved in an analysis out to many different servers. Each of the servers runs the analysis on its own block from the file. Results are collated and digested into a single result after each piece has been analyzed.

2.4 MapReduce supports distributed processing of the common map and reduction operations. In the map step, a master node divides a query or request into smaller problems. It distributes each query to a set of map tasks scheduled on a worker node within a cluster of execution nodes. The output of the map steps is sent to nodes that combine or reduce the output and create a response to the query. Because both the map and reduce functions can be distributed to clusters of commodity hardware and performed in parallel, MapReduce techniques are appropriate for larger datasets.

Many people think big data is about Hadoop technology. It is and it is not. It is about a lot more than Hadoop. One of the key requirements is to understand and navigate federated sources of big data – to discover data in place. New technology has emerged that discovers, indexes, searches, and navigates diverse sources of big data. Of course, big data is also about Hadoop. Hadoop is a collection of open source capabilities. Two of the most prominent ones are Hadoop File System for storing a variety of information, and MapReduce – a parallel processing engine. Data warehouses also manage big data- the volume of structured data is growing quickly. The ability to run deep analytic queries on huge volumes of structured data is a big data problem. It requires massive parallel processing data warehouses and purpose-built appliances for deep analytics. Big data is not just at rest – it is also in motion. Streaming data represents an entirely different big data problem – the ability to quickly analyze and act upon data while it is still moving. This new technology opens a world of possibilities – from processing volumes of data that were just not practical to store, to detecting insight and responding quickly. As much of the worlds, big data is unstructured and in textual content, text analytics is a critical component to analyze and derive meaning from text. Integration and governance technology establishes the veracity of big data, and is critical in determining whether information is trusted or not.

3. Cloud Technology in Big Data

A recent survey conducted by GigaSpaces found that 80 percent of those IT executives who think big data processing is important are considering moving their big data analytics to one or more cloud delivery models. Cloud delivery models offer exceptional flexibility, enabling IT to evaluate the best approach to each business user’s request. For example, organizations that already support an internal private cloud environment can add big data analytics to their in-house offerings, use a cloud services provider, or build a hybrid cloud that protects certain sensitive data in a private cloud, but takes advantage of valuable external data sources and applications provided in public clouds. Using cloud infrastructure to analyze big data makes sense because Investments in big data analysis can be significant and drive a need for efficient, cost-effective infrastructure. Private clouds can offer a more efficient, cost-effective model to implement analysis of big data in-house, while augmenting internal resources with public cloud services. This hybrid cloud option enables companies to use on-demand storage space and computing power via public cloud services for certain analytics initiatives (for example, short-term projects), and provide added capacity and scale as needed. Big data may mix internal and external sources. While enterprises often keep their most sensitive data in-house, huge volumes of big data (owned by the organization or generated by third-party and public providers) may be located externally—some of it already in a cloud environment. Data services are needed to extract value from big data. Depending on requirements and the usage scenario, the best use of your IT budget may be to focus on analytics as a service supported by your internal private cloud, a public cloud, or a hybrid model.

4. Security

Aggregating data by nature increases the risk that a cybercriminal or insider (malicious or otherwise) can compromise sensitive information. Therefore, organizations should strictly limit the number of people who can access repositories like Hadoop. Big data environments should include basic security and controls as a way to defend and protect data. First, access control ensures that the right user gets access to the right data at the right time. Second, continuously monitoring user and application access is highly important especially as individuals changes roles or leave the organization. Monitoring data access and usage patterns can alert security teams to potential abuse or security policies violations like an administrator altering log files. Typically, internal attackers or cybercriminals will leave clues or artifacts about their breach attempts that can be detected through careful monitoring. Monitoring helps ensure security policies are enforced and effective. Organizations can secure data using data abstraction techniques such as encryption or masking. Generally, cybercriminals cannot easily decrypt or recover data after it has been encrypted or masked. The unfortunate reality is that organizations need to adopt a zero trust policy to ensure complete protection.

5. Conclusion

Organizations do not have to feel overwhelmed when it comes to securing big data environments. The same security fundamentals for securing databases, data warehouses and file share systems can be applied to securing Hadoop implementations these solutions scale to protect both traditional data management architectures and big data environments and protect against a complex threat landscape including insider fraud, unauthorized changes and external attacks while remaining focused on business goals and automating compliance.

References

[1] An Updated Forecast of Worldwide Information Growth Through 2011," IDC, March 2008.

[2] Hortonworks TeraData(Best decision possible) http://hortonworks.com/partner/teradata/

Tim Kraska • kraskat@cs.brown.edu Published by the IEEE Computer Society 1089-7801/13/$31.00 © 2013 IEEE IBM Big Data Platform Overview Martin Pavlík +420 731 435 691 martin_pavlik@cz.ibm.com

[3] ERP Enterprise Resource Planning http://www.webopedia.com/TERM/E/ERP.html

[4] Customer_relationship_managementen.wikipedia.org/wiki/Customer_relationship_managemen Big Data: Issues and Challenges Moving Forward

[5] 2013 46th Hawaii International Conference on System Sciences1530-1605/12 $26.00 © 2012 IEEE

[6] Rob Peglar Introduction Analytics BigData Hadoop SNIAEDUCATIONtrackvirtualizationapplication@snia.org

[7] BIG DATA ANALYSIS http://en.wikipedia.org/wiki/Big_data

[8] Big Data Processing in Cloud Computing Environments 2012 International Symposium on Pervasive Systems, Algorithms and Networks 1087-4089/12 $26.00 © 2012 IEEE.