IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Security Challenges in Big Data

Dinesh Singh[1], Dayanand[2], Arushi Arya[3]

Dept. of Computer Applications[1], Dept. of CSE[2], Dept. of CS[3]
TERI PG College, Ghazipur[1], KIET Group of Institutions[2], University of South California[3]

Available online at: www.ijcseonline.org

Accepted: 19/July/2018, Published: 31/Jul/2018

Abstract- We have entered in the 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, attacks against data are also increasing. We need to protect the data we collect and store, securely transfer data over the network and keep output derived from these data confidential. In this paper, we discuss about various challenges big data technology face today and tools to overcome those challenges.

Keywords: Big Data, Security, Hadoop Technology

I. INTRODUCTION

Big Data analysis is one of the most emerging fields in today’s technology driven world. With the boom in connectivity through internet via various social networking sites like LinkedIn , Facebook, Twitter , Tinder etc and also various ecommerce websites , it won’t be wrong to say that we have come to an era in which we have theme specific online stores and hangouts(social network sites).

These sites generate large chunks of data, usually in Terabytes and Petabytes, which at most of the times, gets almost impossible to analyze by standard RDBMS software and database management tools. These type datasets are termed as ‘BIG DATA’.

Big data is defined in terms of ‘3V’ which are Volume, Velocity and Veracity. Big data may be new for stratus, but most of the large firms have been struggling with it for quite a long time, but it cannot be denied that analysis of big data is in today’s world one of the most necessary task, be it for advertising, SEO, attracting for clients or even the analysis of present data-sets.

Hadoop, developed by apache is an open source framework, which works on the principle of OLAP. Analysis of big data is usually done through this. It has various components which serve various purposes like PIG, YARN, SCOOP, MongoDB, FLUME all running on Hadoop Distributed File System. Hadoop enables distributed parallel processing of huge amounts of data among inexpensive, industry level servers that both store and process the data. Hadoop is implemented on Linux; it also has its separate environment given by Cloudera.

Apache Hadoop is 100% open source and pioneered new way of fundamentally storing and processing data relying on commodity hardware. With Hadoop, no data is too big.

II. BACKGROUND

Big data is a popular term used to describe the exponential growth and availability of data, both structured, unstructured and Semi Structured. And big data may be as important to business – and society – as the Internet has become.

1. Lots of Data (Terabytes or Gigabytes)

2. Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, storage, search, sharing, transfer, analysis, and visualization.

3. Systems / Enterprises, Internet users, generate huge amount of data from Gigabytes to and even Terabytes of information.

4. Volume. 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.

5. Velocity. 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.

6. Variety. 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.

III. **BIG DATA CHALLENGES**

1. Need for speed - Now This Time hypercompetitive business environment, companies not only have to find and analyze the relevant data they need, they must find it quickly.

2. Visualization helps organizations perform analyses and make decisions much more rapidly, but the challenge is going through the sheer volumes of data and accessing the level of detail needed, all at a high speed.

3. Data quality - Analyze data quickly and put it in the proper context for the audience that will be consuming the information, the value of data for decision-making purposes if the data is not accurate or timely. This is a challenge with any data analysis, but when considering the volumes of information involved in big data projects, it becomes even more pronounced.

4. Understanding the data - It takes a lot of understanding to get data in the right shape so that you can use visualization as part of data analysis For example, if the data comes from social media content, you need to know who the user is in a general sense – such as a customer using a particular set of products – and understand what it is you’re trying to visualize out of the data. Without some sort of context, visualization tools are likely to be of less value to the user.

IV. TYPES OF DATA

1. Unstructured Data

Unstructured data files often include text and multimedia content. Note that while these sorts of files may have an internal structure, they are still considered "unstructured" because the data they contain doesn't fit neatly in a database. Experts estimate that 80 to 90 percent of the data in any organization is unstructured. And the amount of unstructured data in enterprises is growing significantly often many times faster than structured databases are growing. Photos and graphic images, videos, streaming instrument data, webpages, PDF files, PowerPoint presentations, emails, blog entries, wikis and word processing documents.

2. Structured data

Data that resides in a fixed field within a record or file is called structured data. This includes data contained in relational databases and spreadsheets. Structured data first depends on creating a data model – a model of the types of business data that will be recorded and how they will be stored, processed and accessed. This includes defining what fields of data will be stored and how that data will be stored: data type (numeric, currency, alphabetic, name, date, address).

3. Semi-Structured Data

Semi-structured data is a cross between the two. It is a type of structured data, but lacks the strict data model structure. In 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. In object-oriented databases, one often finds semi-structured data. XML is an example.

V. SECURITY CHALLENGES

Applications of big data introduce security problems such as the following: -

1. Issue of storing all data
   Large amount of data is being generated. Traditional RDBMS tools are not able to store or process this data. It is a good choice to compress the data using compression technology. Organizations can look for options like data lakes and collect and store unstructured data. But, data has to be stored in these lakes wisely or it will never be retrieved.

2. Veracity of data
   With the massive amount of data being collected and stored, it is difficult to be sure that data satisfies the level of authenticity needed by the analysis algorithms to produce accurate output. Hence, authenticity and integrity of data used in the tools should be checked using machine learning and statistics.
3. Privacy of data
   There should be limitations to the use of data sets containing sensitive data. However, tools and emerging technologies have made it easier to extract and correlate data, making it easier to violate the privacy of data. We need designs aimed at creating suitable safeguards to prevent data violations.
4. Analysis of data

As the volume of data increases, it becomes difficult to analysis it all at once. So, data is disseminated, then processed. Finally, processed data is aggregated.

5. Encryption of data
   Traditional encryption algorithms are not suitable for encrypting big datasets. They only cover one aspect while big data gets exported to different environments such as vendor database and combined with different sources.

6. Reporting of data
   Reports generated using big data should be easy to understand and interpret by people. They should be represented in a form that can be easily recognised by looking at them.

VI. PROPOSED SYSTEM DISCRIPTION

Data is constantly at risk by attacks launched through the internet or using social engineering techniques. Some steps that can be followed to keep big data safe are as follows: -

1. Restrictions at each source – Big data extract and collect data from different sources. It is important that these sources have their own access restrictions and policies. It will provide appropriate security to all data sources.
2. Infrastructure of big data – The distributive nature of big data environments, security controls need to be standardised across all locations. It should allow access to scientists or analytical tools to data while protecting the system from possible attacks.
3. Big data tools do not keep security as priority – Programming tools including Hadoop and NoSQL databases do not focus on security. Originally hadoop did not authenticate users or encrypt data being transmitted in the environment, making data vulnerable to attacks.

HADOOP –BIG DATA TOOL

Introduction to Hadoop

Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware.

All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines, or racks of machines) are commonplace and thus should be automatically handled in software by the framework.

1. Hadoop provides a reliable shared storage (HDFS) and analysis system (MapReduce).

2. Hadoop is highly scalable and unlike the relational databases, Hadoop scales linearly. Due to linear scale, a Hadoop Cluster can contain tens, hundreds, or even thousands of servers.

3. Hadoop is very cost effective as it can work with commodity hardware and does not require expensive high-end hardware.

Advantages of Hadoop

1. Data Size and Data Diversity - When you are dealing with huge volumes of data coming from various sources and in a variety of formats then you can say that you are dealing with Big Data. In this case, Hadoop is the right technology for you.

2. Future Planning - It is all about getting ready for challenges you may face in future. If you anticipate Hadoop as a future need then you should plan accordingly. To implement Hadoop on you data you should first understand the level of complexity of data and the rate with which it is going to grow.

3. Multiple Frameworks for Big Data - There are various tools for various purposes. Hadoop can be integrated with multiple analytic tools to get the best out of it, like Mahout for Machine-Learning, R and Python for Analytics and visualization, Python, Spark for real time processing, MongoDB and Hbase for Nosql database, Pentaho for BI etc.

4. Lifetime Data Availability - When you want your data to be live and running forever, it can be achieved using Hadoop's salability. There is no limit to the size of cluster that you can have. You can increase the size anytime as per your need by adding data nodes to it.

Architecture of Hadoop

1. Hadoop works in a master-worker / master-slave fashion.

2. Hadoop has two core components: HDFS and MapReduce.

3. HDFS (Hadoop Distributed File System) offers a highly reliable and distributed storage, and ensures reliability, even on commodity hardware, by replicating the data across multiple nodes. Unlike a regular file system, when data is pushed to HDFS, it will automatically split into multiple blocks (configurable parameter) and stores/replicates the data across various

data nodes. This ensures high availability and fault tolerance.

4. Map Reduce offers an analysis system which can perform complex computations on large datasets. This component is responsible for performing all the computations and works by breaking down a large complex computation into multiple tasks and assigns those to individual worker/slave nodes and takes care of coordination and consolidation of results.

5. The master contains the Name node and Job Tracker components.

5.1 Name node holds the information about all the other nodes in the Hadoop Cluster, files present in the cluster, constituent blocks of files and their locations in the cluster, and other information useful for the operation of the Hadoop Cluster.

5.2 Job Tracker keeps track of the individual tasks/jobs assigned to each of the nodes and coordinates the exchange of information and results.

6. Each Worker / Slave contains the Task Tracker and Data node components.

6.1 Task Tracker is responsible for running the task / computation assigned to it.

6.2 Data node is responsible for holding the data.

7. The computers present in the cluster can be present in any location and there is no dependency on the location of the physical server.

VII. VARIOUS TOOLS USED IN HADOOP-

HIVE - Hive provides a warehouse structure and SQL-like access for data in HDFS and other Hadoop input sources (e.g. Amazon S3). Hive's query language, HiveQL, compiles to MapReduce. It also allows user-defined functions (UDFs). Hive is widely used, and has itself become a "subplatform" in the Hadoop ecosystem.

PIG - Pig is a framework consisting of a high-level scripting language (Pig Latin) and a run-time environment that allows users to execute MapReduce on a Hadoop cluster. Like HiveQL in Hive, Pig Latin is a higher-level language that compiles to MapReduce.

MAHOUT – It is a scalable machine-learning and data mining library. There are currently four main groups of algorithms in Mahout:

1. Recommendations, a.k.a. collective filtering

2. Classification, a.k.a categorization

3. Clustering

4. Frequent itemset mining, a.k.a parallel frequent pattern mining

MAP REDUCE

1. The Map Reduce paradigm for parallel processing comprises two sequential steps: map and reduce.

2. In the map phase, the input is a set of key-value pairs and the desired function is executed over each key/value pair in order to generate a set of intermediate key/value pairs.

HBASE

Based on Google's Bigtable, HBase "is an open-source, distributed, versioned, column-oriented store" that sits on top of HDFS. HBase is column-based rather than row-based, which enables high-speed execution of operations performed over similar values across massive data sets, e.g. read/write operations that involve all rows but only a small subset of all columns. HBase does not provide its own query or scripting language, but is accessible through Java, Thrift, and REST

APIs.

In the reduce phase, the intermediate key/value pairs are grouped by key and the values are combined together according to the reduce code provided by the user; for example, summing. It is also possible that no reduce phase is required, given the type of operation coded by the user.

CONCLUSION

Many businesses already use Big Data for marketing and research, yet may not have the fundamentals right – particularly from a security perspective. As with all new technologies, security seems to be an afterthought at best.

Big Data breaches will be big too, with the potential for even more serious reputational damage and legal repercussions than at present.A growing number of companies are using the technology to store and analyse petabytes of data including web logs, click stream data and social media content to gain better insights about their customers and their business. As a result, information classification becomes even more critical; and information ownership must be addressed to facilitate any reasonable classification.

Most organisations already struggle with implementing these concepts, making this a significant challenge. We will need to identify owners for the outputs of Big Data processes, as well as the raw data. Thus data ownership will be distinct from information ownership – perhaps with IT owning the raw data and business units taking responsibility for the outputs.

Very few organisations are likely to build a Big Data environment in-house, so cloud and Big Data will be inextricably linked. As many businesses are aware, storing data in the cloud does not remove their responsibility for protecting it - from both a regulatory and a commercial perspective.

REFERENCES

[1] Sameer Walker Affiliated with, Madhu Siddalingaiah "Motivation for Big Data-Pro Apache Hadoop" http://link.springer.com/chapter/10.1007/978-1-4302-48644_1#page-2

[2] Jonathan Stuart Ward and Adam Barker-"Undefined By Data: A Survey of Big Data Definitions" School of Computer Science- University of St Andrews, UK{jonthan.stuart.ward, adam.barker}@standrews.ac.ukhttp://arxiv.org/pdf/1309.5821v1.pdf

[3] BIGDATA ANALYTICS - 5th QUARTER BY PHILIP RUSSOM -TDWI research. http://www.tableau.com/sites/default/files/whitepapers/tdwi _bpreport_q411_big_data_analytics_tableau.pdf.

[4] Alvaro A. Cárdenas, Pratyusa K. Manadhata, Sreeranga "Big Data Analytics for Security"Posted by P. Rajanhttp://www.infoq.com/articles/bigdata-analytics-for-security.

[5] data electronically available at http://www.umuc.edu/cybersecurity/about/cybersecurityasics.cfm umuc[71]http://whatis.techtarget.com/definition/cybersecurity

[6] data electronically available at http://whatis.techtarget.com/glossary/Security-Threats-anduntermeasures

[7] data electronically available at teradata-http://www.teradata.com/Cyber-Security-http://bigdata.teradata.com/US/Success-Stories/Innovationsand-Insights/

[8] Securing Big Data - Part 1-Posted by Steve Jones at Tuesday, January 06, 2015

[9]data electronically available at http://service-architecture.blogspot.com/2015/01/securingbig-data-part-2-understanding.html.

[10] unstructured data in big data environment- data electronically available at http://www.dummies.com/how-to/content/unstructureddata-in-a-big-data-environment.html

[11] data electronically available at http://ictactjournals.in/paper/IJSC_Paper_6_pp_1035_1049. pdf

Authors Profile

Dr. Dinesh Singh has completed B.Sc. from Allahabad University, India , Masters in Computer Applications from KNIT SultanPur, India and Doctorate from Samhiggin bottom Institute of Agricultural, Technology and Science. Currently he is teaching in TERI PG College in Dept. MCA. He has organised various conferences and has rich experience of more than 20 years in academics. He has guided several post graduate projects and thesis.

Dayanand has completed bachelors in Technology in Computer Science Engineering from SHIATS, Allahabad, Masters from Birla Institute of Technology, Ranchi in 2013 and currently he is pursuing Doctorate from Sam Higginbottam University of agricultural technology and Sciences, State University, Uttar Pradesh. He has worked as manager IT in Govt. of Delhi and has done a number of govt. projects . He has an experience of 4 years in academics and currently working with KIET group of Institutions, Ghaziabad. He has authored books namely Foundation of Computer Science, Discrete Mathematics and Information Security. He has been awarded Dr. Rajendra Prasad Teachers award 2016. He has published more than 40 research papers in various conferences and journals.

Arushi Arya has completed her Bachelors in Technology in Computer Science Engineering from HMRITM, Guru Gobind Singh Indraprastha University, Delhi. She has Studied Statistical analysis of dataset at University of Rome Tor Vergata. She also Studied Network Security at Sofia University, Europe. Currently she is pursuing her Masters in Computer Science from University of Southern California. She has published 6 research papers and has completed various industrial projects.