Security and Privacy Issues in Big Data

Mr. Manohar Mungare

Asst. Professor
RMD Sinhgad School of Computer Studies, Warje, Pune

Abstract - In this paper, mainly focus is on security and privacy issues in Big data and Hadoop environment. Security and privacy concerns are growing as big data becomes more and more accessible. Big data applications are a great benefit to organizations, business, companies and many large scale and small scale industries. The biggest challenge for big data from a security point of view is the protection of user's privacy. Big data frequently contains huge amounts of personal identifiable information and therefore privacy of users is a huge concern. A popular data processing engine for big data is Hadoop. The overall problem of data security within Hadoop becomes even more difficult when you consider its implementation. Hadoop, and its underlying file system, is a complex distributed system with many points of contact.

I. INTRODUCTION

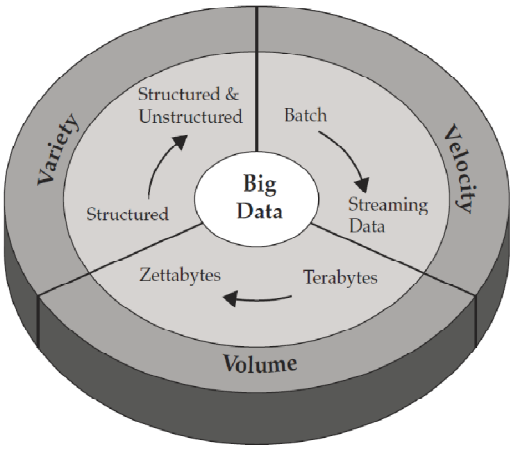
What is Big Data and Hadoop: So what is big data? One perspective is that big data is more and different kinds of data than is easily handled by traditional relational database management systems (RDBMSs). Some people consider 10 terabytes to be big data, but any numerical definition is likely to change over time as organizations collect, store, and analyze more data. Another useful perspective is to characterize big data as having high volume, high velocity, and high variety.

The three Vs-

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

II. CHARACTERIZATION OF BIG DATA:

Volume, Velocity and Variety (V3)



In short, "big data" means there is more of it. It comes more quickly, and comes in more forms.

Today, many organizations are collecting, storing, and analyzing massive amounts of data. This data is commonly referred to as "big data" because of its volume, the velocity with which it arrives, and the variety of forms it takes. Big data is creating a new generation of decision support data management. Businesses are recognizing the potential value of this data and are putting the technologies, people, and processes in place to capitalize on the opportunities.

For many, big data has become synonymous with Hadoop—an open source framework for parallel computing that runs on a distributed file system (such as the Hadoop Distributed File System, or HDFS.) For the first time, Hadoop enables organizations to cost-effectively store all their data as well as multi-structured data, such as Web server logs, sensor data, email and extensible markup language (XML) data. Because multi-structured data consists of 80% of all data by most estimates, the advent of Hadoop heralds the dawn of a new age of data processing in which organizations can pluck the needle out of the data haystack, which consists of terabytes, if not petabytes, of information.

III. SECURITY AND PRIVACY ISSUES

Security and privacy issues are magnified by the three V's of big data: Velocity, Volume, and Variety. These factors include variables such as large-scale cloud infrastructures, diversity of data sources and formats, streaming nature of data acquisition and the increasingly high volume of inter-cloud migrations. Consequently, traditional security mechanisms, which are tailored to securing small-scale static (as opposed to streaming) data, often fall short.

Security and privacy concerns are growing as big data becomes more and more accessible. The collection and aggregation of massive quantities of heterogeneous data are now possible. Large-scale data sharing is becoming routine among scientists, clinicians, businesses, governmental agencies, and citizens. However, the tools and technologies that are being developed to manage these massive data sets are often not designed to incorporate adequate security or privacy measures, in part because we lack sufficient training and a fundamental understanding of how to provide large-scale data security and privacy. We also lack adequate policies to ensure compliance with current approaches to security and privacy. Furthermore, existing technological approaches to security and privacy are increasingly being breached, whether accidentally or intentionally, thus necessitating the continual reassessment and updating of current approaches to prevent data leakage.

IV. WHAT ARE THE PRIVACY CONSTRAINTS?

The privacy of data is another huge concern, and one that increases in the context of Big Data. For electronic health records, there are strict laws governing what can and cannot be done. For other data, regulations, particularly in the US, are less forceful. However, there is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data.

Consider, for example, data gleaned from location-based services. These new architectures require a user to share his/her location with the service provider, resulting in obvious privacy concerns. Note that hiding the user's identity alone without hiding her location would not properly address these privacy concerns. An attacker or a (potentially malicious) location-based server can infer the identity of the query source from its (subsequent) location information. For example, a user's location information can be tracked through several stationary connection points (e.g., cell towers). After a while, the user leaves "a trail of packet crumbs" which could be associated to a certain residence or office location and thereby used to determine the user's identity. Several other types of surprisingly private information such as health issues (e.g. presence in a cancer treatment center) or religious preferences (e.g. presence in a church) can also be revealed by just observing anonymous users' movement and usage pattern over time. Note that hiding a user location is much more challenging than hiding his/her identity. This is because with location-based services, the location of the user is needed for a successful data access or data collection, while the identity of the user is not necessary.

There are many additional challenging research problems. For example, we do not know yet how to share private data while limiting disclosure and ensuring sufficient data utility in the shared data. The existing paradigm of differential privacy is a very important step in the right direction, but it unfortunately reduces information content too far in order to be useful in most practical cases. In addition, real data is not static but gets larger and changes over time; none of the prevailing techniques results in any useful content being released in this scenario. Yet another very important direction is to rethink security for information sharing in Big Data use cases. Many online services today require us to share private information (think of Facebook applications), but beyond record-level access control we do not understand what it means to share data, how the shared data can be linked, and how to give users fine-grained control over this sharing.

V. CHALLENGES OF SECURITY FROM THE PRODUCTION, STORAGE AND USE OF BIG DATA?

The biggest challenge for big data from a security point of view is the protection of user's privacy. Big data frequently contains huge amounts of personal identifiable information and therefore privacy of users is a huge concern.

Because of the big amount of data stored, breaches affecting big data can have more devastating consequences than the data breaches we normally see in the press. This is because a big data security breach will potentially affect a much larger number of people, with consequences not only from a reputational point of view, but with enormous legal repercussions.

When producing information for big data, organizations have to ensure that they have the right balance between utility of the data and privacy. Before the data is stored it should be adequately anonymized, removing any unique identifier for a user. This in itself can be a security challenge as removing unique identifiers might not be enough to guarantee that the data will remain anonymous. The anonymized data could be cross-referenced with other available data following de-anonymization techniques.

When storing the data organizations will face the problem of encryption. Data cannot be sent encrypted by the users if the cloud needs to perform operations over the data. A solution for this is to use "Fully Homomorphic Encryption" (FHE), which allows data stored in the cloud to perform operations over the encrypted data so that new encrypted data will be created. When the data is decrypted the results will be the same as if the operations were carried out over plain text data. Therefore, the cloud will be able to perform operations over encrypted data without knowledge of the underlying plain text data.

While using big data a significant challenge is how to establish ownership of information. If the data is stored in the cloud a trust boundary should be establish between the data owners and the data storage owners.

Adequate access control mechanisms will be key in protecting the data. Access control has traditionally been provided by operating systems or applications restricting access to the information, which typically exposes all the information if the system or application is hacked. A better approach is to protect the information using encryption that only allows decryption if the entity trying to access the information is authorized by an access control policy.

An additional problem is that software commonly used to store big data, such as Hadoop, doesn't always come with user authentication by default. This makes the problem of access control worse, as a default installation would leave the information open to unauthenticated users. Big data solutions often rely on traditional firewalls or implementations at the application layer to restrict access to the information.

VI. CONCLUSION

Big Data environment is widely used in industry and research aspects; therefore security is an important aspect for organizations running on these Big Data environments. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore

not cost-effective to address in the context of one domain alone. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

VII. REFERENCES*:*

1. *The Search for Analysts to Make Sense of Big Data. Yuki Noguchi. National Public Radio, Nov.30, 2011. http://www.npr.org/2011/11/30/142893065/the-search-for-analysts-to-make-sense-of-big-data[NYT2012]*

2. *The Age of Big Data. Steve Lohr. New York Times, Feb 11, 2012. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html*

3. *Hadoop IN Action- By Chuck Lam.*
4. *http://www.isaca.org/Groups/Professional-English/big-data/GroupDocuments/Big_Data_Top_Ten_v1.pdf*
5. *BIG DATA – Opportunities and Challenges*
   - *By BCS Learning & Development Limited.*
6. *www.amazon.in/**Big-Data-Hadoop**-WAGmob-e**book**/dp/B00E5GSS5E*