ISSN:0975-9646

Big Data Security Analytics

K. Radhika,
Lecturer in Computer science

G. Vijaya ,
Lecturer in Computer Science,

***Abstract*—The primary goal of big data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modelers and other analytics professionals to analyze large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence programs. Big data can be analyzed with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics, data mining, text analytics and statistical analysis.**

Big data is now a reality: the volume, variety and velocity of data coming into your organization continue to reach unprecedented levels. This phenomenal growth means that not only must you understand big data in order to decipher the information that truly counts, but you also must understand the possibilities of big data analytics.

Big Data analytics refers to the process of collecting, organizing and analyzing large sets of data to discover patterns and other useful information. Big data analytics can help organizations to better understand the information contained within the data and will also help identify the data that is most important to the business and future business decisions. Big Data analysts basically want the knowledge that comes from analyzing the data

INTRODUCTION:

The term Big Data refers to-large-scale information management and analysis technologies that exceed the capability of traditional data processing technologies. Big Data is differentiated from traditional technologies in three ways:

The amount of data (volume), the rate of data generation and transmission (velocity), and the types of structured and unstructured data the rate of data creation has increased so much that 90%of the data in the world today has been created in the last two years alone. This acceleration in the production of information has created a need for new technologies to analyze massive data sets.

Big Data analytics – the process of analyzing and mining Big Data – can produce operational and business knowledge at an unprecedented scale and specificity. The need to analyze and leverage trend data collected by businesses is one of the main drivers for Big Data analysis tools. The technologies advances in storage, processing, and analysis of Big Data include the rapidly decreasing cost of storage and CPU power in recent years the flexibility and cost-effectiveness of data centers and cloud computing for elastic computation and storage; and the development of new framework such as Hadoop, which allow users to take advantage of these distributed computing system storing large quantities of data through flexible parallel processing

Big Data technologies can be divided into two groups batch processing, which are analytics on data at rest, and stream processing, which are analytics on data in motion. Real-time processing does not always need to reside in memory, and mew interactive analyses of large-scale data sets through new technologies like drill and dermal provide new paradigms for data analysis.

Hadoop is one of the most popular technologies for batch processing. The Hadoop frame work provides developers with the Hadoop Distributed File System for storing large files and the Map-Reduce programming model, which is tailored for frequently occurring large-scale data processing problems that can be distributed and parallelized.

Several tools can help analysts create complex queries and run machine learning algorithms on top of Hadoop. This tool includes Pig, Hive and Mahout and RHadoop. New frame works such as Spark were designed to improve the efficiency of data mining and machine learning algorithms that repeatedly reuse a working set of data, thus improving the efficiency of advanced data analytics algorithms.

There are also several databases designed specifically for efficient storage and query Big Data, including Cassandra , CouchDB , Greenplum Database, HBase , MangoDB and vertica one of the models for stream processing is complex Event processing , which considered information flows as notification of events that need to be aggregated and combined to produce high-level events. Others particular implementation of stream technologies includes infoSphere Streams, jubatus and storm.

The preservation of privacy largely relies on technological limitations on the ability to extract, analyze and correlate potentially data sets. Advance in Big data analytics provide tools to extract and utilize this data, making violations of the privacy easier. In addition to privacy, data used for analytics may include regulated information or intellectual property. System architects must ensure that the data is protected and used only according to regulations.

Data-driven information security dates back to bank fraud detection and anomaly-based intrusion detection systems. Fraud detection is one of the most visible uses for Big Data analytics. The custom-built infrastructure to mine Big Data for fraud detection was not economical to adapt for other fraud detection uses. Off-the-self big data tools and techniques are now bringing attention to analytics for fraud detection in healthcare, insurance and other fields.

In the context of data analytics for intrusion detection the following evolution is anticipated:

- 1st Generation: Intrusion detection systems–Security architects realized the need for layered security because a system with 100% protective security is impossible.
- 2nd Generation: Security information and event management (SIEM) – Managing alerts from different intrusion detection sensors and rules was a big challenge in enterprise settings. SIEM Systems aggregate and filter alarms from many sources and present actionable information to security analysts.

- 3rd Generation: Big Data analytics in security- Big Data tools have the potential to provide a significant advance in actionable security intelligence by reducing the time for correlating, consolidating and contextualizing diverse security event information and also for correlating long-term historical data for forensic purposes.

Analyzing logs, network packets and system events for forensics and intrusion detection has traditionally been a significant problem however, Traditional technologies fail to provide the tools to support long-term, large-scale analytics for several reasons.

1. Storing and retaining a large quantity of data was not economically feasible.
2. Performing analytics and complex, queries on large, structured data sets was inefficient because traditional tools didn't leverage Big data technologies.
3. Traditional tools were not designed to analyze and manage unstructured data.
4. Big Data systems use cluster computing infrastructure.

For example:

For network security: The quantity of Data and frequency analysis of events are too much for traditional systems to handle alone. In traditional system, searching among a month's load of data could take between 20 minutes and an hour. In new Hadoop system running queries with Hive, they get the same results in about one minute. The security data warehouse driving this implementation not only enables users to mine meaningful security information from sources such as firewalls and security devices, but also from website traffic, business process and other day-to-day transactions. This incorporation of unstructured data and multiple disparate data sets in a single analytical framework is one of the main promises of Big Data.

The Big data specific security and privacy challenges are:

1. Secure computations in distributed programming frameworks.
2. Security best practices for non-real data stores.
3. Secure data storage and tractions logs.
4. End- Point input validation/filtering.
5. Real-time security/compliance monitoring.
6. Scalable and composable and privacy- preserving Data mining and analytics.
7. Cryptographically enforced access control and secure communication.
8. Granular access control.
9. Granular audits.
10. Data provenance
11. Data Integration
12. Data Volume
13. Skills Availability
14. Solution cost
15. Understanding the data
16. Displaying meaningful results
17. Dealing with outliers

1. **Secure computations in distributed programming frameworks**:
   Distributed programming frameworks utilize parallelism in computation and storage to process massive amounts of data. With large data sets, it's next to impossible to identify, resulting in significant damage, especially for scientific and financial computations.
2. **Security best practices for non-real data stores**:
   Non- relational data stores popularized by NoSQL databases are still evolving with respect to security infrastructure. Each NoSQL DBs were built to tackle different challenges posed by the analytic world and hence security was never part of the model at any point of its design stage.
3. **Secure data storage and tractions logs:** Data and transactions logs are stored in multi-tiered storage media. Manually moving data between tiers gives the IT manage direct control over exactly what data is moved and when. As the size of data set has been and continues to be growing exponentially, scalability and availability have necessitated auto-tiering for big data storage management. Auto-tiering solution do not keep track of where the data is stored, which poses new challenges to secure data storage.
4. **End- Point input validation/filtering:** Many Big data use cases enterprise setting require data collection from many sources, such as end-point devices. A key challenge in the data collection process is input validation, how can we trust the data? How can we validate that a source of input data is not malicious and how can we filter malicious input from our collection. Input validation and filtering is a daunting challenge posed by untrusted input sources, especially with bring your model.
5. **Real-time security/compliance monitoring:** Real-time security/ compliance monitoring has always been a challenging, given the number of alerts generated by (Security) devices. These alerts (Correlated or not) lead to many false positives, which are mostly ignored or simply "clicked away", as humans cannot cope with the shear amount. This problem might even increase with big data given the volume and velocity of data streams. Big data technologies might also provide an opportunity, in the sense that these technologies do allow for fast processing and analytics of different types of data. This in its turn can be used to provide, for instance, real-time anomaly detection on scalable security analytics.
6. **Scalable and composable and privacy- preserving Data mining and analytics:** Big data can be seen as a troubling manifestation of Big data Brother by potentially enabling invasions of privacy invasive marking, decrease civil freedoms and increase state and corporate control.
7. **Cryptographically enforced access control and secure communication:** To ensure that the most sensitive privative data is end-to-end secure and only accessible to the authorized entities, data has to be encrypted based on the access control policies. To

ensure authentication, agreement and fairness among the distributed entities, a cryptographically secure communication framework has to be implemented.

8. **Granular access control**: The security property that matters from the perspective of access control is security-preventing access to data by people that should not have access. Big data analysis and cloud computing are increasingly focused on handling diverse data sets, both in terms of variety of schemas and security requirements. Legal and policy restrictions on data come from numerous sources.
9. **Granular audits:** With real time security monitoring, we try to be notified at the moment an attack takes place.
10. **Date Provenance:** Provenance metadata will grow in complexity due to large provenance graphs generated from provenance enabled programming environments in big data applications. Analysis of such large provenance graphs to detect metadata dependencies for security/confidentiality applications is computationally intensive.
11. **Data integration**: The ability to combine data that is not similar in structure or source and to do so quickly and at reasonable cost.
12. **Data volume**: The ability to process the volume at an acceptable speed so that the information is available to decision makers when they need it.
13. **Skills availability**: big data is being harnessed with new tools and is being looked at in different ways. There a shortage of people with the skills to bring together the data, analyze it and publish the results or conclusions.
14. **Solution Cost**: since big data has opened up a world of possible business improvements, there is a great deal of experimentation and discovery taking place to determine the patterns that matter and the insights that turn to value. it is crucial to reduce the cost of the solutions used to find that value.
15. **Understanding the data**: It takes a lot of understanding to get data in the right shape so that we can use visualization as part of data analysis. One solution to this challenge is to have the proper domain expertise in place.
16. **Displaying meaningful results**: Plotting points on a graph for analysis becomes difficult when dealing with extremely large amounts of information or a variety of categories of information. One way to resolve this is to cluster data into a higher-level view where smaller groups of data become visible.
17. **Dealing with outliers**: The graphical representations of data made possible by visualization can communicate trends and outliers much faster than tables containing numbers and text.

CONCLUSION:

There are many different types of data feeds, documents and communications protocols that contain diverse clues to data breaches or ongoing attacks. Users demand analysis of a broader data set, in hope of detecting advanced attacks.

To analyze such a large volume of Big Data analytics is typically performed using specialized software tools and applications for predictive analytics, data mining, text mining, and forecasting and data optimization. Collectively these processes are separate but highly integrated functions of high-performance analytics. Using big data tools and software enables an organization to process extremely large volumes of data that a business has collected to determine which data is relevant and can be analyzed to drive better business decisions in the future.

The goal of Big Data analytics for security is to obtain actionable intelligence in real time. Although Big Data analytics have significant promise, there are a number of challenges that must be overcome to realize its true potential. It is practically impossible to imagine the next application without it consuming data, producing new forms of data, and conducting data driven algorithms. As compute environments become cheaper, applications environments become networked and system and analytics environments become shared over the cloud, security, access control, compression and encryption and compliance introduced challenges that have to be addressed in a systematic way.

REFERENCES

[1] Alperovitch, D.(2011). Revealed: Operation Shady RAT. Santa Clara, CA: McAfee.

[2] Bilge, L. & T. Dumitras. (2012, October) Before We Knew It: An empirical study of zero-day attacks in the real world. Paper presented at the ACM Conference on Computer and Communications Security (CCS), Raleigh, NC.

[3] François, J. et al. (2011, November). BotCloud: Detecting botnets using MapReduce. Paper presented at the Workshop on Information Forensics and Security, Foz do Iguaçu, Brazil.

[4] Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety.

[5] Stamford, CT: META Group. Luckham, D. (2002). The Power of Events. Vol. 204. Addison-Wesley.

[6] Reichman, O.J.; Jones, M.B.; Schildhauer, M.P. (2011). "Challenges and Opportunities of Open Data in Ecology". *Science* 331 (6018): 703–5. doi:10.1126/science.1197962.PMID 21311007

[7] Jump up Hellerstein, Joe (9 November 2008). "Parallel Programming in the Age of Big Data