Software Architectures for Big Data and its applications in Cybersecurity

By – Glorin Sebastian CISSP, CISA

This article looks at the rise of Big Data and its implications in the cybersecurity world.

Big data refers to data that is so large, fast, or complex that it’s difficult or impossible to process using traditional methods [8]. With the increased use of social media and digitization, massive amounts of big data are getting generated daily. This has opened vast opportunities with regards to artificial intelligence and associated applications such as robotics and virtual reality. The proliferation of big data is expected to have a huge impact on the field of cybersecurity. Big data and automation can greatly help increase the efficiency of some of the labor-intensive cybersecurity applications such as log and incident management. This article introduces the reader to some of the software architectures that is being used for big data and expands on the applications of big data in the field of cybersecurity.

Introduction

Increasing digitization within various industries such as finance, healthcare, retail etc. has led to big data revolution. As per a 2018 Forbes article, 2.5 quintillion bytes of data a created each day at our current pace, but that pace is only accelerating with the growth of the Internet of Things (IoT). Over the last two years alone, 90 percent of the data in the world was generated [1]. This big data generated includes structured, semi-structured, and unstructured data collected by organizations that can be mined for information and used in machine learning projects, predictive modeling, and other advanced analytics applications [2].

The main types of big data include:

- **Structured data**: characterized by pre-defined data models, mostly text based and easy to search and reside in relational databases and data warehouses. Examples of structured data include transactions, names, geo locations, phone numbers, etc.
- **Unstructured data**: characterized by no pre-defined data models and reside in applications and data warehouses, Examples of unstructured data include text, video, and multimedia files.
- **Semi-structured data**: characterized by loosely organized meta-level structure with unstructured data in HTML, XML, and JSON formats and are usually stored in relational databases. Examples of semi-structured data include web server logs and streaming data from sensors.

Big data is characterized by the 3V’s—Volume, Velocity, and Variety. Big data also includes a wide variety of data types. Some examples for these data types include transaction processing systems, customer databases, geo locations, emails, medical records, internet clickstream logs, financial transactions, mobile applications, and social networks. It also includes machine-generated data, such as network and server log files, and data from sensors on manufacturing machines, industrial equipment, and IoT devices.

This big data is used by companies to continuously analyze and derive meaningful insights which can improve operations and increase revenue. These insights are further used by companies to better their operations, including marketing, advertising, and promotions to increase customer engagement and conversion rates. Big data is having tremendous impacts on most industries; some of the examples include:

- In the energy industry, big data helps oil and gas firms locate potential drilling locations and monitor pipeline operations; similarly, utility companies use big data analytics to track electrical grids. Big data is also used to analyze seismic and microseismic data, improve reservoir characterization and simulation, reduce drilling time and increase drilling safety [3].
- Financial services firms use big data systems for risk management and real-time analysis of market data. It also provides these firms enhanced capabilities for fraud detection and prevention as well as market trading analysis.
- Manufacturers and transportation companies rely on big data to manage their supply chains and optimize delivery routes, to ensure goods are produced in a timely manner.
and dispatched to consumers via the various distribution channels.

- Government uses include emergency response, crime prevention, and smart city initiatives.
- Big data is transforming healthcare with increased digitization of hospitals and medical care. Some of the best examples of application of big data in healthcare includes use of big data for medical chatbots, medical screening, medical screening, and diagnosis.
- **Cybersecurity:** Big data is of major significance in the field of cybersecurity. With proliferation of IoT networks, systems are generating huge amounts of logs which need to be analyzed to identify malicious users or incidents and patterns.

So, what is it that data analysts do with big data? [4] Data analysts extract, clean, load, and analyze data, while data scientists work on new ways of capturing and analyzing data to be used by the analysts. The main steps followed by data analysts include:

- **Data transformation (extract, transform, load):** ETL is a process that extracts, transforms, and loads data from multiple sources to a data warehouse or other unified data repository. ETL cleanses and organizes data in a way which addresses specific business intelligence needs, like monthly reporting. Several versions of regression analysis and statistical modeling is also performed to ensure the data being loaded is clean and ready to be analyzed. For implementation of the above results, it is required to have proper architecture for the big data system. Listed below are some of the examples of software architectures for big data.

**MapReduce**

MapReduce is the programming framework developed by Google and released in a paper in 2003 which is used for processing large data sets with a parallel, distributed algorithm on a cluster. It helps in the division of work among an arbitrary number of computers. This programming model’s steps are explained below:

- **INPUT:** Starts with a list of (key, value) pairs.
- **SPLITTING:** Applies the ‘map’ function.
- **MAPPING:** Results in a new list of (key, value) pairs.
- **SHUFFLING:** These pairs with the same keys are collected to get (key, list of values).
- **REDUCING:** The ‘reduce’ function produces list of values corresponding to each key

The master node role, called “Task tracker,” does out a list to follower “Job Tracker” node, which does the mapping. The master node collects and aggregates this information. The input is taken and then split and then is applied mapping that transforms the data and then shuffles the data to be sent to different computers. This data is then reduced and recombined. The below diagram illustrates the steps in the MapReduce framework.

The weakness of MapReduce is it is not very good with realtime data processing such as online transaction processing, (OLTP) and online analytical processing (OLAP). Some of the common applications of MapReduce include index building for Google Search, spam detection for emails, digital ad optimization, supply chain analytics including product clustering, and statistical machine translation.

**Hadoop**

Hadoop is an open-source, commonly used system used for MapReduce. Hadoop helps to parallelize tasks very cheaply and can be very redundant. Certain core components are behind the ability of Hadoop to capture as well as manage and process data. The core components of Hadoop include the Hadoop Distributed File System (HDFS), YARN, MapReduce, Hadoop Common, Hadoop Ozone, and Hadoop Submarine. The Hadoop distributed file system, like the name suggests, is the component that is responsible for the basic distribution of data across the system of storage, which is a data node. This component is behind the directory of file storage as well as the file system that directs the storage of data within nodes. Applications run concurrently on the Hadoop framework; the YARN component oversees the framework ensuring that resources are appropriately distributed to running applications. This component of the Hadoop framework is also responsible for creating the schedule of jobs that run concurrently. The MapReduce component of Hadoop tools directs the order of batch applications [5].

**Hadoop Weaknesses**

- No partial data reuse: If you are doing this word count and the data changes, there is no way to be partially done. It should be restarted again.
- Also, it parallelizes very parallel tasks.
- There’s also a fair amount of overhead with every MapReduce job.
- Finally, it is a low-level programming paradigm so it cannot do linear regression, etc.

**Hadoop Applications**

Hadoop has multiple applications for big data and is used in many industries and sectors, some of which are noted below [6]:

- Hadoop is used in many finance applications where it is used for reducing risk, identifying rogue traders, as well as for detecting fraud.
- Big data tools are also used by police for identifying criminals and predicting criminal activity.
• The most important application of Hadoop is understanding customer requirements. Different industries such as retail, finance, etc., use Hadoop for analyzing customer requirements and completing sentiment analysis to understand and predict customer spending.

• Advertising targeting platforms use Hadoop to capture and analyze videos, transactions, and social media data. They also analyze the data generated by various social media websites such as Facebook, Twitter, Instagram, etc., to understand the requirements of their target audience.

• Hadoop is also used to improve personal life. It provides many ways for improving our day-to-day life by monitoring the whole daily routine, sleep pattern, and diet plan of healthy people.

• Hadoop plays a vital role in healthcare sectors for improving public health. It analyzes huge volumes of data from medical devices, lab results, clinical data, imaging reports, etc., to provide support to healthcare sectors to treat their patients efficiently.

**Spark**

Apache Spark is a unified analytics engine for large scale data processing. The paradigm for Spark is a “directed acyclic graph” computations. It has data distributed across nodes like Hadoop; however instead of one programming paradigm, it can do arbitrary programming paradigms with arbitrary dependencies. Also, it has data storage for each of its substeps. Spark has higher level tools and is accessible in many languages.

Spark supports Java, Scala, and Python. R users can access Spark using sparklyr. Data frames are key data constructs for analysis. It is like a spreadsheet and exists in R and Python. Data frames are key data constructs for analysis.

Spark is usually used in scenarios with millions of records, such as the Census bureau data interviews which usually contains at least 1,000 questions with millions of households or insurance companies that have over 50 million people interactions. Some other applications of Spark include:

• **Drone service checking for broken infrastructure:** increasing number of companies are using drone technology for visual inspection, as it is a cost-effective and efficient way to inspect inaccessible areas. The use of drones for inspection of power lines, civil infrastructure, etc., offer advantages such as thorough evaluation of the conditions and detailed photographs of defects. This creates huge amounts of data which can be effectively analyzed using big data frameworks.

• **Google and other web scale companies:** Web scale companies create huge amounts of data which can be effectively analyzed using big data frameworks. The data from various web sources is compartmentalized and used for carrying out data processing operations, which include log analysis, web indexing, data warehousing, financial analysis, scientific simulation, etc.

• **Used by cybersecurity folks:** The applications for big data are noted below. It is to be noted that big data has multiple cybersecurity issues with the main impact being related to data protection and privacy. The study about privacy impacts on big data is still at a nascent stage. The privacy issues on big data can be addressed to a large extent by effective security controls and governance strategy. For example, allocating appropriate data owners and custodians, adopting the right cryptography, etc. However, the discussion of cybersecurity issues associated with big data is outside the scope of this article.

**Big Data Applications in Cybersecurity:**

• **IoT (Internet of Things) device data analysis:** The global IoT market is expected to reach USD 1,463.19 billion by 2027 from USD 250.72 billion in 2019, exhibiting a CAGR of 24.9% during the forecast period [7]. IoT devices create a huge amount of data which could be compromised and needs to be constantly monitored. This big data monitoring for IoT devices helps to check if the given IoT device is being maliciously controlled. Big data platforms like Spark are best suited to address this proliferation of system data from IoT devices, effectively analyzing them to understand system behavior and predict any malicious activities or compromise.

• **System log monitoring and cybersecurity incident analysis:** Big data platforms are also used for analyzing and monitoring the system logs to identify any security incidents which is a violation of computer security policies, acceptable user policies, or standard security practices. Big data analysis would be used in every step of the incident lifecycle including preparation, detection and analysis, containment, eradication, and recovery as well as post incident activities. With the capabilities of big data, the incident analysis and intrusion detection monitoring can also be reported in real time.

• **Intrusion detection predictions:** Big data along with machine learning can analyze historical and current data to make cybersecurity predictions on threat patterns. This approach can help in finding touch points of attackers before any attacks are executed. Moreover, it can also help with real-time responses to data breaches.
Conclusion:

With digitization and widespread adoption of artificial intelligence and virtual reality, the amount of data being produced and consumed is expected to increase exponentially. With improved AI and automation opportunities, big data, coupled with other advances in telecommunications industry such as 5G, is expected to form the backbone of the future economy. This is also going to greatly impact and transform the cybersecurity field, which has been struggling with huge shortages of skilled labor. As discussed in the article, it would specifically improve cybersecurity applications which process huge amounts of log data such as incident management and intrusion detection. Overall, the advances in the field of big data will also have huge impacts and positively impact the adoption of cybersecurity defenses and applications.

References:


About the Author

Glorin is a Senior Cybersecurity consultant at a big-4, with over 7 years of cybersecurity and information risk management experience. Certified CISSP, CISA. Published and also reviewer for international cybersecurity and privacy journals and conferences.