International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 04 | April - 2023 Impact Factor: 8.176 ISSN: 2582-3930

Validating Big Data through Software Testing

Gangula Akshaya, Bandi Thanu Sathwik Reddy, Assoc Prof. Surendra Bandi
Department of Computer Science and Engineering
Hyderabad Institute of Technology and Management, Hyderabad, Telangana, India.

Abstract - Big data has become an indispensable part of many industries due to its numerous benefits, which range from statistical analysis to predictive systems and smart cities. With an overwhelming amount of data generated daily from multiple sources such as the internet, mobile phones, and websites, big data presents a unique set of challenges for testing, as the data is frequently unstructured and highly volatile. To handle the massive amounts of data, complex data combinations, and functional errors that can affect performance and degrade application quality, the testing process for big data applications necessitates specialized knowledge and skills. This poses a significant challenge for organizations that rely on big data for various activities, particularly from a business standpoint. Unlike testing for other types of data, big data testing necessitates a thorough understanding of the technical details and minute changes that occur in seconds during data processing. Big data testing differs from other types of testing due to its complexity.[1]

This paper explores the characteristics of big data, as well as various testing methodologies used for big data, and the challenges that accompany it.

Keywords: Big Data, software testing.

I. INTRODUCTION

Before delving into the complexities of big data testing, let's define the terms "Data" and "Big Data." Data is generated and captured in every aspect of our lives today, from supermarkets and e-commerce websites to web browsing and traffic on the roads. This captured information is in its raw form, and after processing, it becomes data. Big data is more than just data that is bigger than a certain size. The volume of big data, which is typically measured in petabytes, zettabytes, or exabytes, distinguishes it not only by its size but also by its volume. The rapid growth of data and its increasing rate of generation has resulted in a new trend of unavoidable processing, where big data computation frames a new pattern. Discovering knowledge from big data has had a huge impact on businesses, scientific studies, governments, and other fields. Due to the rapidly increasing demand for data and a data-driven mindset, Big Data has become a buzzword that all businesses are considering. It includes initiatives and technologies where data is too fast and changing. Data growth coincides with business expansion in every organization with which data engineering companies work. The primary advantage of big data is that it enables us to identify patterns in data that indicate fraud and aggregate large amounts of information to speed up regulatory reporting.

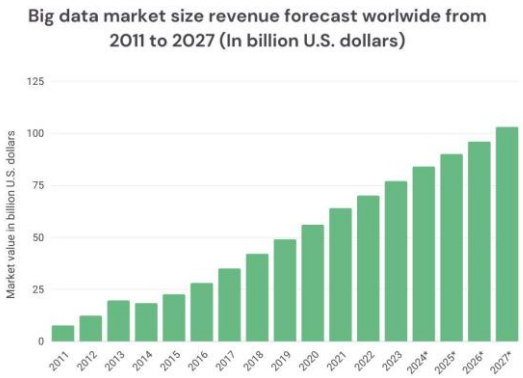

Fig 1. Big Data statistics in recent years

- Since 2012, the rise of Big Data has resulted in the creation of 6 million jobs globally, with approximately 8 million jobs created in the United States alone.
- From 2010 to 2020, the volume of data generated, collected, duplicated, and used worldwide increased significantly, rising in a range of 1.2 trillion gigabytes to fifty nine trillion gigabytes.. This represents a massive 5,000% increase in just a decade.

- By 2020, the global number of smart, networked devices used for data collection, analysis, and sharing would have surpassed 50 billion. In addition, 220 million connected cars were in use that same year.

  Undoubtedly, the big data sector will continue to grow and progress. The increased attention and focus on big data show that it is heading in the right direction. Organizations are expected to fully leverage the potential of big data in the future, resulting in lower business inefficiencies. Today, big data systems are an essential component of the information industry.

  Testing big data presents unique challenges that differ from traditional software evaluation and is critical in addressing the corresponding challenges. The dynamic nature of big data can introduce new challenges, necessitating the exploration of new testing methodologies. The current literature delves into the specifics of big data, the various testing methodologies used to assess it, and the associated challenges.

II. METHODOLOGY

1. Defining Big Data

Big Data refers to huge quantities of records from numerous sources. Traditional tools cannot be used to collect, store, or analyze these data sets because they are too complex.

1.1 Classic V's of Big Data

Big data application systems are software systems that collect, process, analyze, or predict large amounts of data using various platforms, tools, and mechanisms. These applications are closely related to the four main characteristics of big data, which are volume, velocity, variety, and veracity.These are the distinguishing features of Big Data.[2]

- ***Variety*** *(many types of data)*: When big data is received for processing, it can take various forms and formats. This can include various file formats such as.txt,.csv,.xlsx, as well as other types of data such as SMS, audio, video, pdf or other doc formats, and more. Organizations must manage this diversity of data effectively because information can be extracted in a variety of formats. Smart phones transmit a variety of data to network infrastructure, and data collected from surveys, feedback forms, and other sources adds to the vast amount of data that must be accurately analyzed.
- ***Volume*** *(data output)***:** Data is generated from a variety of sources, including machines, networks, and media. These sources provide a massive amount of information that must be extracted and aggregated before it can be used by organizations. Smart phones, for example, transmit a variety of data to network infrastructure, whereas surveys, feedback forms, and other sources collect various types of information. To be useful to organizations, this large and diverse set of data must be thoroughly examined.
- ***Veracity*** *(data uncertainty)***:** The term "veracity" refers to the accuracy and dependability of data, which is crucial in the decision-making process of an organization. With so many data sources available, a large volume of data is generated, making it susceptible to outliers or noise. As a result, the data's nature or function may change.
- ***Velocity*** *(rate of change/data arrival)***:** The velocity characteristic of big data refers to the rate at which data is generated from various sources such as social media, networks, and business operations. This high-speed, real-time data arrives in a continuous stream, and its processing necessitates quick actions. Furthermore, data may change over time, making it more difficult to manage.

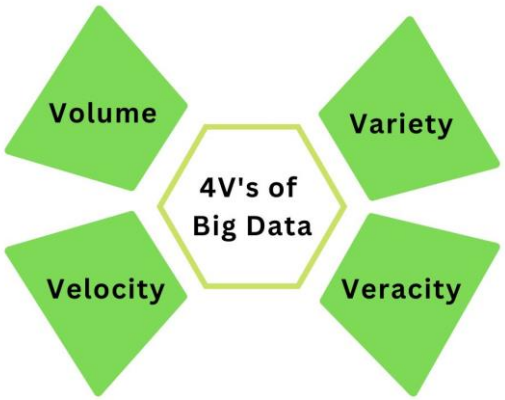

Fig 2. Classic V's of Big Data

1.2 Types of Big Data

Based on the aforementioned characteristics, data comes in various sizes, formats, rates, and so on, giving rise to the following data format categories: structured, unstructured, and semi-structured.[3]

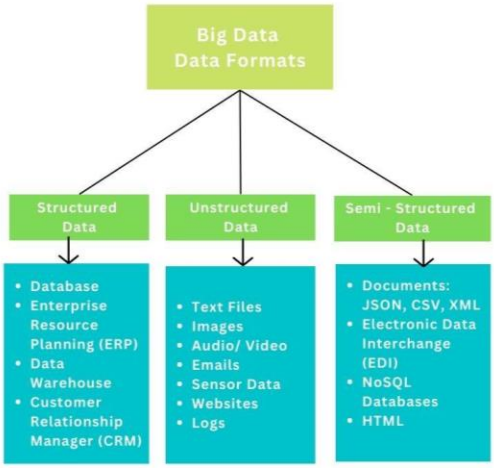

Fig 3. Big Data Classifications

- ***Structured Data***: *(high degree of organization)* Structured data is characterized by a clear and organized design that makes it easy to process and analyze. One example of structured data is the relational database structure, which uses predefined standards for storing and retrieving information. This type of data format includes a relational key and can be easily mapped to pre-designed fields, which facilitates the efficient analysis and processing of data. By conforming to a well-defined structure, structured data can be quickly and accurately analyzed, making it a valuable resource for businesses and organizations.

- ***Unstructured Data***: *(low degree of organization)* Due to the lack of a clear pattern or structure, unstructured data, such as streaming data, web pages, electronic mails, videos, and other forms of unorganized data, can be difficult to analyze and process. As a result, extracting useful information from it necessitates the use of specialized tools and techniques. Identifying hidden patterns within unstructured data can be time-consuming, but it is possible to do so in order to use the data for the intended purpose. Because the majority of real-time data streams are unstructured, processing and analyzing them is a difficult and time-consuming task. However, natural language processing and deep learning, on the other hand, can be used in real-time to extract meaningful insights from large amounts of unstructured data. Overall, while analyzing unstructured data can be difficult, there are a number of tools and techniques available to help extract valuable insights from this type of data.

- ***Semi-Structured Data***: *(partially organized data)* Simple operations like conversions and shifting can help to organize this type of data. This type of data can be handled by a variety of software programmes, including Apache Hadoop. This type of data, which lacks a clear sequence or pattern and is accessible in an unorganized manner, is sometimes referred to as structured data. Tab-delimited text files, CSV files, BibTex files, XML, JSON files, and other markup languages are all examples of structured data. Structured data is less difficult to handle than completely unstructured data because it requires less effort to identify patterns and process information.[4]

2. Testing Big Data

2.1 Need for testing Big Data

Big Data is a collection of large and complex data sets that traditional data processing applications and database management tools cannot process, manage, or capture within a given time frame. The size of these datasets is not fixed and is determined by the size of the organization. Big Data is a broad term that encompasses structured, semi-structured, and unstructured data. When Big Data implementation fails, the consequences can be severe, putting additional strain on the testing team to effectively prepare for a large-scale data testing effort. There are some key points that testers should keep in mind to achieve success in Big Data testing. To begin, testers should understand the significance of Big Data and gain a better understanding of data warehousing and business intelligence testing, as well as the differences between them and Big Data. Second, in order to stay ahead of new technologies, developers and designers should collaborate with testers. We are dealing with unstructured and dynamic data, filing systems, and new concepts such as Hadoop, Cassandra, and others in Big Data testing. To effectively test Big Data, testers must be familiar with the framework being used.[5]

2.2 Methodologies of Big Data Testing

Performance Testing

Giant testing for performance The following activities are incorporated into the data:

Data ingestion and Throughout: The tester checks the system's speed in consuming data from multiple data sources during this activity. Testing entails identifying the various messages that the queue can process in a given amount of time. It also includes determining how quickly data can be inserted into the underlying data store, such as the insertion rate of a Mongo or Cassandra database.

Data Processing: This activity also includes testing the speed at which queries or map-reduce jobs are executed and ensuring that they meet the expected level of speed. It also includes testing the data processing in isolation when the data store is contained within the data sets, such as running Map-Reduce jobs on the underlying HDFS.

Sub-Component Performance: These systems are made up of various components, and each of these components must be validated separately. For example, testing the speed at which messages are listed and consumed, map-reduce jobs, query performance, search functionality, and other system components.

Regression Testing

An essential aspect of maintaining big data systems in the business industry, as well as general data management systems, is diagnosing performance regressions noticed by customers during a production environment. However, when attempting to replicate the problematic behavior in a test environment, database system developers often encounter the issue of missing data. While the database schema and problematic queries can be provided by the customer as part of the regression report, the actual database instance cannot typically be obtained due to privacy restrictions. The database catalog often contains a statistical approximation of the reference database in the form of value distributions, cardinalities, and histograms on columns or column groups. As a fallback solution, developers currently trick the optimizer of a test database by feeding customer catalog data to obtain the query access paths of the actual production system since the underlying data is missing, and the database catalog is often lacking crucial information, such as on multivariate distributions. Synthetic datasets generated in the lab are not representative. The lack of a complete and representative regression database, therefore, slows down the maintenance process and causes additional costs. The methods for data generation based on data and workload characterization, as envisioned in Oligos and Myriad, would provide a solution to this problem.[6]

This testing technique focuses on the errors that will occur as a result of the client's improvement request, bug fix, new feature, or any changes to the application; it may be used by the following techniques:

Reset all: Re-executing all of the tests in the current test suite or bucket is one of the regression testing methods. However, when using a standalone system, this approach can be quite costly. A more promising solution is to run these test cases in parallel on a distributed environment, such as Hadoop.

Regression test selection: This method entails running a subset of the test suite, including reusable and obsolete test cases. The test cases can be prioritized based on the importance of the business impact and the frequency with which they are used. The regression suite can be significantly reduced by selecting test cases based on their priority.

Failover Testing:The purpose of failover testing is to ensure that a system can recover from any type of failure. It involves testing the system's ability to recover data, prevent data corruption, and manage edit logs in the event of critical failures or exceeding performance thresholds. Metrics such as Recovery Point Objective (RPO) and Recovery Time Objective (RTO) can be used to evaluate the effectiveness of the failover test suite (RTO).

2.3 Challenges encountered during Big data testing

Due to the vast and intricate nature of the data sets that comprise Big Data, there is a significant risk of encountering corrupt data and quality issues at each stage of processing when managing and distributing the data across multiple nodes.

There are several requirements like:

- ***Growing requirement to integrate the abundant volume of available data:*** The existence of various data sources has made it essential to enable the integration of all data. However, organizations are compelled to maintain consistently clean and reliable data for this integration, which can be ensured by conducting thorough testing of all available data sources.

- ***Challenges associated with immediate data collection and deployment:*** Data collection and real-time deployment are critical for meeting the business requirements of a company. Obstacles such as accurate data collection, on the other hand, are frequently overcome only by testing the application prior to its live deployment.

- ***Challenges related to the ability to scale in real-time:*** Large-scale data applications are designed to accommodate the level of scalability required for any given situation. However, fundamental errors in the structural components that make up the configuration of Big Data Applications can lead to worst-case scenarios. Therefore, rigorous testing that includes data analysis techniques and exceptional performance testing capabilities is essential for addressing the scalability issues posed by Big Data Applications.

2.4 Big Data Testing Challenges

There are a number of challenges associated with Big Data Platform Testing:

- ***Performance***: Big Data is defined by highly unpredictable and often unstructured information generated from various sources such as weblogs, sensors embedded in devices, GPS systems, and so on. This information is critical for businesses to make informed decisions quickly. However, accessing and analyzing the required information from such a massive volume of data, especially with increased granularity, presents significant challenges.

- ***Scalability***: Scalability refers to a system's ability to handle increasing workloads by expanding the system. However, predicting situations that necessitate scalable systems can be difficult. Workloads can increase as a result of business expansion, new application features, and usage patterns. To process the large volume of data, distributed components of the problem should be distributed to multiple machines or nodes for parallel processing. If necessary, the data should be able to scale quickly across multiple data centers and clouds. When multiple machines work together, the likelihood of failure increases. Failure is the main concern in a multi-machine environment because there is no way for the system to recover if a machine crashes.Furthermore, synchronization between multiple machines continues to be a significant challenge.

- ***Data Security and Continuous Availability***: Big Data frequently contains a large amount of sensitive information, such as personal identifiers, account details, credit card data, and other confidential information. As a result, protecting this sensitive data is critical. However, because of the massive volume of data involved, securing all of this sensitive data can be a daunting task. Many NoSQL Big Data solutions have limited mechanisms for securing Big Data, which presents a significant challenge to ensuring data privacy and security.

- ***Meeting knowledge speed, comprehending it, and addressing data quality***: Extracting relevant information from Big Data can be a difficult task because the data must be in the proper format for effective analysis. Visualization techniques are frequently used to aid in this process. However, ensuring data quality is another major challenge, because even if the data is analyzed quickly, it must be accurate and in the proper context for the intended audience to easily understand it. Poor data quality can have a negative impact on a company's decision-making capabilities.[7]

3. Big Data Validation

Testing Big data is commonly related to varied varieties of testing, for instance, functional testing, performance testing, database testing, and infrastructure testing. Together with these, it's critical to own an unmistakable test plan that allows an easy rendering of massive data testing. When performing big data testing, comprehend that the thought is especially about checking the application’s capability to handle thousands of gigabytes of knowledge. Testing of Big Data for CPS is commonly generally partitioned into three vital stages that incorporate:

Data staging validation: Also observed as a preHadoop stage, the tactic of huge data testing starts with process validation, which aids in guaranteeing if the correct data is pushed into the “Hadoop Distributed classification system (HDFS)”. Validation testing is conducted on data obtained from various sources, such as RDBMS,

webisodes, and social media. Then the information is coordinated with the info utilized within the Hadoop process to test if the 2 coordinate with each other.

MapReduce validation: It refers to the programming concept that enables extensive scalability across numerous servers within a Hadoop cluster. During the testing of Big Data, the second stage of MapReduce involves validation, wherein a tester examines the integrity of the business logic at each join after validating the previous data by running it against various joins.

This aids in assuring that:

- The procedure of MapReduce functions without any flaws.
- Date accumulation or isolation rules are accurately performed on the information.
- Value key sets are produced accurately.

Output Validation: On effectively performing the initial two stages, the last stage is "output validation". It incorporates processed files which are prepared to be passed to an "Enterprise Data Warehouse (EDW)" or another system within the view of particular necessities. The output validation stage incorporates the below-mentioned steps:

- Have to validate change rules are effectively applied.
- Have to validate the info respectability and additionally effective data lading into the resulting organization.
- Guarantee any data defilement by differentiating the target data and also the HDFS file organization data.

III. CONCLUSION

Big data testing is quite different from regular software testing and plays a crucial role to cope up with various corresponding challenges. Conventional software testing methods are inadequate for testing Big Data. Due to the important characteristics that big data exhibit, testing them effectively and generating optimal values is a demanding task. So the best way of doing it is to incorporate data quality maintenance, data sampling technique, and automation of the test suite. Since data is available in different formats which can be structured, semi-structured or unstructured, it requires some pre-processing. With the help of some tools and techniques, we can format the unstructured data into semi-structured data, and furthermore processing can make this semi-structured to structured data. The ever-changing nature of Big Data can pose numerous new challenges, and exploring alternative tools and techniques can be viewed as a potential area for future development.

REFERENCES

[1] Ji, S., Li, Q., Cao, W., Zhang, P. and Muccini, H., 2020. Quality assurance technologies of big data applications: A systematic literature review. *Applied Sciences*, *10*(22), p.8052.

[2] Garg, N., Singla, S. and Jangra, S., 2016. Challenges and techniques for testing of big data. *Procedia Computer Science*, *85*, pp.940-948.

[3] Punn, N.S., Agarwal, S., Syafrullah, M. and Adiyarta, K., 2019, September. Testing big data application. In *2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)* (pp. 159-162). IEEE.

[4] Sneed, H.M. and Erdoes, K., 2015, April. Testing big data (Assuring the quality of large databases). In *2015 IEEE Eighth International Conference on Software Testing, Verification and Validation Workshops (ICSTW)* (pp. 1-6). IEEE.

[5] Staegemann, D., Volk, M., Pohl, M., Häusler, R., Nahhas, A., Abdallah, M. and Turowski, K., 2021. A Preliminary Overview of the Situation in Big Data Testing. *IoTBDS*, pp.296-302.

[6] Alexandrov, A., Brücke, C. and Markl, V., 2013, June. Issues in big data testing and benchmarking. In *Proceedings of the Sixth International Workshop on Testing Database Systems* (pp. 1-5).

[7] Staegemann, D., Volk, M., Nahhas, A., Abdallah, M. and Turowski, K., 2019, November. Exploring the specificities and challenges of testing big data systems. In *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)* (pp. 289-295). IEEE.