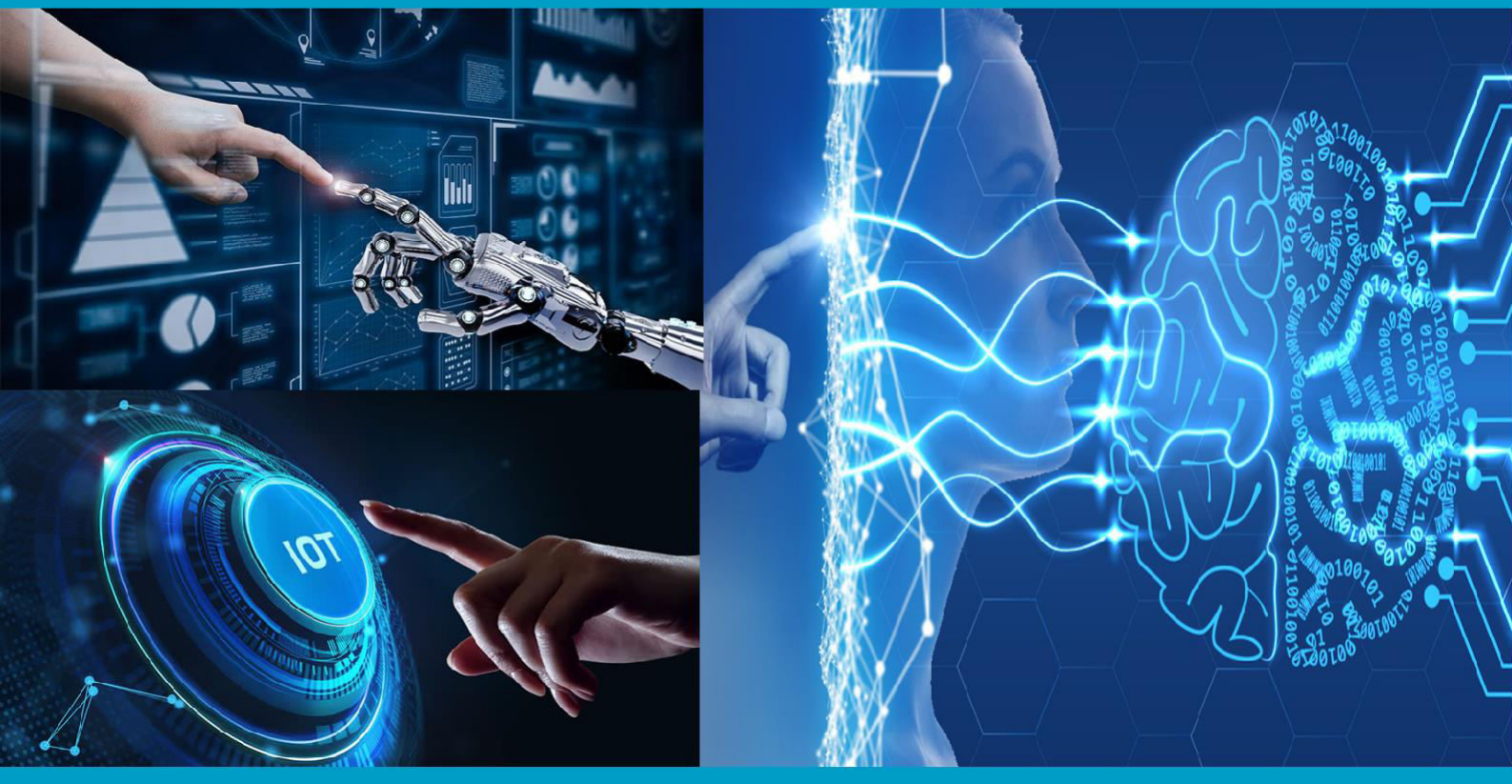
IJMRSETM

e-ISSN: 2395 - 7639



INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

Volume 9, Issue 9, September 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.580

| ISSN: 2395-7639| www.ijmrsetm.com| Impact Factor: 7.580 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 9, Issue 9, September 2022 |

| DOI: 10.15680/IJMRSETM.2022.0909033 |

Handling Quality Challenges Using Big Data Testing Approach

Dr. Archana Verma

Assistant Professor, Computer Science & Engineering, Bipin Tripathi Kumaon Institute of Technology, Dwarahat, Uttarakhand, India

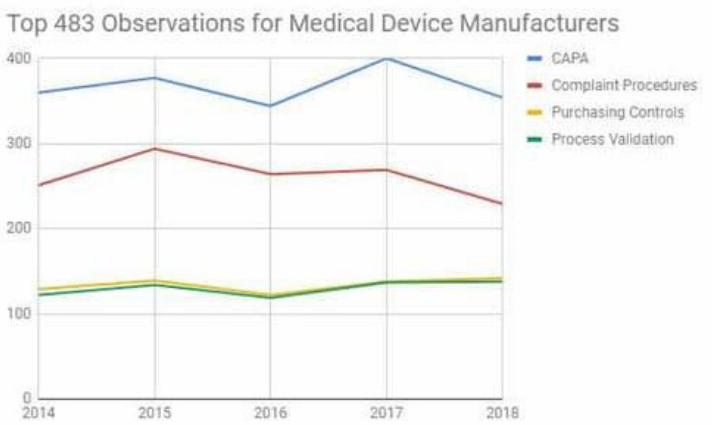
ABSTRACT: Excellence in quality management helps your organization produce the best medical devices on the market. It helps you get your device to market faster and gain a competitive advantage. Quality management can build your brand reputation and keep you from costly customer complaints and product recalls.We've talked to companies of all sizes who got themselves into a bind with poor quality management practices and tools. Sometimes these issues were caused by scaling too quickly. Other times, it was a breakdown of people, processes, or technology. In every case, quality management shouldn't be viewed as an expense. It's an investment into your company's long-term success.Being able to recognize these four common quality management issues empowers you to excel where your competitors fail.

KEYWORDS: quality, challenges, handling, big data, testing, approach, empowers, management

I.INTRODUCTION

How to Avoid the Four Most-Common Quality Management Issues

Inspectors issued 966 483 observations to device manufacturers, including 354 warnings for inadequate Corrective and Preventive Action (CAPA). Inadequate complaint procedures were the second most common with 229 observations. There were 142 instances of purchasing control issues, and 138 observations of insufficient process validation. In total, these four categories comprised 89 percent of 483 warning letters to device manufacturers.[1]



Focusing your quality management efforts on CAPA, complaint handling, purchasing controls, and process validation is a wise investment of your quality management resources. You can gain a regulatory and competitive advantage by avoiding the most common compliance pitfalls that plague the majority of life science organizations who find themselves on the bad list of regulatory bodies. [2]

Problem #1: CAPA

CAPA is the top reason for 483 observations. CAPA quality management issues are generally caused by one of the following quality management snags:

| ISSN: 2395-7639| www.ijmrsetm.com| Impact Factor: 7.580 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 9, Issue 9, September 2022 |

| DOI: 10.15680/IJMRSETM.2022.0909033 |

- The organization isn't following CAPA processes
- The organization's CAPA process doesn't comply with cGMP
- The organization lacks a defined CAPA process or eQMS capabilities[3]

This can occur for many reasons. Fast-growing pharmaceutical startups and scale-ups may delay building a CAPA process during the research & development phase of the product life cycle. A CAPA process may be built as an afterthought, instead of as a core part of the quality management system. In other cases, CAPA issues can arise due to under-involved management, underuse of CAPA, or ineffective root cause determination.[4]

Absent management

CAPA is generally the domain of the quality management team. However, CAPA should not be the responsibility of just the quality team. CAPA is a cross-functional project which should involve the leadership team and all members of the workforce who are involved with the process being investigated. When management isn't highly involved, the quality management team can struggle to understand the bigger picture or create meaningful organizational change.[5]

Underused CAPA

Life sciences organizations need to be judicious about applying CAPA processes. CAPA should be applied to address every breakdown in the quality management system or procedures. If you fail to apply CAPA procedures to issues, you could face major failures. CAPA should be a proactive tool for risk management to mitigate and correct issues before you're facing customer complaints. FDA inspectors evaluate whether organizations consistently apply CAPA to systemic failures.[6]

Missing the root cause

All too often, life sciences organizations conclude CAPA investigations by determining "human error" was the root cause of an issue. While human error may be the true root cause, the leg work to determine exactly why needs to happen—was it a failure of SOPs, training, or management? In other cases, there's even less effort to determine the root. Organizations may identify the symptom of the true issue instead of the real reason the quality management system broke down. If you don't get to the root, you risk facing the same issue repeatedly because the cause hasn't been corrected. Use eQMS data and a structured approach, like Pareto charts, to consistently get to the true basis of CAPA.A CAPA procedure and systems aren't optional.[7] Once these foundational elements are in place, be sure to follow your process and perform the most thorough root cause investigation possible. Don't let your team jump to a conclusion because you will miss the real fix 99% of the time if you rush through CAPA. Ineffective investigations threaten compliance, quality, and your ability to become the market leader.[8]

Problem #2: Complaint Procedures

Creating a defined procedure for handling complaints is a business and regulatory requirement for life sciences organizations. Your procedures must follow the requirements listed in FDA 21 CFR Part 820.198. Some of the most common issues with complaint procedures include a lack of standardized processes and systems, poor management documentation, and ineffective customer care.[9]

No standard procedure

Without a standard approach and system, such as cloud-based QMS software, complaints can easily get lost or overlooked in an email inbox or paper files. Organizations need a standardized procedure for collecting complaint data to follow-up appropriately, and when necessary, escalate complaints to adverse events.[10]

Ineffective management documentation

Documentation is critical to using customer complaints for quality management. Documenting management's response to customer complaints is critical for quality improvement and compliance. Transparent quality management software

| ISSN: 2395-7639| www.ijmrsetm.com| Impact Factor: 7.580 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 9, Issue 9, September 2022 |

| DOI: 10.15680/IJMRSETM.2022.0909033 |

can be crucial to effectively track complaints and resolution. If management can't see the big picture, it's likely leadership will struggle to fully address the root cause of customer grievances.[11]

Customer care

Customer feedback is vital to the success of your organization. If your customers never receive a follow-up or appropriate customer care, your brand satisfaction and reputation metrics can suffer.

A defined procedure is critical, and organizations need a system for collecting and analyzing complaints. Your quality management system should support analysis and comprehensive investigation. You should avoid incomplete analysis of complaints that tie issues to a common root cause. Sometimes, there are nuances that make a new customer complaint different from other complaints in some way. Avoid jumping to conclusions, and ensure your complaint handling process is compliant and viewed as an opportunity for improvement.[12]

Problem #3: Purchasing Controls

"A supply chain is only as strong as its weakest link," said former FDA commissioner Scott Gottlieb. "Every link in the chain must be secure and reliable." The widespread heparin contamination in 2018 serves as a lasting lesson of the importance of purchasing controls. According to a late 2018 statement by Deputy Commissioner Anna Abram, purchasing risks are ever-present and pervasive at every stage of the global supply chain.

Unknown suppliers are a huge risk, according to Abram. FDA investigations have revealed that some manufacturers and compounders are unaware of the true identity of their suppliers, since they source from middlemen. Another risk frequently occurs when suppliers change processes, and this change goes undetected by manufacturers. Ineffective supplier oversight can mean unknown risks when your supplier's quality suddenly drops.

Effective purchasing controls start with a standardized process for qualifying suppliers. Abram urged life sciences organizations to think of supplier relationships as "partnerships," and to view supplier oversight as a lifecycle responsibility.[13]

It's more important than ever to use known suppliers. Ensure if you change suppliers based on price, you're not sacrificing the quality of your APIs or absorbing product risks. Sometimes, saving a few dollars can cost your company tenfold in a recall.

Problem #4: Process Validation

Testing every device or drug isn't possible or practical, which is why organizations test the quality management process instead. Process validation is the act of subjecting a process to such intense evaluation that all outputs are likely to meet design and quality standards every single time.

The US FDA defines process validation as a "means [to] establishing by objective evidence that a process consistently produces a result or product meeting its predetermined specifications." [14] Section 820.75 specifies in part that:

- "The process shall be validated with a high degree of assurance and approved according to established procedures."
- "Each manufacturer shall establish and maintain procedures for monitoring and control of process parameters for validated procedures."

Doing process validation right the first time isn't simple or cheap. However, performing a comprehensive process validation the first time is worth the investment by saving you time and money later. Validation can feel like a waste of precious revenue on your first lots. However, not putting your process through the paces under "actual commercial use conditions" will mean missing out on an opportunity to identify and fix last-minute issues before they become costly production problems.[15]

| ISSN: 2395-7639| www.ijmrsetm.com| Impact Factor: 7.580 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 9, Issue 9, September 2022 |

| DOI: 10.15680/IJMRSETM.2022.0909033 |

"There are process validation issues that have been common this year – they just continuously seem to happen," said FDA investigator Ben Dastoli in a presentation at MedCon 2019 in Cincinnati, Ohio. According to Dastoli, the most common process validation pitfalls include failure to identify all processes which require validation, ignoring variables or worst-case conditions, and insufficient statistical sampling.

Identify all processes which require validation

As FDA investigators walk through a facility, they review manufacturing instructions. Dastgoli has found that gluing, layer-crimping, and reagent mixing are commonly missed processes in pharma. “I always encourage all the managers at a company to walk the entire manufacturing floor,” he says.

Examine variables

Firms don’t always consider variables during validation studies, which means validation is ineffective.[16]

Use sound statistics for sampling

Statistically significant sample sizes are determined by the risk level presented by a failure. FDA inspectors like Dastoli look closely to determine risk during an inspection, and that validation has been repeated enough times for “meaningful and consistent” results.

Test worst-case conditions

Identify variation, reliability, and repeatability across all components of the manufacturing process, including raw materials, components, and equipment.

Address validation failures

“As investigators, we often come across problems that are simply dismissed for reasons such as operator error,” says Dazstoli. If failures arise during the validation process, it’s crucial to investigate the true root cause and take corrective actions, including process revalidation if necessary.[17]

Finding the Best eQMS for Your Company

Quality management issues can have a significant impact on the long-term success of your company. Organizations risk a 483 observation during an FDA inspection, but this isn’t the only potential outcome. Failure to follow CAPA procedures and identify the true root cause of issues can cause repeated problems and avoidable costs. Ineffective complaint handling can jeopardize brand reputation and customer relationships. Purchasing control breakdowns can be linked to unreliable input quality, and speeding through process validation won’t get you trustworthy results.

A strong enterprise quality management system (eQMS) is a necessity to avoid the most common quality management concerns which plague device manufacturers of all sizes. By adopting the right solution during the startup or scale-up phase, you can build your organization on a strong baseline for effective CAPA, supply chain management, and complaint handling. Qualio is the first cloud-based QMS specifically designed for fast-growing life sciences organizations and based on the latest FDA and ISO guidance.[18]

II.DISCUSSION

Earlier, we were only dealing with well-structured data, hosted in large data warehouses, and investing a cost in maintaining those data warehouses and hiring expert professional to maintain and secure information hosted in that data warehouse. Data was structured and can be queried anything as per the needs. But now, this exponential growth of data generates a new vision for data science along with some major challenges.Big Data is something which is growing exponentially with time, and carry raw but very valuable information inside that can change the future of any enterprise. It is a collection, which represents a large dataset, may be collected from multiple sources, or stored in an

| ISSN: 2395-7639| www.ijmrsetm.com| Impact Factor: 7.580 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 9, Issue 9, September 2022 |

| DOI: 10.15680/IJMRSETM.2022.0909033 |

organization. Let‟s understand a real-time example, Big companies like Ikea and Amazon are leveraging the benefit of big data by collecting data from customer‟s buying patterns at their stores, their internal stock information, and their inventory demand-supply relations and analyze all, in seconds even in real-time to add value to its customer experience.

So, extracting information from a large dataset somewhat calls a concept of Data mining which is an analytic process originally designed to explore large datasets. The ultimate goal of data mining is to search for consistency in a pattern or systematic relationship between variables, which helps in predicting the next pattern or behavior.Now, if we take concepts of data mining forward along with large data set, to some extent it becomes a blocker for our existing approach, because big data may contain structured or unstructured data even it may contain data in multiple formats also.[19]

Big Data Testing

Testing is an art of achieving quality in your software product, in terms of perfection of functionality, performance, user experience, or usability. But for big data testing, you need to keep your focus more on the functional and performance aspects of an application. Performance is the key parameter in any big data application which is meant to process terabytes of data. Successful processing of terabytes of data using a commodity cluster with a number of other supportive components needs to be verified. Processing should be faster and accurate which demands a high level of testing.[20]

Processing may be of three types:–

And based on which, we need to integrate different components along with NoSQL data store as per the needs.

Big Data Testing - Test Data

Data plays a vital role in the testing of big data applications. Application is meant to process data and provide an expected output based on implemented logic. The logic needs to be verified before moving to production, as the implementation of logic is completely based on business requirements and data.

1. Test Data Quality
Good quality test data is as important as the test environment. In the big data world, data can have any format or size, it may be in the form of a document, XML, JSON, PDF, etc. at the same time data size may go up to terabytes of petabytes. Hence, test data should also have multiple formats and size should be large enough to ensure the handling of large data processing. In big data testing, it needs data with logical values as per the application requirement and format which is supported by the application.Along with it, data quality is another aspect of big data testing. Ensuring the quality of data before processing through application ensures the accuracy of the final output. Data quality testing itself is a huge domain and covers a lot of best practices which include – data completeness, conformity, accuracy, validity, duplication, and consistency, etc. It should be included in the big data testing and this ensures the level of accuracy application is supposed to provide.[21]

2. Test Data Generation
The generation of test data is again a challenging job, there are multiple parameters, which have to be taken care of while generating test data. It needs a tool, which can help to generate data and should have functions or logic can also be applied over it. Tools like Talend (an open studio) is the best candidate to fulfill the requirements of data generation.

3. Data Storage
After the generation of test data along with quality, it needs to host on a file system. For testing big data applications, data should be stored in the system similar to the production environment. As we are working in big data space, there should have a different number of nodes, and data must be in a distributed environment.

| ISSN: 2395-7639| www.ijmrsetm.com| Impact Factor: 7.580 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 9, Issue 9, September 2022 |

| DOI: 10.15680/IJMRSETM.2022.0909033 |

Big Data Testing - Test Environment

In Big data testing, the test environment should be efficient enough to process a large amount of data as done in the case of a production environment. Real-time production environment clusters generally have 30-40 nodes of cluster and data is distributed on the cluster nodes. There must have some minimum configuration for each node used in the cluster. A cluster may have two modes, in-premise or cloud. For testing in big data, it needs the same kind of environment with some minimum configuration of node.Scalability is also desired to be there in the test environment of big data testing, it helps to study the performance of application with the increase in the number of resources. That data can be used to define SLA (service level agreement) for that particular application.[18]

Big Data Testing can be categorized into three stages:

Step 1: Data Staging Validation
The first stage of big data testing, also known as a Pre-Hadoop stage, is comprised of process validation.

1. Validation of data is very important so that the data collected from various source like RDBMS, weblogs, etc are verified and then added to the system.
2. To ensure data match you should compare source data with the data added to the Hadoop system.
3. Make sure that the right data is taken out and loaded into the accurate HDFS location

Step 2: “Map Reduce” Validation
Validation of “Map Reduce” is the second stage. Business logic validation on every node is performed by the tester. Post that authentication is done by running them against multiple nodes, to make sure that the:

- The process of Map Reduce works perfectly.
- On the data, the data aggregation or segregation rules are imposed.
- Creation of key-value pairs is there.
- After the Map-Reduce process, Data validation is done.

Step 3: Output Validation Phase
The output validation process is the final or third stage involved in big data testing. The output data files are created and they are ready to be moved to an EDW (Enterprise Data Warehouse) or any other such system as per requirements. The third stage consisted of:

- Checking on the transformation rules is accurately applied.
- In the target system, it needs to ensure that data is loaded successfully and the integrity of data is maintained.
- By comparing the target data with the HDFS file system data, it is checked that there is no data corruption.[17]

Big Data - Performance Testing

Big data applications are meant to process a large amount of data, and it is expected that it should take minimum time to process maximum data. Along with it, application jobs should consume a considerable amount of memory and CPU. In big data testing, performance parameter plays an important role and helps to define SLA's. It covers the performance of the base machine and cluster. Also, for example, In the case of Hadoop, map-reduce jobs should be written with proper coding guidelines, to perform better in the production environment. Profiling can also be done on map-reduce jobs before integration, to ensure their optimized execution.

| ISSN: 2395-7639| www.ijmrsetm.com| Impact Factor: 7.580 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 9, Issue 9, September 2022 |

| DOI: 10.15680/IJMRSETM.2022.0909033 |

Tools used in Big Data Scenarios

Big Data Testing - Challenges

In big data testing, certain challenges are involved which needs to be addressed by the big data testing approach.

1. Test Data
Exponential growth had been observed in the growth of data in the last few years. A huge amount of data are being generated daily and stored in large data centers or data marts. So, there is a demand for efficient storage and a way to process it in an optimized way. If we consider the telecom industry, it generates a large number of call logs daily and they need to be processed for better customer experience and compete in the market. The same goes with the test data, test data should be similar to production data and should contain all the logically acceptable fields in it.

This becomes a challenge for testing big data application, generating test data similar to production data is a real challenge. Test data should also be large enough to verify proper working big data application.[16]

2. Environment
The processing of data highly depends on the environment and its performance. An optimized environment setup gives high performance and fast data processing results. Distributed computing is used for the processing of big data which has data hosted in a distributed environment. The testing environment should have multiple numbers of nodes and data should be distributed over the nodes. At the same time, it also needs to monitor those nodes, to ensure the highest performance with minimum CPU and memory utilization. Nodes should be monitored and there should have a graphical presentation of node performance. So, the test environment has two aspects – distributed nodes and their monitoring, which should be covered in the testing approach.

3. Performance
Performance is the key requirement of any big data application, and of course because of which enterprises are moving towards NoSQL technologies, technologies that can handle their big data and process in the minimum time frame. A large dataset should be processed in a minimum considerable time frame. In big data testing, performance testing is a challenge, it requires monitoring of cluster nodes during execution and also time is taken for every iteration of execution.Big Data is the trend that is revolutionizing society and its organizations due to the capabilities it provides to take advantage of a wide variety of data, in large volumes and with speed. Keeping the challenges in mind we have defined the approach of testing big data applications. This approach of big data testing will make it easy for a test engineer to verify and certify the business requirement implementations and for stack holders, it saves a huge amount of cost, which has to be invested to get the expected business returns.

III.RESULTS

Big Data has gone beyond the realms of merely being a buzzword. It is now vastly adopted among companies and corporates, irrespective of size. Even so, many companies are still grappling with the huge plethora of information that's coming their way.Forrester researchers have come up with the eye-opening statistic that Big Data technology has been adopted by 40% of the firms in 2016, and more companies (prediction says 30%) are joining the bandwagon in the next 12 months.New data technologies are entering the market, but the old ones are also going strong. However, it is important to note here that the adoption of Big Data technologies will never slow down, at least not anytime in the near future. So it is important to note how we can overcome the challenges that come with this technology. This article sheds more light on this aspect.With the huge onslaught of data that's coming in, businesses have tried different methods to handle them all. Conventional database systems have been replaced with horizontal database, columnar designs and cloud enabled schemas. Even so, the role of quality analysis is still teetering on toddler legs because, in order to Big Data applications, you need a specific mindset, skills and knowledge, followed by a knowledge of data science.[15]

| ISSN: 2395-7639| www.ijmrsetm.com| Impact Factor: 7.580 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 9, Issue 9, September 2022 |

| DOI: 10.15680/IJMRSETM.2022.0909033 |

As per Gartner, "Big Data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."
In simple words, Big Data refers to huge quantities of data. True, there is no particular size parameter to define the size of this technology. It is safe to assume that the standard way to measure it comes in terabytes or petabytes. Data comes in from all directions, and the volume and velocity would be monstrous.Data gets replaced at a rapid pace and hence the need for processing becomes higher, especially when it is related to social media feeds. But that is not the only medium through which information comes in. It could come from a variety of sources and number of formats. If you check a data repository, you can see text files, images, audio files, presentations, video files, spreadsheets, databases and email messages. Depending on the requirement, the format and structure for the data could vary.The data collected from mobile devices, digital repositories and social media can be unstructured and structured. Structure data, of course, is very easy to analyze. Unstructured data like voice, emails, video and documents are difficult to analyze and take up a lot of resources.One of the most reliable solutions for database management system is the relational database management system or RDBMS, the leading players in the solution being Microsoft, IBM and Oracle. RDBMS uses structured query language or SQL to manage the data, define, query and update it.Still, when the data size is tremendous, RDBMS finds it difficult to handle, and even if it does, the process will be very expensive. This proves that relational databases do not have the ability to handle big data, and new technologies are required. Traditional databases are perfect for structured data, but not for unstructured data.[17]

Big Data Characteristics and Data Formats

The wave of information that keeps coming in is characterised by the 3V's:

- Volume
- Variety and
- Velocity

Volume

Volume is perhaps the most associated V with respect to big data. Here, you deal with incomprehensible options of data. An example would be Facebook with its immense capacity to store photos. It is believed that Facebook is storing more than 250 billion images (as of 2015 records), and the numbers will only keep growing. 250 billion images a day is just a drop in the bucket. Considering the fact that people upload more than 900 million photos a day, the numbers are staggering.And in YouTube, about 300 hours worth of videos are uploaded since almost everybody has a smartphone these days. Vendors have begun to manage their app data in the cloud, and SaaS app vendors have huge amounts of data to store.

Velocity

Velocity is all about the measure of data that's coming in. Take the example of Facebook. Look at the amount of photos it has to process, store and then retrieve. In the initial stages of data acquiring, companies used to do batch processing of the information that's coming in.The data is fed into the server, and then it waits for the result. However, this can work only when the incoming data is slow, and with the speed at which data can be acquired now, it is possible to have a waiting period, however short that maybe. Information keeps coming in from all directions in real time, and you need to make them coherent enough to analyze and draw conclusions.[18]

Variety

There are all kinds of data - structured, unstructured and semi-structured. You will have photographs, encrypted packets, tweets, sensor data and plenty more of this kind. Data is not something given to you in a spreadsheet any more. When data comes in the structured format, things become easier, but when they come in the form of photos, videos,

| ISSN: 2395-7639| www.ijmrsetm.com| Impact Factor: 7.580 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 9, Issue 9, September 2022 |

| DOI: 10.15680/IJMRSETM.2022.0909033 |

audio recordings, books, geospatial data, presentations, email messages, posts & comments, documents, tweets and ECG strips, you have them unstructured and overflowing.Finding a pattern from this insane flow of data is not easy, and the process to make them coherent is known as data analytics. And each of the data collected would require particular technologies and analytic methods to get a clear picture of what they indicate, making the information valuable.As explained earlier, data is something that provides information and this is used for analysis and drawing conclusions. Data comes in different sizes, formats, etc. Hence, the three different categories:

- Structured
- Unstructured and
- Semi-structured data

Structured data

Structured data can be easily utilised, because they do have a definite structure and are well organised. An example would be a spreadsheet where information is available in a tabulated form. Identifying patterns and extracting useful and actionable insights would be easier. The data is also stored in a relational DB, and pulled easily.Understandably, structured data is the most processed of all information and can be managed easily. They have a relational key and can be easily mapped into predesigned fields.[19]

Unstructured data

They are huge amounts of data stored in no particular pattern. But to gauge an understandable pattern from it, you need the help of sophisticated tools and technologies. Examples would be videos, web pages, emails, PowerPoint presentations, location coordinates, streaming data, etc. This lack of structure makes it difficult to manage the data from a relational DB.80% of the data that's found online is unstructured. However, these kinds of data do have some sort of internal structure, but they don't neatly fit in any database. Exploiting unstructured data to its full advantage would help you make critical business decisions.[20]

Semi Structured data

This kind of data is not rigidly organised and can be utilised after a little bit of sifting, processing & conversion. Softwares like Apache Hadoop are used for this. However, these are not stored in the relational DB. In fact, semi structured data can be called structured data that is available in an unorganised manner.Examples of such kind of information comes in the form of web data such JSON (JavaScript Object Notation) files, tab-delimited text files, .csv files, BibTex files, XML and other markup languages. Having a semi-structure makes it easier to ease up space, clarify and compute data. They come with organisational properties that makes analysis easy.

Big Data Application Testing – Key Components

As Big Data is described through the above-mentioned three Vs, you need to know how to process all this data through its various formats at high speed. This processing can be split into three basic components. To be successful, testers will have to be aware of these components.

1. Data Validation: Understandably, this is one of the most important components of data collection. To ensure the data is not corrupted or is accurate, it is important that it is validated. For this purpose, the sources will be checked. The information procured is validated against actual business requirements. The initial data will be fed into Hadoop Distributed File System (HDFS), and this will also be validated. The file partition will be checked thoroughly, followed by copying them into different data units. Tools like Datameer, Talent and Informatica are used for step-by-step validation.

| ISSN: 2395-7639| www.ijmrsetm.com| Impact Factor: 7.580 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 9, Issue 9, September 2022 |

| DOI: 10.15680/IJMRSETM.2022.0909033 |

Data validation is also known as pre-Hadoop testing, and makes it certain that the collected data is from the right resources. Once that step is completed, it is then pushed into the Hadoop testing system for tallying with the source data.

2. Process Validation: Once the data and the source are matched, they will be pushed to the right location. This would be the Business Logic validation or Process Validation, where the tester will verify the business logic, node by node, and then verify it against different nodes. Business Logic Validation is the validation of Map Reduce, the heart of Hadoop.
The tester will validate the Map-Reduce process and check if the key-value pair is generated correctly. Through "reduce" operation, the aggregation and consolidation of data is checked out.

3. Output Validation: Output validation is the next important component. Here the generated data is loaded into the downstream system. This could be a data repository and the data goes through analysis and further processing. This is then further checked to make sure the data is not distorted, by comparing HDFS file system with target data.
Architecture testing is another crucial part of Big Data testing, as having poor architecture will make the whole effort go wasted. Luckily, Hadoop is highly resource intensive, and is capable of processing huge amounts of data and for this, architectural testing becomes mandatory. It is also important to ensure that there is no data corruption, and compare the HDFS file system data with target UI or business intelligence system.

ETL Testing

ELT is the acronym for Extract, Transform and Load, and has been around for a very long time because it is associated with traditional batch processing in data environment. The function of data warehouses is to provide businesses with data that they can consolidate, analyze and make coherent ideas of data that is relevant to their focus/goals. There are ETL tools through which the raw data is converted into a meaningful format. The tool also helps them to convert the data into a format that could be used by the businesses. Software vendors like IBM, Pervasive, Pentaho, etc. provide ETL software tools.[21]

Extract - Once the data is collected, it would be extracted/read from the source database. This is done to all the databases.Transform - Transformation of the data is done next. The data format is changed into usable chunks, and must conform to the requirements of the target database.Load - This is the final stage where you write data to the target database.To ensure that the data procured in this manner is trustworthy, tools for data integration processes like data profiling, cleansing, auditing are all integrated with data quality tools. This entire process will ensure that you have extracted actual data. ETL tools are also important for loading and converting both structured and unstructured data into Hadoop. It also depends on the kind of ETL tools that you use. Highly advanced ones lets you convert multiple simultaneously.The data processing segment in a data warehouse, follows a three-layer architecture during ETL process.

Data Warehouse Staging Layer

The staging area is a temporary location or a landing zone where data from all the resources is stored. This zone ensures that all the data is available before they are integrated into the Data Warehouse. It is imperative for the data to be placed somewhere because of varying business cycles, hardware limitations, network resource limitations and data processing cycles. You cannot extract all the data from all the databases at the same time. Hence, data in the data warehouse is transient.

Data Integration Layer

This is the foundation of next-generation analytics and contributes to business intelligence. The data integration layer is a combination of semantic, reporting and analytical technologies, based on the semantic knowledge framework. Data is

| ISSN: 2395-7639| www.ijmrsetm.com| Impact Factor: 7.580 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 9, Issue 9, September 2022 |

| DOI: 10.15680/IJMRSETM.2022.0909033 |

arranged in hierarchical groups known as facts and converted into facts and aggregated facts. This layer is the link between the staging layer and the database.

Access Layer

Using common business terms, users will be able to access the data from the warehouse. The Access Layer is what the users can access, and the users themselves know what to make of the data. It is almost like a virtual layer, because it doesn't store information. The layer contains data targeted to a specific population, making access and usage easier.[20]

Benefits of Using Big Data Testing

Through Big Data testing, you can ensure the data in hand is qualitative, accurate and healthy. The data you had collected from different sources and channels are validated, aiding in better decision making. There are several benefits to Big Data testing.Better Decision Making - When data gets in the hands of the right people, it becomes an asset. So when you have the right kind of data in hand, it would help you make sound decisions. It lets you analyse all the risks and make use of only the data that will contribute to the decision making process.Data Accuracy - Gartner says that data volume is likely to expand by 800% in the next 5 years, and 80% of this data will be unstructured. Imagine the volume of data that you have to analyse. You need to convert all this data into a structured format before it can be mined. Armed with the right kind of data, businesses can focus on their weak areas, and be better prepared to beat the competition.Better Strategy and Enhanced Market Goals - You can chart a better decision making strategy or automate the decision making process with the help of big data. Collect all the validated data, analyze it, understand user behavior and ensure all of them are realised in the software testing process, so you can deal out something they need. Big data testing helps you optimise business strategies by looking at this information.Increased Profit and Reduced Loss - Loss in business will be minimal or even a thing of past, if data is correctly analysed. If the accumulated data is of poor quality, the business suffers terrible losses. Isolate valuable data from structured and semi-structured information so no mistakes are made when dealing with customers.[21]

IV. CONCLUSIONS

Transforming data with intelligence is a huge concern. As big data is integral to a company's decision making strategy, it is not even possible to begin asserting the importance of arming yourself with reliable information.Big Data processing is a very promising field in today's complex business environment. Applying the right dose of test strategies, and following best practices would help ensure qualitative software testing. The idea is to recognise and identify the defects in the early stages of testing and rectify them. This helps in cost reduction and better realization of company goals. Through this process, the problems that testers faced during software testing are all solved now because the testing approaches are all driven by data.[21]

REFERENCES

[1]. Rokke, C. and O. Yadav. Challenges and Barriers to Total Quality Management: An Overview. International Journal of Performability Engineering, 2012; 8(6): 653-665. 55
[2]. Lemak, D., R. Reed, and P. Satish. Commitment to Total Quality Management: Is There a Relationship with Firm Performance? Journal of Quality Management, 1997; 2(1): 67- 86.
[3]. Goodman, J., G. Bargatze, and C. Grimm. The Key Problem with TQM. Quality Progress, 1994; January: 45-48.
[4]. Kanji, G. Implementation and pitfalls of total quality management. Total Quality Management, 1996; 7(3): 331-343.
[5]. Adeoti, J. Total Quality Management (TQM) Factors: An Empirical Study of Kwara State Government Hospitals. Ethno Med, 2011; 5(1): 17-23
[6]. Tamimi, N., and R. Sebastianelli. The Barriers to Total Quality Management. Quality Progress, 1998; June: 57-60.
[7]. Oakland, J. Interdependence and cooperation: The essentials of total quality management. Total Quality Management, 1997; 8(2&3): S31-S35.
[8]. Masters, R. Overcoming the barriers to TQM"s success. Quality Progress, 1996; 29(5): 53.

| ISSN: 2395-7639| www.ijmrsetm.com| Impact Factor: 7.580 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 9, Issue 9, September 2022 |

| DOI: 10.15680/IJMRSETM.2022.0909033 |

[9]. Salegna, G. and F. Fazel. Obstacles to Implementing Quality. Quality Progress, 2000; July: 53-57.
[10]. Coulson-Thomas, C. Quality: Where Do We Go from Here? The International Journal of Quality & Reliability Management, 1992; 9(1): 38-56.
[11]. Lakhe, R. and R. Mohanty. Understanding TQM. Production Planning & Control, 1994; 5(5): 426-441
[12]. Grant, R., R. Shani, and R. Krishnan. TQM‟s Challenge to Management Theory and Practice. Sloan Management Review, 1994; 35(2): 25-35.
[13]. Reger, R., L. Gustafson, S. Demarie, and J. Mullane. Reframing the Organization: Why Implementing Total Quality is Easier Said than Done. Academy of Management Review, 1994; 19(3): 565-584.
[14]. Fok, L., S. Hartman, A. Patty, J. Razek. Human factors affecting the acceptance of total quality management. The International Journal of Quality & Reliability Management, 2000; 17(7): 714-729.
[15]. Sebastianelli, R. and N. Tamimi. Understanding the Obstacles to TQM Success. The Quality Management Journal, 2003; 10(3): 45-56.
[16]. Burli, S., V. Bagodi, and B. Kottursheṭtar. TQM dimensions and their interrelationships in ISO certified engineering institutes of India. Emerald Publishing Group Limited, 2012: 1-17.
[17]. Cheng, T. and E. Ngai. Identifying potential barriers to total quality management using principal component analysis and correspondence analysis. International Journal of Quality & Reliability Management, 1997; 14(4-5): 391.
[18]. Kotter, J. Leading Change: Why Transformation Efforts Fail. Harvard Business Review, March-April 1995; 59-67.
[19]. Wayhan, V. and E. Balderson. TQM and Financial Performance: A Research Standard. Total Quality Management, 2007; 18(4): 393-401.
[20]. Oakland, J. Leadership and Policy Deployment: The Backbone of TQM. Total Quality Management & Business Excellence, 2011; 22(5): 517-534.
[21]. Fuchsberg, G. Management: Baldrige Awards May Be Losing Some Luster. The Wall Street Journal, April 1992; B.1-B.2.

Impact Factor: 7.580

INTERNATIONAL JOURNAL
OF MULTIDISCIPLINARY RESEARCH
IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

+91 99405 72462 +91 63819 07438 ijmrsetm@gmail.com