Cosmos Multidisciplinary Research E-Journal
Online Available at www.cmrj.in
Recognized International Peer Reviewed Journal
Impact Factor 4.94
ISSN No. 2456-1665

Big Data Analytics

Rahul Laxmikant Gajre
Research Student

Abstract -

A huge repository of terabytes of data is generated each day from modern information systems and digital technologies such as Internet of Things and cloud computing. This enormous amounts of data have become available on hand to decision makers. Big data is originally associated with three key concepts: volume, variety, and velocity. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets. Furthermore, decision makers need to be able to gain valuable insights from such varied and rapidly changing data, ranging from daily transactions to customer interactions and social network data. Such value can be provided using big data analytics, which is the application of advanced analytics techniques on big data. Current usage of the term big data tends to refer to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from big data, and occasionally to a particular size of data set. Data with many fields (columns) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate. Big data is a field that treats ways to analyze, systematically extract information from or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.

INTRODUCTION -

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Simply definition of big data is data that contains greater variety, arriving in increasing volumes and with more velocity. This is also known as the three Vs. Put simply, big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can't manage them. But these massive volumes of data can be used to address business problems we wouldn't have been able to tackle before.

The three Vs of big data

Volume	The amount of data matters. With big data, you'll have to process high volumes of low-density, unstructured data. This can be data of unknown value, such as Twitter data feeds, clickstreams on a web page or a mobile app, or sensor-enabled equipment. For some organizations, this might be tens of terabytes of data. For others, it may be hundreds of petabytes.
Velocity	Velocity is the fast rate at which data is received and (perhaps) acted on. Normally, the highest velocity of data streams directly into memory versus being written to disk. Some internet-enabled smart products operate in real time or near real time and will require real-time evaluation and action.

Variety	Variety refers to the many types of data that are available. Traditional data types were structured and fit neatly in a relational database. With the rise of big data, data comes in new unstructured data types. Unstructured and semistructured data types, such as text, audio, and video, require additional preprocessing to derive meaning and support metadata.

The value—and truth—of big data

Two more Vs have emerged over the past few years: value and veracity. Data has intrinsic value. But it's of no use until that value is discovered. Equally important: How truthful is your data—and how much can you rely on it? Today, big data has become capital. Think of some of the world's biggest tech companies. A large part of the value they offer comes from their data, which they're constantly analyzing to produce more efficiency and develop new products.Recent technological breakthroughs have exponentially reduced the cost of data storage and compute, making it easier and less expensive to store more data than ever before. With an increased volume of big data now cheaper and more accessible, you can make more accurate and precise business decisions. Finding value in big data isn't only about analyzing it (which is a whole other benefit). It's an entire discovery process that requires insightful analysts, business users, and executives who ask the right questions, recognize patterns, make informed assumptions, and predict behavior.

Big data benefits:

- Big data makes it possible for you to gain more complete answers because you have more information.
- More complete answers mean more confidence in the data—which means a completely different approach to tackling problems.

Big data uses -

Companies use big data in their systems to improve operations, provide better customer service, create personalized marketing campaigns and take other actions that, ultimately, can increase revenue and profits. Businesses that use it effectively hold a potential competitive advantage over those that don't because they're able to make faster and more informed business decisions.

For example, big data provides valuable insights into customers that companies can use to refine their marketing, advertising and promotions in order to increase customer engagement and conversion rates. Both historical and real-time data can be analyzed to assess the evolving preferences of consumers or corporate buyers, enabling businesses to become more responsive to customer wants and needs.

Big data is also used by medical researchers to identify disease signs and risk factors and by doctors to help diagnose illnesses and medical conditions in patients. In addition, a combination of data from electronic health records, social media sites, the web and other sources gives healthcare organizations and government agencies up-to-date information on infectious disease threats or outbreaks.

Here are some more examples of how big data is used by organizations:

- In the energy industry, big data helps oil and gas companies identify potential drilling locations and monitor pipeline operations; likewise, utilities use it to track electrical grids.
- Financial services firms use big data systems for risk management and real-time analysis of market data.

- Manufacturers and transportation companies rely on big data to manage their supply chains and optimize delivery routes.
- Other government uses include emergency response, crime prevention and smart city initiatives.

Big data can help you address a range of business activities, from customer experience to analytic, more specifically they are as

Product development	Companies like Netflix and Procter & Gamble use big data to anticipate customer demand. They build predictive models for new products and services by classifying key attributes of past and current products or services and modelling the relationship between those attributes and the commercial success of the offerings. In addition, P&G uses data and analytics from focus groups, social media, test markets, and early store rollouts to plan, produce, and launch new products.
Predictive maintenance	Factors that can predict mechanical failures may be deeply buried in structured data, such as the year, make, and model of equipment, as well as in unstructured data that covers millions of log entries, sensor data, error messages, and engine temperature. By analyzing these indications of potential issues before the problems happen, organizations can deploy maintenance more cost effectively and maximize parts and equipment uptime.
Customer experience	The race for customers is on. A clearer view of customer experience is more possible now than ever before. Big data enables you to gather data from social media, web visits, call logs, and other sources to improve the interaction experience and maximize the value delivered. Start delivering personalized offers, reduce customer churn, and handle issues proactively.
Fraud and compliance	When it comes to security, it's not just a few rogue hackers—you're up against entire expert teams. Security landscapes and compliance requirements are constantly evolving. Big data helps you identify patterns in data that indicate fraud and aggregate large volume.
Machine learning	Machine learning is a hot topic right now. And data—specifically big data—is one of the reasons why. We are now able to teach machines instead of program them. The availability of big data to train machine learning models makes that possible.
Operational efficiency	Operational efficiency may not always make the news, but it's an area in which big data is having the most impact. With big data, you can analyze and assess production, customer feedback and returns, and other factors to reduce outages and anticipate future demands. Big data can also be used to improve decision-making in line with current market demand.
Drive innovation	Big data can help you innovate by studying interdependencies among humans, institutions, entities, and process and then determining new ways to use those insights. Use data insights to improve decisions about financial and planning considerations. Examine trends and what customers want to deliver new products and services. Implement dynamic pricing. There are endless possibilities.

So in short 6 big data benefits for businesses are

1) Better customer insight.
2) Improved operations.
3) More insightful market intelligence.
4) Agile supply chain management.
5) Data-driven innovation.
6) Smarter recommendations and targeting.

How big data is stored and processed

Big data is often stored in a data lake. While data warehouses are commonly built on relational databases and contain structured data only, data lakes can support various data types and typically are based on Hadoop clusters, cloud object storage services, NoSQL databases or other big data platforms.

Many big data environments combine multiple systems in a distributed architecture; for example, a central data lake might be integrated with other platforms, including relational databases or a data warehouse. The data in big data systems may be left in its raw form and then filtered and organized as needed for particular analytics uses. In other cases, it's preprocessed using data mining tools and data preparation software so it's ready for applications that are run regularly.

Big data processing places heavy demands on the underlying compute infrastructure. The required computing power often is provided by clustered systems that distribute processing workloads across hundreds or thousands of commodity servers, using technologies like Hadoop and the Spark processing engine.

Getting that kind of processing capacity in a cost-effective way is a challenge. As a result, the cloud is a popular location for big data systems. Organizations can deploy their own cloud-based systems or use managed big-data-as-a-service offerings from cloud providers. Cloud users can scale up the required number of servers just long enough to complete big data analytics projects. The business only pays for the storage and compute time it uses, and the cloud instances can be turned off until they're needed again.

How big data analytics works

To get valid and relevant results from big data analytics applications, data scientists and other data analysts must have a detailed understanding of the available data and a sense of what they're looking for in it. That makes data preparation, which includes profiling, cleansing, validation and transformation of data sets, a crucial first step in the analytics process.

Once the data has been gathered and prepared for analysis, various data science and advanced analytics disciplines can be applied to run different applications, using tools that provide big data analytics features and capabilities. Those disciplines include machine learning and its deep learning offshoot, predictive modelling, data mining, statistical analysis, streaming analytics, text mining and more.

Using customer data as an example, the different branches of analytics that can be done with sets of big data include the following:

- **Comparative analysis.** This examines customer behavior metrics and real-time customer engagement in order to compare a company's products, services and branding with those of its competitors.
- **Social media listening.** This analyzes what people are saying on social media about a business or product, which can help identify potential problems and target audiences for marketing campaigns.

- **Marketing analytics.** This provides information that can be used to improve marketing campaigns and promotional offers for products, services and business initiatives.
- **Sentiment analysis.** All of the data that's gathered on customers can be analyzed to reveal how they feel about a company or brand, customer satisfaction levels, potential issues and how customer service could be improved.

Big data management technologies

Hadoop, an open source distributed processing framework released in 2006, initially was at the center of most big data architectures. The development of Spark and other processing engines pushed MapReduce, the engine built into Hadoop, more to the side. The result is an ecosystem of big data technologies that can be used for different applications but often are deployed together.

Big data platforms and managed services offered by IT vendors combine many of those technologies in a single package, primarily for use in the cloud. Currently, that includes these offerings, listed alphabetically:

- Amazon EMR (formerly Elastic MapReduce)
- Cloudera Data Platform
- Google Cloud Dataproc
- HPE Ezmeral Data Fabric (formerly MapR Data Platform)
- Microsoft Azure HDInsight

For organizations that want to deploy big data systems themselves, either on premises or in the cloud, the technologies that are available to them in addition to Hadoop and Spark include the following categories of tools:

- storage repositories, such as the Hadoop Distributed File System (HDFS) and cloud object storage services that include Amazon Simple Storage Service (S3), Google Cloud Storage and Azure Blob Storage;
- cluster management frameworks, like Kubernetes, Mesos and YARN, Hadoop's built-in resource manager and job scheduler, which stands for Yet Another Resource Negotiator but is commonly known by the acronym alone;
- stream processing engines, such as Flink, Hudi, Kafka, Samza, Storm and the Spark Streaming and Structured Streaming modules built into Spark;
- NoSQL databases that include Cassandra, Couchbase, CouchDB, HBase, MarkLogic Data Hub, MongoDB, Neo4j, Redis and various other technologies;
- data lake and data warehouse platforms, among them Amazon Redshift, Delta Lake, Google BigQuery, Kylin and Snowflake; and
- SQL query engines, like Drill, Hive, Impala, Presto and Trino.

Big data challenges

In connection with the processing capacity issues, designing a big data architecture is a common challenge for users. Big data systems must be tailored to an organization's particular needs, a DIY undertaking that requires IT and data management teams to piece together a customized set of technologies and tools. Deploying and managing big data systems also require new skills compared to the ones that database administrators and developers focused on relational software typically possess.

Both of those issues can be eased by using a managed cloud service, but IT managers need to keep a close eye on cloud usage to make sure costs don't get out of hand. Also,

migrating on-premises data sets and processing workloads to the cloud is often a complex process.

Other challenges in managing big data systems include making the data accessible to data scientists and analysts, especially in distributed environments that include a mix of different platforms and data stores. To help analysts find relevant data, data management and analytics teams are increasingly building data catalogs that incorporate metadata management and data lineage functions. The process of integrating sets of big data is often also complicated, particularly when data variety and velocity are factors.

Keys to an effective big data strategy

In an organization, developing a big data strategy requires an understanding of business goals and the data that's currently available to use, plus an assessment of the need for additional data to help meet the objectives. The next steps to take include the following:

- prioritizing planned use cases and applications;
- identifying new systems and tools that are needed;
- creating a deployment roadmap; and
- evaluating internal skills to see if retraining or hiring are required.

To ensure that sets of big data are clean, consistent and used properly, a data governance program and associated data quality management processes also must be priorities. Other best practices for managing and analyzing big data include focusing on business needs for information over the available technologies and using data visualization to aid in data discovery and analysis.

Big data collection practices and regulations

As the collection and use of big data have increased, so has the potential for data misuse. A public outcry about data breaches and other personal privacy violations led the European Union to approve the General Data Protection Regulation (GDPR), a data privacy law that took effect in May 2018. GDPR limits the types of data that organizations can collect and requires opt-in consent from individuals or compliance with other specified reasons for collecting personal data. It also includes a right-to-be-forgotten provision, which lets EU residents ask companies to delete their data.

While there aren't similar federal laws in the U.S., the California Consumer Privacy Act (CCPA) aims to give California residents more control over the collection and use of their personal information by companies that do business in the state. CCPA was signed into law in 2018 and took effect on Jan. 1, 2020.

To ensure that they comply with such laws, businesses need to carefully manage the process of collecting big data. Controls must be put in place to identify regulated data and prevent unauthorized employees from accessing it.

The human side of big data management and analytics

Ultimately, the business value and benefits of big data initiatives depend on the workers tasked with managing and analyzing the data. Some big data tools enable less technical users to run predictive analytics applications or help businesses deploy a suitable infrastructure for big data projects, while minimizing the need for hardware and distributed software know-how. Big data can be contrasted with small data, a term that's sometimes used to describe data sets that can be easily used for self-service BI and analytics. A commonly quoted axiom is, "Big data is for machines; small data is for people."

Big data challenges

While big data holds a lot of promise, it is not without its challenges. First, big data is…big. Although new technologies have been developed for data storage, data volumes are doubling in size about every two years. Organizations still struggle to keep pace with their data and find ways to effectively store it. But it's not enough to just store the data. Data must be used to be valuable and that depends on curation. Clean data, or data that's relevant to the client and organized in a way that enables meaningful analysis, requires a lot of work. Data scientists spend 50 to 80 percent of their time curating and preparing data before it can actually be used. Finally, big data technology is changing at a rapid pace. A few years ago, Apache Hadoop was the popular technology used to handle big data. Then Apache Spark was introduced in 2014. Today, a combination of the two frameworks appears to be the best approach. Keeping up with big data technology is an ongoing challenge.

REFERENCES

1] https://en.wikipedia.org/wiki/Big_data
2] https://www.oracle.com/in/big-data/what-is-big-data/
3] https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0268-2
4] https://www.guru99.com/what-is-big-data.html
5] https://www.sciencedirect.com/science/article/pii/S014829631630488X
6] https://searchdatamanagement.techtarget.com/definition/big-data
7]https://thesai.org/Downloads/Volume7No2/Paper_67A_Survey_on_Big_Data_Analytics_Challenges.pdf