CHENNAI
INSTITUTE OF TECHNOLOGY
(Autonomous)

UNIT I UNDERSTANDING BIG DATA

Introduction to big data – convergence of key trends – unstructured data – industry examples of bigdata – web analytics – big data applications– big data technologies – introduction to Hadoop –open source technologies – cloud and big data – mobile business intelligence – Crowd sourcing analytics – inter and trans firewall analytics.

1.1 Introduction to big data

Big data refers to extremely large and complex sets of data that cannot be easily managed, processed, or analyzed using traditional data processing techniques. It encompasses vast volumes of information generated from various sources such as social media, sensors, digital devices, financial transactions, and more. Big data is characterized by its volume, velocity, variety, veracity and value.

Volume is the amount of data which we deal with is of very large size of peta bytes

Velocity refers to the speed at which data is generated and needs to be processed.

Variety refers to the diverse types and formats of data available in big data. It includes structured data (e.g., databases), semi-structured data (e.g., XML, JSON), and unstructured data (e.g., text, images, videos)..

Veracity refers to the reliability and accuracy of the data. Since big data comes from various sources, it may contain inconsistencies, errors, or missing values. Ensuring data quality and reliability is essential to obtain meaningful and reliable insights.

Value refers to the usefulness of the collected data

How does bigdata works?

Big data works through a series of processes that involve data collection, storage, processing, analysis, and visualization. Here's an overview of how big data works:

1. **Data Collection:** Big data is sourced from various channels such as social media platforms, websites, sensors, transactional systems, and more. This data can be structured (e.g., databases), semi-structured (e.g., XML, JSON), or unstructured (e.g., text, images, videos). Data collection can be done in real-time or in batch mode.

2. **Data Storage:** Once collected, the data needs to be stored in a way that allows for efficient processing and analysis. Traditional relational databases may not be suitable for handling the large volumes and variety of big data.

3. **Data Processing:** Big data processing involves transforming and manipulating the collected data to extract valuable insights. Distributed computing frameworks like Apache Spark are often used for processing big data. Data processing tasks may include filtering, aggregating, joining, cleaning, and transforming the data.

4. **Data Analysis:** Once the data is processed, it can be analyzed to uncover patterns, trends, correlations, and other valuable insights. Big data analytics techniques include statistical analysis, machine learning, data mining, and predictive modelling

5. **Data Visualization:** Big data insights are often presented visually through data visualizations such as charts, graphs, and dashboards. Data visualization helps to communicate complex information in a clear and understandable manner, allowing stakeholders to grasp insights quickly and make informed decisions.

1.2 Convergence of key trends

The convergence of key trends refers to the coming together of different technological, social, and economic trends that influence and shape various aspects of our lives. The convergence of key trends in big data has a significant impact on how data is collected, managed, analyzed, and utilized. Here are some examples of the convergence of key trends in big data:

1. **Internet of Things (IoT) and Big Data**: The rapid increase of IoT devices generates massive amounts of data from various sources such as sensors, connected devices, and machines. This data is often unstructured and requires advanced big data analytics techniques to extract valuable insights. The integration of IoT and big data enables realtime monitoring, predictive maintenance, and optimized decision-making in industries such as manufacturing, healthcare, and smart cities.
2. **Artificial Intelligence (AI) and Machine Learning (ML):** AI and ML algorithms play a crucial role in big data analytics. They enable organizations to analyze large datasets, identify patterns, and make intelligent predictions. AI techniques like natural language processing (NLP) and computer vision enhance the understanding and analysis of unstructured data, such as text and images, contributing to advanced analytics and automated decision-making in various domains.
3. **Cloud Computing and Big Data:** Cloud computing provides scalable and costeffective infrastructure for storing and processing big data. Cloud platforms offer distributed storage systems and powerful computing resources that can handle largescale data processing tasks. The convergence of cloud computing and big data enables organizations to store, process, and analyze massive datasets without investing in onpremises infrastructure, allowing for more flexibility and agility.
4. **Edge Computing and Big Data:** Edge computing brings processing and analysis capabilities closer to the data source, reducing latency and enabling real-time decisionmaking. This convergence is particularly relevant in applications where data

needs to be processed and analyzed at the edge in real-time, such as autonomous vehicles, industrial IoT, and smart surveillance. Edge computing offloads some of the processing burden from centralized systems and enhances the efficiency and responsiveness of big data analytics.

5. **Privacy and Ethics in Big Data:** The convergence of big data with privacy and ethics is becoming increasingly important. As the amount of data collected and analyzed grows, concerns around data privacy, security, and ethical use of data intensify. Organizations need to address these concerns by implementing robust data governance practices, ensuring compliance with regulations, and adopting ethical frameworks for handling and analyzing big data.
6. **Data Visualization and Big Data:** With the massive volume and complexity of big data, effective data visualization becomes crucial for understanding and communicating insights. Visualization techniques and tools help in representing large datasets in a visually intuitive manner, enabling decision-makers to grasp patterns, trends, and relationships quickly. The convergence of data visualization with big data enhances the ability to gain actionable insights and make informed decisions.

The convergence of these key trends in big data provides organizations with new opportunities to derive value from their data, optimize operations, and drive innovation. However, it also poses challenges in terms of data management, security, privacy, and the need for skilled professionals with expertise in big data analytics and emerging technologies.

1.3 Unstructured data

Unstructured data refers to data that does not conform to a specific data model or format and does not fit neatly into traditional rows and columns like structured data. It is typically in a

human-readable form and can come from various sources such as text documents, social media posts, emails, images, videos, audio recordings, sensor data, and more.

Unlike structured data that can be easily organized and queried in a database, unstructured data lacks a predefined structure, making it challenging to analyze using conventional methods. Unstructured data presents unique characteristics and complexities, including varying lengths, free-form text, multiple languages, and context-dependent meaning.

Examples of unstructured data:

1. Textual data: This includes documents, reports, emails, articles, social media posts, chat logs, and other forms of written or typed text.
2. Multimedia data: Images, videos, and audio recordings are examples of unstructured data. Analyzing and extracting information from multimedia data often involves techniques such as computer vision and speech recognition.
3. Sensor data: Unstructured data can also come from sensors embedded in various devices and systems. For example, data collected from IoT devices, such as temperature sensors, motion sensors, or GPS location data, is typically unstructured.

Challenges and importance of unstructured data:

Unstructured data presents several challenges for organizations, but it also holds immense value and insights. Some of the challenges include:

1. **Volume:** Unstructured data can be voluminous, making it difficult to store and process efficiently.
2. **Variety:** Unstructured data comes in various formats, making it challenging to integrate and analyze different types of data together.
3. **Complexity:** Unstructured data often requires advanced techniques, such as natural language processing, machine learning, and text mining, to extract meaningful information and uncover patterns.

Despite these challenges, unstructured data is valuable for organizations because it can contain hidden insights, customer sentiment, market trends, and other valuable information that cannot be easily derived from structured data alone. Analyzing unstructured data can provide a deeper understanding of customer behavior, improve decision-making, enable sentiment analysis, facilitate personalized marketing, and support various research and business objectives.

To analyze unstructured data, organizations employ techniques such as text mining, sentiment analysis, entity recognition, topic modeling, image recognition, and audio analysis. These techniques help in extracting, organizing, and transforming unstructured data into a structured format for further analysis and decision-making.

1.4 Industry examples of bigdata

Big data has found applications across various industries, enabling organizations to gain valuable insights, make informed decisions, and drive innovation. Here are some industry examples of how big data is being utilized:

1. **Healthcare:** Big data analytics is transforming healthcare by improving patient care, optimizing operations, and facilitating medical research. It helps in analyzing large volumes of patient data, electronic health records (EHRs), medical imaging, genomics data, and clinical trials to identify patterns, predict disease outbreaks, personalize treatments, and enhance overall healthcare delivery.
2. **Financial Services:** Big data is extensively used in the financial industry for fraud detection, risk management, customer analytics, and algorithmic trading. It enables organizations to analyze large-scale transactional data, social media sentiment, customer behavior, and market trends to identify anomalies, assess risks, personalize financial services, and make data-driven investment decisions.
3. **Retail and E-commerce:** Big data analytics is leveraged by retailers and e-commerce platforms to gain insights into customer preferences, optimize inventory management, personalize marketing campaigns, and improve supply chain efficiency. It involves analyzing customer transaction data, browsing behavior, social media interactions, and competitor pricing to drive customer engagement and enhance the overall shopping experience.
4. **Manufacturing:** Big data plays a vital role in the manufacturing industry by enabling predictive maintenance, optimizing production processes, and improving quality control. Analyzing sensor data, machine logs, and production line data helps in identifying potential equipment failures, reducing downtime, optimizing resource allocation, and enhancing overall operational efficiency.

5. **Telecommunications:** Big data analytics helps telecommunications companies in managing network performance, optimizing customer experience, and offering personalized services. Analyzing call detail records (CDRs), network logs, social media interactions, and customer feedback enables organizations to detect network issues, predict customer churn, tailor service offerings, and improve network infrastructure planning.
6. **Transportation and Logistics:** Big data is used in transportation and logistics for route optimization, predictive maintenance of vehicles, supply chain management, and demand forecasting. Analyzing data from GPS trackers, traffic sensors, weather forecasts, and shipment records helps organizations optimize logistics operations, reduce costs, improve delivery times, and enhance overall customer satisfaction.

These are just a few examples of industries where big data is making a significant impact. The applications of big data extend to many other sectors, including energy and utilities, government, education, marketing, cybersecurity, and more. The key is to collect, store, process, and analyze relevant data to extract actionable insights that can drive innovation, efficiency, and competitive advantage in a particular industry.

1.5 Web analytics

Web analytics refers to the collection, measurement, analysis, and reporting of data related to website usage and visitor behavior. It involves tracking and analyzing various metrics and data points to understand how users interact with a website, optimize its performance, and improve user experience. Here's an overview of web analytics:

1. **Data Collection**: Web analytics relies on the collection of data about website visitors and their interactions. This data is gathered through various methods such as tracking codes embedded in web pages, cookies, log files, and integration with other tools and platforms. Key data collected includes page views, visitor demographics, referral sources, device types, and user actions.
2. **Metrics and Key Performance Indicators (KPIs):** Web analytics involves tracking and measuring specific metrics and KPIs to assess website performance and user behavior. Common metrics include:
    - Visits/Sessions: The number of times users visit the website. o Pageviews: The number of times pages are viewed by users.
    - Unique Visitors: The number of distinct individuals who visit the website. o Bounce Rate: The percentage of visitors who leave the website after viewing only one page. o Conversion Rate: The percentage of visitors who complete a desired action, such as making a purchase or filling out a form.
    - Average Session Duration: The average time users spend on the website.
3. **Data Analysis and Insights**: Once the data is collected, it is analyzed to gain insights into user behavior, website performance, and marketing effectiveness. Data analysis may involve segmentation of visitors based on demographics, traffic sources, or behavior patterns. It also includes identifying trends, patterns, and correlations in the data to understand user preferences, optimize website design and content, and improve conversion rates.
4. **Reporting and Visualization**: Web analytics tools provide reporting and visualization capabilities to present the analyzed data in a user-friendly and actionable format.

Reports can include dashboards, charts, graphs, and tables that summarize key metrics, trends over time, and performance comparisons. Visualization aids in understanding data patterns and communicating insights effectively.

5. **Continuous Optimization:** Web analytics is an ongoing process that involves continuously monitoring, analyzing, and optimizing website performance. By identifying areas for improvement, such as high bounce rates or low conversion rates, website owners can implement changes to enhance user experience, increase engagement, and drive desired actions.

Web analytics tools and platforms: Numerous web analytics tools and platforms are available, ranging from free to enterprise-level solutions. Some popular web analytics tools include

Google Analytics: which monitors website traffic, behaviors and conversions.

Optimizely: is a customer experience ad A/B testing platform that helps businesses test ad optimize their online experiences.

Kissmetrics: is a customer analytics platform that gathers website data and presents it in an easy to read format.

Crazy egg: is tool that tracks where customers click on a page. This information can help organizations understand how visitors interact with cotet and why they leave the site.

Web analytics is essential for businesses and website owners to understand their online audience, evaluate the effectiveness of marketing campaigns, optimize website performance, and make data-driven decisions to improve overall online presence and achieve business goals.

1.6 Big data applications

Big data applications refer to the use of large and complex datasets to extract meaningful insights and drive decision-making processes. These applications leverage advanced data processing techniques and technologies to analyze massive volumes of structured and unstructured data from various sources. Here are some common areas where big data applications are extensively used:

1. Big data and marketing :

In marketing, big data comprises gathering, analyzing, and using massive amounts of digital information to improve business operations, such as:

- **Getting a 360-degree view of their audiences.** The concept of "know your customer" (KYC) was initially conceived many years ago to prevent bank fraud. KYC provides insight into customer behavior that was once limited to large financial institutions. Now, because of the accessibility of big data, the benefits of KYC are available to even small and medium businesses, thanks to cloud computing and big data.
- **Customer engagement**, specifically how your customers view and interact with your brand, is a key factor in your marketing efforts. Big data analytics provides the business intelligence you need to bring about positive change, like improving existing products or increasing revenue per customer.

- **Brand awareness** is another way big data can have a significant impact on marketing. Aberdeen Group's Data-Driven Retail study showed that "data-driven retailers enjoy a greater annual increase in brand awareness by 2.7 times (20.1% vs. 7.4%) when compared to all others."
- The 360-degree view from big data allows marketers to present customer-specific content when and where it is most effective to improve online and in-store brand recognition and recall. Big data allows you to be the Band-Aid of your product category even if you don't have the marketing budget of Johnson & Johnson.
- **Improved customer acquisition** is another great benefit that big data brings to marketing. A McKinsey survey found that "intensive users of customer analytics are 23 times more likely to clearly outperform their competitors in terms of new customer acquisition." Leveraging the cloud allows for the gathering and analysis of consistent and personalized data from multiple sources, such as web, mobile applications, email, live chat, and even in-store interactions.
- Big data can help marketers **leverage real-time data in cloud computing environments.** The ability of big data to acquire, process, and analyze real-time data quickly and accurately enough to take immediate and effective action cannot be matched by any other technology. This is critical when analyzing data from GPS, IoT sensors, clicks on a webpage, or other real-time data.
- Big data analytics is an essential component of big data. It provides business intelligence that **results in time and cost savings** by optimizing marketing performance.

Three types of big data for marketers

Marketers are interested in three types of big data: customer, financial, and operational. Each type of data is typically obtained from different sources and stored in different locations.

1. **Customer data** helps marketers understand their target audience. The obvious data of this type are facts like names, email addresses, purchase histories, and web searches. Just as important, if not more so, are indications of your audience's attitudes that may be gathered from social media activity, surveys, and online communities.
2. **Financial data** helps you measure performance and operate more efficiently. Your organization's sales and marketing statistics, costs, and margins fall into this category. Competitors' financial data such as pricing can also be included in this category.
3. **Operational data** relates to business processes. It may relate to shipping and logistics, customer relationship management systems, or feedback from hardware sensors and other sources. Analysis of this data can lead to improved performance and reduced costs.

2. Fraud and Big Data

Fraud is intentional deception made for personal gain or to damage another individual. One of the most common forms of fraudulent activity is credit card fraud. In order to prevent the fraud, credit card transactions are monitored and checked in near real time. If the checks identify pattern inconsistencies and suspicious activity, the transaction is identified for review and escalation. Big Data technologies provide an optimal technology solution based on the following three Vs:

1. High volume. Years of customer records and transactions (150 billion1 records per year)
2. High velocity. Dynamic transactions and social media information
3. High variety. Social media plus other unstructured data such as customer emails, call center conversations, as well as transactional structured data.

The Capgemini team pointed out that they use an open-source weapon named Elastic Search, which is a distributed, free/open-source search server based on Apache Lucene (see Figure). It can be used to search all kind of documents at near real-time. They use the tool to index new transactions, which are sourced in real-time, which allows analytics to run in a distributed fashion utilizing the data specific to the index. Using this tool, large historical data sets can be used in conjunction with real-time data to identify deviations from typical payment patterns. This Big Data component allows overall historical patterns to be compared and contrasted, and allows the number of attributes and characteristics about consumer behavior to be very wide, with little impact on overall performance.

Once the transaction data has been processed, the percolator query then performs the functioning of identifying new transactions that have raised profiles. Percolator is a system for incrementally processing updates to large data sets.

Percolator is the technology that Google used in building the index—that links keywords and URLs—used to answer searches on the Google page. Percolator query can handle both structured and unstructured data. This provides scalability to the event processing framework, and allows specific suspicious transactions to be enriched with additional unstructured information.

Another approach to solving fraud with Big Data is social network analysis (SNA). SNA could reveal all individuals involved in fraudulent activity, from perpetrators to their asso- ciates, and understand their relationships and behaviors to identify a bust out fraud case.

The Importance of Big Data Analytics in Terms of Fraud Prevention:

As online purchase, payment and money transfer transactions increase, the risks of fraud that may occur through these transactions also increase.It was very difficult for companies to process and analyze the huge amount of data that emerged from these transactions and use it in fraud detection. At this point, we come across an indispensable facilitating tool: big data analytics for fraud detection. Using big data analytics in some points of fraud detection provides many advantages.

One of the most important points when detecting fraud is to take actions quickly. It may take a long time to identify the suspicious ones among this large number of irregular data resulting from transactions.

Some transactions may be perceived as suspicious by misinterpretations as a result of these long analyzes. During this evaluation process, there will still be a need for people, namely a manual workload, to analyze the data and check for suspicious transactions or misinterpretations.

In order to protect the company and customers from harm, it is necessary to draw up rules based on this data and looking at past fraudulent activities, so that we can establish systems that can prevent possible damages and frauds that may occur.

All these mean more cost, time, manual work. Big data analytics plays the biggest helper role in solving these issues. By using data analyzed with techniques in big data analytics can provide:

- Low costs
- More accurate and precise detections
- Optimized workflows and efficiency of systems
- Better services to customers

In addition data mining and machine learning made by big data analytics are used in fraud analytics. These tools enable the implementation of payment fraud analytics, financial fraud analytics, and insurance fraud detection analytics.

What are the Common Problems in Big Data Analytics in Fraud Detection?

We mentioned the importance of big data analytics in detecting fraud. Although it makes it easier to detect fraud, it can also bring some problems with it. Some of these problems can be listed as:

- **Unrelated or Insufficient Data:** The data from the transactions may come from many different sources. In some cases, false results can be obtained in fraud detection due to these insufficient or irrelevant data. Detection can be based on the inappropriate rules used in the algorithm. Because of this risk of failure, companies may be hesitant to use big data analytics and machine learning.
- **High Costs:** Big data analytics and fraud detection systems may cause some costs such as the cost of software, hardware systems, the cost of components used for sustainability of these systems and the time spent.
- **Dynamic Fraud Methods:** As technology develops, fraud methods develop at the same pace. In order to catch this speed and detect fraud, it is necessary to constantly monitor the data and give rules to the algorithms with new and accurate data analytics.
- **Data Security:** While processing the data and making decisions with this data analytics system, the security of the data is also a problem to be considered. That means the security of data should be checked.

3. Risk and bigdata

Many of the world’s top analytics professionals work in risk management. It would be an understatement to say that risk management is data-driven without advanced data analytics, modern risk management would simply not exist. The two most common types of risk management are credit risk management and market risk management.

Credit risk analytics focus on past credit behaviors to predict the likelihood that a borrower will default on any type of debt by failing to make payments which they obligated to do.

Market risk analytics focus on understanding the likelihood that the value of a portfolio will decrease due to the change in stock prices, interest rates, foreign exchange rates, and commodity prices.

Credit risk management is a critical function that spans a diversity of businesses across a wide range of industries. Traditionally, credit risk management was rooted in the philosophy of minimizing losses. However, over time, credit risk professionals and business leaders came to understand that there are acceptable levels of risk that can boost profitability beyond what would normally have been achieved by simply focusing on avoiding write-offs.

Credit risk professionals are stakeholders in key decisions that address all aspects of a business, from finding new and profitable customers to maintaining and growing relationships with existing customers.

As businesses grow, what starts out as a manual and judgmental process of making credit decisions gives way to a more structured and increasingly automated process in which data-driven decisions becomes the core.

The vast amount of both qualitative and quantitative information available to credit risk professionals can be overwhelming to digest and can slow down a process with potential sales at risk. With advanced analytical tools,these abundant and complex data sources can be distilled into simple solutions that provide actionable insights and are relatively easy to implement. As Figure 2.3 illustrates, there are four critical parts of the typical credit risk framework: planning, customer acquisition, account management, and colections. All four parts are handled in unique ways through the use of Big Data.

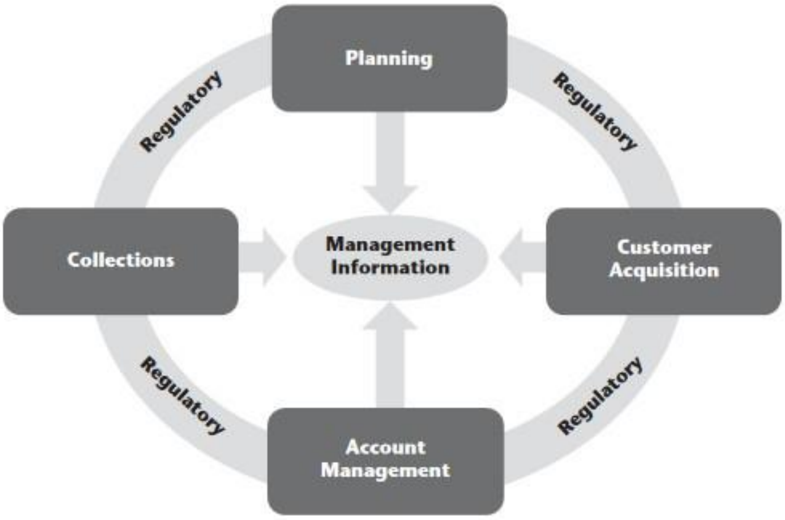


Figure 2.3 Credit Risk Framework
Source: Ori Peled.

How Does Big Data Help Manage Risks?

The benefits of big data may help a company improve sales, lower costs, streamline staffing, and more. Applying big data to risk management is helpful for improving financial, digital, and other parts of the business. Here are some common examples of where big data and risk management come together to help businesses like yours.

Identify fraud

One of the most obvious uses of big data science in risk management is fraud prevention. Using large datasets, it’s possible to zero in on suspicious activity in real-time to prevent future losses. You may have seen this in action when a potentially fraudulent credit card transaction was blocked at the point of sale. Big data is deeply integrated into payment processing systems. This is just one example of the benefits of big data for business.

Financial risk management

Your business may offer credit to current and new customers. If you are willing to deliver a product or service before taking payment, your business is extending credit. Big data may help your business understand credit risks and the likelihood of full payment before delivering a product or service on credit.

Reduce customer churn rate

Repeat business is the lifeblood of many companies. If you are losing customers and don't understand why it could be detrimental to long-term revenue. Part of financial risk management is preventing customer losses. Big data could help your sales or operations teams identify which types of customers are most at risk and opportunities to avoid customer churn.

Reduce employee attrition

Losing employees is a major cost for some businesses. Each time someone leaves for any reason, the business loses that person's knowledge and productivity. It takes time and money to post a job, interview and hire a replacement, and train the new hire. Human-resources focused analytics may help understand why employees leave. This gives you the ability to take action to lower the risk that your valued team members will want to search for a new job elsewhere.

Operational risk assessment

Most business managers understand the biggest risks in their operations, but it's impossible to have eyes on every part of your business at once. The power of big data could allow you to identify new risk areas in your operations that might have cost your business but can be solved once you better understand the risks.

Digital risk management

According to a study by IBM, the average cost of a data breach is $3.86 million USD ($5.17 million CAD) and it takes an average of 280 days to identify and contain a breach. Big data analysis could help identify vulnerabilities before the digital bad actors do. In the event of a breach, a data analytics risk assessment could help patch the security flaw and put your business back on track as quickly as possible.

Lower risk for future growth

If you plan to grow your business, big data can be a valuable tool. For example, a retail or restaurant chain can tap into big data to learn about prospective locations. While managers used to have to rely on the look and feel of a location to make a gut decision, these days they can look at demographic and population data to greatly increase the odds of picking a profitable location every time.

4. Bigdata and algorithmic trading:

Big data and algorithmic trading are closely interconnected, with big data playing a significant role in shaping algorithmic trading strategies. Algorithmic trading refers to the use of computer algorithms to execute trades in financial markets. These algorithms leverage large amounts of data, including historical market data, real-time market feeds, news, and other relevant information, to make informed trading decisions.

Here's how big data impacts algorithmic trading:

1. Data Collection: Big data technology allows for the collection and storage of massive volumes of financial market data. This includes historical price and volume data, order book data, news feeds, social media sentiment, and other relevant information. This vast amount of data is essential for building accurate and robust trading models.
2. Data Processing and Analysis: Big data analytics techniques are employed to process and analyze the collected financial market data. This involves handling structured and unstructured data, cleaning and preprocessing the data, and applying advanced analytics algorithms to identify patterns, correlations, and anomalies that can be used to generate trading signals.

3. Trading Strategy Development: Big data analytics enables traders and quantitative analysts to develop sophisticated trading strategies. By analyzing historical and realtime market data, big data techniques can be used to identify market inefficiencies, detect patterns, and generate trading signals for buying or selling securities. These signals can be based on technical indicators, statistical models, machine learning algorithms, or a combination of approaches.
4. Risk Management: Big data analytics plays a crucial role in risk management for algorithmic trading. By analyzing various data sources, including market data, portfolio positions, and risk factors, big data techniques can help quantify and manage risks associated with algorithmic trading strategies. This includes monitoring risk exposure, identifying potential anomalies or deviations from expected behavior, and implementing risk mitigation measures.
5. Real-Time Decision Making: Big data enables real-time decision making in algorithmic trading. By processing and analyzing real-time market data, news feeds, and other relevant information, algorithms can quickly identify trading opportunities, react to market events, and execute trades in fractions of a second. This speed and efficiency are critical in high-frequency trading and other fast-paced trading environments.
6. Backtesting and Performance Analysis: Big data analytics allows traders to backtest and evaluate the performance of algorithmic trading strategies. By using historical market data, traders can simulate the execution of trades and assess the strategy's performance under different market conditions. This helps in refining and optimizing trading algorithms to improve profitability and reduce risk.
7. Market Surveillance and Compliance: Big data technology is also used for market surveillance and compliance in algorithmic trading. Regulatory bodies and exchanges employ big data analytics to monitor trading activities, detect market manipulation or abusive practices, and ensure compliance with regulations.

Overall, big data plays a crucial role in algorithmic trading by enabling the collection, processing, and analysis of vast amounts of financial market data. It helps in developing trading strategies, risk management, real-time decision making, performance analysis, and compliance. The combination of big data and algorithmic trading has revolutionized the financial industry, making trading faster, more efficient, and data-driven.

5. Big data and advances in healthcare:

Big data has brought about significant advances in healthcare by enabling the collection, analysis, and utilization of large and diverse healthcare datasets. Here are some ways in which big data is transforming the healthcare industry:

1. Precision Medicine: Big data analytics allows healthcare professionals to analyze vast amounts of genomic, clinical, and lifestyle data to develop personalized treatment plans. By identifying patterns and correlations within large datasets, healthcare providers can tailor medical interventions, medications, and therapies to individual patients, resulting in more effective and targeted healthcare.
2. Disease Prevention and Early Detection: Big data analytics helps identify early signs and risk factors for diseases. By analyzing patient data, including electronic health records (EHRs), genetic information, wearable devices, and lifestyle data, healthcare professionals can identify patterns and markers that indicate disease susceptibility. This enables proactive interventions, early detection, and prevention of diseases.

3. Real-Time Patient Monitoring: The integration of big data analytics with real-time patient monitoring systems allows healthcare providers to monitor patients' health continuously. Wearable devices, sensors, and mobile apps collect patient data, such as vital signs, activity levels, and sleep patterns. Big data analytics can process this data in real-time, providing insights for remote monitoring, early warning systems, and personalized interventions.
4. Healthcare Operational Efficiency: Big data analytics helps healthcare institutions improve operational efficiency. By analyzing large datasets on patient flow, resource allocation, and clinical processes, hospitals and clinics can optimize workflows, reduce wait times, and enhance resource utilization. This leads to improved patient experiences, reduced costs, and better overall healthcare delivery.
5. Drug Discovery and Development: Big data plays a vital role in accelerating drug discovery and development processes. By analyzing large datasets structures, genomics, clinical trials, and patient outcomes, researchers can identify potential drug targets, predict drug efficacy, and optimize clinical trial design. This reduces the time and cost involved in bringing new drugs to market.
6. Public Health Surveillance: Big data analytics enables the monitoring and surveillance of public health trends and outbreaks. By analyzing diverse data sources, such as social media, internet searches, electronic health records, and environmental data, public health agencies can detect and respond to disease outbreaks more effectively. This helps in early intervention, resource allocation, and implementation of preventive measures.
7. Predictive Analytics and Decision Support: Big data analytics supports healthcare decision-making by providing predictive models and decision support systems. By analyzing large datasets, healthcare professionals can predict disease progression,

patient outcomes, and healthcare resource needs. This aids in treatment planning, resource allocation, and healthcare policy development.

8. Patient Safety and Quality Improvement: Big data analytics contributes to patient safety and quality improvement initiatives. By analyzing patient data, adverse events, and clinical outcomes, healthcare providers can identify patterns and factors contributing to errors, adverse events, or suboptimal outcomes. This knowledge helps in implementing targeted interventions, best practices, and quality improvement on molecular
9. programs.

It's important to note that while big data brings immense opportunities, it also poses challenges regarding data privacy, security, and ethical considerations. Healthcare organizations need to ensure appropriate data governance frameworks, secure data handling practices, and compliance with regulations to protect patient privacy and confidentiality.

Overall, big data has the potential to revolutionize healthcare by improving patient care, driving medical advancements, enhancing operational efficiency, and transforming public health practices.

6. Advertising and bigdata

Big data has had a significant impact on the field of advertising, transforming the way advertisers target, engage, and measure the effectiveness of their campaigns. Here are some key ways in which big data is influencing advertising:

1. Audience Targeting: Big data enables advertisers to understand their target audiences better. By analyzing large volumes of data on consumer behavior, demographics, interests, and preferences, advertisers can create more precise and personalized audience segments. This allows for highly targeted advertising campaigns that are more relevant to individual consumers, leading to improved engagement and conversion rates.
2. Behavioral Tracking: Big data analytics allows advertisers to track and analyze consumer behavior across various digital channels. By collecting and analyzing data on website visits, ad interactions, search history, and social media activity, advertisers can gain insights into consumer preferences and tailor their ad placements and messaging accordingly. This helps in delivering more personalized and contextually relevant advertisements.
3. Real-Time Ad Optimization: Big data enables real-time monitoring and optimization of advertising campaigns. By analyzing data on ad impressions, click-through rates, conversions, and other metrics, advertisers can make immediate adjustments to their campaigns. This includes optimizing ad placements, creative elements, and targeting parameters to improve performance and maximize return on investment (ROI).
4. Ad Creative Personalization: Big data allows for dynamic ad creative personalization. By leveraging data on consumer preferences, browsing history, and demographics, advertisers can deliver customized and relevant ad creative to individual consumers. This includes personalized product recommendations, tailored messages, and dynamic ad content that resonates with each user, increasing the chances of engagement and conversion.

5. Attribution and Measurement: Big data analytics provides insights into the effectiveness of advertising campaigns. By analyzing data on customer journeys, conversion paths, and various touchpoints, advertisers can attribute conversions or sales to specific advertising efforts. This helps in understanding the impact of different advertising channels and optimizing marketing budgets accordingly.
6. Social Media Advertising: Social media platforms generate vast amounts of usergenerated data. Big data analytics enables advertisers to leverage this data to target specific audience segments with relevant ads. By analyzing social media conversations, sentiment analysis, and user profiles, advertisers can identify trends, interests, and influencers to optimize their social media advertising strategies.
7. Predictive Analytics: Big data analytics helps advertisers make predictions about consumer behavior and campaign performance. By analyzing historical data and patterns, advertisers can anticipate future trends, identify high-value customers, and optimize their advertising strategies accordingly. Predictive analytics can assist in areas such as customer segmentation, churn prediction, and campaign performance forecasting.
8. Ad Fraud Detection: Big data analytics plays a critical role in detecting and preventing ad fraud. By analyzing large datasets and monitoring patterns, algorithms can identify suspicious activities, fraudulent impressions, and invalid clicks. Advertisers can leverage big data techniques to enhance ad fraud detection systems and protect their advertising investments.

Overall, big data has revolutionized advertising by enabling more targeted, personalized, and data-driven campaigns. It empowers advertisers to reach the right audiences with the right messages, optimize their campaigns in real-time, and measure the impact of their advertising efforts. However, it is important to balance the benefits of big data with ethical considerations and respect for user privacy.

1.7 Big data technologies

The term “big data” refers to a collection of extensive, complicated data collections and volumes that include enormous quantities of data, data management capabilities, social media analytics, and real-time data.

Big data is about data volume, and extensive data set is measured typically in terms of terabytes or petabytes.

The vast collection of data is growing at an exponential pace over time and has led to the development of various big data-based technologies. The use of these techniques for analyzing enormous volumes of data is known as big data analytics. Big Data Technology on the other hand is simply the collection of software utilities meant to analyze, process, and extract information from very complex and massive data sets that traditional data processing software cannot handle.

What is Big Data Technology

There is a large number of technologies that constitute big data technologies. While these technologies are distinct from each other, all these technologies involve dealing with a huge amount of data i.e. Big Data. When you sit to study the big data architecture, you will see how vast it is and how little you might know.

On an overall level, big data technologies comprise a number of software and tools that allow the user to analyze, extract, manage, and process a huge amount of data. This data is often complex, huge in size and volume, and is sometimes unstructured.

Big data tools help in managing large and complex data that most traditional tools cannot do. This is because the traditional tools function on local machines or servers, and are designed for small to medium-sized structural data.

Before getting into categorizing the big data technologies, let us first take a look at the **four primary objectives** of these tools. This is lead us to the final answer to what you mean by big data technologies.

1. Integration

Big Data Technologies' biggest objective is to integrate a large amount of data from the source to the user. This includes identifying mechanisms and procedures to stream data seamlessly in the day-to-day work framework. Incorporating a huge amount of data in an effortless manner is the objective of many big data technologies & techniques.

2. Processing

In a way an extension of the Integration aspect of big data technologies, a number of big data technologies are responsible for creating the infrastructure that helps us deal with a large amount of data This is done often by dividing the data into manageable pieces, using sophisticated hardware placed in a remote location and then with the help of integration based tools, giving a user an experience they have when working with small amount of data on their local machine.

3. Storage and Management

While Big Data technologies and tools help integrate a large amount of data in the day-today working, they are primarily responsible for the storage of this huge amount of data.

These big data technologies help us manage this data, make sure we don’t lose track of it, and ensure that this data is available for use when required.

4. Solutions

A large number of big data technologies and tools provide the users with the necessary options to find the precious insights that can help the organization’s leaders make key decisions and solve business problems. These tools help have the option to integrate with other big data technologies and people with coding prowess can then use these tools to mine, find patterns, run statistical tests, create predictive models, etc on large quantities of data. This can provide the users with information that otherwise would have been easily lost if such big data technologies were not available.

Types of Big Data Technologies

As discussed above, Big Data technologies try to help users store, manage, integrate, process data, and find solutions. The various big data technologies and techniques can be by and large divided into **two categories**– Operational and Analytical.

Operational

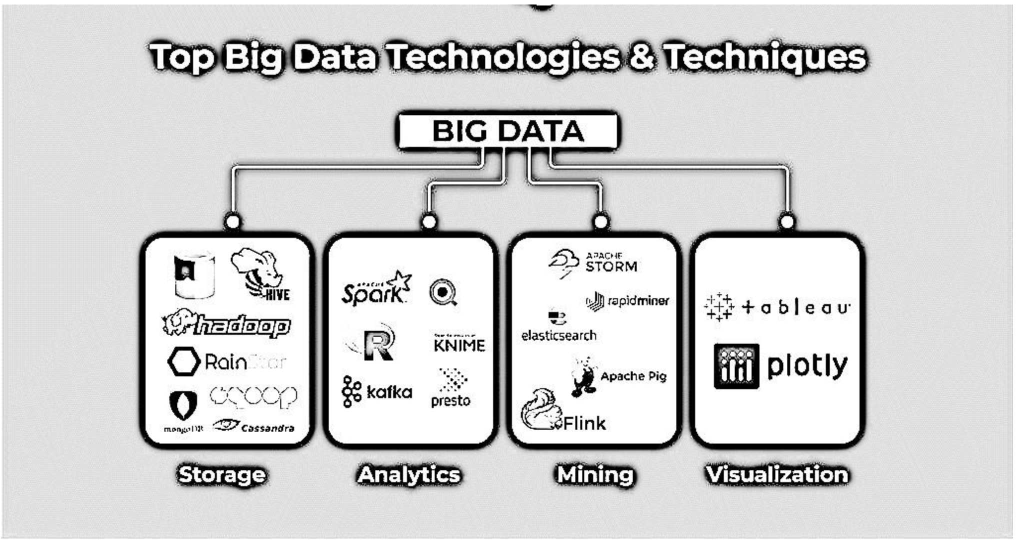
It is this side of Big Data Technologies where a large number of Big Data Engineers, IT technicians, and Networking experts come together and are involved in big data processing and storage.

These are those technologies responsible for generating the big data and storing them in an efficient manner that is not responsible for providing any solutions. This data can be generated from social media platforms, shopping sites, hospitalities-based businesses such as ticketing data, or some other data-generating from within the organization.

Analytics

This is where Data Analysts and Data scientists come into play. They use those types of Big data technologies that help find business solutions by analyzing a large amount of data. Here common analytics methodologies come into play for example stock marketing predictions, weather forecasting, early identification of disease for insurance companies, etc.

Top Big Data Technologies & Techniques



Big data tools are usually categorized according to their utility. However, putting it more generally – all big data technologies & techniques can be divided into *4 major categories* viz. storage, analytics, mining, and visualization.

Data Storage Tools

1. Hadoop

As one of the most common big data tools, Hadoop is used to deal with the data in a clump mode i.e. it breaks down the data into smaller pieces which the user then consumes. Hadoop basically is a system that uses a peculiar distribution system to deal with Big Data. Important components of the HADOOP ecosystem include HDFS (Hadoop Distributed File System), YARN (Yet Another Resource Negotiator), MapReduce, and HadoopCommon.

2. HIVE

HIVE is an important tool in the Big Data tools domain as it allows the user to easily extract, read and write data. It uses a SQL methodology and interface and consequently uses a similar query-based language known as HQL (Hive Query language), supports all SQL data types, and has drivers (JDBC Drivers) and a command line (HIVE Command-Line).

3. MongoDB

This is another tool important for big data as far as storage is concerned. MongoDB is a NoSQL-based database and thus is a bit different from typical RDBMS databases. It doesn’t use the traditional schema and data structure, so it becomes easy to store large amounts of data. MongoDB is particularly famous for its flexibility and performance with the various distributed architectures.

4. Apache Sqoop

This tool is used to move big data to structured data stores such as MYSQL, Oracle from Apache Hadoop, and vice-a-versa. It provides connectors for all the major RDBMS.

5. RainStor

RainStor is a database management software developed by a company with the same name. It provides an enterprise online data archiving solution that works on top of Hadoop and runs natively on the HDFS. It stores large amounts of data using a technique called the Deduplication Techniques.

6. Data Lakes

Data Lakes is used as a consolidated repository for storing all kinds of data i.e. untransformed data is stored using Data Lakes. As there is no prerequisite to convert the data into a structured format before storing it, it makes the process of data accumulation easy and one can use various big data-based tools to perform analytics.

7. Cassandra

It is another big data tool for data storage. It is a NoSQL database and can deal with data spread across several clusters. Due to its scalability, query-based language property, integration with MapReduce, and distributed mechanism, it is often the top choice in the NoSQL databases.

Data Analytics Based Big Data Tools

1. Apache Spark

Spark is one of the most important aspects of Big Data as it allows the user to perform tasks such as batch, interactive or iterative processing, visualization, manipulation, etc. As it uses RAM, it is much quicker than many other older big data technologies. Its contender Hadoop is commonly used with structured and batch processing whereas Spark is often used with real-time data making both the tools to be used by organizations simultaneously.

2. R language

R is a programming language often known as a language made by the statisticians for the statisticians. It helps data analysts and scientists to perform statistical tests and create statistical models. However, R can also be used for creating Machine Learning and Deep Learning-based models. R made to this list because it can now be integrated with big databased technologies.

3. QlikView

A highly powerful BI tool, QlikView helps in generating quick and detailed insights from big data. It has a simple, straightforward user interface helping the user to simply click on data points and perform unrestricted data analytics. It helps in identifying the relationship between features in datasets or among datasets and uses associative data modeling.

4. Qlik Sense

Similar to QlikView, Qlik Sense has a drag and drop User Interface that helps users to create story-based reports quickly.

5. Hunk

Hunk is a Big Data tool that helps explore, analyze, and visualize data in the Hadoop Ecosystem. It allows for drag and Drops analytics, a rich developer environment, customer dashboards, integrated analytics, and fast deployment. In order to analyze the data, it uses the Splunk Search Processing Language.

6. Platfora

It is a subscription-based big data tool that helps in Big Data Analytics. It is an interactive tool that helps the user with raw data and early identification of patterns

7. Kafka

Kafka is a stream-processing platform and is similar to an Enterprise Messaging System. It deals with large amounts of real-time data feeds and publishes them in a distributed format. It is used for data processing, data delivery in real-time, etc.

8. Presto

This tool is often used in big data analytics. It uses the Distributed SQL Query engine which is optimized for running Interactive Analytics Queries. Presto can be used with different data sources like Hive, Cassandra, MYSQL, etc. It helps in pipeline executions, creating user-defined functions, and simple debugging of code. Lastly, the biggest advantage is that it scales large velocity of data and is the reason that companies such as Facebook have built Presto-based applications for their data analytics needs.

9. KNIME

Based on Java, this big data tool helps the user in workflow and data analytics by using sophisticated data mining techniques, data extraction using SQL style queries, predictive analytics, etc. It also has the capability to connect to the common Hadoop distributions and use machine learning algorithms via MLlib integration.

10. Splunk

This big data tool is used to provide quick insights into the data. The user can access Hadoop clusters from virtual indexes, and connect to tools such as Tableau and different databases including Oracle, MYSQL, etc. It helps in dealing with real-time streaming data by helping the user in indexing, capturing, and correlating such data. Additionally, the users can easily gain insights by generating reports, graphs dashboards, etc.

11. Mahout

Mahout allows for the implementation of Machine Learning to Big Data and performs MLbased applications such as Segmentation, Classification, Collaborative Filtering, etc.

Data Mining Based Big Data Tools

1. MapReduce

It is because of MapReduce that it is possible to deal with the volume aspect of Big Data as it employs parallel and distributed algorithms to apply logic to the huge data. Thus MapReduce that plays a crucial role in transforming the otherwise unmanageable data into a manageable structure. The word MapReduce is a portmanteau, as the word represents a combination of two methodologies – Map and Reduce. While Map is responsible for sorting and filtering data, Reduce is pivotal for summarizing and aggregating data.

2. Apache PIG

Developed by Yahoo, PIG works on a query-based language known as Pig Latin which is similar to SQL and is used for structuring, processing, and analyzing big data. It provides ease of programming as it allows the user to create their own custom functions while translating them into MapReduce program in the background.

3. RapidMiner

It can perform a range of functions from ETL to data mining and even predictive modeling and machine learning. RapidMiner is a useful tool as it helps in solving problems faced in business analytics, supports multiple languages, and provides the users with a simple user interactive interface. It is open-source which is one of its biggest advantages as even while being open-sourced it is still generally considered secure.

4. Apache Storm

Storm is an open-source, scalable distributed real-time computational framework. Based on Clojure and Java, it helps in dealing with processing unbounded streams of data.

5. Flink

Apache Flink is designed to deal with both unbounded and bounded streams of data. While bounded data has a definite start and end, its processing is also equated to batch processing. As a result, Flink is a highly scalable distributed processing engine. You can use Flink with data pipelines and data analytics applications.

7. ElasticSearch

Written in Java, it is a common search engine tool used by a number of companies (such as Accenture, StackOverflow, Netflix, etc.) It provides a full-text search engine that is based on the Lucene library and provides an HTTP Web Interface and Schema-free JSON documents.

Data Visualization Based Big Data Tools

1. Tableau

A powerful data visualization tool, Tableau can now be integrated with spreadsheets and even with big data platforms, cloud, relational databases, etc., providing quick insights from raw data. As it is a secure proprietary tool that even allows sharing of dashboards in realtime, the popularity of this tool is ever increasing.

2. Plotly

Plotly is a library that allows the user to create interactive dashboards. It provides API libraries for all the common languages including Python, R MATLAB, and Julia. For example, Plotly Dash is something that helps in creating interactive graphs easily using python.

Trending Big Data Technologies and Tools

There are a number of latest big data technologies that are going to be in high demand. i)

Docker

Docker allows the users to build, test, deploy and run their applications. It uses the concept of containers which is a mechanism for dividing the software into standardized units. It helps the user to package their application such that all aspects of the application such as dependent libraries, system tools, etc are all in one place. Docker helps to standardize the application operations, cut costs and increase the reliability of the whole deployment process.

ii) **TensorFlow**

The use of AI is trending in Big Data and one of the leading tools in this race is TensorFlow, an ecosystem of recourses, modules, and tools in itself. TensorFlow can be used to create

Machine Learning and Deep Learning models that can be applied to Big Data. iii) **Apache Beam**

It is used for creating Parallel data processing pipelines. This helps in simplifying largescale data processing. The user has to build a program defining the pipeline, and these programs can be written in Java, Python, Go, etc. iv) **Kubernetes**

This big data tool helps the user to automate, deploy and scale containerized applications across clusters of hosts. Container applications deployed in the cloud can easily be tracked by using Kubernetes.

v) **Blockchain**

This Big Data technology helps in creating encrypted blocks of data and chains them together making the transaction of digit currency such as bitcoin extremely secure. Blockchain's potential application in the world of big data is huge – so much that the BFSI domain is all set to encompass the benefits of implementing it. vi) **Airflow**

Apache Airflow is a scheduling and pipeline tool. It helps in scheduling and managing complex data pipelines stemming from various sources. In this tool Directed Acyclic Graphs (DAG) are used to represent these workflows. Airflow helps make and handle the data pipelines, which makes ML models become more accurate and efficient. **Where are Big Data technologies used?**

A large number of successful companies use big data. Common Examples of the use of Big data technologies include-

1. Netflix, Spotify, and other streaming platforms for recommending movies and songs
2. Amex and other Credit Card companies to understand consumer behavior
3. Amazon, Flipkart, Walmart, etc for online trading and purchasing
4. Hospitality companies for online ticket booking

What are the advantages of big data technologies?

While there are a number of advantages, the most common advantages include cost optimization, improving operational efficiency, and fostering competitive pricing.

1.7 Introduction to Hadoop

Hadoop is an open-source framework for distributed storage and processing of large-scale datasets across clusters of computers. It provides a scalable and reliable platform for handling big data. Hadoop was initially developed by the Apache Software Foundation and is widely used in various industries for storing, processing, and analyzing massive volumes of data.

The key components of the Hadoop framework are:

1. Hadoop Distributed File System (HDFS): HDFS is a distributed file system that allows for the storage of large datasets across multiple machines. It is designed to handle data replication, fault tolerance, and high-throughput access to data. HDFS breaks data into blocks and distributes them across a cluster of nodes.
2. MapReduce: MapReduce is a programming model and processing framework used for distributed data processing in Hadoop. It enables the parallel processing of data across the cluster by dividing the work into map and reduce tasks. The map tasks process data and generate intermediate key-value pairs, which are then consolidated and reduced to produce the final results.
3. Yet Another Resource Negotiator (YARN): YARN is the resource management framework in Hadoop. It manages and allocates resources in a Hadoop cluster to different applications, ensuring efficient resource utilization and workload management. YARN allows for the integration of various data processing engines, such as MapReduce and Apache Spark, within the same cluster.

Why Is Hadoop Important?

Hadoop is a beneficial technology for data analysts. There are many essential features in Hadoop which make it so important and user-friendly.

1. The system is able to store and process enormous amounts of data at an extremely fast rate. A semi-structured, structured and unstructured data set can differ depending on how the data is structured.
2. Enhances operational decision-making and batch workloads for historical analysis by supporting real-time analytics.
3. Data can be stored by organisations, and it can be filtered for specific analytical uses by processors as needed.
4. A large number of nodes can be added to Hadoop as it is scalable, so organisations will be able to pick up more data.

5. A protection mechanism prevents applications and data processing from being harmed by hardware failures. Nodes that are down are automatically redirected to other nodes, allowing applications to run without interruption.

Based on the above reasons, we can say that Hadoop is important.

Hadoop provides several **advantages** for big data processing:

1. Scalability: Hadoop allows organizations to scale their data storage and processing capabilities by adding more nodes to the cluster. It can handle petabytes or even exabytes of data, making it suitable for large-scale data storage and analysis.
2. Fault Tolerance: Hadoop ensures high availability and fault tolerance through data replication. Data is automatically replicated across multiple nodes in the cluster, so if a node fails, the data remains accessible from other nodes. This fault-tolerant design ensures data reliability and minimizes downtime.
3. Cost-Effectiveness: Hadoop can run on commodity hardware, which is more affordable compared to specialized infrastructure. It eliminates the need for expensive, high-end servers and storage systems, making big data processing more accessible and costeffective for organizations.
4. Data Processing Flexibility: Hadoop can handle various types of data, including structured, semi-structured, and unstructured data. It supports batch processing as well as real-time streaming data processing, allowing for diverse data processing needs.
5. Ecosystem: Hadoop has a vast ecosystem of tools and technologies that complement its core components. These include Apache Hive for SQL-like querying, Apache Pig for data processing, Apache Spark for in-memory analytics, Apache HBase for real-time database operations, and many more. This ecosystem provides additional functionalities and expands the capabilities of Hadoop.

Who Uses Hadoop?

Hadoop is a popular big data tool, used by many companies worldwide. Here’s a brief sample of successful Hadoop users:

- British Airways
- Uber
- The Bank of Scotland
- Netflix
- The National Security Agency (NSA), of the United States
- The UK’s Royal Mail system
- Expedia
- Twitter

Now that we have some idea of Hadoop’s popularity, it’s time for a closer look at its components to gain an understanding of what is Hadoop.

Components of Hadoop

Hadoop is a framework that uses distributed storage and parallel processing to store and manage Big Data. It is the most commonly used software to handle Big Data. There are three components of Hadoop.

1. Hadoop HDFS - Hadoop Distributed File System (HDFS) is the storage unit of Hadoop.
2. Hadoop MapReduce - Hadoop MapReduce is the processing unit of Hadoop.
3. Hadoop YARN - Hadoop YARN is a resource management unit of Hadoop.

Let us take a detailed look at Hadoop HDFS in this part of the What is Hadoop article.

Hadoop HDFS

Data is stored in a distributed manner in HDFS. There are two components of HDFS - name node and data node. While there is only one name node, there can be multiple data nodes.

HDFS is specially designed for storing huge datasets in commodity hardware. An enterprise version of a server costs roughly $10,000 per terabyte for the full processor. In case you need to buy 100 of these enterprise version servers, it will go up to a million dollars.

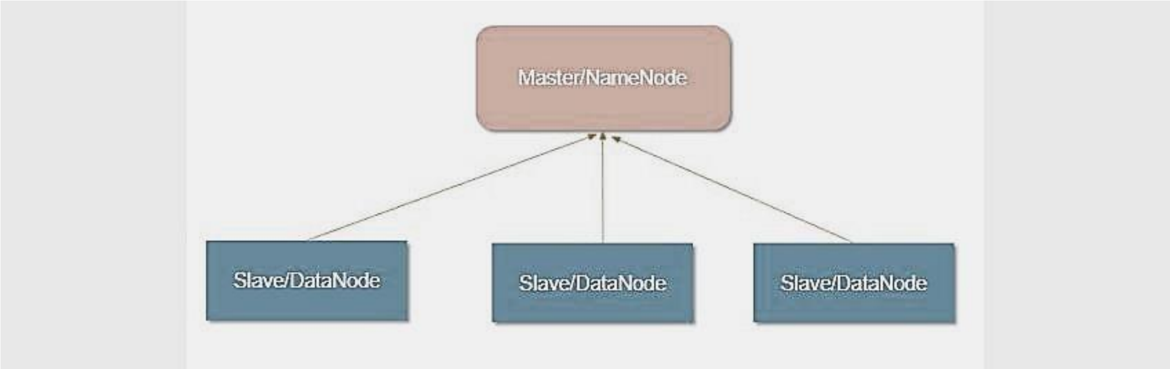
Hadoop enables you to use commodity machines as your data nodes. This way, you don’t have to spend millions of dollars just on your data nodes. However, the name node is always an enterprise server.

Features of HDFS

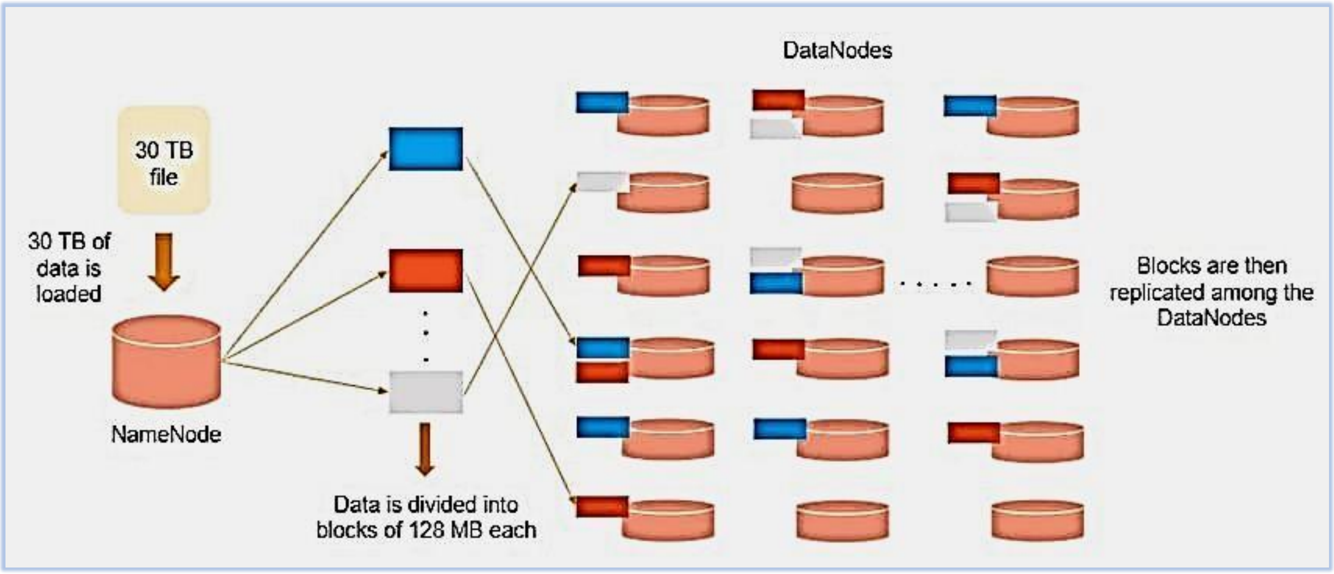
- Provides distributed storage
- Can be implemented on commodity hardware
- Provides data security
- Highly fault-tolerant - If one machine goes down, the data from that machine goes to the next machine

Master and Slave Nodes

Master and slave nodes form the HDFS cluster. The name node is called the master, and the data nodes are called the slaves.



The name node is responsible for the workings of the data nodes. It also stores the metadata. The data nodes read, write, process, and replicate the data. They also send signals, known as heartbeats, to the name node. These heartbeats show the status of the data node.


DataNodes
30 TB file
30 TB of data is loaded
NameNode
Data is divided into blocks of 128 MB each
Blocks are then replicated among the DataNodes

Consider that 30TB of data is loaded into the name node. The name node distributes it across the data nodes, and this data is replicated among the data notes. You can see in the image above that the blue, grey, and red data are replicated among the three data nodes.

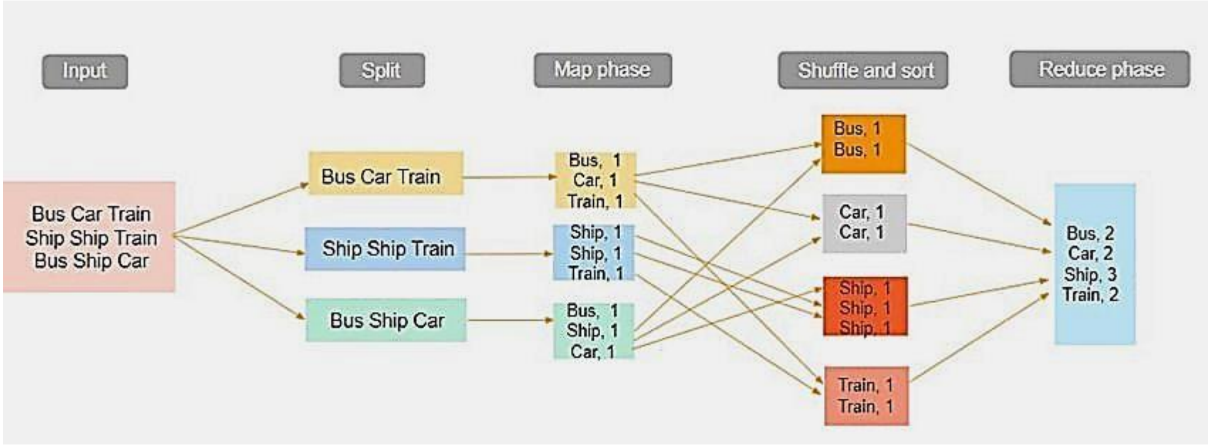
Replication of the data is performed three times by default. It is done this way, so if a commodity machine fails, you can replace it with a new machine that has the same data.

Let us focus on Hadoop MapReduce in the following section of the What is Hadoop article.

Hadoop MapReduce

Hadoop MapReduce is the processing unit of Hadoop. In the MapReduce approach, the processing is done at the slave nodes, and the final result is sent to the master node.

A data containing code is used to process the entire data. This coded data is usually very small in comparison to the data itself. You only need to send a few kilobytes worth of code to perform a heavy-duty process on computers.



The input dataset is first split into chunks of data. In this example, the input has three lines of text with three separate entities - “bus car train,” “ship ship train,” “bus ship car.” The dataset is then split into three chunks, based on these entities, and processed parallelly.

In the map phase, the data is assigned a key and a value of 1. In this case, we have one bus, one car, one ship, and one train.

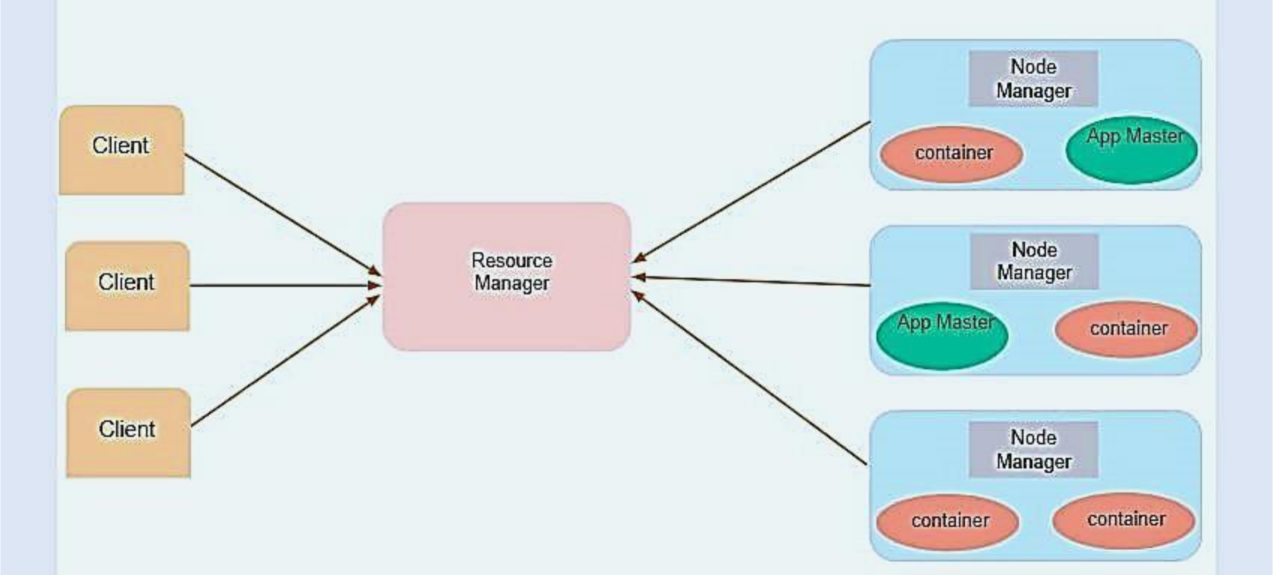
These key-value pairs are then shuffled and sorted together based on their keys. At the reduce phase, the aggregation takes place, and the final output is obtained.

Hadoop YARN is the next concept we shall focus on in the What is Hadoop article.

Hadoop YARN

Hadoop YARN stands for Yet Another Resource Negotiator. It is the resource management unit of Hadoop and is available as a component of Hadoop version 2.

- Hadoop YARN acts like an OS to Hadoop. It is a file system that is built on top of HDFS.
- It is responsible for managing cluster resources to make sure you don't overload one machine.
- It performs job scheduling to make sure that the jobs are scheduled in the right place



Suppose a client machine wants to do a query or fetch some code for data analysis. This job request goes to the resource manager (Hadoop Yarn), which is responsible for resource allocation and management.

In the node section, each of the nodes has its node managers. These node managers manage the nodes and monitor the resource usage in the node. The containers contain a collection of physical resources, which could be RAM, CPU, or hard drives. Whenever a job request comes in, the app master requests the container from the node manager. Once the node manager gets the resource, it goes back to the Resource Manager.

How Does Hadoop Work?

The primary function of Hadoop is to process the data in an organised manner among the cluster of commodity software. The client should submit the data or program that needs to be processed. Hadoop HDFS stores the data. YARN, MapReduce divides the resources and assigns the tasks to the data. Let's know the working of Hadoop in detail.

- The client input data is divided into 128 MB blocks by HDFS. Blocks are replicated according to the replication factor: various DataNodes house the unions and their duplicates.
- The user can process the data once all blocks have been put on HDFS DataNodes.
- The client sends Hadoop the MapReduce programme to process the data.
- The user-submitted software was then scheduled by ResourceManager on particular cluster nodes.
- The output is written back to the HDFS once processing has been completed by all nodes.

Hadoop Distributed File System:

HDFS is known as the Hadoop distributed file system. It is the allocated File System. It is the primary data storage system in Hadoop Applications. It is the storage system of Hadoop that is spread all over the system. In HDFS, the data is once written on the server, and it will continuously be used many times according to the need. The targets of HDFS are as follows.

- The ability to recover from hardware failures in a timely manner
- Access to Streaming Data
- Accommodation of Large data sets
- Portability

Hadoop Distributed File System has two nodes included in it. They are the Name Node and Data Node.

Name Node:

Name Node is the primary component of HDFS. Name Node maintains the file systems along with namespaces. Actual data can not be stored in the Name Node. The modified data, such as Metadata, block data etc., can be stored here.

Data Node:

Data Node follows the instructions given by the Name Node. Data Nodes are also known as 'slave Nodes'. These nodes store the actual data provided by the client and simply follow the commands of the Name Node.

Job Tracker:

The primary function of the Job Tracker is resource management. Job Tracker determines the location of the data by communicating with the Name Node. Job Tracker also helps in finding the Task Tracker. It also tracks the MapReduce from Local Node to Slave Node. In Hadoop, there is only one instance of Job Trackers. Job Tracker monitors the individual Task Tracker and tracks the status. Job Tracker also helps in the execution of MapReduce in Hadoop.

Task Tracker:

Task Tracker is the slave daemon in the cluster which accepts all the instructions from the Job Tracker. Task Tracker runs on its process. The task trackers monitor all the tasks by capturing the input and output codes. The Task Tracker helps in mapping, shuffling and reducing the data operations. Task Tracker arranges different slots to perform different tasks. Task Tracker continuously updates the status of the Job Tracker. It also informs about the number of slots available in the cluster. In case the Task Tracker is unresponsive, then Job Tracker assigns the work to some other nodes.

How Hadoop Improves on Traditional Databases

Understanding what is Hadoop requires further understanding on how it differs from traditional databases.

Hadoop uses the HDFS (Hadoop Data File System) to divide the massive data amounts into manageable smaller pieces, then saved on clusters of community servers. This offers scalability and economy.

Furthermore, Hadoop employs MapReduce to run parallel processings, which both stores and retrieves data faster than information residing on a traditional database. Traditional databases are great for handling predictable and constant workflows; otherwise, you need Hadoop's power of scalable infrastructure.

How Is Hadoop Being Used?

Hadoop is being used in different sectors to date. The following sectors have the usage of Hadoop.

1. Financial Sectors:

Hadoop is used to detect fraud in the financial sector. Hadoop is also used to analyse fraud patterns. Credit card companies also use Hadoop to find out the exact customers for their products.

2. Healthcare Sectors:

Hadoop is used to analyse huge data such as medical devices, clinical data, medical reports etc. Hadoop analyses and scans the reports thoroughly to reduce the manual work.

3. Hadoop Applications in the Retail Industry:

Retailers use Hadoop to improve their sales. Hadoop also helped in tracking the products bought by the customers. Hadoop also helps retailers to predict the price range of the products.

Hadoop also helps retailers to make their products online. These advantages of Hadoop help the retail industry a lot.

4. Security and Law Enforcement:

The National Security Agency of the USA uses Hadoop to prevent terrorist attacks. Data tools are used by the cops to chase criminals and predict their plans. Hadoop is also used in defence, cybersecurity etc.

5. Hadoop Uses in Advertisements:

Hadoop is also used in the advertisement sector too. Hadoop is used for capturing video, analysing transactions and handling social media platforms. The data analysed is generated through social media platforms like Facebook, Instagram etc. Hadoop is also used in the promotion of the products.

There are many more advantages of Hadoop in daily life as well as in the Software sector too.

Challenges of Using Hadoop

Despite Hadoop's awesomeness, it's not all hearts and flowers. Hadoop comes with its own issues, such as:

- There's a steep learning curve. If you want to run a query in Hadoop's file system, you need to write MapReduce functions with Java, a process that is non-intuitive. Also, the ecosystem is made up of lots of components.

- Not every dataset can be handled the same. Hadoop doesn't give you a "one size fits all" advantage. Different components run things differently, and you need to sort them out with experience.

- MapReduce is limited. Yes, it's a great programming model, but MapReduce uses a file-intensive approach that isn't ideal for real-time interactive iterative tasks or data analytics.

- Security is an issue. There is a lot of data out there, and much of it is sensitive. Hadoop still needs to incorporate the proper authentication, data encryption, provisioning, and frequent auditing practices..

1.7.1 Open source technologies

There are several open-source technologies available for big data analytics. These technologies provide the tools and frameworks necessary to process, analyze, and derive insights from large volumes of data. Here are some popular open-source technologies used in big data analytics:

1. Apache Hadoop: Apache Hadoop is one of the most widely used open-source frameworks for distributed storage and processing of big data. It provides the Hadoop Distributed File System (HDFS) for distributed storage and the MapReduce programming model for distributed processing.

2. Apache Spark: Apache Spark is a fast and general-purpose cluster computing system that provides in-memory processing capabilities for big data analytics. It supports various programming languages and offers a rich set of libraries for data analysis, machine learning, and graph processing.

3. Apache Kafka: Apache Kafka is a distributed streaming platform used for building realtime data pipelines and streaming applications. It is designed to handle highthroughput, fault-tolerant, and scalable streaming of data.
4. Apache Flink: Apache Flink is a powerful open-source stream processing framework that supports both batch and real-time processing. It provides low-latency and highthroughput processing of streaming data and supports advanced analytics, event time processing, and state management.
5. Apache Cassandra: Apache Cassandra is a highly scalable and distributed NoSQL database designed to handle large amounts of data across multiple commodity servers. It provides high availability and fault tolerance and is optimized for write-heavy workloads.
6. Elasticsearch: Elasticsearch is a distributed, real-time search and analytics engine built on top of Apache Lucene. It allows you to store, search, and analyze large volumes of data in near real-time, making it suitable for various big data analytics use cases.
7. Apache Druid: Apache Druid is a high-performance, real-time analytics database designed for sub-second query response times. It is optimized for OLAP (Online Analytical Processing) workloads and can handle large-scale data exploration and analysis.
8. R and Python: While not specific technologies, R and Python are popular open-source programming languages commonly used for big data analytics. They provide rich libraries and frameworks, such as R's "tidyverse" and Python's "pandas" and "NumPy," for data manipulation, analysis, and visualization.

These are just a few examples of the open-source technologies available for big data analytics. Each technology has its own strengths and use cases, so it's important to evaluate your specific requirements before choosing the right tool for your big data analytics needs.

1.7.2 The Cloud and Big Data

Cloud computing and big data are two interconnected technologies that work together to provide scalable, flexible, and efficient solutions for managing and analyzing large datasets.

Cloud Computing and Big Data Integration:

Let's dive into each component:

1. **Cloud Infrastructure**: Cloud computing offers a virtualized infrastructure that provides computing resources, such as virtual machines, storage, and networks, ondemand. This infrastructure can be accessed and managed remotely over the internet. Cloud infrastructure can be provided by public cloud providers like AWS, Azure, or GCP, or it can be a private cloud within an organization's data center.
2. **Data Storage and Processing Services**: Cloud platforms provide various services for storing and processing data. These services include cloud-based storage solutions like

Amazon S3, Azure Blob Storage, or Google Cloud Storage, which offer scalable and durable storage for big data. Additionally, cloud providers offer managed data processing services like AWS Glue, Azure Data Lake Analytics, or Google BigQuery, which allow for efficient data processing and querying at scale.

3. **Big Data Processing Frameworks**: Big data processing frameworks, such as Apache Hadoop, Apache Spark, or Apache Flink, are designed to handle large volumes of data and perform distributed processing. These frameworks enable parallel computation across clusters of machines and provide tools and libraries for tasks like data ingestion, transformation, analysis, and machine learning.
4. **Scalable Computing Resources**: Cloud computing offers the advantage of elastic scalability, allowing users to scale computing resources up or down based on their needs. This scalability is particularly valuable in big data scenarios, where processing large datasets can require substantial computing power. Cloud providers can dynamically allocate resources to accommodate the varying demands of big data workloads.
5. **Data Analytics Applications**: Cloud-based big data solutions enable the development and deployment of data analytics applications. These applications leverage the processing power and scalability of the cloud infrastructure to perform advanced analytics, including data mining, machine learning, real-time analytics, and predictive analysis on large datasets. Cloud platforms offer the necessary tools, APIs, and services to build and deploy these applications seamlessly.

By integrating cloud computing and big data, organizations can leverage the advantages of both technologies. Cloud infrastructure provides the necessary resources and scalability to store and process large datasets, while big data processing frameworks enable efficient distributed computing and advanced analytics. This integration empowers organizations to derive valuable insights, make data-driven decisions, and effectively manage their big data workloads.

1.7.3 Mobile Business Intelligence

The ability to access analytics and data on mobile devices or tablets rather than desktop computers is referred to as mobile business intelligence.

Here's how Mobile BI leverages big data technology:

1. **Mobile Device Accessibility**: Mobile BI enables users to access and interact with big data analytics platforms and tools through mobile devices such as smartphones and tablets. These devices provide a convenient and portable means of accessing large datasets, reports, and visualizations anytime, anywhere.
2. **Data Integration**: Mobile BI integrates with big data platforms and data sources to retrieve, process, and analyze large volumes of structured and unstructured data. It can connect to data warehouses, data lakes, cloud-based storage systems, and real-time data streams, ensuring that users have access to comprehensive and up-to-date information.
3. **Real-time Data Processing**: Big data technologies such as Apache Hadoop, Apache Spark, or cloud-based analytics platforms are utilized to process and analyze large datasets in real-time. Mobile BI leverages these technologies to provide real-time insights and analytics, enabling users to monitor key performance indicators, track trends, and respond to business events in real-time.

4. **Data Visualization and Dashboards**: Mobile BI utilizes big data visualization techniques to present complex data in a user-friendly and visually appealing format on mobile devices. It offers interactive dashboards, charts, graphs, and maps that allow users to explore data, drill-down into details, and gain insights from visual representations of the data.
5. **Advanced Analytics and Machine Learning**: Mobile BI leverages big data analytics techniques such as data mining, predictive analytics, and machine learning to provide advanced insights and predictive capabilities on mobile devices. Users can perform complex analytics tasks, such as forecasting, clustering, anomaly detection, and predictive modeling, leveraging the power of big data analytics algorithms.
6. **Data Security and Privacy**: Mobile BI in big data technology emphasizes data security and privacy. It incorporates measures such as user authentication, data encryption, access controls, and data anonymization to protect sensitive business information. Compliance with data protection regulations is also ensured.
7. **Collaboration and Sharing**: Mobile BI facilitates collaboration and sharing of insights among users. It enables users to share reports, dashboards, and analysis results with team members, fostering collaboration, and enabling collective decision-making.

By combining mobile devices and big data technology, Mobile BI in big data enables organizations to leverage the power of real-time analytics, advanced insights, and data visualization on mobile devices. It empowers users with the flexibility and agility to access, analyze, and act upon large volumes of data, leading to more informed and data-driven decision-making.

Need for mobile BI?

Mobile phones' data storage capacity has grown in tandem with their use. You are expected to make decisions and act quickly in this fast-paced environment. The number of businesses receiving assistance in such a situation is growing by the day.

To expand your business or boost your business productivity, mobile BI can help, and it works with both small and large businesses. Mobile BI can help you whether you are a salesperson or a CEO. There is a high demand for mobile BI in order to reduce information time and use that time for quick decision making.

As a result, timely decision-making can boost customer satisfaction and improve an enterprise's reputation among its customers. It also aids in making quick decisions in the face of emerging risks.

Data analytics and visualisation techniques are essential skills for any team that wants to organise work, develop new project proposals, or wow clients with impressive presentations.

Advantages of mobile BI

1. Simple access

Mobile BI is not restricted to a single mobile device or a certain place. You can view your data at any time and from any location. Having real-time visibility into a firm improves production and the daily efficiency of the business. Obtaining a company's perspective with a single click simplifies the process.

2. Competitive advantage

Many firms are seeking better and more responsive methods to do business in order to stay ahead of the competition. Easy access to real-time data improves company opportunities and raises sales and capital. This also aids in making the necessary decisions as market conditions change.

3. Simple decision-making

As previously stated, mobile BI provides access to real-time data at any time and from any location. During its demand, Mobile BI offers the information. This assists consumers in obtaining what they require at the time. As a result, decisions are made quickly.

4. Increase Productivity

By extending BI to mobile, the organization's teams can access critical company data when they need it. Obtaining all of the corporate data with a single click frees up a significant amount of time to focus on the smooth and efficient operation of the firm. Increased productivity results in a smooth and quick-running firm.

1.7.3 Crowd sourcing analytics

Crowdsourcing is a sourcing model in which an individual or an organization gets support from a large, open-minded, and rapidly evolving group of people in the form of ideas, microtasks, finances, etc. Crowdsourcing typically involves the use of the internet to attract a large group of people to divide tasks or to achieve a target. Crowdsourcing can help different types of organizations get new ideas and solutions, deeper consumer engagement, optimization of tasks, and several other things.

Crowdsourcing analytics in big data technology involves harnessing the collective intelligence and efforts of a crowd of individuals to analyze and extract insights from large and complex datasets. It leverages the power of distributed human intelligence to tackle data analysis challenges that may be difficult or time-consuming for traditional approaches. Here's how crowdsourcing analytics can be applied in big data technology:

1. **Problem Decomposition**: Crowdsourcing analytics breaks down complex data analysis tasks into smaller, more manageable components. These tasks can include data labeling, categorization, sentiment analysis, data verification, data cleaning, image or text recognition, and other similar activities.
2. **Crowdsourcing Platforms**: Online crowdsourcing platforms, such as Amazon Mechanical Turk, Kaggle, or specialized platforms, are used to connect organizations or individuals with a crowd of contributors. These platforms provide a marketplace where individuals can participate in data analysis tasks and earn rewards or incentives for their contributions.

3. **Diverse Skillsets**: Crowdsourcing analytics allows organizations to tap into a diverse pool of talent with varying backgrounds, skillsets, and expertise. This diversity brings different perspectives and knowledge to the data analysis process, potentially leading to more comprehensive and accurate insights.
4. **Data Quality Control**: Ensuring the quality and reliability of crowd-contributed data analysis results is critical. Quality control mechanisms, such as validation by multiple contributors, consensus-based decision-making, expert reviews, and statistical techniques, are employed to filter out errors, noise, or biased responses and enhance the accuracy and reliability of the aggregated results.
5. **Scalability and Efficiency**: Crowdsourcing analytics can provide scalability and efficiency in handling large volumes of data. By distributing tasks to a crowd, organizations can tap into the collective power of many contributors simultaneously, allowing for faster data analysis and processing.

6. **Gamification and Incentives**: Engaging and motivating the crowd to participate actively in crowdsourcing analytics is essential. Gamification elements, competitions, leaderboards, rewards, and reputation systems are often employed to incentivize contributors, encouraging them to invest their time and effort in performing accurate and high-quality data analysis.
7. **Challenges and Considerations**: Crowdsourcing analytics in big data technology also poses challenges. Ensuring data privacy and security, managing the quality of contributed results, maintaining communication and coordination within the crowd, addressing potential biases or inconsistencies, and providing clear instructions are some considerations that need to be addressed.

Crowdsourcing analytics in big data technology enables organizations to tap into a vast pool of distributed human intelligence to tackle complex data analysis tasks. It offers scalability, diversity of expertise, and the potential for faster and more cost-effective data analysis. However, careful planning, quality control mechanisms, and proper management of the crowd are necessary to ensure the accuracy, reliability, and integrity of the results obtained through crowdsourcing analytics in the context of big data.

Where Can We Use Crowdsourcing?

Crowdsourcing is touching almost all sectors from education to health. It is not only accelerating innovation but democratizing problem-solving methods. Some fields where crowdsourcing can be used.

1. Enterprise
2. IT
3. Marketing
4. Education
5. Finance
6. Science and Health

Examples Of Crowdsourcing

1. **Doritos**: It is one of the companies which is taking advantage of crowdsourcing for a long time for an advertising initiative. They use consumer-created ads for one of their 30-Second Super Bowl Spots(Championship Game of Football).
2. **Starbucks**: Another big venture which used crowdsourcing as a medium for idea generation. Their white cup contest is a famous contest in which customers need to decorate their Starbucks cup with an original design and then take a photo and submit it on social media.
3. **Lays**:" Do us a flavor" contest of Lays used crowdsourcing as an idea-generating medium. They asked the customers to submit their opinion about the next chip flavor they want.
4. **Airbnb**: A very famous travel website that offers people to rent their houses or apartments by listing them on the website. All the listings are crowdsourced by people.

There are several examples of businesses being set up with the help of crowdsourcing.

Crowdsourcing Sites

Here is the list of some famous crowdsourcing and crowdfunding sites.

1. Kickstarter
2. GoFundMe
3. Patreon
4. RocketHub

Advantages Of Crowdsourcing

1. Evolving Innovation: Innovation is required everywhere and in this advancing world innovation has a big role to play. Crowdsourcing helps in getting innovative ideas from people belonging to different fields and thus helping businesses grow in every field.
2. Save costs: There is the elimination of wastage of time of meeting people and convincing them. Only the business idea is to be proposed on the internet and you will be flooded with suggestions from the crowd.
3. Increased Efficiency: Crowdsourcing has increased the efficiency of business models as several expertise ideas are also funded.

1.7.5 Inter and trans firewall analytics

Inter and trans firewall analytics in big data technology involves the analysis of network traffic data collected from multiple firewalls to gain insights into security threats, network behavior, and patterns. It utilizes big data analytics techniques to process and analyze large volumes of firewall logs and other network data for improved security and network management. Here's a breakdown of inter and trans firewall analytics in big data technology:

1. **Data Collection**: Inter firewall analytics involves collecting network traffic data, including firewall logs, from multiple firewalls deployed across different locations or

network segments. The data can include information about incoming and outgoing network connections, traffic volume, protocols used, source and destination IP addresses, port numbers, and timestamps.

2. **Data Integration**: The collected firewall logs and network data are integrated into a central repository or big data platform that can handle the large volumes of data. This integration allows for a comprehensive view of network traffic across multiple firewalls, enabling cross-firewall analysis and correlation.
3. **Data Processing and Analysis**: Big data analytics techniques, such as data mining, machine learning, and statistical analysis, are applied to the integrated firewall data. These techniques help identify patterns, anomalies, and potential security threats. The data can be analyzed in real-time or through batch processing to uncover hidden insights and security risks.
4. **Threat Detection and Prevention**: Inter and trans firewall analytics aim to detect and prevent security threats by analyzing the network traffic data. By applying advanced analytics algorithms, it becomes possible to identify suspicious activities, malware, intrusion attempts, and other security incidents that may span multiple firewalls. This enables proactive threat mitigation and faster incident response.
5. **Network Behavior Analysis**: Inter firewall analytics also focuses on understanding the network behavior and identifying abnormal network activities. By analyzing the aggregated traffic data, it becomes possible to detect anomalies, unusual traffic patterns, and potential network performance issues. This analysis helps optimize network management, capacity planning, and resource allocation.
6. **Visualization and Reporting**: Big data analytics platforms provide visualization and reporting capabilities to present the analyzed data in a clear and intuitive manner. Dashboards, charts, graphs, and reports can help security analysts and network administrators understand the insights derived from the inter and trans firewall analytics. Visual representations aid in identifying trends, outliers, and potential security risks.
7. **Scalability and Performance**: Big data technologies allow for the scalability and performance needed to handle the high volume, velocity, and variety of firewall logs and network traffic data. Distributed storage systems, parallel processing, and cluster computing frameworks enable efficient processing and analysis of large-scale firewall data.

Inter and trans firewall analytics in big data technology enhance security monitoring and network management by providing a holistic view of network traffic across multiple firewalls. It enables the detection of complex threats, identification of network anomalies, and optimization of network performance. By leveraging big data analytics techniques, organizations can improve their security posture and enhance their overall network infrastructure.