International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Big Data Analytics for Business Intelligence and Decision-Making

Shriram Srinivasan

Cambridge A-Level Student

Abstract:

Big data - a boon or a bane?

In today's digital age, big data refers to the large amount of data companies can access for their daily activities. It encompasses data collected from various sources, which can later be interpreted, analyzed, and utilized for the company's benefit. The demand for large volumes arises from the market's need to scrutinize and leverage current data. [1]

With the increasing demand and availability of vast data also comes the need to find trends, patterns, and correlations in enormous amounts of unprocessed data to support data-driven decision-making, this process is known as big data analytics. With the advent of recent tools, these processes apply well-known statistical techniques—like regression and clustering—on larger datasets, so that they can be utilized to their utmost extent.

Big data analytics is a subset of advanced analytics, involving complex applications that include predictive models, statistical algorithms, and what-if analysis powered by analytical systems. Proper and effective utilization and analysis of big data have emerged as a significant challenge for companies worldwide. It involves the intricate process of examining and perusing this large dataset, alongside insightful analysis of hidden patterns, correlations, market trends, and customer preferences. These insights help organizations make informed and intelligent business decisions.

Data analytics technologies and techniques help organizations with the correct and necessary means to analyze datasets and extract new information. Simultaneously, all queries related to Business Intelligence (BI) are also addressed by the various business operations and actions performed.

A prime example of big data analytics is the education industry, where a vast amount of student records, admission progress, assessments taken, and the overall growth of a student, along with other data, needs to be collected, aggregated, processed, and analyzed. In the field of education, big data analytics refers to the utilization of data for accounting, decision-making, predictive analytics, and numerous other purposes. Although this data varies to a colossal extent in type, quality, and accessibility, presenting substantial challenges, it also offers immense benefits. [2]

SOURCES OF BIG DATA

With the emergence of the internet, gathering data on any topic has now become a facile task, and can be done with a single click, making the availability of several sources handy. Businesses obtain information from both internal and external sources. Data obtained from internal sources and collected by the researchers themselves sometimes also referred to as primary data, is often easily accessible; however, data obtained from external references or a third party, also referred to as secondary data, requires effort and time to be scrutinized.

PRIMARY DATA

Primary data is original, independent, and reliable, however its reliability depends on the accuracy of the source of information making it either completely reliable, biased or unreliable. Moreover, such data can also be time-consuming and if collected by an inexperienced researcher this data might also lead to undesirable consequences.

The collection of data and the type of source chosen by the researcher usually depend on the attributes of the respondents and the appropriateness of the circumstances of the field of study.

The main sources from where primary data can be gathered are:

Transactional data refers to the data collected by businesses at different points during the sale of the company's products. Data collected during the point of sale, online purchases, and inventory management is known as retail transactions, and data collected during credit card purchases, ATM withdrawals, and online banking activities is known as banking transactions. The main sources of transactional data include payment orders, invoices, storage records, and e-receipts.[2]

Transaction data if used at the correct point in time brings favorable results for the company; such data is time-sensitive and highly volatile. Companies using transactional data can promptly always succeed in gaining the upper hand in the market. However, it is also worth noting that, transactional data also demands a separate set of expertise to process, analyze, interpret, and manage data. At times businesses might find the interpretation of information and proper utilization of this data challenging and out of their scope.

Machine data is the data that is automatically generated by various machines, being used in a company, during the company's routine work. This data usually comes out in response to specific events or on fixed schedules. Sources of this kind of data include smart sensors, SIEM logs, medical devices, road cameras, IoT devices, satellites, desktops, mobile phones, and industrial machinery. Data extracted from machine sources has risen exponentially along with changes in the external market environment. Various sensors that record this type of data include servers, user applications, websites, and cloud programs. The benefits of machine data also involve an easy understanding of all customer behaviors and less human intervention.[2]

However, machine data comes with its challenges, which include the high cost of storing unlimited data generated automatically by machines. The quality of machine data might also be inconsistent at times. Without proper expertise, researchers might not be able to integrate the data within the specified time, thus reducing the efficiency of such data.

Social Data is one of the most customary data in today's day and age. With the increasing trend in social media, and more than half of the world using social media platforms, data collection from these sources has become conventional. Any data derived from social media platforms such as video uploads, and comments shared on Facebook, Instagram, Twitter, YouTube, LinkedIn, etc. is referred to as Social Data. Social data's quantitative and qualitative insights help companies interact with their customers without any barriers. Social media spreads like wildfire and reaches an extensive audience base within seconds, gauging important insights regarding customer behavior, and their sentiments. Brands advertising their products on social media can build strong connections with their online targeted demographics. And later on, businesses can harness this data to get the best and desired results.[2]

However, issues like spam, fake accounts, or skewed user-generated content can make social media data noisy and inaccurate. Furthermore, because social media data and other types of data (such as transactional and machine data) differ in their formats, structures, and semantics, integrating such data for a thorough analysis can be challenging.

Survey and Feedback Data is a type of data gathered from a small group of people, who represent a larger population. This method of data collection is used to gather insights, opinions, attitudes, behaviors, and demographics, through face-to-face interviews, online surveys, focus groups, etc. The data collected can be used to prepare statistical analysis. Such data can be descriptive and numerical, depending on the questions asked in the survey. Survey data typically uses multiple-choice questions, and open-ended questions, to gather a range of data. In contrast, feedback data depends on reviews, customer support interactions, or responses provided by the focus groups.[2]

To a significant extent, this data is dependent on the personal choices of the person answering the questionnaire. There could be sampling bias, dishonest answers, and privacy issues. Also, long or confusing surveys may lead to fake answers, especially if the respondents are offered a reward for completing the survey.

Experimental Data - The experimental method of collecting data is the process that involves user experiments, and research to collect data systematically and scientifically, it also ensures, the accuracy, reliability, and validity of data. The most frequently used experimental data methods are-[3]

CRD (Completely Randomized design) involves using data analytics based on randomization and replication. This kind of data is mostly used for comparing experiments.[3]

RBD (Randomized Block Design) involves the division of data into small blocks and then the undertaking of random experiments on each block. In this results are derived through the application of the analysis of variance (ANOVA) technique.[3]

LSD (Latin Square Design) involves an NxN of rows and columns where each row and column, contains a unique letter appearing only once per row. This experimental design is similar to CRD and RBD. This structure allows the identification of differences with minimal errors in experimentation. The sudoku puzzle is a common example of a Latin square design.[3]

FD (Factorial design) involves an experimental setup in which each experiment includes two factors and each factor has its own set of values. Thus, various combinations of these factors are generated by conducting trials.[3]

Furthermore, the use of the experimental method of data collection is costly and time-consuming. It also includes practical problems like difficulty in the manipulation and measurement of variables, controlling extraneous variables or errors like sampling bias or experimental bias. At times the companies might also encounter the issue of the pseudoreplication (a statistical error that occurs when data points are treated as independent when they are not) of the data. Such factors to a very large extent influence the accuracy and validity of results.[3]

SECONDARY DATA

Secondary data, which is collected from external sources, is a manageable, time efficient and cost-effective technique of data collection. However, it comes with its own risk of having been previously utilized by the collector. Thus it cannot be referred to as first-hand or original data. Various types of secondary data are -

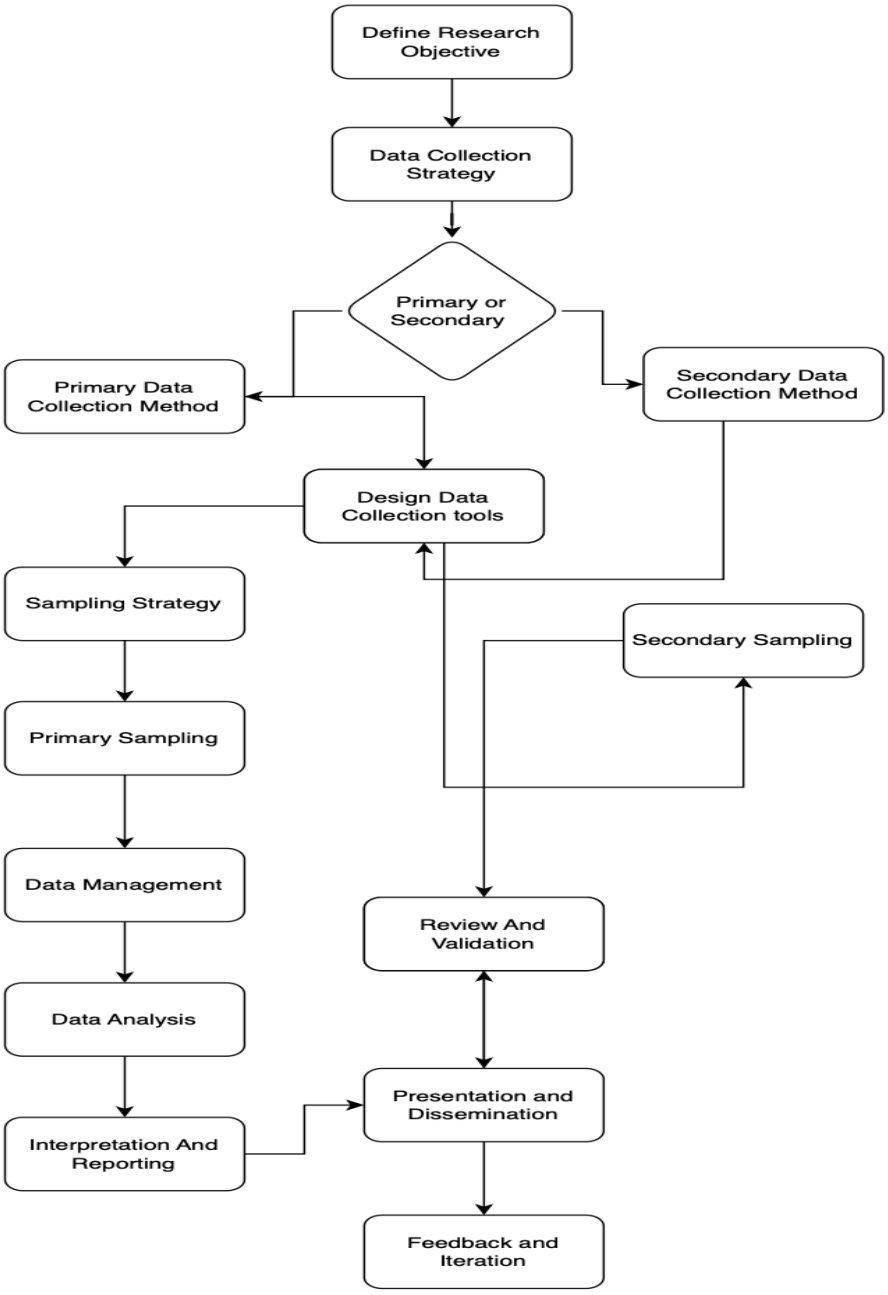
Blogs are one of the most common sources of online information, but their reliability can sometimes be questionable. Nowadays, almost everyone has a blog, which helps attract traffic to their website and allows them to earn money through advertisements. The credibility of a blog depends on the quality, the content and the blogger's positive or negative views about the product.[4]

Publications by renowned organizations, Prominent national and international organizations such as the

ICMR, WHO, and others conduct regular surveys and create their case studies, which they subsequently post on their websites. Anyone can access the data and statistics of all surveys from the official website of the organization conducting the survey.[4]

Books, Magazines, Newspapers, and articles, Most of the data collected through primary sources, gets transformed into secondary data after being published in newspapers, magazines, books, etc. This also includes research papers published on various websites by academics and scientists in related fields of medicine, finance, economics, etc.[4]

Government records are a vital and authentic source of secondary data, providing valuable information for research in almost all fields, like, marketing, management, humanities, and social sciences. The most common examples of government records include census data, health records, and educational institution records. These records are typically gathered to assist in effective planning, fund allocation, and project prioritization.[4]



A diagram explaining the flow of process that a business might go through before and after choosing their data collection type.

The Role of Big Data in Business Intelligence

Businesses all over the globe use big data to gain a competitive edge and insights into customer behavior, market trends, and operational efficiency. These insights help businesses to make faster, smarter, and correct decisions with minimal risk. Based on the analysis of big data, companies prepare predictive models, patterns, and trends on how people interact or respond. Various steps involved in business intelligence include collecting, analyzing, and interpreting data. Big data plays a critical role in business intelligence by providing the raw material needed for any kind of analysis. Also, analysis of big data helps organizations to identify trends, patterns, correlations, etc. It is only through the means of big data that hidden data turns out to be beneficial.[5][6]

Big data helps not only retailers but also wholesalers to successfully understand market trends and decide their course of action. Such analysis also enables them to understand customer behavior across multiple channels within a short period. Based on the collected information, businesses can organize personalized marketing campaigns, improve product quality, and gain customer loyalty.[5]

Another approach that big data impacts business decisions is by enabling predictive analysis. Analyzing historical data helps businesses identify patterns and predict preferences. This involves decision-making at every step, from product development to inventory management (the whole supply chain). By correctly utilizing appropriate tools and strategies, businesses can harness big data to achieve a competitive edge and foster growth.[6]

With the growing advent of big data, multiple technologies have been introduced. The most common technologies include big data infrastructure, big data management, big data processing, and big data analytics.[6]

Features of Big Data Analytics in Business Intelligence

1. Enhanced data integration–

One of the most important features of Big Data analysis is the correct and appropriate integration of data so that the desired results can be achieved. Data integration involves collecting and combining heterogeneous data from various sources in one place. Data integration encompasses data replication, ingestion, and transformation to standardize different data types for storage in target repositories such as data warehouses, data lakes, or data lakehouses. The next pivotal step is to ensure that the data is coherent and viable for meaningful insights and resolutions.[6][7]

There are five different approaches, or patterns, to execute data integration: ETL (Extract, Transform, and Load), ELT(Extract, Load, and Transform), Data streaming, API (application integration), and data virtualization.

ETL—---For decades, **ETL** has been one of the most common methods to extract, cleanse, and consolidate data into a data warehouse. Due to its feature to ensure comprehensive data extraction, it also remains the most viable option. The software used for ETL is capable of extracting data from source systems, transforming it into the desired format, and loading it into the destination system. Often, this process is largely automated and guided by data maps that facilitate seamless data flows. With changing times and trends, modern ETL software has been simplified to enable serverless ETL processing. Today, instead of spending days developing connectors between applications, ETL teams can create data maps and automate

tasks within seconds. This significantly enhances the efficiency and reliability of results achieved through this method.[6][7]

One of the biggest downsides to modern ETL processes (streamlined and automated as they can be) is the complexity that such a setup can have in managing and debugging data pipelines. As ETL systems get more automated and sophisticated, the processes entailing them, together with the data flow, become correspondingly intricate, and more complex to trace. This can then further create a problem when problems do arise and are diagnosed, as it is difficult to resolve them without data inconsistencies or errors that may not always be so immediately apparent. Moreover, automated systems may introduce other types of errors or dependencies that are not so easily predictable or controllable.[6][7]

ELT—--Another competitive method of analyzing data is ELT. ELT includes the extraction, loading, and transformation of data to get the desired results. This approach is comparatively quicker because in this the data is directly loaded into the destination data warehouse which is followed by the transformation of data. This approach is more appropriate when data sets are large and timelines need to be followed since loading is often faster. The ELT approach is primarily used for streaming data rather than batch processing, it also transforms chunks of data rather than large volumes at once making it an excellent choice for the financial, medical, and engineering sectors.[6][7]

It is worth noting that the negative side of the ELT approach includes a high load on the target data warehouse. Since data goes into the repository before being transformed, the system has to bear the initial volume of the raw data. This implies higher storage costs and possible lag in the data warehouse performance if the system has not been prepared for such loads. Also, since the transformation of data happens right after loading, its temporary consumption of storage space for raw data is extremely high and requires high computational resources to handle the processing, which eventually may influence the overall efficiency and performance of the system.[6][7]

Data Streaming —- Near-real-time data is valuable for operational intelligence and event-driven systems. Data streaming plays a crucial role in this by continuously collecting, transmitting, and updating information, ensuring constant availability. Data streaming represents a steady flow of information generated by various sources.[7][8]

In the process of data streaming, data is continuously collected and transmitted in high volumes. Organizations typically have multiple data sources that record data simultaneously. The size of streaming data can range from a few bytes to several megabytes, partly due to the inclusion of events and sensor data. This type of data is invaluable for real-time analytics, providing businesses with insights into various aspects of their operations. For example, by continuously analyzing clickstreams and customer posts from social media, companies can track changes in public sentiment regarding their brands or products.[7][8][9]

Data streaming is characterized by its chronological significance, continuous flow, uniqueness, non-homogeneity, and imperfections.[8]

One of the drawbacks of data streaming is how complex it can get to manage and process this continuous stream of data; it strains infrastructures and resources. For this, streaming systems require robust and scalable solutions that will handle high velocities of data. Ensuring quality in a real-time data stream could be really tricky, considering different sources and formats. This level of complexity brings a host of latency, integration, and scalability challenges along with it, making the expectation of consistent and accurate insights hard to keep up with operational demands in a streaming environment.[7][8][9]

Application Integration

The main aim of Application integration is to allow smooth transfer and synchronization of data between various applications, this type of application is most commonly used to ensure that the data in the human resource system and finance system, of any organization, matches with each other. Application integration systems need to maintain consistency between the various data sets in an organization. Most of the time it is observed that various applications like API(Application Programming Interfaces) for data exchange and, SaaS(Software as a Service) application automation tools assist in creating and managing native API integrations efficiently and at scale. Application integration involves the integration approach in which several chips, in close proximity and fine alignment, are coupled together. In this, each process works in collaboration with another in order to get increased operational efficiency. It involves four levels; presentation-level integration, business process integration, data integration, and communications-level integration.[7][10]

Data Virtualization

Data virtualization technology is a new way of managing data that completely changes everything and is more flexible and comprehensive. With the use of data virtualization techniques, instead of having to extract, transform, and load data (ETL) into separate storage systems using conventional methods; organizations can create virtual representations called views that show the information in real time across various environments such as cloud solutions, data warehouses even big data repositories.[7][11]

Moreover, underneath these views, there is a semantic abstraction layer that simplifies the complexities associated with underlying data structures and allows applications, business intelligence tools, and analytic platforms to access data seamlessly. This technology directly puts together data from different sources, without moving it physically or duplicating it, which speeds up time-to-insight, secures centralized access controls for improved security of information, and facilitates quick integration of new sources for agile business processes.[7][11]

However, data virtualization can occasionally cause latency problems because it entails accessing and integrating data from multiple sources in real time. Network latency, processing overhead from data integration or transformation, or competition for resources within the virtualization layer itself can all contribute to these delays. Therefore, in comparison to direct data integration techniques, queries or operations that significantly rely on data virtualization may encounter slower response times.[7][11]

2. Real-time analytics

Another crucial feature of big data analytics in BI is real-time analytics as it allows immediate insights and actions on the data that has been collected from various sources. This process analyzes data as soon as it is generated making the results prompt. Since the data collected is processed and analysed instantly it enables users to make faster decisions based on the current available information. With the help of real-time analytics, businesses can remain proactive and make decisions as soon as data is generated and received by their system. Various real-time apps installed by the companies respond to queries within seconds. An important feature of real-time analytics is the ability to gather large amounts of data with high velocity and low reaction time. Depending on the type of analytics it can be on-demand and uninterrupted. On-demand Analytics notifies the result only when the user requests for it. This type of analytics is used when employees or executives need to search for information about an ongoing problem or search for areas that are niché. Continuous real-time analytics is a type of analysis where the conditions

for notifications are an important issue for observing events such as tracking sales of various branches of a company on a daily, weekly, or monthly basis.[7][12][13]

However, it is noticeable that due to the high pace of real-time analytics sometimes it gets difficult to maintain a high quality of accuracy and scalability. Also, it is crucial to know that to ensure minimal delay in data processing and analysis with such high volume and velocity of big data, achieving true real-time processing requires robust infrastructure and optimized algorithms.[7][12][13]

3. The four main pillars of analytics

With advanced algorithms and machine learning, Big Data Analytics has also developed the ability to predict future trends and prescribe actions accordingly. Predictive and prescriptive analytics work with two more types of Analytics known as Descriptive and Diagnostic Analytics. Together these forces work as the four pillars of the process of Analytics.

1. **Descriptive Analytics—--** Observing and understanding the trends of what happened in the past is the basic step to be taken by any organization to be able to decide their future course of action. Descriptive analytics helps organizations and businesses to understand the trends that have been going on in the past.[14][15][16]

Scrutinizing the data using descriptive Analytics can help to set patterns and trends, making it easier to analyze and study them. Descriptive analytics also helps businesses answer questions like how much sales have been made and the reason behind the success or failure to achieve their goals.

Additionally, the application of descriptive analytics also helps the non-data analysts to understand the metrics that they want to study. It turns the stream of facts collected by businesses into information that can be acted upon. Examples of descriptive analytic techniques include

- Annual revenue reports
- Survey response summaries
- Year-over-year sales reports

Regardless of the various positive aspects, the biggest limitation that an analyst faces while using descriptive analytics is that it remains limited to analyzing data from past events. Also, brainstorming and the ability to develop possible responses, and solutions and choosing the correct path to move forward with the analysis depends on the ability of the team handling the data.[14][15][16]

2. **Diagnostic Analytics**—--- Diagnostic analytics is the analysis that goes beyond the description of what has already happened to understand the reason why events occurred the way they had. This is done by building upon descriptive analytics, which is more or less a report based on past events, to give the 'why' behind trends and patterns. Diagnostic Analytics identifies various patterns and examines them to explain their occurrence. Once that is done, strategies can be developed with more effectiveness rather than trial & error. This type of analytics is useful for businesses trying to reproduce positive results and steer clear of the negative aspects.[14][15][16]

Examples of diagnostic analytic techniques include

- The reason behind the loss of customers in a particular quarter of the year?
- The reasons behind the better performance of a product or service?
- Why did the annual sales of a particular product increase or decrease?

One major limitation of diagnostic analytics is that the information attained concerns past events and, cannot help in coming up with very relevant actionable insights for improving the plan of action. While some businesses may find this analytics sufficient to understand the causal relationships and sequences,

others might need advanced analytics to come up with a deeper interpretation. For such businesses, managing huge datasets normally calls for more advanced solutions like predictive or prescriptive analytics, along with other accompanying tools for the extraction of useful insights.[14][15][16]

3. **Predictive analytics**—- aims to forecast future outcomes by analyzing historical data and identifying trends. Unlike descriptive analytics, which interprets current and past data directly, Predictive Analytics uses statistical modeling and machine learning techniques to estimate probabilities and make educated guesses about what might happen in the future.[14][15][16]

The process starts by defining the specific outcomes to make predictions. Analysts then aggregate relevant data and use various models to project potential future scenarios. Outlining a range of possible future outcomes helps businesses set realistic goals, mitigate risks, and make informed decisions.[14][15][16]
Predictive Analytics enables organizations to plan more effectively and address past shortcomings by focusing on potential future scenarios, ultimately contributing to more strategic and result-oriented decision-making.[14][15][16]
Examples of predictive analytics include:

- It allows marketers to analyze the elasticity of demand amongst the customers for a new campaign or product.
- The use of customer's browsing and purchasing history to recommend products to them for their next purchase.
- Helping financial organizations decide whether a customer will or will not pay their credit card bill on time (based on their payment history).

The main challenge with predictive analytics is that its insights are limited to the available data. Also, a matter of concern is that smaller or incomplete data sets lead to less accurate predictions. While sufficient data is needed for reliable business intelligence (BI), what qualifies as "sufficient" varies by industry, business, audience, and use case. Moreover, predictive analytics lacks the capability to consider intangible or human factors, such as sudden economic changes or weather variations, which can also significantly impact outcomes.[14][15][16]

4. **Prescriptive analytics**—-- Prescriptive Analytics uses different sources such as statistics, machine learning, and data mining to provide specific recommendations to businesses so that they can make effective business decisions ("What one should do") and propose guidelines on the best course of action for future outcomes. This being the most advanced type of analytics, offers actionable insights rather than just raw data. Data-driven recommendations remove human intervention and also the risk of personal bias. This approach focuses on what should happen instead of what could have happened.[14][15][16]

With prescriptive analytics, entrepreneurs can anticipate future outcomes and understand why they will occur. It predicts the consequences or impacts of decisions and the domino effects across various business areas, regardless of the logical and chronological order of decisions made.[14][15][16]
Examples of prescriptive analytics include:

- Assessing the economic conditions to determine appropriate interest rates for bank loans.
- Recognising the new features (by analyzing data like customer surveys and market research to identify what features are most desirable for customers and prospects) that should be included in a product to ensure its success in the market.
- Developing strategies to enhance patient care in healthcare, such as evaluating the likelihood of future health issues and tailoring treatment plans to mitigate those risks.

Despite its positives, prescriptive analytics is complex, involves multiple variables and tools, and includes algorithms and big data, making the involvement of effective data infrastructure crucial to be able to manage it successfully. Training and evaluating such models precisely is time-consuming. Even using the most efficacious AutoML tools, which get analytics up and running with a default model, businesses might still have to tune parameters to fit the use case better and test & evaluate the model with new data to see whether the recommendations generated are according to the expectations of businesses.

Various business organizations can choose between these four types of business analysis when they are ready to make the best of the data that has been collected from various sources. The choice of the right type of analytics can also depend on the questions businesses want to be answered and the decisions they need to make.[14][15][16]

Challenges in Big Data Analytics

Along with the numerous benefits of Big data, come multiple huge risks, that cannot be ignored, by any user. However, a thorough understanding of the various challenges and the correct strategies to overcome them is a crucial factor that needs to be taken into consideration by organizations to be able to get fruitful results from the data being analyzed. These risks stem from the fact that analytical tools involve storing, managing, and analyzing diverse data collected from numerous sources.

1. Data Privacy and Security—Cybersecurity and Privacy

Data privacy and security are significant challenges in big data analytics because big data systems often collect and store large amounts of personal data from multiple sources. Collection, storage, and scrutiny of data make individuals vulnerable to exposure due to the aggregation and analysis of specific behavioral data. This can lead to the collection of more information than necessary, increasing the likelihood of privacy and security breaches. Such a breach is usually not acceptable to people whose data is being collected and used by companies. Prioritizing data privacy helps build and maintain customer trust, especially in an era where data breaches are common. Focusing on protecting data enhances the reputation and credibility of companies.

Additionally, strict regulations across various jurisdictions mandate proper data collection and usage, with non-compliance leading to significant legal and financial consequences. Beyond fines, data breaches can result in costly damage control, compensation, and loss of business to companies. Moreover, safeguarding user data is an ethical responsibility, as it respects individuals' rights to privacy, recognizing that data represents real people, thus helping in bringing in more business and success in the future.

Thus, some of the most common challenges that companies come across are—

- Collection of more data than intended, leading to security and privacy violations.
- Lack of authenticity due to the reason that big data can come from a variety of sources, including online and offline activities.
- A lack of strong data confidentiality principles and use of compliant cloud access management services.
- A lack of Centralized key management systems.

2. Data Quality

High data quality means the data is accurate, consistent, comprehensive, and updated, but it's also context-sensitive. Before a firm develops its data quality management process, it's important to identify potential data issues, understanding these common problems will help businesses manage data quality more effectively within the organization. Different roles and applications require different quality criteria. Data

quality issues might occur due to various factors, like, human error, system entropy, or natural forces. Like any software or data application, big data analytics systems are also occasionally prone to failure. As data flows through production pipelines, there are numerous chances for its quality to be compromised. These issues can arise whenever data is missing, corrupted, or erroneous during the typical operation of a pipeline. Problems may occur during ingestion, within the data warehouse, during transformation, or at any stage in between. Data is considered low quality if it is stale, inaccurate, duplicated, incomplete, or if the model does not accurately reflect reality. Additionally, anomalies affecting transformed data at any point in the production process can also degrade the quality of the data in hand. For instance, a dataset of credit card transactions with many cancellations might be unsuitable for sales analysis but valuable for fraud detection.[17][18]

Some of the most common Data quality issues are:

- NULL values- The NULL value or missing data is a general data quality problem where a field is left blank because of intentional omission or an error in an API outage. Businesses can perform a NULL value test if they want to handle NULL values; this might be dbt's (dbt™ is a SQL-first transformation workflow) not_null test, whereby values are expected to be present within an indicated column after model execution. This will ensure that data integrity is made possible by bringing about proper decision-making.[17][18]
- Distribution errors - The distribution errors provide insights into whether the collected data reflects reality. Deviations from expected ranges indicate distribution errors. Inaccurate data and incorrect data representation, such as a misentered revenue figure or erroneous patient weight, can lead to significant data quality issues. Industries with stringent regulatory requirements, such as healthcare and finance, must be particularly vigilant about these errors. While outlier data points might indicate emerging trends, they are often signs of anomalies that could indicate broader issues. Monitoring data variety is crucial for maintaining data quality and anticipating potential problems.[17][18]
  Redundant data- Businesses today face severe challenges in managing data streaming from local databases, data lakes, and streaming platforms. This most often leads to numerous duplications of records that adversely affect customer experiences, misrepresentation of analytical insights, and skew machine learning models. These duplicate records can drive marketing inefficiencies, increase database costs, and put a company's reputation and bottom line at risk. Such issues are very common in data transfers and are usually due to loose aggregations or human errors. Testing for uniqueness should be implemented by businesses so that these problems are found long before the data hits production pipelines, using the capability afforded by dbt for such tests.[17][18]
  Outdated data- Data novelty management becomes one of the crucial tasks of a company dealing with batch or streaming data. When the data stops refreshing on time, it becomes stale and its value decreases for the users downstream, who are dependent on real-time accuracy. Ensuring timely refreshes in data means that ETL processes are effectively utilized; businesses can monitor them by employing SQL rules that track such indicators of data novelty. Another challenge is that if data turns stale, businesses might miss great opportunities; the insights might go wrong, and so will the decisions. Thus, data requires periodic monitoring and timely inspections to keep it updated and useful.[17][18]

3. Data Storage

—--Big data encompasses structured, semi-structured, and unstructured data (text, images, videos, logs, etc.). Storage of huge data comes as another challenge to businesses as all organizations might not have

the capacity to hold huge data in their systems. Moreover, traditional storage systems often lack the capacity and scalability to handle huge volumes of data, pushing the need for distributed storage systems like Hadoop Distributed File System (HDFS), Amazon S3, and other cloud storage solutions.[19]
Solutions such as NoSQL databases (e.g., MongoDB, Cassandra) and object storage are used to address this challenge. However, choosing and managing the right storage solution for each data type adds complexity to the storage strategy.
—----Data velocity also comes as a part and parcel of the challenges of storing data. The generated data needs to be stored at high speeds (real-time or near real-time). Handling high-velocity data requires storage systems capable of fast data ingestion and retrieval.
For example, streaming data from IoT devices or financial markets requires a storage infrastructure that supports low-latency operations, often used in-memory databases (e.g., Apache Kafka, Redis) or specialized storage systems.[20][22]
—-------At times companies might not also be in a position to afford the high cost of acquiring, maintaining, and scaling storage infrastructure. High-performance storage solutions (SSD, NVMe) are expensive, and cloud storage costs can accumulate with large data volumes.[22]
Efficient data management strategies, including data compression, tiered storage (moving less-used data to cheaper storage), and using cost-effective storage solutions (like Amazon Glacier for archival), become essential to manage costs.[21][22]
—------Data itself has a life span, so it is not incorrect to say that, not all data remains equally important over time. Proper data lifecycle management, including data archiving, deletion, and retention policies, is essential to keep storage systems optimized. At times companies might have to spend huge amounts to keep data up to date and relevant to the changing environment, as deciding which data to keep in high-performance storage and which to archive is crucial for both performance and cost management.[22]

4. Skilled Workforce Data Interpretation and Analysis. ...

Rapid growth in Big Data analytics has turned rather challenging, mainly because of the shortage of skilled professionals that such growth has precipitated. That means there is a high demand for data scientists, engineers, and analysts in the market, yet only a few are available, hence this gap in skills has affected organizations in reaping from Big Data. An ideal professional in Big Data would be one with a rare combination of skills in mathematics, statistics, advanced analytics, and solid business context knowledge. This multi-dimensional skill is pretty crucial in converting the insights gained from data into actionable business strategies.[23][24][25]
The dearth of skilled data scientists is worsened by the limited number of universities offering data science specializations and a lack of computer science education at earlier levels, such as IGCSE and A-levels. However, IT leaders can slightly mitigate this by using advanced analytics and AI tools like Google's AutoML from AWS, Google, and Microsoft, which reduce in-house development needs. Secondly, CIOs can outsource "bridgers" or "shapers," professionals who, besides their technical competencies, understand and communicate business needs and can act as an intermediary between IT and business units to drive efficiency.[23][24][25]
This in turn makes data analysis and decision-making by enterprises slower, as a shortage of skilled personnel may lead to slower problem-solving. To cope with the shortage, companies invest in training or try to recruit talent from international markets, although it is also noticeable that this can only slightly mitigate the problem.[23][24][25]

5. Cost and Infrastructure

Cost and infrastructure are significant challenges in big data analysis due to the complex nature and scale of big data environments.

—---Setting up the infrastructure for the collection of big data analysis at the initial stages involves a substantial initial investment. This includes purchasing high-capacity servers, storage solutions, network equipment, and other hardware components and distributed computing environments (like Hadoop clusters) and specialized storage systems, which can be expensive to procure and maintain for many organizations.

—----Once the infrastructure is in place it requires ongoing costs for maintenance, upgrades, and repairs. Regular hardware upgrades are necessary to keep up with the rapid growth of data, advances in technology, and the need for higher processing speeds. Infrastructure maintenance also includes monitoring system health, handling data backups, and ensuring data security, all of which require skilled personnel and additional software tools.

—---While cloud services like AWS, Google Cloud, and Microsoft Azure offer flexibility and scalability, they introduce their own set of costs. Cloud-based big data analysis incurs expenses for data storage, data transfer, compute instances, and other services. This can lead to quick escalation of costs with cloud services, especially when processing large datasets, utilizing high-performance computing resources, or transferring data in and out of the cloud. Managing cloud costs efficiently requires careful monitoring, optimization of resource usage, and potentially using features like auto-scaling and tiered storage options.

—--Big data analysis requires specialized software tools (e.g. Apache Spark, NoSQL databases, data visualization tools) that may have licensing fees. It also includes the requirement of a team of skilled professionals, including data engineers, data scientists, system administrators, and security experts. In today's day and time, these roles are in high demand, making it expensive for firms to hire and retain professionals with these skills. Additionally, continuous training is necessary to keep up with evolving technologies and best practices in big data management, adding to operational costs.[23]

6. Integration and Data Silos.

A data silo is a collection of data that's isolated from the rest of the organization and is only accessible to a specific group of people or departments.

Data silos are one big problem the organizations face, as these isolated data pools do take a lot of time, money, and effort to settle. It also discourages organizational learning since businesses cannot use their experiences in the best possible way. Silos may further result in inaccurate datasets and misunderstandings because fragmented pieces of information do not allow the data to be viewed as a whole. This fragmentation leads to less-than-ideal customer experiences, given that keeping track of and integrating customer interactions across teams becomes an overbearing process.

Furthermore, there are security and compliance risks with data silos, since they make the establishment of an effective data governance framework cumbersome and thus might lead to data breaches and cyber threats. They also propagate an organizational setup that creates disunity within teams in sharing a common idea. Overcoming these pitfalls means breaking down the barriers that data silos create to improve data accuracy, customer experiences, and overall organizational cohesion.[23]

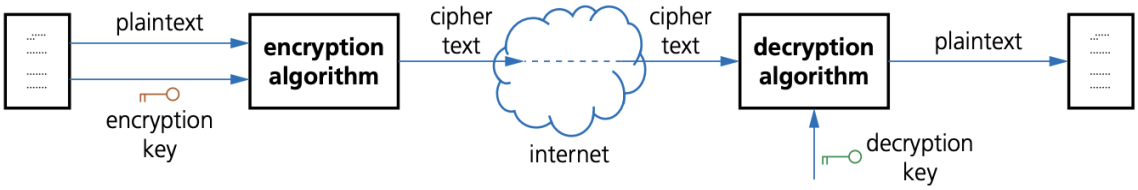
Common big data security practices

Implementing robust big data security measures is crucial for safeguarding sensitive information and mai-

ntaining data integrity. Here are eight best security practices businesses can implement to secure big data.

1. Encryption

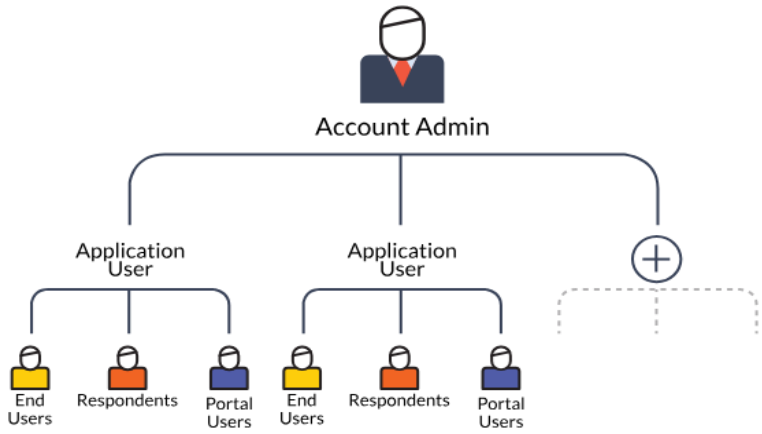
Encryption is the backbone of big data security whereby data is transformed into unreadable formats commonly known as cipher text, which can only be deciphered by users having the right key. This way, encryption guarantees confidentiality for sensitive data while being transmitted and during storage by protecting data in computers, servers, or within the network. In case unauthorized parties intercept the data, they cannot decode it. In addition to that, encryption preserves the integrity of data by making it hard to tamper with the data.[26]



[26]

2. Effective user access control

Big data contains sensitive and valuable information, and protecting it is critical for businesses. Effective user access control ensures that only authorized users can access, modify, or delete this data, reducing the risk of unauthorized access, data breaches, or theft. There are multiple ways to implement user access control for big data. One common approach is using role-based access control, as it allows authorities to create roles and assign access according to those roles. User accounts control access rights. This often involves levels of access. For example, in a hospital, it would not be appropriate for a cleaner to have access to data about any of the patients.[27]



[27]

3. Monitoring cloud security (Big Data Analytics in Cloud -Comparative Study)

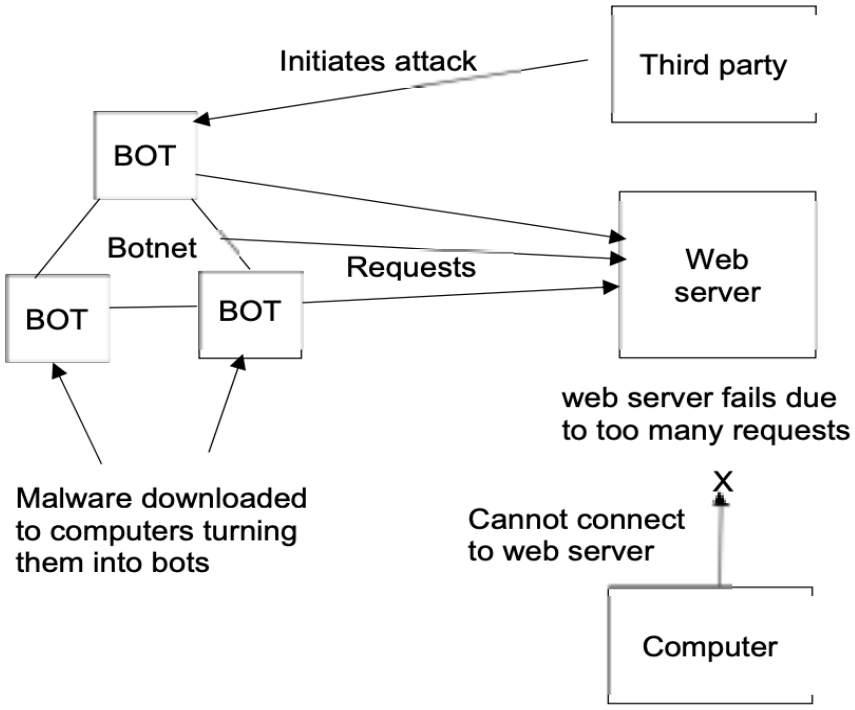
Cloud platforms make several business advantages available to big data analytics; because of scalability on demand, they enable scaling up their big data infrastructure to meet the growth of data volume. This flexibility is important to manage the variable workloads of big data analytics.

However, the cloud infrastructure exposed via API keys, tokens, and other misconfigurations opens it to the cyber-attacks. Therefore, threat detection becomes crucial to secure big data assets. Monitoring tools

for cloud security can detect unauthorized attempts to access data or exfiltration to help a business ensure cloud security.[28]

4. Network traffic analysis

Network traffic analysis can expose anomalies in network behavior, such as unusual data transmissions or unexpected surges in network traffic, which may indicate the occurrence of imminent security threats like data breaches or insider attacks. In addition, the patterns related to specific attack types may also involve malware attacks, phishing, DDoS, or even MitM attacks, thus allowing for better detection and mitigation of risks before severe damage may occur. It also helps in real-time monitoring of compliance to industry regulations and security standards. Organizations can identify suspicious activities by analyzing traffic patterns and behaviors to mitigate any potential cyberattacks and hence assure smooth and secure network performance.[26][28]



5. Employee training and awareness

Training employees in the organization is an important aspect of big data security as it allows them to understand threats to big data and how to prevent such threats. According to a Security Today report, in a joint study conducted by Stanford University Professor Jeff Hancock and cybersecurity firm Tessian, 88 percent of incidents of data breaches in any organization are because of employee errors.[29]

Proper training helps employees understand how to minimize this risk, such as learning the best practices of big data security: password creation, identification of phishing emails, and reporting suspicious activities. Moreover, with training employees are fully aware of the set regulations regarding data protection.[30]

6. Prompt incident response plan

A prompt incident response plan is essential for big data security. It provides a framework of guidelines

and procedures for organizations to swiftly and efficiently address cyberattacks, minimizing potential damage and expediting data recovery. Additionally, it plays a key role in validating and restoring data post-incidents, ensuring its accuracy and reliability. To create an immediate incident response plan, organizations should consider the following steps:[29]

- Identify the types of incidents that can occur
- Develop a specific response plan for each type of incident
- Assign roles and responsibilities
- Test the response plan regularly
- Eradicate the threat from affected systems
- Restore systems to normal operations securely

7. Real-time compliance and security monitoring

Real-time compliance and security monitoring remain very crucial in the context of big data security in this digital era wherein businesses are dealing with voluminous amounts of data. This allows organizations to detect suspicious activities and take appropriate action before actual damage is caused by enabling continuous monitoring of safety and adherence to compliance issues.[29]

Because big data very often deals with sensitive information concerning businesses and customers, they become liable under a number of regulations and compliance standards, such as GDPR, HIPAA, or PCI-DSS. Real-time monitoring ensures big data processing activities remain compliant with the regulations through alerts in case of any violation that helps an organization avoid associated penalties and reputational damage.[29]

8. Regular data backup

Security incidents like data breaches and malware attacks can lead to data loss that may be challenging to recover without proper backups. Even with proactive measures to prevent cyberattacks, it is essential to be prepared for unexpected events or breaches. Regular data backups are vital for ensuring data security, allowing organizations to restore lost or corrupted data while minimizing operational disruptions and financial losses. Having reliable backups also reinforces confidence, ensuring that critical data can be recovered promptly, helping to maintain trust and protect the organization's reputation.[29][30]

Conclusion:

In this era of information, massive and diverse data is generated at high velocities daily, holding within it patterns and insights of immense value. This research has examined the transformative potential of big data, which, through advanced analytic techniques, enables organizations to uncover hidden knowledge that can drive business change and enhance decision-making. By reviewing the literature, I have analyzed key concepts of big data analytics, including definitions, characteristics, sources, challenges, and emerging technologies. The Internet of Things, cloud computing, and data centers were also discussed as closely related technologies that propel big data's development and progress.

Despite significant advancements, challenges persist in areas such as standardization, real-time performance, storage, management, and security. These are critical to unlocking the full potential of big data and require innovative solutions from researchers and developers. Moreover, big data analytics offers broad applications across various fields, such as customer intelligence, fraud detection, supply chain management, and healthcare, thereby benefiting sectors like retail, telecom, and manufacturing.

My research has highlighted essential big data tools, methods, and techniques, providing a roadmap for both users and developers interested in implementing big data solutions. However, managing big data presents complexities, particularly with issues of data volume, velocity, and variety, which demand robust storage, integration, processing, and analytical frameworks. Future research could focus on developing comprehensive frameworks that address these challenges, further enhancing big data analytics' role in decision-making. When applied correctly, big data holds the promise of substantial scientific, technological, and humanitarian advancements, offering unforeseen insights and benefits to decision-makers and contributing to a brighter future across multiple domains.

References:

Hashemi-Pour, C., Botelho, B., & Bigelow, S. J. (n.d.). (March, 2024) Big data: Definition. TechTarget. Retrieved from https://techtarget.com/searchdatamanagement/definition/big-data

Sharma, R. (2024, March 4). Sources of big data: Where does it come from? upGrad. Retrieved from https://www.upgrad.com

Akhaleqh02. (2024, July 30). Different sources of data for data analysis. GeeksforGeeks. Retrieved from https://www.geeksforgeeks.org/different-sources-of-data-for-data-analysis/

Vedantu. (n.d.). Sources of secondary data. 2024 from https://www.vedantu.com/commerce/sources-of-secondary-data

Sheedy, J. (2023, May 30). The role of big data in business intelligence: Trends and opportunities. Retrieved https://www.jasonsheedy.com

Adaga, E. M., Okorie, G. N., Egieya, Z. E., & Ikwue, U. (2024). The role of big data in business strategy: A critical review. Computer Science & IT Research Journal, 4(3), 327–350. Retrieved from https://www.researchgate.net/publication/377071984_THE_ROLE_OF_BIG_DATA_IN_BUSINESS_STRATEGY_A_CRITICAL_REVIEW

Qlik. (n.d.). Data integration. Qlik. Retrieved from https://www.qlik.com/us/data-integration

AltexSoft. (2021, September 10). Data integration: Approaches, techniques, tools, and best practices for implementation. Retrieved from https://www.altexsoft.com/blog/data-integration/

Ashraf, S. (2020, July 30). Data integration approaches – Which one is right for business? Data Integration Info. Retrieved from https://dataintegrationinfo.com/data-integration-approaches/

Amazon Web Services, Inc. (n.d.). Analytics and data lake use cases. AWS. Retrieved from https://aws.amazon.com/free/analytics/

GeeksforGeeks. (2022, December 9). Data virtualization. GeeksforGeeks. Retrieved from https://www.geeksforgeeks.org/data-virtualization/

GeeksforGeeks. (2024, July 5). Real-time analytics in big data. GeeksforGeeks, from https://www.geeksforgeeks.org/real-time-analytics-in-big-data/

BIGWORK. (n.d.). Real-time analytics BIGWORK. Retrieved from https://bigworkthailand.com/blog/real-time-analytics

Adobe Communications Team. (2022, November 8). Types of analytics explained — descriptive, predictive, prescriptive, and more. Adobe. Retrieved from https://business.adobe.com/blog/basics/descriptive-predictive-prescriptive-analytics-explained

Qlik. (n.d.). What is prescriptive analytics? Qlik. Retrieved from https://www.qlik.com/us/augmented-analytics/prescriptive-analytics

Segal, T. (2022, December 31). Prescriptive analytics. Investopedia. Retrieved from

https://www.investopedia.com/terms/p/prescriptive-analytics.asp
Osborn, T. (2024, May 8). 8 data quality issues and how to solve them. Monte Carlo Data. Retrieved from https://www.montecarlodata.com/blog-8-data-quality-issues
Novogroder, I. (2023, December 21). 14 most common data quality issues and how to fix them. Monte Carlo Data. Retrieved from https://www.montecarlodata.com/blog/data-quality-issues
Geeks for Geeks. (2022, June 17). Introduction to Hadoop Distributed File System (HDFS). Geeks for Geeks. Retrieved from https://www.geeksforgeeks.org/introduction-to-hadoop-distributed-file-systemhdfs/
Apache Software Foundation. (n.d.). Apache Kafka. Retrieved from https://kafka.apache.org/
Amazon Web Services. (n.d.). Amazon S3. Retrieved from https://aws.amazon.com/s3/
Geeks for Geeks. (2024, May 10). What is data lifecycle management? Geeks for Geeks. Retrieved from https://www.geeksforgeeks.org/what-is-data-lifecycle-management/
Kumar, N., Hema, K., Hordiichuk, V., & Menon, R. (2023). Harnessing the power of big data: Challenges and opportunities in analytics. Retrieved from https://www.researchgate.net/publication/373770538_Harnessing_the_Power_of_Big_Data_Challenges_and_Opportunities_in_Analytics
Express Computer. (n.d.). Big data talent shortage: How to bridge the gap? Fractal. Retrieved from https://fractal.ai/news/big-data-talent-shortage-bridge-gap/
Rae, S. (2018, June 6). Big data skills shortages – and how to work around them. ComputerWeekly. Retrieved from https://www.computerweekly.com/opinion/Big-data-skills-shortages-and-how-to-work-around-them
Watson, D., & Williams, H. (n.d.). Cambridge International AS & A Level Computer Science Boost eBook. Hodder Education. Retrieved from https://www.hoddereducation.com/cambridge-international-as-a-level-computer-science
Key Survey. (n.d.). Access levels & permissions. Retrieved from https://www.keysurvey.com/survey-software/access-levels-permissions/
Miryala, N. K., Meta, & Gupta, D. (2023). Big data analytics in cloud - Comparative study. Retrieved from https://www.researchgate.net/publication/376892490_Big_Data_Analytics_in_Cloud_-Comparative_Study
Dutta, S. (n.d.). Big data security: Advantages, challenges, and best practices. Turing. Retrieved from https://www.turing.com/resources/big-data-security
Ackerman, R. (2023, August 2). Just why are so many cyber breaches due to human error? Security Today. Retrieved from https://securitytoday.com/articles/2022/07/30/just-why-are-so-many-cyber-breaches-due-to-human-error.aspx