Introduction to Big Data

1. Introduction

We live in a digital world where data is increasing rapidly because of the increasing use of the internet, sensors and heavy machines at a very high rate. The sheer volume, variety, velocity and veracity of such data is called **big data**. Big Data is structured, unstructured, semi-structured or heterogeneous in nature. Data is everywhere in the form of numbers, text, images and video. As the data continues to grow, there comes the necessity to organize it. The need to sort, organize and analyze this Big Data is called **Big Data Analytics**. The process of capturing or collecting Big Data is known as **datafication**. Big Data is datafied so that it can be used productively.

2. Features of Big Data:

The basic features of Big Data include

- It is a new data challenge that requires leveraging the existing system differently.
- It is often unstructured and qualitative in nature
- It is denoted by 4 V's i.e. Volume, Variety, Velocity and Veracity

3. Classification of Big Data:

a. Structured data

Structured Data is used to refer to the data which is already stored in databases, in an ordered manner. It accounts for about 20% of the total existing data, and is used the most in programming and computer-related activities.

There are two sources of structured data- machines and humans. All the data received from sensors, web logs and financial systems are classified under **machine-generated data.** These include medical devices, GPS data, data of usage statistics captured by servers and applications and the huge amount of data that usually move through trading platforms, to name a few.

Human-generated structured data mainly includes all the data a human input into a computer, such as his name and other personal details. When a person clicks a link on the internet, or even makes a move in a game, data is created- this can be used by companies to figure out their customer behaviour and make the appropriate decisions and modifications.

b. Unstructured data

While structured data resides in the traditional row-column databases, unstructured data is the opposite- they have no clear format in storage. The rest of the data created, about 80% of the total account for unstructured big data. Most of the data a person encounters belongs to this category- and until recently, there was not much to do to it except storing it or analysing it manually.

The Unstructured data is further divided into – Captured and User Generated data.

Captured data is passively based on user's behavior. For instance, if someone types something on the search bar through Google, it is captured at the moment to have a basic research on what's on trend and case studies in future. Another example could be the GPS via smartphone that captures each moment someone searches for something and gets a real-time output.
User-generated data is that kind of unstructured data which is put on internet each and every moment by the users themselves. For instance, the Likes, Shares, Tweets, Re-tweets, Comments, on Facebook posts/photos/videos, YouTube, Twitter, etc. are all user-generated.

c. Semi-structured data.

The line between unstructured data and semi-structured data has always been unclear, since most of the semi-structured data appear to be unstructured at a glance. Information that is not in the traditional database format as structured data, but contain some organizational properties which make it easier to process, are included in semi-structured data. For example, NoSQL documents are considered to be semi-structured, since they contain keywords that can be used to process the document easily.

4. Introduction to analytics:

Analytics is a broad term that encompasses the processes, technologies, frameworks and algorithms to extract meaningful insights from data. Raw data in itself does not have a meaning until it is contextualized and processed into useful information. Analytics is the process of extracting and creating information from raw data by filtering, processing, categorizing, condensing and contextualizing the data. This information obtained is then organized and structured to infer knowledge about the system and/or its users, its environment, and its operations and progress towards its objectives, thus making the systems smarter and more efficient.

The choice of the technologies, algorithms, and frameworks for analytics is driven by the analytics goals of the application. The goals of the analytics task may be:

a) to predict something (for example whether a transaction is a fraud or not, whether it will rain on a particular day, or whether a tumor is benign or malignant)

b) to find patterns in the data (for example, finding the top 10 days with heavy rainfall in the year, finding which pages are visited the most on a particular website, or finding the most searched celebrity in a particular year)

c) finding relationships in the data (for example, finding similar news articles, finding similar patients in an electronic health record system, finding related products on an ecommerce website, finding similar images, or finding correlation between news items and stock prices).

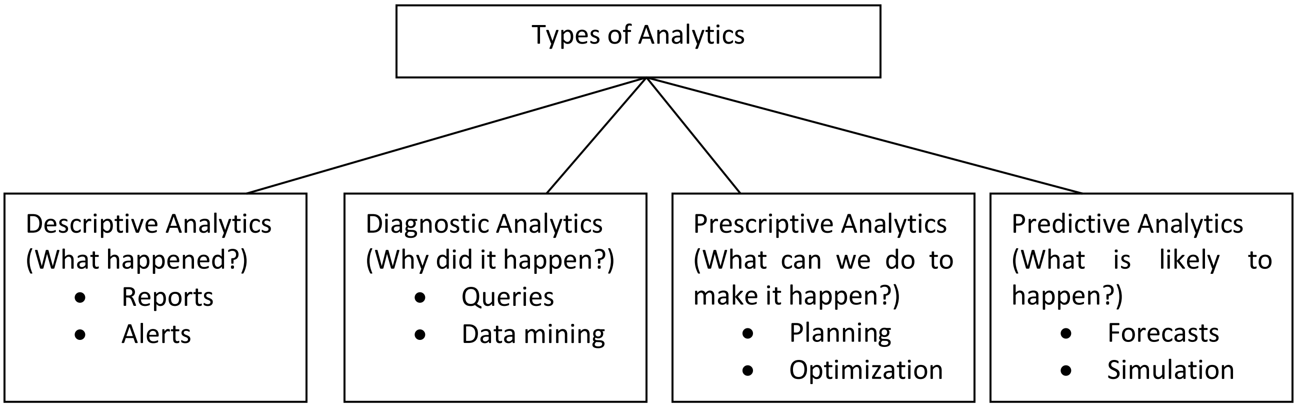
5. Types of Analytics:

Big data analytics reformed the ways of conducting business such as it improves decision making, business process management. Hence it is also known as business analytics. Business analytics makes use of data along with different other techniques like information technology, statistics, quantitative methods and other models to provide the results. There are different types of business analytics. They are:

a. Descriptive Analytics
b. Diagnostic Analytics
c. Predictive Analytics
d. Prescriptive Analytics

Descriptive Analytics

Descriptive analytics comprises analyzing past data to present it in a summarized form which can be easily interpreted. Descriptive analytics aims to answer – "What has happened?" A major portion of analytics done today is descriptive analytics through use of statistics functions such as counts, maximum, minimum, mean, top-N, percentage, for instance. These statistics help in describing patterns in the data and present the data in a summarized form. Examples for this type include – computing the total number of likes for a particular post, computing the average monthly rainfall or finding the average number of visitors per month on a website. Descriptive analytics is useful to summarize the data. It analyses the past or current business events and helps managers to develop a road map for future actions. It performs an in depth analysis of data to reveal details such as frequency of events, operation costs, and underlying reason for failure. It helps to identify the root cause of the problem.



Diagnostic Analytics

Diagnostic analytics comprises analysis of past data to diagnose the reasons as to why certain events happened. Diagnostic analytics aims to answer - Why did it happen? Let us consider an example of a system that collects and analyzes sensor data from machines for monitoring their health and predicting failures. While descriptive analytics can be useful for summarizing the data by computing various statistics (such as mean, minimum, maximum, variance, or top-N), diagnostic analytics can provide more insights into why certain a fault has occurred based on the patterns in the sensor data for previous faults.

Predictive Analytics

Predictive analytics comprises predicting the occurrence of an event or the likely outcome of an event or forecasting the future values using prediction models. Predictive analytics aims to answer - What is likely to happen? For example, predictive analytics can be used for predicting when a fault will occur in a machine, predicting whether a tumor is benign or malignant, predicting the occurrence of natural emergency (events such as forest fires or river floods) or forecasting the pollution levels. Predictive Analytics is done using predictive models which are trained by existing data. These models learn patterns and trends from the existing data and predict the occurrence of an event or the likely outcome of an event (classification models) or forecast numbers (regression models). The accuracy of prediction models depends on the quality and volume of the existing data available for training the models, such that all the patterns and trends in the existing data can be learned accurately. Before a model is used for prediction, it must be validated with existing data. The typical approach adopted while developing prediction models is to divide the existing data into training and test data sets.

Prescriptive Analytics

While predictive analytics uses prediction models to predict the likely outcome of an event, prescriptive analytics uses multiple prediction models to predict various outcomes and the best course of action for each outcome. Prescriptive analytics aims to answer - What can we do to make it happen? Prescriptive Analytics can predict the possible outcomes based on the current choice of actions. We can consider prescriptive analytics as a type of analytics that uses different prediction models for different inputs. Prescriptive analytics prescribes actions or the best option to follow from the available options. For example, prescriptive analytics can be used to prescribe the best medicine for treatment of a patient based on the outcomes of various medicines for similar patients. Another example of prescriptive analytics would be to suggest the best mobile data plan for a customer based on the customer’s browsing patterns.

6. Overview of Big Data:

Big data is defined as collections of datasets whose volume, velocity or variety is so large that it is difficult to store, manage, process and analyze the data using traditional databases and data processing tools. In the recent years, there has been an exponential growth in the both structured and unstructured data generated by information technology, industrial, healthcare, Internet of Things, and other systems.

According to an estimate by IBM, 2.5 quintillion bytes of data is created every day. A recent report by DOMO estimates the amount of data generated every minute on popular online platforms. Given below are some facts of usage of data by users in different domains:

- Facebook users share nearly 4.16 million pieces of content
- Twitter users send nearly 300,000 tweets
- Instagram users like nearly 1.73 million photos
- YouTube users upload 300 hours of new video content
- Apple users download nearly 51,000 apps
- Skype users make nearly 110,000 new calls
- Amazon receives 4300 new visitors
- Uber passengers take 694 rides
- Netflix subscribers stream nearly 77,000 hours of video

Big Data has the potential to power next generation of smart applications that will leverage the power of the data to make the applications intelligent. Applications of big data span a wide range of domains such as web, retail and marketing, banking and financial, industrial, healthcare, environmental, Internet of Things and cyber-physical systems.

Big Data analytics deals with collection, storage, processing and analysis of this massive-scale data. Specialized tools and frameworks are required for big data analysis when:

a. the volume of data involved is so large that it is difficult to store, process and analyze data on a single machine
b. the velocity of data is very high and the data needs to be analyzed in real-time
c. there is variety of data involved, which can be structured, unstructured or semi-structured, and is collected from multiple data sources
d. various types of analytics need to be performed to extract value from the data such as descriptive, diagnostic, predictive and prescriptive analytics.

Big data analytics involves several steps starting from data cleansing, data munging (or wrangling), data processing and visualization. Big data analytics life-cycle starts from the collection of data from multiple data sources. Specialized tools and frameworks are required to ingest the data from different sources into the dig data analytics backend. The data is stored in specialized storage solutions (such as distributed filesystems and non-relational databases) which are designed to scale. Based on the analysis requirements (batch or real-time), and type of analysis to be performed (descriptive, diagnostic, predictive, or predictive) specialized frameworks are used. Big data analytics is enabled by several technologies such as cloud computing, distributed and parallel processing frameworks, non-relational databases, in-memory computing.

Some examples of big data are listed as follows:

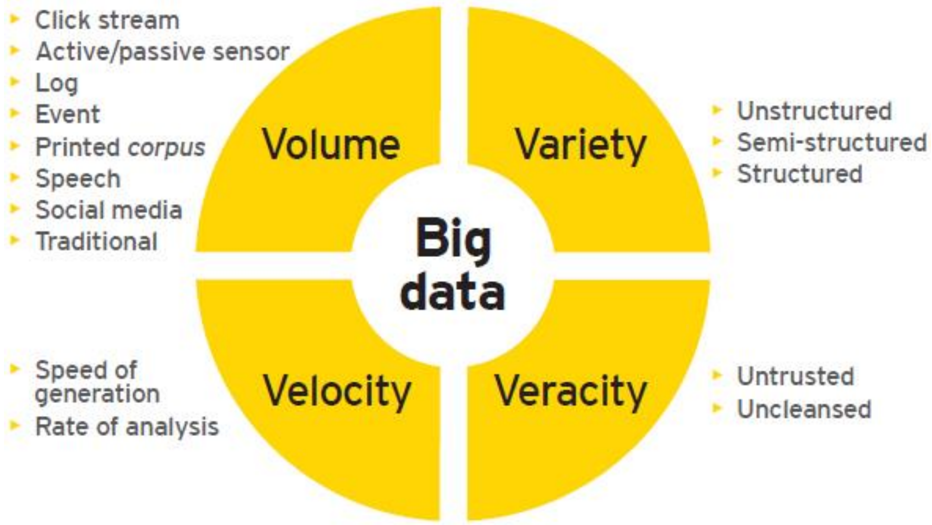
- Data generated by social networks including text, images, audio and video data
- Click-stream data generated by web applications such as e-Commerce to analyze user
- behavior
- Machine sensor data collected from sensors embedded in industrial and energy systems
- for monitoring their health and detecting failures
- Healthcare data collected in electronic health record (EHR) systems
- Logs generated by web applications
- Stock markets data
- Transactional data generated by banking and financial applications

7. Characteristics of Big Data:

According to Gartner, the growth of Big data can be depicted in terms of 4V's

- Volume
- Velocity
- Variety
- Veracity

Volume: It is the amount of data generated by the organization or individuals. The volume of data in most organizations is approaching exabytes. The main characteristic that makes data “big” is the sheer volume. It makes no sense to focus on minimum storage units because the total amount of information is growing exponentially every year. Organizations are doing their best to handle this increasing volume of data. For example: Every minute over 571 websites are being created. The exact size of the Internet will never be known.



Velocity: Velocity describes the rate at which data is generated, captured and shared. Enterprises can capitalize on data only if it is capture and shares on real time. Information processing system lie CRM and ERP face problems associated with data, which keeps adding up but cannot be processed quickly. These systems are able to handle data in batches every few hours. Even this time lag causes the data to lose its importance as new data is constantly being generated. Velocity is the frequency of incoming data that needs to be processed. The following are the sources of high velocity data – Face book status updates, or credit card swipes are being sent on a particular telecom carrier every minute of every day, and the other activities on social media. A streaming application like Amazon Web Services is an example of an application that handles the velocity of data.

Variety: Data is generated at a fast pace. The data is generated from different sources such as internal, external, social or behavioral. Data is available in different formats such as video, image, text, audio etc. Even a single source can generate varied format.

Variety is one the most interesting developments in technology as more and more information is digitized. Traditional data types (structured data) include things on a bank statement like date, amount, and time. Structured data is augmented by unstructured data, which is where things like Twitter feeds, audio files, MRI images, web pages, web logs are put — anything that can be captured and stored.

With unstructured data, on the other hand, there are no rules. A picture, a voice recording, a tweet — they all can be different but express ideas and thoughts based on human understanding. One of the goals of big data is to use technology to take this unstructured data and make sense of it.

Veracity: Veracity refers to uncertainty of data. i.e. whether the data obtained is correct or consistent. Veracity refers to the trustworthiness of the data. Out of the huge amount of data that is generated, only correct data can be used for further analysis. Data when processed becomes information and a lot of effort goes into processing it. Big Data especially unstructured and semi-structured forms are messy in nature. It takes a lot of time and expertise to clean that data and make it suitable for analysis.

Value: Value of data refers to the usefulness of data for the intended purpose. The end goal of any big data analytics system is to extract value from the data. The value of the data is also related to the veracity or accuracy of the data. For some applications value also depends on how fast we are able to process the data.

8. Domain Specific Examples of Big Data

The applications of big data span a wide range of domains including (but not limited to) homes, cities, environment, energy systems, retail, logistics, industry, agriculture, Internet of Things, and healthcare. This section provides an overview of various applications of big data for each of these domains.

8.1 Web

Web Analytics: Web analytics deals with collection and analysis of data on the user visits on websites and cloud applications. Analysis of this data can give insights about the user engagement and tracking the performance of online advertisement campaigns. For collecting data on user visits, two approaches are used. In the first approach, user visits are logged on the web server which collects data such as the date and time of visit, resource requested, user's IP address, HTTP status code, for instance. The second approach, called page tagging, uses a JavaScript which is embedded in the web page. Whenever a user visits a web page, the JavaScript collects user data and sends it to a third party data collection server. A cookie is assigned to the user which identities the user during the visit and the subsequent visits. The benefit of the page tagging approach is that it facilitates real-time data collection and analysis. This approach allows third party services, which do not have access to the web server (serving the website) to collect and process the data. These specialized analytics service providers (such as Google Analytics) are offer advanced analytics and summarized reports. The key reporting metrics include user sessions, page visits, top entry and exit pages, bounce rate, most visited page, time spent on each page, number of unique visitors, number of repeat visitors, for instance.

Performance Monitoring: Multi-tier web and cloud applications such as such as e-Commerce, Business-to-Business, Health care, Banking and Financial, Retail and Social Networking applications, can experience rapid changes in their workloads. To ensure market readiness of such applications, adequate resources need to be provisioned so that the applications can meet the demands of specified workload levels and at the same time ensure that the service level agreements are met.

Ad Targeting & Analytics: Search and display advertisements are the two most widely used approaches for Internet advertising. In search advertising, users are displayed advertisements ("ads"), along with the search results, as they search for specific keywords on a search engine. Advertisers can create ads using the advertising networks provided by the search engines or social media networks. These ads are setup for specific keywords which are related to the product or service being advertised. Users searching for these keywords are shown ads along with the search results. Display advertising, is another form of Internet advertising, in which the ads are displayed within websites, videos and mobile applications who participate in the advertising network. Display ads can either be text-based or image ads. The ad-network matches these ads against the content on the website, video or mobile application and places the ads. The most commonly used compensation method for Internet ads is Pay-per-click (PPC), in which the advertisers pay each time a user clicks on an advertisement. Advertising networks use big data systems for matching and placing advertisements and generating advertisement statistics reports. Advertisers can use big data tools for tracking the performance of advertisements, optimizing the bids for pay-per-click advertising, tracking which keywords link the most to the advertising landing pages and optimizing budget allocation to various advertisement campaigns.

Content Recommendation: Content delivery applications that serve content (such as music and video streaming applications), collect various types of data such as user search patterns and browsing history, history of content consumed, and user ratings. Such applications can leverage big data systems for recommending new content to the users based on the user preferences and interests. Recommendation systems use two broad category approaches - user-based recommendation and item based recommendation. In user-based recommendation, new items are recommended to a user based on how similar users rate those items. Whereas in item-based recommendation, new items are recommended to a user based on how the user rated similar items.

8.2 Financial

Credit Risk Modeling: Banking and Financial institutions use credit risk modeling to score credit applications and predict if a borrower will default or not in the future. Credit risk models are created from the customer data that includes, credit scores obtained from credit bureaus, credit history, account balance data, account transactions data and spending patterns of the customer. Credit models generate numerical scores that summarize the creditworthiness of customers. Since the volume of customer data obtained from multiple sources can be massive, big data systems can be used for building credit models. Big data systems can help in computing credit risk scores of a large number of customers on a regular basis.

Fraud Detection: Banking and Financial institutions can leverage big data systems for detecting frauds such as credit card frauds, money laundering and insurance claim frauds. Real-time big data analytics frameworks can help in analyzing data from disparate sources and label transactions in real-time. Machine learning models can be built for detecting anomalies in transactions and detecting fraudulent activities. Batch analytics frameworks can be used for analyzing historical data on customer transactions to search for patterns that indicate fraud.

8.3 Healthcare

The healthcare ecosystem consists of numerous entities including healthcare providers (primary care physicians, specialists, or hospitals), payers (government, private health insurance companies, employers), pharmaceutical, device and medical service companies, IT solutions and services firms, and patients. The process of provisioning healthcare involves massive healthcare data that exists in different forms (structured or unstructured), is stored in disparate data sources (such as relational databases, or file servers) and in many different formats. To promote more coordination of care across the multiple providers involved with patients, their clinical information is increasingly aggregated from diverse sources into Electronic Health Record (EHR) systems. EHRs capture and store information on patient health and provider actions including individual-level laboratory results, diagnostic, treatment, and demographic data.

Some healthcare applications that can benefit from big data systems:

Epidemiological Surveillance: Epidemiological Surveillance systems study the distribution and determinants of health-related states or events in specified populations and apply these studies for diagnosis of diseases under surveillance at national level to control health problems. EHR systems include individual-level laboratory results, diagnostic, treatment, and demographic data. Big data frameworks can be used for integrating data from multiple EHR systems and timely analysis of data for effectively and accurately predicting outbreaks, population-level health surveillance efforts, disease detection and public health mapping.

Patient Similarity-based Decision Intelligence Application: Big data frameworks can be used for analyzing EHR data to extract a cluster of patient records most similar to a particular target patient. Clustering patient records can also help in developing medical prognosis applications that predicts the likely outcome of an illness for a patient based on the outcomes for similar patients.

Adverse Drug Events Prediction: Big data frameworks can be used for analyzing EHR data and predict which patients are most at risk for having an adverse response to a certain drug based on adverse drug reactions of other patients.

Detecting Claim Anomalies: Heath insurance companies can leverage big data systems for analyzing health insurance claims to detect fraud, abuse, waste, and errors.

Evidence-based Medicine: Big data systems can combine and analyze data from a variety of sources, including individual-level laboratory results, diagnostic, treatment and demographic data, to match treatments with outcomes, predict patients at risk for a disease. Systems for evidence-based medicine enable providers to make decisions not only based on their own perceptions but also from the available evidence.

Real-time health monitoring: Wearable electronic devices allow non-invasive and continuous monitoring of physiological parameters. These wearable devices may be in various forms such as belts and wrist-bands. Healthcare providers can analyze the collected healthcare data to determine any health conditions or anomalies. Big data systems for real-time data analysis can be used for analysis of large volumes of fast-moving data from wearable devices and other in-hospital or in-home devices, for real-time patient health monitoring and adverse event prediction.

8.4 Internet of Things

Internet of Things (IoT) refers to things that have unique identities and are connected to the Internet. The "Things" in IoT are the devices which can perform remote sensing, actuating and monitoring. IoT devices can exchange data with other connected devices and applications (directly or indirectly), or collect data from other devices and process the data either locally or send the data to centralized servers or cloud-based application back-ends for processing the data, or perform some tasks locally and other tasks within the IoT infrastructure, based on temporal and space constraints.

Some IoT applications that can benefit from big data systems:

Intrusion Detection: Intrusion detection systems use security cameras and sensors (such as PIR sensors and door sensors) to detect intrusions and raise alerts. Alerts can be in the form of an SMS or an email sent to the user. Advanced systems can even send detailed alerts such as an image grab or a short video clip sent as an email attachment.

Smart Parking: Smart parkings make the search for parking space easier and convenient for drivers. Smart parkings are powered by IoT systems that detect the number of empty parking slots and send the information over the Internet to smart parking application back-ends. These applications can be accessed by the drivers from smart-phones, tablets and in-car navigation systems. In a smart parking, sensors are used for each parking slot, to detect whether the slot is empty or occupied. This information is aggregated by an on-site smart parking controller and then sent over the Internet to cloud-based big data analytics backend.

Smart Roads: Smart roads equipped with sensors can provide information on driving conditions, travel time estimates and alerts in case of poor driving conditions, traffic congestions and accidents. Such information can help in making the roads safer and help in reducing traffic jams. Information sensed from the roads can be communicated via Internet to cloud-based big data analytics applications. The analysis results can be disseminated to the drivers who subscribe to such applications or through social media.

Structural Health Monitoring: Structural Health Monitoring systems use a network of sensors to monitor the vibration levels in the structures such as bridges and buildings. The data collected from these sensors is analyzed to assess the health of the structures. By analyzing the data it is possible to detect cracks and mechanical breakdowns, locate the damages to a structure and also calculate the remaining life of the structure. Using such systems, advance warnings can be given in the case of imminent failures of the structures.

Smart Irrigation: Smart irrigation systems can improve crop yields while saving water. Smart irrigation systems use IoT devices with soil moisture sensors to determine the amount of moisture in the soil and release the flow of water through the irrigation pipes only when the moisture levels go below a predefined threshold. Smart irrigation systems also collect moisture level measurements in the cloud where the big data systems can be used to analyze the data to plan watering schedules.

8.5 Environment

Environment monitoring systems generate high velocity and high volume data. Accurate and timely analysis of such data can help in understanding the current status of the environment and also predicting environmental trends. Let us look at some environment monitoring applications that can benefit from big data systems:

Weather Monitoring: Weather monitoring systems can collect data from a number of sensor attached (such as temperature, humidity, or pressure) and send the data to cloud-based applications and big data analytics backend. This data can then be analyzed and visualized for monitoring weather and generating weather alerts.

Air Pollution Monitoring: Air pollution monitoring systems can monitor emission of harmful gases (CO2, CO, NO, or NO2) by factories and automobiles using gaseous and meteorological sensors. The collected data can be analyzed to make informed decisions on pollution control approaches.

Noise Pollution Monitoring: Due to growing urban development, noise levels in cities have increased and even become alarmingly high in some cities. Noise pollution can cause health hazards for humans due to sleep disruption and stress. Noise pollution monitoring can help in generating noise maps for cities. Urban noise maps can help the policy makers in urban planning and making policies to control noise levels near residential areas, schools and parks. Noise pollution monitoring systems use a number of noise monitoring stations that are deployed at different places in a city. The data on noise levels from the stations is sent to cloud-based applications and big data analytics backend. The collected data is then aggregated to generate noise maps.

Forest Fire Detection: Forest fires can cause damage to natural resources, property and human life. There can be different causes of forest fires including lightening, human negligence, volcanic eruptions and sparks from rock falls. Early detection of forest fires can help in minimizing the damage. Forest fire detection systems use a number of monitoring nodes deployed at different locations in a forest. Each monitoring node collects measurements on ambient conditions including temperature, humidity, light levels, for instance.

River Floods Detection: River floods can cause extensive damage to the natural and human resources and human life. River floods occur due to continuous rainfall which causes the river levels to rise and flow rates to increase rapidly. Early warnings of floods can be given by monitoring the water level and flow rate. River flood monitoring system use a number of sensor nodes that monitor the water level (using ultrasonic sensors) and flow rate (using the flow velocity sensors). Big data systems can be used to collect and analyze data from a number of such sensor nodes and raise alerts when a rapid increase in water level and flow rate is detected.

Water Quality Monitoring: Water quality monitoring can be helpful for identifying and controlling water pollution and contamination due to urbanization and industrialization. Maintaining good water quality is important to maintain good health of plant and animal life. Water quality monitoring systems use sensors to autonomously and continuously monitor different types contaminations in water bodies (such as chemical, biological, and radioactive). The scale of data generated by such systems is massive. Big data systems can help in real-time analysis of data generated by such systems and generate alerts about any degradation in water quality, so that corrective actions can be taken.

8.6 **Logistics & Transportation**

Real-time Fleet Tracking: Vehicle fleet tracking systems use GPS technology to track the locations of the vehicles in real-time. Cloud-based fleet tracking systems can be scaled up on demand to handle large number of vehicles. Alerts can be generated in case of deviations in planned routes. Big data systems can be used to aggregate and analyze vehicle locations and routes data for detecting bottlenecks in the supply chain such as traffic congestions on routes, assignment and generation of alternative routes, and supply chain optimization.

Shipment Monitoring: Shipment management solutions for transportation systems allow monitoring the conditions inside containers. For example, containers carrying fresh food produce can be monitored to detect spoilage of food. Shipment monitoring systems use sensors such as temperature, pressure, humidity, for instance, to monitor the conditions inside the containers and send the data to the cloud, where it can be analyzed to detect food spoilage. The analysis and interpretation of data on the environmental conditions in the container and food truck positioning can enable more effective routing decisions in real time. Therefore, it is possible to take remedial measures such as - the food that has a limited time budget before it gets rotten can be re-routed to a closer destinations, alerts can be raised to the driver and the distributor about the transit conditions, such as container temperature exceeding the allowed limit, humidity levels going out of the allowed limit, for instance, and corrective actions can be taken before the food gets damaged.

Route Generation & Scheduling: Modern transportation systems are driven by data collected from multiple sources which is processed to provide new services to the stakeholders. By collecting large amount of data from various sources and processing the data into useful information, data-driven transportation systems can provide new services such as advanced route guidance, dynamic vehicle routing, anticipating customer demands for pickup and delivery problem, for instance. Route generation and scheduling systems can generate end-to-end routes using combination of route patterns and transportation modes and feasible schedules based on the availability of vehicles. As the transportation network grows in size and complexity, the number of possible route combinations increases exponentially. Big data systems can provide fast response to the route generation queries and can be scaled up to serve a large transportation network.

Hyper-local Delivery: Hyper-local delivery platforms are being increasingly used by businesses such as restaurants and grocery stores to expand their reach. These platforms allow customers to order products (such as grocery and food items) using web and mobile applications and the products are sourced from local stores (or restaurants). As these platforms scale up to serve a large number of customer (with thousands of transactions every hour), they face various challenges in processing the orders in real-time. Big data systems for real-time analytics can be used by hyper-local delivery platforms for determining the nearest store from where to source the order and finding a delivery agent near to the store who can pickup the order and deliver to the customer.

Cab/Taxi Aggregators: On-demand transport technology aggregators (or cab/taxi aggregators) allow customers to book cabs using web or mobile applications and the requests are routed to nearest available cabs (sometimes even private drivers who opt-in their own cars for hire). The cab aggregation platforms use big data systems for real-time processing of requests and dynamic pricing. These platforms maintain record of all cabs and match the trip requests from customers to the nearest and most suitable cabs. These platforms adopt dynamic pricing models where the pricing increases or decreases based on the demand and the traffic conditions.

8.7 Industry

Machine Diagnosis & Prognosis: Machine prognosis refers to predicting the performance of a machine by analyzing the data on the current operating conditions and the deviations from the normal operating conditions. Machine diagnosis refers to determining the cause of a machine fault. Industrial machines have a large number of components that must function correctly for the machine to perform its operations. Sensors in machines can monitor the operating conditions such as (temperature and vibration levels). The sensor data measurements are done on timescales of few milliseconds to few seconds, which leads to generation of massive amount of data. Machine diagnostic systems can be integrated with cloud-based storage and big data analytics backend for storage, collection and analysis of such massive scale machine sensor data. A number of methods have been proposed for reliability analysis and fault prediction in machines. Case-based reasoning (CBR) is a commonly used method that find solutions to new problems based on past experience. This past experience is organized and represented as cases in a case-base. CBR is an effective technique for problem solving in the fields in which it is hard to establish a quantitative mathematical model, such as machine diagnosis and prognosis. Since for each machine, data from a very large number of sensors is collected, using such high dimensional data for creation of a case library reduces the case retrieval efficiency. Therefore, data reduction and feature extraction methods are used to find the representative set of features which have the same classification ability as the complete set of features.

Risk Analysis of Industrial Operations: In many industries, there are strict requirements on the environment conditions and equipment working conditions. Monitoring the working conditions of workers is important for ensuring their health and safety. Harmful and toxic gases such as carbon monoxide (CO), nitrogen monoxide (NO), Nitrogen Dioxide (NO2), for instance, can cause serious health problems. Gas monitoring systems can help in monitoring the indoor air quality using various gas sensors. Big data systems can also be used to analyze risks in industrial operations and identify the hazardous zones, so that corrective measures can be taken and timely alerts can be raised in case of any abnormal conditions.

Production Planning and Control: Production planning and control systems measure various parameters of production processes and control the entire production process in real-time. These systems use various sensors to collect data on the production processes. Big data systems can be used to analyze this data for production planning and identifying potential problems.

8.8 Retail

Retailers can use big data systems for boosting sales, increasing profitability and improving customer satisfaction. Let us look at some applications of big data analytics for retail:

Inventory Management: Inventory management for retail has become increasingly important in the recent years with the growing competition. While over-stocking of products can result in additional storage expenses and risk (in case of perishables), under-stocking can lead to loss of revenue. RFID tags attached to the products allow them to be tracked in real-time so that the inventory levels can be determined accurately and products which are low on stock can be replenished. Tracking can be done using RFID readers attached to the retail store shelves or in the warehouse. Big data systems can be used to analyze the data collected from RFID readers and raise alerts when inventory levels for certain products are low. Timely replenishment of inventory can help in minimizing the loss in revenue due to out-of-stock inventory. Analysis of inventory data can help in optimizing the re-stocking levels and frequencies based on demand.

Customer Recommendations: Big data systems can be used to analyze the customer data (such as demographic data, shopping history, or customer feedback) and predict the customer preferences. New products can be recommended to customers based on the customer preferences and personalized offers and discounts can be given. Customers with similar preferences can be grouped and targeted campaigns can be created for customers.

Store Layout Optimization: Big data systems can help in analyzing the data on customer shopping patterns and customer feedback to optimize the store layouts. Items which the customers are more likely to buy together can be placed in the same or nearby racks.

Forecasting Demand: Due to a large number of products, seasonal variations in demands and changing trends and customer preferences, retailers find it difficult to forecast demand

and sales volumes. Big data systems can be used to analyze the customer purchase patterns and predict demand and sale volumes.

9. Analytics flow for Big Data:

A generic flow for big data analytics, detailing the steps involved in the implementation of a typical analytics application and the options available at each step, is presented. Figure 1.2 shows the analytics flow with various steps. For an application, selecting the options for each step in the analytics flow can help in determining the right tools and frameworks to perform the analyses.

9.1 Data Collection

Data collection is the first step for any analytics application. Before the data can be analyzed, the data must be collected and ingested into a big data stack. The choice of tools and frameworks for data collection depends on the source of data and the type of data being ingested. For data collection, various types of connectors can be used such as publish-subscribe messaging frameworks, messaging queues, source-sink connectors, database connectors and custom connectors.

9.2 Data Preparation

Data can often be dirty and can have various issues that must be resolved before the data can be processed, such as corrupt records, missing values, duplicates, inconsistent abbreviations, inconsistent units, typos, incorrect spellings and incorrect formatting. Data preparation step involves various tasks such as data cleansing, data wrangling or munging, de-duplication, normalization, sampling and filtering. Data cleaning detects and resolves issues such as corrupt records, records with missing values, records with bad formatting, for instance. Data wrangling or munging deals with transforming the data from one raw format to another. For example, when we collect records as raw text files form different sources, we may come across inconsistencies in the field separators used in different files. Some file may be using comma as the field separator, others may be using tab as the field separator. Data wrangling resolves these inconsistencies by parsing the raw data from different sources and transforming it into one consistent format. Normalization is required when data from different sources uses different units or scales or have different abbreviations for the same thing. For example, weather data reported by some stations may contain temperature in Celsius scale while data from other stations may use the Fahrenheit scale. Filtering and sampling may be useful when we want to process only the data that meets certain rules. Filtering can also be useful to reject bad records with incorrect or out-of-range values.

9.3 Analysis Types

The next step in the analysis flow is to determine the analysis type for the application. In Figure 1.2 we have listed various options for analysis types and the popular algorithms for each analysis type.

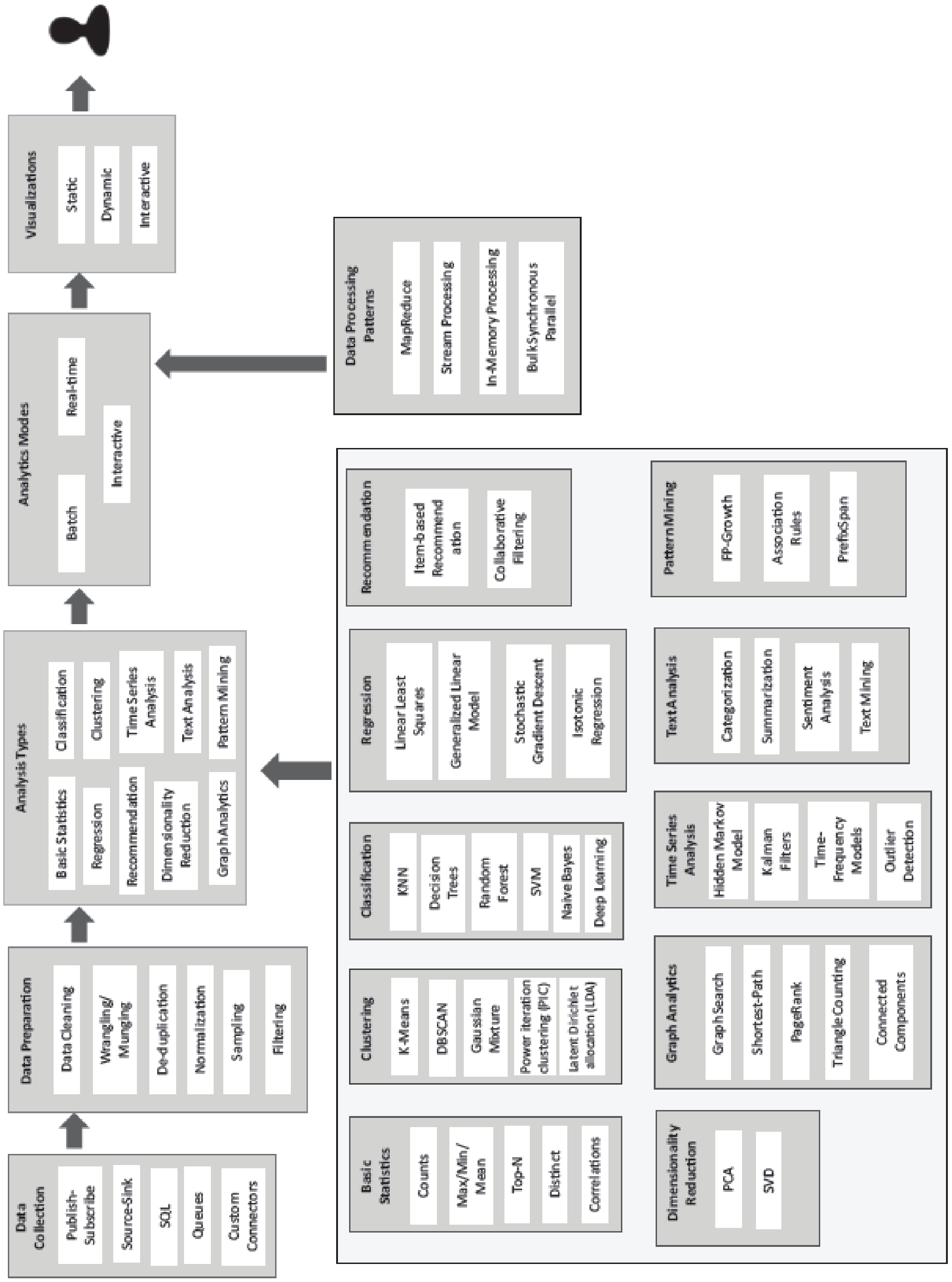


Figure 1.2: Big Data analytics flow

9.4 Analysis Modes

With the analysis types selected for an application, the next step is to determine the analysis mode, which can be either batch, real-time or interactive. The choice of the mode depends on the requirements of the application. If your application demands results to be updated after short intervals of time (say every few seconds), then real-time analysis mode is chosen. However if your application only requires the results to be generated and updated on larger timescales (say daily or monthly), then batch mode can be used. If your application demands flexibility to query data on demand, then the interactive mode is useful. Once you make a choice of the analysis type and the analysis mode, you can determine the data processing pattern that can be used. For example, for basic statistics as the analysis type and the batch analysis mode, MapReduce can be a good choice. Whereas for regression analysis as the analysis type and real-time analysis mode (predicting values in real-time), the Stream Processing pattern is a good choice. The choice of the analysis type, analysis mode, and the data processing pattern can help you in shortlisting the right tools and frameworks for data analysis.

9.5 Visualization

The choice of the visualization tools, serving databases and web frameworks is driven by the requirements of the application. Visualizations can be static, dynamic or interactive. Static visualizations are used when you have the analysis results stored in a serving database and you simply want to display the results. However, if your application demands the results to updated regularly, then you would require dynamic visualizations (with live widgets, plots, or gauges). If you want your application to accept inputs from the user and display the results, then you would require interactive visualizations.

10. Big Data Stack: The big data stack comprises of the following elements. They are

- ✓ Raw data sources
- ✓ Data Access Connectors
- ✓ Data Storage
- ✓ Batch Analytics
- ✓ Real time Analytics
- ✓ Interactive Querying

10.1. Raw Data Sources: In any big data analytics application or platform, before the data is processed and analyzed, it must be captured from the raw data sources into the big data systems and frameworks. Some of the examples of raw big data sources include:

- o Logs: Logs generated by web applications and servers which can be used for performance monitoring
- o Transactional Data: Transactional data generated by applications such as eCommerce, Banking and Financial
- o Social Media: Data generated by social media platforms
- o Databases: Structured data residing in relational databases
- o Sensor Data: Sensor data generated by Internet of Things (IoT) systems
- o Clickstream Data: Clickstream data generated by web applications which can be used to analyze browsing patterns of the users
- o Surveillance Data: Sensor, image and video data generated by surveillance systems
- o Healthcare Data: Healthcare data generated by Electronic Health Record (EHR) and other healthcare applications
- o Network Data: Network data generated by network devices such as routers and firewalls

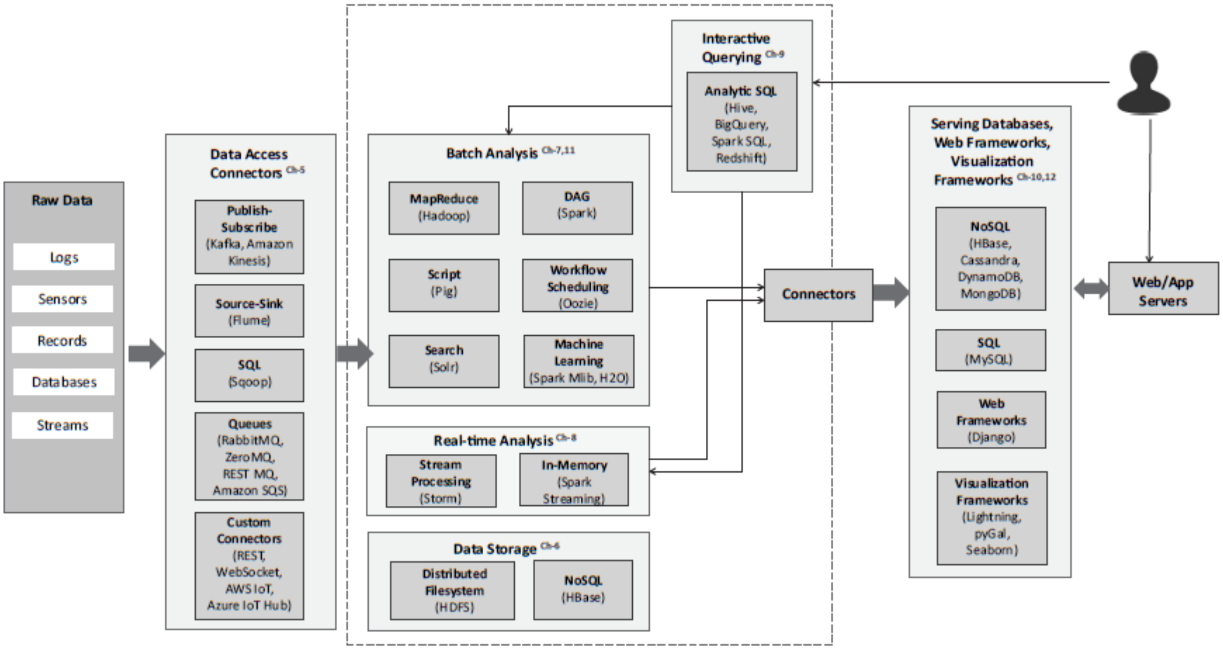

Raw Data
Logs
Sensors
Records
Databases
Streams
Data Access Connectors Ch-5
Publish-Subscribe (Kafka, Amazon Kinesis)
Source-Sink (Flume)
SQL (Sqoop)
Queues (RabbitMQ, ZeroMQ, REST MQ, Amazon SQS)
Custom Connectors (REST, WebSocket, AWS IoT, Azure IoT Hub)
Batch Analysis Ch-7,11
MapReduce (Hadoop)
DAG (Spark)
Script (Pig)
Workflow Scheduling (Oozie)
Search (Solr)
Machine Learning (Spark Mlib, H2O)
Real-time Analysis Ch-8
Stream Processing (Storm)
In-Memory (Spark Streaming)
Data Storage Ch-6
Distributed Filesystem (HDFS)
NoSQL (HBase)
Interactive Querying Ch-9
Analytic SQL (Hive, BigQuery, Spark SQL, Redshift)
Connectors
Serving Databases, Web Frameworks, Visualization Frameworks Ch-10,12
NoSQL (HBase, Cassandra, DynamoDB, MongoDB)
SQL (MySQL)
Web Frameworks (Django)
Visualization Frameworks (Lightning, pyGal, Seaborn)
Web/App Servers

Figure 1.3: Big Data Stack

10.2 Data Access Connectors

The Data Access Connectors includes tools and frameworks for collecting and ingesting data from various sources into the big data storage and analytics frameworks. The choice of the data connector is driven by the type of the data source.

- Publish-Subscribe Messaging: Publish-Subscribe is a communication model that involves publishers, brokers and consumers. Publishers are the source of data. Publishers send the data to the topics which are managed by the broker.
- Source-Sink Connectors: Source-Sink connectors allow efficiently collecting, aggregating and moving data from various sources (such as server logs, databases, social media, streaming sensor data from Internet of Things devices and other sources) into a centralized data store
- Database Connectors: Database connectors can be used for importing data from relational database management systems into big data storage and analytics frameworks for analysis.
- Messaging Queues: Messaging queues are useful for push-pull messaging where the producers push data to the queues and the consumers pull the data from the queues. The producers and consumers do not need to be aware of each other.

- Custom Connectors: Custom connectors can be built based on the source of the data and the data collection requirements. Some examples of custom connectors include:custom connectors for collecting data from social networks, custom connectors for NoSQL databases and connectors for Internet of Things (IoT).

10.3 Data Storage

The data storage block in the big data stack includes distributed filesystems and non-relational (NoSQL) databases, which store the data collected from the raw data sources using the data access connectors. Hadoop Distributed File System (HDFS), a distributed file system that runs on large clusters and provides high-throughput access to data. With the data stored in HDFS, it can be analyzed with various big data analytics frameworks built on top of HDFS. For certain analytics applications, it is preferable to store data in a NoSQL database such as HBase. HBase is a scalable, non-relational, distributed, column-oriented database that provides structured data storage for large tables.

10.4 Batch Analytics

The batch analytics block in the big data stack includes various frameworks which allow analysis of data in batches. These include the following:

- Hadoop-MapReduce: Hadoop is a framework for distributed batch processing of big data. The MapReduce programming model is used to develop batch analysis jobs which are executed in Hadoop clusters.
- Pig: Pig is a high-level data processing language which makes it easy for developers to write data analysis scripts which are translated into MapReduce programs by the Pig compiler.
- Oozie: Oozie is a workflow scheduler system that allows managing Hadoop jobs. With Oozie, you can create workflows which are a collection of actions (such as MapReduce jobs) arranged as Direct Acyclic Graphs (DAG).
- Spark: Apache Spark is an open source cluster computing framework for data analytics. Spark includes various high-level tools for data analysis such as Spark Streaming for streaming jobs, Spark SQL for analysis of structured data

10.5 Real-time Analytics:

The real-time analytics block includes the Apache Storm and Spark Streaming frameworks. Apache Storm is a framework for distributed and fault-tolerant real-time computation. Storm can be used for real-time processing of streams of data. Storm can consume data from a variety of sources such as publish-subscribe messaging frameworks, messaging queues as RabbitMQ or ZeroMQ) and other custom connectors. Spark Streaming is a component of Spark which allows analysis of streaming data such as sensor data, click stream data, web server logs, for instance. The streaming data is ingested and analyzed in micro-batches. Spark Streaming enables scalable, high throughput and fault-tolerant stream processing.

10.6 Interactive Querying

Interactive querying systems allow users to query data by writing statements in SQL-like languages. Some systems used for querying include:

- Spark SQL: Spark SQL is a component of Spark which enables interactive querying. Spark SQL is useful for querying structured and semi-structured data using SQL-like queries.
- Hive: Apache Hive is a data warehousing framework built on top of Hadoop. Hive provides an SQL-like query language called Hive Query Language, for querying data residing in HDFS.
- Amazon Redshift: Amazon Redshift is a fast, massive-scale managed data warehouse service. Redshift specializes in handling queries on datasets of sizes up to a petabyte or more parallelizing the SQL queries across all resources in the Redshift cluster.
- Google BigQuery: Google BigQuery is a service for querying massive datasets. BigQuery allows querying datasets using SQL-like queries.

10.7 Serving Databases, Web & Visualization Frameworks

While the various analytics blocks process and analyze the data, the results are stored in serving databases for subsequent tasks of presentation and visualization. These serving databases allow the analyzed data to be queried and presented in the web applications.

- MySQL: MySQL is one of the most widely used Relational Database Management System (RDBMS) and is a good choice to be used as a serving database for data analytics applications where the data is structured.
- Amazon DynamoDB: Amazon DynamoDB is a fully-managed, scalable, high-performance NoSQL database service from Amazon. DynamoDB is an excellent choice for a serving database for data analytics applications as it allows storing and retrieving any amount of data and the ability to scale up or down the provisioned throughput.
- Cassandra: Cassandra is a scalable, highly available, fault tolerant open source non-relational database system.
- MongoDB: MongoDB is a document oriented non-relational database system. MongoDB is powerful, flexible and highly scalable database designed for web applications and is a good choice for a serving database for data analytics applications.

The following tools are used for visualization tools and frameworks:

- Lightning: Lightning is a framework for creating web-based interactive visualizations.
- Pygal: The Python Pygal library is an easy to use charting library which supports charts of various types.
- Seaborn: Seaborn is a Python visualization library for plotting attractive statistical plots.

Mapping Analytics Flow to Big Data Stack

For data collection tasks, the choice of a specific tool or framework depends on the type of the data source (such as log files, machines generating sensor data, social media feeds, records in a relational database, for instance) and the characteristics of the data. If the data is to ingested in bulk (such as log files), then a source-sink such as Apache Flume can be used. However, if high-velocity data is to be ingested at real-time, then a distributed publish-subscribe messaging framework such as Apache Kafka or Amazon Kinesis can be used. For ingesting data from relational databases, a framework such as Apache Sqoop can be used. Custom connectors can be built based on HTTP/REST, WebSocket or MQTT, if other solutions don’t work well for an application or there are additional constraints. For example, IoT devices generating sensor data may be resource and power constrained, in which case a light-weight communication protocol such as MQTT may be chosen and a custom MQTT-based connector can be used.

For data cleaning and transformation, tools such as Open Refine and Stanford DataWrangler can be used. These tools support various file formats such as CSV, Excel, XML, JSON and line-based formats. With these tools you can remove duplicates, filter records with missing values, trim leading and trailing spaces, transpose rows to columns, transform the cell values, cluster similar cells and perform various other transformations. For filtering, joins, and other transformations, high-level scripting frameworks such as Pig can be very useful. The benefit of using Pig is that you can process large volumes of data in batch mode, which may be difficult with standalone tools. When you are not sure about what transformation should be applied and want to explore the data and try different transformations, then interactive querying frameworks such as Hive, SparkSQL can be useful.With these tools, you can query data with queries written in an SQL-like language.

For the basic statistics analysis type (with analysis such as computing counts, max, min, mean, top-N, distinct, correlations, for instance), most of the analysis can be done using the Hadoop-MapReduce framework or with Pig scripts. Both MapReduce and Pig allow data analysis in batch mode. For basic statistics in batch mode, the Spark framework is also a good option. For basic statics in real-time mode, Spark Streaming and Storm frameworks can be used. For basic statistics in interactive mode, a framework such as Hive and SparkSQL can be used.

Data Collection

Analysis Type	Framework (Mode)
Publish-Subscribe	Kafka, Kinesis
Source-Sink	Flume
SQL	Sqoop
Queues	SQS, RabbitMQ, ZeroMQ, RESTMQ
Custom Connectors	REST, WebSocket, MQTT

Data Preparation

Analysis Type	Framework
Data Cleaning	Open Refine
Data Wrangling	Open Refine DataWrangler
De-Duplication	Open Refine, Pig, Hive, Spark SQL
Normalization Sampling, Filtering	MapReduce, Pig, Hive, Spark SQL

Basic Statistics

Analysis Type	Framework (Mode)
Counts, Max, Min, Mean, Top-N, Distinct	Hadoop-MapReduce (Batch), Pig (Batch), Spark (Batch), Spark Streaming (Realtime), Spark SQL (Interactive), Hive (Integrative), Storm (Real-time)
Correlations	Hadoop-MapReduce (Batch), Spark Mlib (Batch)

Clustering

Analysis Type	Framework (Mode)
K-Means	Hadoop-MapReduce (Batch), Spark Mlib (Batch & Real-time) H2O (Batch)
DBSCAN	Spark (Batch)
Gaussian Mixture	Spark Mlib (Batch)
PIC	Spark Mlib (Batch)
LDA	Spark Mlib (Batch)

Classification

Analysis Type	Framework (Mode)
KNN	Spark Mlib (Batch , Realtime)
Decision Trees	Spark Mlib (Batch, Realtime)
Random Forest	Spark Mlib (Batch , Realtime), H2O (Batch)
SVM	Spark Mlib (Batch , Realtime)
Naïve Bayes	Spark Mlib (Batch, Realtime), H2O (Batch)
Deep Learning	H2O (Batch)

Regression

Analysis Type	Framework (Mode)
Linear Least Squares	Spark Mlib (Batch, Realtime)
Generalized Linear Model	H2O (Batch)
Stochastic Gradient Descent	Spark Mlib (Batch, Realtime)
Isotonic Regression	Spark Mlib (Batch, Realtime)

Assignment Questions:

1. Explain the four V’s of Big Data.
2. Explain the various sources of Big Data.
3. Explain the various types of analytics in Big Data
4. Explain the classification of Big Data
5. Explain the applications of Big Data.

Essay Question (10 marks)

1. Explain the various elements of Big Data Stack with a neat diagram
2. Explain the analytics flow of big data with a neat diagram