(2½ Hours)

[Total Marks: 60]

N. B.: (1) **All** questions are **compulsory**.
(2) Make **suitable assumptions** wherever necessary and **state the assumptions** made.
(3) Answers to the **same question** must be **written together**.
(4) Numbers to the **right** indicate **marks**.
(5) Draw **neat labelled diagrams** wherever **necessary**.
(6) Use of **Non-programmable** calculators is **allowed**.

I	**Choose the correct alternative and rewrite the entire sentence with the correct alternative.** **(30)**			
1.	Big data is high-velocity and high-variety information assets that demand cost effective, innovative forms of information processing for enhanced insight and decision making.			
	a.	Data discovery	**b.**	Data Science
	c.	Big Data	**d.**	Statistical analysis
2.	Data science is the science of extracting knowledge from data.			
	a.	Data analyst	**b.**	Data Science
	c.	Big Data	**d.**	Statistical analysis
3.	Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware.			
	a.	Mapreduce	**b.**	Spark
	c.	HBase	**d.**	Hadoop
4.	Data in TERA bytes size is called Big Data.			
	a.	TERA	**b.**	PETA
	c.	GIGA	**d.**	META
5.	In Data Analytics Lifecycle, Data preparation requires the presence of an analytic sandbox, in which the team can work with data and perform analytics.			
	a.	Data preparation	**b.**	Data discovery
	c.	distributed file system	**d.**	Data analytics
6.	Database such as Oracle, DB2, Teradata, MySQL,PostgreSQL, preadsheets, OLTP systems are the source of Structured Data			
	a.	Structured Data	**b.**	Semi-Structured Data
	c.	UnStructured Data	**d.**	Simple Data
7.	K-means clustering algorithm determines the distance between an object and its cluster centroid by Euclidean Distance measure.			
	a.	Hamming Distance	**b.**	Manhattan Distance
	c.	Euclidean Distance	**d.**	Minkowski Distance

8.	Association rule mining uses models to analyze data for patterns, transactions that consist of one or more items in a database.			
	a.	K-Means	**b.**	Association rule
	c.	KNN Clustering	**d.**	SVM Clustering
9.	Partitioning is one of the approaches to improve Apriori's efficiency, according to which, any item set that is potentially frequent in a transaction database, must be frequent in at least one of the partitions of the transaction database.			
	a.	Partitioning	**b.**	Sampling
	c.	Hash-based item set counting	**d.**	Dynamic item set counting
10.	Linear regression is an analytical technique used to model the relationship between several input variables and a continuous outcome variable.			
	a.	Regression	**b.**	Logistic regression
	c.	Classification	**d.**	Linear regression
11.	In Regression, there are relationship between dependent variable and independent variables.			
	a.	Logistic regression	**b.**	Regression
	c.	Linear regression	**d.**	Association Rule
12.	In regression model, when the outcome variable is a categorical variable, then Logistic regression can be used.			
	a.	Association Rule	**b.**	Logistic regression
	c.	K-Means	**d.**	Linear regression
13.	A Decision tree is also called prediction tree , which uses a tree structure to specify sequences of decisions and consequences.			
	a.	Naïve Bayes	**b.**	K-Means
	c.	Decision tree	**d.**	Clustering
14.	In Naïve Bayes, the input variables are generally convert continuous variables into categorical ones and this process is referred as the discretization of continuous variables			
	a.	discontinues	**b.**	different
	c.	attributes	**d.**	discretization
15.	Raw text is converted into collections of tokens after the tokenization, where each token is generally a word.			
	a.	bag-of-words	**b.**	case folding
	c.	word	**d.**	tokenization
16.	In Classification methods, Bagging uses the bootstrap technique that repeatedly samples with replacement from a dataset according to a uniform probability distribution.			
	a.	SVM	**b.**	sampling
	c.	bootstrap	**d.**	AdaBoost

17.	Random forest is a class of ensemble methods using decision tree classifiers			
	a.	AdaBoost	**b.**	k-means
	c.	Logistic regression	**d.**	Random forest
18.	ID3 (or Iterative Dichotomiser 3) is one of the first decision tree algorithms, and it was developed by John Ross Quinlan			
	a.	ID3 - Iterative Dichotomiser 3	**b.**	AdaBoost
	c.	Random forest	**d.**	Logistic regression
19.	YARN allows the data stored in HDFS (Hadoop Distributed File System) to be processed and run by various data processing engines such as batch processing, stream processing, interactive processing, graph processing			
	a.	Mapreduce	**b.**	Data mining
	c.	YARN	**d.**	DataWarehousing
20.	Hadoop Streaming allows to create and run MapReduce jobs with any script or executable as the mapper or the reducer.			
	a.	streaming	**b.**	MapReduce
	c.	HIVE	**d.**	YARN
21.	**Apache Hive** is a Data warehousing tool that is built on top of the Hadoop, and is one of the best tools used for data analysis on Hadoop.			
	a.	ETL	**b.**	Data mining
	c.	YARN	**d.**	Apache Hive
22.	The NameNode is the Hadoop Distributed File System, which keeps the directory tree of all files in the file system, and tracks the locations of all files in the cluster.			
	a.	NameNode	**b.**	DataNode
	c.	MapReduce	**d.**	ResourceManager
23.	By default in Spark, the Python API uses the Pickle module for serialization.			
	a.	JSON	**b.**	complex
	c.	NameNode	**d.**	Pickle
24.	An inverted index is a mapping from an index term to locations in a set of documents.			
	a.	DROP index	**b.**	inverted index
	c.	forward index	**d.**	alter index
25.	HBase is classified as Column family			
	a.	Complex	**b.**	Row family
	c.	Column family	**d.**	Tree family
26.	Sqoop is a tool designed to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases.			
	a.	HIVE	**b.**	HBase
	c.	Sqoop	**d.**	SQL

27.	A Flume agent is a JVM process.			
	a.	JRE	**b.**	JAVA
	c.	JVM	**d.**	SDK
28.	Load function is used to read data in Pig.			
	a.	Show	**b.**	Scan
	c.	Read	**d.**	Load
29.	Spark SQL automatically infers the schema of a JSON datasets.			
	a.	SQL	**b.**	JSON
	c.	XML	**d.**	HIVE
30.	In flume the data flow from many sources to one channel is known as Fan-in Flow			
	a.	Fan-in Flow	**b.**	fan out
	c.	Single agent	**d.**	Multiple flow

II	**Attempt *any one* of the following:**		**6**
	a)	Define big data. Why is big data required? How does traditional BI environment differ from big data environment? ANS: Big data is high-velocity and high-variety information assets that demand cost effective, innovative forms of information processing for enhanced insight and decision making. Big data requires storage. To improve operations, provide better customer service, create personalized marketing campaigns and take other actions that, ultimately, can increase revenue and profits. The analytical accuracy will lead a greater positive impact in terms of enhancing operational efficiencies, reducing cost and time, and originating new products, new services, and optimizing existing services. In traditional BI environment, all the enterprise's data is housed in a central server whereas in a big data environment data resides in a distributed file system. The distributed file system scales by scaling in(decrease) or out(increase) horizontally as compared to typical database server that scales vertically. In traditional BI, data is generally analysed in an offline mode whereas in big data, it is analysed in both real-time streaming as well as in offline mode. Traditional Bl is about structured data and it is here that data is taken to processing functions (move data to code) whereas big data is about variety: Structured, semi-structured, and unstructured data and here the processing functions are taken to the data (move code to data).	
	b)	Write a short note on Classification of Analytics. ANS: Classification analysis is a data analysis task within data-mining, that identifies and assigns categories to a collection of data to allow for more accurate analysis. The classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and	

		statistics.Classification analysis can be used to question, make a decision, or predict behavior through the use of an algorithm. Basic analytics is slicing and dicing of data to help with basic business insights. This is about reporting on historical data, basic visualization, etc. There are four Analytics: **Descriptive analytics** is used to understand the overall performance at an aggregate level and is by far the easiest place for a company to start as data tends to be readily available to build reports and applications. **Diagnostic analytics**, tends to be more accessible and fit a wider range of use cases than machine learning/predictive analytics. It might even find that it solves some business problems you earmarked for predictive analytics use cases. **Predictive analytics** is a form of advanced analytics that determines what is likely to happen based on historical data using machine learning. Historical data that comprises the bulk of descriptive and diagnostic analytics is used as the basis of building predictive analytics models. Predictive analytics helps companies address use cases such as: Predicting maintenance issues and part breakdown in machines. Determining credit risk and identifying potential fraud. Predict and avoid customer churn by identifying signs of customer dissatisfaction. **Prescriptive analytics** requires strong competencies in descriptive, diagnostic, and predictive analytics which is why it tends to be found in highly specialized industries (oil and gas, clinical healthcare, finance, and insurance to name a few) where use cases are well defined. Prescriptive analytics help to address use cases such as: Automatic adjustment of product pricing based on anticipated customer demand and external factors. Flagging select employees for additional training based on incident reports in the field. Prescriptive analytics primary aim is to take the educated guess or assessment out of data analytics and streamline the decision-making process.	
	c)	Explain different phases of the Data Analytics Lifecycle each in detail ANS: **Phase 1**- Discovery: In Phase 1, the team learns the business domain, including relevant history such as whether the organization or business unit has attempted similar projects in the past from which they can learn. The team assesses the resources available to support the project in terms of people, technology, time and data. Important activities in this phase include framing the business problem as an analytics challenge that can be addressed in subsequent phases and formulating initial hypotheses (IHs) to test and begin learning the data. **Phase 2**- Data preparation: Phase 2 requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project. The team needs to execute extract, load, and transform (ELT) or extract, transform and load (ETL) to get data into the sandbox. The ELT and ETL are sometimes abbreviated as ETLT. Data should be transformed in the ETLT process so the team can work with it and analyze it. In this phase, the team also needs to familiarize itself with	

		the data thoroughly and take steps to condition the data. **Phase 3**-Model planning: Phase 3 is model planning, where the team 29 determines the methods, techniques and workflow it intends to follow for the subsequent model building phase. The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models. **In Phase 4** Model building: , the team develops data sets for testing, training, and production purposes. In addition, in this phase the team builds and executes models based on the work done in the model planning phase. The team also considers whether its existing tools will suffice for running the models, or if it will need a more robust environment for executing models and workflows (for example, fast hardware and parallel processing, if applicable). **Phase 5**-Communicate results: In Phase 5, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in Phase 1. The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders. **Phase 6**-0perationalize: In Phase 6, the team delivers final reports, briefings, code and technical documents. In addition, the team may run a pilot project to implement the models in a production environment.	
2	**Attempt *<u>any one</u>* of the following:**		**6**
	a)	What is Linear regression? Explain in detail. Also explain any two of its applications. ANS: Linear regression is an analytical technique used to model the relationship between several input variables and a continuous outcome variable. A key assumption is that the relationship between an input variable and the outcome variable is linear. Although this assumption may appear restrictive, it is often possible to properly transform the input or outcome variables to achieve a linear relationship between the modified input and outcome variables. A linear regression model is a probabilistic one that accounts for the randomness that can affect any particular outcome. Based on known input values, a linear regression model provides the expected value of the outcome variable based on the values of the input variables, but some uncertainty may remain in predicting any particular outcome. <u>Application:</u> • Real estate: A simple linear regression analysis can be used to model residential home prices as a function of the home's living area. Such a model helps set or evaluate the list price of a home on the market. The model could be further improved by including other input variables such as number of bathrooms, number of bedrooms, lot size, school district rankings, crime statistics, and property taxes • Demand forecasting: Businesses and governments can use linear regression models to predict demand for goods and services. For example, restaurant chains can appropriately prepare for the predicted type and quantity of food that customers will consume based upon the weather, the day of the week, whether an item is offered as a special, the time of day, and the reservation volume. Similar models can be built to predict retail sales, emergency room visits, and ambulance dispatches. • Medical: A linear regression model can be used to analyze the effect of a proposed radiation treatment on reducing tumor sizes. Input variables might include duration of	

		a single radiation treatment, frequency of radiation treatment, and patient attributes such as age or weight.	
	b)	Write a short note on association rules ANS: An unsupervised learning method called association rules. This is a descriptive, not predictive, method often used to discover interesting relationships hidden in a large dataset. The disclosed relationships can be represented as rules or frequent item sets. Association rules are commonly used for mining transactions in databases. Using association rules, patterns can be discovered from the data that allow the association rule algorithms to disclose rules of related product purchases. The uncovered rules are listed on the right side of Figure. The first three rules suggest that when cereal is purchased, 90% of the time milk is purchased also. When bread is purchased, 40% of the time milk is purchased also. When milk is purchased, 23% of the time cereal is also purchased. In the example of a retail store, association rules are used over transactions that consist of one or more items. In fact, because of their popularity in mining customer transactions, association rules are 42 sometimes referred to as market basket analysis. Each transaction can be viewed as the shopping basket of a customer that contains one or more items. This is also known as an itemset. The term itemset refers to a collection of items or individual entities that contain some kind of relationship. This could be a set of retail items purchased together in one transaction, a set of hyperlinks clicked on by one user in a single session, or a set of tasks done in one day. An itemset containing k items is called a k-itemset denoted by {item1,item 2, . . . item k}.	
	c)		
3	**Attempt *<u>any one</u>* of the following:**		**6**
	a)	Write a short note on decision tree. ANS: A decision tree (also called prediction tree) uses a tree structure to specify sequences of decisions and consequences. Given input X = {x1 ,x2 ,...xn}, the goal is to predict a response or output variable Y. Each member of the set {x1 ,x2 ,...xn} is called an input variable. The prediction can be achieved by constructing a decision tree with test points and branches. At each test point, a decision is made to pick a specific branch and traverse down the tree. Eventually, a final point is reached, and a prediction can be made. Due to its flexibility and easy visualization, decision trees are commonly deployed in data mining applications for classification purposes. The input values of a decision tree can be categorical or 63 continuous. A decision tree employs a structure of test points (called nodes) and branches, which represent the decision being made. A node without further branches is called a leaf node. The leaf nodes return class labels and, in some implementations, they return the probability scores. A decision tree can be converted into a set of decision rules. In the following example rule, income and mortgage_amount are input variables, and the response is the output variable default with a probability score. IF income 100K THEN default = True WITH PROBABILITY 75% Decision trees have two varieties: classification trees and regression trees. Classification trees usually apply to output variables that are categorical—often	

		binary—in nature, such as yes or no, purchase or not purchase, and so on. Regression trees, on the other hand, can apply to output variables that are numeric or continuous, such as the predicted price of a consumer good or the likelihood a subscription will be purchased.	
	b)	Explain briefly the Naïve Bayes. ANS: Naive Bayes is a probabilistic classification method based on Bayes' theorem. Bayes' theorem gives the relationship between the probabilities of two events and their conditional probabilities. A naive Bayes classifier assumes that the presence or absence of a particular feature of a class is unrelated to the presence or absence of other features. For example, an object can be classified based on its attributes such as shape, color, and weight. The input variables are generally categorical, but variations of the algorithm can accept continuous variables, There are also ways to convert continuous variables into categorical ones. This process is often referred to as the discretization of continuous variables. For an attribute such as income, the attribute can be converted into categorical values as shown below. • Low Income: income < $10,000 • Working Class: $10,000 < income < $50,000 • Middle Class: $50,000 < income < $1,000,000 • Upper Class: income >$1,000,000 The output typically includes a class label and its corresponding probability score. The probability score is not the true probability of the class label, but it's proportional to the true probability. Because naive Bayes classifiers are easy to implement and can execute efficient. Spam filtering is a classic use case of naive Bayes text classification. Bayesian spam filtering has become a popular mechanism to distinguish spam e-mail from legitimate e-mail. Naive Bayes classifiers can also be used for fraud detection. In the domain of auto insurance, for example, based on a training set with attributes such as driver's rating, vehicle age, vehicle price, historical claims by the policy holder, police report status, and claim genuineness, naive Bayes can provide probability-based classification of whether a new claim is genuine.	
	c)	Explain the steps involved in Text analysis with example. ANS: Text analysis, sometimes called text analytics, refers to the representation, processing, and modeling of textual data to derive useful insights. An important component of text analysis is text mining, the process of discovering relationships and interesting patterns in large text collections. Text analysis often deals with textual data that is far more complex. A corpus (plural: corpora) is a large collection of texts used for various purposes in Natural Language Processing (N LP). Another major challenge with text analysis is that most of the time the text is not structured. A text analysis problem usually consists of three important steps: • Parsing • Search and Retrieval • Text Mining. **Parsing** is the process that takes unstructured text and imposes a structure for further analysis. The unstructured text could be a plain text file, a weblog, an Extensible Markup Language (XML) file, a Hyper Text Markup Language (HTML) file, or a Word document. Parsing deconstructs the provided text and renders it in a more structured way for the subsequent steps.	

		Search and retrieval is the identification of the documents in a corpus that contain search items such as specific words, phrases, topics, or entities like people or organizations. These search items are generally called key terms. Search and retrieval originated from the field of library science and is now used extensively by web search engines. **Text mining** uses the terms and indexes produced by the prior two steps to discover meaningful insights pertaining to domains or problems of interest. With the proper representation of the text, many of the techniques such as clustering and classification, can be adapted to text mining.	
4	**Attempt *any one* of the following:**		**6**
	a)	Explain the 4 stages of the Big Data Pipeline. ANS: The 4 stages of the Big Data Pipeline are: • staging, • ingestion, • computation, • workflow management • In its most basic form, this model, like the data science pipeline, takes raw data and transforms it into insights. • This stage generates a reusable data product as the output by transforming the ingestion, staging, and computation phases into an automated workflow. • A feedback system is often needed during the workflow management stage, that gives the output of one job can be automatically fed in as the data input for the next, allowing for self-adaptation. • The ingestion phase involves both the model's initialization and the model's device interaction with users. 101 • Users may define data source locations or annotate data during the initialization process • While interacting, users will receive the predictions given by the model and in turn give important feedback to strengthen the model • The staging step requires executing transformations on data to make it usable and storeable, allowing it to be processed. • The tasks of staging include data normalization, standardization & data management. • The computation phase takes maximum time while executing the key responsibilities of extracting insights from data, conducting aggregations or reports, and developing machine learning models for recommendations, regressions, clustering, or classification. • The workflow management phase involves tasks such as abstraction, orchestration, and automation, which enables to operationalize the performance of workflow steps. The final output is supposed to be an program that is automated that can be run as desired.	
	b)	Explain the concept of MapReduce. ANS: MapReduce (also known as MR) was the first Hadoop computational system. • MapReduce is a straightforward but effective computational system for fault-tolerant distributed computing across a cluster of centrally controlled machines. It	

		accomplishes this by using a "functional" programming style that is essentially parallelizable, allowing several independent tasks to perform a function on local groups of data and then combining the results. • Functional programming is a programming methodology that guarantees stateless evaluation of unit computations. This implies that functions are closed, in the sense that they do not exchange state and depend solely on their inputs. Data is transferred between functions by using the output of one function as the input of a completely different function. • Map Reduce provides the two functions that distribute work and aggregate results called map and reduce **Map Function** • MapReduce offers the map and reduce functions, which distribute work and aggregate results. • A map function takes a list of key/value pairs as input and works on each pair separately. • The map operation is where the core analysis or processing takes place, as this is the function that sees each individual element in the dataset **Reduce Function** • Any emitted key/value pairs will be grouped by key after the map phase, and those key/value groups will be used as input for per-key minimization functions. • When a reduce function is applied to an input set, the output is a single, aggregated value.	
	c)	Write a short note on Hadoop Distributed File System ANS: • HDFS doubles the amount of storage space available from a single computer by storing it via a cluster of low-cost, unreliable devices. • HDFS is a layer of software that sits on top of a native file system, allowing it to communicate with local file systems and generalising the storage layer. • HDFS was built with the aim of storing large files while still allowing for real-time access to data. • For storing raw input data for computation, intermediate results between computational phases, and overall job results, HDFS is the best choice. • HDFS is not good as a data backend for applications requiring realtime updates, interactive analysis and record based transactional support. Following are few characteristics of HDFS: • HDFS is best suited to a small number of very large files—for example, millions of large files (greater than 100 MB in size) 107 rather than billions of smaller files that would otherwise occupy the same amount of space. • HDFS follows the WORM (write once, read many) pattern and does not permit random file appends or writes. • HDFS is designed for large-scale, continuous file reading rather than random reading or collection.	
5	**Attempt *any one* of the following:**		**6**
	a)	What are the Distributed Analysis and Patterns. ANS: MapReduce and Spark allow developers and data scientists the ability to easily conduct data parallel operations, where data is distributed to multiple processing nodes	

		and computed upon simultaneously, then reduced to a final output. YARN provides simple task parallelism by allowing a cluster to perform multiple different operations simultaneously by allocating free computational resources to perform individual tasks. Parallelism reduces the amount of time required to perform a single computation, thereby unlocking datasets that are measured in petabytes, analyzed at thou- sands of records per second, or composed of multiple heterogeneous data sources. The primary principle of conducting large-scale analytics can be summarized by the quip from Creighton Abrams: "When eating an elephant, take one bite at a time." Whereas single operations take many small bites of the data, these operations must be composed into a step-by- step sequence called a data flow to be organized into more meaningful results. Data flows may fork and merge, allowing for both task and data parallelism if two operations can be computed simultaneously, but the sequence must maintain the property that data is fed sequentially from an input data source to a final output. For that reason, data flows are described as directed acyclic graphs (DAGs). It is important, therefore, to realize that if an algorithm, analysis, or other non-trivial computation can be expressed as a DAG, then it can be parallelized on Hadoop. Unfortunately, it also quickly becomes apparent that many algorithms aren't easily converted into DAGs, and are therefore unsuitable for this type of parallelism. Algo- rithms that cannot be described as a directed data flow include those that maintain or update a single data structure throughout the course of computation (requiring some shared memory) or computations that are dependent on the results of another at intermediate steps (requiring intermediate interprocess communication). Algorithms that introduce cycles, particularly iterative algorithms that are not bounded by a finite number of cycles, are also not easily described as DAGs. There are tools and techniques that address requirements for cyclicity, shared mem- ory, or interprocess communication in both MapReduce and Spark, but to make use of these tools, algorithms must be rewritten to a distributed form. Rather than rewrite algorithms, a less technical but equally effective approach is usually employed: design a data flow that decomposes the input domain into a smaller output that fits into the memory of a single machine, run the sequential algorithm on that output, then vali- date that analysis across the cluster with another data flow (e.g., to compute error).	
	b)	Explain with example Projection in context of Pig ANS: Pig, like Hive, is an abstraction of MapReduce, allowing users to express their data processing and analysis operations in a higher-level language that then compiles into a MapReduce job. Pig is now a top-level Apache Project that includes two main platform components: • Pig Latin, a procedural scripting language used to express data flows. • The Pig execution environment to run Pig Latin programs, which can be run in local or MapReduce mode and includes the Grunt commandline interface. Pig Latin is procedural in nature and designed to enable programmers to easily implement a series of data operations and transformations that are applied to datasets to form a data pipeline. While Hive is great for use cases that translate well to SQL-based scripts, SQL can become unwieldy when multiple complex data transformations are required. Pig Latin is ideal for implementing these types of multistage data flows, particularly in cases where we need to aggregate data from multiple sources and perform subsequent transformations at each stage of the data processing flow. Pig Latin scripts start with data, apply transformations to the data	

		until the script describes the desired results, and execute the entire data processing flow as an optimized MapReduce job. Additionally, Pig supports the ability to integrate custom code with user-defined functions (UDFs) that can be written in Java, Python, or JavaScript, among other supported languages. Pig thus enables us to perform near arbitrary transformations and ad hoc analysis on our big data using comparatively simple constructs. It is important to remember the earlier point that Pig, like Hive, ultimately compiles into MapReduce and cannot transcend the limitations of Hadoop's batch-processing approach. However, Pig does provide us with powerful tools to easily and succinctly write complex data processing flows, with the fine-grained controls that we need to build real business applications on Hadoop.	
	c)	Explain the functionality of the Design pattern. ANS: Design patterns are a special term in software design: generic, reusable solutions for a particular programming challenge. We can explore functional design patterns for solving parallel computations in both MapReduce and Spark. These patterns show a generic strategy and principle that can be used in more complex or domain-specific roles. Donald Miner and Adam Shook explore 23 design patterns for common MapReduce jobs. They loosely categorize them as follows: Summarization: Provide a summary view of a large dataset in terms of aggregations, grouping, statistical measures, indexing, or other high-level views of the data. Filtering: Create subsets or samples of the data based on a fixed set of criteria, without modifying the original data in any way. 135 Data Organization: Reorganize records into a meaningful pattern in a way that doesn't necessarily imply grouping. This task is useful as a first step to further computations. Joins: Collect related data from disparate sources into a unified whole. Metapatterns: Implement job chaining and job merging for complex or optimized computations. These are patterns associated with other patterns. Input and output: Transform data from one input source to a different output source using data manipulation patterns, either internal to HDFS or from external sources	