UNIT–1 : Introduction to Data Analytics & Data Analytics Lifecycle (PYQ Solution)

AKTU · Data Analytics (BADS601) · Solutions by **Chirag Jain**, Ycotes

SECTION A Answers (≈70 Words Each)

Q1. Differentiate between Predictive and Prescriptive Data Analytics. (Jan 2025, 2022)

Predictive analytics uses historical data, statistics and machine-learning models to forecast *what is likely to happen* (for example, next month's sales or customer churn). Prescriptive analytics goes one step further and recommends *what action should be taken* to achieve a goal, using optimisation, simulation and rules. Predictive gives a forecast; prescriptive gives a decision. Prescriptive analytics normally builds on the output of predictive models.

Q2. Define the term data lake, data base and data warehouse. (Jan 2025)

A database stores current, structured, operational data for day-to-day transactions (OLTP), e.g. a bank's account table. A data warehouse is a subject-oriented, integrated, time-variant store of cleaned historical data used for reporting and analysis (OLAP). A data lake is a large repository that keeps raw data of every kind (structured, semi-structured, unstructured) in its native format, with the schema applied only when the data is read.

Q3. Explain the concept of Outliers. (Jan 2025)

An outlier is a data point that lies far away from the other observations in a data set. It may be caused by measurement or entry errors, or it may be a genuine rare event such as fraud. Outliers distort the mean, variance and model results, so they are detected using box plots, the IQR rule (below Q1 −

1.5 IQR or above Q3 + 1.5 IQR) or Z-scores, and then removed, capped or investigated.

Q4. Differentiate between Analysis and Reporting. (2022)

Reporting organises data into summaries, tables and dashboards to show *what happened*; it is routine, scheduled and descriptive. Analysis interprets the data to find *why it happened* and what to do next, using questions, statistics and models; it is exploratory and customised. Reporting gives information; analysis gives insight and recommendations. Reporting is the input to analysis, and analysis adds value by turning facts into decisions.

Q5. What are the common open source tools for the model building phase? (2023)

In the model-building phase of the analytics life cycle, common open-source tools are R and Python (with scikit-learn, pandas, NumPy, TensorFlow, Keras) for statistics and machine learning, Weka for data mining, Octave and Apache Spark MLlib for large-scale modelling, and KNIME and RapidMiner Community for visual workflows. Jupyter Notebook is used to write and share the code, and SQL/PostgreSQL is used for in-database analytics.

SECTION B Answers (≈250 Words Each)

Q1. Explain the different categories of data analytics with examples. (Jan 2025)

Data analytics is the process of examining data to draw useful conclusions. It is divided into four categories, each answering a different question and building on the previous one.



Four categories of data analytics

1. Descriptive analytics – What happened?

- Summarises past data using counts, averages, charts and dashboards.
- Example: monthly sales report, website traffic summary, a cricketer's batting average.

2. Diagnostic analytics – Why did it happen?

- Drills down into the data to find causes using correlation, drill-down, data mining and root-cause analysis.
- Example: finding that sales fell in March because a competitor cut prices and stock-outs occurred in two cities.

3. Predictive analytics – What is likely to happen?

- Uses statistical models and machine learning (regression, classification, time series) on historical data to forecast.
- Example: predicting which customers will leave, forecasting demand, credit-risk scoring.

4. Prescriptive analytics – What should we do?

- Recommends the best action using optimisation, simulation and decision rules.
- Example: route optimisation for delivery trucks, dynamic price recommendation, suggesting the best marketing offer to each customer.

Category	Question	Typical techniques	Example
Descriptive	What happened?	Aggregation, visualisation	Sales dashboard
Diagnostic	Why?	Drill-down, correlation	Reason for a sales drop
Predictive	What will happen?	Regression, ML	Churn prediction
Prescriptive	What to do?	Optimisation, simulation	Best delivery route

As we move from descriptive to prescriptive, the **value and complexity** of the analysis both increase. Organisations normally use all four together.

Q2. Explain the Process Model and Computation Model of Big Data Platform (2022)

A **big data platform** combines storage, processing, management and analytics tools to handle very large, fast and varied data. Its working is described by a process model and a computation model.

Process model (how data flows through the platform)

- **Data acquisition:** collect data from sensors, logs, social media, databases and applications, in batch or real time.
- **Storage:** keep raw and processed data in distributed systems such as **HDFS**, NoSQL stores or a data lake.
- **Pre-processing:** clean, filter, integrate and transform the data (ETL).
- **Analysis / mining:** apply statistics, machine learning and queries to extract patterns.
- **Visualisation and decision:** present results through dashboards and reports; feed back to the business.



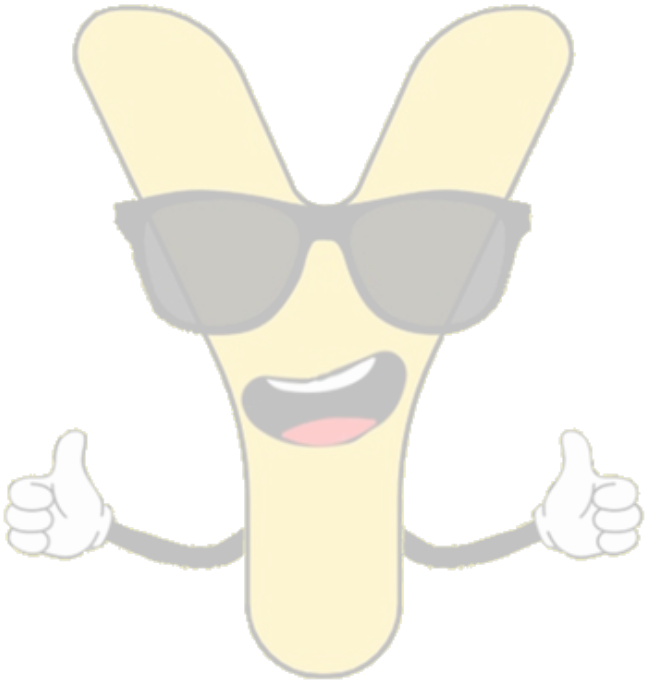
Process model of a big data platform

Computation model (how the work is executed)

- **Batch computation:** large stored data sets are processed in bulk, e.g. **MapReduce** on Hadoop. High throughput, high latency.
- **Stream computation:** data is processed as it arrives, e.g. **Storm, Spark Streaming, Flink**. Low latency.

- **Interactive / query computation:** fast ad-hoc queries using Hive, Impala or Spark SQL.
- **Graph and iterative computation:** for networks and machine-learning algorithms, e.g. Pregel, GraphX.
- Work is split into tasks that run in parallel on a cluster of commodity machines close to the data (**data locality**), with fault tolerance by replication.

In short, the process model tells *what stages* data passes through, while the computation model tells *how* each stage is computed on a distributed cluster.



SECTION C Answers (≈500 Words Each)

Q1. Differentiate between Structured data, Semi-structured data and Unstructured Data. (Jan 2025)

Data is classified by how well it follows a fixed format or schema.

Structured data is organised in rows and columns with a predefined schema, so it can be stored in relational databases and queried with SQL. Examples: bank transactions, employee tables, spreadsheets. About 10-20% of enterprise data is structured.

Semi-structured data has no rigid table schema but carries tags, keys or markers that separate fields and give a hierarchy. Examples: **JSON, XML**, HTML, e-mail headers, log files. It is stored in NoSQL or document stores and is self-describing.

Unstructured data has no predefined model or organisation. Examples: text documents, images, audio, video, social-media posts. It forms most (about 80%) of all data and needs special tools such as text mining, image processing and NLP.

Basis	Structured	Semi-structured	Unstructured
Schema	Fixed, predefined	Flexible, self-describing	None
Format	Tables (rows, columns)	JSON, XML, tags	Text, image, audio, video
Storage	RDBMS, data warehouse	NoSQL, document store	Data lake, HDFS, object store
Query	SQL	XPath, JSON query	Search, text mining, ML
Ease of analysis	Easy	Moderate	Difficult
Example	Student marks table	JSON API response	YouTube video



Data from no schema to full schema

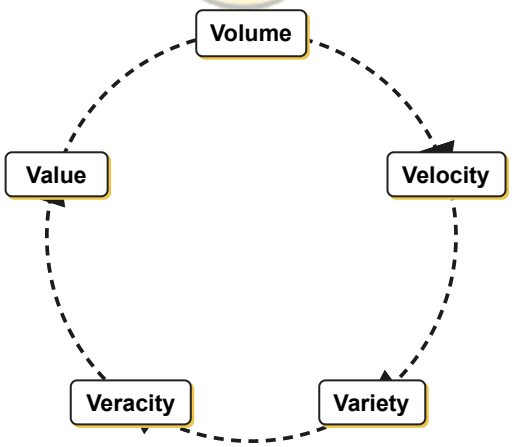
Modern analytics platforms have to handle all three types together, usually by storing raw data in a data lake and converting it to a structured form when it is needed for analysis.

Q2. Describe Big Data and its characteristics. (Jan 2025)

Big Data means data sets that are so large, fast-growing and varied that traditional database systems and tools cannot capture, store, manage or analyse them within a reasonable time. Sources include social media, sensors and IoT devices, e-commerce transactions, mobile phones and server logs. Typical examples are Facebook posts, Google searches and airline sensor data.

Characteristics (the Vs of Big Data)

1. **Volume:** the enormous size of data, measured in terabytes, petabytes and zettabytes. Example: Facebook generates several petabytes daily.
2. **Velocity:** the speed at which data is generated and must be processed, often in real time. Example: stock-market ticks, sensor streams.
3. **Variety:** the many forms of data – structured, semi-structured and unstructured (text, images, video, logs).
4. **Veracity:** the quality, accuracy and trustworthiness of data. Noise, bias and missing values make data uncertain.
5. **Value:** the useful business insight that can be extracted; data is worthless unless it leads to decisions.
6. **Variability (and Visualisation):** data flow and meaning change over time, so peaks and context shifts must be handled.



The Vs of Big Data

Importance and challenges

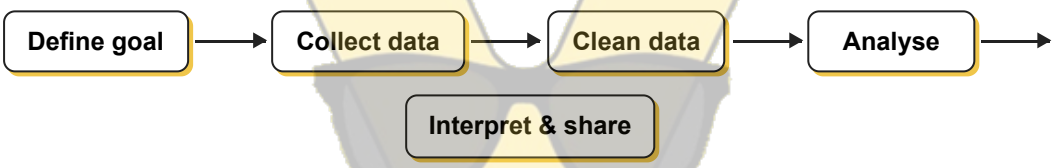
- Better decisions, personalised marketing, fraud detection, predictive maintenance and new products.
- Challenges: storage cost, processing speed, data quality, security and privacy, and shortage of skills.

Technologies: Hadoop (HDFS, MapReduce), Spark, NoSQL databases, Kafka, Hive and Pig.

Thus Big Data is defined not only by size but by the combination of volume, velocity and variety, and its real purpose is to create **value** from data.

Q3. Discusses the steps involved in Data Analysis Process. (2022)

The **data analysis process** is a systematic sequence of steps that converts raw data into useful conclusions and decisions.



Steps of the data analysis process

1. Define the question / objective. State clearly the business problem, what is to be measured and what decision will use the result. Example: "Why did sales fall in the last quarter?"

2. Data collection. Gather relevant data from primary sources (surveys, experiments, sensors) and secondary sources (databases, files, web, APIs). Decide sample size and period.

3. Data cleaning (preparation). Raw data is rarely perfect. This step handles:

- missing values (delete, fill with mean/median, predict),
- duplicates and inconsistent formats,
- outliers and errors,
- transformation, normalisation and integration of several sources.
  This is usually the most time-consuming step.

4. Data exploration. Use summary statistics, charts and correlations to understand distributions, trends and relationships; form hypotheses.

5. Data analysis / modelling. Apply suitable techniques – descriptive statistics, regression, classification, clustering, time series – and build and validate a model.

6. Interpretation. Explain what the results mean, check them against the original objective, and state limitations and assumptions.

7. Visualisation and reporting. Present findings with charts, dashboards and a clear story for decision makers.

8. Action and feedback. Implement the decision, monitor the outcome and repeat the cycle if needed.

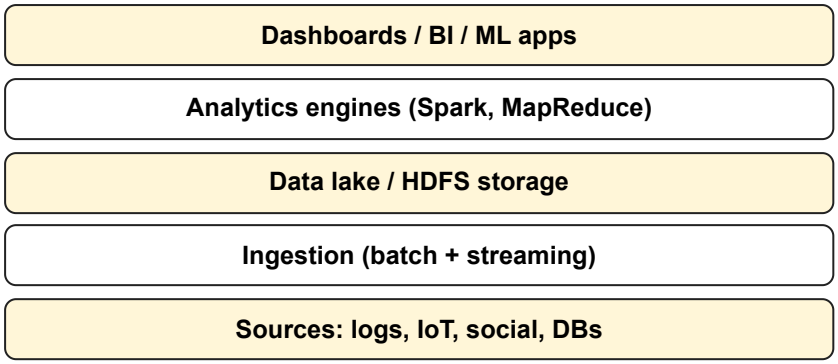
Step	Output
Define goal	Problem statement
Collect	Raw data set
Clean	Reliable data set
Explore	Patterns, hypotheses
Analyse	Model / results
Interpret & report	Insights, decisions

The process is **iterative**: findings often send the analyst back to collect more data or clean it further.

Q4. Compare and contrast traditional analytics structure to modern analytics architecture. (2023, 2022)

Traditional analytics architecture was built for small, structured, mostly historical data. Data from operational systems is extracted, transformed and loaded (ETL) into a central **enterprise data warehouse (EDW)**. Analysts then pull a small extract into local tools (Excel, SAS, SPSS) and make reports. It is a centralised, schema-first design.

Modern analytics architecture is designed for big, fast and varied data. Raw data from many sources (logs, social media, sensors, apps) lands in a **data lake / Hadoop (HDFS)** cluster, is processed in parallel (MapReduce, Spark) close to where it is stored, and is analysed in place using in-database and in-memory analytics, machine learning and real-time streaming. A data warehouse may still exist as one component.



Modern analytics architecture

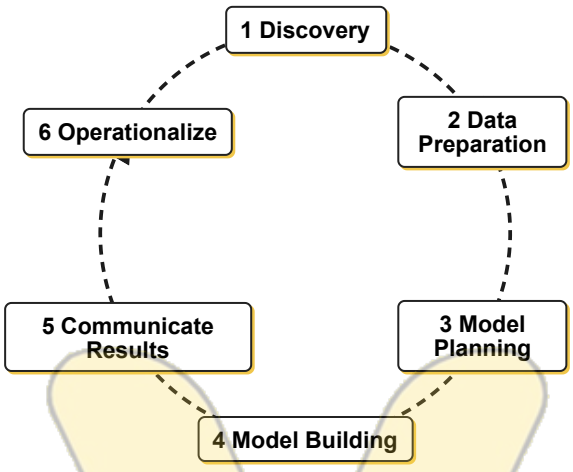
Basis	Traditional	Modern
Data type	Structured only	Structured, semi- and unstructured
Data size	GB to a few TB	TB to PB and beyond
Storage	Central EDW (RDBMS)	Distributed data lake, HDFS, NoSQL
Schema	Schema-on-write	Schema-on-read
Processing	Data moved to the analytic tool; sampling	Processing moved to the data; full data set
Speed	Batch, periodic	Batch plus real-time streaming
Scalability	Vertical (costly big servers)	Horizontal (commodity clusters)
Cost	High licence and hardware cost	Low cost, open source
Analytics	Descriptive reports, BI	Predictive and prescriptive, machine learning
Users	IT-controlled, analysts wait	Self-service, data scientists
Flexibility	Rigid, slow to change	Flexible and agile

Similarities: both aim to turn data into decisions, both need data integration, cleaning, security and governance, and the warehouse still serves trusted reporting in modern set-ups.

Conclusion: modern architecture overcomes the limits of the traditional one in volume, variety and velocity, and is usually combined with the warehouse in a hybrid design.

Q5. Explain various phases of Data Analytics Life Cycle. (2023)

The **Data Analytics Lifecycle** is a six-phase, iterative framework used by data scientists to run an analytics project from business problem to operational deployment. Work can move back to an earlier phase when new findings appear.



Data Analytics Lifecycle

Phase 1 – Discovery. Understand the business domain and the problem. Identify stakeholders, resources (people, tools, time, data), and frame the problem. Define success criteria and formulate initial hypotheses to test.

Phase 2 – Data Preparation. Build an **analytic sandbox**, then extract, load and transform data (ETLT). Explore, clean and condition the data; identify missing values, outliers and inconsistencies. Usually the longest phase (about 50-80% of the effort).

Phase 3 – Model Planning. Explore the variables, find relationships between them, and select the key variables and suitable techniques (regression, classification, clustering, etc.). Tools: R, SQL Analysis Services, SAS/ACCESS.

Phase 4 – Model Building. Develop training and testing data sets, then build, run and tune the models. Check whether the model is accurate enough and robust. Open-source tools: R, Python, Weka, Octave.

Phase 5 – Communicate Results. Compare the outcome with the success criteria set in Phase 1, decide whether the project succeeded or failed, identify key findings and business value, and present them to stakeholders in a clear narrative.

Phase 6 – Operationalize. Deliver final reports, briefings, code and technical documents. Run a **pilot** in a production environment, monitor performance, and then roll out fully.

Phase	Main output
Discovery	Problem framing, hypotheses
Data preparation	Clean analytic data set
Model planning	Chosen variables and methods
Model building	Trained and tested model
Communicate results	Findings, business value
Operationalize	Deployed model, reports

Conclusion: the lifecycle gives a repeatable and structured path, and its iterative nature lets the team refine the work until the business goal is met.

