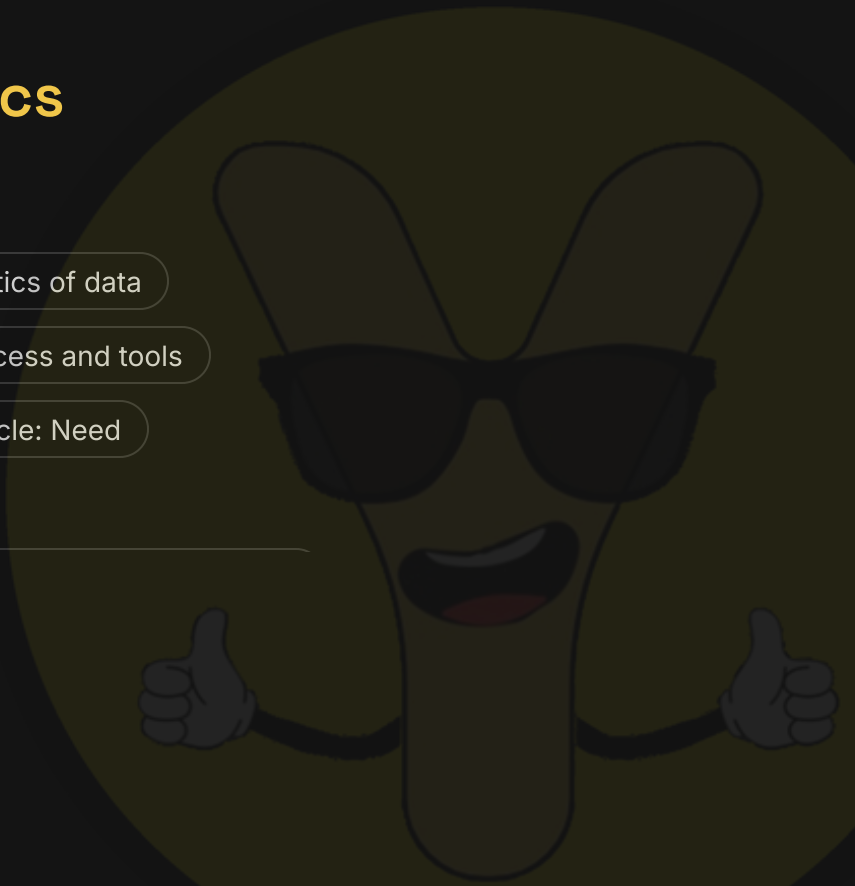
AKTU · 3RD YEAR · SEMESTER 6 · BADS601

Data Analytics

Unit 1 — Introduction to Data Analytics & Data Analytics Lifecycle

- Sources and nature of data
- Classification of data (structured, semi-structured, unstructured)
- Characteristics of data
- Introduction to Big Data platform
- Need of data analytics
- Evolution of analytic scalability
- Analytic process and tools
- Analysis vs reporting
- Modern data analytic tools
- Applications of data analytics
- Data Analytics Lifecycle: Need
- Key roles for successful analytic projects



DECODED BY

Chirag Jain

Founder, Ycotes · Gold Medalist in Artificial Intelligence, IIT Madras · taught 10,000+ students

Sources and Nature of Data

Data originates from operational, transactional, and environmental systems in continuous streams or discrete batches.

- Internal Data Sources
  - Transactional databases (ERP, CRM), enterprise audit logs, and operational history.
  - Proprietary customer transactions and enterprise financial metrics.
- External Data Sources
  - Public web scraped data, social media feeds, third-party API services, and open datasets.
  - Syndicated market research data and financial market exchange feeds.
- Sensor and Machine Data
  - Internet of Things (IoT) devices, RFID tags, smart meters, and industrial sensors.
  - Continuous real-time telemetry streams generated without human intervention.
- Nature of Data
  - Static vs. Dynamic: Historical archived records versus real-time streaming data.
  - Continuous vs. Discrete: Quantitative continuous metrics vs. distinct categorical attributes.

Classification of Data

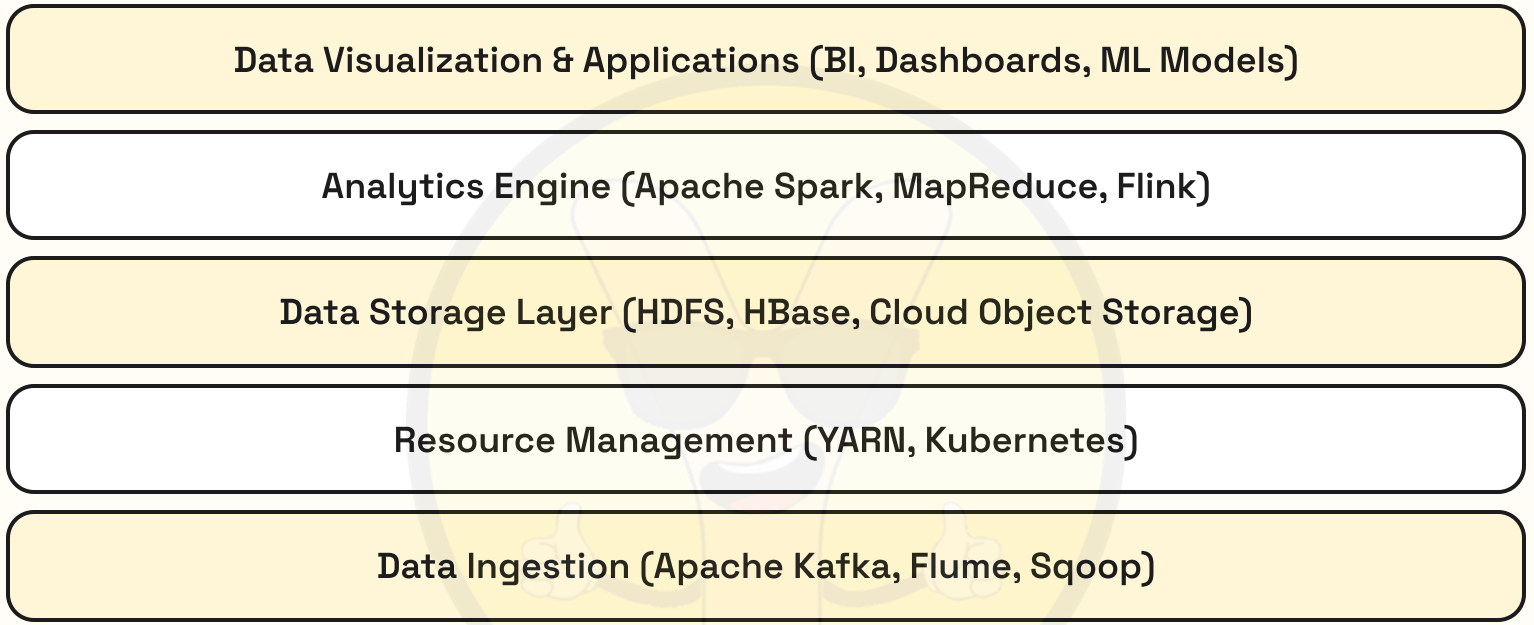
Criteria	Structured Data	Semi-Structured Data	Unstructured Data
Format	Rigid tabular schema (rows & columns)	Self-describing schema (tags/keys)	No predefined data model or schema
Storage	Relational Databases (RDBMS), CSV	JSON, XML, NoSQL (MongoDB)	Object Storage, HDFS, Data Lakes
Querying	Standard SQL queries	Specialized queries (XPath, JSONPath)	NLP, Computer Vision, ML models
Flexibility	Low flexibility; schema-on-write	Moderate flexibility; hybrid schema	High flexibility; schema-on-read
Examples	SQL tables, Excel spreadsheets	Web API responses, server log files	Video, audio, PDF documents, raw text

Characteristics of Data: The 5 V's

Modern enterprise data is defined by five key dimensions that determine processing requirements.

- Volume
  - Refers to the scale of data generated, ranging from Terabytes to Zettabytes.
  - Requires distributed file systems like HDFS for scalable storage.
- Velocity
  - Speed at which data is accumulated and must be processed in real time.
  - Examples include stock trading ticks and streaming sensor telemetry.
- Variety
  - Heterogeneity of data formats, spanning structured, semi-structured, and unstructured types.
  - Demands multi-model processing architectures.
- Veracity
  - Trustworthiness, quality, and precision of collected raw data.
  - Requires rigorous data cleansing, noise filtering, and validation.
- Value
  - Actionable insights extracted from raw data to drive business decisions.
  - The primary goal of applying analytics algorithms to enterprise data.

Big Data Platform Architecture



Layered Architecture of a Modern Big Data Processing Platform

Need of Data Analytics

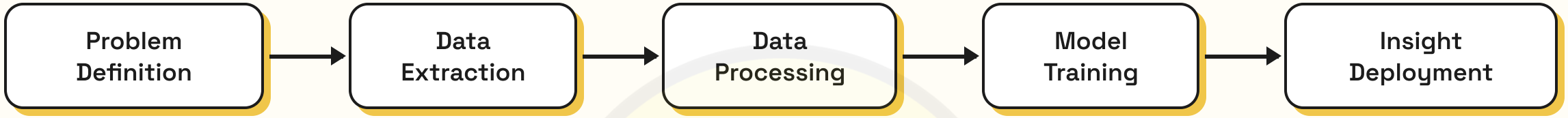
Organizations leverage data analytics to transform raw operational data into competitive strategy.

- Data-Driven Decision Making
  - Replaces intuition with evidence-based quantitative insights.
  - Minimizes financial risk and optimizes enterprise workflows.
- Predictive Capabilities
  - Identifies trends and patterns to forecast future market shifts.
  - Enables proactive maintenance and churn risk reduction.
- Operational Efficiency
  - Pinpoints bottlenecks in supply chains and business processes.
  - Automates repetitive analytical tasks to reduce overhead.
- Customer Personalization
  - Analyzes user behavior to deliver targeted product recommendations.
  - Enhances customer retention and lifetime business value.

Evolution of Analytic Scalability

Analytical Era	Architecture	Data Handling	Analytical Focus
Traditional Analytics	Single desktop server, RDBMS	Small datasets (Gigabytes), batch mode	Descriptive reporting, historical SQL queries
Enterprise Data Warehouse	Centralized MPPs, appliance clusters	Moderate datasets (Terabytes), structured	Business Intelligence (BI), OLAP cubes
In-Database Analytics	Analytic functions integrated into DBMS	Large enterprise datasets, schema-on-write	Scoring models directly inside database engine
Distributed Big Analytics	Hadoop/Spark clusters, cloud computing	Massive datasets (Petabytes+), all types	Real-time streaming, Deep Learning, Predictive ML

The Core Analytic Process



Sequential flow of the end-to-end analytical process

Comparison: Data Reporting vs. Data Analysis

Parameter	Data Reporting	Data Analysis
Primary Goal	Translates raw data into informative summaries	Transforms data into actionable business insights
Core Question	What happened? (Historical factual state)	Why did it happen? What will happen next?
Approach	Pulls metrics, formats tables, builds dashboards	Explores patterns, tests hypotheses, builds models
Human Involvement	High automation, standardized static templates	Requires critical thinking, domain knowledge, ML
Data Used	Structured transactional data	Structured, semi-structured, and unstructured data
Output	Static reports, charts, monthly scorecards	Recommendations, predictive algorithms, decisions

Modern Data Analytic Tools

Modern analytics relies on a rich ecosystem of specialized tools across storage, compute, and visualization.

- Data Processing & Compute
  - Apache Spark: Memory-centric distributed computing framework.
  - Apache Hadoop: Distributed storage (HDFS) and MapReduce execution.
- Statistical & ML Programming
  - Python: Rich ecosystem including Pandas, NumPy, and Scikit-Learn.
  - R: Comprehensive statistical computing and graphics software.
- Business Intelligence & Visualization
  - Tableau & PowerBI: Interactive dashboarding and enterprise reporting.
  - Looker: Data exploration engine with semantic data modeling.
- Data Warehousing & Querying
  - Snowflake & Google BigQuery: Serverless cloud data warehouses.
  - Apache Hive & Trino: Distributed SQL query engines for data lakes.

Applications of Data Analytics

Data analytics drives innovation and process optimization across major industry verticals.

- Healthcare & Pharmaceuticals
  - Predictive diagnostic models and early disease outbreak detection.
  - Personalized medicine based on genomic sequence analysis.
- Finance & Banking
  - Real-time credit card fraud detection algorithms using anomaly scoring.
  - Automated algorithmic trading and portfolio risk assessment.
- Retail & E-Commerce
  - Market basket analysis using association rule mining for recommendations.
  - Dynamic pricing engines and customer churn forecasting.
- Manufacturing & Supply Chain
  - Predictive equipment maintenance using IoT sensor telemetry.
  - Supply chain route optimization to minimize transit latency.

Data Analytics Lifecycle Overview

DATA ANALYTICS LIFECYCLE

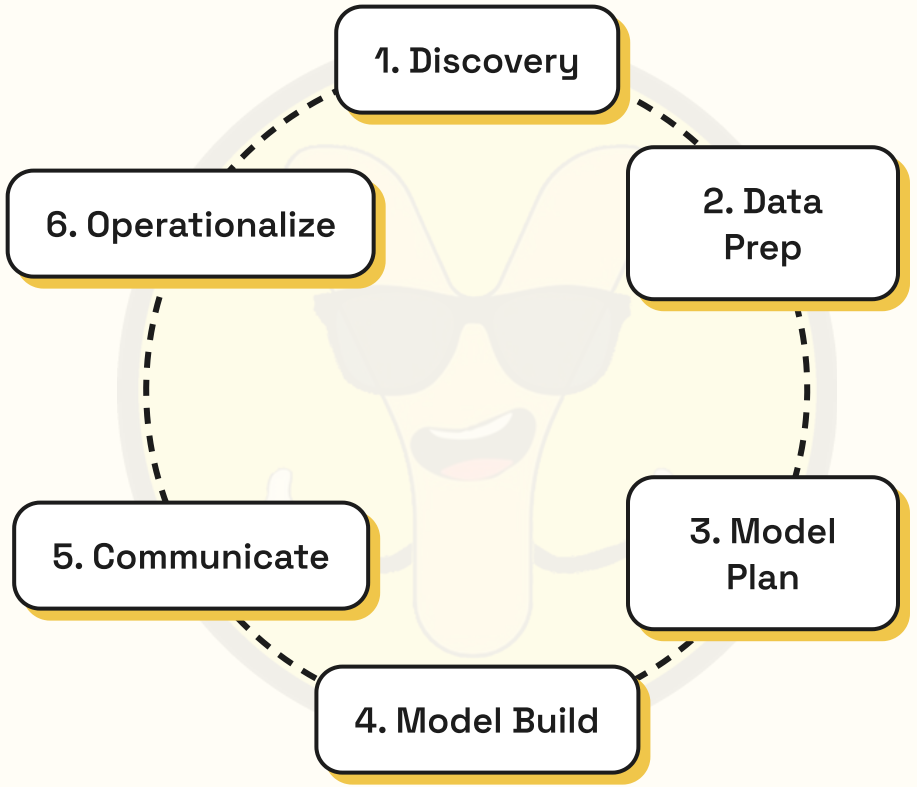
A structured, iterative methodology that defines the best practices and step-by-step process for executing analytics projects from initial business understanding to deployment.

- Ensures rigorous scientific methodology rather than ad-hoc data exploration.
- Provides clear milestones and repeatable frameworks for multidisciplinary teams.
- Promotes continuous validation, feedback loops, and operational integration.

Key Roles in Analytic Projects

Role	Primary Responsibilities	Key Skill Set
Business User / Sponsor	Defines business problem, vision, and project budget	Domain expertise, strategic planning, KPI definition
Project Manager	Tracks project execution, deadlines, and resources	Agile methodologies, communication, risk mitigation
Data Engineer	Builds data pipelines, ingestion pipelines, architectures	SQL, Spark, ETL tools, database systems, HDFS
Data Scientist	Develops predictive models, algorithms, and statistical tests	Python, R, Machine Learning, Linear Algebra, Statistics
Data Analyst	Performs exploratory analysis and designs dashboards	SQL, Excel, Tableau, descriptive statistics
Database Administrator	Configures database access, security, and performance	DB tuning, access control, backup & recovery

Phases of Data Analytics Lifecycle



The six iterative phases of the Data Analytics Lifecycle

Lifecycle Phase 1 & Phase 2

Discovery and Data Preparation set the foundational business context and clean data pipeline.

- Phase 1: Discovery
  - Learn business domain, evaluate available resources, and frame business problem.
  - Formulate hypotheses to test and define success metrics.
  - Assess data sources, budget, hardware, and analytical software needs.
- Phase 2: Data Preparation
  - Extract, transform, and load (ETL) raw data into an analytical sandbox.
  - Perform data cleaning: handle missing values, outliers, and corrupt records.
  - Conduct Exploratory Data Analysis (EDA) to understand distributions and features.

Lifecycle Phase 3 & Phase 4

Model Planning and Model Building form the core phase where mathematical models are designed and trained.

- Phase 3: Model Planning
  - Determine techniques, algorithms, and workflow methods (e.g., classification, regression).
  - Select candidate variables and feature subsets relevant to target outcome.
  - Choose software tools and set up training/testing data splits.
- Phase 4: Model Building
  - Develop datasets for testing, training, and validation of candidate models.
  - Execute algorithms using tools like Python, R, or Spark MLlib.
  - Assess model performance using metrics like Accuracy, Precision, Recall, and R^2.

Lifecycle Phase 5 & Phase 6

Communicating Results and Operationalizing bring analytical insights into real-world business production.

- Phase 5: Communicate Results
  - Evaluate project outputs against original business criteria defined in Phase 1.
  - Identify key findings, quantify business impact, and document limitations.
  - Present insights tailored to executive stakeholders via visual dashboards.
- Phase 6: Operationalization
  - Deliver final analytical reports, code, and technical documentation.
  - Deploy the model into production environments for live decision-making.
  - Implement continuous monitoring to track model drift and performance over time.

Important Exam Questions (Unit 1)

- **Q1** Differentiate between Structured, Semi-Structured, and Unstructured data with appropriate examples and storage solutions.
- **Q2** Explain the 5 V's of Big Data and discuss how Big Data platforms handle velocity and volume challenges.
- **Q3** Draw and explain the 6 phases of the Data Analytics Lifecycle in detail with suitable industry examples.
- **Q4** Compare Data Reporting and Data Analysis across key parameters like goal, question focus, and automation.
- **Q5** Discuss the evolution of analytic scalability from traditional RDBMS to modern distributed in-database analytics.
- **Q6** Explain the key roles and responsibilities required for executing a successful enterprise data analytics project.
- **Q7** Describe the data preparation phase of the analytics lifecycle, highlighting the importance of data cleaning and EDA.

Unit 1 done. You are exam-ready.

Revise the Exam corner once more, then open Unit 2 on theycotes.com. Every AKTU subject, unit by unit, free. Notes by Chirag Jain, Ycotes.

