



Course Name

BUSINESS ANALYTICS

Course Code

BAD714B

Course Coordinator

Contact Hours

40

Mrs. MADHU NAGARAJ
Assistant Professor
Dept of CSEData Science
ATMECE, Mysuru

CIE Marks

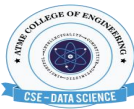
50

SEE Marks

50



A T M E
College of Engineering



Descriptive Analytics I -Nature of Data, Big Data, and Statistical Modeling Module 2

Mrs. Madhu Nagaraj
Assistant Professor
Dept of CSE-Data Science
ATMECE

THE NATURE OF DATA IN ANALYTICS

- Data: a collection of Raw facts & figures processed into meaningful insights.
- usually obtained as the result of experiences, observations, or experiments
- Data may consist of numbers, words, images,...
- Data is the lowest level of abstraction (from which information and knowledge are derived)
- Data can be of different types
 - Structured: Organized rows/columns (e.g., sales transactions).
 - Semi-structured: Tagged/marked format (e.g., XML, JSON).
 - Unstructured: Free-form (e.g., social media posts, videos).
- 4Vs: Volume (big size), Variety (formats), Velocity (speed), Veracity (trust).
- Example: Social media sentiment analysis from tweets & posts.

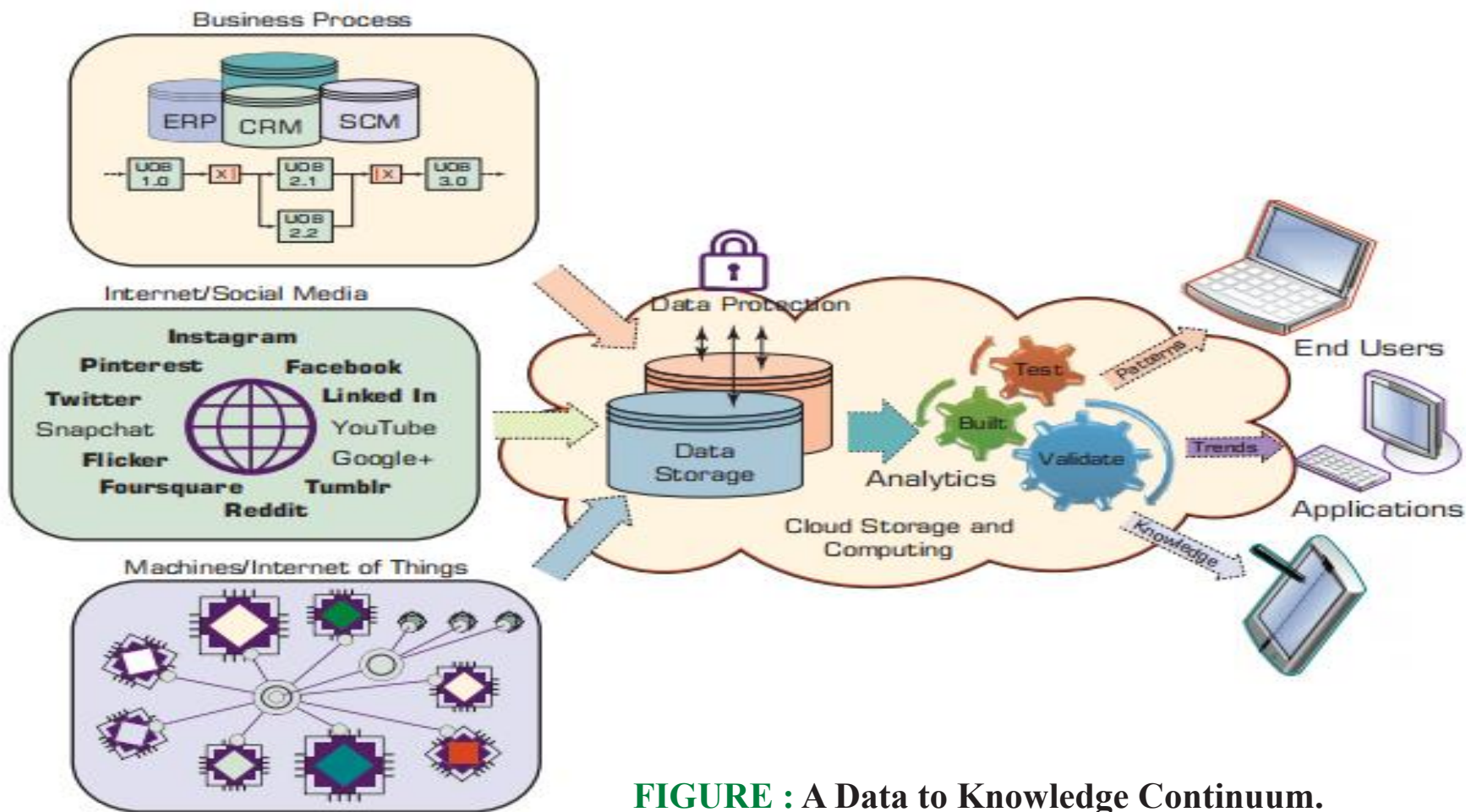


FIGURE : A Data to Knowledge Continuum.

Data Sources

- Business Process: Data comes from systems like ERP (Enterprise Resource Planning), CRM (Customer Relationship Management), and SCM (Supply Chain Management).
- Internet/Social Media: Platforms like Facebook, Twitter, YouTube, Instagram, etc., generate massive amounts of unstructured data.
- Machines/Internet of Things (IoT): Devices, sensors, and machines continuously generate operational and usage data.

Data Storage and Protection

- All collected data is stored in centralized or cloud-based storage systems.
- Data protection ensures privacy, security, and compliance with regulations.

Analytics (Processing Layer)

- Data goes through stages:
 - Built: Developing models and analytical systems.
 - Test: Ensuring accuracy and reliability.
 - Validate: Checking and refining results for correctness.
- Cloud storage and computing provide scalability and power to handle large volumes of data.

Output (Applications & End Users)

- Processed data insights are delivered through applications (dashboards, reports, mobile apps).
- End users (businesses, customers, organizations) use this information for decision-making, trend analysis, and knowledge discovery.



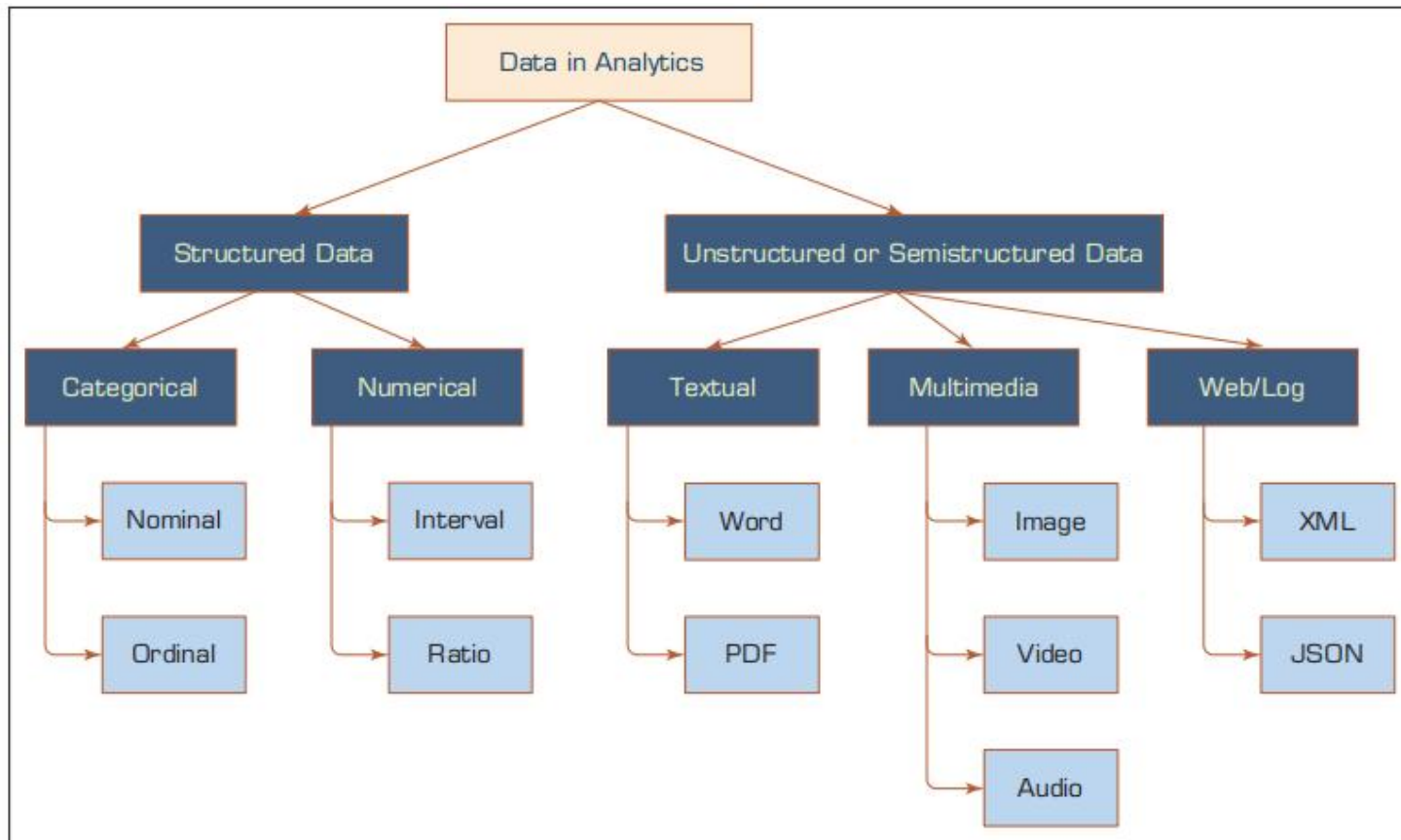
The diagram illustrates how data flows from multiple sources (business, social media, IoT) into storage and analytics systems, gets processed securely, and finally provides actionable insights to end users through applications.

Following are some of the characteristics (metrics) that define the readiness level of data for an analytics study (Delen, 2020; Kock, McQueen, & Corner, 1997).

- Data Source Reliability – Trustworthiness of the original source; reduces risks of misrepresentation during transfers.
- Data Accuracy – Correctness and alignment with intended meaning (e.g., recorded details match what was provided).
- Data Accessibility – Ease of retrieval, especially with diverse storage systems (databases, data lakes, Hadoop, etc.).
- Data Security & Privacy – Controlled access, legal compliance (e.g., HIPAA), and protection against unauthorized use.
- Data Richness – Inclusion of sufficient variables and dimensions for meaningful analytics.
- Data Consistency – Proper integration of data from multiple sources to avoid errors (e.g., mixing patient records).

- Data Currency/Timeliness – Data should be up-to-date and recorded close to the event to ensure reliability.
- Data Granularity – Level of detail must be sufficient for analysis (e.g., medical results with decimals, detailed demographics).
- Data Validity – Values must fall within defined acceptable ranges (e.g., gender = male/female/other).
- Data Relevancy – Only variables pertinent to the study should be included to avoid misleading results.

Simple Taxonomy of Data



Simple Taxonomy of Data

- **Categorical Data**– Labels used to divide variables into groups (e.g., race, sex, age group). Also called discrete data; values are finite and symbolic, not fractional.
 - **Nominal**: A type of categorical data where codes are labels only, without order (e.g., marital status: single, married, divorced). Can be binomial (yes/no) or multinomial (blue/green/brown).
 - **Ordinal**: Categorical data with rank order (e.g., credit score: low, medium, high;). Order matters, but differences between values are not precisely measurable. Eg: customer satisfaction levels).
- **Numerical Data**: Variables with measurable numeric values (e.g., age, income, distance). Can be integer (whole numbers) or real (fractions). Also called continuous data.

Interval Data – Numeric scale with meaningful intervals but no true zero (e.g., temperature in Celsius or Fahrenheit). Zero does not indicate absence of the quantity.

Ratio Data – Numeric scale with a true, nonarbitrary zero (e.g., length, mass, time, Kelvin temperature). Ratios are meaningful, allowing full mathematical operations.

Other Data Types – Text, spatial, images, video, and audio must be converted into categorical or numeric form for analytics.

Static vs. Dynamic Data – Static data does not change; dynamic (temporal/time series) data changes over time.

Model Requirements – Some machine learning algorithms (e.g., neural networks, SVMs, logistic regression) require numeric data, so categorical variables must be converted into numeric form (e.g., one-hot encoding).

Conversion Challenges – Converting categorical variables with many unique values can create too many variables, so care is needed.

Algorithms Needing Categorical Data – Some (e.g., ID3 decision trees, rough sets) require categorical inputs. Earlier versions required manual discretization of numeric data, but modern tools can auto-convert.

Tool Support – Business analytics tools increasingly automate data transformation, making it easier to prepare data for different predictive models.

Data Type	Definition	Examples	Order?	True Zero?	Key Notes
Categorical (Discrete)	Groups/labels dividing variables into classes	Race, sex, age group	No	No	Values are finite, symbolic; not fractional
Nominal	Labels without order	Marital status (single/married/divorced), colors, yes/no	No	No	Used for naming only; binomial or multinomial
Ordinal	Categories with meaningful rank order	Credit score (low/medium/high), education level	Yes	No	Order matters, but intervals not measurable
Numeric (Continuous)	Measurable numbers (integer or real)	Age, income, distance, temperature	Yes	Depends on scale	Allows fractional values
Interval	Numeric scale with equal intervals, no true zero	Celsius/Fahrenheit temperature	Yes	No	Zero is arbitrary; differences make sense
Ratio	Numeric scale with meaningful zero	Length, mass, time, Kelvin temperature	Yes	Yes	Ratios meaningful; full mathematical use

Art and Science of Data Preprocessing

- Raw data is often incomplete, inconsistent, or noisy and needs preprocessing.
 - The next image shows the main steps in the data preprocessing endeavor.
1. In the **first phase** of data preprocessing, the relevant data is collected from the identified sources, and are merged together. Such an integration or merging process is called **data blending**
 2. **Data cleaning** (or data cleansing) is the process of identifying and correcting (or removing) errors and inconsistencies in data to improve its quality.

Key Tasks in Data Cleaning:

- Handling Missing Values
- Removing Duplicates
- Correcting Errors

3. Data Transformation

- Data transformation involves converting data into a suitable format or structure for analysis.
- This is where data is cleaned up further and reshaped to make machine learning or statistical analysis more effective.

Common Transformation Techniques:

- Normalization/Scaling
- Aggregation
- Binning/Discretization
- Date/Time Features

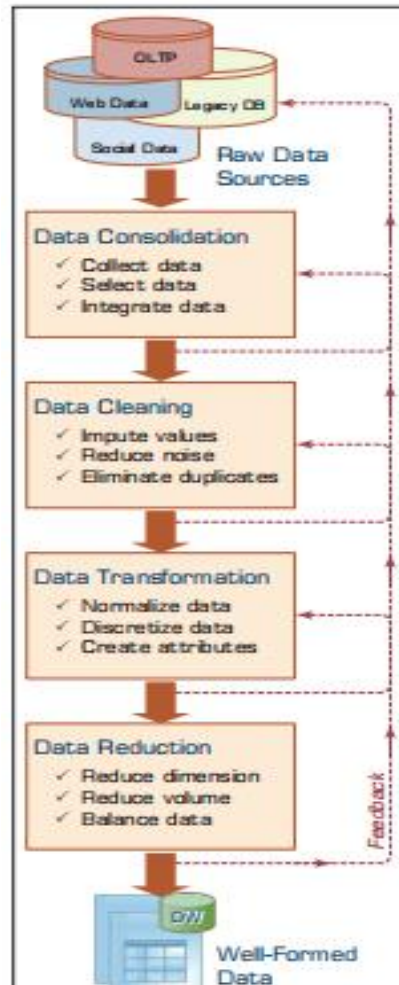
4. Data Reduction

Data reduction aims to reduce the size of the data while preserving as much useful information as possible. This improves processing time and helps avoid overfitting.

Common Data Reduction Techniques:

- Sampling
- Dimensionality Reduction
- Numerosity Reduction

Step	Purpose	Key Techniques
Data Consolidation	Combine data from multiple sources	Schema matching, entity resolution
Data Cleaning	Improve data quality by fixing errors and omissions	Imputation, deduplication, correction
Data Transformation	Convert data into suitable formats for analysis	Scaling, encoding, aggregation
Data Reduction	Reduce size while preserving insights	Sampling, PCA, feature selection



DEFINITION OF BIG DATA

- Big Data means different things to people with different backgrounds and interests.
- Traditionally, the term Big Data has been used to describe the massive volumes of data analyzed by huge organizations like Google or research science projects at NASA.
- Big Data is typically defined by three “V”s: **volume, variety, velocity**.
- In addition to these three, we see some of the leading Big Data solution providers adding other “V”s, such as **veracity** (IBM), **variability** (SAS), and **value** proposition.

The V's of Big Data :

- **Volume** – The massive amount of data generated every second.
- Example: Facebook generates over 4 petabytes of data per day.

- **Velocity** – The speed at which data is created, collected, and processed. **Example:** Stock market data or sensor data from IoT devices is generated in real time.
- **Variety** – Different types of data (structured, semi-structured, unstructured). **Example:** Tweets (text), YouTube videos (video), and Excel files (structured).
- **Veracity** – The quality, accuracy, and trustworthiness of data. **Example:** User reviews on e-commerce sites may contain fake or misleading information.
- **Value** – The usefulness of data in generating insights and decisions. **Example:** Netflix uses viewing data to recommend shows that match user preferences.
- **Variability** – Inconsistency in the meaning or format of data. **Example:** A word like “Apple” could mean the fruit or the tech company, depending on context.



FUNDAMENTALS OF BIG DATA ANALYTICS

Big Data by itself, regardless of the size, type, or speed, is worthless unless business users do something with it that delivers value to their organizations.

In the case with any other large IT investment, the success in Big Data analytics depends on a number of factors.

The following are the Keys to Success with Big Data Analytics

1. A clear business need

Big Data initiatives should start with a specific problem to solve, not just collecting data randomly.

Example: An e-commerce company wants to reduce shopping cart abandonment rates. Analytics can help identify at which stage customers drop off and why.

2. Strong, committed sponsorship

Top management must support and fund analytics projects.

Example: A hospital's leadership invests in predictive analytics to improve patient care and reduce readmissions.

3. Alignment between the business and IT strategy

Big Data projects succeed when business goals and IT capabilities are aligned.

Example: A bank wants to improve fraud detection. The business team sets the goal, while IT provides real-time monitoring systems to detect unusual transactions.



4. A fact-based decision-making culture

Organizations must trust data-driven insights over gut feelings.

Example: Netflix uses viewing data to recommend shows instead of relying only on human intuition about what audiences may like.

5. A strong data infrastructure

Requires reliable data storage, processing, and management systems.

Example: Amazon uses AWS cloud infrastructure to process huge volumes of customer and transaction data worldwide.



6. The right analytics tools

Choosing appropriate tools and technologies for analysis.

Example: Using Hadoop and Spark for big data processing, or Tableau/Power BI for visualization.

7. Personnel with advanced analytical skills

Skilled data scientists, engineers, and analysts are crucial.

Example: Google hires top AI researchers to develop models like Google Translate and recommendation engines.

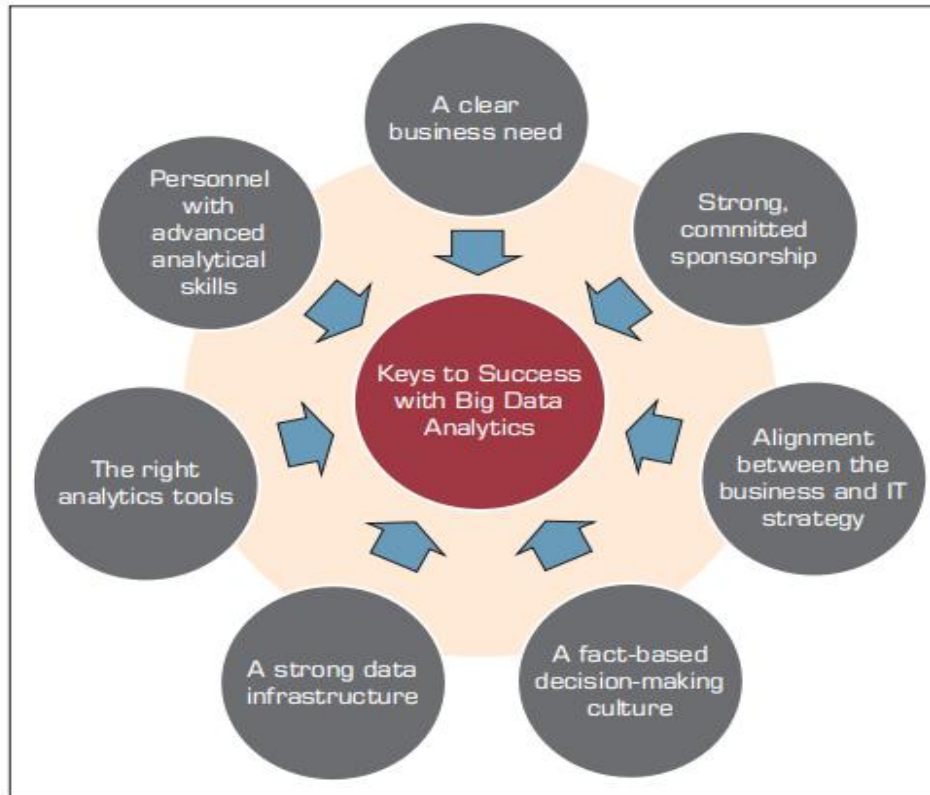


Figure 3.8- graphical depiction of the most critical success factors.



Business Problems Addressed by Big Data Analytics

The top business problems addressed by Big Data overall are process efficiency and cost reduction. Here is a partial list of problems that can be addressed using Big Data analytics:

- Process efficiency and cost reduction
- Brand management
- Revenue maximization, cross-selling, and up-selling
- Enhanced customer experience
- Churn identification, customer recruiting
- Improved customer service
- Identifying new products and market opportunities
- Risk management
- Regulatory compliance
- Enhanced security capabilities



Business Problems Addressed by Big Data Analytics

The top business problems addressed by Big Data overall are process efficiency and cost reduction. Here is a partial list of problems that can be addressed using Big Data analytics:

- Process efficiency and cost reduction
- Brand management
- Revenue maximization, cross-selling, and up-selling
- Enhanced customer experience
- Churn identification, customer recruiting
- Improved customer service
- Identifying new products and market opportunities
- Risk management
- Regulatory compliance
- Enhanced security capabilities

BIG DATA TECHNOLOGIES

There are a number of technologies for processing and analyzing Big Data. The three Big Data technologies that stand out that most believe will transform the business analytics and data management markets are **Hadoop, MapReduce, and NoSQL.**

Hadoop

- Hadoop is an open-source framework for processing, storing, and analyzing massive amounts of distributed, unstructured data.
- Originally created by Doug Cutting at Yahoo!, inspired by MapReduce, a user-defined function developed by Google in the early 2000s for indexing the Web.
- Hadoop is designed to process large volumes of data (Big Data) across many machines without relying on a single machine.
- It is built to be scalable, fault-tolerant and cost-effective.
- Instead of relying on expensive high-end hardware, Hadoop works by connecting many inexpensive computers (called nodes) in a cluster.

How does Hadoop Works

Data Sources- Structured, semi-structured, and unstructured data (e.g., logs, social media, internal data).

HDFS (Hadoop Distributed File System)

- Splits data into smaller parts and distributes across multiple nodes.
- Replicates parts for fault tolerance (if one node fails, another has a copy).
- NameNode: Keeps track of where data resides and manages node communication.

MapReduce Framework

Map Phase:

- Client submits a job to the JobTracker.
- JobTracker identifies where data is located (via NameNode) and distributes tasks to relevant nodes.
- Each node processes its data locally in parallel.

Reduce Phase:

- Aggregates results from the Map phase across nodes.
- Produces the final output, stored on another node for client access.

MapReduce

- MapReduce is a technique popularized by Google that distributes the processing of very large multistructured data files across a large cluster of machines.
- High performance is achieved by breaking the processing into small units of work that can be run in parallel across the hundreds of nodes in the cluster.
- MapReduce is a programming model, not a programming languageThe easiest way to describe how MapReduce works is through the use of an example—see the Colored Square Counter in Figure next.

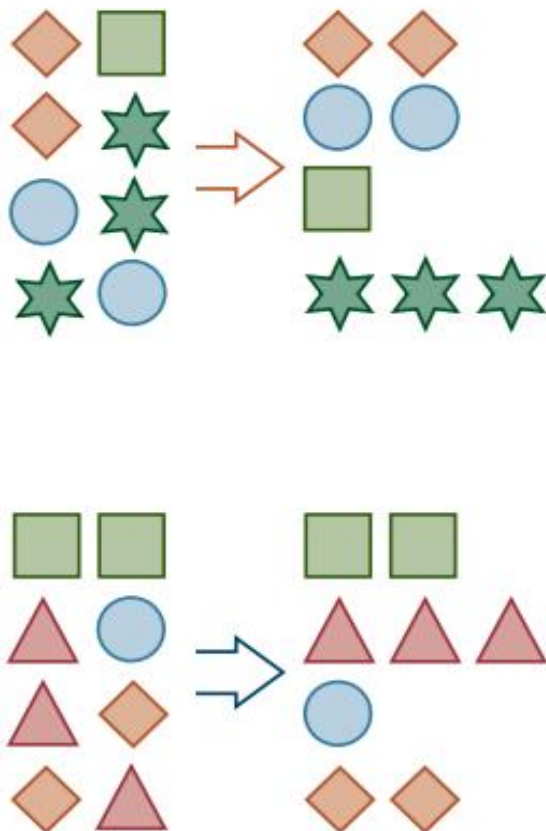
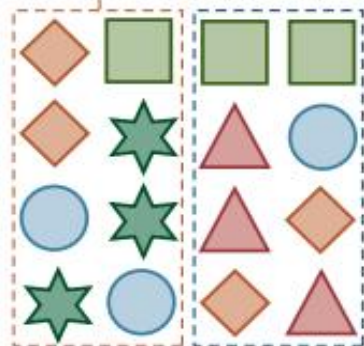
MapReduce Example: Counting Colored Squares

Input

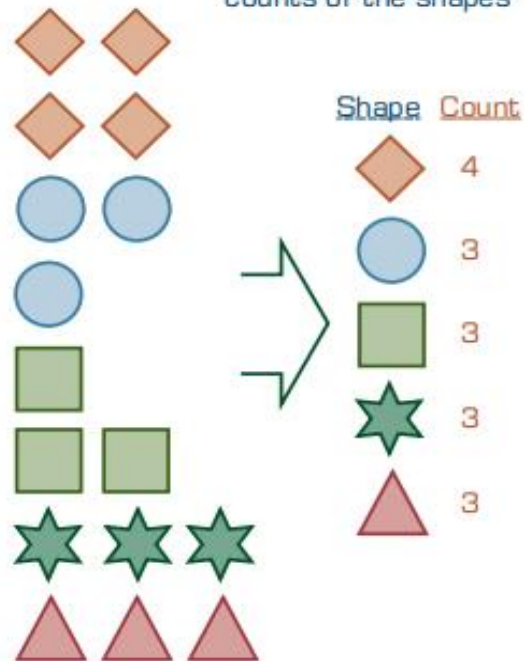
A set of colored squares.

Goal: count how many squares of each color exist.

The Goal: Determining the frequency counts of the shapes



Results: Frequency counts of the shapes



Raw Data

Map Function

Reduce Function



Splitting

Input file is divided into multiple splits (pieces).

Each split is processed independently and in parallel.

Map Phase

Each map program groups squares by color within its split.

Output: key–value pairs (color, count).

Shuffle & Sort

Intermediate results from all map tasks are merged.

Groups all values belonging to the same color together.

Optimizations: combiners (local aggregation) and custom shuffle/sort functions can reduce overhead.

Reduce Phase

Reduce program sums up counts for each color.

Final output: (color, total count).

Why use Map Reduce?

- MapReduce aids organizations in processing and analyzing large volumes of multistructured data.
- Application examples include indexing and search, graph analysis, text analysis, machine learning, data transformation, and so forth.
- These types of applications are often difficult to implement using the standard SQL employed by relational DBMSs.
- It also has the advantage that developers do not have to be concerned with implementing parallel computing—this is handled transparently by the system.

Hadoop Technical Components

In addition to MapReduce, a Hadoop “stack” is made up of a number of components,

- **Hadoop Distributed File System (HDFS):** The default storage layer in any given Hadoop cluster.
- **Name Node:** The node in a Hadoop cluster that provides the client information on where in the cluster particular data is stored and if any nodes fail.
- **Secondary Node:** A backup to the Name Node, it periodically replicates and stores data from the Name Node.
- **Job Tracker:** The node in a Hadoop cluster that initiates and coordinates MapReduce jobs or the processing of the data.
- **Worker Nodes:** The grunts of any Hadoop cluster, worker nodes store data and take direction to process it from the Job Tracker.

In addition to these components, the Hadoop ecosystem is made up of a continuously expanding large number of complementary subprojects.

Hive- It is a Hadoop-based data warehousing-like framework originally developed by Facebook. It allows users to write queries in an SQL-like language called HiveQL, which are then converted to MapReduce.

Pig - Pig is a Hadoop-based query language developed by Yahoo! It is relatively easy to learn and is adept at very deep, very long data pipelines (a limitation of SQL).

HBase- HBase is a nonrelational database that allows for low-latency, quick lookups in Hadoop. It adds transactional capabilities to Hadoop, allowing users to conduct updates, inserts, and deletes. eBay and Facebook use HBase heavily.

Flume- Flume is a framework for populating Hadoop with data. Agents are populated throughout one's IT infrastructure—inside Web servers, application servers, and mobile devices, **example**—to collect data and integrate it into Hadoop.

Hadoop vs Spark

Performance: Spark is faster (in-memory processing) while Hadoop is slower (disk-based, batch processing).

Cost: Hadoop is cheaper (uses disk storage), Spark is costlier (needs lots of RAM).

Processing: Hadoop → best for batch/linear processing. Spark → best for real-time, streaming, and iterative tasks.

Scalability: Both scale well using HDFS.

Security: Hadoop is more secure overall, Spark has basic authentication but can integrate with Hadoop.

Analytics: Spark is stronger (includes MLlib for machine learning, regression, classification, etc.).

Use Hadoop for: very large data beyond memory, batch jobs, low budget, historical/archive analysis.

Use Spark for: real-time stream processing, quick results, iterative algorithms, graph processing, ML applications.

NoSQL Databases

NoSQL (Not Only SQL) databases are designed to handle large-scale, unstructured, and semi-structured data that traditional relational databases (RDBMS) cannot manage efficiently.

Unlike relational databases, NoSQL systems:

- Do not require fixed schemas (flexible structure).
- Scale easily across distributed servers.
- Prioritize high performance, scalability, and low latency over strict ACID compliance..

When to Use NoSQL?

- Real-time applications - gaming leaderboards, stock trading, fraud detection.
- Scalable systems - social media feeds, IoT sensor data, recommendation systems.
- Flexible schema requirements - rapidly changing datasets (e.g., product catalogs).

BIG DATA AND STREAM ANALYTICS

- A stream is defined as a continuous sequence of data elements.
- Stream analytics processes data in real time, as it is created, enabling instant insights and faster decision-making.
- Why important? Time-to-action is critical, and technology now allows real-time capture and analysis of continuous data streams.
- Stream analytics enables real-time insights across industries, turning continuous data streams into instant, actionable decisions.

Applications

Energy/Smart Grids – Real-time power distribution, predicting demand, optimizing renewable energy use.

E-Commerce – Amazon/eBay analyze clicks, searches, and views instantly to recommend products and maximize sales.

Telecommunications – Analyze call detail records (CDR) to prevent churn, identify influencers, improve campaigns



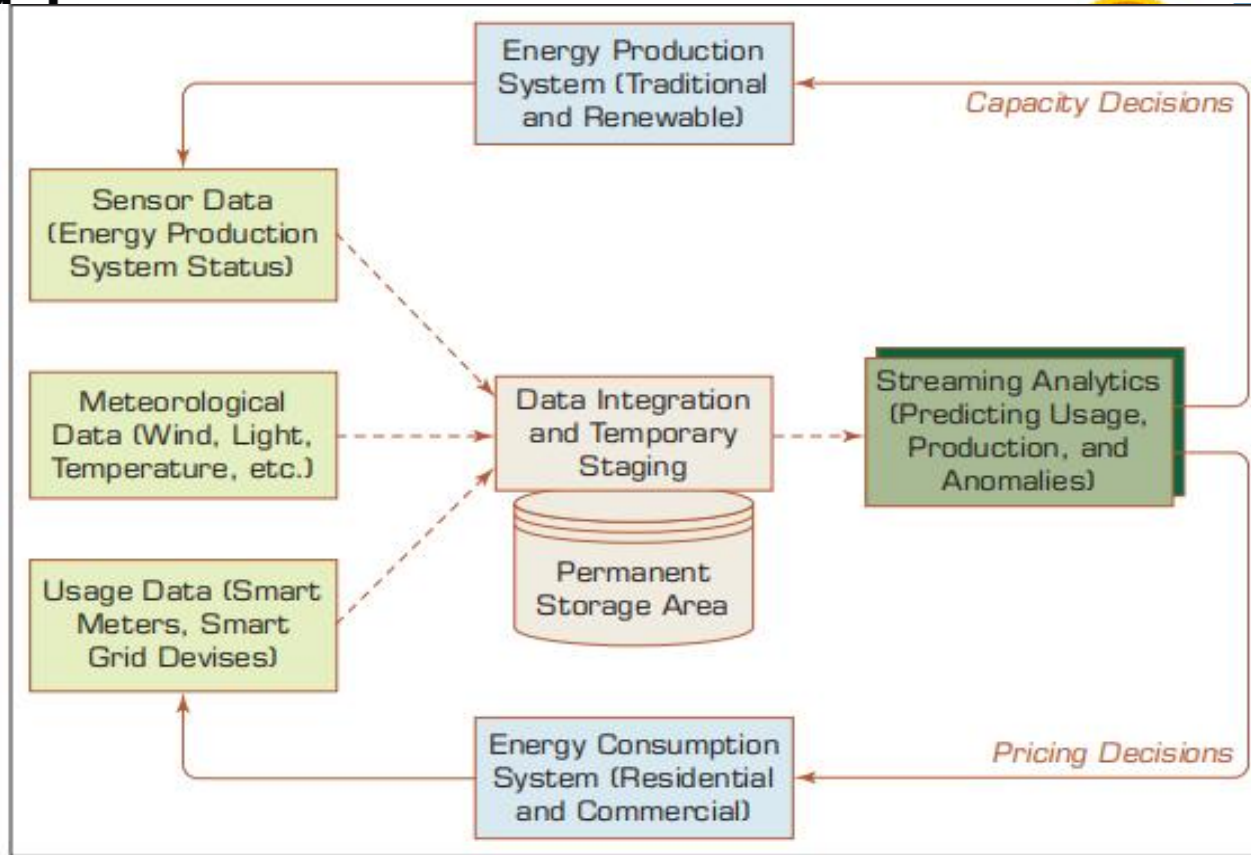
Law Enforcement & Cybersecurity – Real-time crime detection, surveillance, fraud detection, and network intrusion prevention.

Power Industry – Smart meters and sensors provide data for optimizing supply, integrating weather data, and forecasting demand.

Financial Services – Real-time trading decisions, fraud detection, and monitoring illegal activities.

Health Sciences – Medical devices stream patient data (e.g., ECG, blood pressure) for anomaly detection and life-saving interventions.

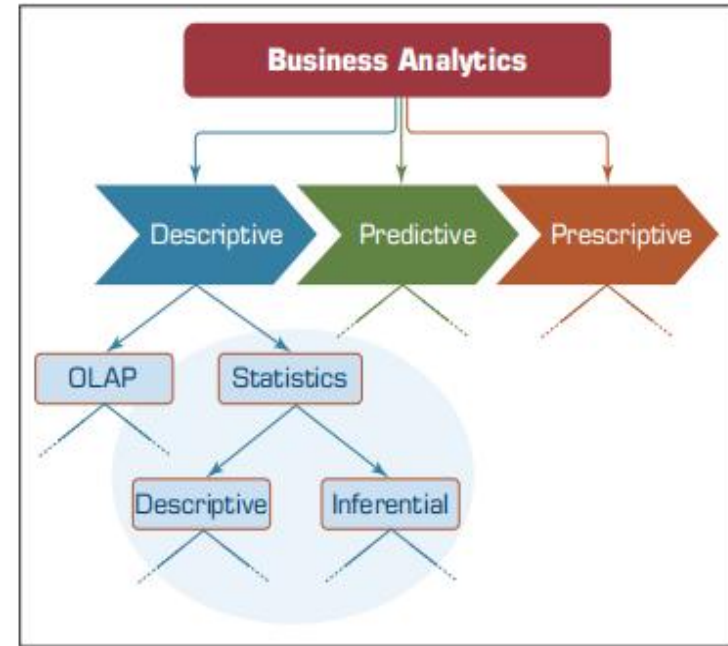
Government – Disaster management, monitoring environment, traffic management using live sensor and GPS data.



A Use Case of Streaming Analytics in the Energy Industry

STATISTICAL MODELING FOR BUSINESS ANALYTICS

- Because of the increasing popularity of business analytics, the traditional statistical methods and underlying techniques are also regaining their attractiveness as enabling tools to support evidence-based managerial decision making.
- Statistics (statistical methods and underlying techniques) is usually considered as part of descriptive analytics.
- Fig shows the relationship



Relationship between Statistics and Descriptive Analytics

- Some of the statistical methods can also be considered as part of predictive analytics such as discriminant analysis.
- As shown in Figure , descriptive analytics has two main branches: **statistics** and **online analytics processing (OLAP)** .
- OLAP is the term used for analyzing, characterizing, and summarizing structured data stored in organizational databases (often stored in a data warehouse or in a data mart).
- The OLAP branch of descriptive analytics has also been called business intelligence.
- Statistics, on the other hand, helps to characterize the data either one variable at a time or multivariables all together, using either descriptive or inferential methods.
- Descriptive Statistics → summarizes data at hand.
- Inferential Statistics → draws conclusions about the population.

Descriptive Statistics

- Helps explain and present data (tables, graphs, charts).
- Identifies patterns, outliers, and unusual values.

Measures of Central Tendency (Centrality):

Mean

- The mean is the sum of all data values divided by the total number of values.
- It represents the "balance point" of the dataset.

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Formula:

For a dataset with n observations:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Median

The middle value when the data is arranged in ascending (or descending) order.

Splits the dataset into two equal halves.

Suitable for ordinal, interval, and ratio data.

Robust to outliers and skewed data (unlike mean).

Mode

The value that appears most frequently in the dataset.

Represents the "most popular" or common observation.

Formula:

No standard mathematical formula; it is simply the most frequent value(s).

Useful for categorical (nominal) data.

Measures of Dispersion (Spread):

Show how data values vary around the central point.

Range: Max – Min.

Variance: Average squared deviation from mean.

Standard Deviation: Square root of variance; widely used.

Mean Absolute Deviation: Average absolute differences from mean.

Quartiles & Interquartile Range (IQR): Focus on middle 50% of data; robust to outliers.

Box Plot: Visual tool for showing centrality and spread.



A T M E
College of Engineering



ISO 9001:2015



Shape of Distribution:

Normal Distribution: Symmetric bell curve.

Skewness: Asymmetry of distribution.

Positive $\rightarrow$ long right tail.

Negative $\rightarrow$ long left tail.

Kurtosis: Peakiness of distribution.

3 $\rightarrow$ peaked, $<3 \rightarrow$ flat.

REGRESSION MODELING FOR INFERENCE STATISTICS

- Regression is a statistical method used to model the relationship between one or more independent (explanatory) variables and a dependent (response) variable.
- It helps in understanding, explaining, and predicting outcomes.
- Hypothesis testing: To find relationships between variables.
- Prediction/Forecasting: To estimate future values of the dependent variable.

Correlation vs Regression:

- Correlation: Measures the degree of association between variables (no cause-effect).
- Regression: Determines dependency — how one variable affects another (cause-effect relationship).



Types of Regression

- **Simple Regression:**

One independent variable and one dependent variable.

Example: Predicting a person's weight from their height.

- **Multiple Regression:**

More than one independent variable.

Example: Predicting weight from height, BMI, and gender.

- **Nonlinear Regression:**

Used when the relationship between variables is not linear.

Assumptions in Linear Regression

- Linearity: Relationship between variables is linear and additive.
- Independence: Errors (residuals) are independent of each other.
- Normality: Errors follow a normal distribution.
- Constant variance (Homoscedasticity): Error variance remains the same across all values.
- No multicollinearity: Independent variables are not highly correlated with each other.

Logistic Regression

- A classification technique used to predict categorical outcomes (e.g., Yes/No, Pass/Fail) rather than continuous ones.
- It models the probability of an event occurring using the logit function (log of odds).
- Converts categorical output into a continuous scale (0–1 range).

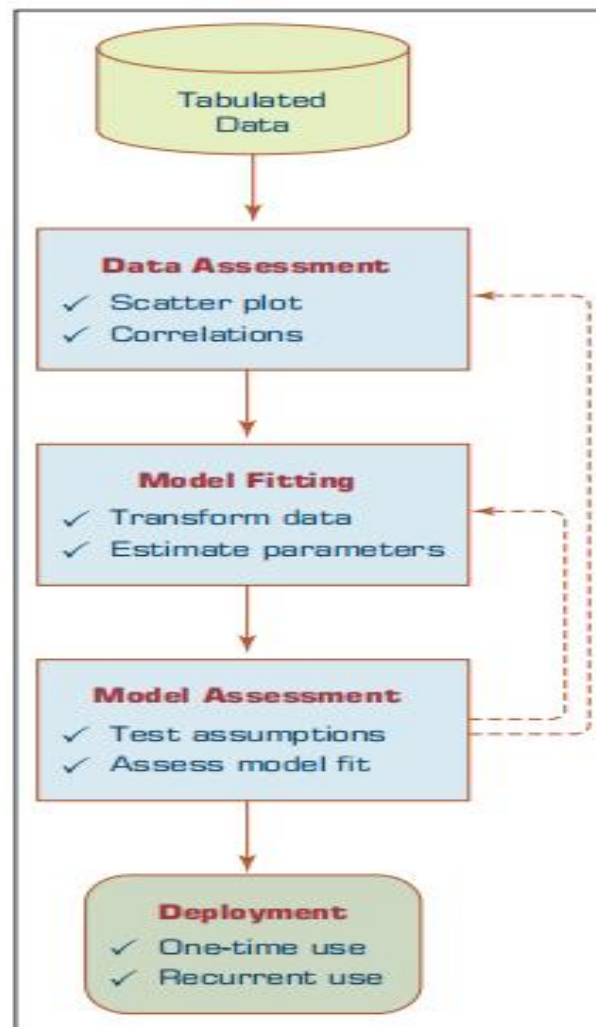
Types:

- Binary Logistic Regression: Two outcome classes (e.g., accept/reject).
- Multinomial Logistic Regression: More than two outcome classes.

Uses: Common in predictive analytics, medical research, and social sciences to classify outcomes.

- Linear regression predicts continuous values (like sales, prices).
- Logistic regression predicts categorical outcomes (like pass/fail, spam/not spam).
- Both help organizations analyze trends, test hypotheses, and make data-driven decisions.

A Process Flow for Developing Regression Models



1. Tabulated Data

- This is the starting point — the raw dataset organized in a table (rows as observations, columns as variables).
- It includes both dependent (target) and independent (predictor) variables.
- Example: A dataset with columns like Hours Studied (X) and Exam Score (Y).

2. Data Assessment

- In this step, the data is explored and analyzed to understand relationships among variables.
- Scatter Plot: Visualize how one variable changes with another (e.g., linear or non-linear relationship).
- Correlations: Measure the strength and direction of relationships between variables.
- The goal is to check whether a linear relationship exists and if the data quality is good enough for modeling.

3. Model Fitting

Here, the linear regression equation is created and parameters (coefficients) are estimated.

Transform Data: If the data is not linear, apply transformations (e.g., log, square root) to achieve linearity.

Estimate Parameters: Use statistical methods (like least squares) to find the best-fit line that minimizes prediction errors.

4. Model Assessment

The model is evaluated to ensure it fits the data well and satisfies regression assumptions.

Test Assumptions: Check for linearity, independence, normality, and equal variance (homoscedasticity) of residuals.

Assess Model Fit: Use performance metrics like R^2 , RMSE, and residual plots to judge how well the model explains the data.

If issues are found, the process loops back to Data Assessment or Model Fitting for improvement (as shown by dotted arrows).

5. Deployment

- Once the model performs well, it can be deployed for practical use.
- Two types of deployment:
 - One-time use: Apply the model to a single dataset for analysis or prediction.
 - Recurrent use: Implement the model in a system for ongoing or automated predictions (e.g., sales forecasting, demand prediction).