AKTU · 3RD YEAR · SEMESTER 6 · BAD601

Data Analytics

Unit 5 — Frameworks, Visualization and R Programming

- Frameworks: MapReduce, Hadoop, Pig, Hive, HBase, MapR, Sharding, NoSQL Databases, S3, Hadoop Distributed File Systems
- Visualization: visual data analysis techniques, interaction techniques, systems and applications
- Introduction to R
- R graphical user interfaces
- Data import and export
- Attribute and data types
- Descriptive statistics
- Exploratory data analysis
- Visualization before analysis
- Analytics for unstructured data



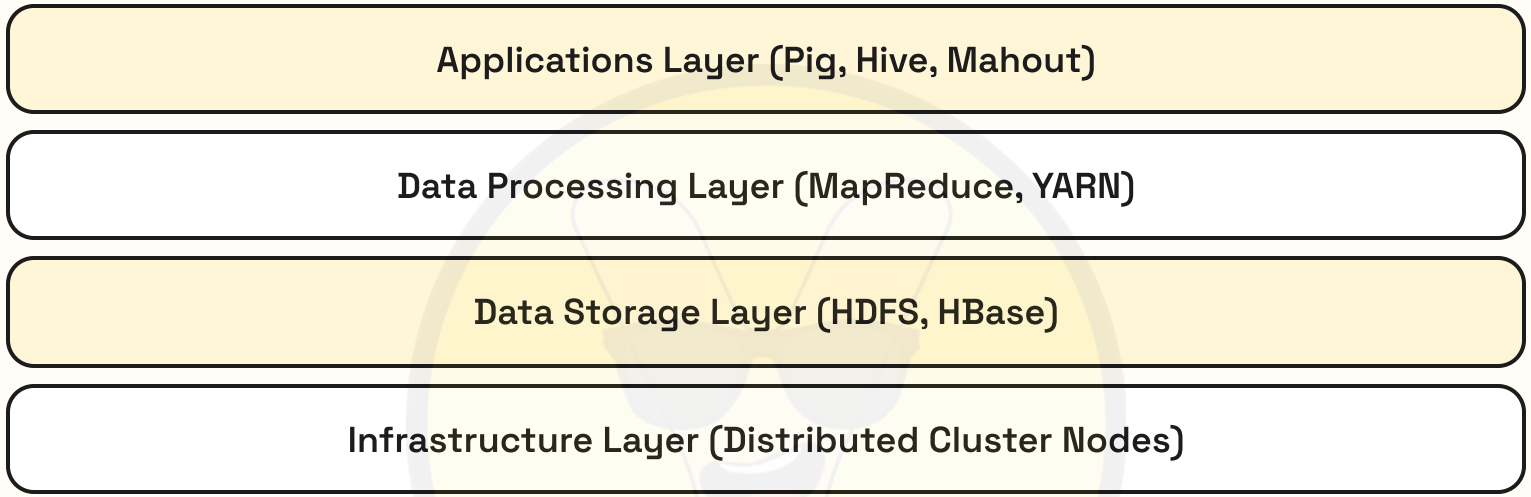
DECODED BY

Chirag Jain

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

theycotes.com

Hadoop Ecosystem Architecture



Stacked abstraction layers of the Apache Hadoop Ecosystem.

Hadoop Distributed File System (HDFS)

HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

A Java-based, highly fault-tolerant distributed file system designed to store and process massive datasets across clusters of commodity hardware.

- NameNode manages system namespace, directory tree, and file block locations metadata.
- DataNodes store actual data blocks and execute read/write commands from clients.
- Files are split into fixed-size blocks (default 128 MB) distributed across nodes.
- Replication factor (default 3) ensures data availability despite node failures.

MapReduce Execution Architecture



Sequential data processing flow of MapReduce framework.

Comparison: Apache Pig, Hive, and HBase

Parameter	Apache Pig	Apache Hive	Apache HBase
Primary Purpose	Data flow execution	Data warehousing	NoSQL key-value DB
Query Language	Pig Latin (Procedural)	HiveQL (Declarative SQL)	Java API / Shell
Abstraction	Dataflow scripts	Tables and Schemas	Sparse Column Rows
Latency	Batch processing (High)	Batch processing (High)	Real-time lookup (Low)
Target Users	Data Engineers	SQL Analysts	Application Developers

Enterprise Distributed Storage: MapR-FS and AWS S3

Modern data architectures leverage high-performance distributed file systems and cloud storage alternatives.

- MapR File System (MapR-FS) provides POSIX compliance and direct read/write file access without NameNode bottlenecks.
- Amazon Simple Storage Service (AWS S3)
  - Object storage solution decoupling storage from compute resources.
  - Offers extreme durability, auto-scaling, and lower operational overhead than HDFS.
- HDFS requires active cluster management, whereas S3 provides serverless object storage.

Sharding and NoSQL Database Architecture

SHARDING

A database scaling technique that horizontally partitions a large database into smaller, independent chunks called shards across multiple servers.

- Horizontal partitioning splits tables by rows, assigning distinct partition keys to shards.
- NoSQL databases trade traditional ACID properties for high scalability and availability (BASE model).
- CAP Theorem states a distributed system can simultaneously guarantee only two of: Consistency, Availability, Partition Tolerance.
- NoSQL types include Document (MongoDB), Key-Value (Redis), Column-Family (Cassandra), Graph (Neo4j).

Visual Data Analysis Techniques

Visual data analysis leverages graphical representations to inspect complex data structures and extract insights.

- Geometric projection techniques map multi-dimensional data into 2D or 3D scatter plots and parallel coordinates.
- Pixel-oriented techniques represent each data value as a colored pixel to maximize visual data density.
- Icon-based techniques map data dimensions to visual features like Chernoff faces or stick figures.
- Hierarchical techniques visualize tree-structured relationships using treemaps and sunburst diagrams.

Interaction Techniques and Visualization Systems

Interactive visual interfaces enable users to dynamically explore, filter, and isolate data patterns.

- Brushing and linking highlights corresponding data points across multiple related visualizations simultaneously.
- Zooming and panning allows users to alter visual scale and navigate large visual datasets smoothly.
- Filtering dynamically constrains displayed data ranges using interactive controls like sliders and checkboxes.
- Popular visualization systems include Tableau, PowerBI, D3.js, and ggplot2 in R.

Importance of Visualization Before Analysis

Visualizing raw data prior to statistical modeling prevents wrong assumptions and analytical mistakes.

- Anscombe's Quartet demonstrates that datasets with identical summary statistics can have drastically different visual distributions.
- Identifies extreme outliers and noise that severely skew parametric statistical metrics like mean and variance.
- Reveals non-linear relationships, clusters, and multi-modal distributions invisible in simple numerical summaries.
- Guides appropriate model selection by validating statistical assumptions like normality and homoscedasticity.

Introduction to R Programming and R GUIs

R is an open-source environment specifically created for statistical computing, data analysis, and graphics.

- Interpreted language with extensive statistical packages available via Comprehensive R Archive Network (CRAN).
- RStudio IDE provides integrated code editor, console, environment manager, and visualization viewer.
- GUI alternatives like R Commander (Rcmdr) offer point-and-click interfaces for introductory statistical tasks.
- Vectorized operations allow efficient computation on entire datasets without explicit loop structures.

R Data Structures and Data Types

Structure	Dimensions	Homogeneous?	Key Characteristics
Vector	1D	Yes	Basic sequence of elements of same type
Matrix	2D	Yes	Two-dimensional grid of homogeneous elements
Array	n-D	Yes	Multi-dimensional extension of matrices
List	1D	No	Heterogeneous collection of arbitrary objects
Data Frame	2D	No	Tabular structure with columns of different types

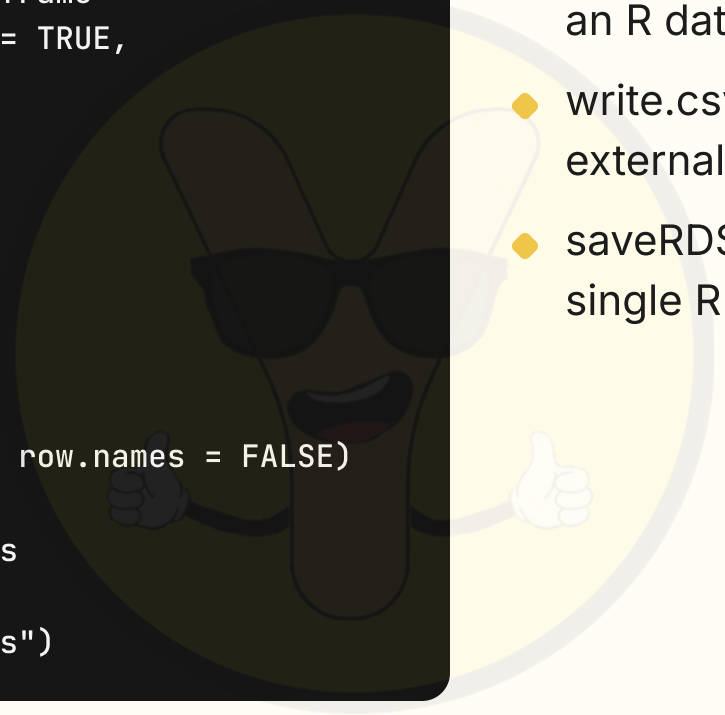
Data Import and Export Operations in R

```
# Reading tabular CSV data into R data frame
data ← read.csv("dataset.csv", header = TRUE,
stringsAsFactors = FALSE)

# Inspecting imported data structure
head(data, 5)
str(data)
summary(data)

# Exporting processed data frame to CSV
write.csv(data, "output_processed.csv", row.names = FALSE)

# Reading and writing RDS R object files
saveRDS(data, file = "data_store.rds")
restored_data ← readRDS("data_store.rds")
```



- read.csv() imports comma-separated data into an R data.frame object.
- write.csv() exports processed tabular data to external disk storage.
- saveRDS() and readRDS() efficiently store single R objects in binary format.

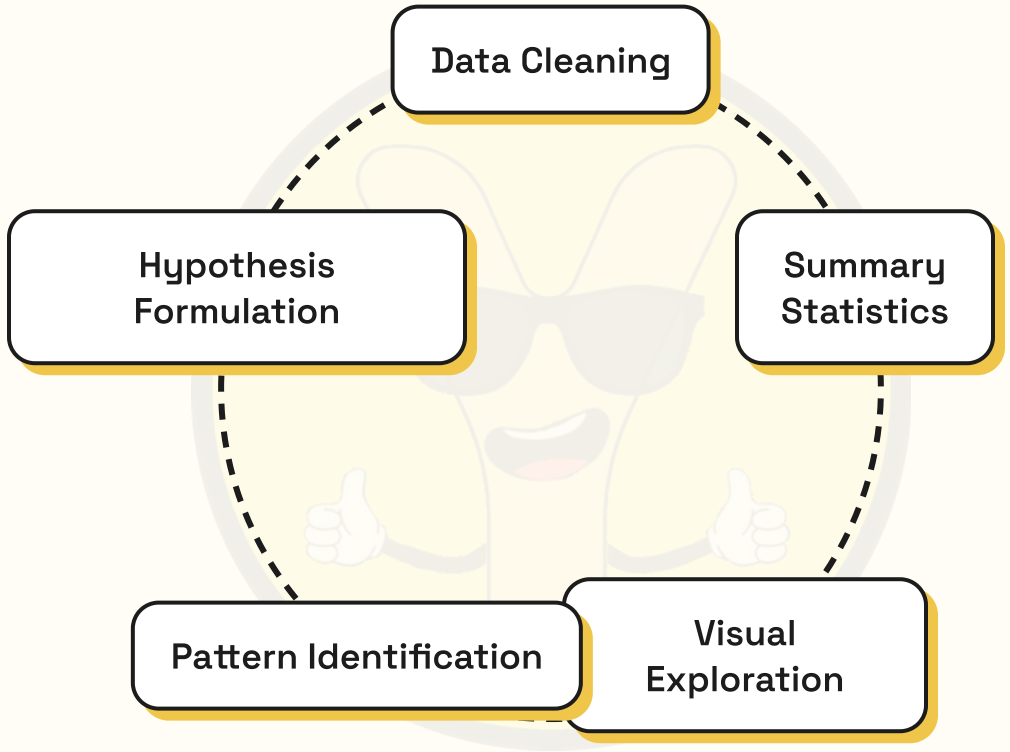
Descriptive Statistics in R: Formulas and Calculation

Given sample observations X = {10, 20, 30, 40, 50}, calculate sample mean $\bar{x}$, sample variance s^2, and sample standard deviation s.

1. Sample Mean formula: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{10+20+30+40+50}{5} = 30$
2. Deviations from mean: $(10-30)^2 = 400, (20-30)^2 = 100, (30-30)^2 = 0, (40-30)^2 = 100, (50-30)^2 = 400$
3. Sum of squared deviations: $\sum(x_i - \bar{x})^2 = 400 + 100 + 0 + 100 + 400 = 1000$
4. Sample Variance formula: $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 = \frac{1000}{4} = 250$
5. Sample Standard Deviation formula: $s = \sqrt{s^2} = \sqrt{250} \approx 15.81$

Sample Mean $\bar{x} = 30$, Sample Variance $s^2 = 250$, Sample Standard Deviation $s \approx 15.81$.

Exploratory Data Analysis (EDA) Workflow



Iterative cycle of Exploratory Data Analysis.

Analytics for Unstructured Data

Unstructured data lacks formal schema and requires transformations to extract structured numerical features.

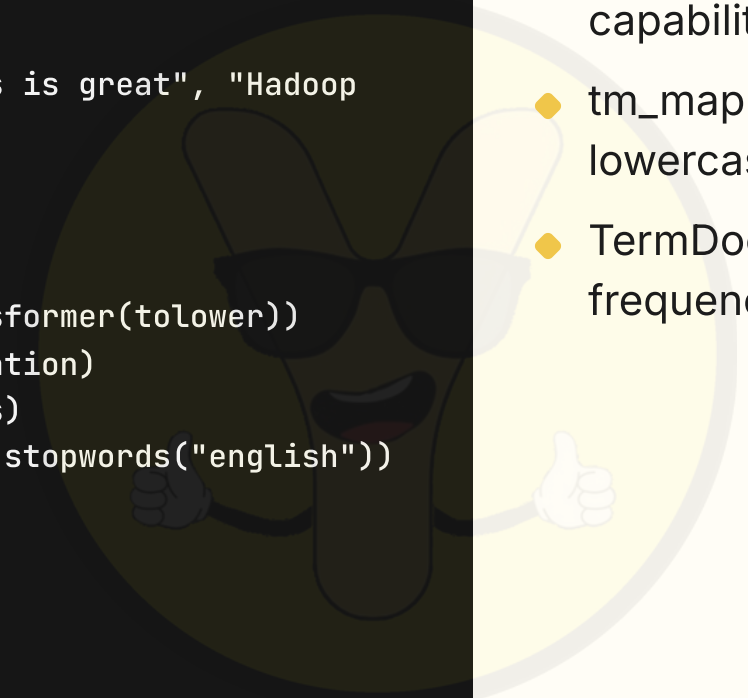
- Sources include raw text logs, social media posts, sensor streams, audio, and video files.
- Text Preprocessing involves tokenization, lowercasing, stop-word removal, and stemming or lemmatization.
- Vector Space Model converts text to document-term matrices using TF-IDF weighting.
- TF-IDF Formula
  - $\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log\left(\frac{N}{\text{DF}(t)}\right)$
  - Balances term frequency with inverse document frequency across corpus.
- Enables downstream ML tasks such as sentiment analysis, text classification, and topic modeling.

Text Analytics Pipeline in R

```
library(tm)
# Create text corpus from vector
docs ← VectorSource(c("Data analytics is great", "Hadoop
and R for analytics"))
corpus ← VCorpus(docs)

# Text Preprocessing
corpus ← tm_map(corpus, content_transformer(tolower))
corpus ← tm_map(corpus, removePunctuation)
corpus ← tm_map(corpus, removeNumbers)
corpus ← tm_map(corpus, removeWords, stopwords("english"))

# Generate Term-Document Matrix (TDM)
tdm ← TermDocumentMatrix(corpus)
m ← as.matrix(tdm)
print(m)
```



- tm package provides core text mining capabilities in R.
- tm_map() applies transformations like lowercasing and stopword removal.
- TermDocumentMatrix creates structured frequency matrix from unstructured text.

Comparative Summary: SQL vs NoSQL vs HDFS

Feature	Relational Databases (SQL)	NoSQL Databases	Hadoop HDFS
Data Structure	Structured tables	Semi-structured/Unstructured	Unstructured files/blocks
Schema Requirement	Schema-on-Write	Schema-on-Read / Dynamic	Schema-on-Read
Scalability	Vertical (Scale-up)	Horizontal (Scale-out)	Horizontal (Scale-out)
Transaction Consistency	ACID compliance	BASE (Eventual consistency)	Append-only write once
Optimal Use Case	OLTP operations	Real-time Web applications	Batch Big Data processing

AKTU Exam Corner: Likely Unit 5 Questions

- **Q1** Explain the HDFS architecture with a neat diagram. Discuss the roles of NameNode and DataNode.
- **Q2** Describe the MapReduce execution workflow with a suitable example like Word Count.
- **Q3** Differentiate between Apache Pig, Apache Hive, and Apache HBase based on architecture and use cases.
- **Q4** What is database sharding? Explain CAP theorem and its implications on NoSQL databases.
- **Q5** Discuss various visual data analysis techniques and explain the significance of visualization before analysis.
- **Q6** Explain data structures available in R. Write an R script to import CSV data and compute descriptive statistics.
- **Q7** Define TF-IDF. How is unstructured text data processed and analyzed using R programming?

Unit 5 done. You are exam-ready.

Revise the Exam corner once more, then open Unit 1 on theycotes.com. Every AKTU subject, unit by unit, free. Notes by Chirag Jain, Ycotes.

