Big Data Technologies in Practice

This presentation guides students through essential big data tools, from Hadoop to data mining techniques. We'll explore practical applications with real datasets.

Hadoop and Spark: The Foundation

Hadoop Ecosystem

Apache Hadoop provides distributed storage and processing of large datasets.

- HDFS for storage
- MapReduce for processing
- YARN for resource management

Apache Spark

Spark offers in-memory processing with greater speed than Hadoop.

- 100x faster than MapReduce
- Supports streaming data
- Includes ML libraries

Data Processing Workflows

Data Collection

Gathering structured and unstructured data from diverse sources.

Data Storage

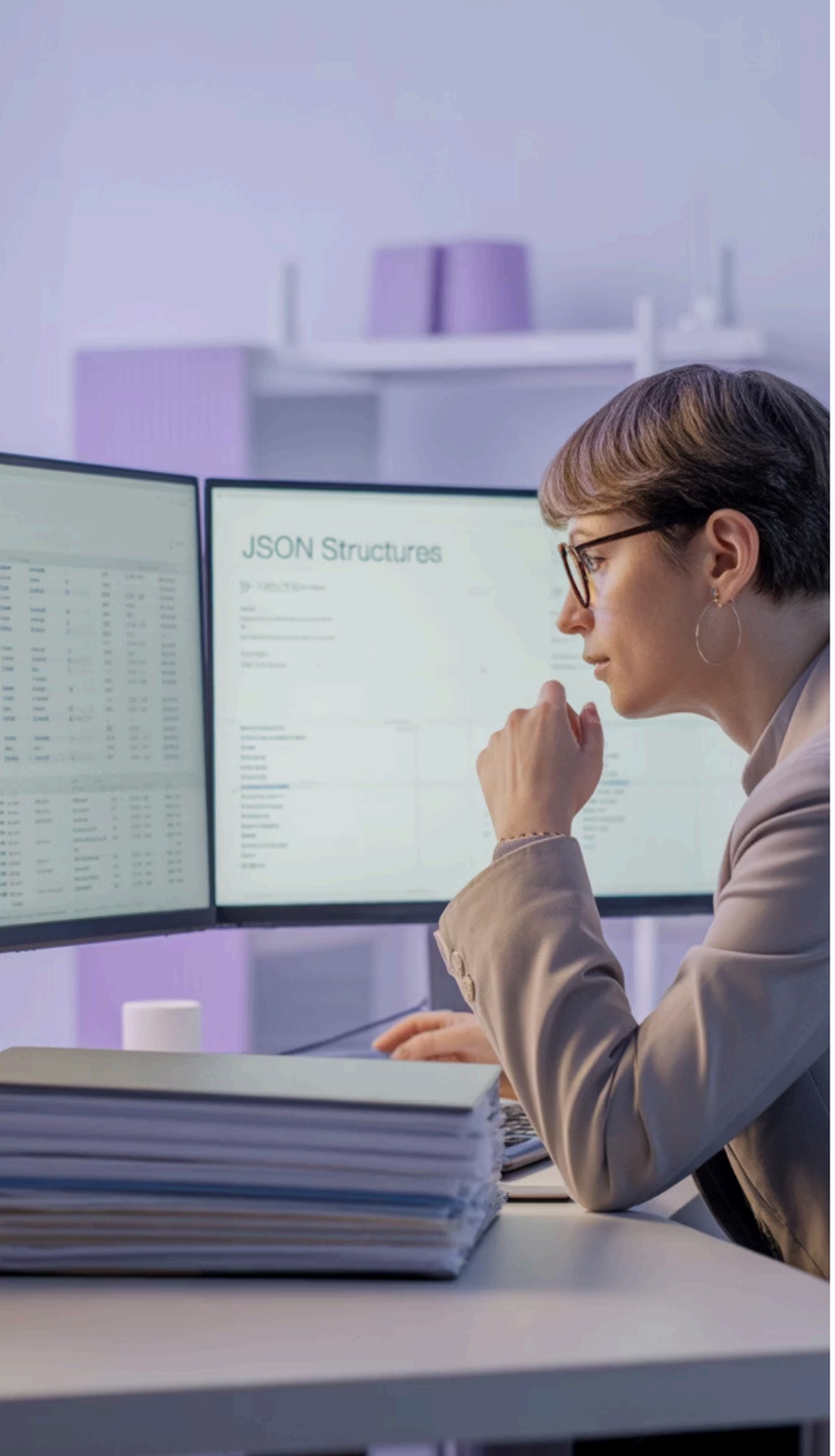
Implementing distributed storage solutions for large volumes.

Data Processing

Applying batch or stream processing to transform raw data.

Analysis & Visualization

Extracting insights and creating visual representations.



Working with Real-World Datasets

CSV Files

Comma-separated values format is widely used for tabular data.

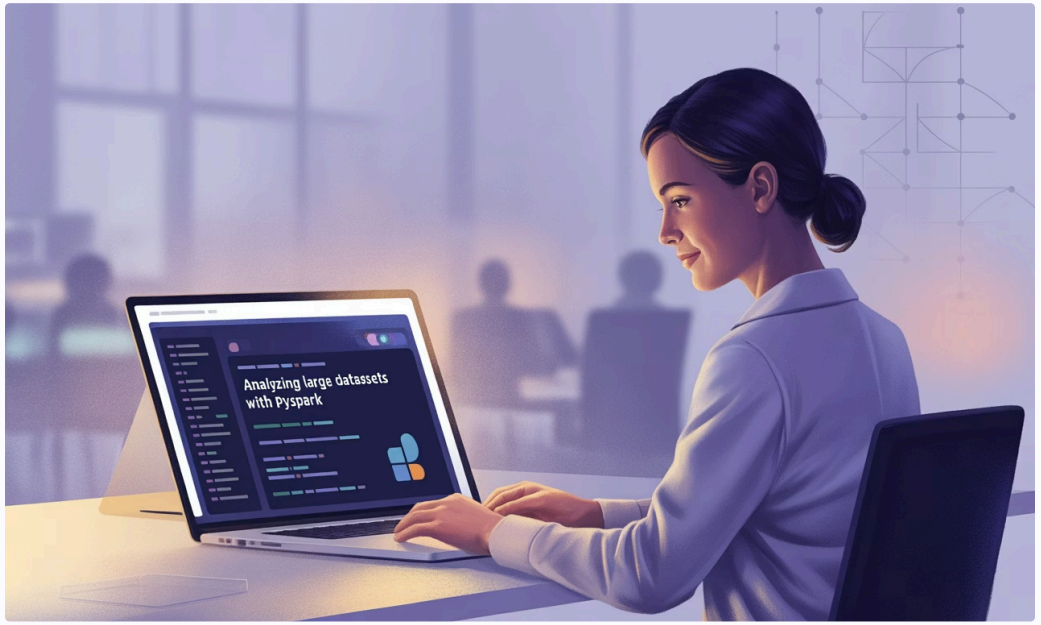
```
name,age,city
john,34,chicago
sam,28,boston
```

JSON Data

JavaScript Object Notation offers flexible hierarchical structure.

```
{
  "name": "John",
  "age": 34,
  "city": "Chicago"
}
```

Essential Big Data Tools



Python

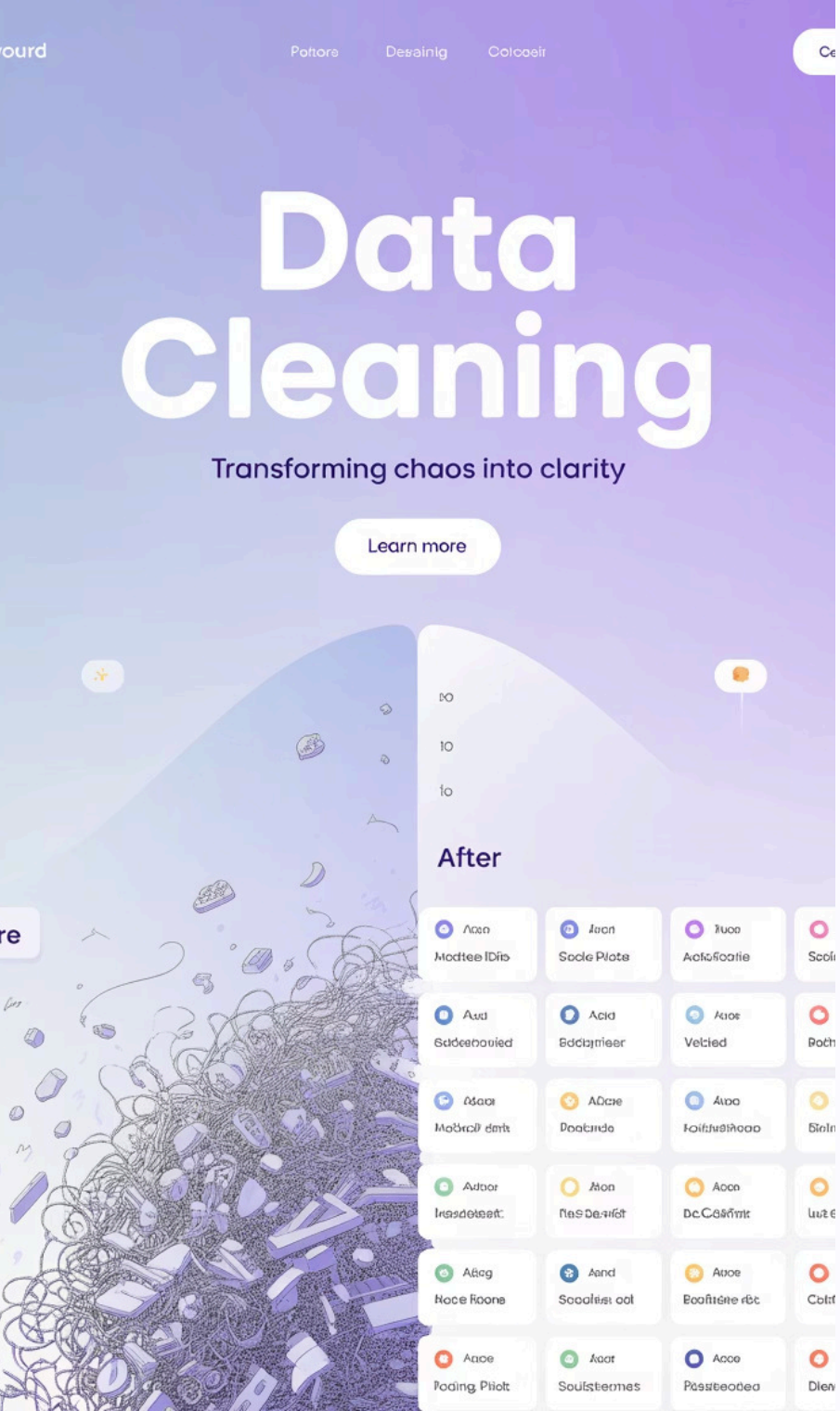
Versatile programming language with powerful data libraries like Pandas and NumPy.

PySpark

Python API for Spark that enables distributed data processing with familiar syntax.

RapidMiner

Visual workflow designer for data science with minimal coding required.



Data Wrangling and Cleaning

Identify Issues

- Missing values
- Outliers
- Duplicates

Clean Data

- Imputation
- Standardization
- Deduplication

Transform

- Normalization
- Feature engineering
- Aggregation

Introduction to Data Mining

What is Data Mining?

Data mining is the process of discovering patterns and knowledge from large datasets.

It combines statistics, machine learning, and database systems to extract insights.

Why It Matters

- Uncovers hidden patterns humans might miss
- Predicts future trends and behaviors
- Automates insight discovery from massive datasets

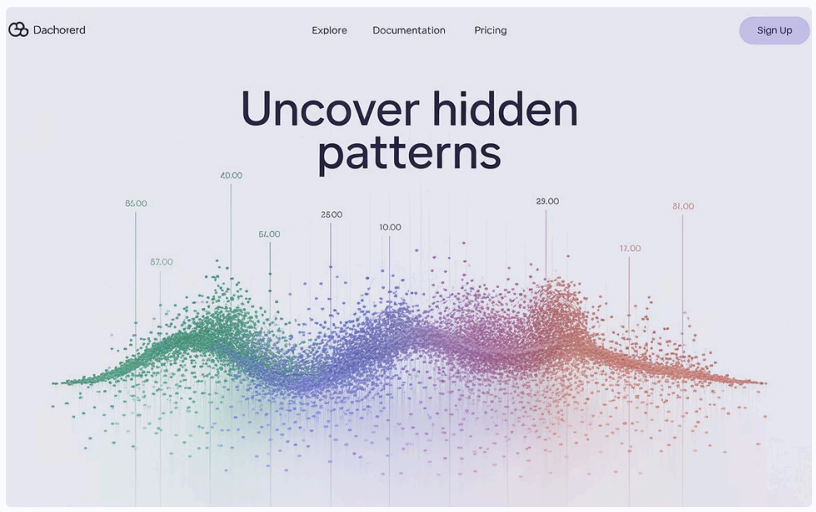
Basic Mining Techniques

Frequency Analysis

Examines how often values appear in a dataset. Helps identify common patterns and outliers.

Pattern Recognition

Identifies relationships between variables. Used in market basket analysis and recommendation systems.



Clustering

Groups similar data points together. Useful for customer segmentation and anomaly detection.