Navigating Big Data and Data Engineering

Understanding Tools, Architecture, and Concepts for Handling Large-Scale Data

ADEJO OJOMIDEJU GIDEON
FlexiSAF Generative AI and Data Science Internship – Advanced Stage
Module 3 Deliverable

Introduction to Big Data Engineering

01

Challenge

Storing, processing, and analyzing massive datasets.

02

Data Engineering

Building systems that can manage, transform, and optimize these data flows efficiently.

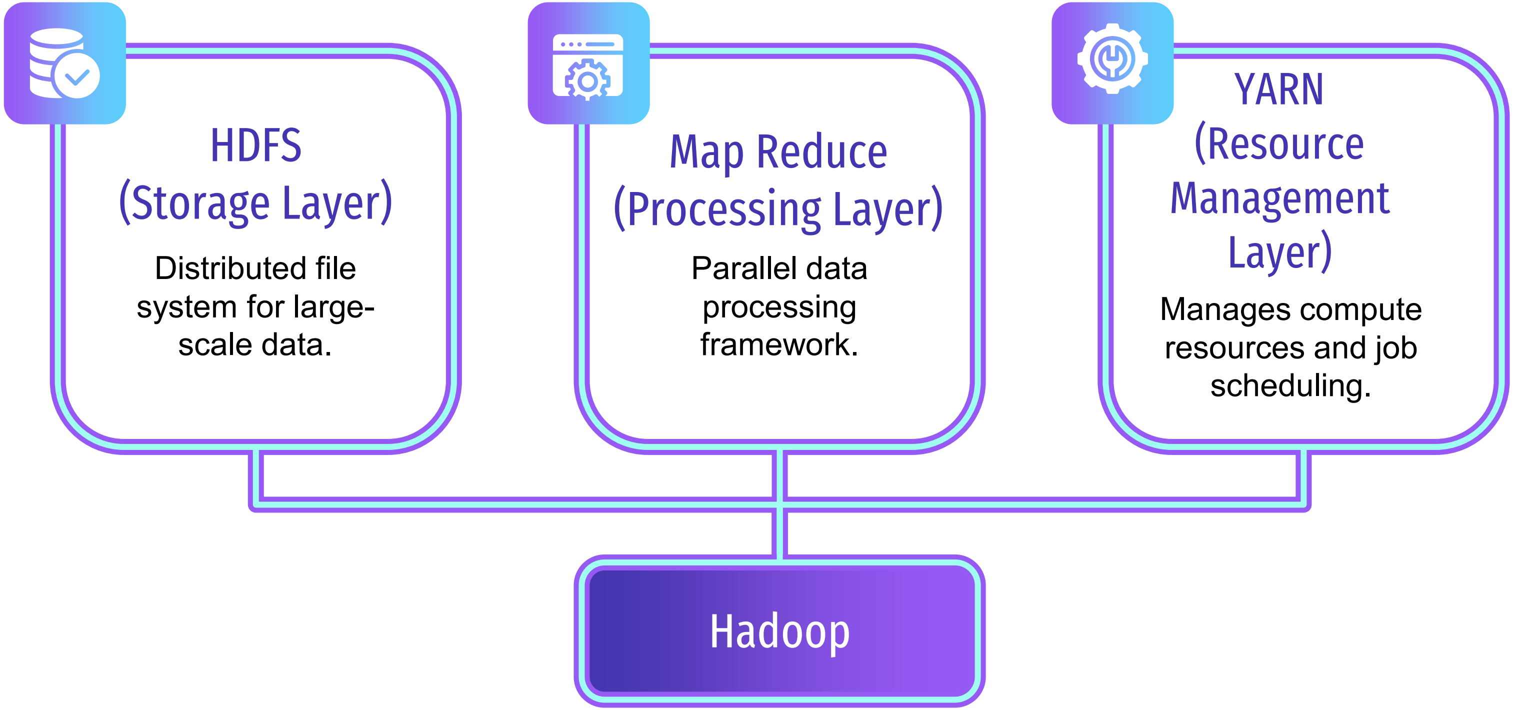
03

Solutions

Hadoop, Spark, and Hive are solutions designed to address these challenges.

Hadoop: The Foundation of Big Data

Built on Google's papers to manage distributed data processing.



HDFS (Storage Layer)

Distributed file system for large-scale data.

Map Reduce (Processing Layer)

Parallel data processing framework.

YARN (Resource Management Layer)

Manages compute resources and job scheduling.

Hadoop

HDFS Architecture & Reliability

- Data is split into blocks and distributed across nodes.
- Replication ensures fault tolerance — copies stored on multiple nodes.
- NameNode stores metadata; DataNodes store actual data blocks.
- If a node fails, replicas guarantee recovery and uninterrupted operation.

Hadoop Ecosystem Tools

Hive: SQL-like interface for data warehousing and querying.

Pig: Scripting interface for data transformation.

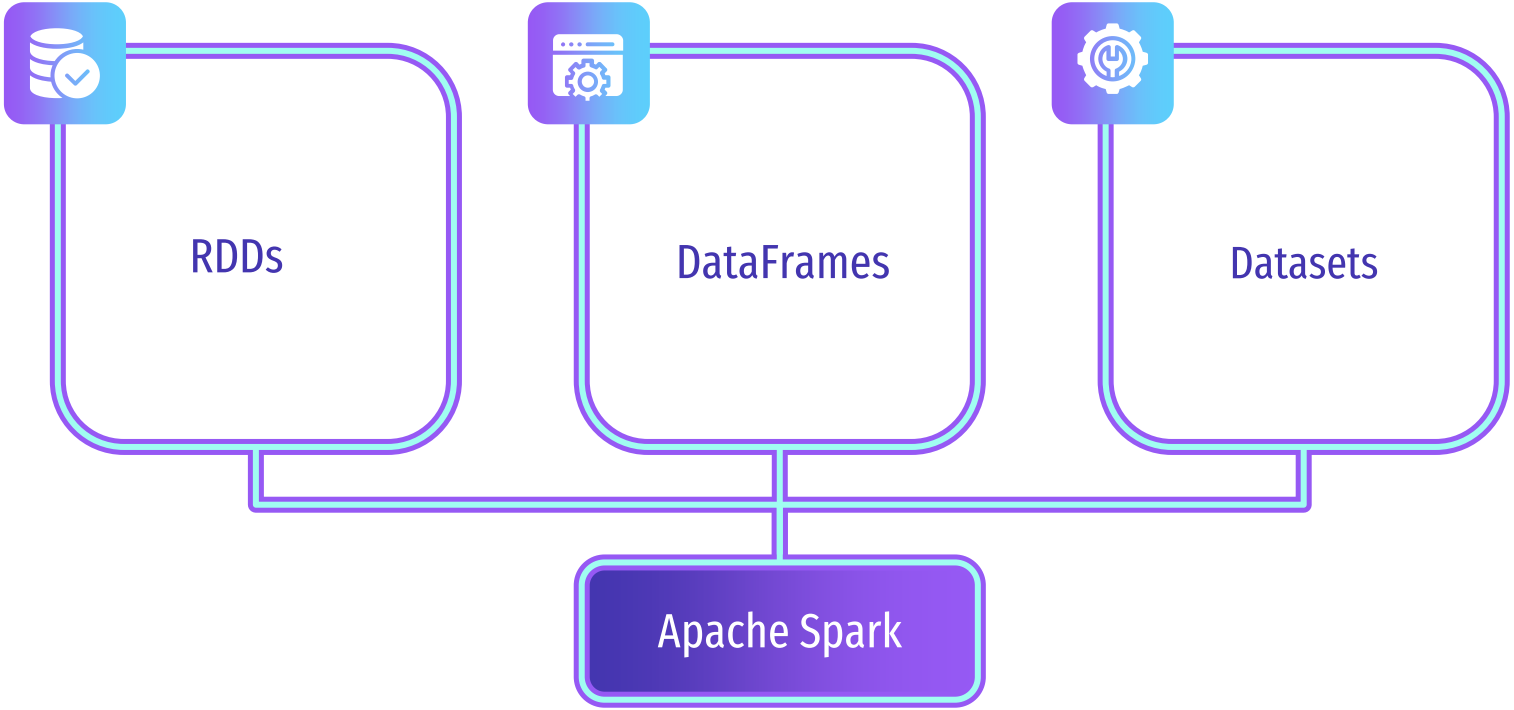
Sqoop: Data transfer between Hadoop and relational databases.

Oozie: Workflow scheduler for job coordination.

The tools above extend Hadoop’s capabilities while relying on the MapReduce or Spark engine.

Apache Spark: The Next Generation

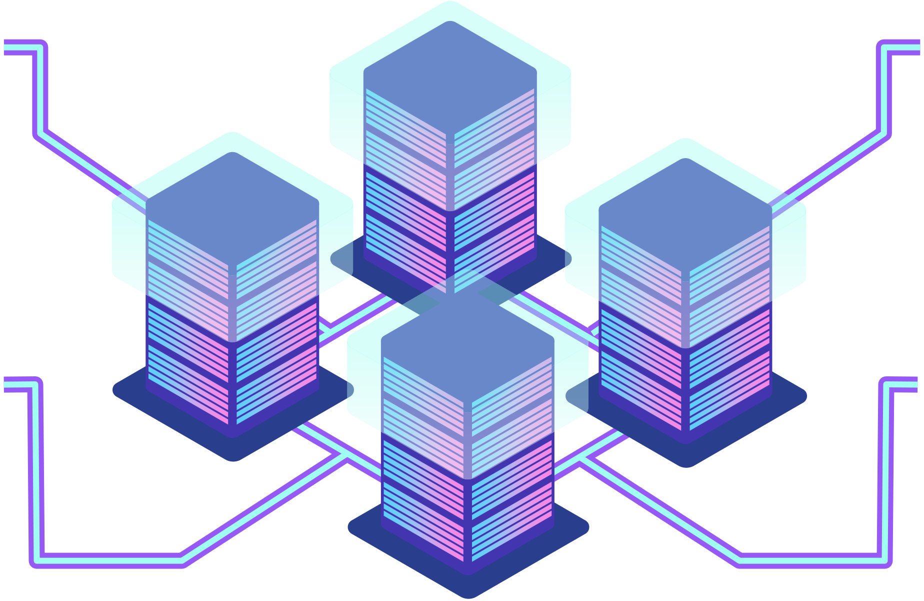
Spark revolutionized Big Data processing through in-memory computation.



Apache Spark: Key Advantages

- 100× faster than MapReduce for iterative jobs.
- Lazy evaluation optimizes resource usage. If a node fails, replicas guarantee recovery and uninterrupted operation.
- Supports multiple languages, like Python, Java, Scala.

Hive: Advanced Features & Optimization



Partitioning

Divides tables logically (e.g., by date) for faster queries.

ACID Tables

Enable INSERT, UPDATE, DELETE operations for transactional consistency

Bucketing

Hash-based grouping to optimize joins and sampling.

File Formats

ORC and Parquet for efficient, compressed, columnar storage.

Skills & Career Path for Data Engineers

CORE TECHNICAL SKILLS

- Linux commands and shell scripting.
- SQL proficiency.
- Programming in Java, Scala, or Python.

CAREER ROADMAP

- Learn Hadoop → Spark → Hive → Cloud Platforms → ETL Tools.
- Build hands-on projects and optimize resumes with practical experience.

Conclusion & Key Takeaways

Big Data Engineering bridges theory and practice in managing data at scale.

Hadoop provides distributed reliability, Spark offers speed and flexibility, and Hive adds data warehousing power.

Strong foundations in Linux, SQL, and programming are essential for success.

Continuous learning and experimentation are vital in navigating the evolving Big Data ecosystem.

THANK YOU!