BIG DATA ANALYTICS

Big Data Enabling Technologies

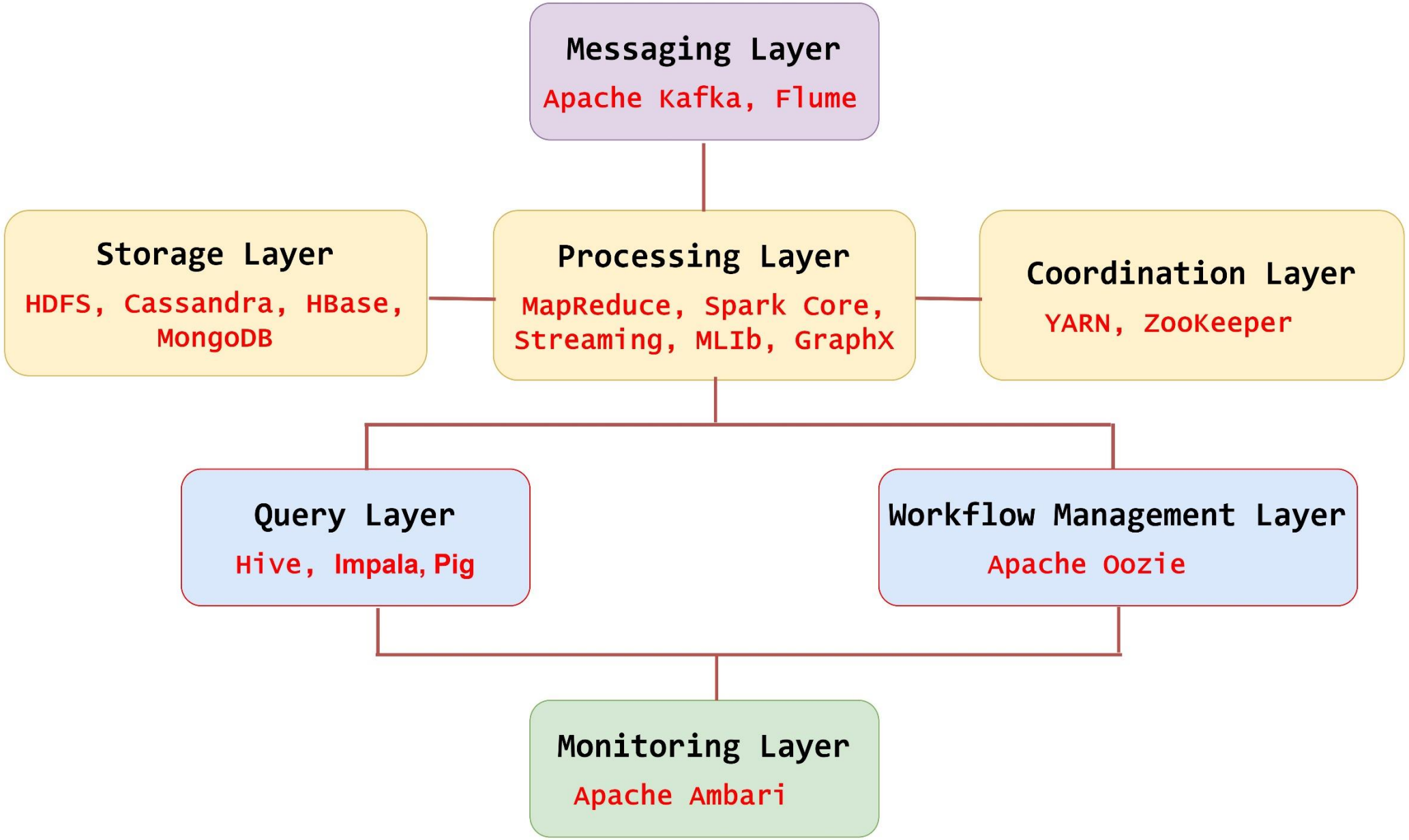
Introduction to Big Data: *Outline*

- Introduction
- Big Data Enabling Technologies 
- Hadoop Stack for Big Data

Big Data Enabling Technologies

- Big Data has transformed the way organizations handle, process, and analyze large volumes of data.
- Below is an overview of key technologies, an integrated diagram, their explanations, and real-time applications.

Integration Diagram of Big Data Enabling Technologies



Big Data Enabling Technologies

The diagram represents the relationships between core Big Data components and technologies:

1. Storage Layer: HDFS, NoSQL Databases (Cassandra, HBase, MongoDB)
2. Processing Layer: MapReduce, Apache Spark (Core, Streaming, MLlib, GraphX, PySpark)
3. Coordination Layer: YARN, ZooKeeper
4. Query Layer: Hive, Impala, Pig
5. Messaging Layer: Apache Kafka, Flume
6. Workflow Management: Apache Oozie
7. Monitoring Layer: Apache Ambari

Big Data Enabling Technologies

Explanations and Real-Time Applications:

1. Apache Hadoop:

- An open-source framework for distributed storage and processing of large datasets using a cluster of commodity hardware.

Applications:

- Log data analysis in IT companies.
- Fraud detection in banking and insurance sectors.

Big Data Enabling Technologies

Explanations and Real-Time Applications:

2. Hadoop Ecosystem:

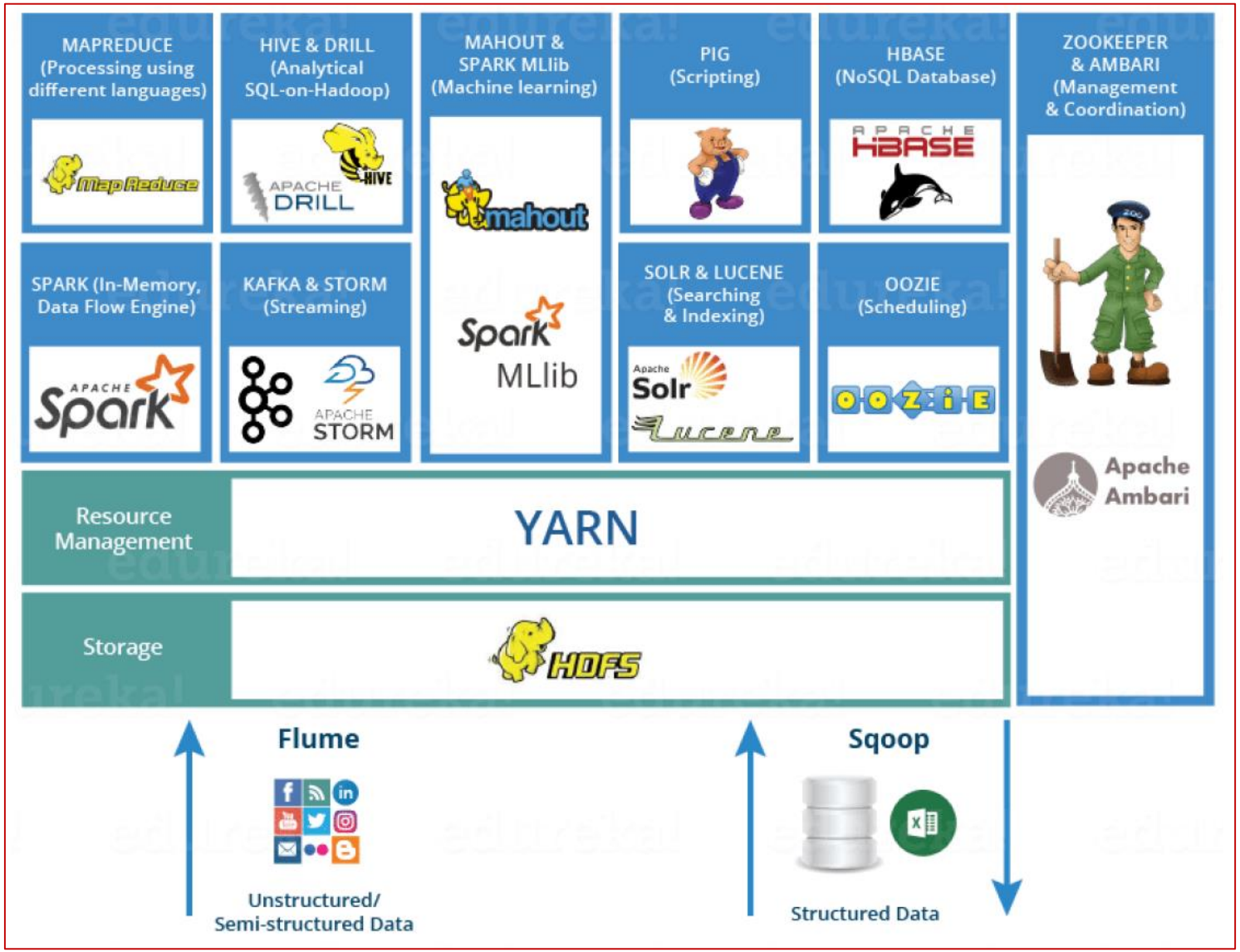
- Includes tools like HDFS, YARN, Hive, and MapReduce that work together to process Big Data efficiently.

Applications:

- Customer behavior analysis in e-commerce.
- Predictive maintenance in manufacturing.

Big Data Enabling Technologies

2. Hadoop Ecosystem:



Big Data Enabling Technologies

Explanations and Real-Time Applications:

3. HDFS (Hadoop Distributed File System) Architecture:

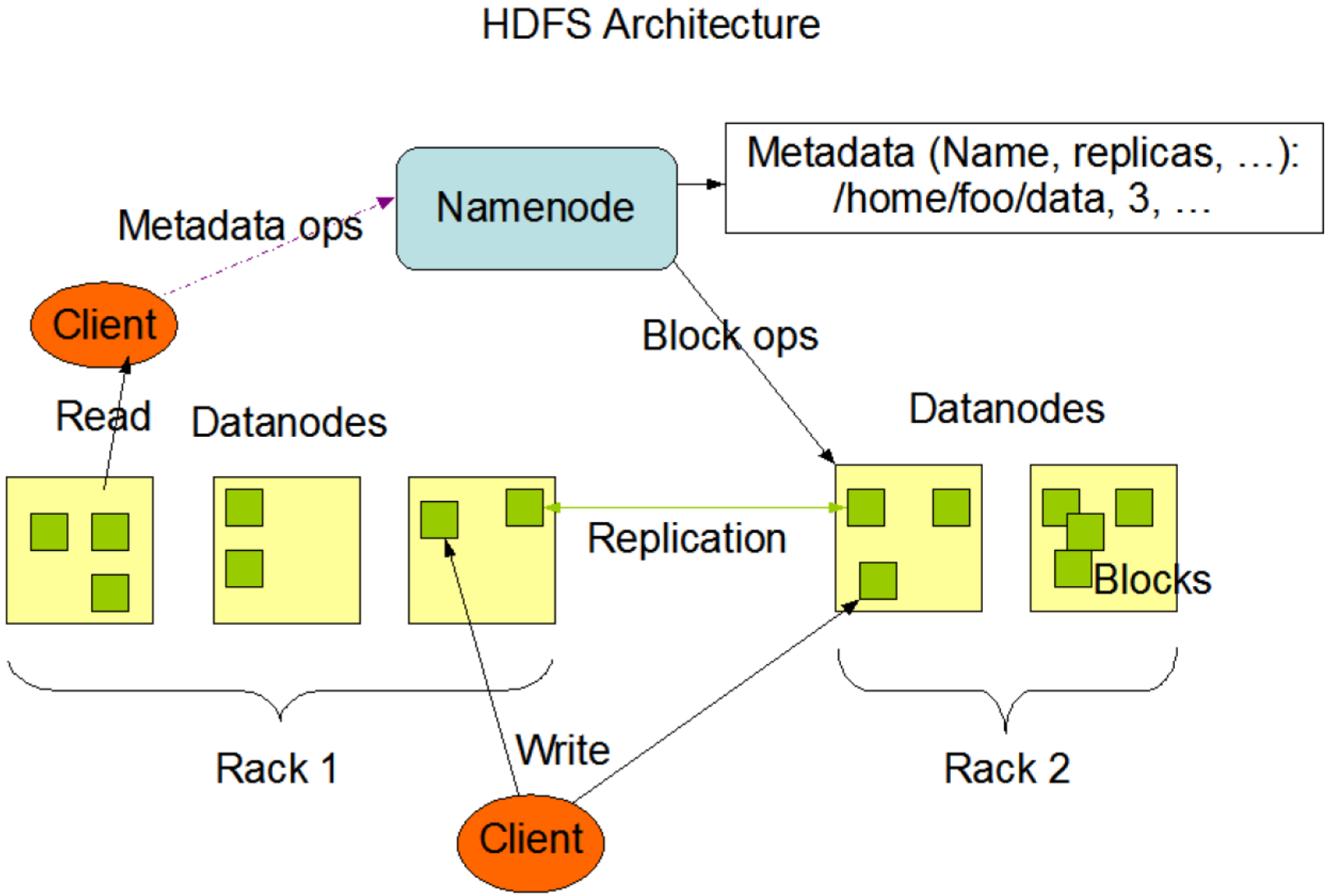
- A distributed file system designed for large-scale data storage and high-throughput access to data.

Applications:

- Video-on-demand services like Netflix.
- Scientific data storage for research institutions.

Big Data Enabling Technologies

3. HDFS (Hadoop Distributed File System) Architecture:



Big Data Enabling Technologies

Explanations and Real-Time Applications:

4. MapReduce:

- A programming model for processing and generating large data sets with a parallel, distributed algorithm on a cluster.

Applications:

- Indexing web pages in search engines.
- Analyzing social media trends.

Big Data Enabling Technologies

4. MapReduce Architecture

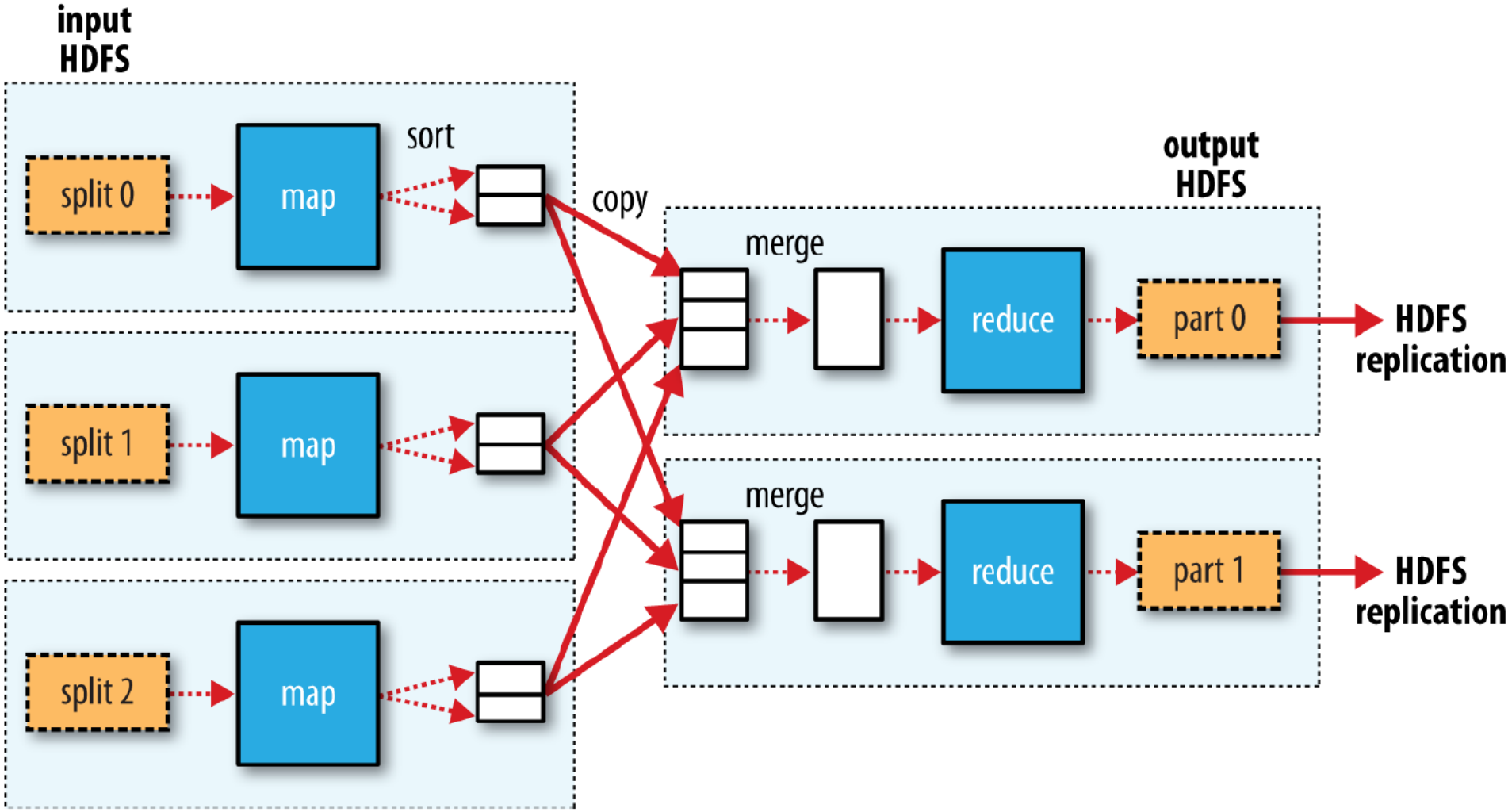


Figure. ***MapReduce data flow with multiple reduce tasks***

Big Data Enabling Technologies

4. MapReduce Architecture

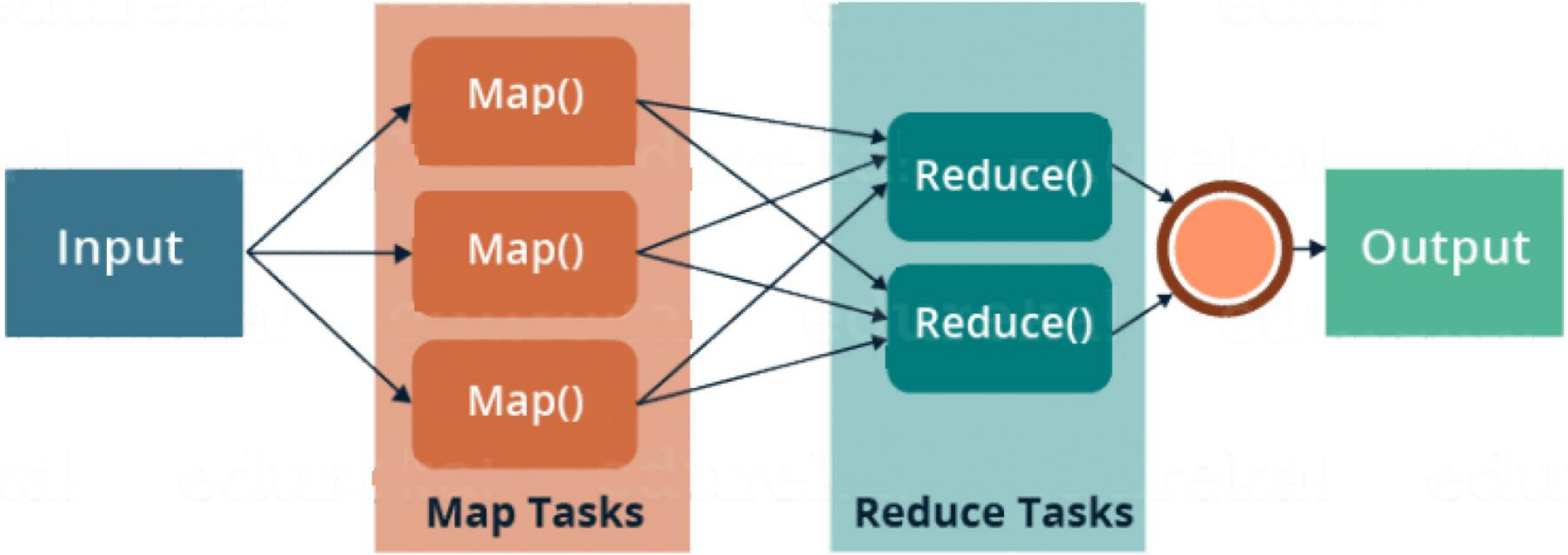


Figure. ***MapReduce data flow with multiple reduce tasks***

Big Data Enabling Technologies

Explanations and Real-Time Applications:

5. YARN (Yet Another Resource Negotiator):

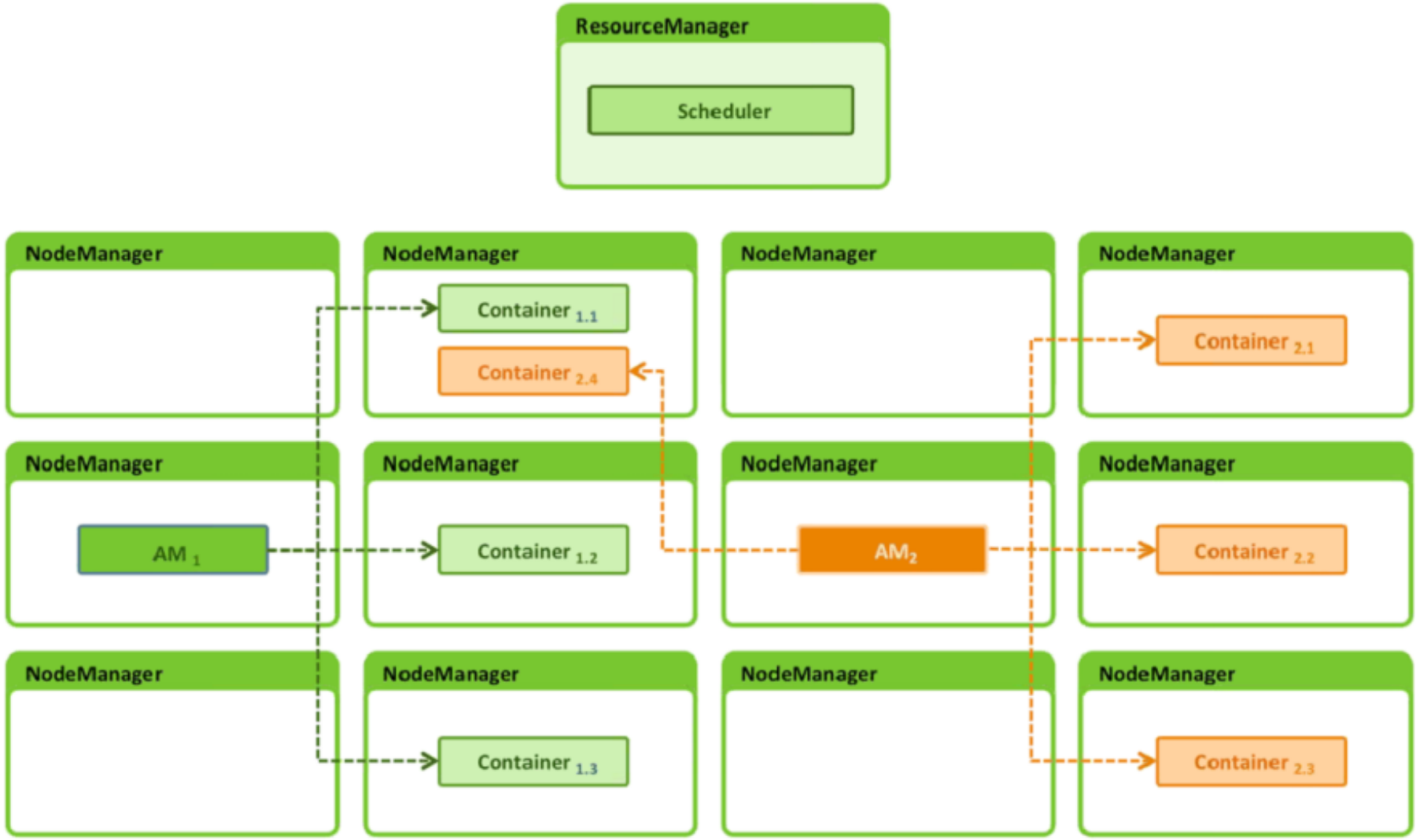
- Manages and schedules resources across Hadoop clusters for efficient processing.

Applications:

- Multi-tenant environments in large organizations.
- Load balancing in cloud services.

Big Data Enabling Technologies

5. YARN (Yet Another Resource Negotiator) Architecture:



AM - Application Master

Big Data Enabling Technologies

Explanations and Real-Time Applications:

6. NoSQL Databases:

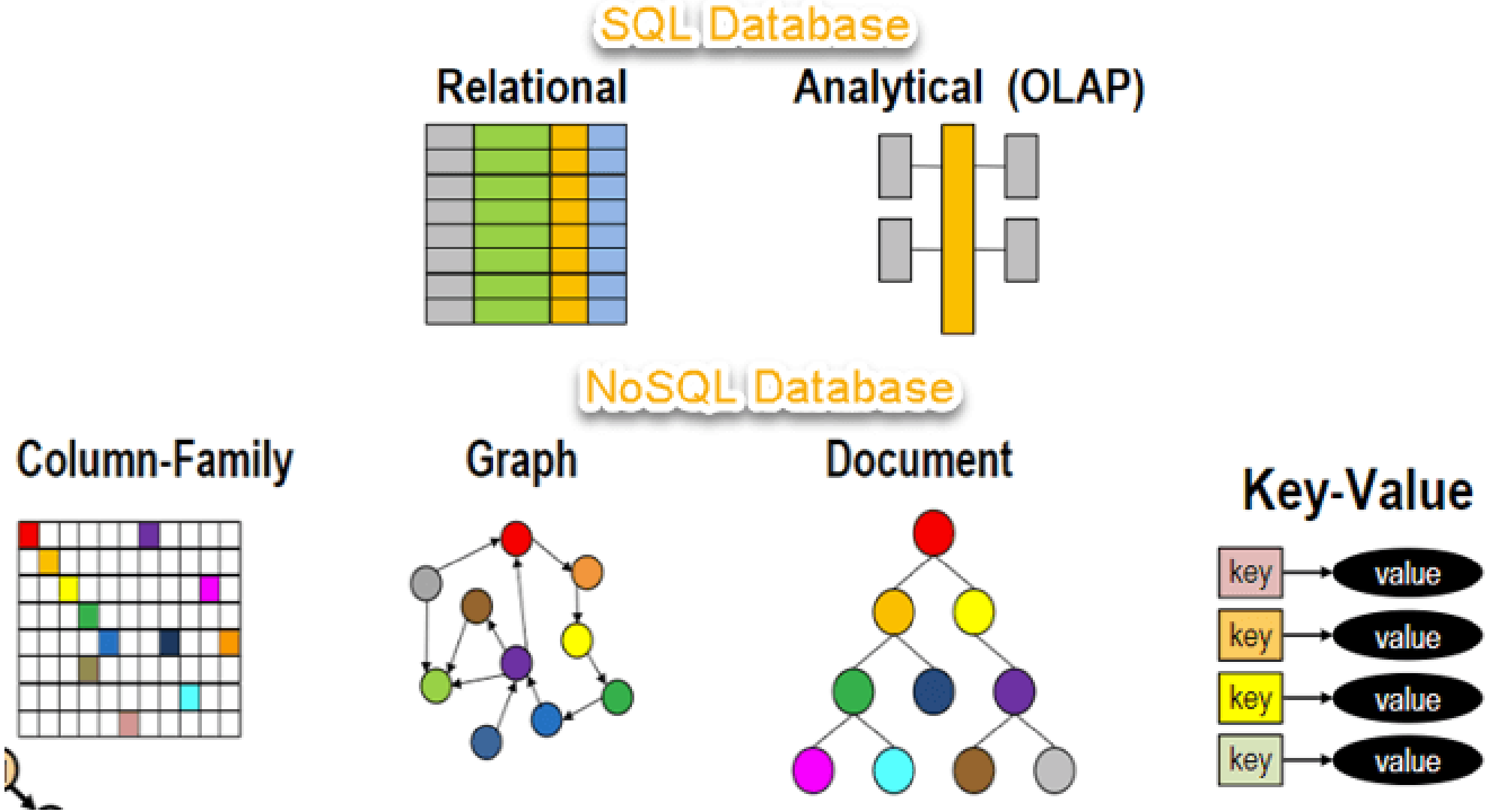
- Databases like Cassandra, HBase, and MongoDB designed to handle unstructured or semi-structured data.

Applications:

- Real-time analytics in IoT.
- Product catalog management in retail.

Big Data Enabling Technologies

6. NoSQL Databases:



Big Data Enabling Technologies

Explanations and Real-Time Applications:

7. Hive:

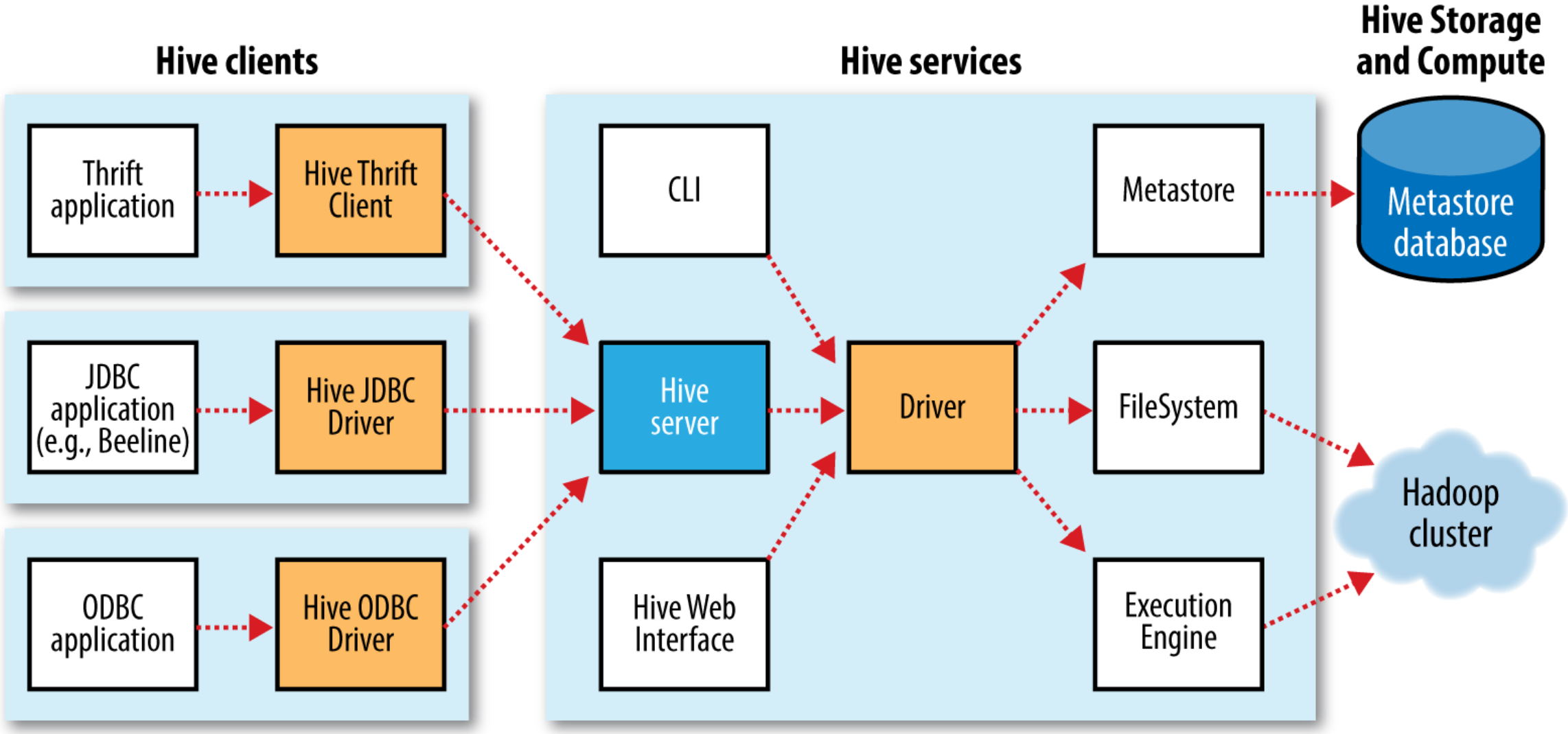
- A data warehouse infrastructure built on Hadoop for querying and analyzing large datasets using SQL-like language.

Applications:

- Data summarization in business intelligence.
- Reporting for ad-hoc queries in telecom.

Big Data Enabling Technologies

7. Hive Architecture:



Big Data Enabling Technologies

Explanations and Real-Time Applications:

8. Apache Spark:

- A unified analytics engine for large-scale data processing with in-memory computation capabilities.

Applications:

- Real-time data analytics in stock markets.
- Fraud detection in credit card transactions.

Big Data Enabling Technologies

8. Apache Spark Architecture:

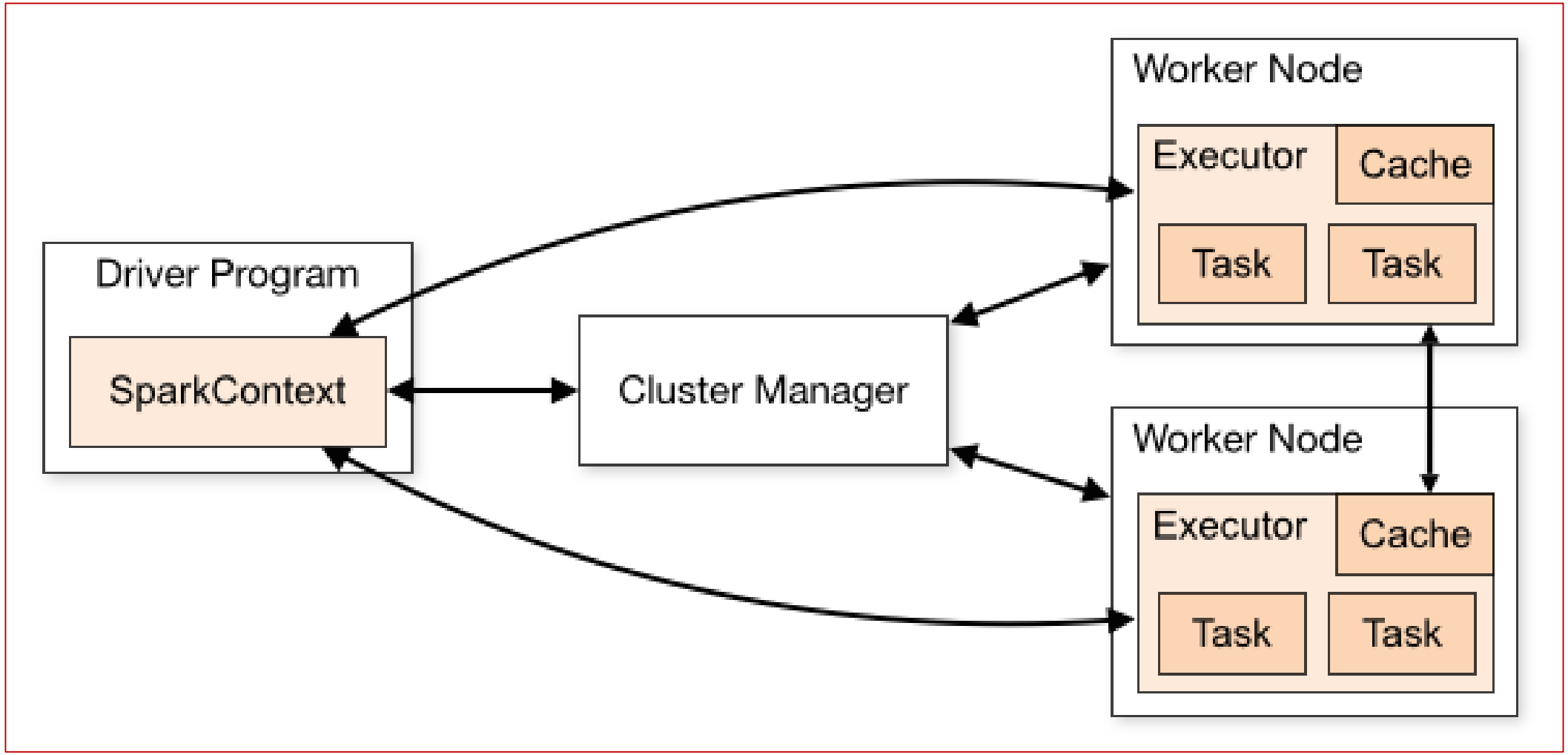


Figure. Apache Spark architecture

Big Data Enabling Technologies

Explanations and Real-Time Applications:

9. ZooKeeper:

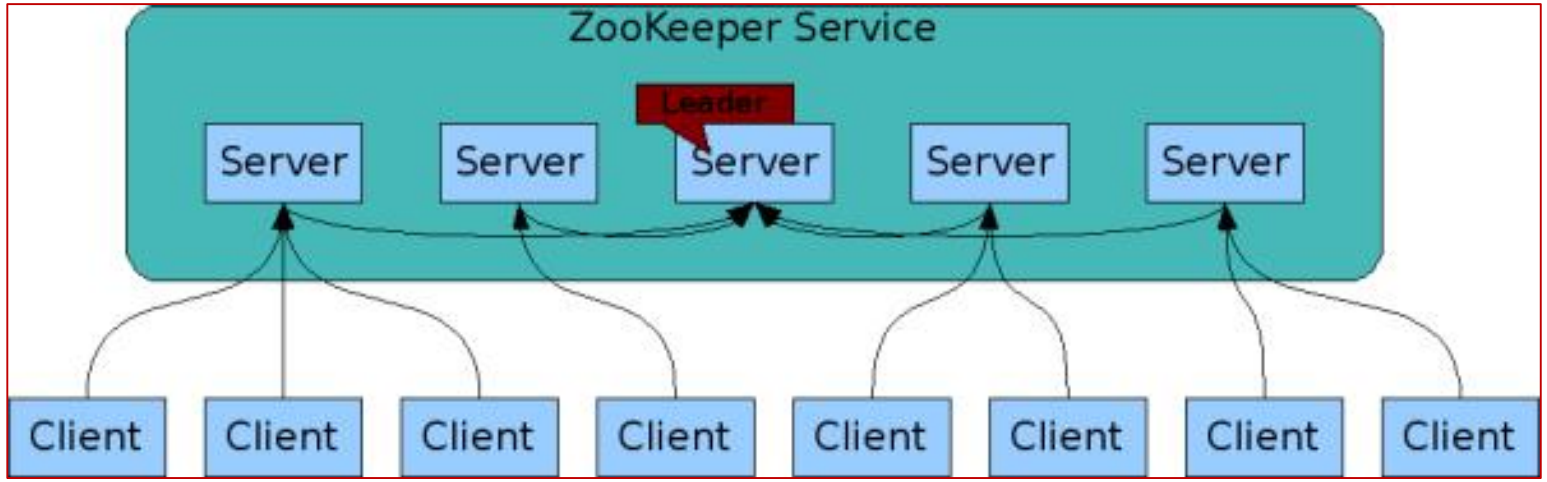
- A centralized service for maintaining configuration information, naming, and synchronization.

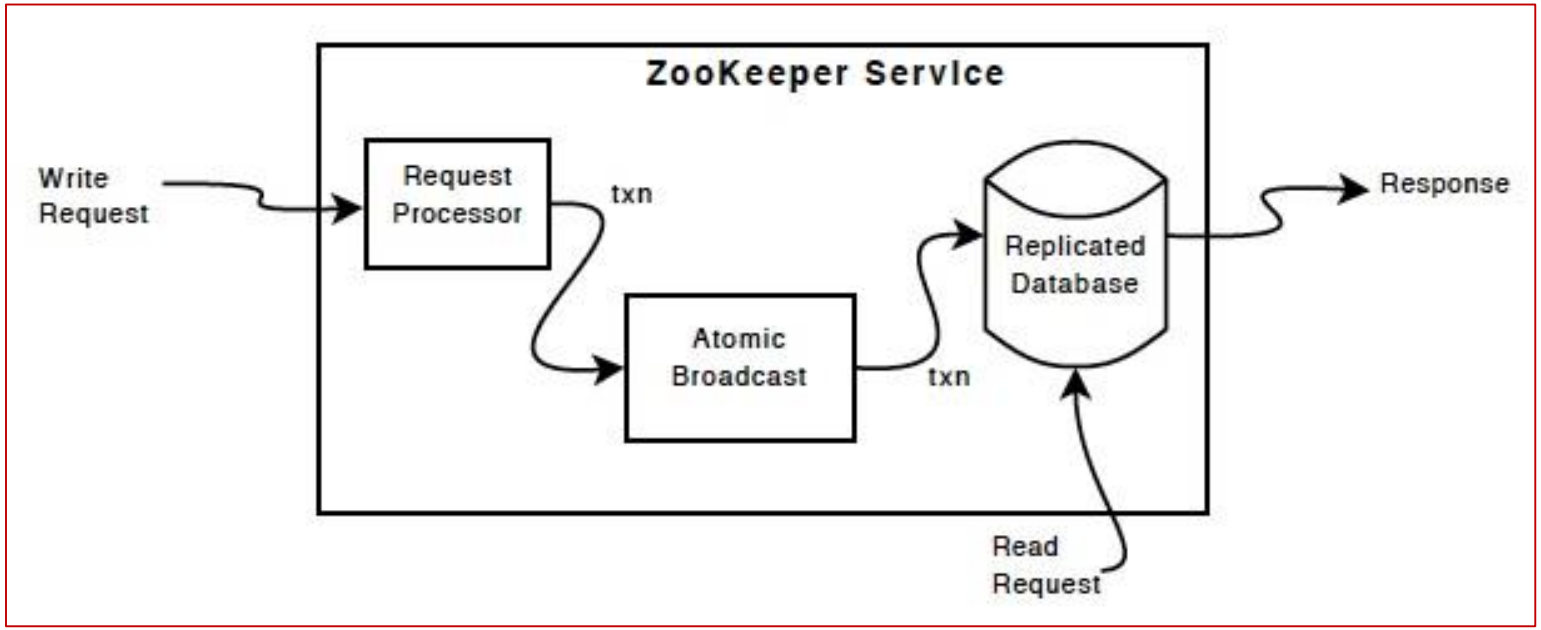
Applications:

- Coordination of microservices in distributed systems.
- Managing leader election in clusters.

Big Data Enabling Technologies

9. ZooKeeper:





Big Data Enabling Technologies

Explanations and Real-Time Applications:

10. Cassandra:

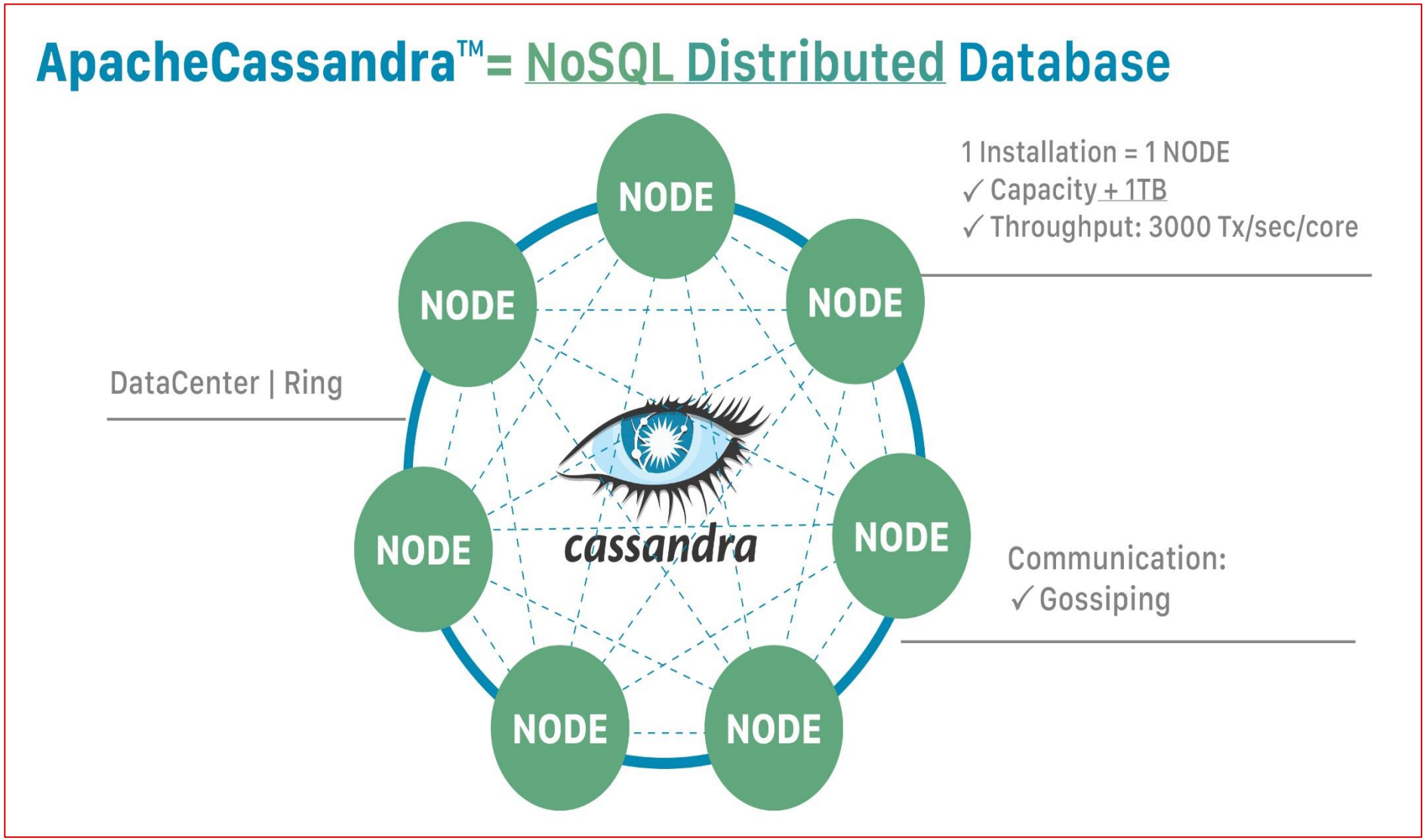
- A NoSQL database that provides high availability and scalability for large datasets.

Applications:

- User activity tracking in social media platforms.
- Time-series data management in IoT.

Big Data Enabling Technologies

10. Cassandra:



Big Data Enabling Technologies

Explanations and Real-Time Applications:

11. HBase:

- A NoSQL database that runs on top of HDFS for random, real-time read/write access to large datasets.

Applications:

- Processing geospatial data.
- Storage for online transaction processing (OLTP) systems.

Big Data Enabling Technologies

11. HBase:

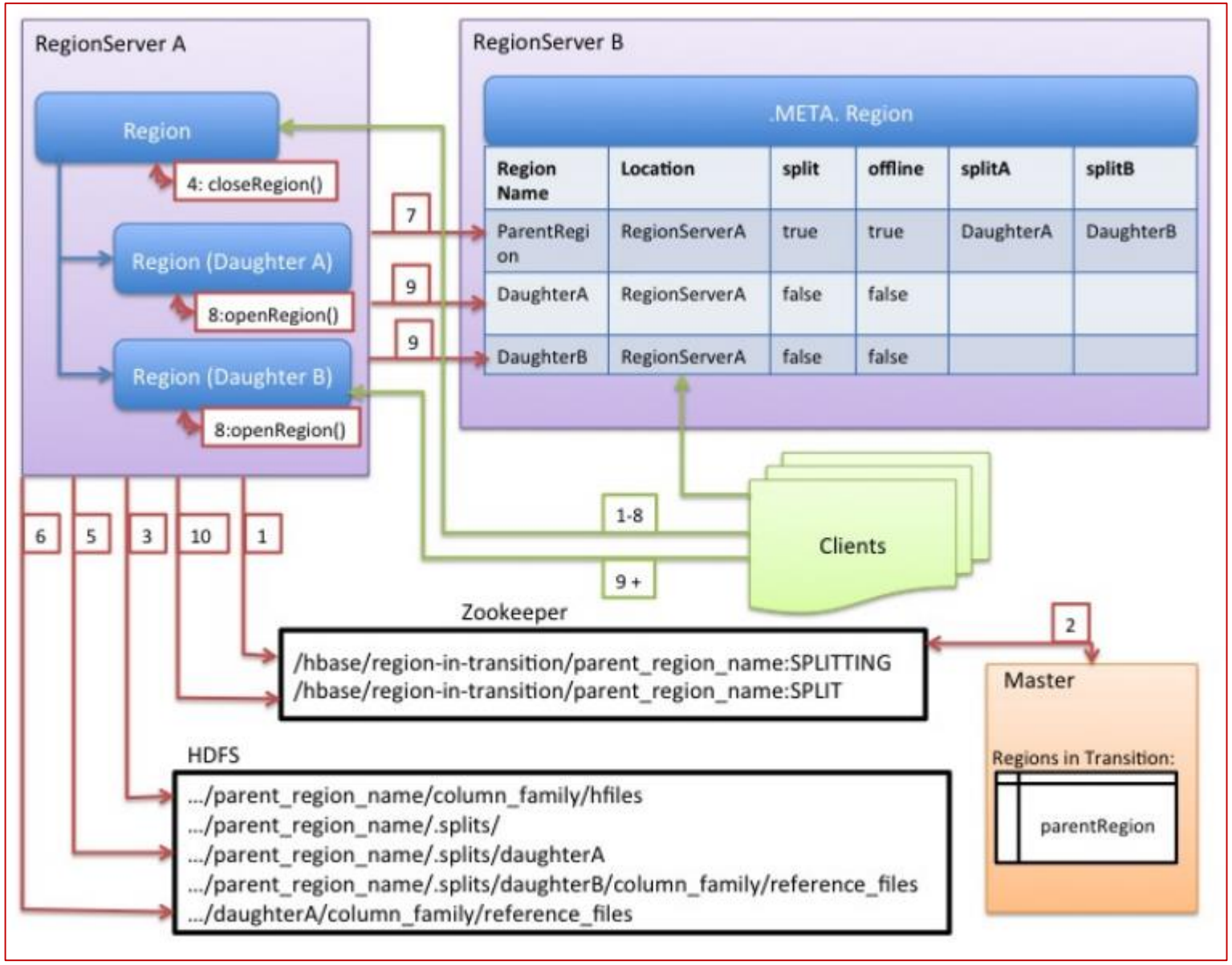


Figure. RegionServer Split Process

Big Data Enabling Technologies

Explanations and Real-Time Applications:

12. MongoDB:

- A document-oriented NoSQL database for flexible and scalable data storage.

Applications:

- Content management systems.
- Catalog data in retail and e-commerce.

Big Data Enabling Technologies

12. MongoDB:

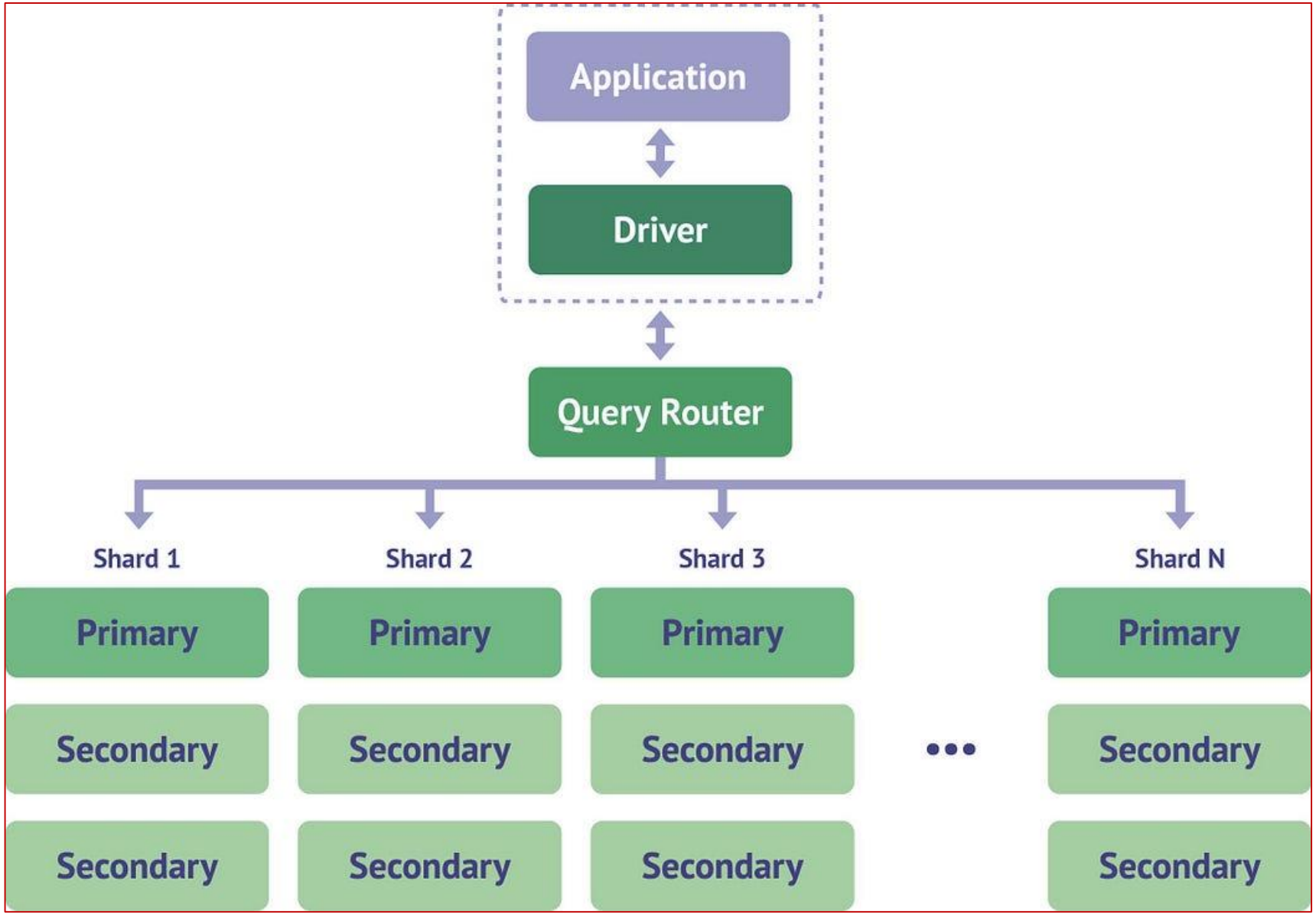


Figure. MongoDB Architecture

Big Data Enabling Technologies

Explanations and Real-Time Applications:

13. Spark Streaming:

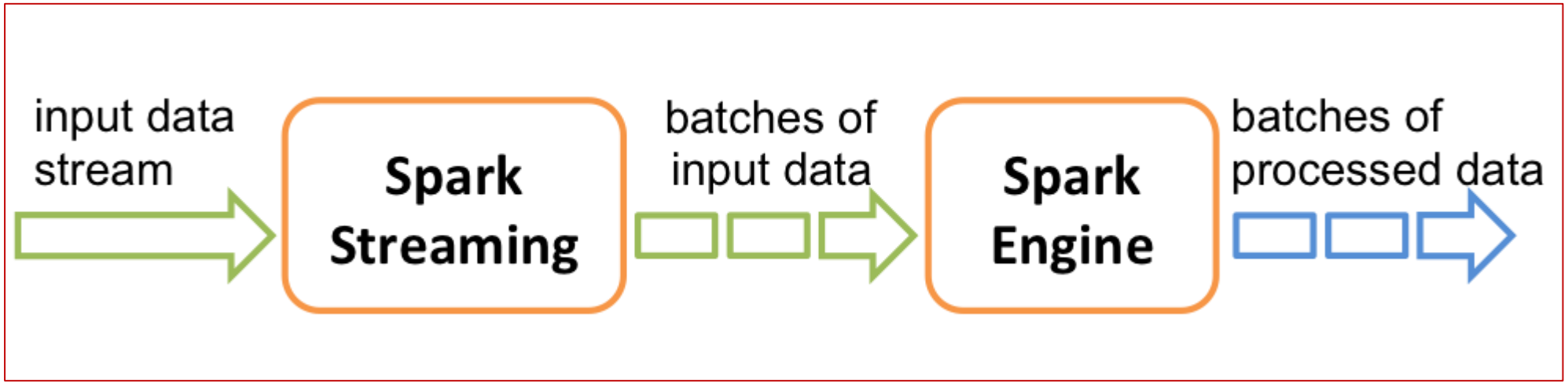
- A component of Apache Spark for processing real-time streaming data.

Applications:

- Real-time sentiment analysis on social media.
- Monitoring network traffic for cybersecurity.

Big Data Enabling Technologies

13. Spark Streaming:



Big Data Enabling Technologies

Explanations and Real-Time Applications:

14. Apache Kafka:

- A distributed messaging system for handling real-time data feeds.

Applications:

- Event-driven architecture in e-commerce platforms.
- Log aggregation in large-scale systems.

Big Data Enabling Technologies

14. Apache Kafka:

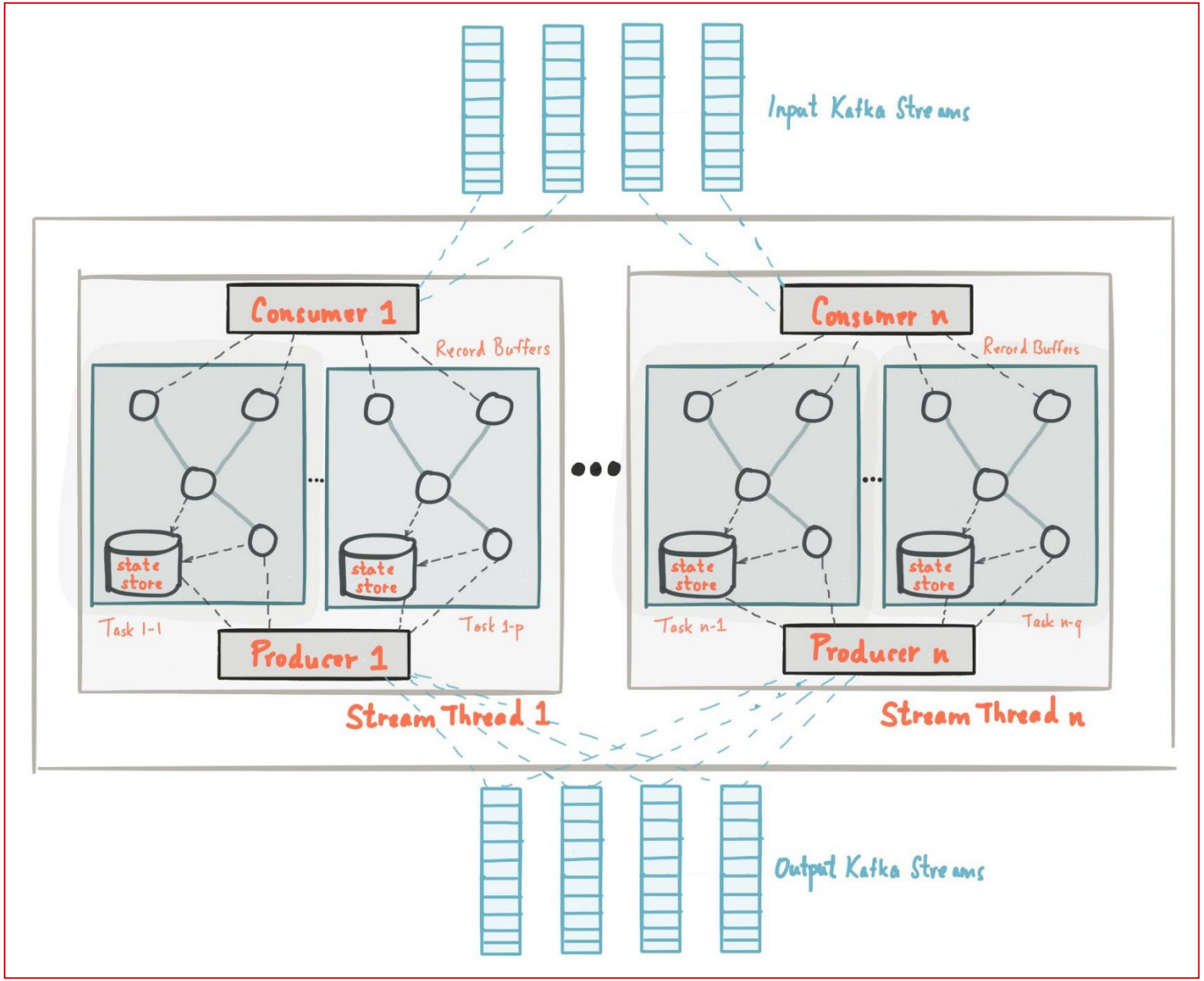


Fig: The picture shows the anatomy of an application that uses the Kafka Streams library

Big Data Enabling Technologies

Explanations and Real-Time Applications:

15. Apache Flume:

- Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data.
- It has a simple and flexible architecture based on streaming data flows.
- It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms.
- It uses a simple extensible data model that allows for online analytic application.

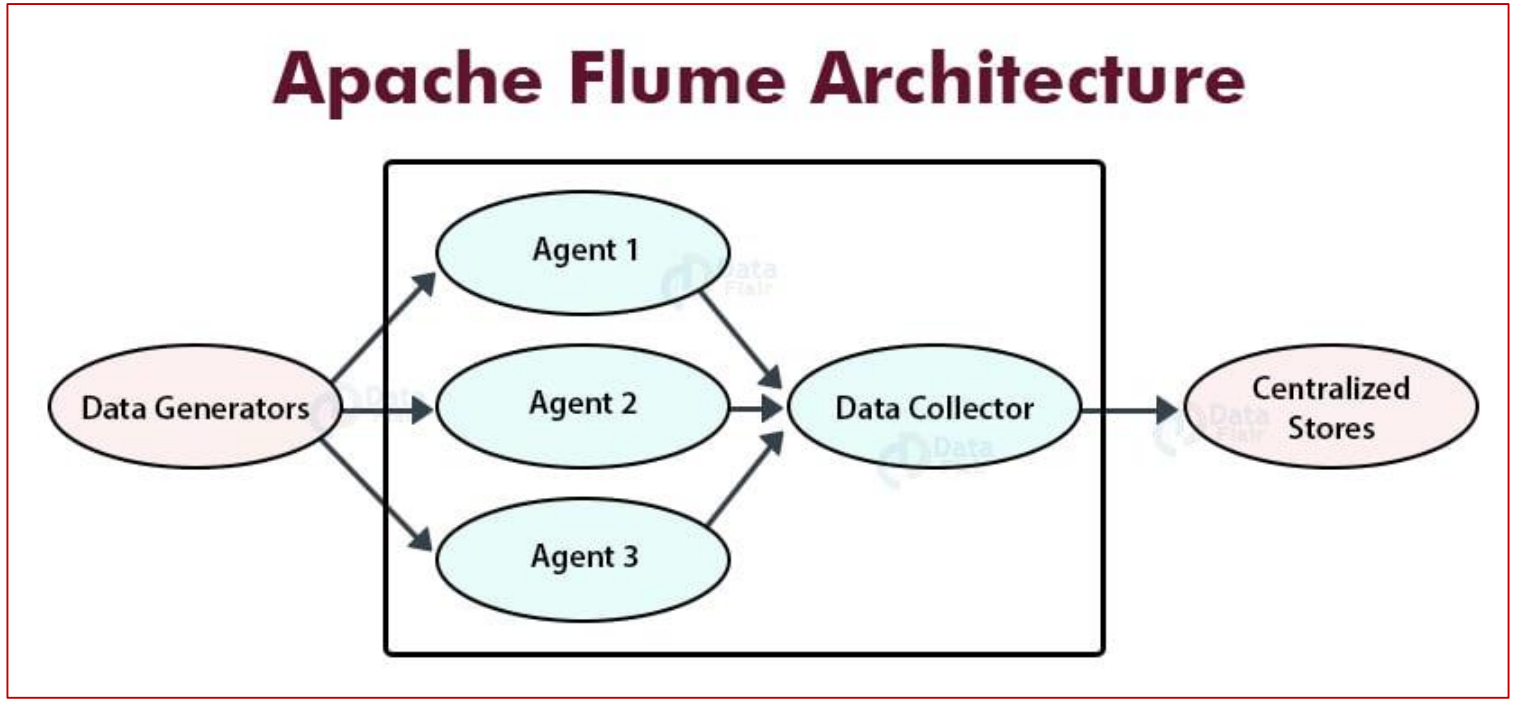
Applications:

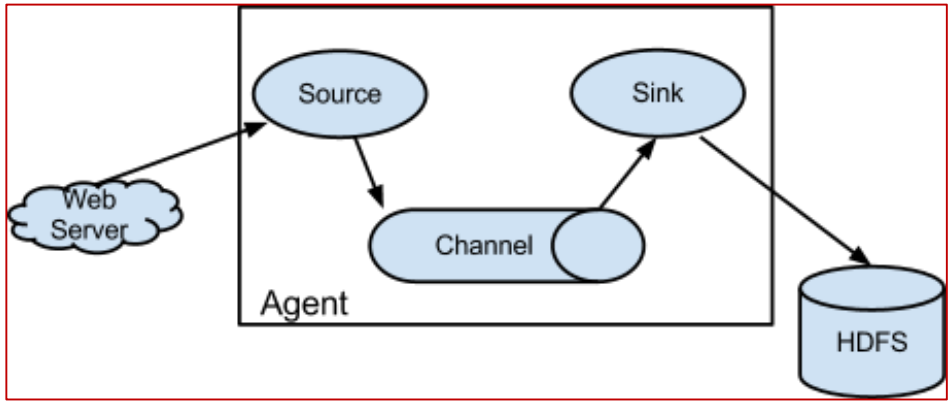
- Real-time log analysis in IT systems.
- Streaming data ingestion for analytics.

Big Data Enabling Technologies

15. Apache Flume:

It has a simple and flexible architecture based on streaming data flows.





Big Data Enabling Technologies

Explanations and Real-Time Applications:

16. Spark MLlib:

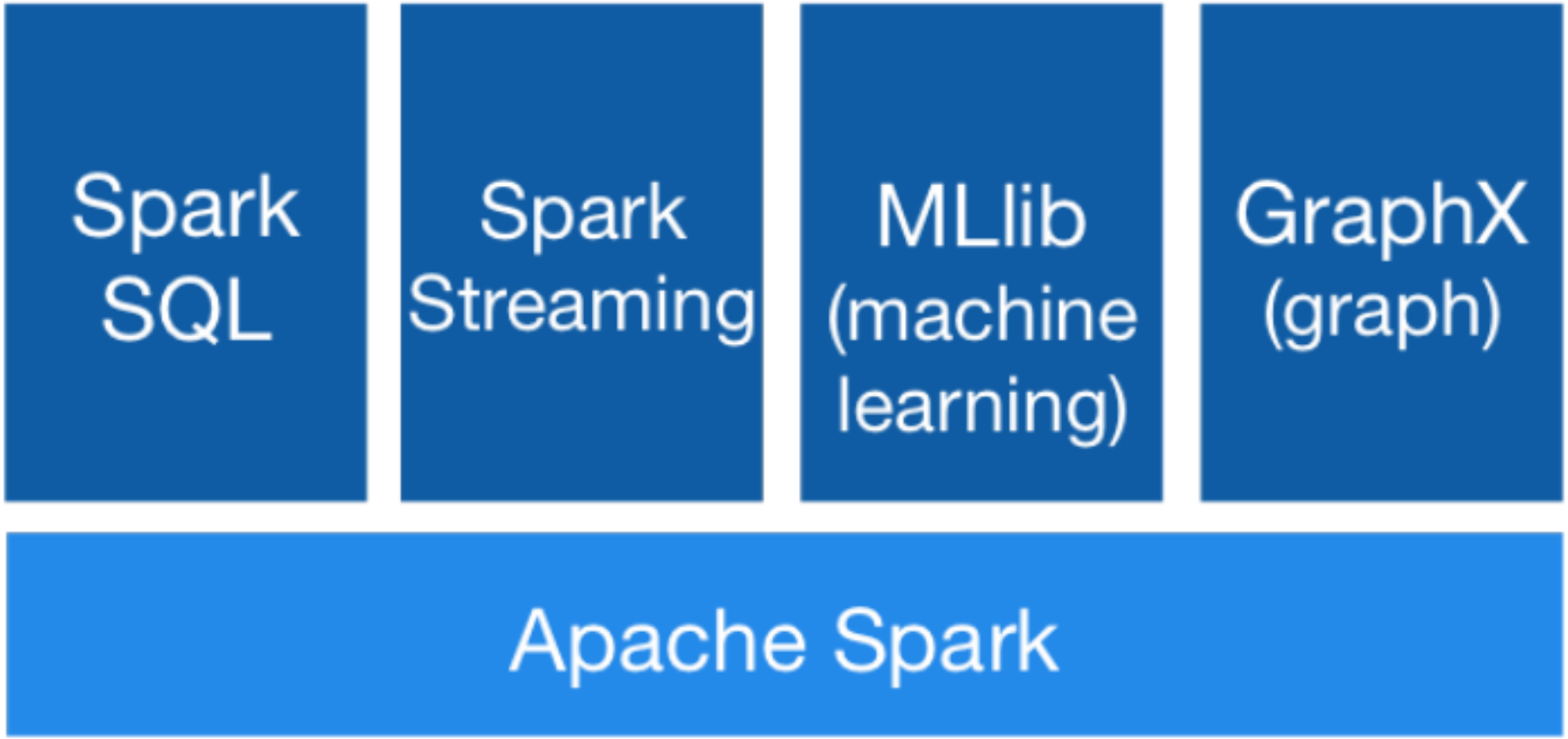
- A machine learning library in Spark for scalable ML algorithms.

Applications:

- Predictive analytics in healthcare.
- Recommendation systems in streaming platforms.

Big Data Enabling Technologies

16. Spark MLlib:



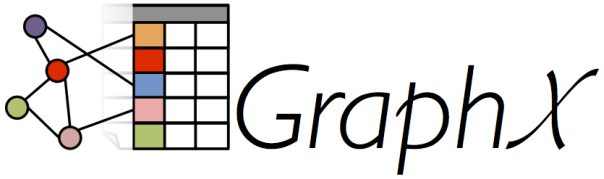
Big Data Enabling Technologies

16. Spark MLlib Components:

Component	Description	Examples	Applications
Data Types	APIs for working with data using RDDs or DataFrames.	RDDs, DataFrames	Switching to DataFrames for better performance and simplified code management.
Classification & Regression	Predictive models for labeled data.	Logistic Regression, Random Forests, Gradient Boosted Trees	Spam detection, loan approval, housing price prediction.
Clustering	Unsupervised grouping of data points.	K-Means, Gaussian Mixture Models	Customer segmentation, image compression, anomaly detection.
Collaborative Filtering	Recommendation systems based on user-item interactions.	Alternating Least Squares (ALS)	Product recommendations, movie suggestions, personalized advertising.
Dimensionality Reduction	Reduces data dimensionality for easier computation.	Principal Component Analysis (PCA), Singular Value Decomposition (SVD)	Feature selection, visualization, noise reduction.
Pipelines	Combines data processing and model building into reusable workflows.	Transformers, Estimators	End-to-end machine learning pipelines.
Feature Extraction & Transformation	Tools for preparing data for machine learning.	Tokenization, TF-IDF (Term Frequency-Inverse Document Frequency), Word2Vec, StandardScaler	Sentiment analysis, text classification, preprocessing for numerical datasets.
Evaluation Metrics	Methods to measure model performance.	Precision, Recall, F1 Score, MSE (Mean Squared Error), RMSE (Root Mean Squared Error)	Model tuning, selecting the best model for deployment.
Model Persistence	Saving and loading machine learning models.	Save/Load APIs	Reusing trained models for batch and real-time predictions.
Distributed Linear Algebra	Scalable matrix and vector operations for ML algorithms.	Matrix Factorization	Graph processing, large-scale simulations, recommendation systems.
Graph Analytics (GraphX)	Integration for graph-based learning algorithms.	PageRank, Connected Components	Social network analysis, influence propagation, community detection.
Streaming & Real-time ML	Machine learning on real-time data streams.	Spark Streaming with MLlib	Fraud detection, dynamic pricing, real-time personalization.

Big Data Enabling Technologies

Explanations and Real-Time Applications:



17. GraphX:

- A distributed graph-processing framework in Spark.

Applications:

- Social network analysis.
- Fraud detection in financial transactions.

Big Data Enabling Technologies

17. GraphX Components:

Component	Description	Examples	Applications
Graph Representation	Represents graphs as RDD-based abstractions (Graph and Edge).	Nodes (vertices), Edges	Social networks, flight routes.
Property Graph	Allows vertices and edges to have attributes (properties).	Vertex attributes: user data; Edge attributes: relationships	Enriched data modeling for graphs.
Graph Operators	Provides operations to manipulate graphs.	Subgraph, joinVertices, aggregateMessages	Filtering graphs, computing graph metrics.
Pregel API	Implements vertex-centric iterative algorithms for parallel processing.	PageRank, Connected Components	Ranking pages in search engines, finding clusters in social graphs.
Built-in Algorithms	Predefined algorithms for common graph problems.	PageRank, Triangle Counting, Shortest Paths	Web link analysis, fraud detection, shortest route optimization.
Graph Construction	Tools for creating graphs from RDDs or loading external datasets.	From RDDs, files, or existing libraries	Building graphs from CSV data or edge-list files.
Graph Queries	Supports querying of graphs for patterns and metrics.	Counting edges, filtering vertices	Analyzing user interactions, detecting anomalies.
Integration with MLlib	Combines machine learning and graph processing for advanced analytics.	Label Propagation for community detection	Social influence modeling, collaborative filtering.
Fault Tolerance	Provides recovery mechanisms for distributed graph processing.	Checkpointing	Ensuring stability of long-running graph computations.
Visualization Support	Facilitates exporting graphs for visualization.	Export to Gephi, Neo4j	Visualizing relationships and networks.

Big Data Enabling Technologies

Explanations and Real-Time Applications:

18. Apache Oozie:

- A workflow scheduler for managing Hadoop jobs.

Applications:

- Automating data pipeline workflows.
- Orchestrating complex data analytics tasks.

Big Data Enabling Technologies

18. Apache Oozie:

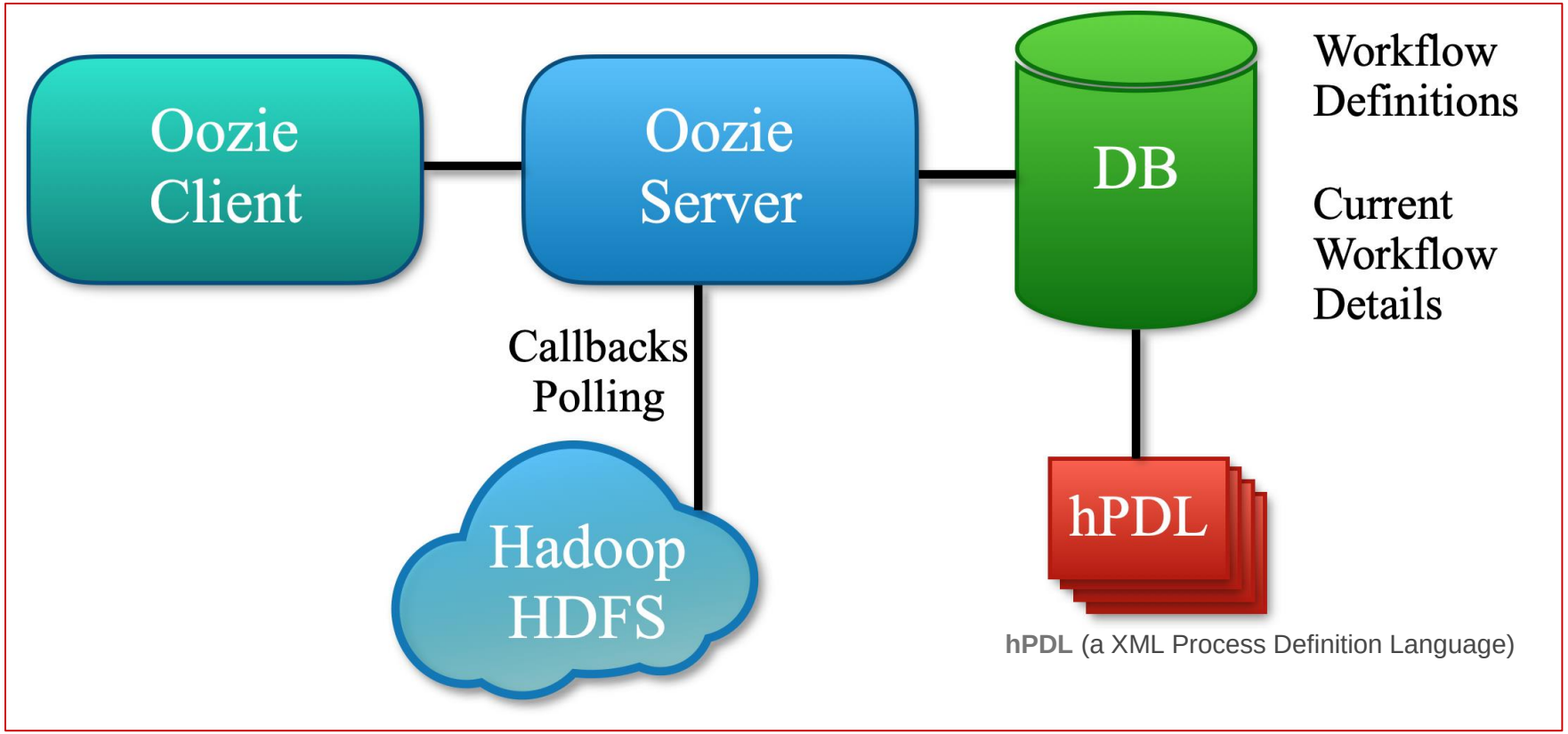


Fig: architecture of Apache Oozie

Big Data Enabling Technologies

Explanations and Real-Time Applications:

19. Apache Ambari:

- A web-based tool for provisioning, monitoring, and managing Hadoop clusters.

Applications:

- Cluster health monitoring in enterprise environments.
- Simplified cluster setup for Big Data platforms.

Big Data Enabling Technologies

19. Apache Ambari Architecture:

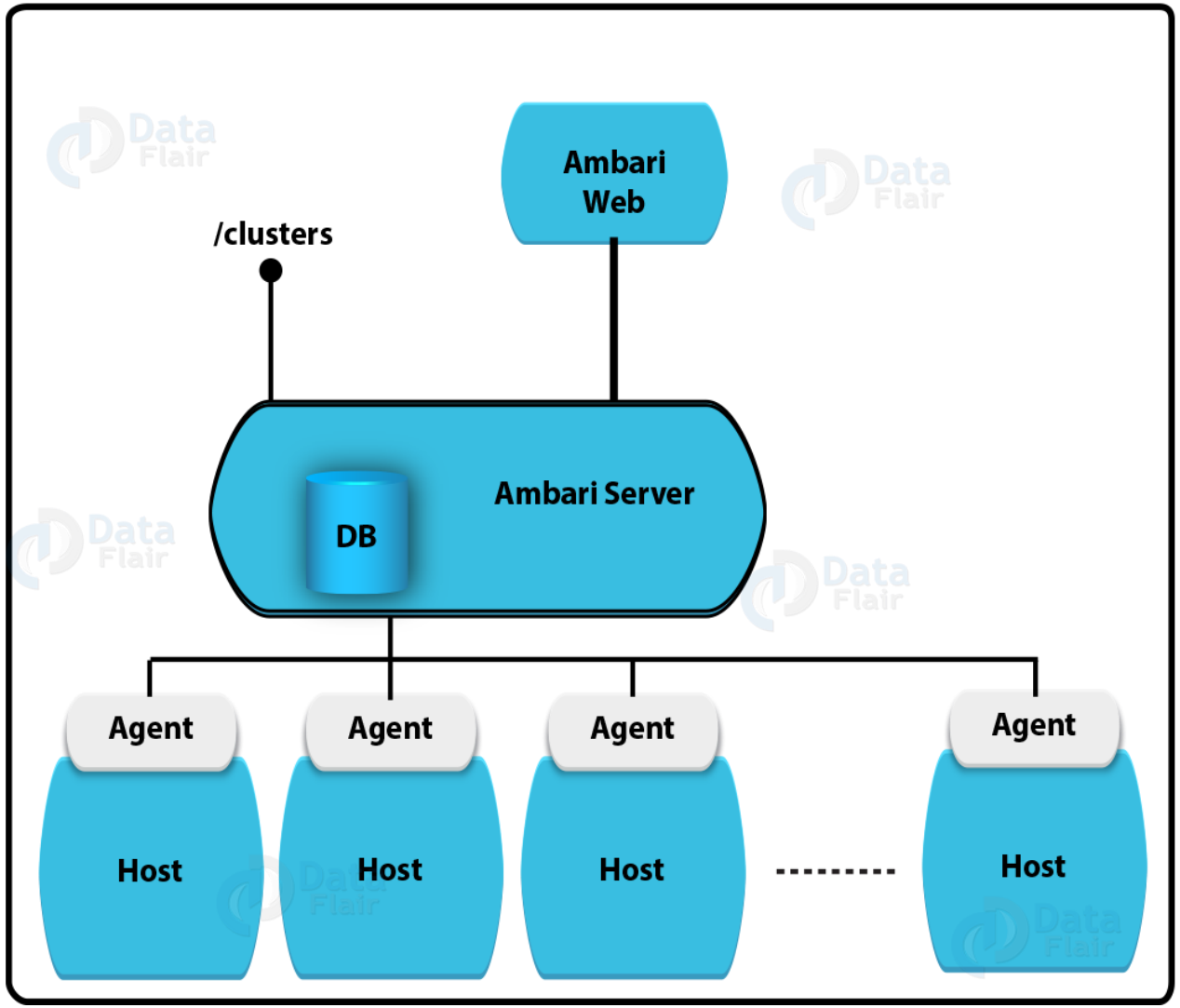


Fig: Apache Ambari Architecture
Src: DataFlair

Big Data Enabling Technologies

Explanations and Real-Time Applications:

20. Apache Pig:

- A high-level platform for creating programs that run on Hadoop, designed to handle large-scale data manipulation.

Applications:

- Data preprocessing for analytics.
- Log file analysis for IT operations.

Big Data Enabling Technologies

20. Apache Pig Architecture: