APACHE SPARK

by Priti Bhardwaj , CDAC NOIDA, India

Real Time Analytics

Banking

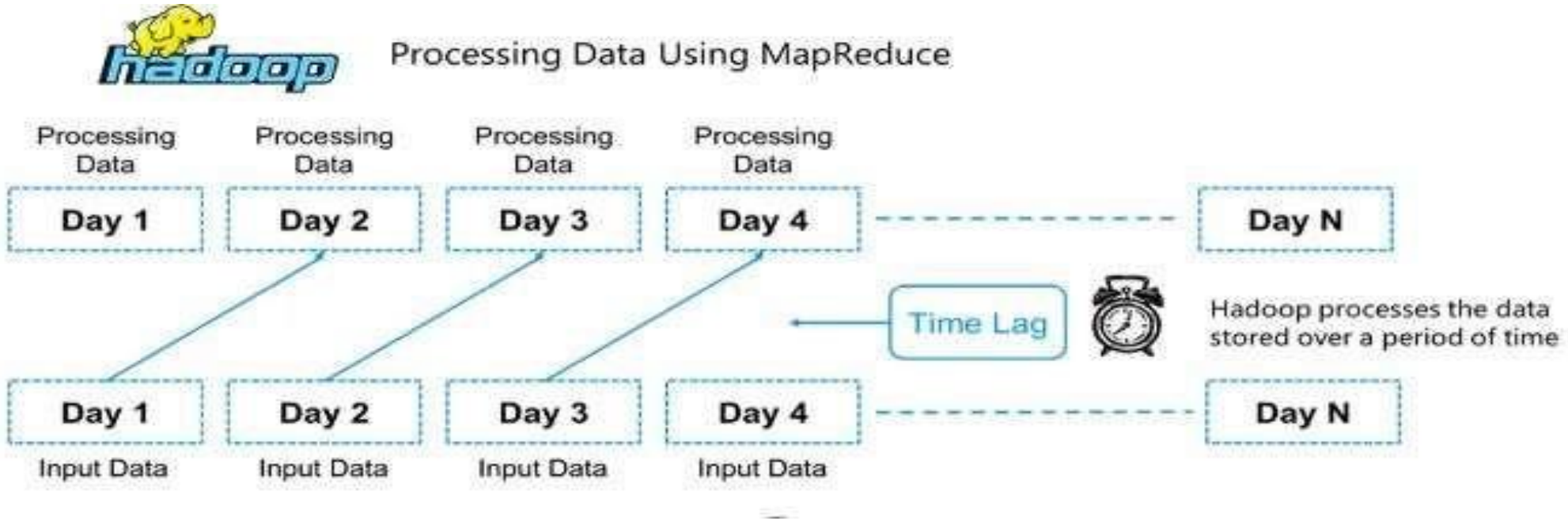
Government

Healthcare

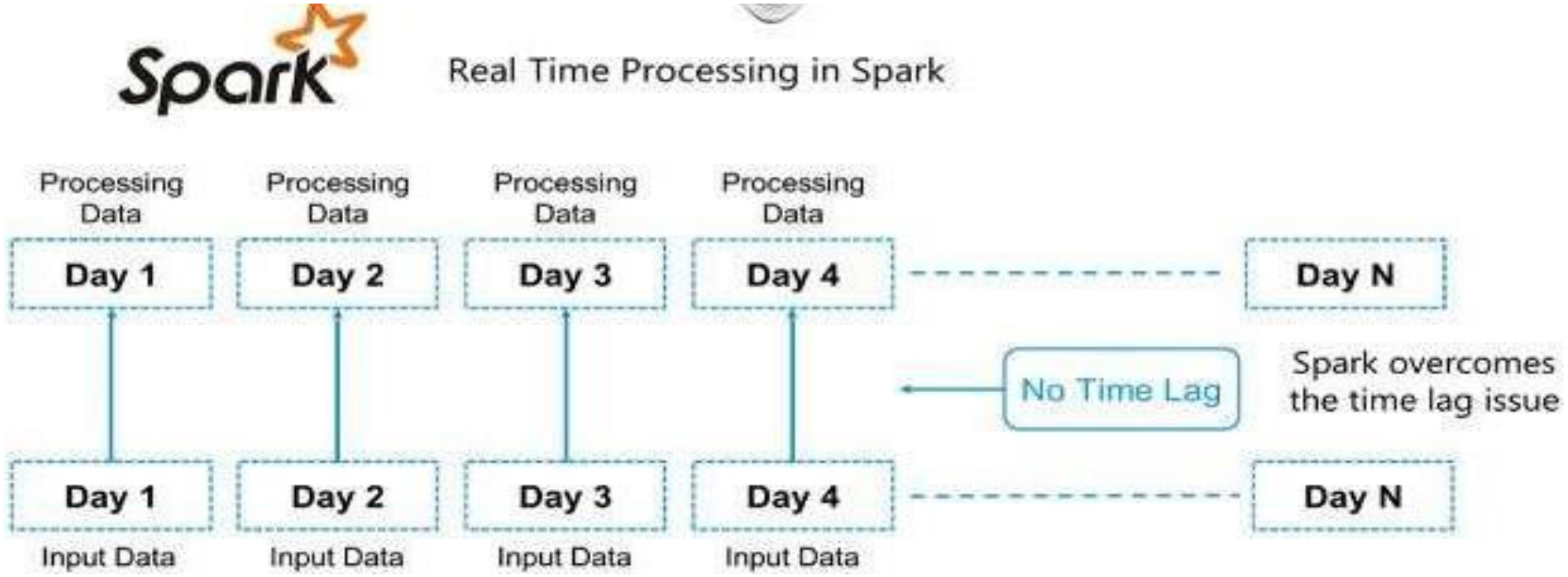
Telecommunications

Stock Market

Why Spark when Hadoop is already there?



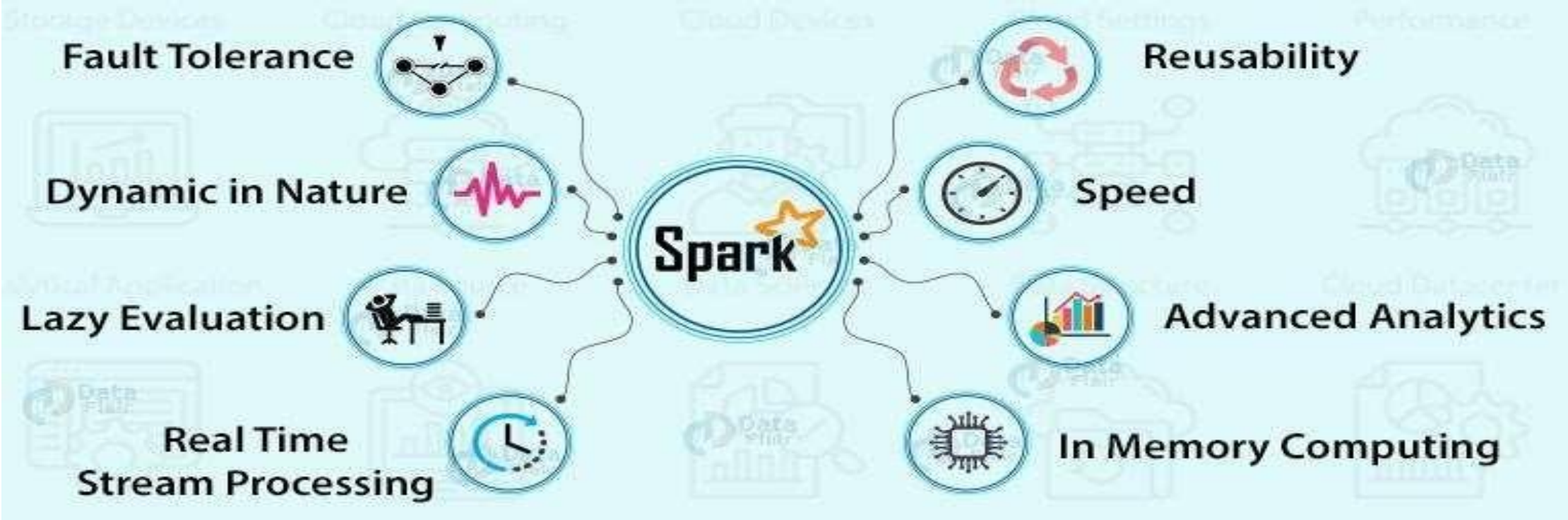
Why Spark when Hadoop is already there?



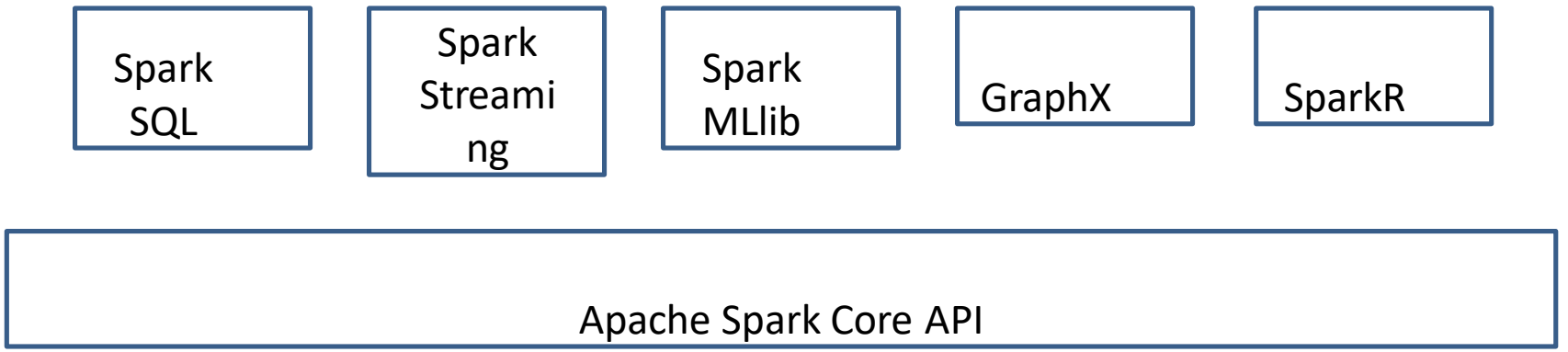
What is Spark?

- Apache Spark is an open source cluster computing framework for real-time data processing.
- Main feature of Apache Spark : in-memory cluster computing that increases the processing speed of an application.
- Spark provides an interface for programming entire clusters with implicit data parallelism and fault tolerance.

Features of Apache Spark



Spark Components



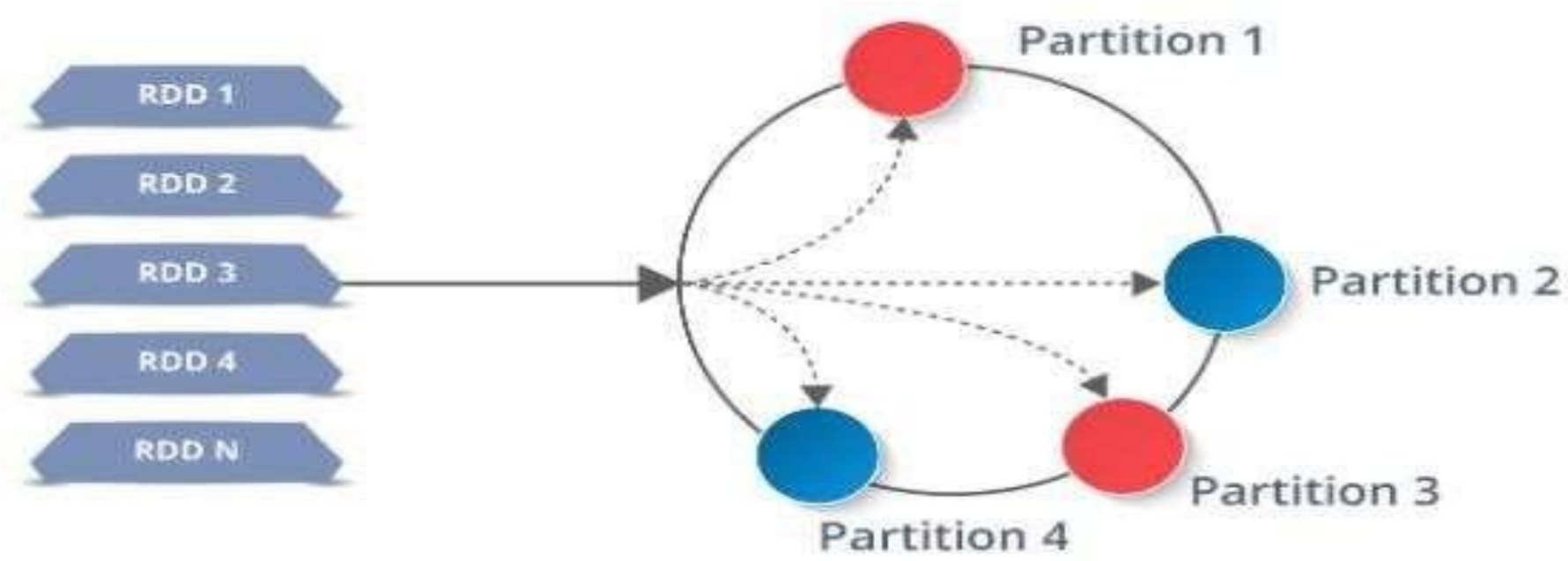
Spark Deployment Modes

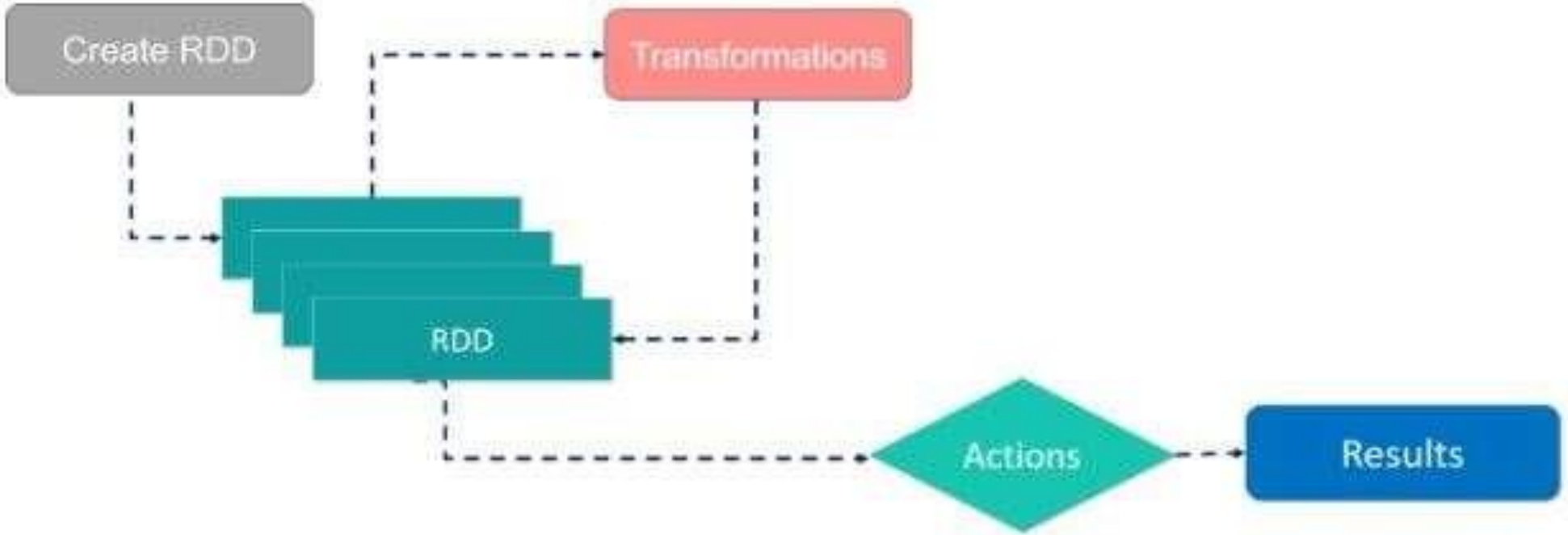
- Standalone (used for learning & development)
- Local mode (used for learning & development)
- Cluster mode (can work with MESOS or YARN)

Resilient Distributed Dataset(RDD)

- RDDs are the building blocks of any Spark application. RDDs Stands for:

- **Resilient:** Fault tolerant and is capable of rebuilding data on failure
- **Distributed**: Distributed data among the multiple nodes in a cluster
- **Dataset:** Collection of partitioned data with values


RDD 1
RDD 2
RDD 3
RDD 4
RDD N
Partition 1
Partition 2
Partition 3
Partition 4

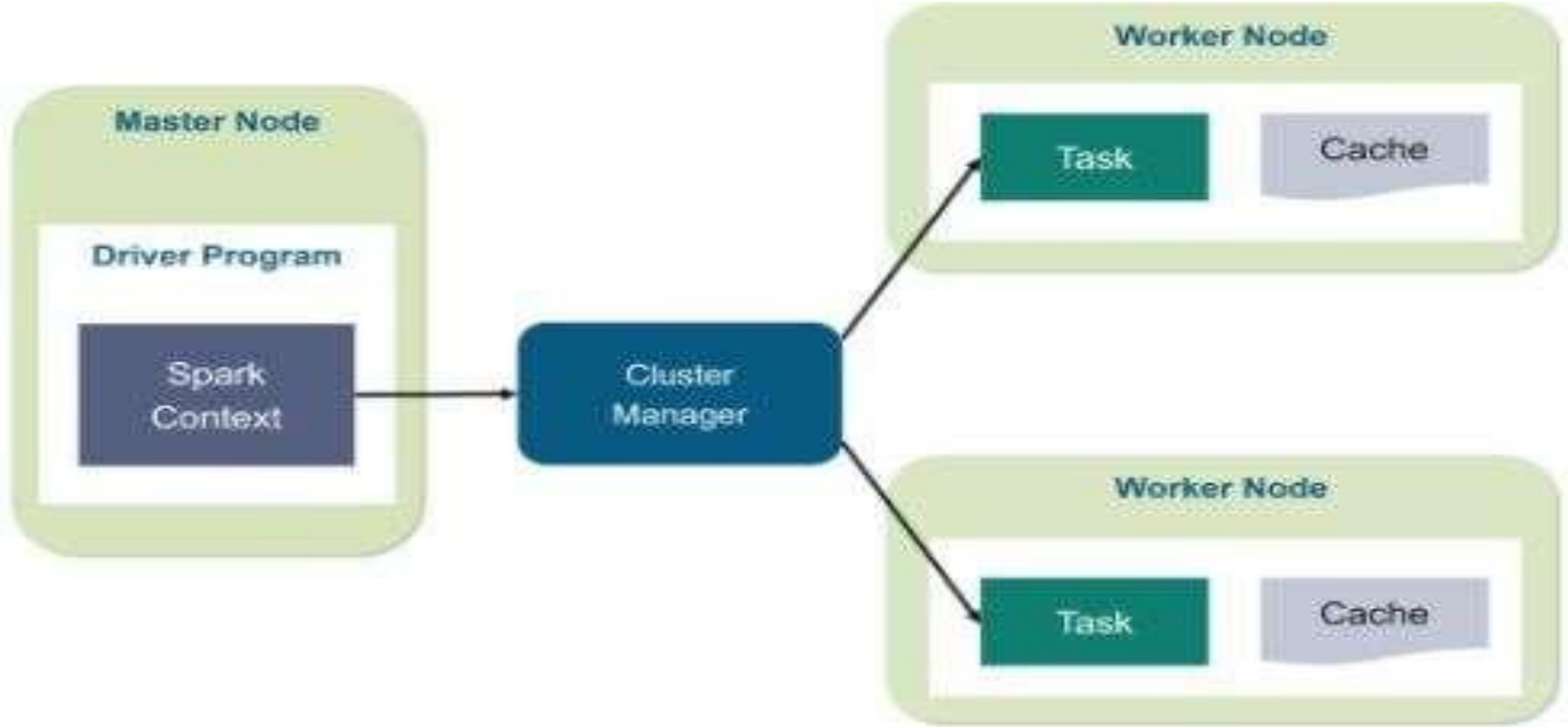

Create RDD
Transformations
RDD
Actions
Results

Resilient Distributed Dataset(RDD)

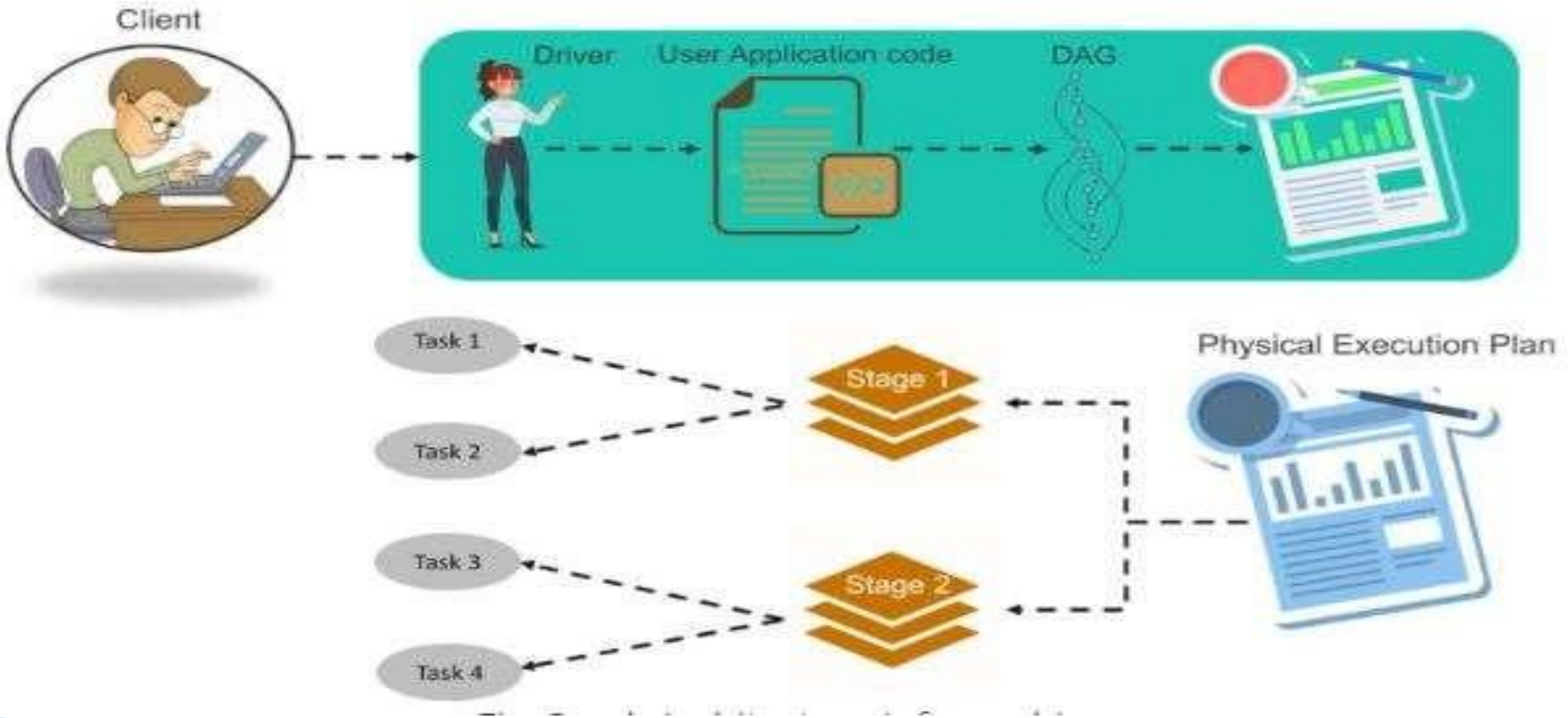
Two types of operations:

- **Transformations:** They are the operations that are applied to create a new RDD.
- **Actions**: They are applied on an RDD to instruct Apache Spark to apply computation and pass the result back to the driver.

Working of Spark Architecture



Working of Spark Architecture



Working of Spark Architecture

How RDD solves the problem ?

by Priti Bhardwaj , CDAC NOIDA, India

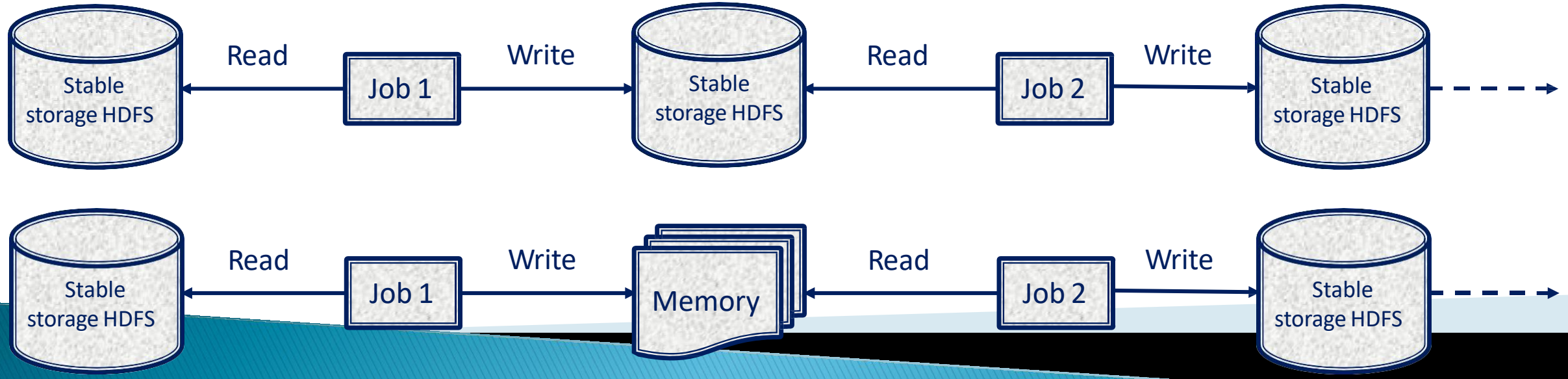
How Spark RDD solve the problem?

- There Exists a Synchronization Barrier between Map and Reduce Tasks

- Data Sharing is slow with Hadoop MapReduce - Spark RDDs solve this.

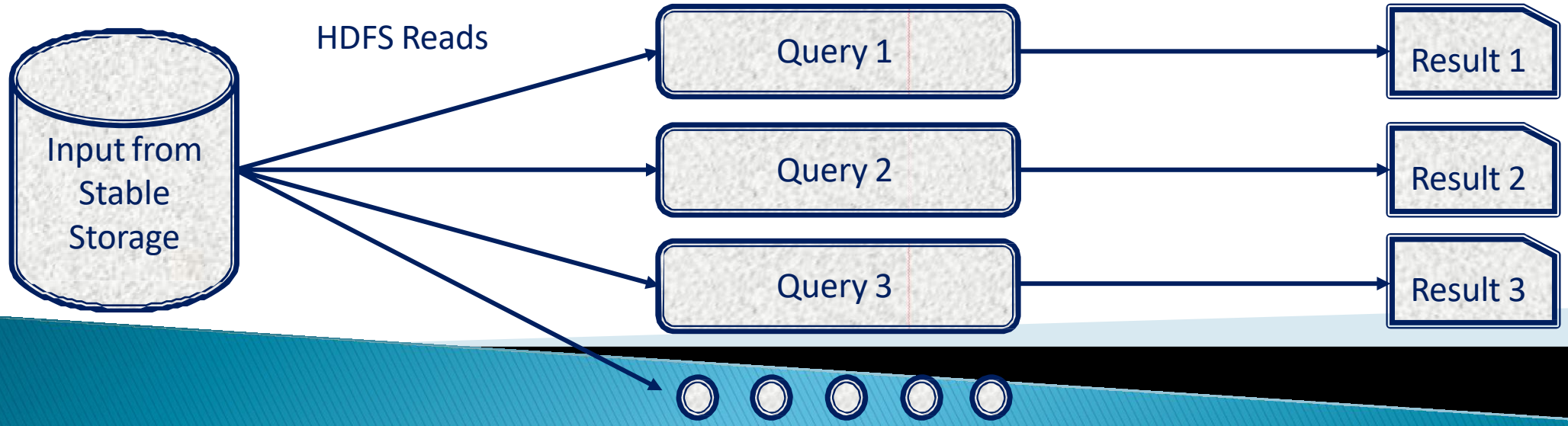
How Spark RDD solve the problem?

- Iterative Operations on Hadoop MapReduce and on Spark RDDs



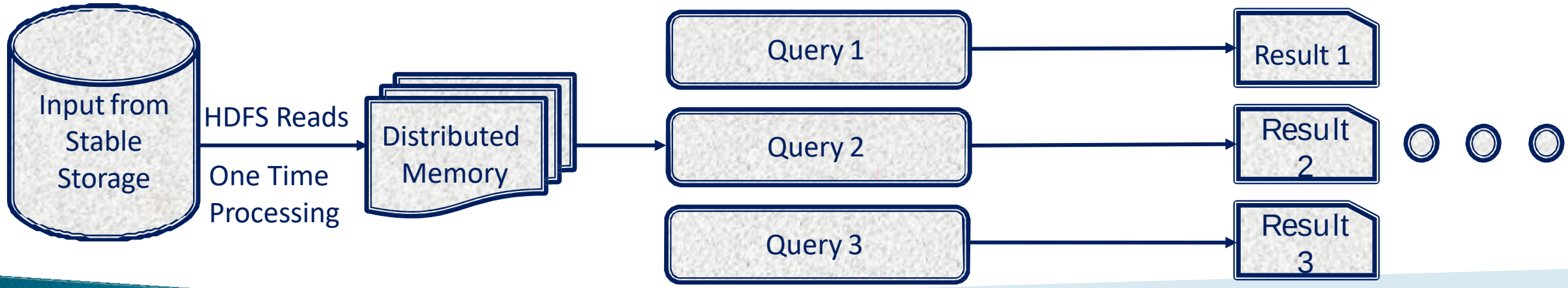
How Spark RDD solve the problem?

- Interactive Operations on Hadoop MapReduce



How Spark RDD solve the problem?

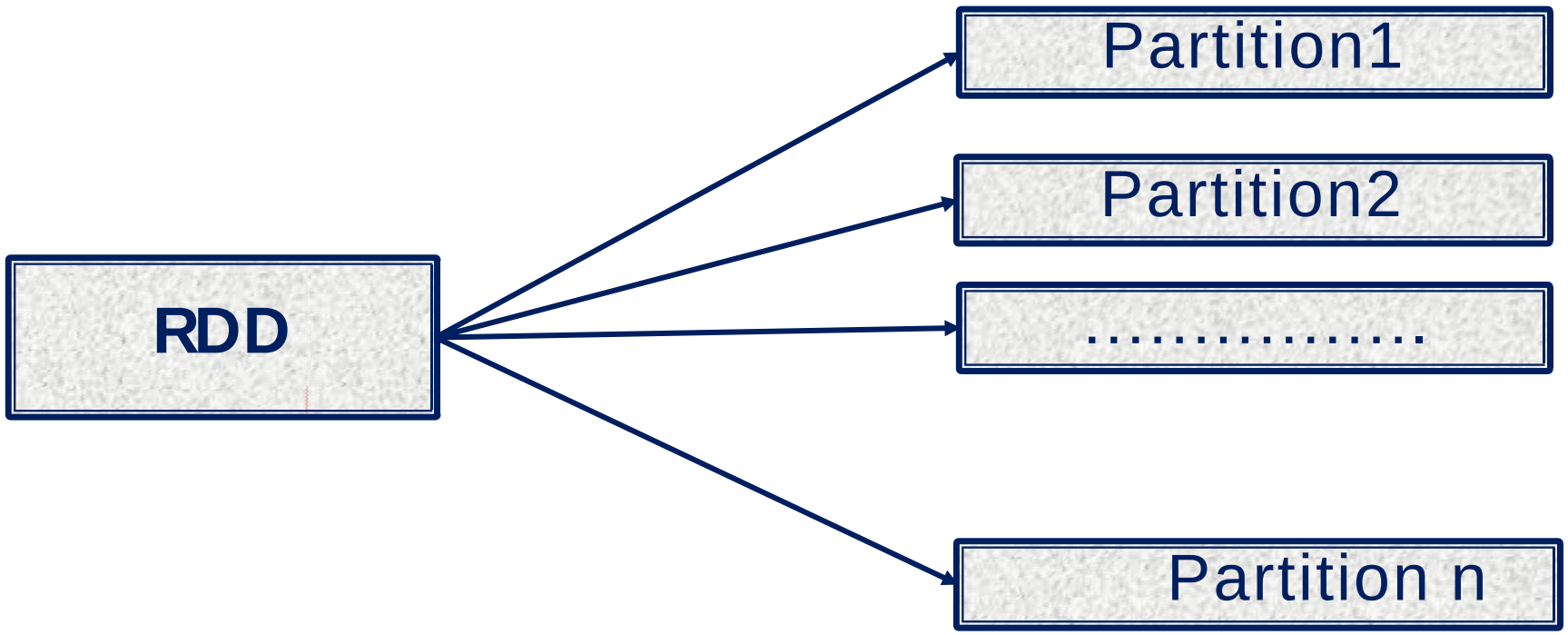
- Interactive Operations on Spark RDDs



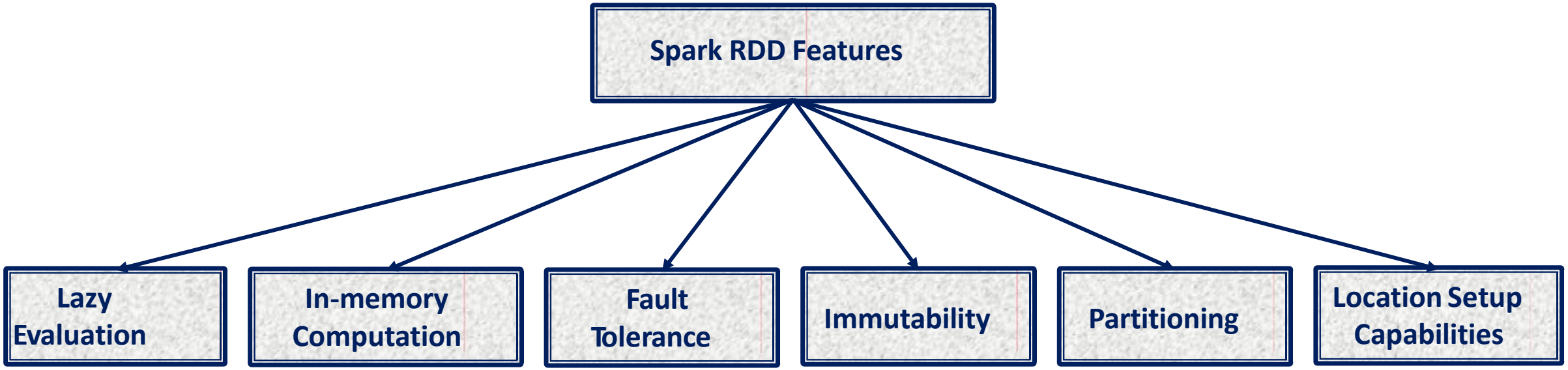
What is Spark RDDs?

- RDD are the building blocks of any spark application.
- **Resilient**: Fault tolerant and capable of rebuilding data on failure.
- **Distributed**: Distributed data among the multiple nodes in a cluster.
- **Dataset**: Collection of partitioned data with values.
- Each dataset in RDD is divided into logical partitions.

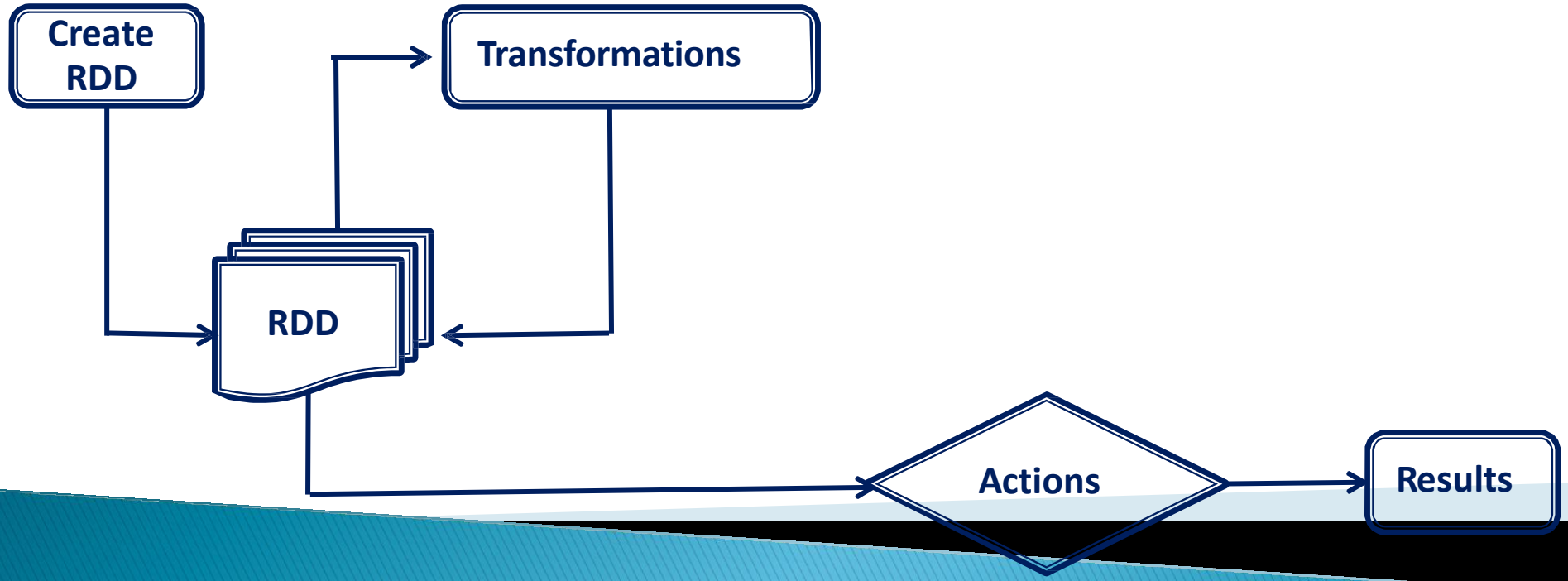
Spark RDD



Spark RDD Features



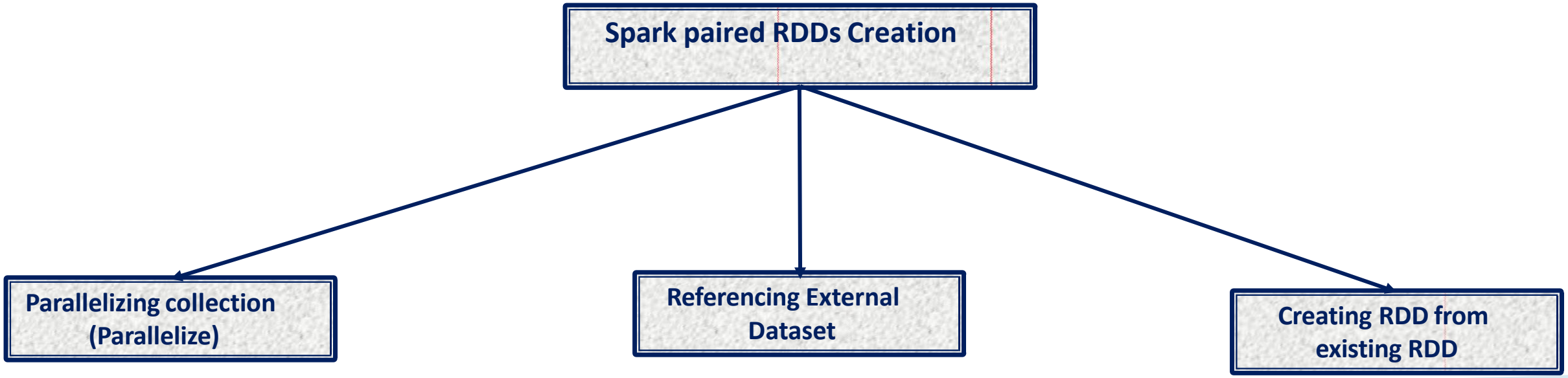
Spark RDD Operations



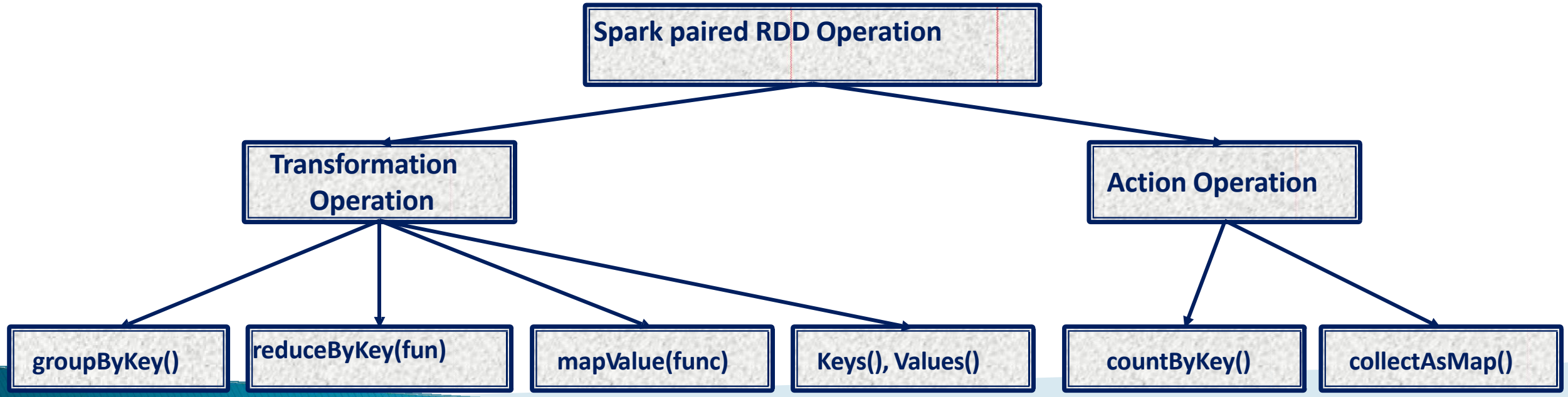
Introduction to Spark paired RDDs

- Spark Paired RDDs contains a key-value pair.
- Special operations like shuffle, grouping or aggregating elements by key.
- Operations for the key-value pair are available in the Pair RDD functions class.
- We can regroup the data across the network.
- Operations like reduceByKey() ,join.

How to create Spark paired RDDs



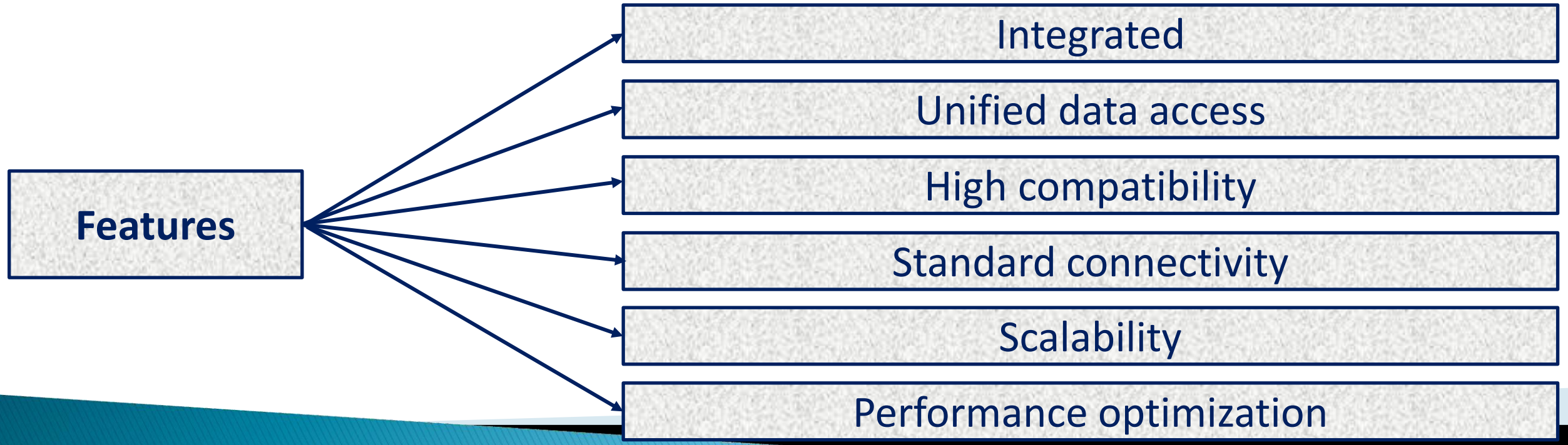
Spark paired RDD Operation



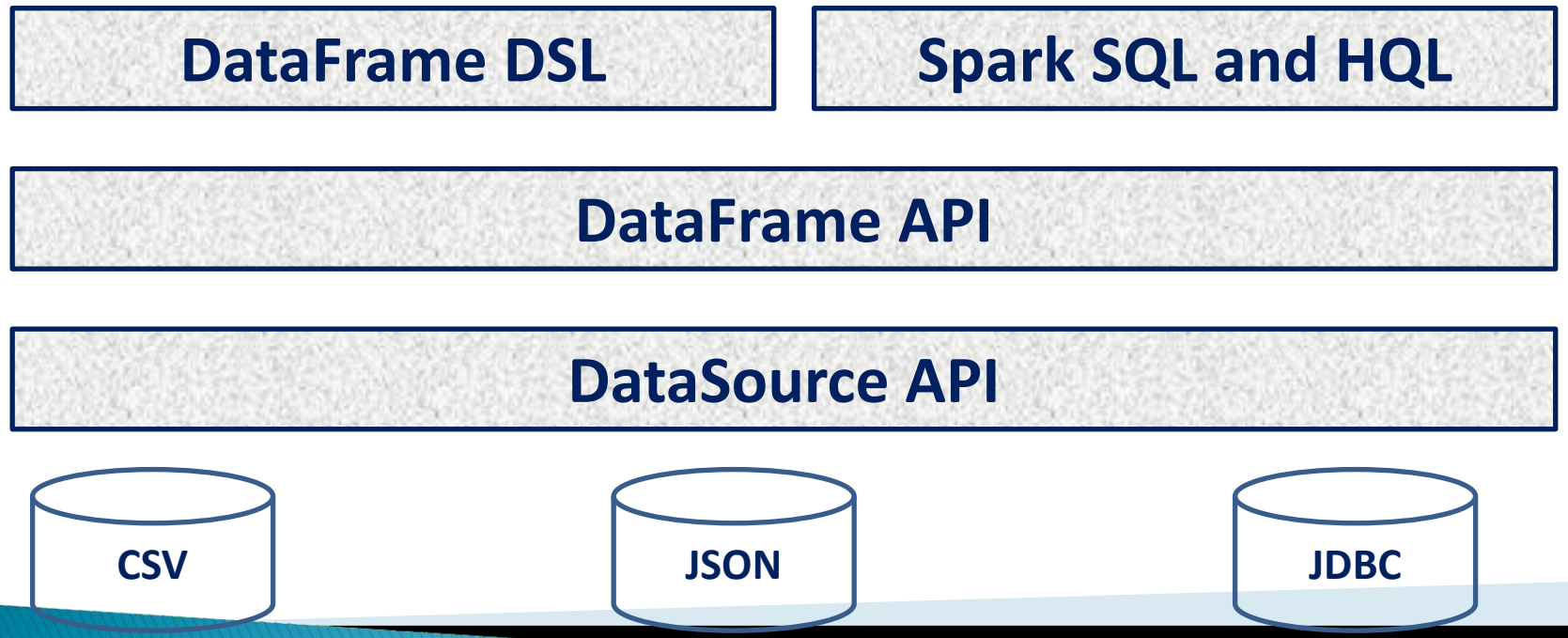
Spark SQL

- Spark module for structured data processing
- DataFrame API and Datasets API
- Spark SQL runs on top of the spark core.
- Extensible optimizer called catalyst

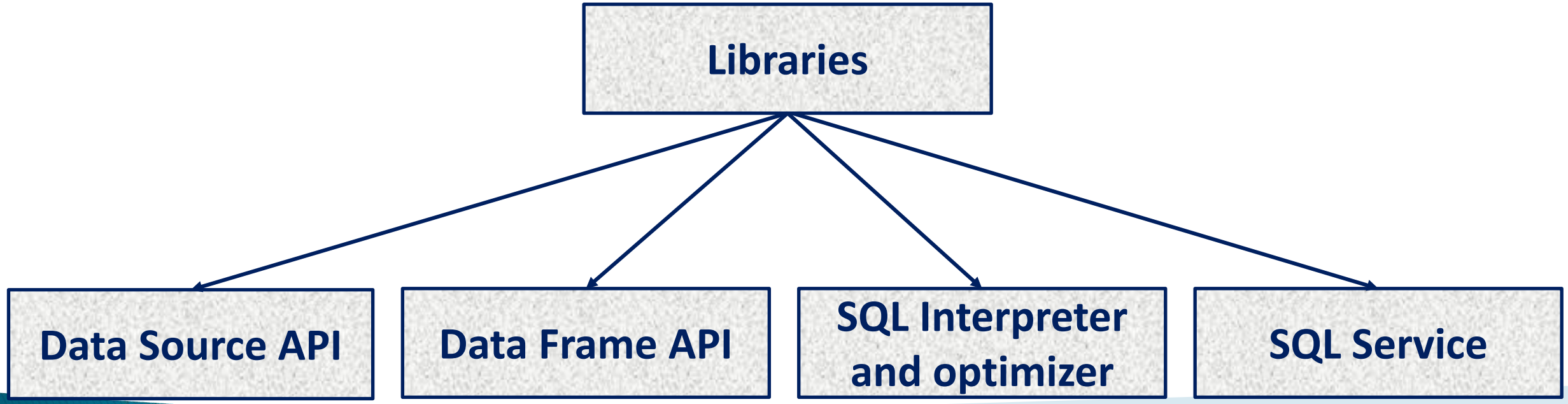
Features of Spark SQL



Spark SQL Architecture



Spark SQL Libraries



SQL Context

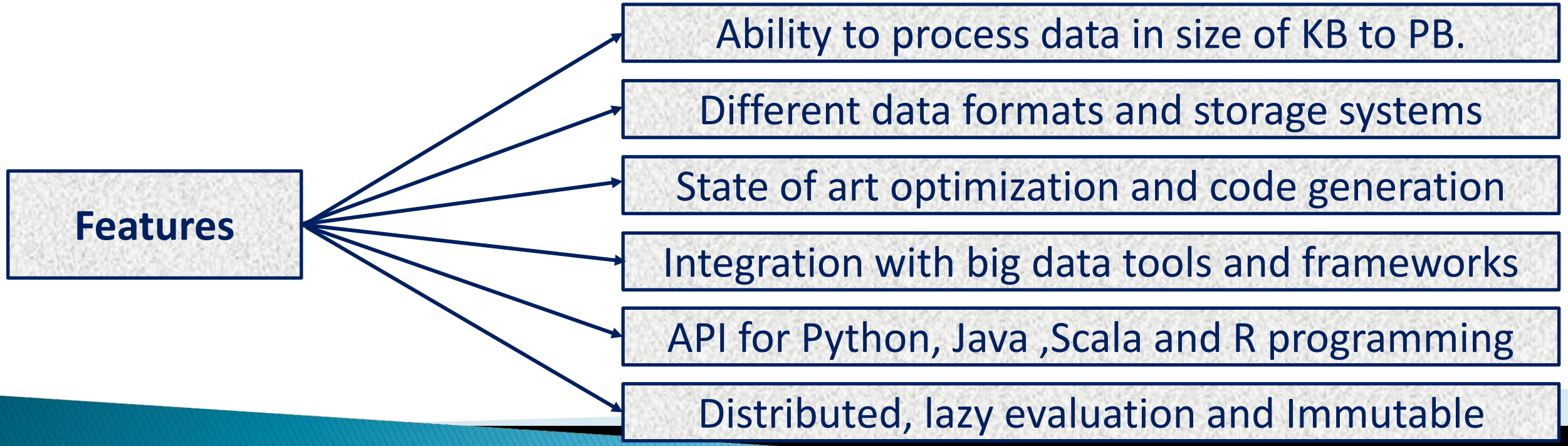
- SQLContext is a class used to initialize functionalities of Spark SQL
- An entry point to Spark SQL
- SparkContext object **sc** is required for initializing SQLContext class.
- val sqlcontext = new org.apache .spark.sql .SQL Context(sc)

DataFrame

- A distributed collection of data, which is organized into named columns

- A DataFrame can be constructed from an array of different sources such as Hive tables, Structured Data files, external databases, or existing RDDs.

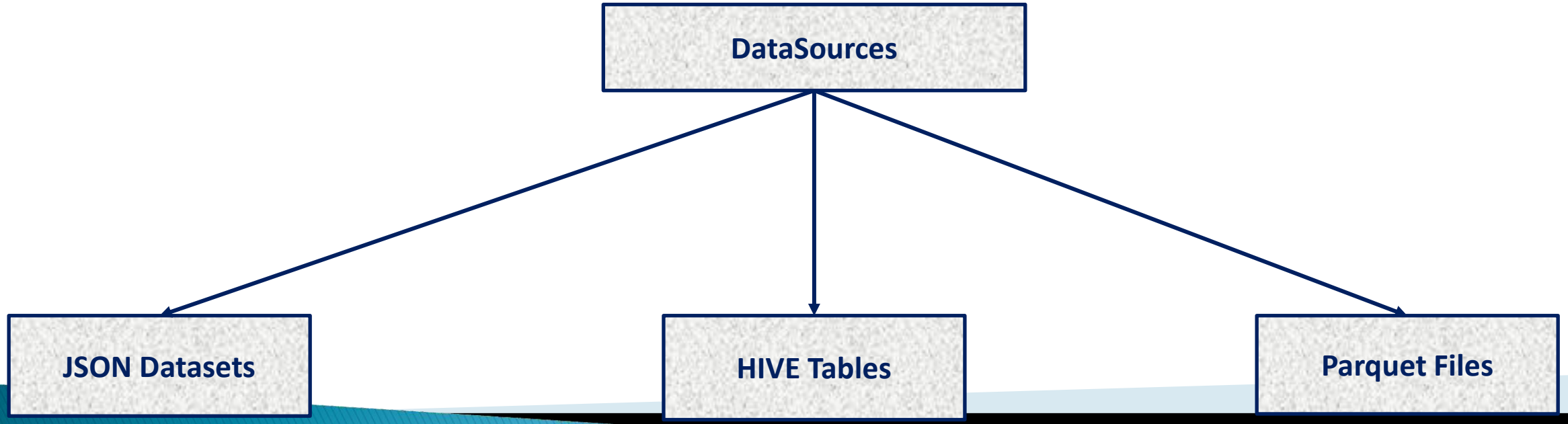
Features of DataFrames



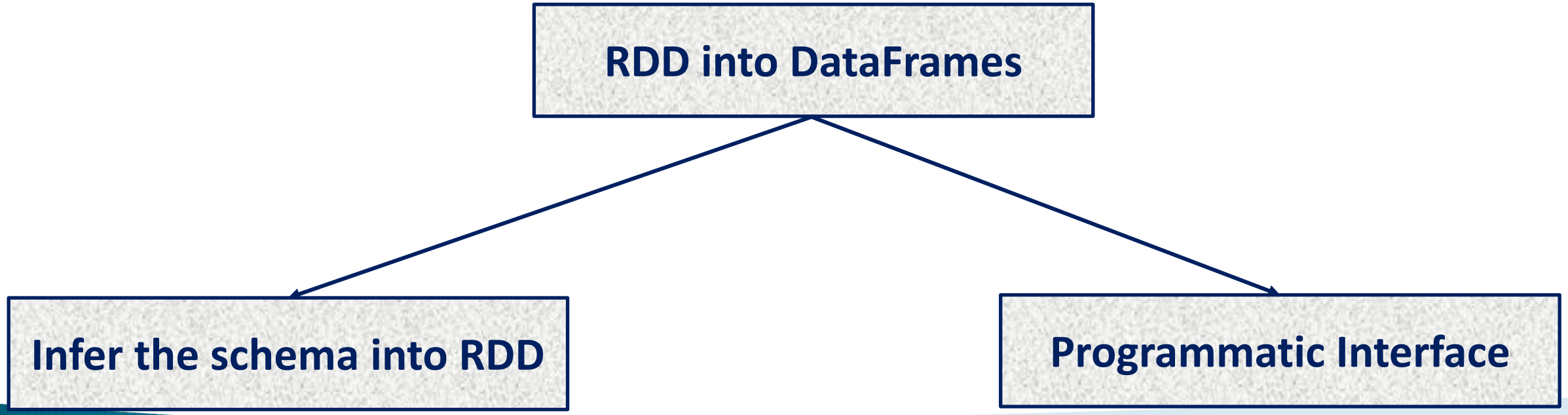
Create DataFrame

- createDataFrame()
- toDF()
- Existing RDD, DataFrame, Dataset, List, Seq data objects
- Sources like Text,CSV,JSON,XML, Binary files, RDBMS Tables, Hive

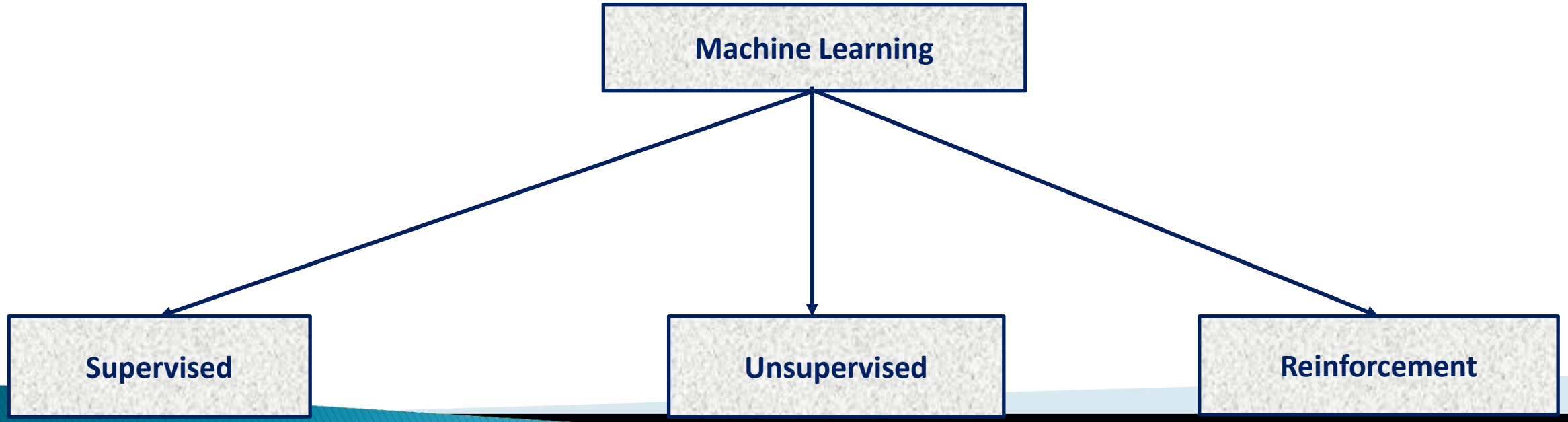
Different Types of DataSources



Converting Existing RDDs into DataFrames



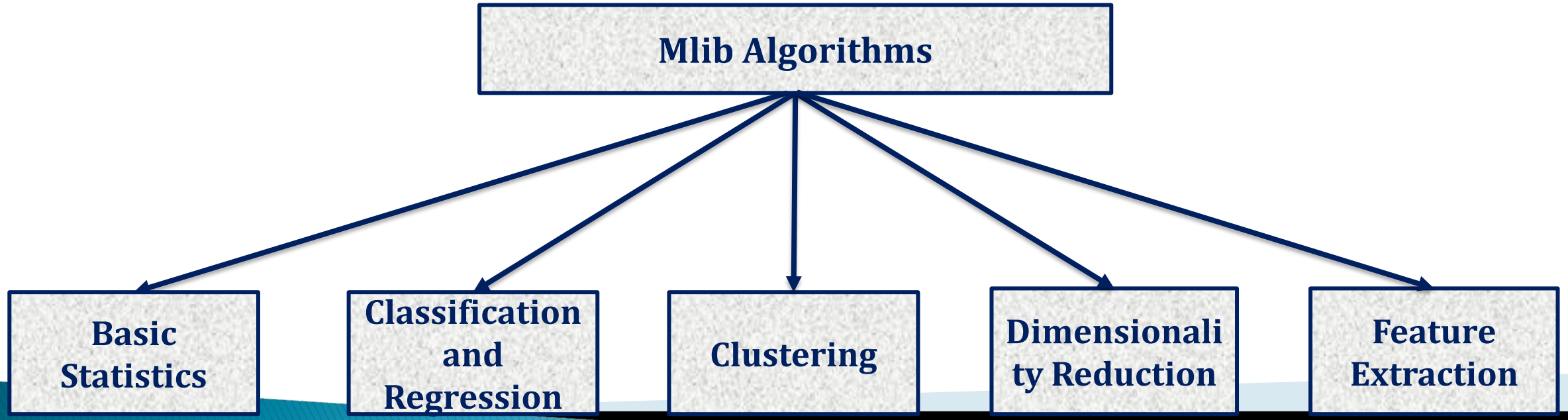
Machine Learning Tasks



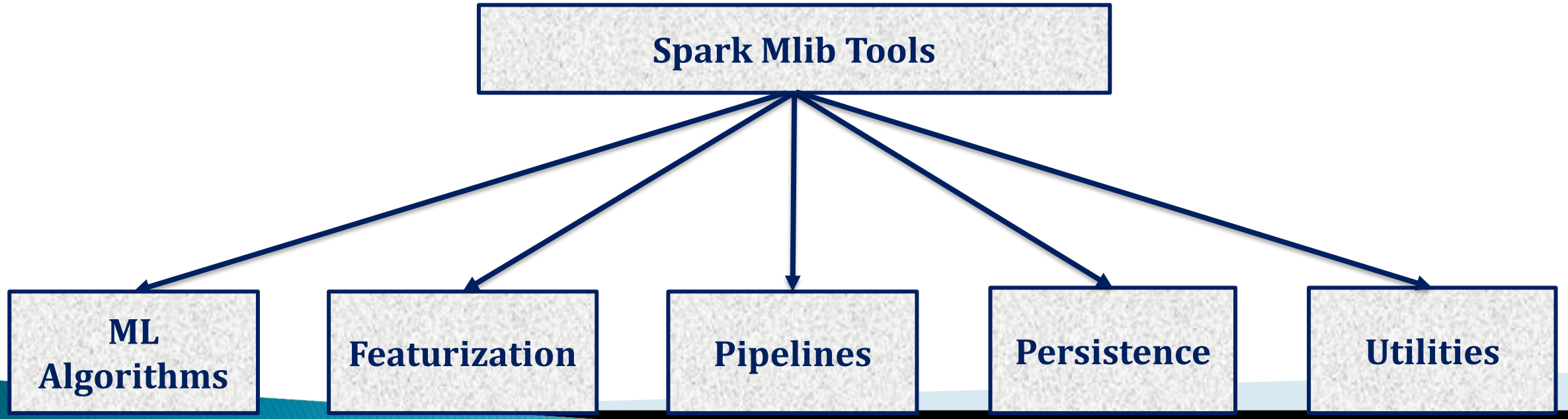
Spark MLib

- Spark MLlib is used to perform machine learning in Apache Spark
- *spark.mllib* : original API built on top of RDDs
- *spark.ml* :higher level API built on top of DataFrames.

Algorithms and Utilities in Spark MLib



Spark Mlib Tools



THANK YOU