What is Apache Spark?

- Apache Spark is an Open source analytical processing engine for large-scale powerful distributed data processing and machine learning applications.
- Spark was Originally developed at the University of California, Berkeley's, and later donated to the Apache Software Foundation.
- In February 2014, Spark became a Top-Level Apache Project and has been contributed by thousands of engineers making Spark one of the most active open-source projects in Apache.

Language supported by Spark

- Apache Spark 3.5 is a framework that is supported in Scala, Python, R Programming, and Java. Below are different implementations of Spark.

- Spark – Default interface for Scala and Java
- PySpark – Python interface for Spark
- SparklyR – R interface for Spark.

Features of Apache Spark

- In-memory computation
- Distributed processing using parallelize
- Can be used with many cluster managers (Spark, Yarn, Mesos e.t.c)
- Fault-tolerant
- Immutable
- Lazy evaluation
- Cache & persistence
- Inbuild-optimization when using DataFrames
- Supports ANSI SQL

Advantages of Apache Spark

- Spark is a general-purpose, **in-memory**, fault-tolerant, **distributed processing** engine that allows you to process data efficiently in a distributed fashion.
- Applications running on Spark are **100x** faster than traditional systems.
- You will get great benefits from using Spark for data ingestion pipelines.
- Using Spark we can process data from Hadoop **HDFS**, **AWS S3**, **Databricks DBFS**, **Azure Blob Storage,** and many file systems.
- Spark also is used to process real-time data using Streaming and Kafka.
- Using Spark Streaming you can also stream files from the file system and also stream from the socket.

BASIC

- RDD- Resilient Distributed Dataset
- DAG – Directed Acyclic graph

What is HDFS

- HDFS is Hadoop Distributed File System
- We all know that hadoop uses distributed storage as well as distributed processing
- In Hadoop the place where all the data is stored can be called as HDFS

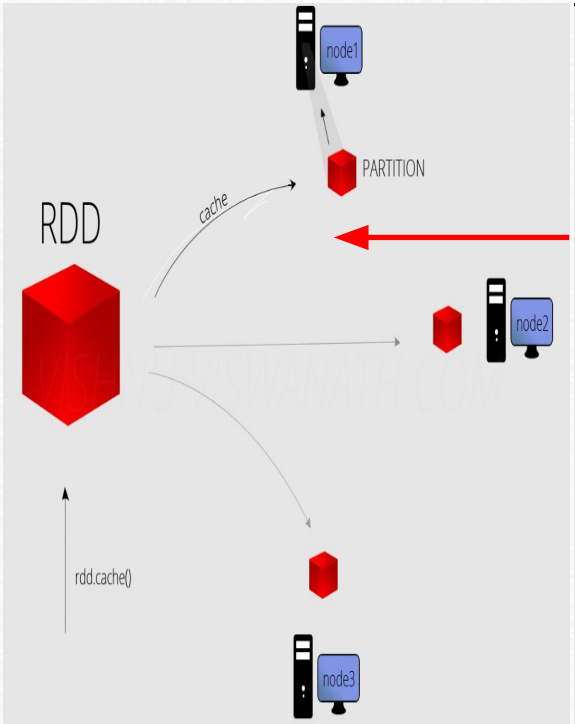
File Blocks

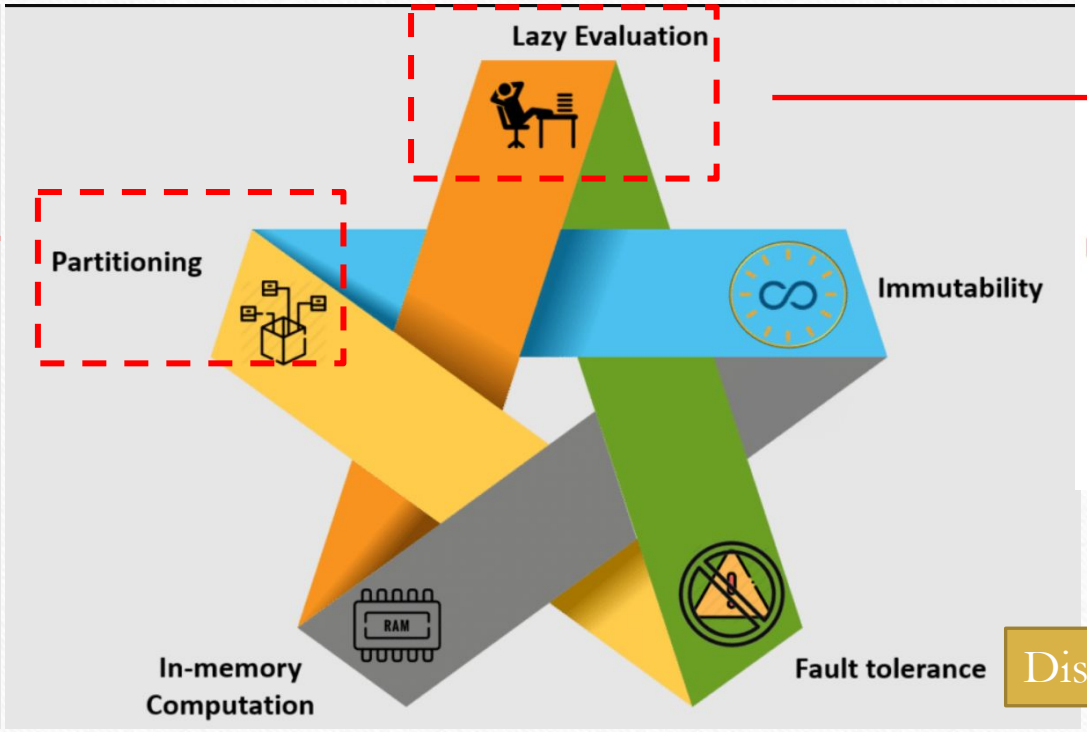
- Hadoop divides data in 128MB Block

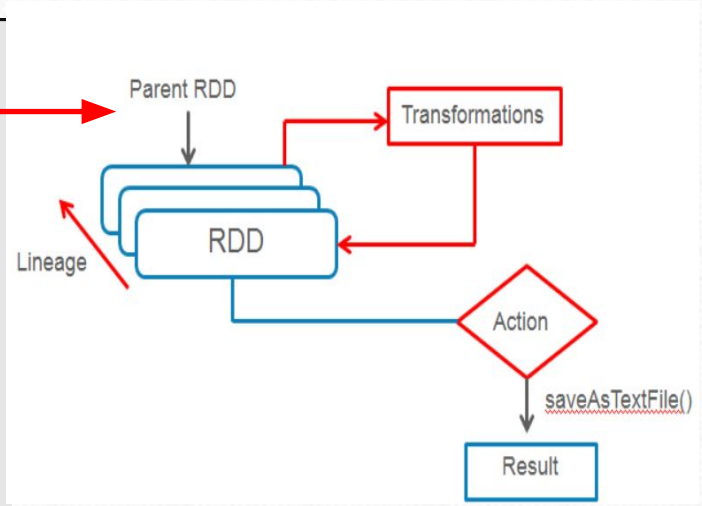
For Hadoop1 → 64MB For Hadoop2 → 128MB

- Configurable

RDD- Resilient Distributed Dataset

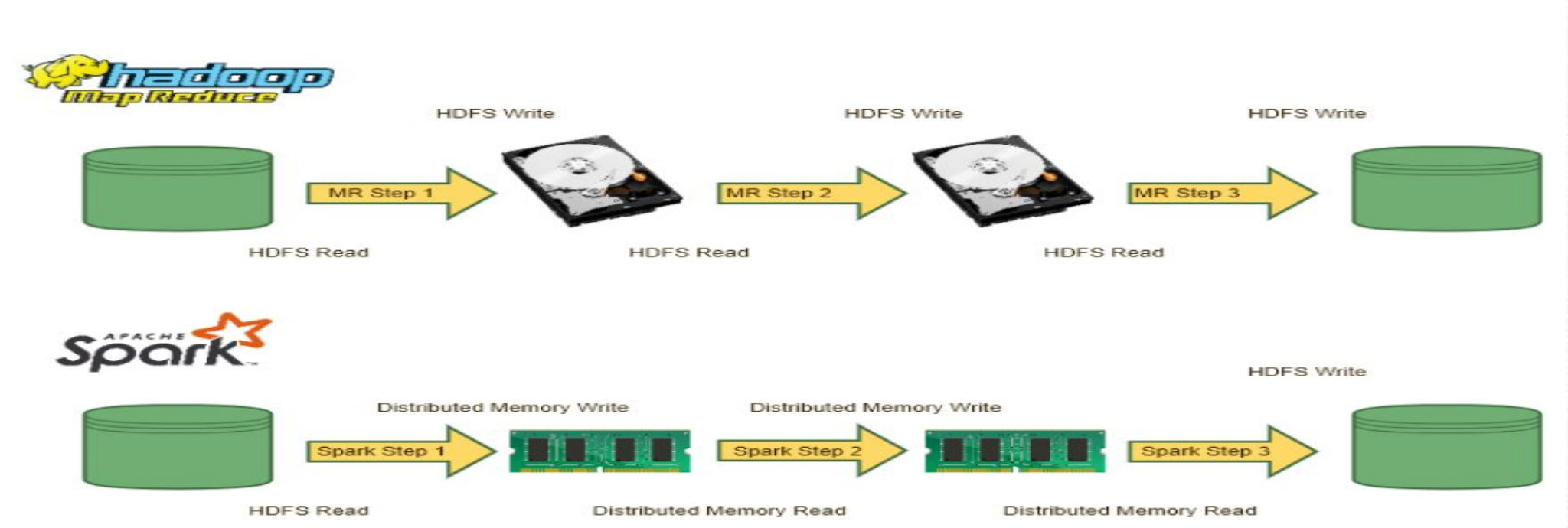






Distributed datasheet

In memory computing



RDD Operation's

Lazy Evaluation

- Rdd1= sc.parallelize(path…,)
- Rdd2= Rdd1.filter()……..

Transformation

- Rdd3.take(2)….

Action

RDD Operation's

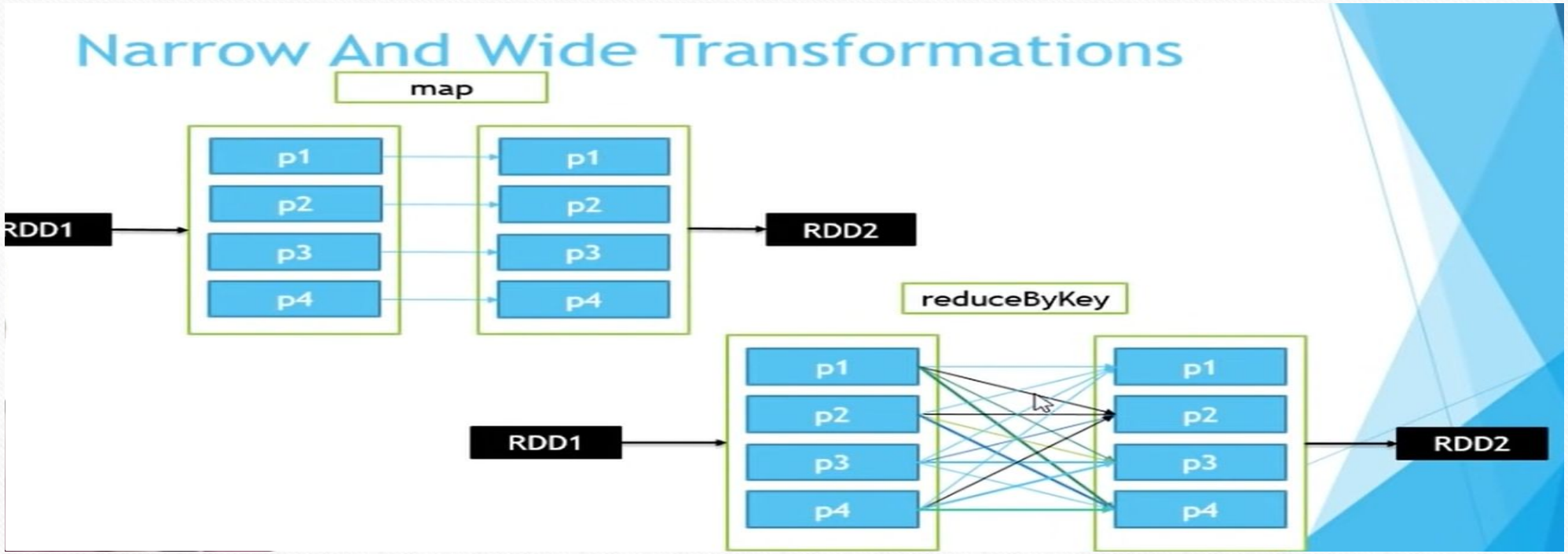
- **Transformation's**

I. Narrow Transformations (map, filter, sample, union…)

II. Wide Transformations (intersection , join…)

- **Action's** (count, collect, take…)

Transformation's



Cluster Manager

- Hadoop YARN
- Apache Mesos
- Standalone scheduler (Apache Spark)

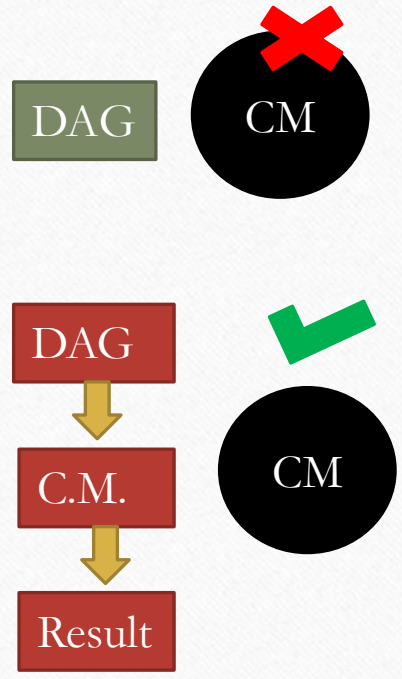
DAG – Directed Acyclic graph

- Rdd1= spark.read.csv(path…,)
- Rdd2= Rdd1.filter()……..

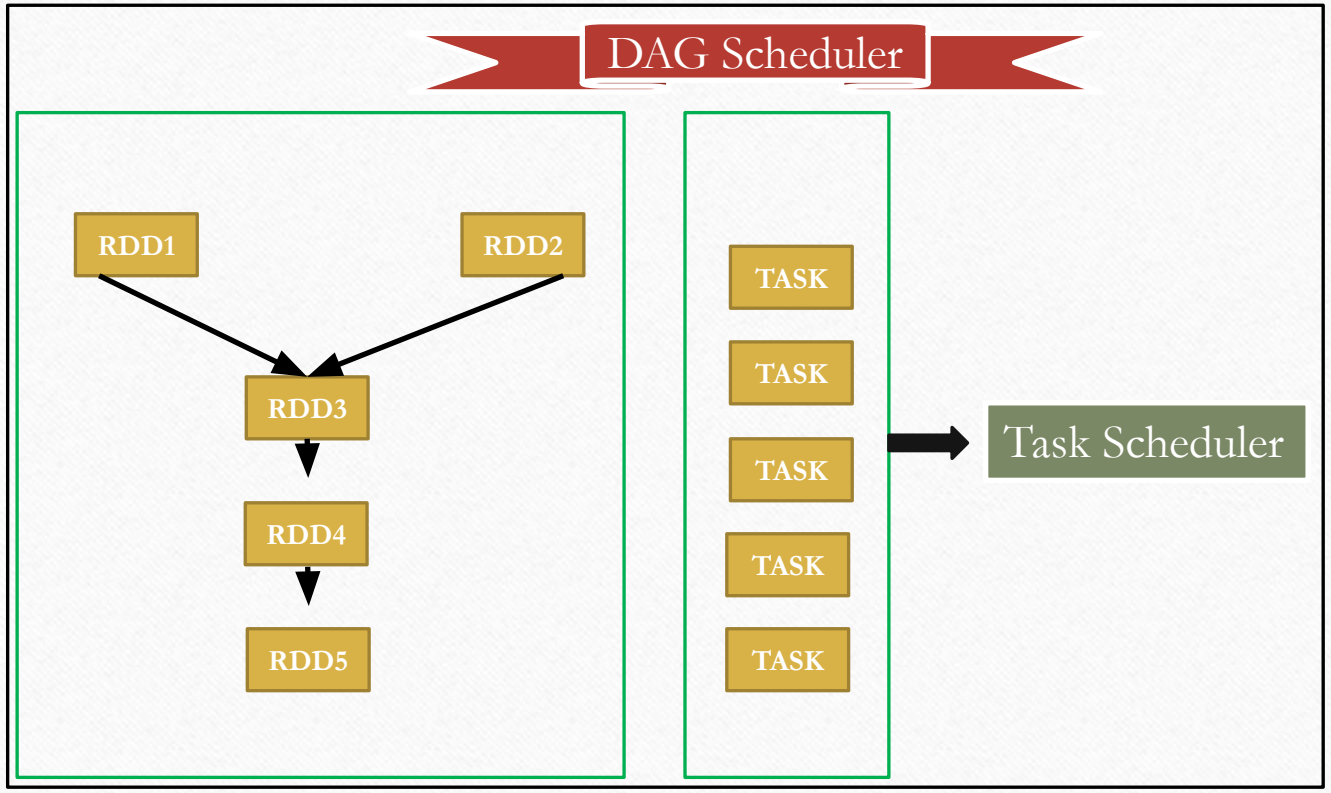
Transformation

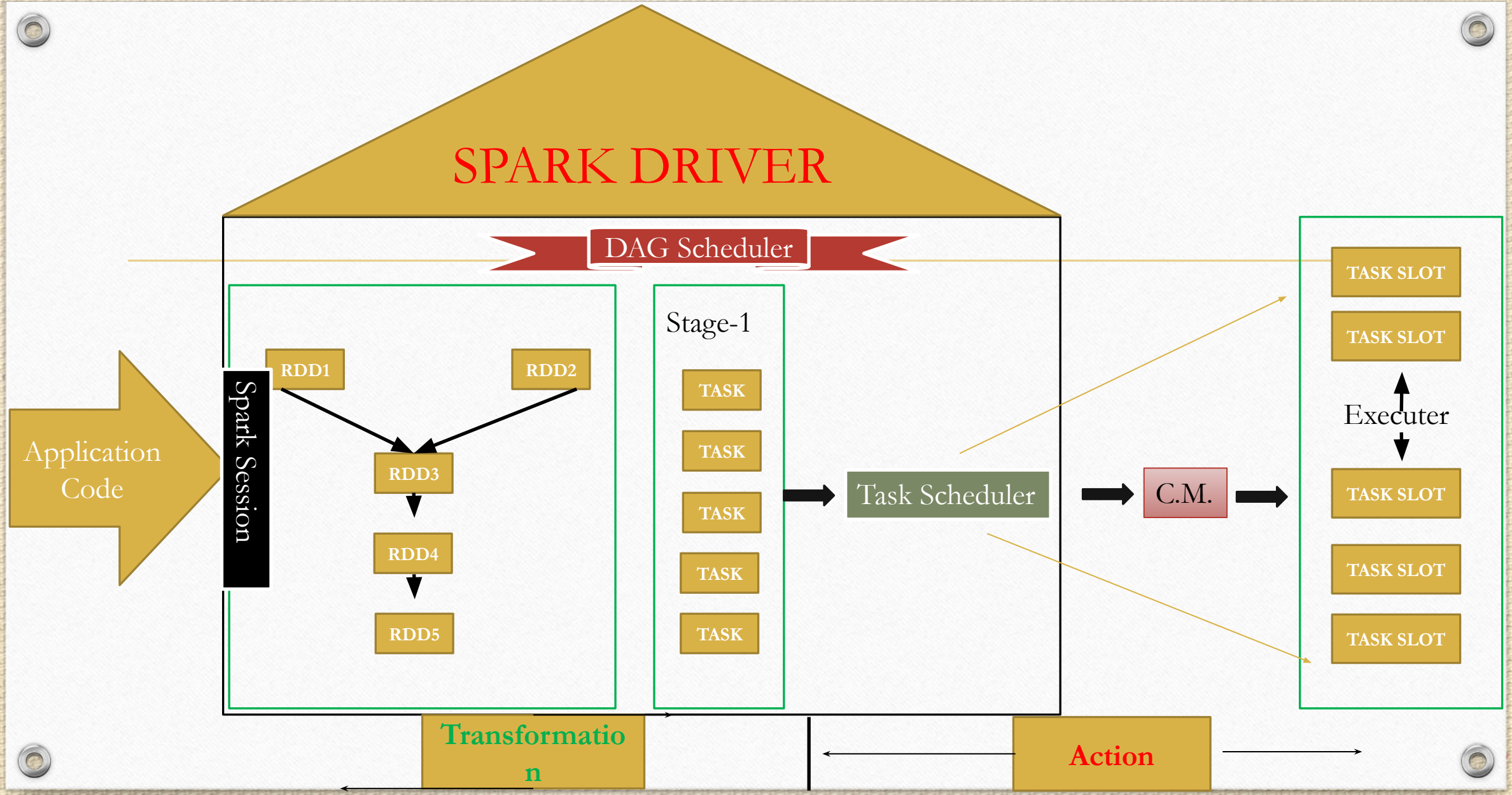
- Rdd3.take(2)….

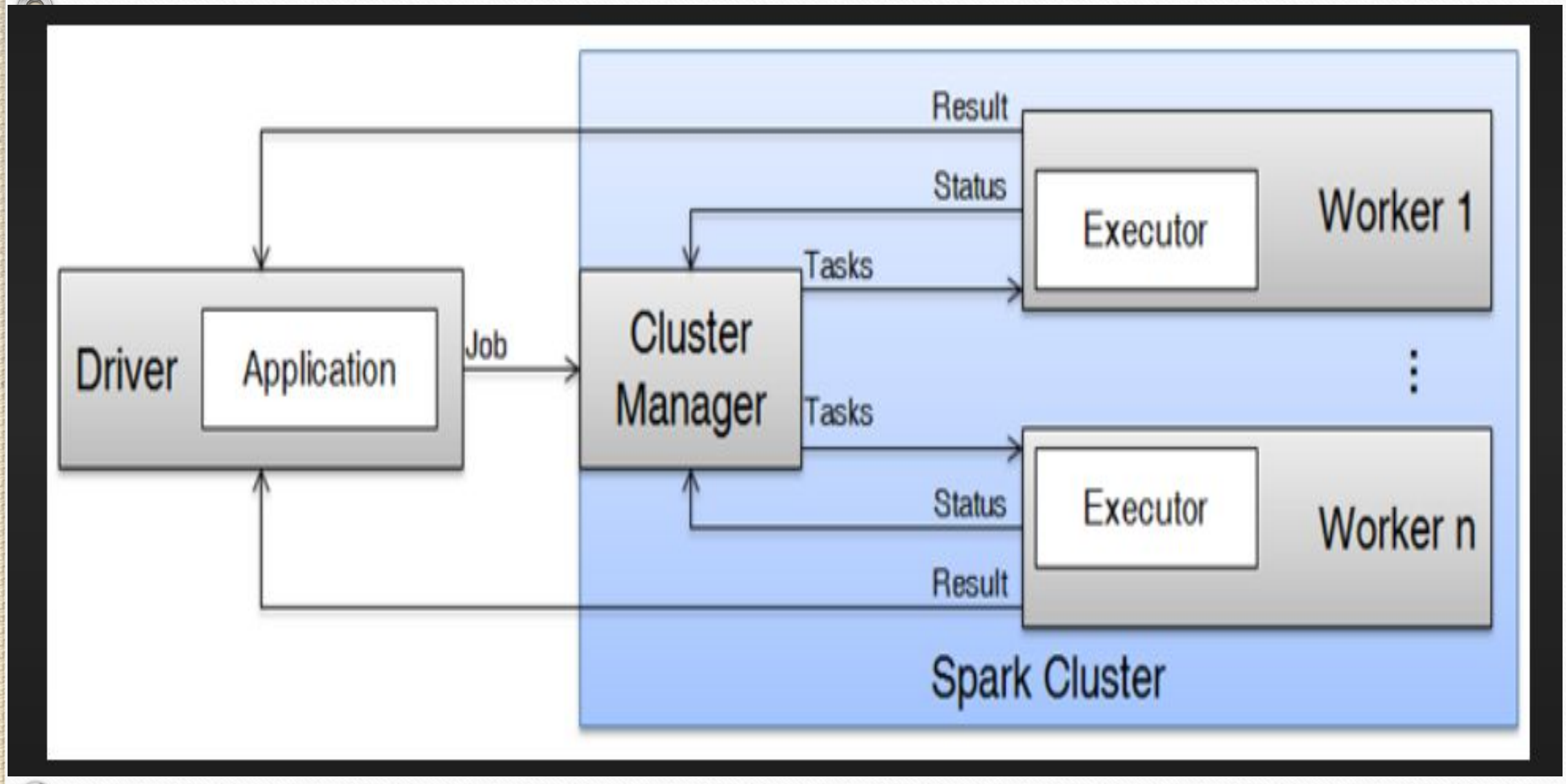
Action



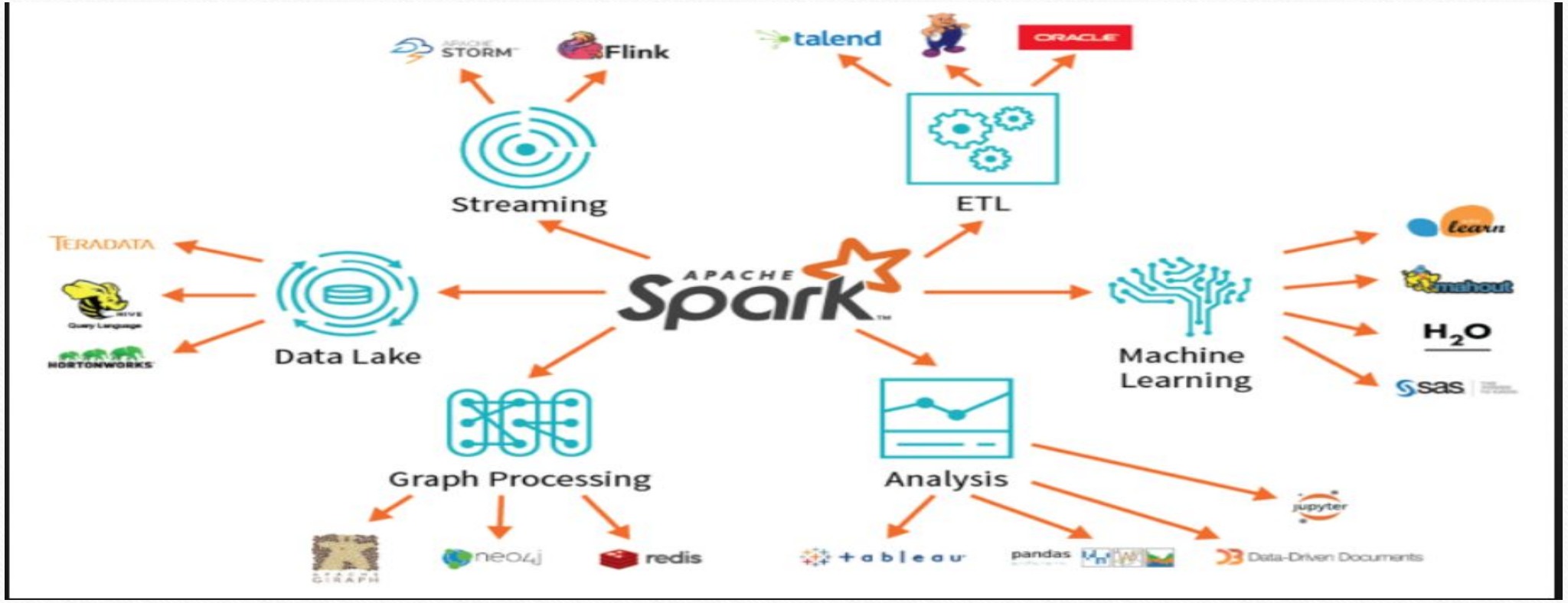
DAG


DAG Scheduler
RDD1
RDD2
RDD3
RDD4
RDD5
TASK
TASK
TASK
TASK
TASK
Task Scheduler


SPARK DRIVER
DAG Scheduler
Application Code
Spark Session
RDD1
RDD2
RDD3
RDD4
RDD5
Stage-1
TASK
TASK
TASK
TASK
TASK
Task Scheduler
C.M.
TASK SLOT
TASK SLOT
Executer
TASK SLOT
TASK SLOT
TASK SLOT
Transformation
Action


Result
Status
Tasks
Driver
Application
Job
Cluster Manager
Executor
Worker 1
Tasks
Status
Executor
Worker n
Result
Spark Cluster

Designed to cover wide range of workload



THANK YOU

YOUR QUE. & SUGGESTION ARE MOST WELCOME