

Introduction to Spark

Internal training



Contents

- I. Overview of relevant concepts
 - 1. Introduction
 - 2. Spark's core components
 - 3. Unified framework
 - 4. RDDs
 - 5. Lazy vs eager evaluation
 - 6. Wide vs narrows transformations
 - 7. Catalyst Optimizer
 - 8. Shuffling & Partitioning
- II. Performance Evaluation
 - 1. Explore query plans
 - 2. Spark UI
 - 3. Types of joins



What is Apache Spark?

A distributed data processing engine designed for big data and large-scale computation

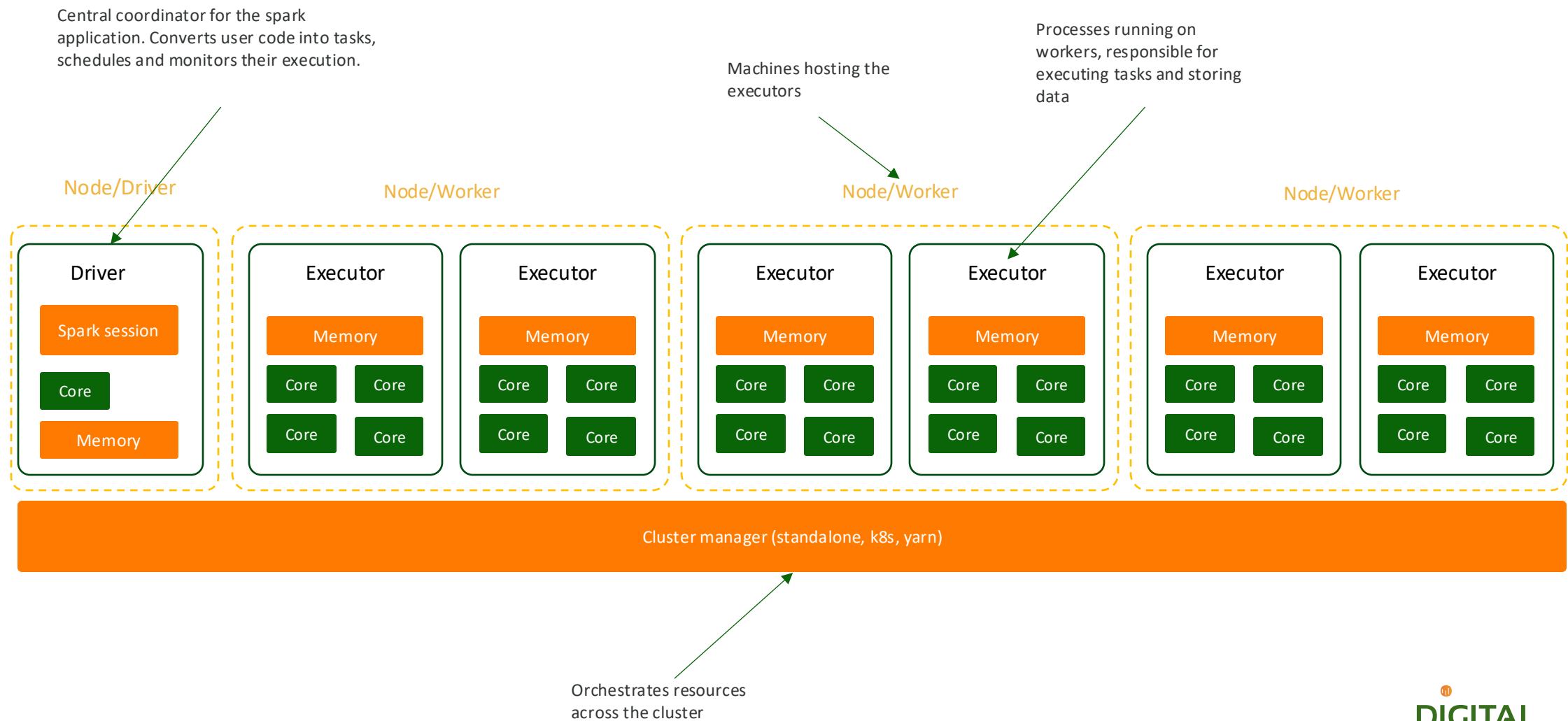
Spark's key features

- Speed: In-memory processing for faster computation
- Scalability: Handles petabytes of data across large clusters
- Versatility: Supports multiple programming languages (API, Java, Scala, R, SQL)

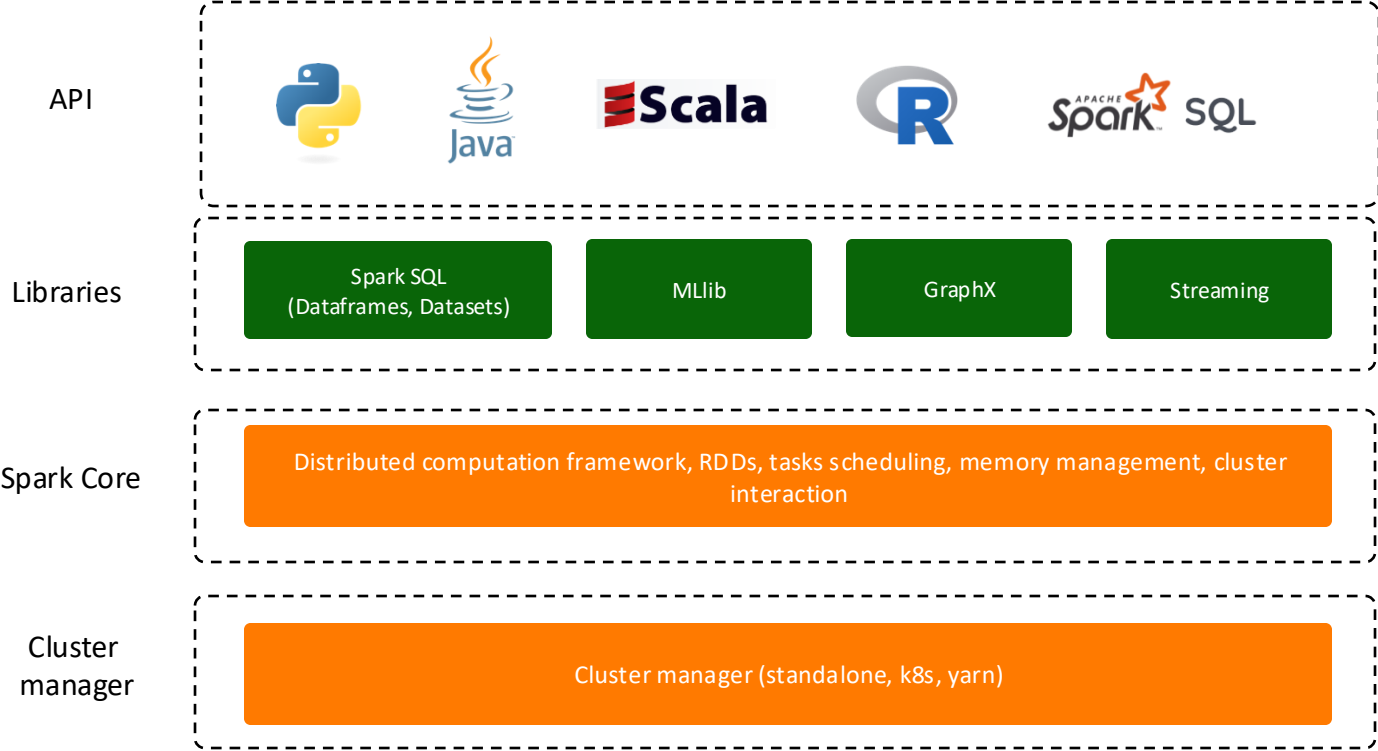
Common use cases:

- ETL pipelines
- Data Analytics
- Machine Learning workflows

Spark's core components



Spark's unified framework

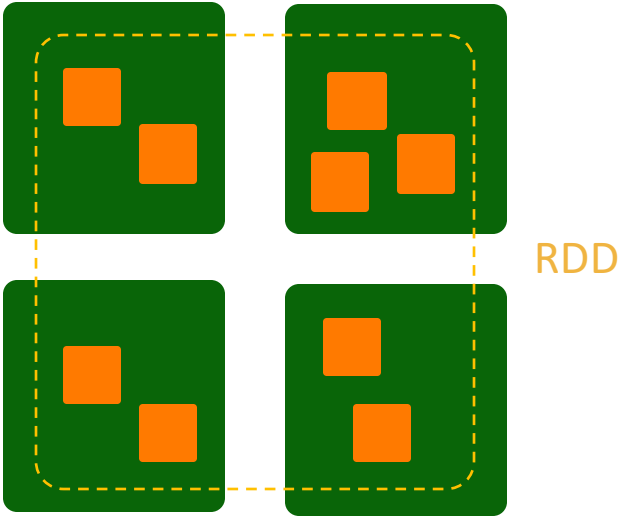
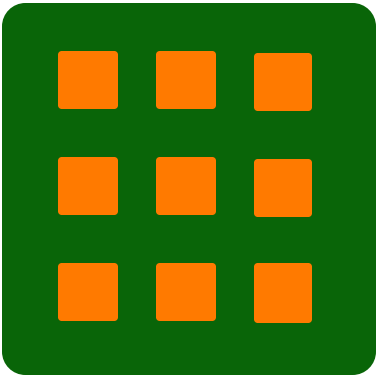




RDDs

Resilient distributed dataset (RDD), is a fault-tolerant, immutable collection of elements that can be operated in parallel across a cluster of machines.

```
my_var = [1, 2, 3, 4, 5, ..., N]
```

```
my_rdd = sc.parallelize([1, 2, 3, 4, 5, ..., N])
```



-  Node
-  Data point

Lazy vs eager evaluation

Lazy

Evaluation of expressions is delayed until their results are needed

Eager

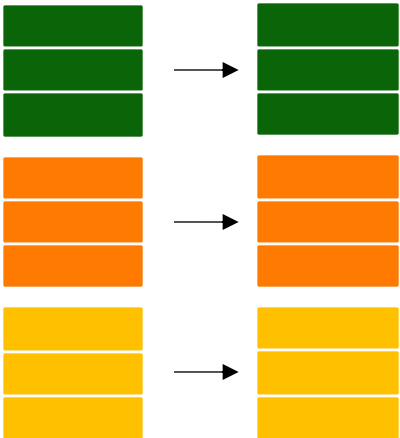
Evaluation of expressions occur every time a new expression is declared.

Feature	Transformations	Actions
Definition	Operations that define a computation plan	Operations that trigger an execution
Evaluation	Lazy – not executed until an actions runs	Eager – triggers the execution
Examples	map(), filter(), select(), flatMap(), withColumn(), sample(), groupBy(), join(), sortBy(), repartition(), coalesce(), etc	count(), collect(), save(), show(), take(), toPandas()
Returns	New RDD/Dataframe	Final result

Narrow vs Wide transformations

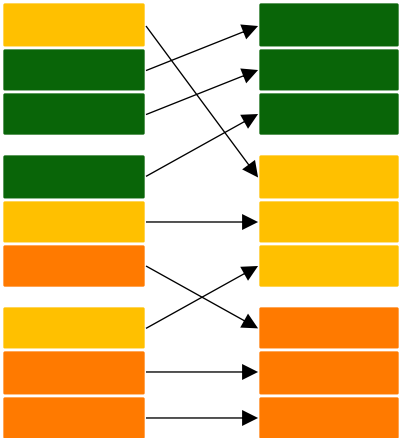
Transformations

Narrow transformations



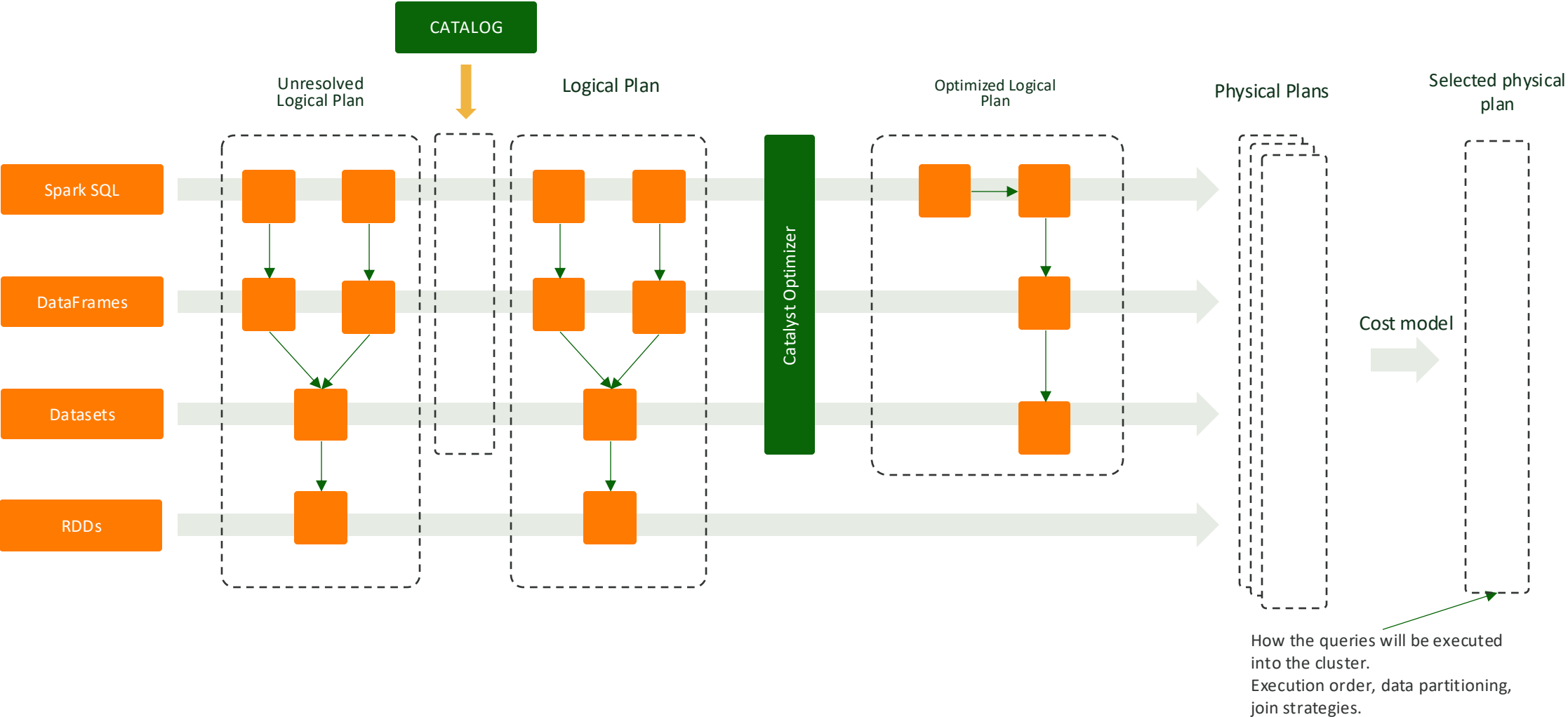
Filter
FlatMap
Map
WithColumn
Sample

Wide transformations



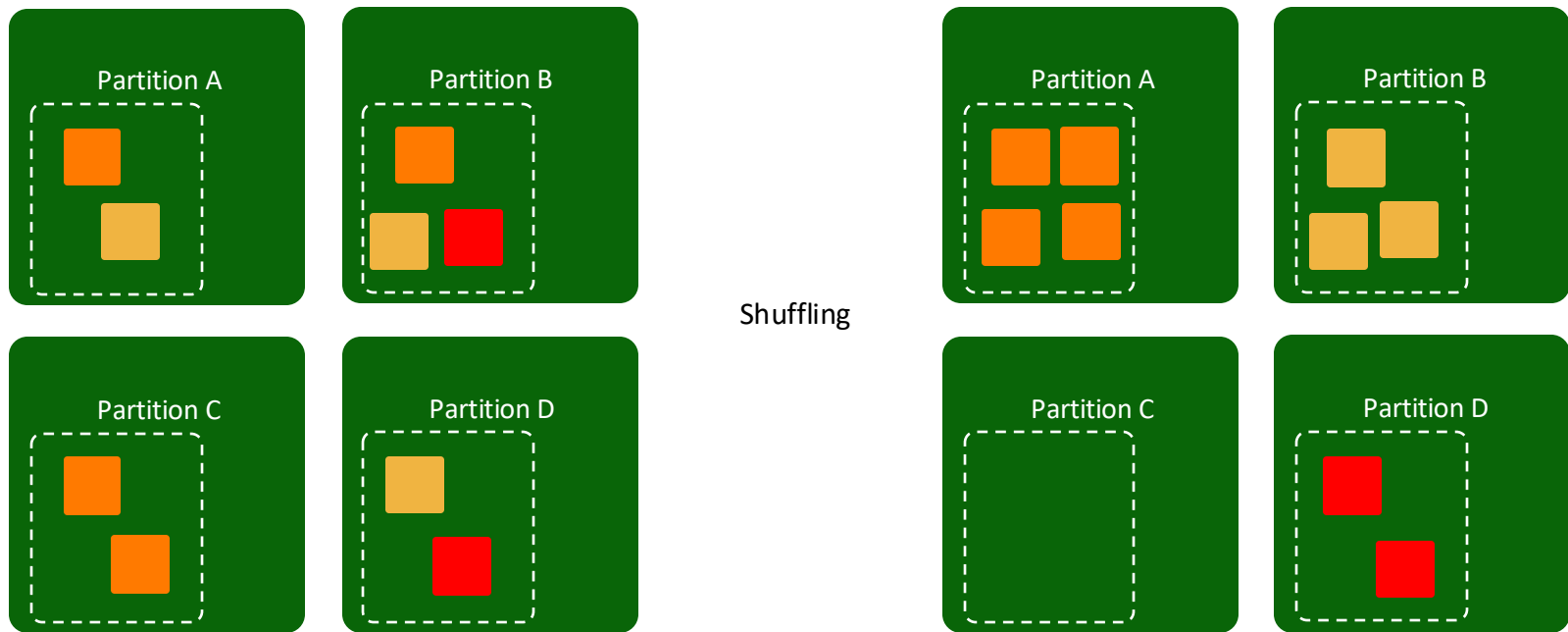
GroupBy
Join
Distinct
Repartition
Coalesce
SortBy
Distinct

Catalyst optimiser and Tungsten engine



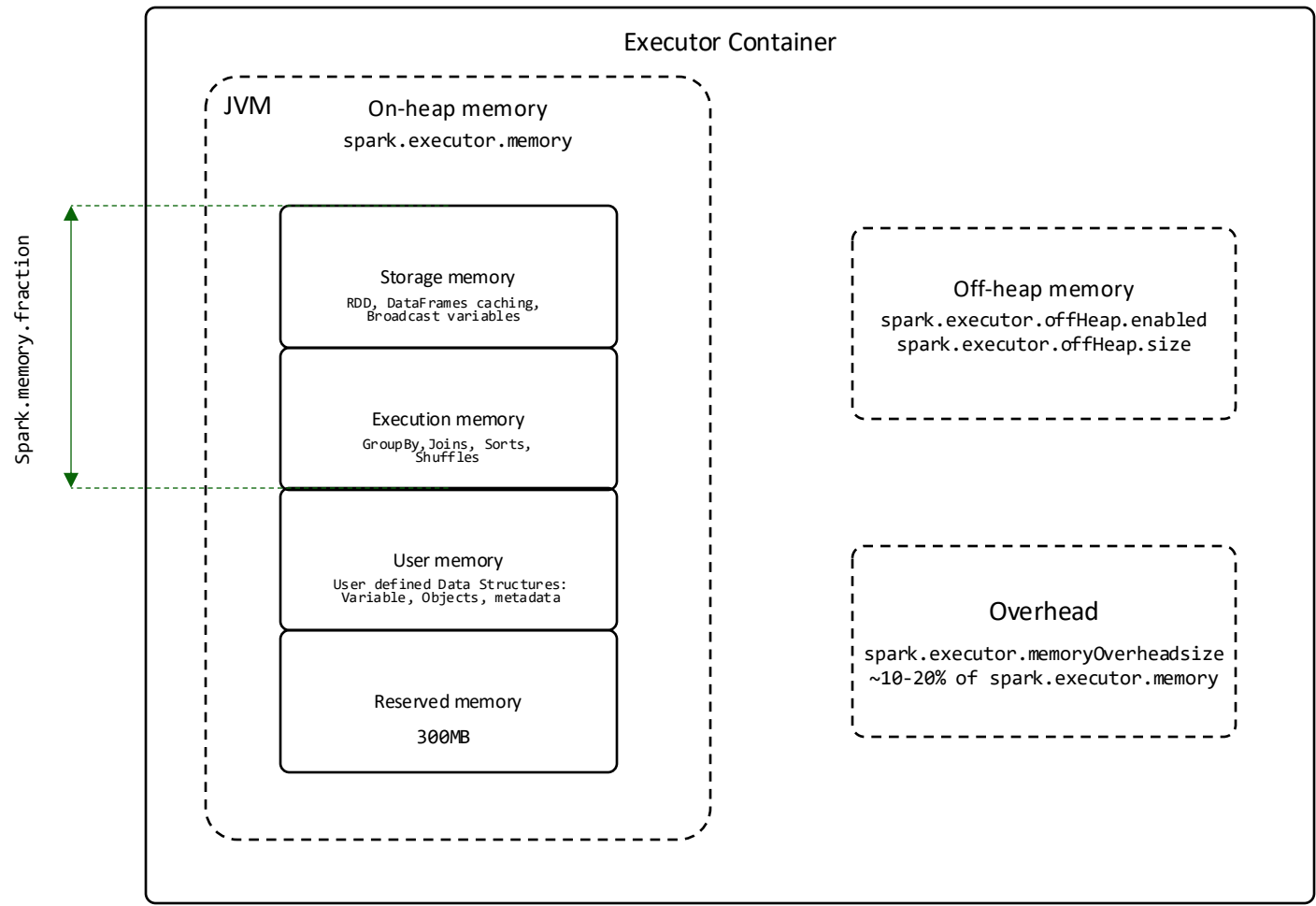
Source: [Key topics in Apache Spark](#), [Catalyst Optimizer](#)

Shuffling and partitioning



- Node
- Data point

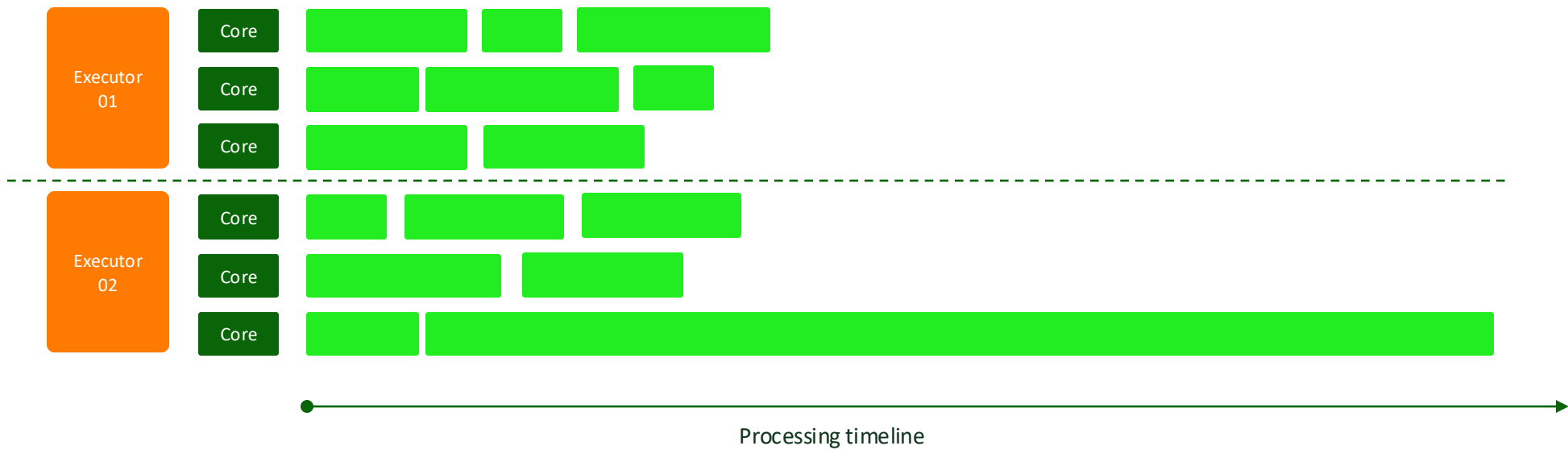
Bonus: Memory management



Source: [Spark's memory management overview](#)

Bonus: Data Skew

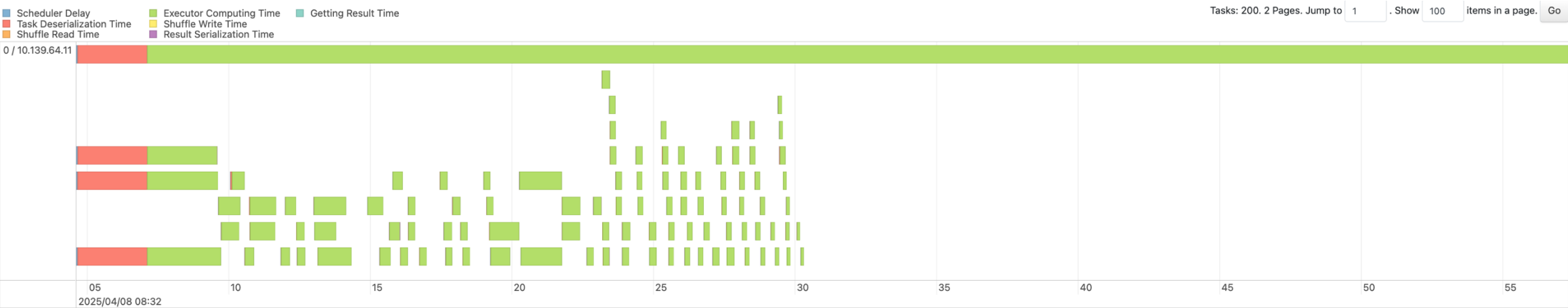
Transformations: groupBy(), join()



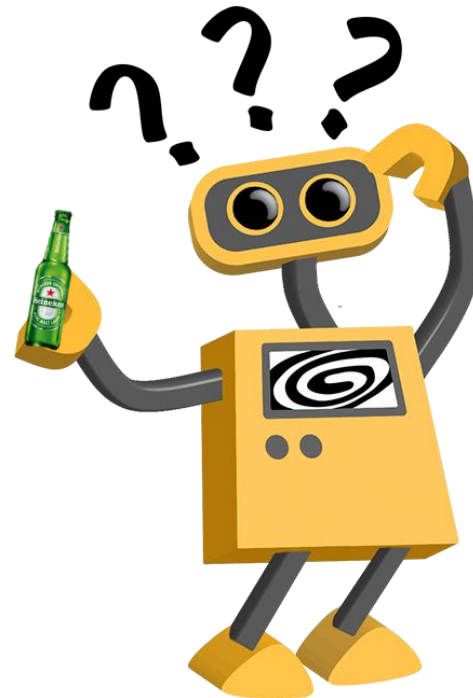
- Worker/Executor
- Core
- Task



Bonus: Data Skew



Summary Metrics for [200 Completed Tasks](#)





Questions?