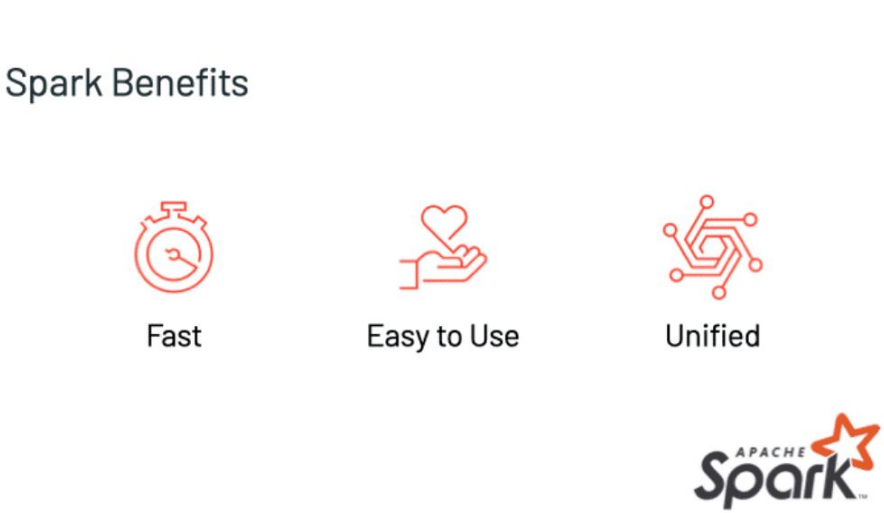
DSCC_202_402

Spark Core Module

Why Spark?

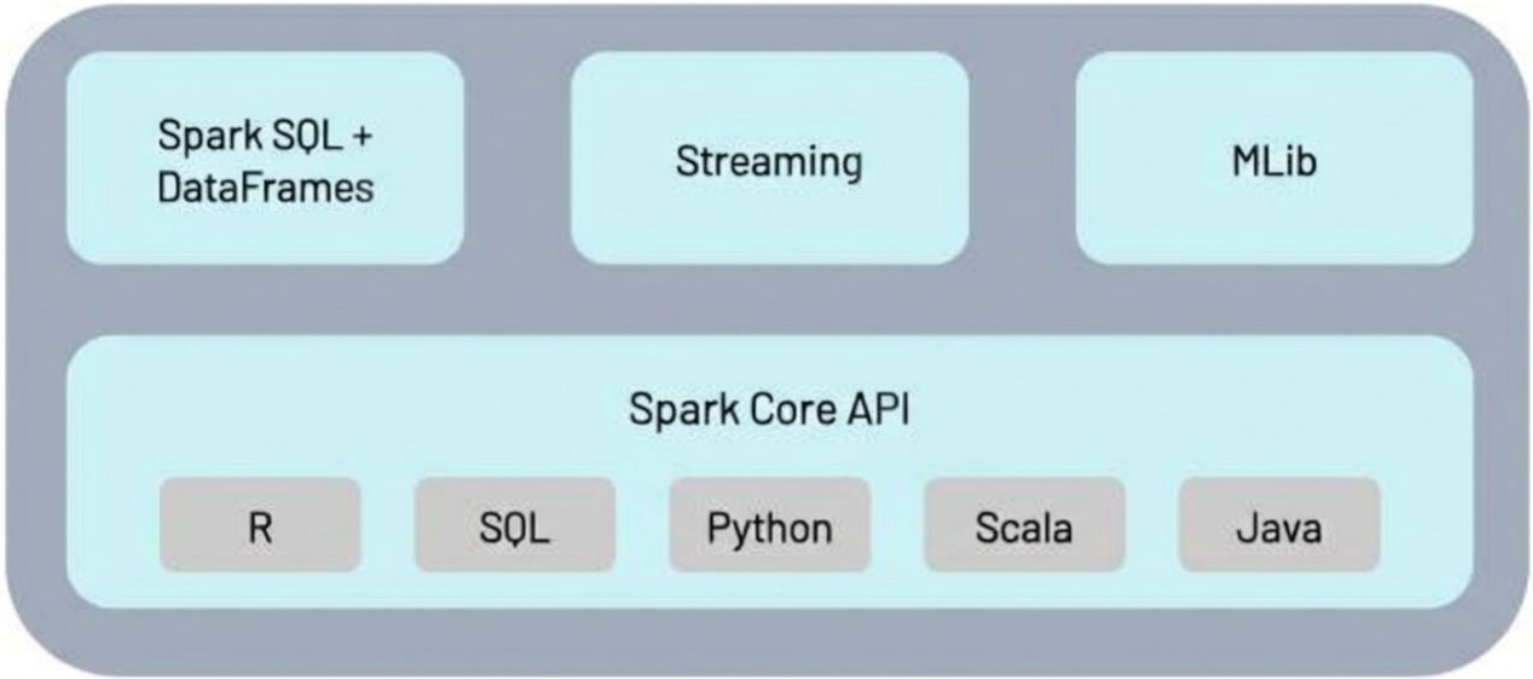


Spark Benefits

- 100x faster than Hadoop for large scale data processing
- Easy-to-use APIs for operating on large datasets
- A unified API and engine for SQL queries, streaming data, machine learning. Spark can be seamlessly combined with other tools to create complex workflows

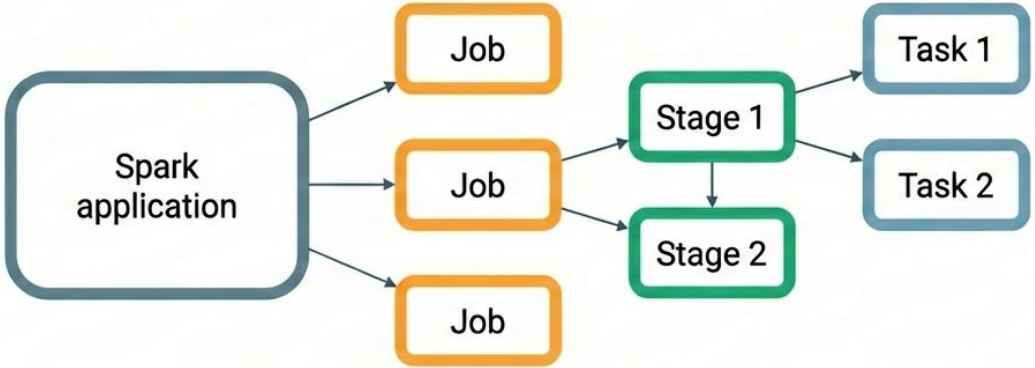
[Ray vs Dask vs Apache Spark™ — Comparing Data Science & Machine Learning Engines]

Spark API

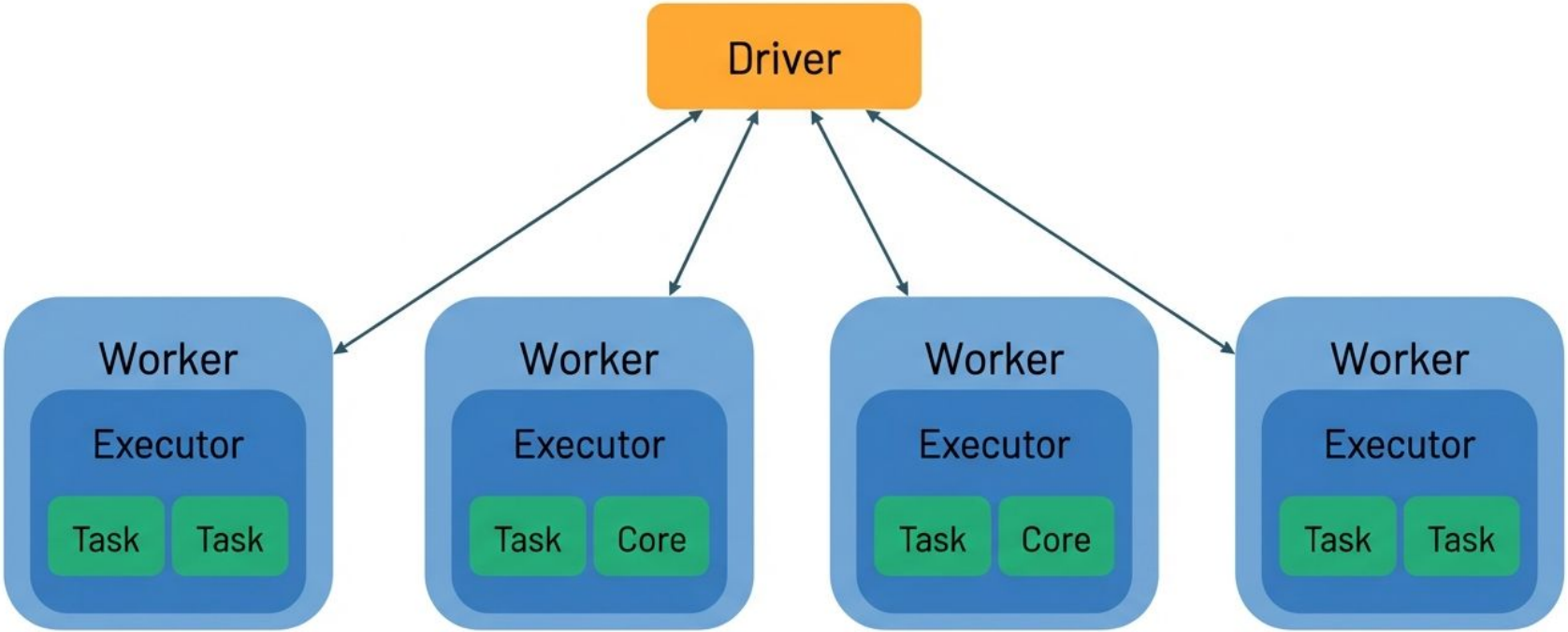

Spark SQL +
DataFrames
Streaming
MLib
Spark Core API
R
SQL
Python
Scala
Java

Spark Execution

Spark uses clusters of machines to process big data by breaking a large task into smaller ones and distributing the work among several machines. Let's look at how spark executes a spark application.



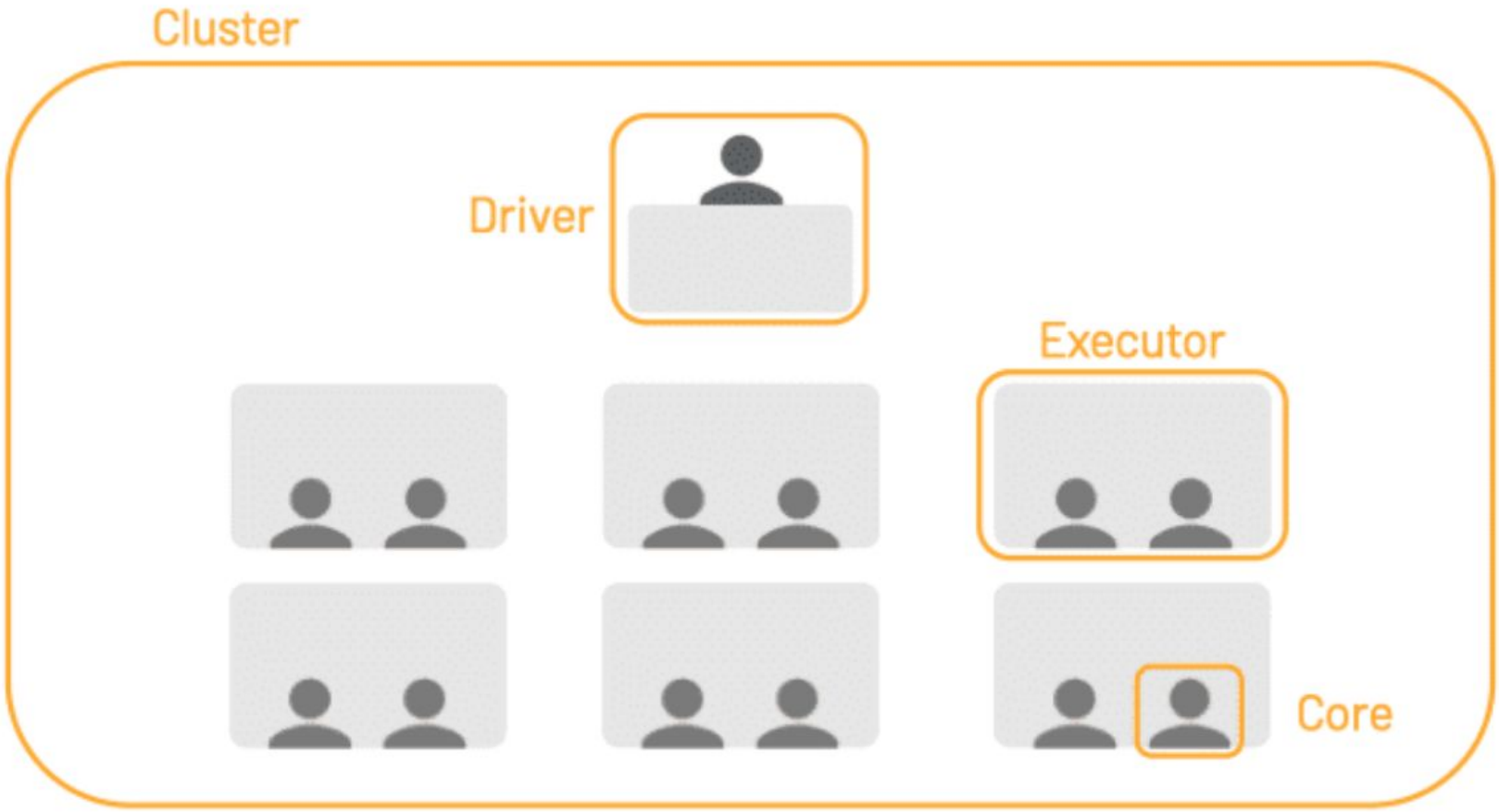
Spark Cluster



Spark Architecture

By the end of this lesson you will be able to:

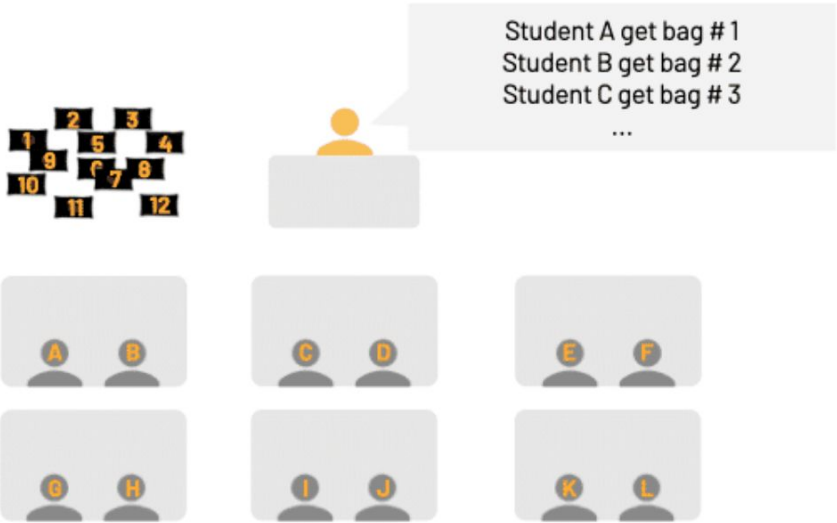
1. Identify the components of a Spark cluster
2. Identify cluster components and concepts in a filtering scenario
3. Identify cluster components and concepts in a counting scenario


Cluster
Driver
Executor
Core

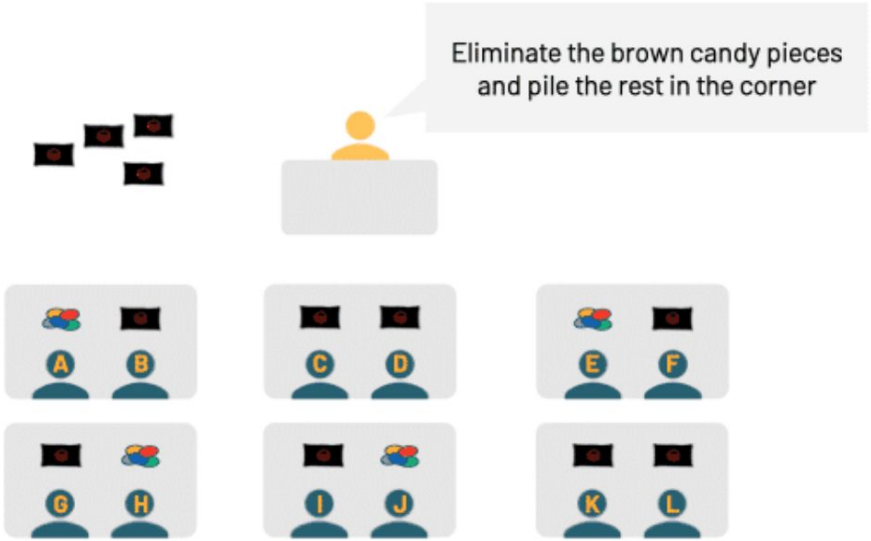

Data
Partition

Scenario 1: Filter out brown pieces from candy bags

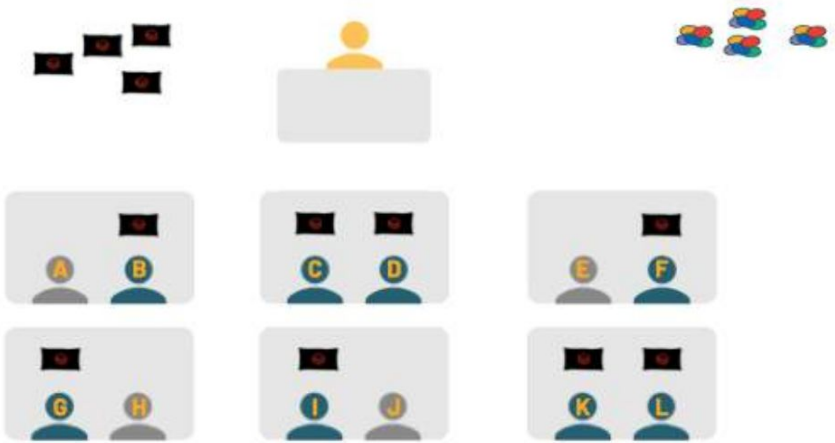
Step 1


Student A get bag # 1
Student B get bag # 2
Student C get bag # 3
...
1
2
3
4
5
6
7
8
9
10
11
12
A
B
C
D
E
F
G
H
I
J
K
L

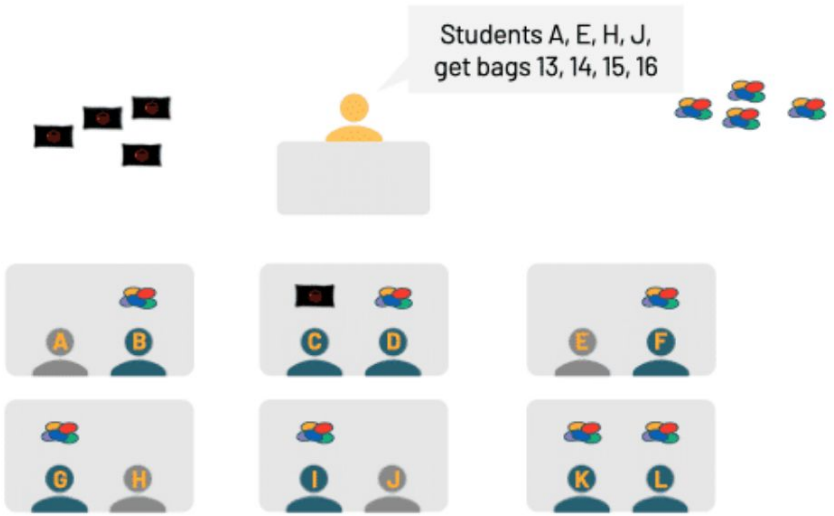
Step 2


Eliminate the brown candy pieces and pile the rest in the corner
A
B
C
D
E
F
G
H
I
J
K
L

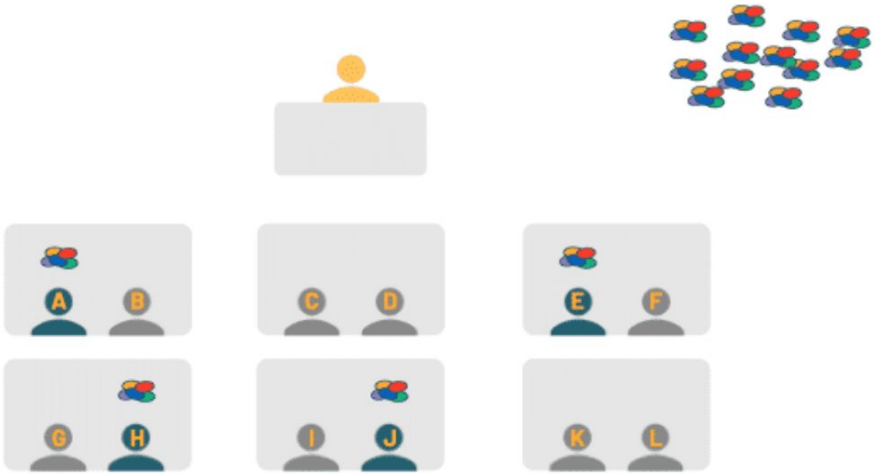
Step 3



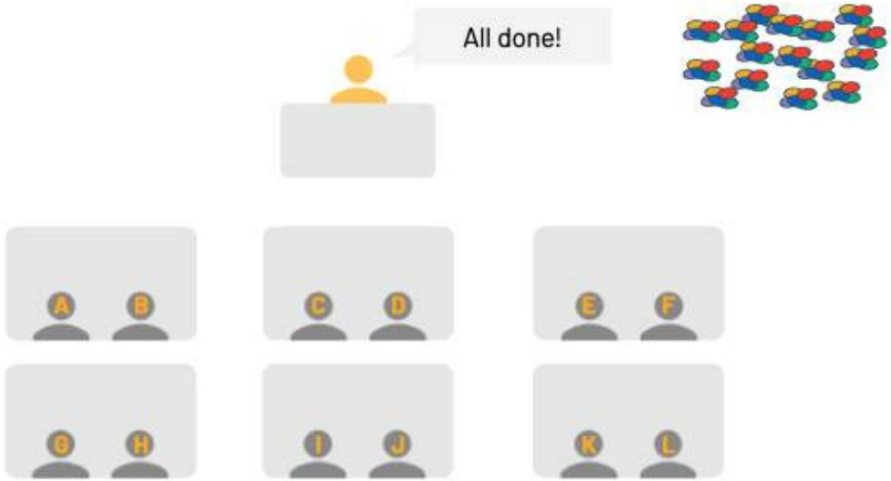
Step 4



Step 5

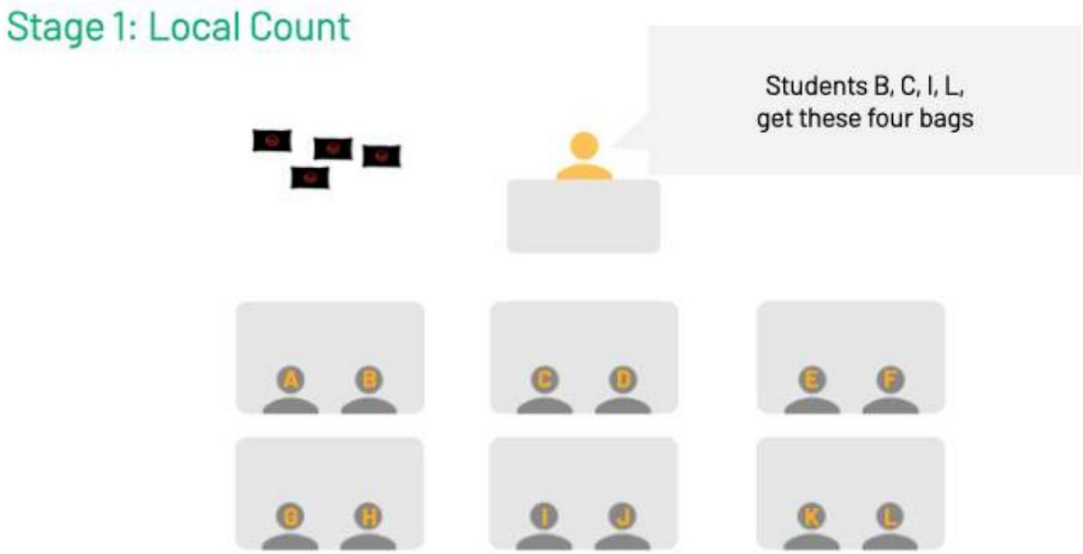


Step 6



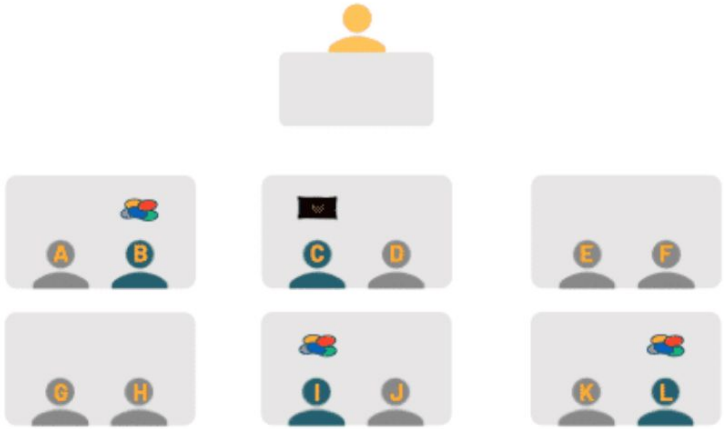
Scenario 2: Count total pieces in candy bags

Step 1


Stage 1: Local Count
Students B, C, I, L,
get these four bags
A
B
C
D
E
F
G
H
I
J
K
L

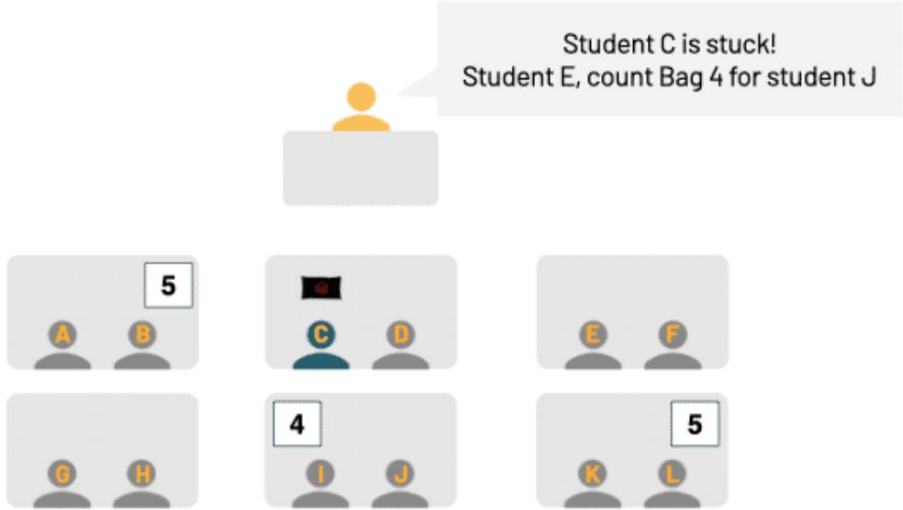
Step 2

Stage 1: Local Count



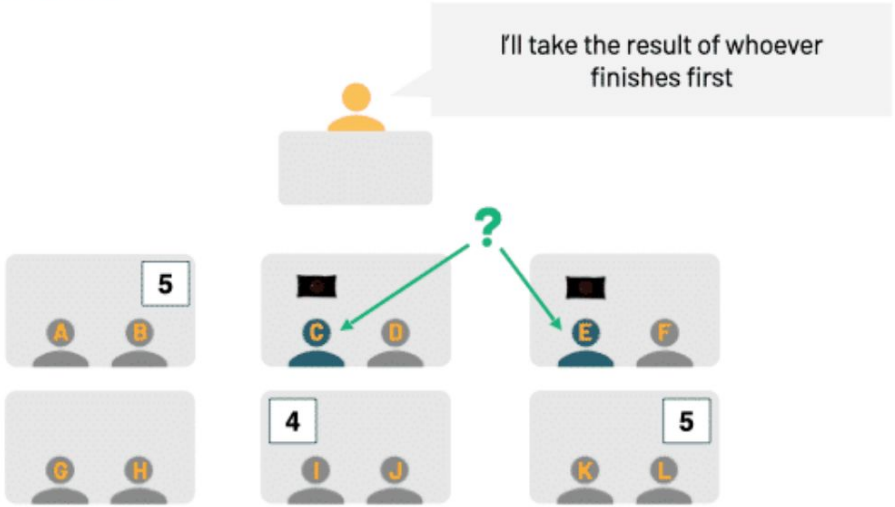
Step 3

Stage 1: Local Count



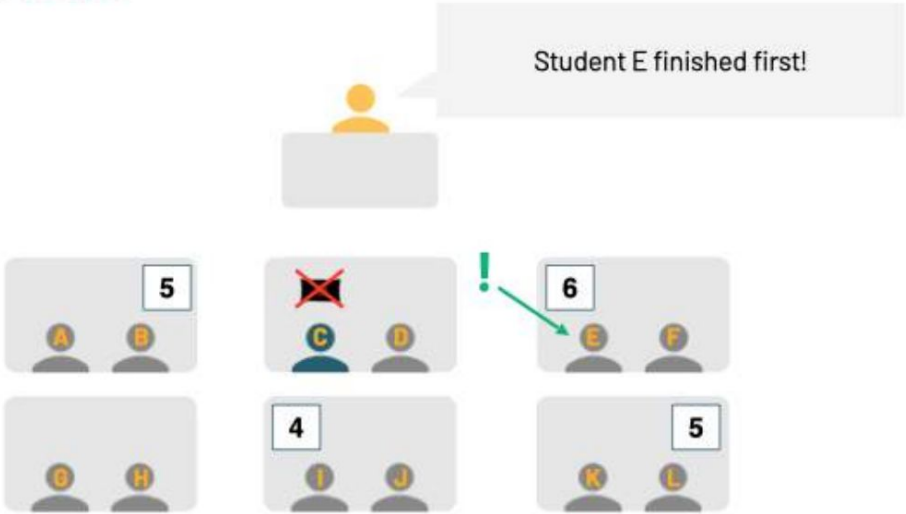
Step 4

Stage 1: Local Count



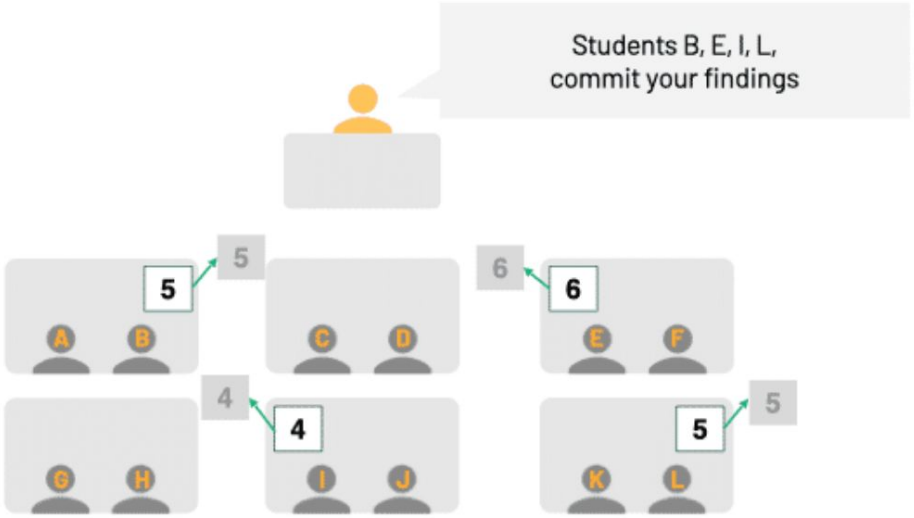
Step 5

Stage 1: Local Count



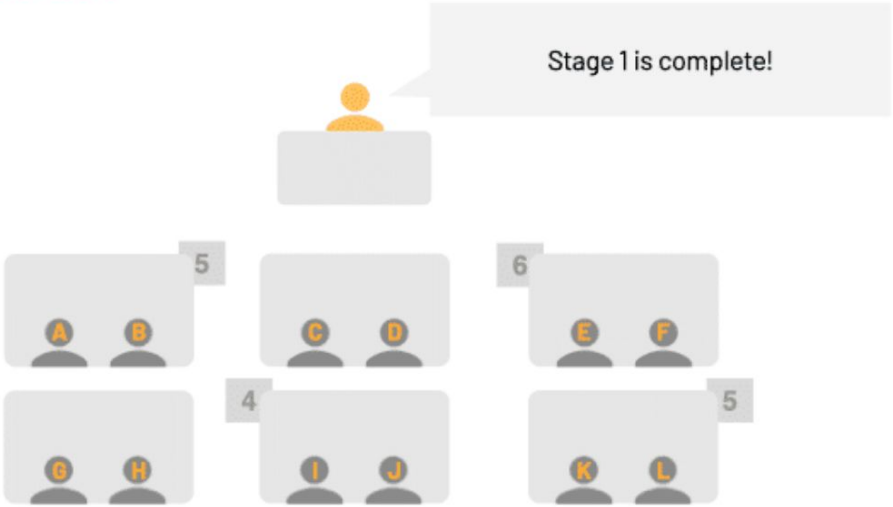
Step 6

Stage 1: Local Count



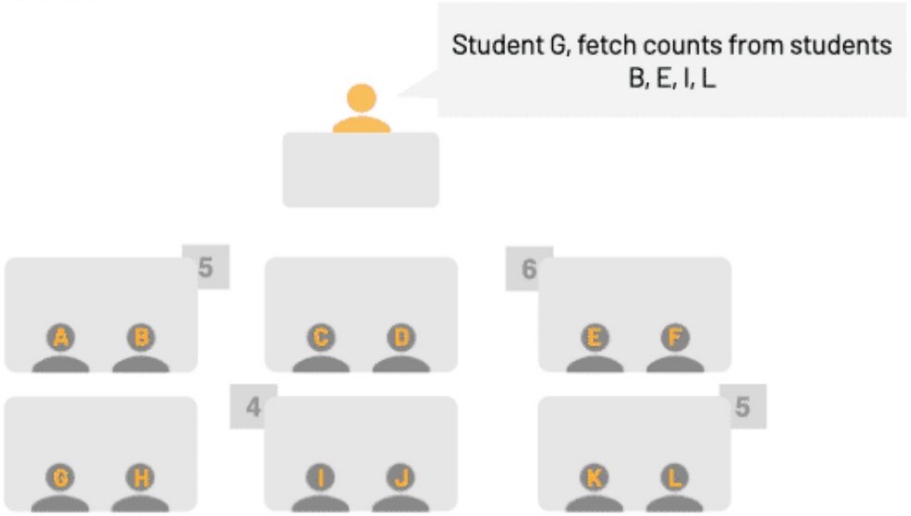
Step 7

Stage 1: Local Count



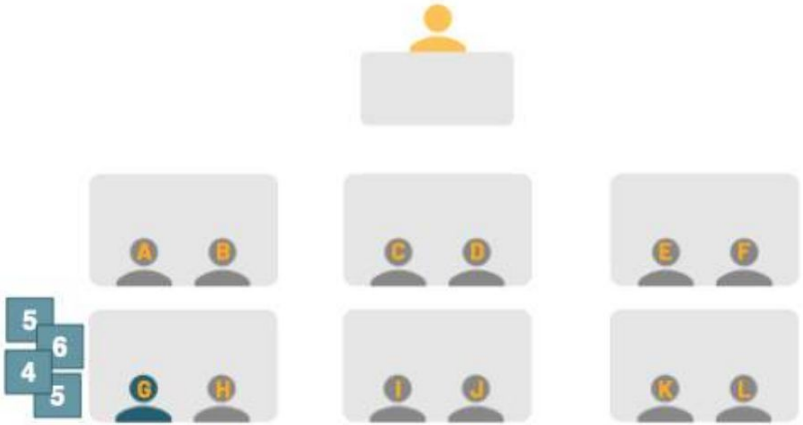
Step 8

Stage 2: Global Count



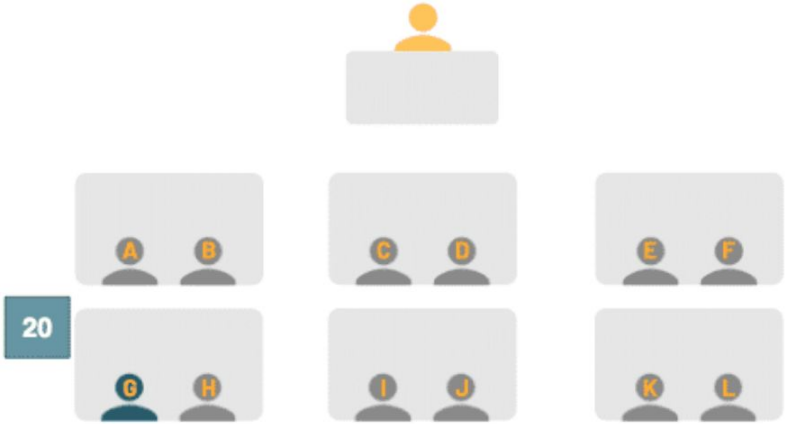
Step 9

Stage 2: Global Count



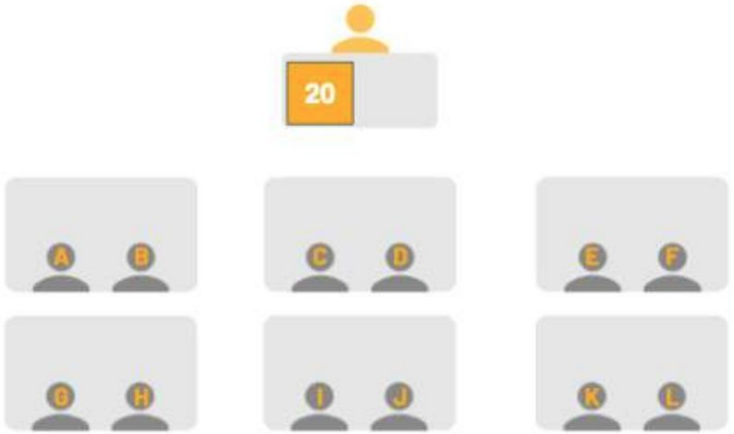
Step 10

Stage 2: Global Count

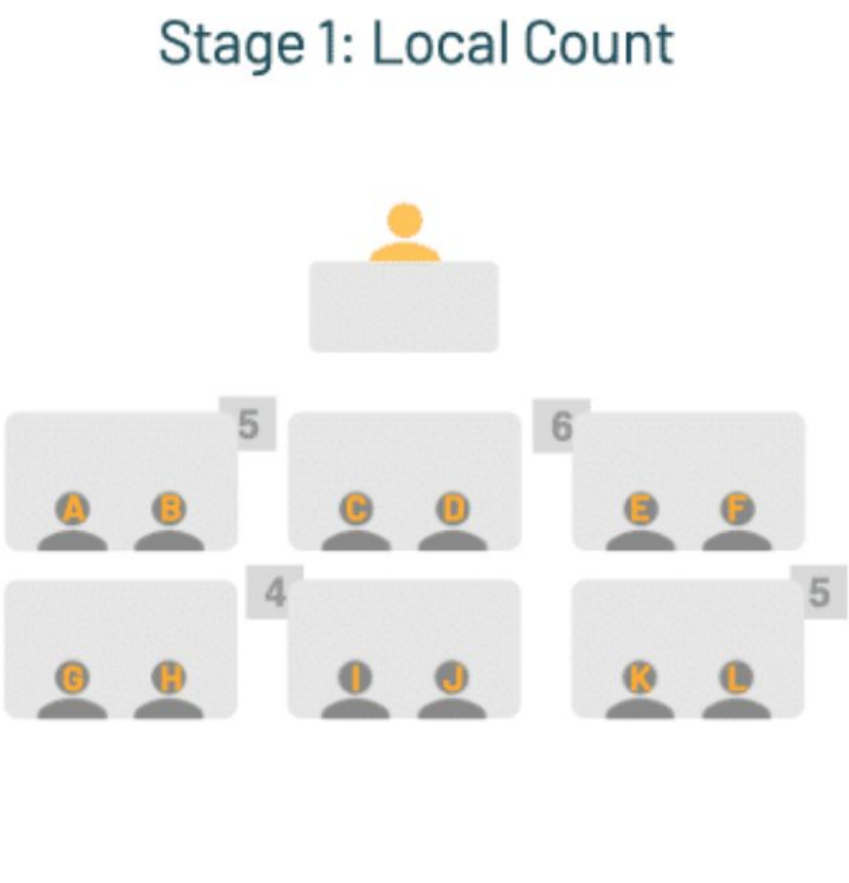


Step 11

Stage 2: Global Count



Summary


Stage 1: Local Count
5
6
A B
C D
E F
4
5
G H
I J
K L


Stage 2: Global Count
20
A B
C D
E F
G H
I J
K L

App □ Jobs □ Stages □Tasks

Spark Cluster