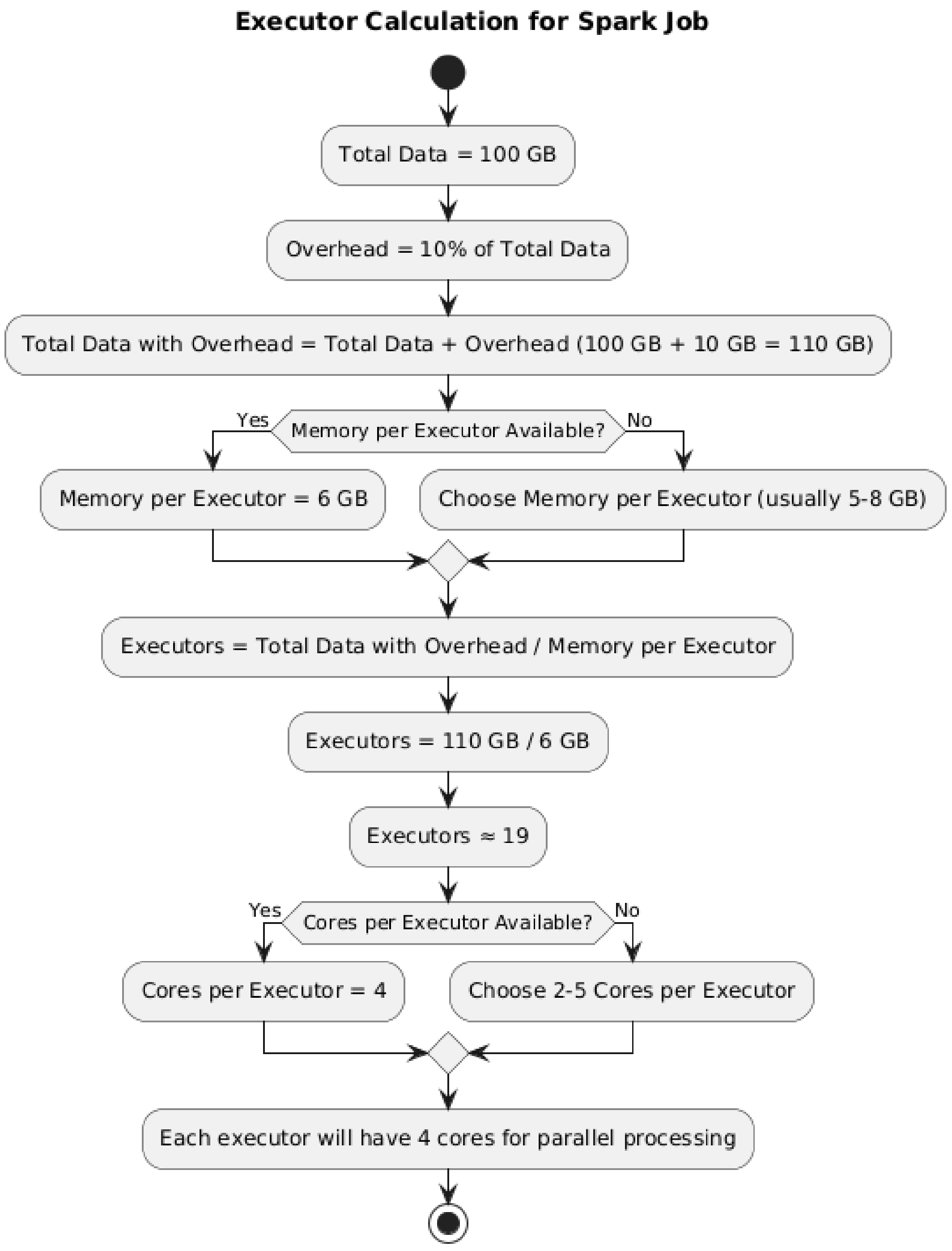
Scenario Series : calculate the number of executors required to process 100 GB

To calculate the number of executors required to process 100 GB of data in Spark, you need to consider several factors like the executor memory, overhead, core count, and how Spark partitions the data. Here's a step-by-step breakdown:



Executor Calculation for Spark Job
Total Data = 100 GB
Overhead = 10% of Total Data
Total Data with Overhead = Total Data + Overhead (100 GB + 10 GB = 110 GB)
Memory per Executor Available?
Yes
No
Memory per Executor = 6 GB
Choose Memory per Executor (usually 5-8 GB)
Executors = Total Data with Overhead / Memory per Executor
Executors = 110 GB / 6 GB
Executors ≈ 19
Cores per Executor Available?
Yes
No
Cores per Executor = 4
Choose 2-5 Cores per Executor
Each executor will have 4 cores for parallel processing

Follow me on LinkedIn – Shivakiran kotur

Determine the Memory per Executor:

The amount of memory allocated to each executor determines how much data it can handle. A general rule is to allocate around 5-8 GB per executor, depending on the workload.

For example, let’s assume 6 GB per executor for your calculation.

Estimate the Data Overhead:

Spark incurs overhead for tasks like serialization, shuffling, etc. A typical overhead estimate is around 10-15% of the data size.

For 100 GB of data, this means:

- Overhead: 10% of 100 GB = 10 GB
- Total data to process: 100 GB + 10 GB overhead = 110 GB

Estimate Data per Executor:

Next, divide the total data by the memory per executor to estimate how much data each executor will process.

For example:

- Total data: 110 GB
- Memory per executor: 6 GB

Number of executors required:

$$\text{Executors} = \frac{\text{Total Data (GB)}}{\text{Memory per Executor (GB)}} = \frac{110}{6} \approx 19 executors$$

Fine-tune Based on Cores:

Each executor should have at least 2-5 cores for optimal performance. Let’s assume 4 cores per executor as a common practice. You’ll need to divide the data processing load by the number of cores.

For example, if your cluster has 4 cores per executor, the number of parallel tasks Spark can handle per executor increases, allowing for better performance and efficiency.

Summary:

To process 100 GB of data:

- Memory per executor: 6 GB
- Estimated overhead: 10 GB (approx.)
- Total executors: ~19 executors
- Each executor should have 4 cores.

This is a general guide. For production workloads, always test and adjust based on your data's specific characteristics (like skewness, partitioning, and complexity of transformations).

Follow me on LinkedIn – Shivakiran kotur