



PySpark Scenario-Based Interview Questions (Complete Notes Series)

DAY 17 — Spark Internals
(DAG, Jobs, Stages, Tasks)



Karthik Kondpak
9989454737

PySpark Scenario-Based Interview

Questions (Complete Notes Series)

DAY 17 — Spark Internals (DAG, Jobs, Stages, Tasks)

Concepts Covered Today

- Spark execution model
- DAG (Directed Acyclic Graph)
- Jobs, Stages, Tasks
- Narrow vs Wide transformations
- Shuffle boundaries
- How actions trigger execution

High-Level Spark Execution Flow

Driver Program



Logical Plan



DAG Scheduler



Stages (Shuffle boundaries)



Tasks (per partition)



Executors

What is a DAG?

A **DAG (Directed Acyclic Graph)** is Spark's internal representation of **transformations and actions**, showing how data flows without cycles.

Spark optimizes execution by analyzing the DAG before running jobs.

⚙️ Question 1: What Triggers a Spark Job?

✅ Correct Answer

Only **actions** trigger Spark jobs.

Examples:

- `count()`
- `show()`
- `collect ()`
- `write()`

Question 2: What is a Job?

◆ Definition

A **Job** is created **for every action** in Spark.

Example:

```
df.count()    # Job 1  
df.write.parquet("/data #)Job 2
```

Question 3: What is a Stage?

◆ Definition

A **Stage** is a set of tasks that can be executed **without shuffle**.

New stage is created **whenever shuffle occurs**.

Narrow vs Wide Transformations

Type

Transformation		Shuffle
map	Narrow	
filter	Narrow	
select	Narrow	
groupBy	Wide	
join	Wide	

Question 4: What is a Task?

◆ Definition

A **Task** is the **smallest unit of execution** in Spark.

- One task per partition per stage
- Executed by executors



Indian Real-Time Scenario

```
df.filter("amount > 1000") \  
  .groupBy("city") \  
  .sum("amount") \  
  .show()
```

Execution Breakdown

- filter → Narrow
- groupBy → Shuffle → New Stage
- show() → Action → Job triggered

Does one Spark job always have one stage?

✓ **Correct Answer**

No. One job can have **multiple stages**, depending on shuffles.

How Many Tasks Will Be Created?

◆ Rule

Number of tasks = number of partitions in that stage.

```
df.rdd.getNumPartitions()
```

Why Spark Jobs Get Stuck at 99%?

✓ Correct Explanation

- Skewed partitions
- One task processing massive data
- Other tasks already finished



**Let's build your Data
Engineering journey
together!**



Call us directly at: 9989454737



<https://seekhobigdata.com/>

