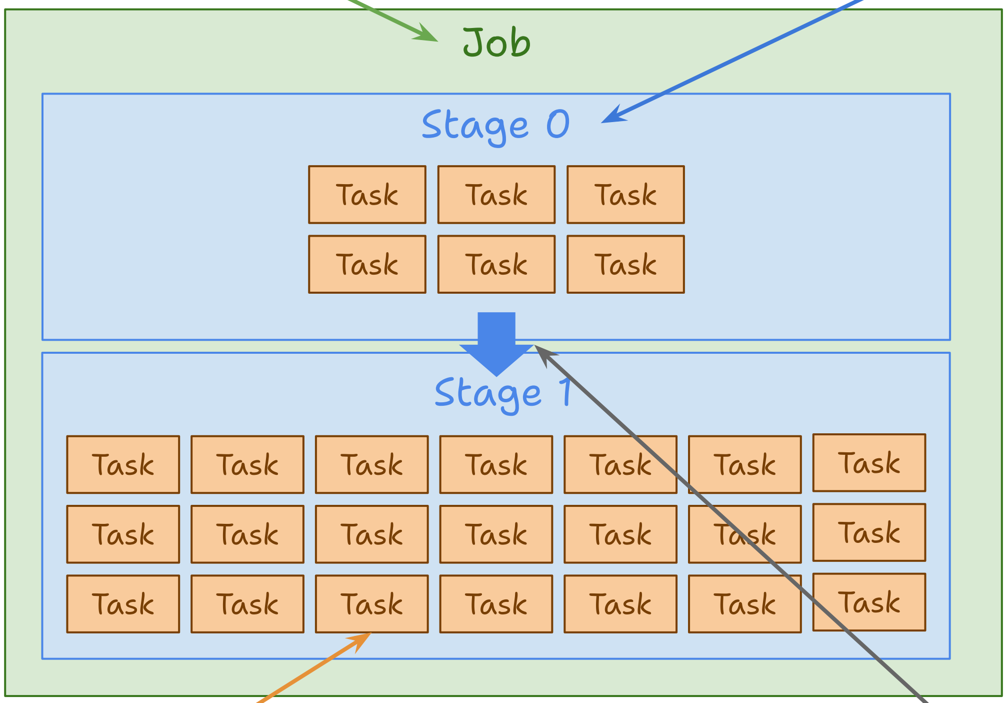
Understanding Jobs, Stages, and Tasks in PySpark

1. What are Jobs, Stages, and Tasks?

A **Job** is triggered by an **action** (like .count(), .collect(), .write(), etc). It represents a complete computation.

A **Stage** represents a set of tasks that can be executed in **parallel**. Each job is divided into stages based on **wide transformations** like a shuffle.



A **Task** is the smallest unit of work. Each stage is divided into tasks, where each task is a **single operation** applied to a **partition** of the data.

A new stage is created whenever there is a **data shuffle**, such as during operations like groupBy, join, distinct, repartition, etc

Let's see an example

Understanding Jobs, Stages, and Tasks in PySpark

2. Given the *customers_df* dataframe:

```
+----------+---------+-----------+------+
|      date|     name|   category|amount|
+----------+---------+-----------+------+
|2024-07-03|Catherine|Electronics|399.99|
|2024-07-01|    Alice|Electronics|1200.5|
|2024-07-04|    Alice|  Furniture| 450.0|
|2024-07-05|      Bob|Electronics|650.75|
|2024-07-02|      Bob|  Furniture| 850.0|
+----------+---------+-----------+------+
```

3. Let's perform some data transformations:

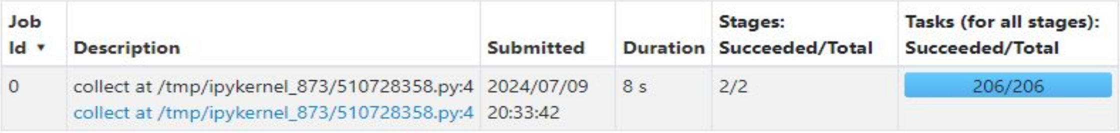
```
df_filtered = (
    customers_df
    .filter(customers_df["Amount"] > 500)
)
category_sales = (
    df_filtered
    .groupBy("category").sum("Amount")
)
result = category_sales.collect()
```

Step 2: A single stage is created to **read** the data and **filter** it.

Step 3: Since **'grouping by'** involves a data shuffle, a new stage is created to group and sum the data.

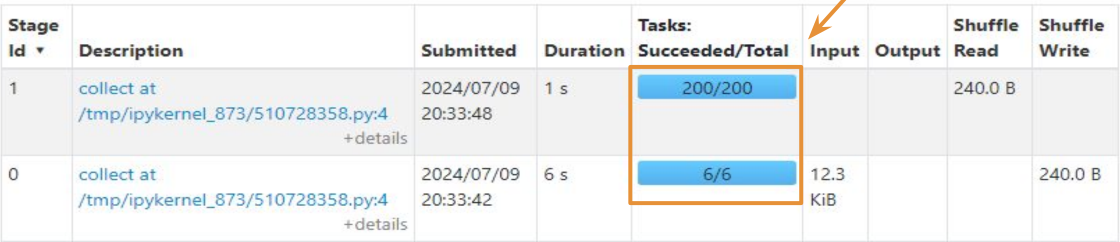
Step 1: The action **.collect()** will trigger a job.

4. How does it look in the Spark UI?

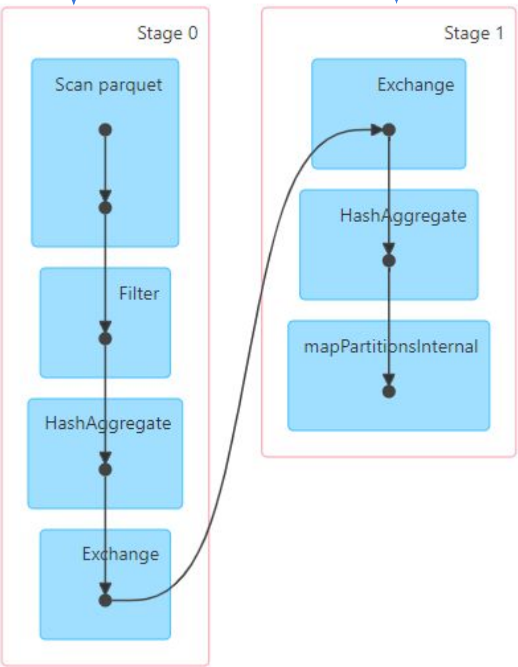


Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0	collect at /tmp/ipykernel_873/510728358.py:4 collect at /tmp/ipykernel_873/510728358.py:4	2024/07/09 20:33:42	8 s	2/2	206/206

Step 4: Within each stage, tasks are created for each partition of the data.



Stage Id ▾	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
1	collect at /tmp/ipykernel_873/510728358.py:4 +details	2024/07/09 20:33:48	1 s	200/200			240.0 B	
0	collect at /tmp/ipykernel_873/510728358.py:4 +details	2024/07/09 20:33:42	6 s	6/6	12.3 KiB			240.0 B



DAG Visualisation