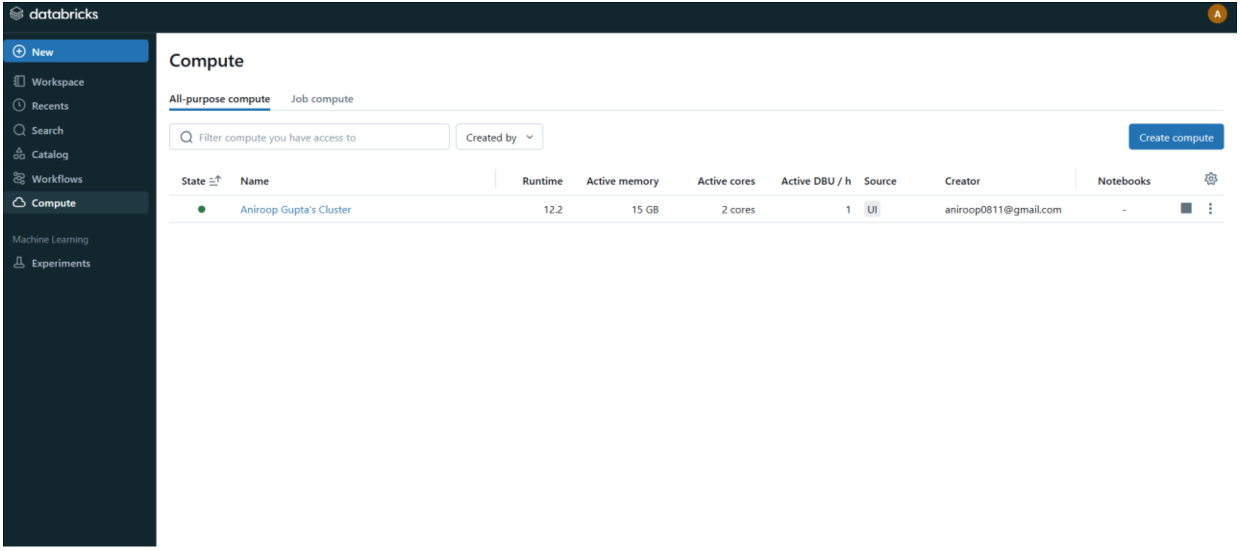
Apache Spark Day1 Assignment

Submitted by - Aniroop Gupta
DE Batch1

Task 1: Write Steps to create a cluster in Databricks.

Steps to create a cluster are:



State	Name	Runtime	Active memory	Active cores	Active DBU / h	Source	Creator	Notebooks
●	Aniroop Gupta's Cluster	12.2	15 GB	2 cores	1	UI	aniroop0811@gmail.com	-

1. Login to the Databricks account and navigate to the dashboard.

2. Click on **New** button from top left corner and click cluster.

3. Give the cluster a name, such as "Subrat Shukla’s Cluster" and configure it to according to requirements.

4. Click on **Create Compute** from the bottom and let the process running.

5. Databricks will initiate the provisioning process, and the cluster will become active once it’s ready.

6. Once the green circle appears, the cluster is ready for use.

Task 2: Explain Spark RDD architecture.

Resilient Distributed Dataset (RDD)-

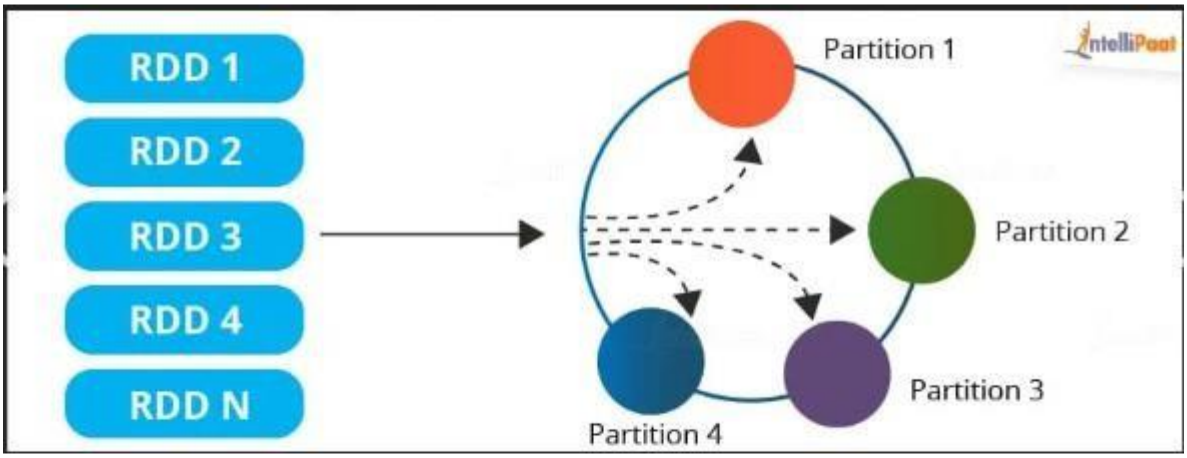
An RDD is an immutable distributed collection of objects that can be processed in parallel. Data in an RDD is divided into logical partitions, each of which is distributed across cluster nodes. Once created, RDDs cannot be modified; transformations generate new RDDs.

RDD supports two types of operations:

- **Transformations:** Operations like map(), filter(), and reduceByKey() that create new RDDs from existing ones.
- **Actions:** Operations like count(), collect(), and saveAsTextFile() that trigger computation and produce results.

Architecture Workflow:

1. **Creation:** RDDs are created from external data (HDFS, S3) or through parallelized collections.
2. **Transformation:** Developers apply transformations to build a computation pipeline.
3. **Execution:** Spark scheduler distributes tasks across nodes, processes partitions in parallel, and applies actions.



--Thank You!