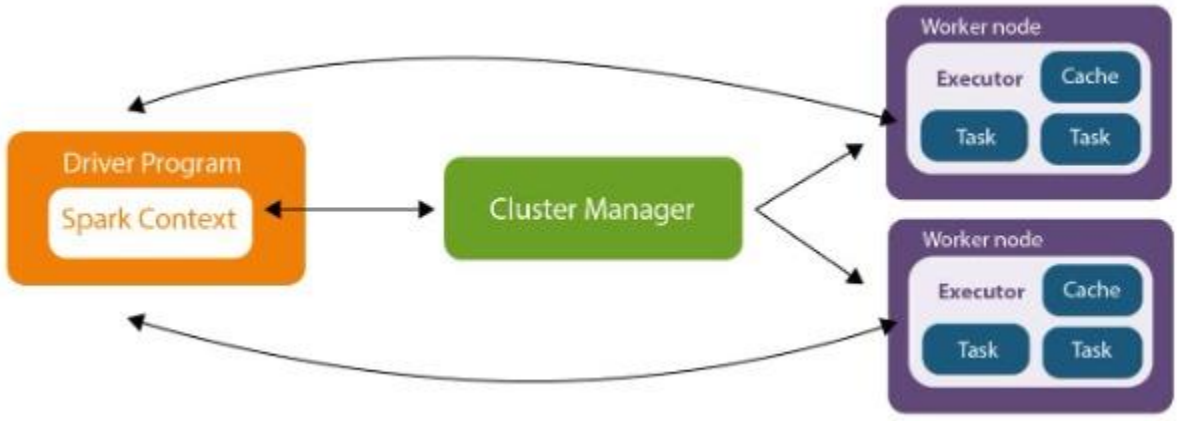
Assignment – Create Cluster

-Sarthak Niranjan Kulkarni (Maverick)

- sarthakkul2311@gmail.com - (+91) 93256 02791

18/11/2024 (Monday)

RDD Architecture:-



- **Driver Program:** It is the central coordinating component in Spark, responsible for managing the lifecycle of an application. It defines RDDs, applies transformations and actions, and constructs the Directed Acyclic Graph (DAG) for execution. The driver communicates with the cluster manager to request resources and schedules tasks on worker nodes for distributed processing.

- **Cluster Manager:** It is responsible for managing resources across the cluster, such as CPU, memory, and executors. It allocates resources to the driver and worker nodes, enabling efficient distributed computation. Examples of cluster managers include YARN, Apache Mesos, and Spark's Standalone Cluster Manager.

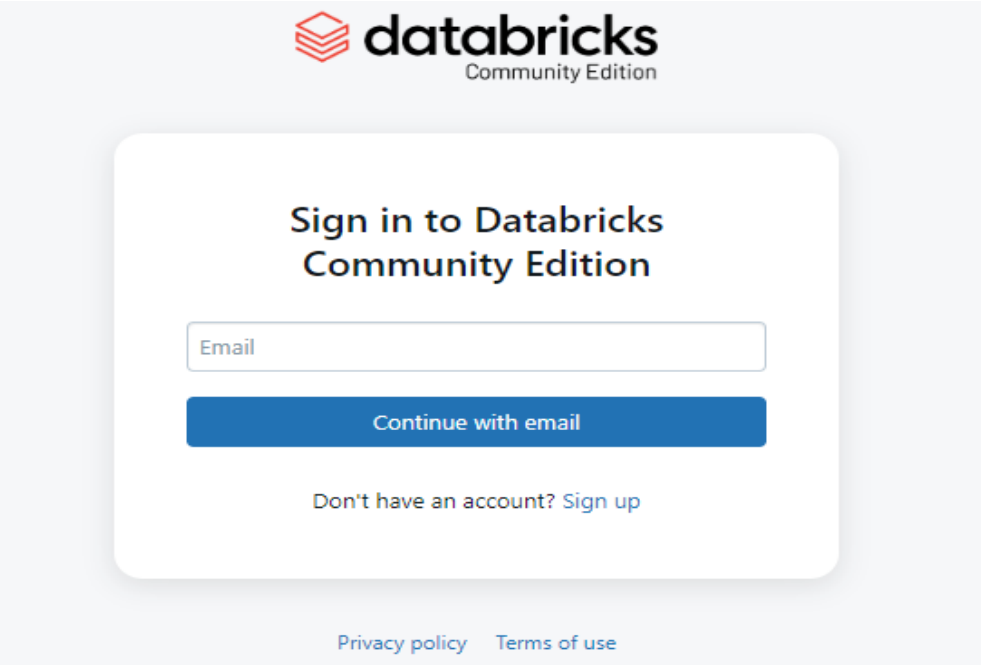
- **Worker Node:** It executes the tasks assigned by the driver program using resources allocated by the cluster manager. Each worker hosts **executors**, which process partitions of RDDs and return results to the driver. Worker nodes also manage intermediate data storage and ensure fault tolerance during computation.

Steps to create Cluster:-

Step 1. :-

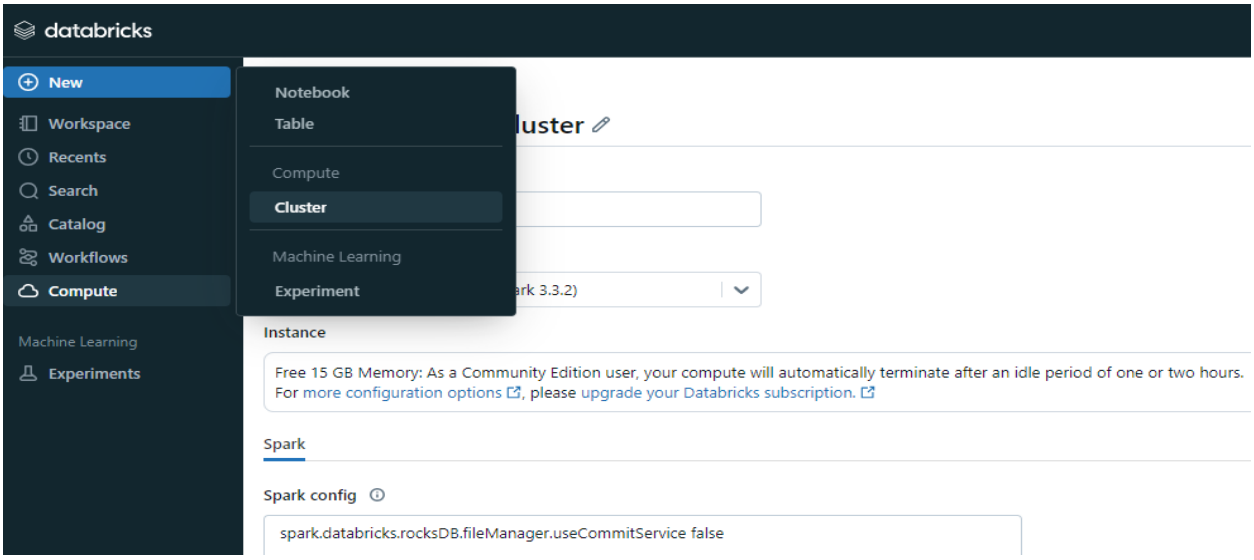
Visit this link to sign up or create an account of Databricks:

https://accounts.cloud.databricks.com/login?tuuid=5c7ca2c3-a2f0-4937-8905-f2edeaa01bab



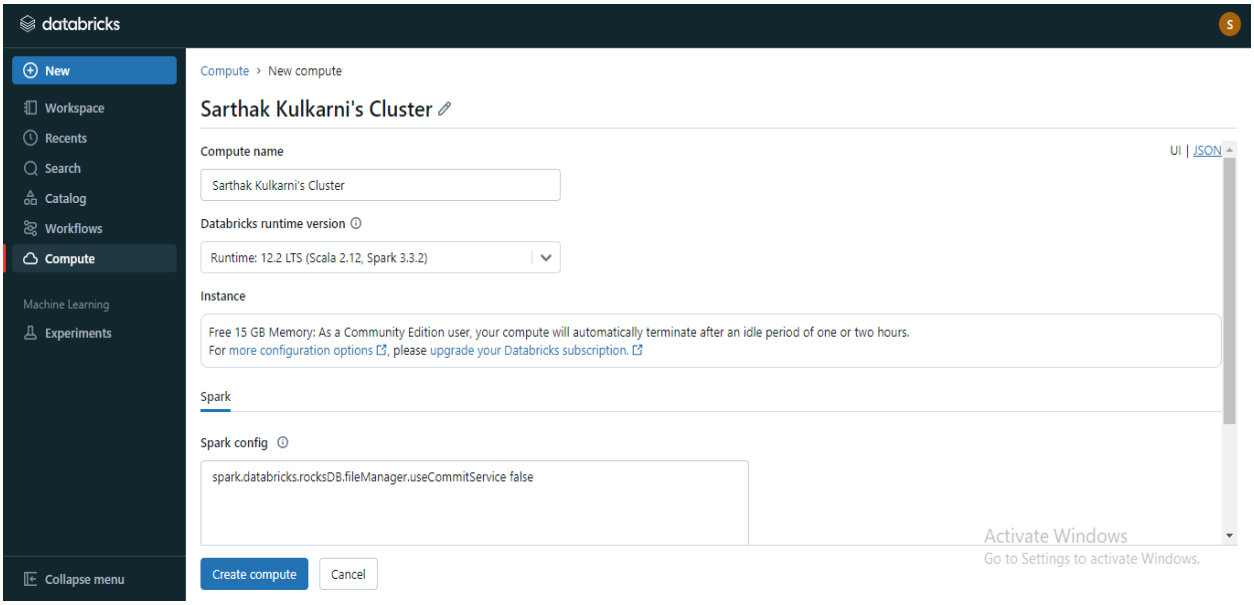
Step 2. :-

Click on sidebar and choose option **NEW** then select **Cluster** option to create a new cluster

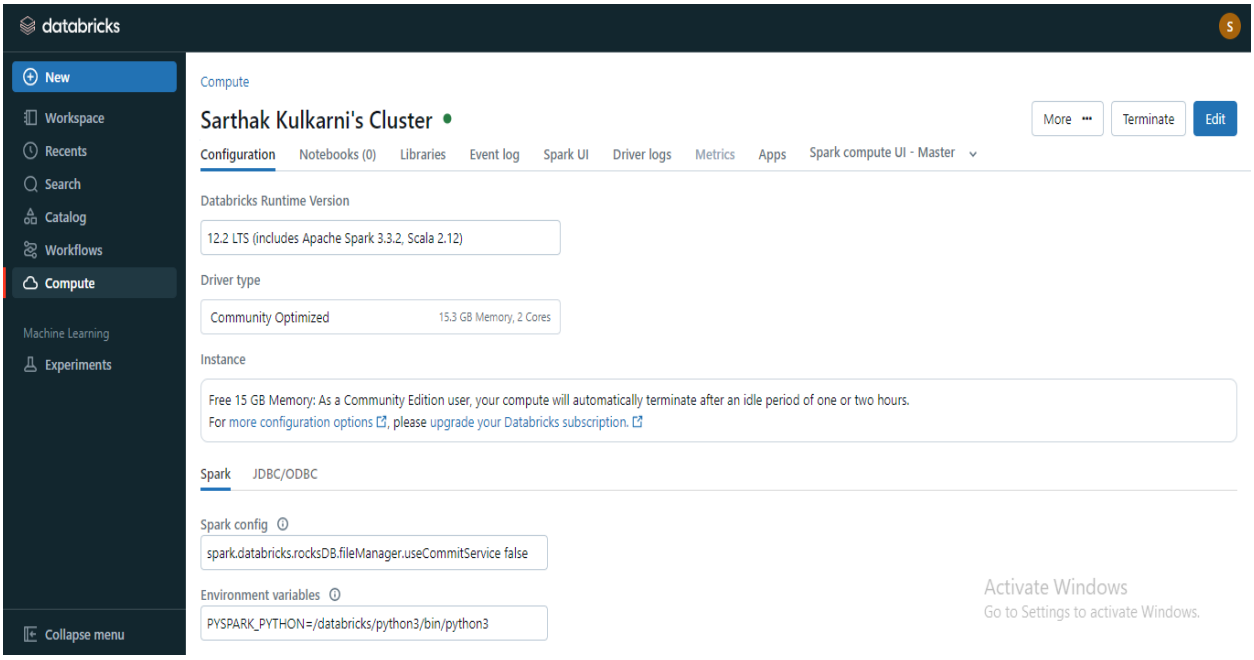


Step 3. :-

Then change the name and select the configurations according to your requirements and click on the option **CREATE COMPUTE** to create the cluster**.**



Below window ensures that your cluster is successfully created: -



Step 4. :-

You can create one Notebook and run a code in it to test functionality of Cluster