

Assignment – Create Cluster

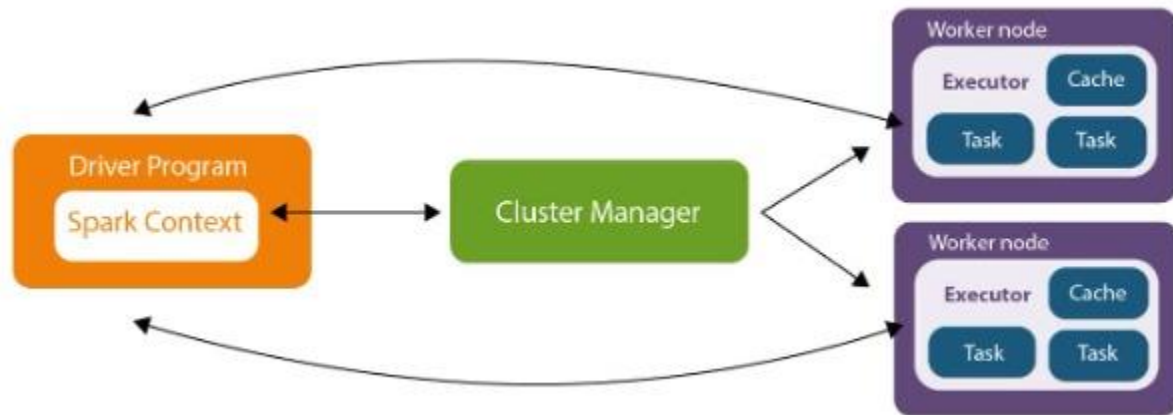
-Sarthak Niranjana Kulkarni (Maverick)

- sarthakkul2311@gmail.com

- (+91) 93256 02791

18/11/2024 (Monday)

RDD Architecture:-



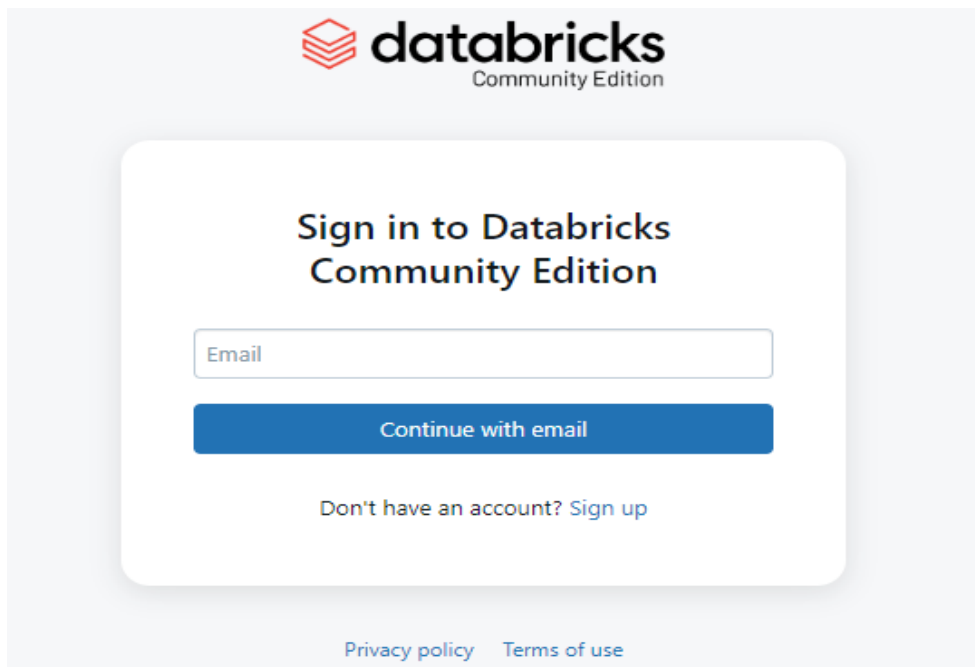
- **Driver Program:** It is the central coordinating component in Spark, responsible for managing the lifecycle of an application. It defines RDDs, applies transformations and actions, and constructs the Directed Acyclic Graph (DAG) for execution. The driver communicates with the cluster manager to request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

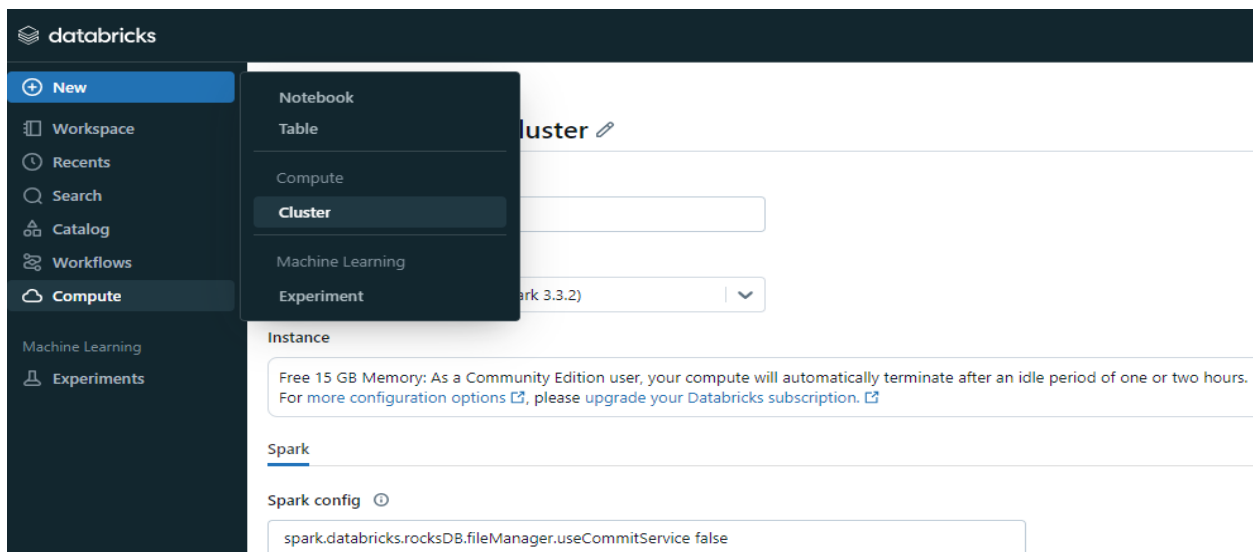
Visit this link to sign up or create an account of Databricks:

<https://accounts.cloud.databricks.com/login?tuuid=5c7ca2c3-a2f0-4937-8905-f2edea01bab>



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.

The screenshot shows the Databricks 'New compute' page. The left sidebar contains navigation links: New, Workspace, Recents, Search, Catalog, Workflows, Compute (selected), Machine Learning, and Experiments. The main content area is titled 'Compute > New compute' and 'Sarthak Kulkarni's Cluster'. It includes a 'Compute name' field with the value 'Sarthak Kulkarni's Cluster', a 'Databricks runtime version' dropdown set to 'Runtime: 12.2 LTS (Scala 2.12, Spark 3.3.2)', and an 'Instance' section with a warning about memory and idle time. Below these is a 'Spark' section with a 'Spark config' field containing 'spark.databricks.rocksDB.fileManager.useCommitService false'. At the bottom, there are 'Create compute' and 'Cancel' buttons. An 'Activate Windows' watermark is visible in the bottom right corner.

Below window ensures that your cluster is successfully created: -

The screenshot shows the Databricks 'Sarthak Kulkarni's Cluster' configuration page. The left sidebar is the same as the previous screenshot. The main content area is titled 'Compute' and 'Sarthak Kulkarni's Cluster'. It includes a 'Configuration' tab and a 'More' menu with 'Terminate' and 'Edit' options. The 'Configuration' tab shows the 'Databricks Runtime Version' as '12.2 LTS (includes Apache Spark 3.3.2, Scala 2.12)', the 'Driver type' as 'Community Optimized' with '15.3 GB Memory, 2 Cores', and the 'Instance' section with the same warning. Below these is a 'Spark' section with 'JDBC/ODBC' as the 'Spark' type, a 'Spark config' field with 'spark.databricks.rocksDB.fileManager.useCommitService false', and an 'Environment variables' field with 'PYSPARK_PYTHON=/databricks/python3/bin/python3'. An 'Activate Windows' watermark is visible in the bottom right corner.

Step 4. :-

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

