

Apache Spark Assignment-1

Name: **Harish Er**

Q) Architecture of Spark.

Apache Spark is a distributed computing framework designed for large-scale data processing. Its architecture ensures high performance, scalability, and versatility, making it suitable for a wide range of data analytics tasks. Here's a breakdown of its key components:

1. Core Components:

- **Driver:**
 - a. The central coordinating entity that manages the SparkContext.
 - b. Converts user-defined code into a Directed Acyclic Graph (DAG) of tasks and schedules their execution on worker nodes.
- **Cluster Manager:**
 - a. Allocates resources to Spark applications.
 - b. Supported managers include Spark Standalone, Apache Mesos, Hadoop YARN, and Kubernetes.

2. Execution Layer:

- **RDDs (Resilient Distributed Datasets):**
 - a. Immutable distributed collections of objects forming Spark's core abstraction.
 - b. Operations on RDDs are lazy and only executed when an action (e.g., collect, count) is invoked.

3. Storage Layer:

- Spark supports in-memory storage for intermediate results, which accelerates iterative algorithms.

4. Execution Modes:

- **Standalone Mode:** Spark runs independently with its cluster manager.

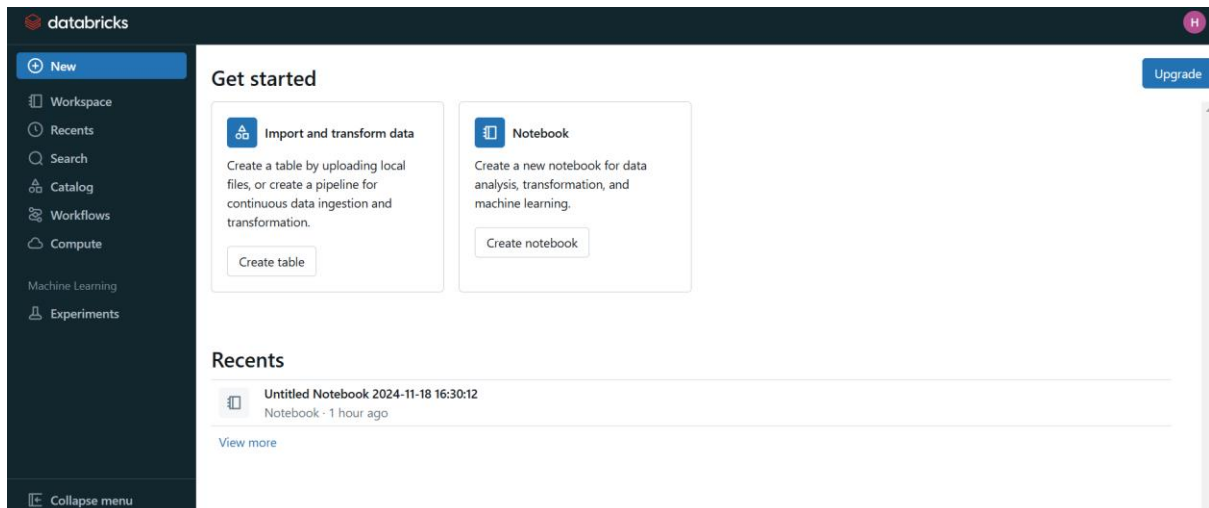
- **Cluster Mode:** Applications are deployed and managed using external cluster managers.
- **Local Mode:** Ideal for development and testing, everything runs on a single machine.

5. Libraries and APIs:

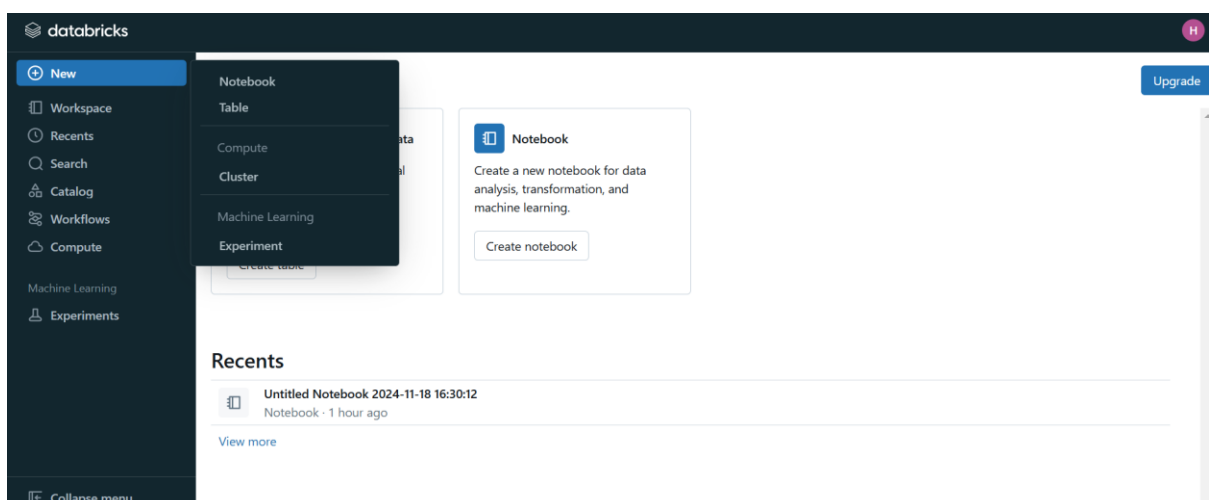
- Built on the Spark Core, various libraries extend its functionality
- Supports APIs in multiple languages (Scala, Python, Java, R).

Q) Steps of building a cluster in Spark.

Step 1: Login in the home page.



Step 2: Click on New -> Cluster:



Step 3: Create Cluster:

The screenshot shows the Databricks 'New compute' page. The left sidebar contains navigation links: New, Workspace, Recents, Search, Catalog, Workflows, Compute (highlighted), Machine Learning, and Experiments. The main content area is titled 'Harish E R's Cluster' and includes the following fields:

- Compute name:** A text input field containing 'Harish E R's Cluster'.
- Databricks runtime version:** A dropdown menu showing 'Runtime: 12.2 LTS (Scala 2.12, Spark 3.3.2)'.
- Instance:** A text area containing the message: 'Free 15 GB Memory: As a Community Edition user, your compute will automatically terminate after an idle period of one or two hours. For more configuration options, please upgrade your Databricks subscription.'
- Spark:** A section header.
- Spark config:** A text input field containing 'spark.databricks.rocksDB.fileManager.useCommitService false'.

At the bottom of the form are two buttons: 'Create compute' and 'Cancel'.

Step 4: View Cluster:

The screenshot shows the Databricks 'Compute' page. The left sidebar is the same as in Step 3. The main content area is titled 'Compute' and has two tabs: 'All-purpose compute' (selected) and 'Job compute'. Below the tabs is a search bar with the placeholder text 'Filter compute you have access to' and a 'Created by' dropdown menu. A 'Create compute' button is located in the top right corner. Below these elements is a table listing the clusters.

State	Name	Runtime	Active me...	Active cores	Active DB...	Source	Creator	Notebooks	
●	Harish E R's Cluster	12.2	15 GB	2 cores	1	UI	harish.er562@gm...	-	

At the bottom right of the page, there is a pagination control showing '1' and '20 / page'.