**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**:
  1. The central coordinating entity that manages the SparkContext.
  2. Converts user-defined code into a Directed Acyclic Graph (DAG) of tasks and schedules their execution on worker nodes.
* **Cluster Manager**:

1. Allocates resources to Spark applications.
2. Supported managers include Spark Standalone, Apache Mesos, Hadoop YARN, and Kubernetes.
3. **Execution Layer:**

* **RDDs (Resilient Distributed Datasets)**:

1. Immutable distributed collections of objects forming Spark’s core abstraction.
2. 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.

1. **Execution Modes**:

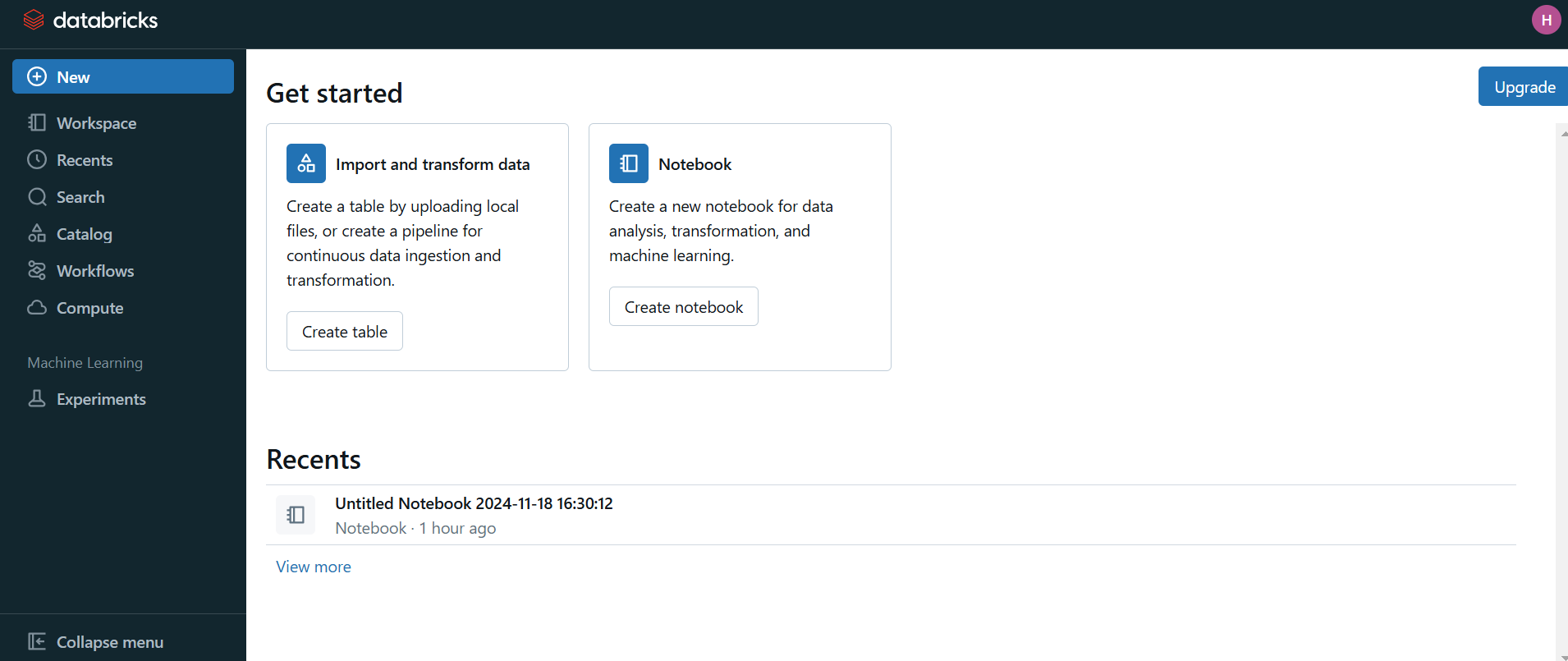
* **Standalone Mode**: Spark runs independently with its cluster manager.
* **Cluster Mode**: Applications are deployed and managed using external cluster managers.
* **Local Mode**: Ideal for development and testing, everything runs on a single machine.

1. **Libraries and APIs**:

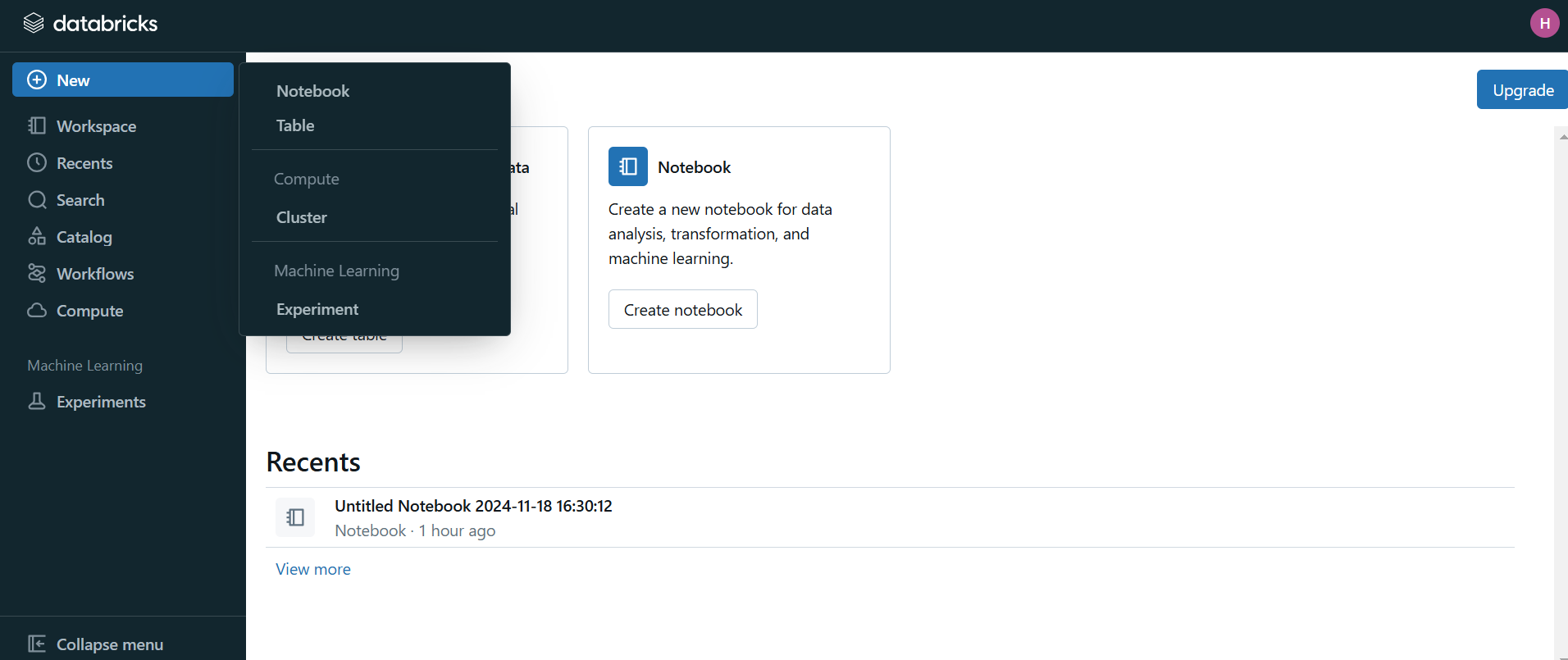
* Built on the Spark Core, various libraries extend its functionality
* Supports APIs in multiple languages (Scala, Python, Java, R).

**Q) Steps of building a cluster in Spark.**

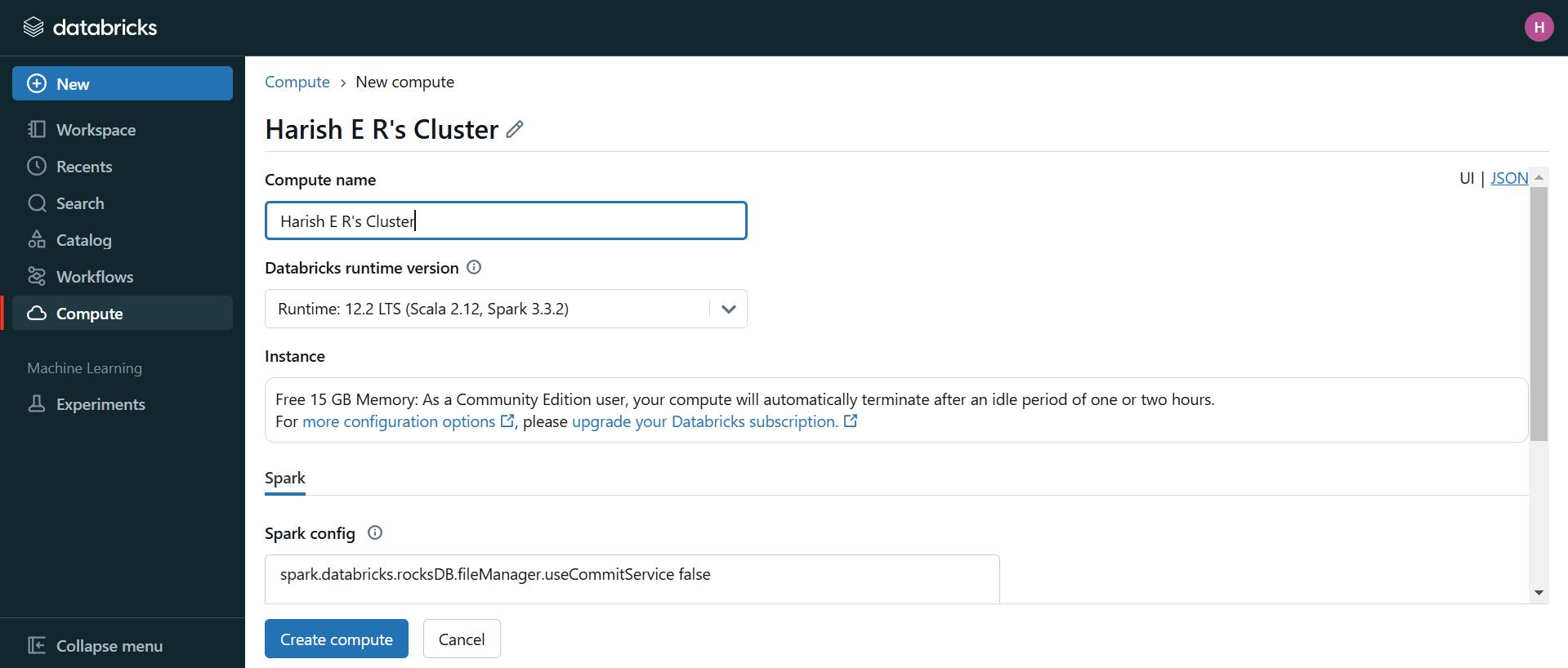
Step 1: **Login in the home page.**



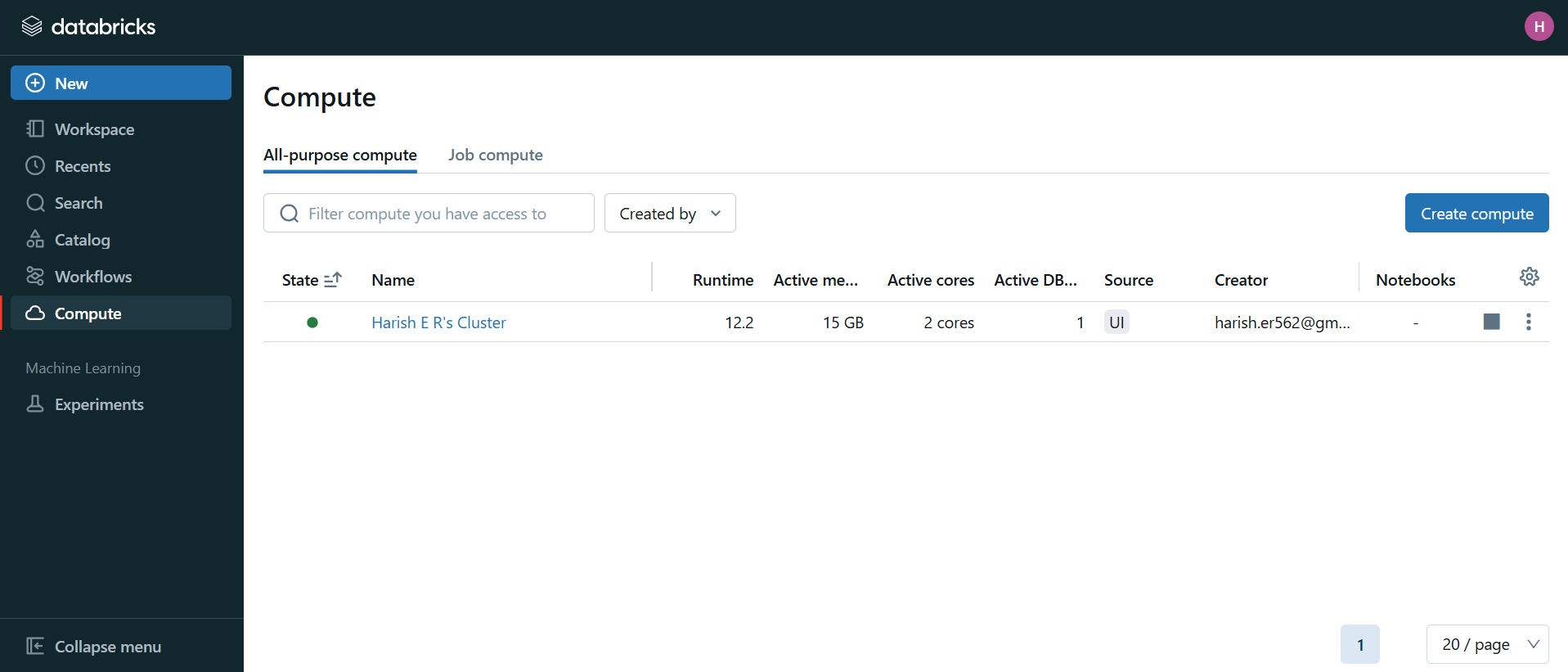
Step 2: **Click on New -> Cluster:**

****

Step 3: **Create Cluster:**



Step 4: **View Cluster:**

****