**SPARK CLUSTER ASSIGNMENT**

**Name: S.Nikhil Sai**

**Date: 18-11-2024**

**1. Spark RDD:**



**Expalnation:**

The image specifically illustrates Apache Spark's RDD (Resilient Distributed Dataset) architecture and its key characteristics. It illustrates the architecture of a distributed data processing system, specifically showing how data flows from storage to results through distributed processing.

**Breakdown:**

* "Data on Disk" represents the raw data storage (shown as a cylinder symbol typically used for databases/storage)
* "HDFS read" indicates the data is being read from a Hadoop Distributed File System (HDFS)
* "One Time Processing" suggests this is a batch processing operation
* The data moves into "Distributed Memory" (shown as a darker rectangular block)
* This represents data being loaded into memory across multiple nodes in a distributed system
* Multiple queries (Query1, Query2, Query3) can be executed in parallel against the same distributed dataset
* Each query produces its own result (Result1, Result2, Result3)

**Process:**

* Data is loaded once from persistent storage
* The loaded data is distributed across multiple nodes' memory
* Multiple queries/transformations can be performed on the same dataset without reloading from disk
* This approach optimizes performance by minimizing I/O operations and enabling parallel processing.

**2. Creating the cluster in Data bricks:**

**Step 1:** After the successful creation of data bricks account, click on ‘compute’ on the left side bar.

**A screenshot of a computer

Description automatically generated**

**Step 2:** After getting into compute section, click on create compute option that appears on right top.

A screenshot of a computer

Description automatically generated

**Step 3:** Now, name the cluster and give appropriate details to create a new cluster**.** And click on create compute option that appears on bottom left.

**A screenshot of a computer

Description automatically generated**

**Step 4:** The cluster might take a couple of minutes to get created. Once the cluster is created, click on the compute option that appears on left sidebar to view the newly created cluster.

A screenshot of a computer

Description automatically generated

Thus, the cluster is created:

A screenshot of a computer

Description automatically generated