**Big Data Systems – Assignment 2**

**Group 124:**

Ashutosh Kumar– 2020fc04341

Rahul Paul – 2020fc04350

Ashwini BR – 2020fc04180

**Solution Implementation Overview:**

***Data Write:***

Diagram

Description automatically generated

**Legend: Thick lines RW access. Thin lines are RO access**

Apache Hadoop clusters, built on Docker containers using Dockerfile and Docker-compose is used to demonstrate the working model. There are a total of 9 Hadoop clusters – 8 for storing data and 1 for storing MetaData (for DataRetrieval).

Each cluster consists of 5 containers., namely - NameNode, DataNode, ResourceManager, NodeManager and HistoryServer. All have unique container names and exteranal ports configured. The sample Docker Cluster setup files for 4 such clusters are included in the assignment submission.

The required storage Optimization is achieved using **Docker Volumes.**

Each cluster has one unique Docker Volume mounted in ReadWrite mode on the datanode in the cluster. The same DockerVolume is accessible by other clusters (based on replication factor) in ReadOnly mode.

Here, Replication factor is assumed to be same as Repetition factor

**For Example:** A data value occurring only once in a column is written to Cluster 1 and in turn to a Docker Volume that is accessible only by cluster1 in RW mode. A data value occuring 5 times in a column is written to Cluster5 and in turn to a Docker Volume that is accessible in RW mode on Cluster 5 but also in RO mode on Cluster1, Cluster2, Cluster3 and Cluster4. Thus, the same data is read accessible from 5 clusters simultaneously, while only 1 cluster can perform the write operation.

The number of Data clusters was decided as 8 based on the data preprocessing and EDA performed on the given dataset. It was observed that the columns “Course Name”, “Course Rating”, “Course Description” and “Skills” have counts ranging from 1 to 8, while the columns - “University” and “Difficulty Level” had a wider range. A modulo operation is performed on the count values of these columns to determine the DataNode on which the data shall reside.

The data written to datanodes consists of Actual\_Data\_Value and Record\_ID\_list ie., list of records in the original dataset having those values.

The metadata is with Record\_ID\_list and Node\_ID is written to the metadata cluster.

**DataFrame Screenshots for Write:**

Graphical user interface, text

Description automatically generated

**Data Written into DockerVolume:**

**Graphical user interface

Description automatically generated with low confidence**

**MetaData Written:**

**Record ID Node ID**

**Graphical user interface, application

Description automatically generated**

***Data Read:***

*Diagram

Description automatically generated*

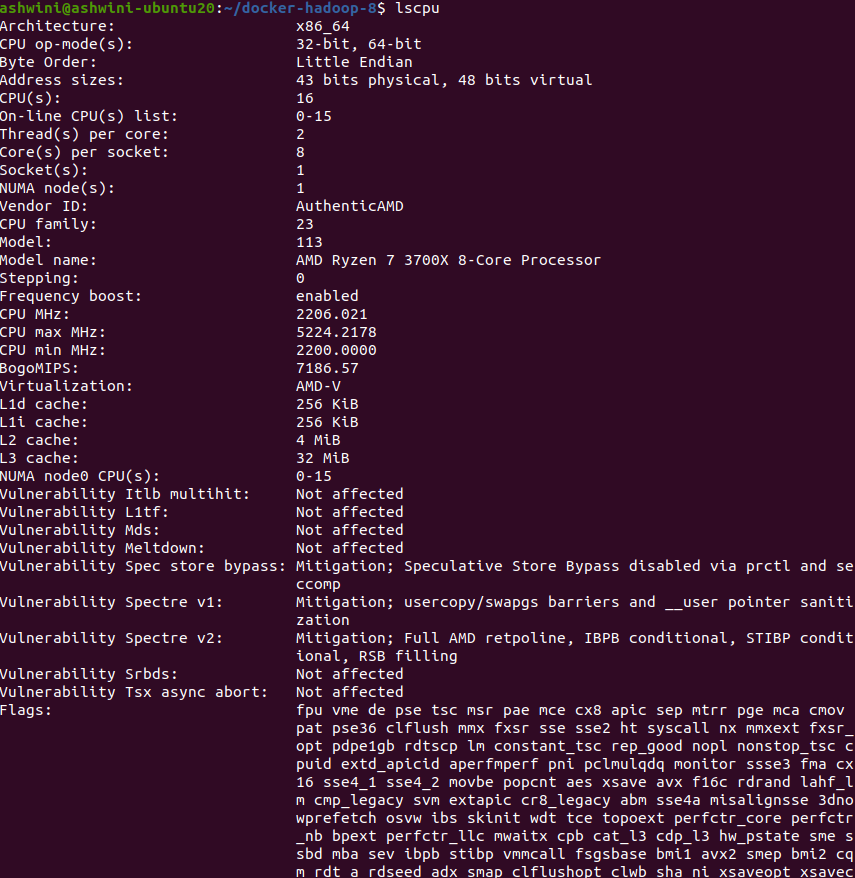
**Legend: Thin lines indicate read operations**

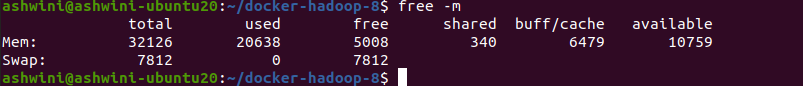
To read the data, the Read program takes input from the user on the record to be fetched. The program then connects to the metadata node to search for the in the metadata from Record\_ID\_list and get the corresponding Node\_ID.

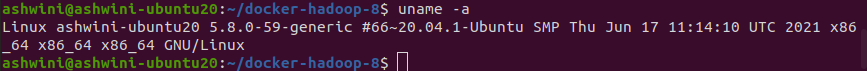
A request is then made to the corresponding Cluster where the record ID is again searched from the Record\_ID\_list to get the corresponding Actual\_Data\_Value. Thus, retrieving back the data.

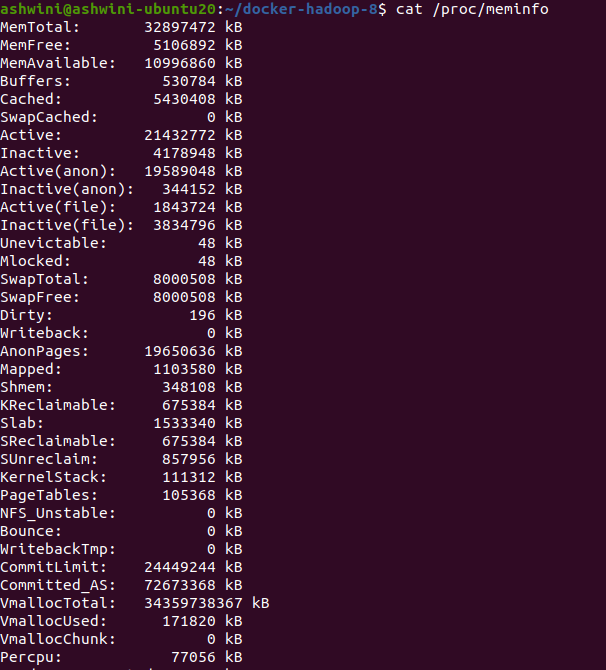
**System Specifications:**

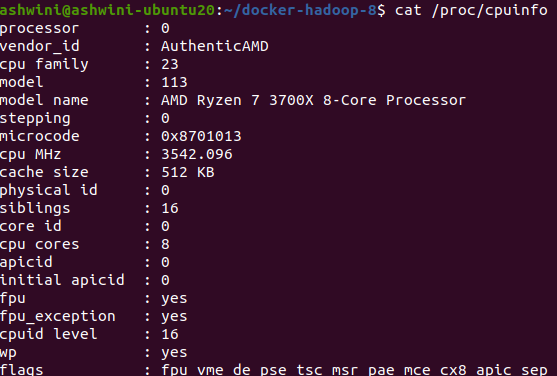
|  |  |
| --- | --- |
| *Operating System* | Ubuntu 20 |
| *CPU* | 32 cores |
| *Memory* | 32GB |
| *Storage* | 200G |

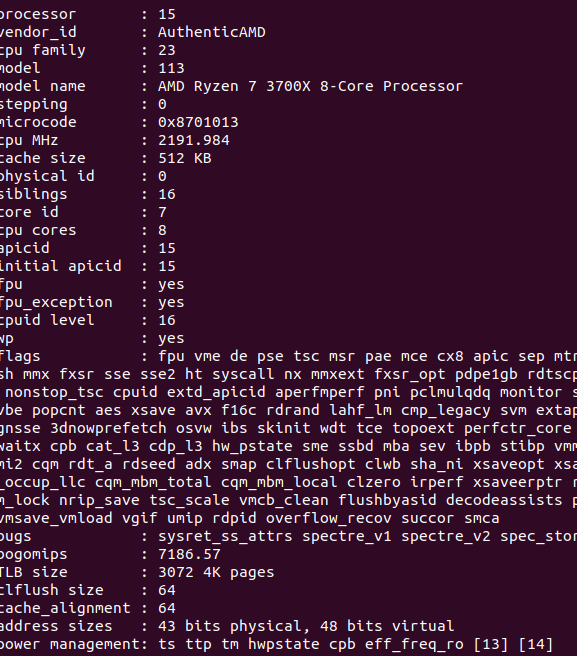




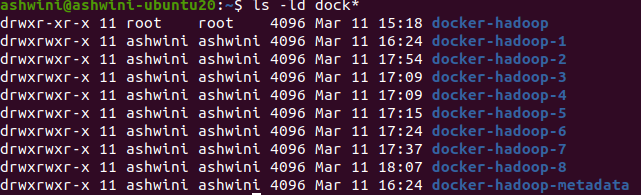


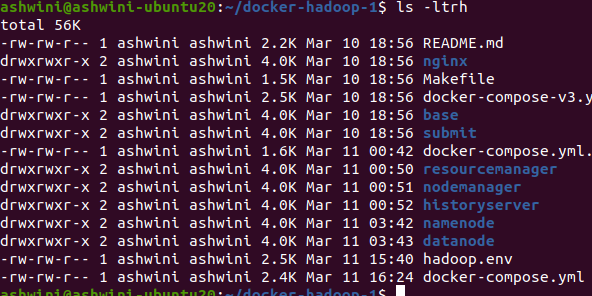




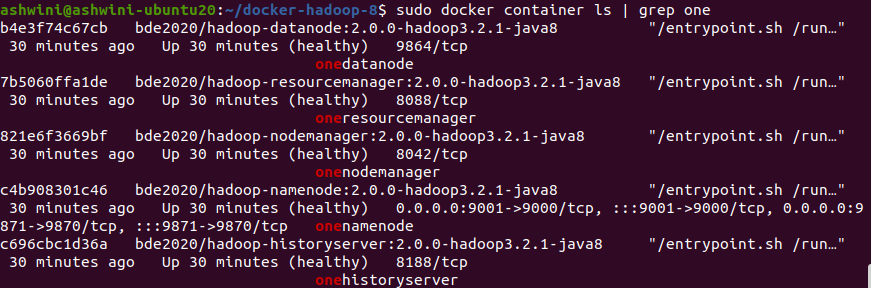


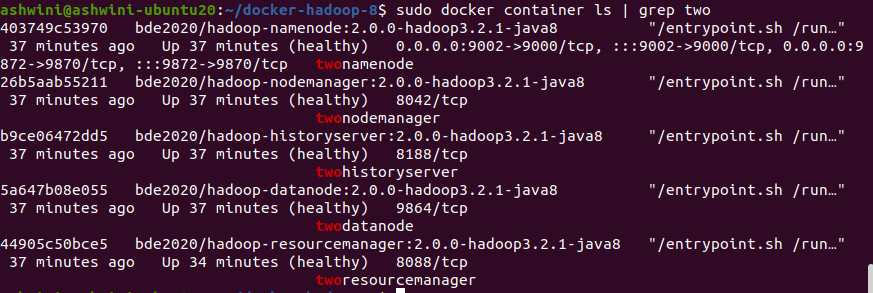
**Docker Clusters Directory Structure:**

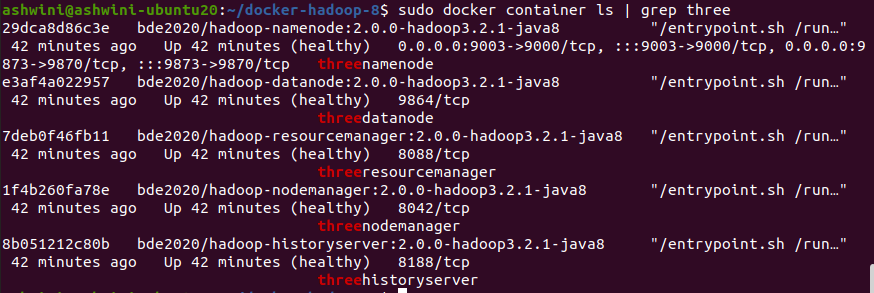


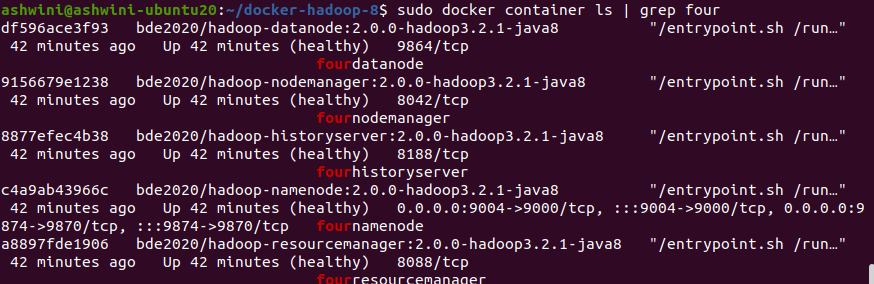


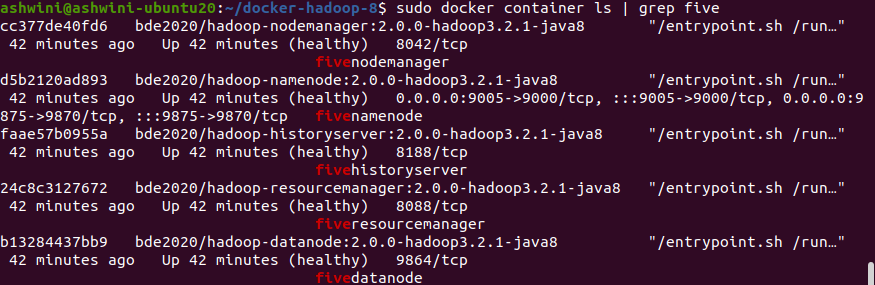
**Docker Daemons running for each Cluster**

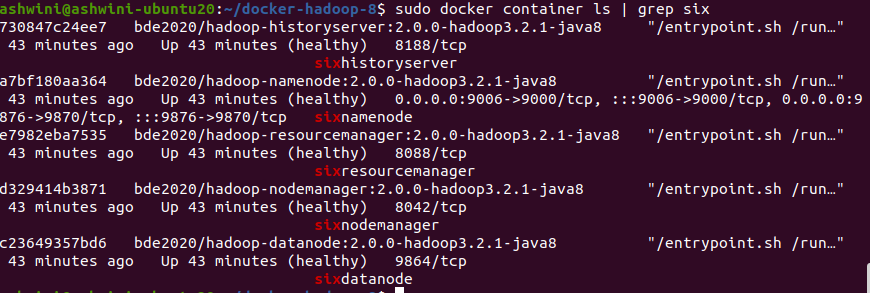


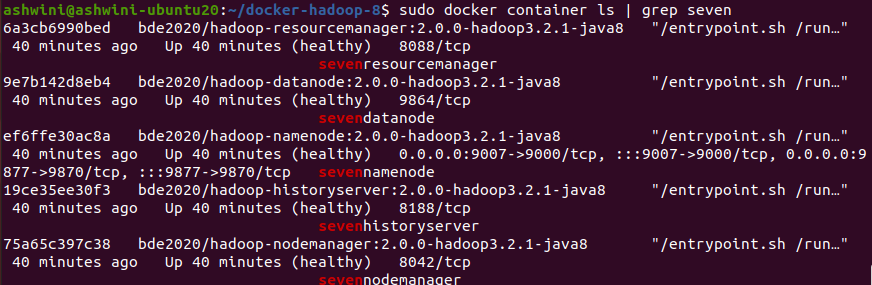


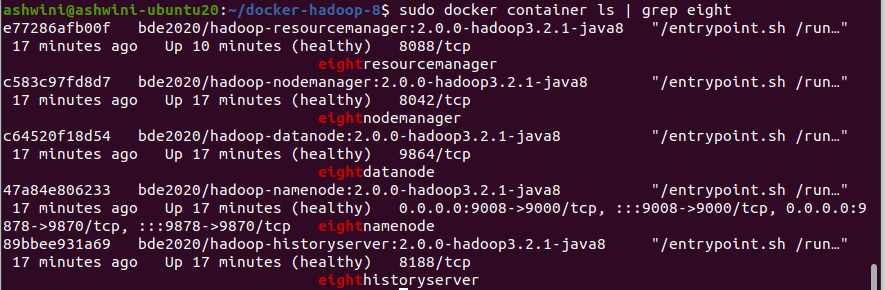




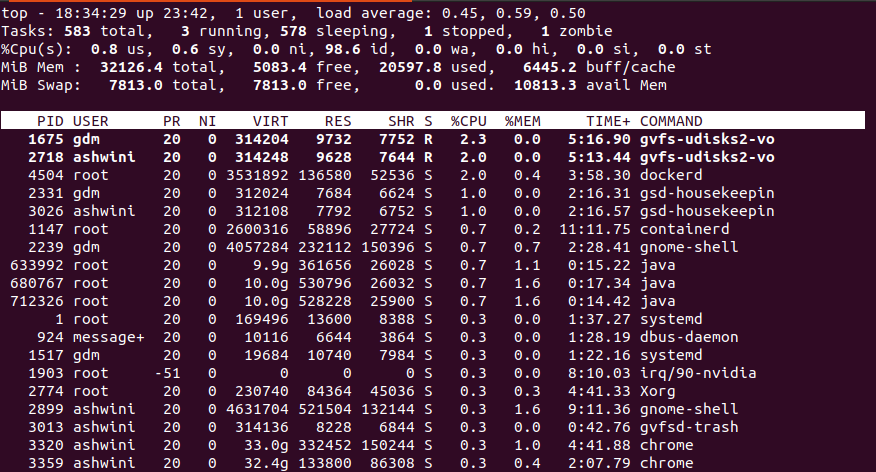








**System Usage when all 8 clusters are up**

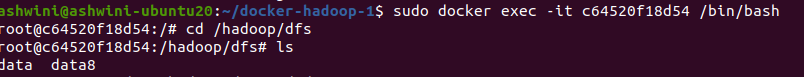
****

**DockerVolumes inside DataNode Containers**

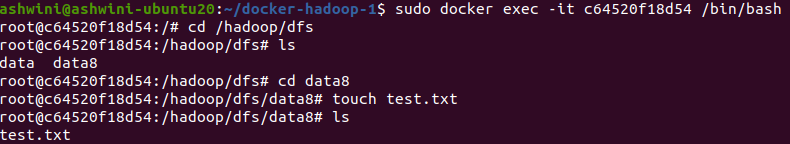
C64520f18d54 – DataNode container of Cluster8

C23649357bd6 – DataNode container of Cluster6

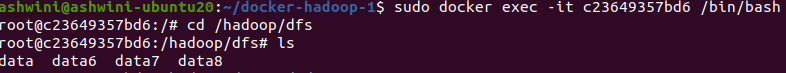
Docker Volumes accessible by Cluster8

****

Docker Volume /data 8 in RW mode on Cluster8

****

Docker Volumes accessible by Cluster6



DockerVolume /data8 of Cluster8 is ReadOnly on Cluster6. While /data6 of Cluster6 is accessible for Write.

