ASSIGNMENT 03

Sabaragamuwa University of Sri Lanka

Faculty of Computing

Software Engineering

SE6103 – Parallel and Distributed Systems

Name: Disara Mapalagama
Reg. No: 19APSE4302
Academic Period: 3rd Year 2nd Semester
Degree Program: BSc (Hons) in Software Engineering
Due Date: 06.01.2024

Assignment 03

Question 2

1. Use the knowledge you gained during the Hadoop lab sessions to analyze the following example of a text and find the number of occurrences of each word in the text.

Hadoop is an open-source framework that stores and processes large amounts of data across a network of computers. Hadoop is a very interesting framework to learn about networking.

Answer:

Confirming Pre-requisites Command: Docker images Output:

```
PS C:\Users\HP> docker images
REPOSITORY                    TAG      IMAGE ID       CREATED         SIZE
nginx                         latest   3b25b682ea82   6 weeks ago     192MB
alpine                        latest   91ef0af61f39   2 months ago    7.8MB
hello-world                   latest   d2c94e258dcb   18 months ago   13.3kB
bde2020/hadoop-namenode       latest   b638307a2119   4 years ago     1.37GB
bde2020/hadoop-historyserver  latest   9d3d45ba75ae   4 years ago     1.37GB
bde2020/hadoop-datanode       latest   9e7b07fcba31   4 years ago     1.37GB
```

Start containers:

```
PS C:\Users\HP> docker run -d --name hadoop-namenode bde2020/hadoop-namenode
d6c54ebc9a48f6fd421b8cf1d1e0f1c9fb8ae6c547da9394dba69a54ffb93160
PS C:\Users\HP> docker run -d --name hadoop-datanode bde2020/hadoop-datanode
16cb27b41b06aac4f47227bcc3aa677852b54136420d4146c04c5d34877b91b2
PS C:\Users\HP> docker run -d --name hadoop-historyserver bde2020/hadoop-historyserver
15d247e5af8c297c8ae3e909d5209915548130abbf763f2cb7224a7c050e7d2c
```

Confirmed running:

```
PS C:\Users\HP> docker ps
CONTAINER ID   IMAGE                         COMMAND              CREATED          STATUS
         PORTS        NAMES
15d247e5af8c   bde2020/hadoop-historyserver  "/entrypoint.sh /run…"  47 seconds ago   Up 46 seconds (h
ealthy)   8188/tcp   hadoop-historyserver
16cb27b41b06   bde2020/hadoop-datanode       "/entrypoint.sh /run…"  55 seconds ago   Up 54 seconds (h
ealthy)   9864/tcp   hadoop-datanode
```

Created docker-compose.yaml and the text file (ex01.txt) in ex01 directory:

```yaml
docker-compose.yaml
1    services:
2      namenode:
3        image: bde2020/hadoop-namenode:latest
4        container_name: namenode
5        environment:
6          - CLUSTER_NAME=ShopSmartCluster
7          - CORE_CONF_fs_defaultFS=hdfs://namenode:8020
8        ports:
9          - "9870:9870"
10         - "9000:9000"
11       volumes:
12         - namenode-data:/hadoop/dfs/namenode
13
14     datanode:
15       image: bde2020/hadoop-datanode:latest
16       container_name: datanode
17       environment:
18         - CORE_CONF_fs_defaultFS=hdfs://namenode:8020
19       volumes:
20         - datanode-data:/hadoop/dfs/data
21       depends_on:
22         - namenode
23
24     historyserver:
25       image: bde2020/hadoop-historyserver:latest
26       container_name: historyserver
27       depends_on:
28         - namenode
29         - datanode
30       ports:
31         - "8188:8188"
32       environment:
33         - CORE_CONF_fs_defaultFS=hdfs://namenode:8020
34       volumes:
35         - namenode-data:/hadoop/dfs/namenode
36         - datanode-data:/hadoop/dfs/data
37
38   volumes:
39     namenode-data:
40     datanode-data:
41
```

```
PS E:\SUSL SE\THIRD YEAR\SECOND SEMESTER\Parallel and Distributed Systems\Practicals\ex01> docker exec
it namenode hdfs dfs -ls /output1
>>
Found 2 items
-rw-r--r--   3 root supergroup          0 2024-11-25 09:53 /output1/_SUCCESS
-rw-r--r--   3 root supergroup        223 2024-11-25 09:53 /output1/part-r-00000
```

Deployed the cluster and uploaded files:

```
PS E:\SUSL SE\THIRD YEAR SECOND SEMESTER\Parallel and Distributed Systems\Practicals\ex01> docker
exec it namenode hdfs dfs -cat /output1/part-r-00000
>>
2025-01-06 14:53:29,846 INFO sasl. SaslDataTransferClient: SASL encryption trust check:
localHostTrusted = false, remoteHostTrusted = false
Hadoop          2
is              2
framework       2
of              2
a               2
an              1
open            1
source          1
that            1
stores          1
and             1
processes       1
large           1
amounts         1
data            1
across          1
network         1
computers       1
very            1
interesting     1
to              1
learn           1
about           1
networking      1
```