Introduction to Hadoop

Short description

Lab focus: Install and test Hadoop distributed file system and Apache Pig on a Docker container.
Submission guidelines: This is not a graded assignment.

Exercise One - Set up a single Hadoop cluster using Docker

Task 1: Clone your docker-hadoop repository from [3]. The "docker-compose.yml" file has Hadoop installation and is ready to deploy in Docker containers. You can use the below command to clone the repository:

```
git clone https://github.com/big-data-europe/docker-hadoop.git
```

Task 2: Explore the docker-hadoop directory and docker-compose. This will download the images from Docker Hub and run the containers in detached mode. Check the running containers and access ports (Name Node - 9870, Data Node - 9864, Node Manager - 8042, Resource Manager - 8088, History Server - 8188).
Start the container using the command below :

```
docker-compose up -d
```

NOTE! This may take some time...

Task 3: Check the user interface dashboard of the Name Node on your computer by visiting the address: http://localhost:9870.

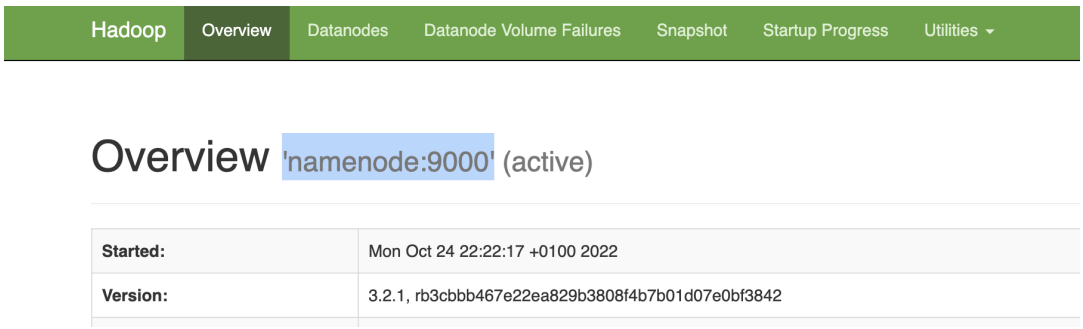

Figure 1: hdfs hostname.

Exercise Two - HDFS

NOTE! Keep the container running from the previous Exercise.

Task 1: Enter into the Name Node using interactive terminal in bash mode. Check that the HDFS is working. Follow the below steps to do so:

1. `docker exec -it namenode /bin/bash` To enter container.
2. `hdfs dfs -ls /` Checking if HDFS is working.

Task 2: Create a directory called 'P5' in HDFS and copy the given csv file from local file system into HDFS. Copy file from your computer into docker container. Make sure file is readable.

To do so run the commands below:

1. inside the containers terminal run:
   - hdfs dfs -mkdir /P5
2. Open another terminal and run:
   - docker cp Employee.csv namenode:/
3. Back to the container's terminal:
   - hdfs dfs -put Employee.csv /P5
   - hdfs dfs -ls /P5
   - hdfs dfs -cat /P5/Employee.csv

Exercise Three - Apache Pig

Task 1: Download the pig installation from [4].

NOTE! Now Enter again inside the Name Node container.

Task 2: Copy tar.gz from your local folder into Name Node and uncompress it. Move in to "bin" directory and check if you can run pig (enter inside bin/pig-0.15.0/bin):

```
root@namenode:/pig-0.17.0/bin#./pig -version
```

Task 3: Run pig and start Pig grunt shell. Load the Employee.csv file and store the data. You can do so using the command:
You can find your hostname in the http://localhost:9870, as is shown in first figure.
employee = LOAD 'hdfs://<hostname>/P5/Employee.csv' USING PigStorage(',') as (County:chararray, Salary:int);

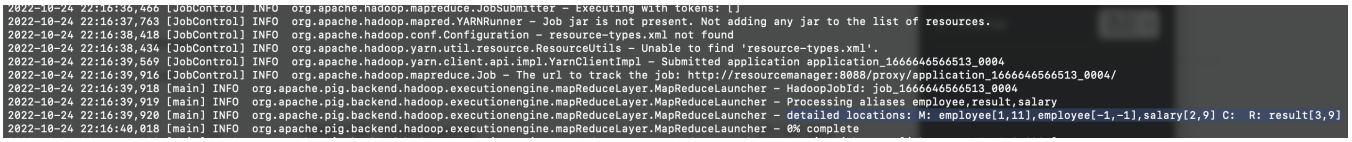

Figure 2: Result.

Task 4: Make a group command in Pig to calculate the average salary of a person in each county. Show the output.

To do so, follow the steps:

1. salary = GROUP employee BY County;
2. result = FOREACH salary GENERATE employee.County, AVG(employee.Salary);
3. DUMP result;

You can view the results in your terminal as is shown in the second figure.

Links

[1] https://hadoop.apache.org
[2] https://pig.apache.org
[3] https://github.com/big-data-europe/docker-hadoop
[4] https://dlcdn.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz

---