School of Science, Computing and Emerging Technologies

COS80023 Big Data

Pass Task 4: Parallellisation with MapReduce

Overview

Practise the principle of MapReduce and try out an example on Azure HDInsight.

Purpose

Demonstrate an understanding of the potential of MapReduce in speeding up tasks on big data sets.

Task

Carry out the tasks described below and answer the questions in your submission.

Time

This task should be completed in the fourth tutorial.

This task should take no more than 2 hours to complete.

Resources

- Presentation (from Blackboard)
- MS Azure tutorial for the creation of clusters: https://docs.microsoft.com/en-us/azure/hdinsight/hadoop/apache-hadoop-linux-tutorial-get-started (If you use the link to quickstart on this page, you will not see all the options discussed in the task)
- This may help clarify the connection between the Hadoop cluster and the Azure Storage: https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hadoop-use-blob-storage
- Any other online material
- genAI – Allowed for research. You must formulate the answers in your own words and be able to answer questions about the topics.

Feedback

Discuss your answers with the tutorial instructor.

Next

Get started on module 5.

Pass Task 4 — Submission Details and Assessment Criteria

Write down the questions and answers in a text or Word document, convert to pdf and upload to Canvas. Your tutor will mark the submission on line. If the submission is not marked as '1', it is considered as incomplete and must be resubmitted.

Subtask 4.1

Run the wordcount MapReduce code already on Azure to count the words of a file you choose and upload with the following steps:

1. Create a Hadoop cluster and storage.

2. Use Azure Storage account and upload your file.

3. Use ssh to connect to the cluster and analyse the file using MapReduce.

Use it to count the words in a file that you choose.

> **Hint**: Find some text in a newspaper, e.g. www.guardian.com. Do not use the same text as another student.

1. Deciding your location

Our subscription has access to several locations, but the resources in each location are limited. As we all work at the same time, please choose your own location as follows:

Student number is odd (ends in a 1, 3, 5 etc.): <yourlocation> = Australia East

Student number is even (ends in 2, 4, 6 etc.): <yourlocation> = Australia Southeast

It is fine to have your resource group in a different location, but all resources (storage, clusters) should be chosen from your location to ensure nothing goes wrong.

2. Creating an HDInsight Cluster for MapReduce

As you know, in Hadoop tasks run on a cluster of nodes. First, you have to create the cluster.

Assuming you are already logged in, go to your dashboard. In the search field top centre of the page, type HDInsight. Choose HDInsight clusters from the options.

Click on '+Create'.
Select the correct subscription (containing COS80023) and your resource group.

Set the cluster name to s<yourstudentnumber>cluster (no upper case letters allowed), e.g. s12345678cluster.

Choose <yourlocation> as location.

Choose Hadoop 3.1 as cluster type. Leave the default cluster username and ssh user. Choose a password with upper case, lower case, numbers and a special character.

Question 1: Do you think the choice of location matters? Why/why not?

Click Next to proceed to Storage. Select Azure Storage. Click Create new. Name your new storage <yourstudentnumber>storage.

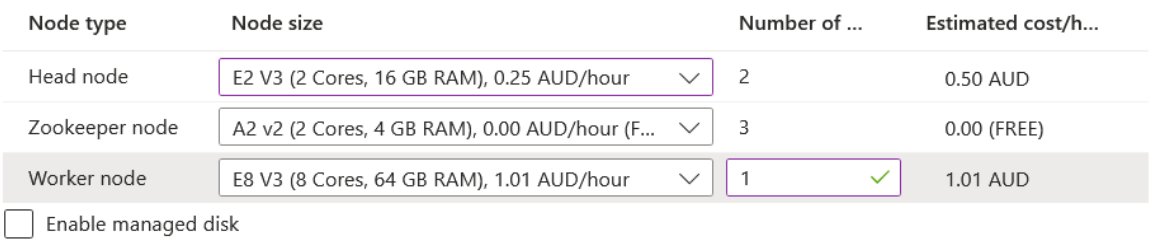
For the container choose <yourstudentnumber>container.

Leave the other options as default.

Click Next to proceed to Security and Networking. Do not change the default options.

Click Next to proceed to Configuration+pricing. Examine the default resources for the cluster. There are head nodes, Zookeeper nodes and worker nodes.

Choosing the smallest possible options might lead to using resources that are scarce in Australian locations. Please change the Head node to E2 V3. Leave the other two, but reduce the Worker node to 1 instead of 4 (observe the pricing that adjusts with the choice).



No need to add script action.

Click Review+create. On the summary page, you get to create the cluster. It typically takes a few minutes for the cluster to deploy.

To find out about the progress (and possible errors), click on notifications on the top right (bell-shaped icon).

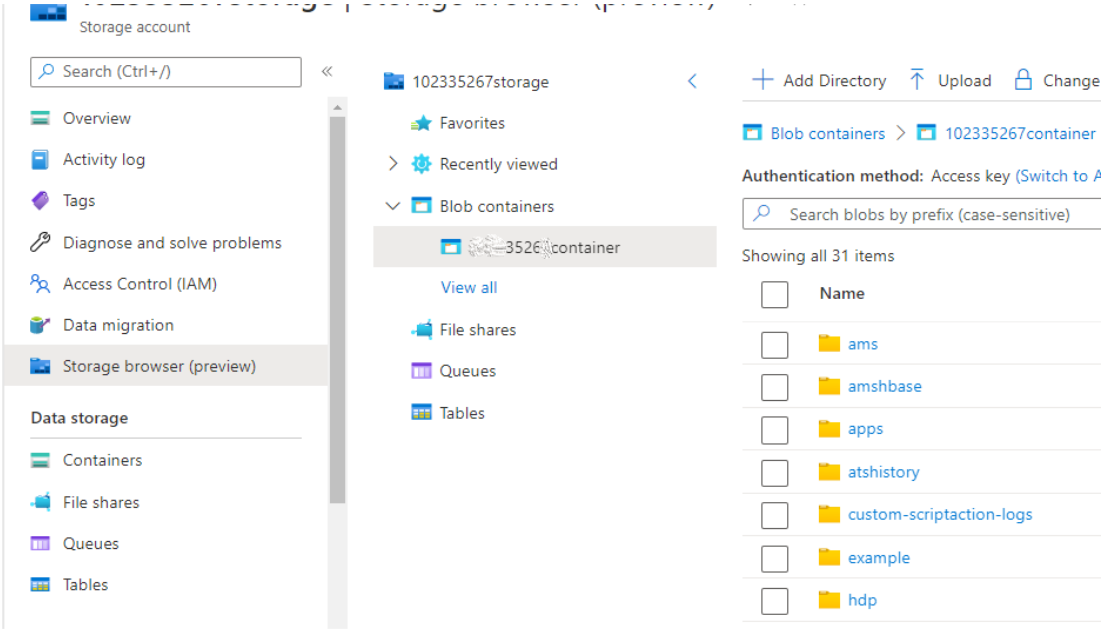
3. Using Storage on Azure

To analyse a file using MapReduce, you have to put the file where MapReduce can find it. There are two options, Data Lakes and Azure Storage. We will use Azure Storage that we have created beforehand.

Go to the storage account when it has been created.

Click on Storage browser (preview).

Click on Blob Container. You should see container you created earlier. Click on it. Click upload and find the file you want to use to count the words of on your file system. This is what the dashboard should look like:



Question 2: What is the purpose of the ssh protocol? How does it implement security?

4. Running wordcount

When the deployment has completed, click on Go to Resource. Find the connection string for ssh. (Hint: Look under Settings). When you have found it, click the 'copy to clipboard' option.

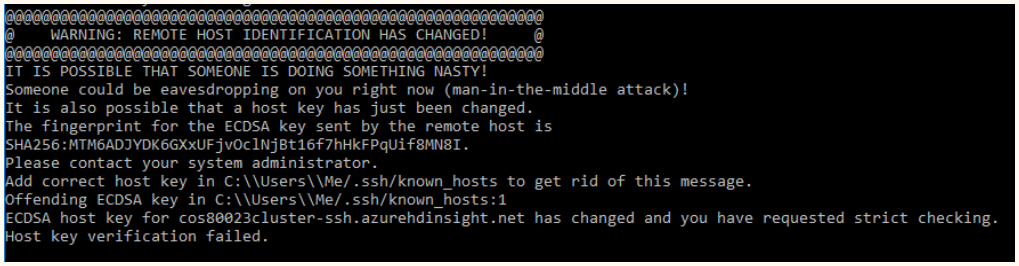
Open a command window ('cmd' in the Windows search bar). Right-click anywhere on the Command Prompt. This pastes the string from the clipboard. Press Enter to run the command.

If your connect string was correct, you will see:

```
C:\Users\Me>ssh sshuser@cos80023cluster-ssh.azurehdinsight.net
The authenticity of host 'cos80023cluster-ssh.azurehdinsight.net (13.70.81.153)' can't be established.
ECDSA key fingerprint is SHA256:MTM6ADJYDK6GXxUFjvOclNjBt16f7hHkFPqUif8MN8I.
Are you sure you want to continue connecting (yes/no)? yes
```

This is to tell you that your computer has never had any dealings with this host and does not recognise its signature. You can safely say yes. The reply will be that your address has been permanently added to the list of known hosts.

Warning: If you create the cluster the second time (after deleting it the first time), the signature will have changed. The prompt will no longer ask you, but tell you connecting is too dangerous:



You have to find the known_hosts file and erase the entry for this connection (asurehdinsight.net) before you can continue.

Having been added to the list of known hosts means the encryption key has been stored and from now on, messages can be securely sent from client to server and back.

Run the connect string again (push the 'up' arrow to return to the previous command). This time you will be prompted for the password. If you type it correctly, you are ready for the next step.

Invoke the wordcount example already on HDInsight:

```
yarn jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-mapreduce-examples.jar wordcount
```

Question 3: What does the interface tell you (specifically, the last row that starts with 'Usage')? How do you think you can fix this?

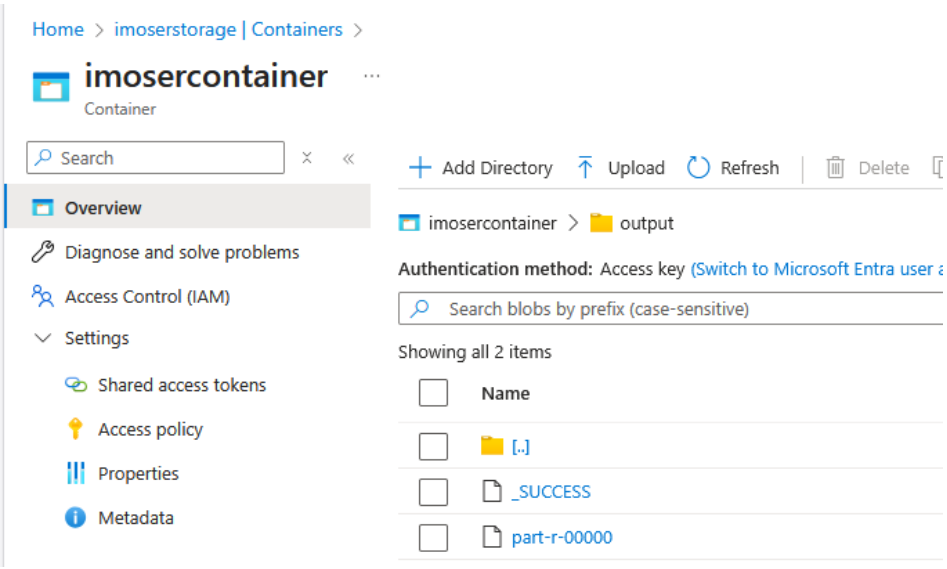
The file you are using should be here:

wasb://<yourstudentnumber>container@<yourstudentnumber>storage.blob.core.windows.net/<yourfilename>

You can use this as an output directory:

Question 4: What does the wasb prefix mean, and how does it relate to HDFS?

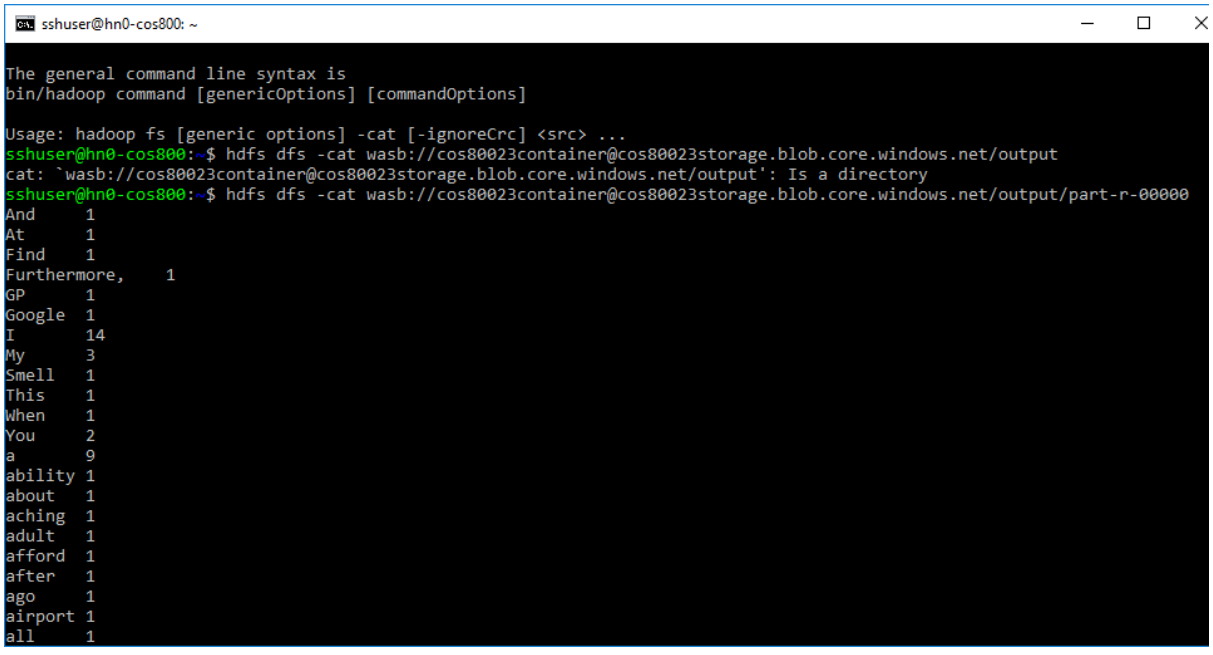
If the wordcount example runs successfully, it creates a file called part-r-00000 (it would create more files with different numbers if the input file was bigger).



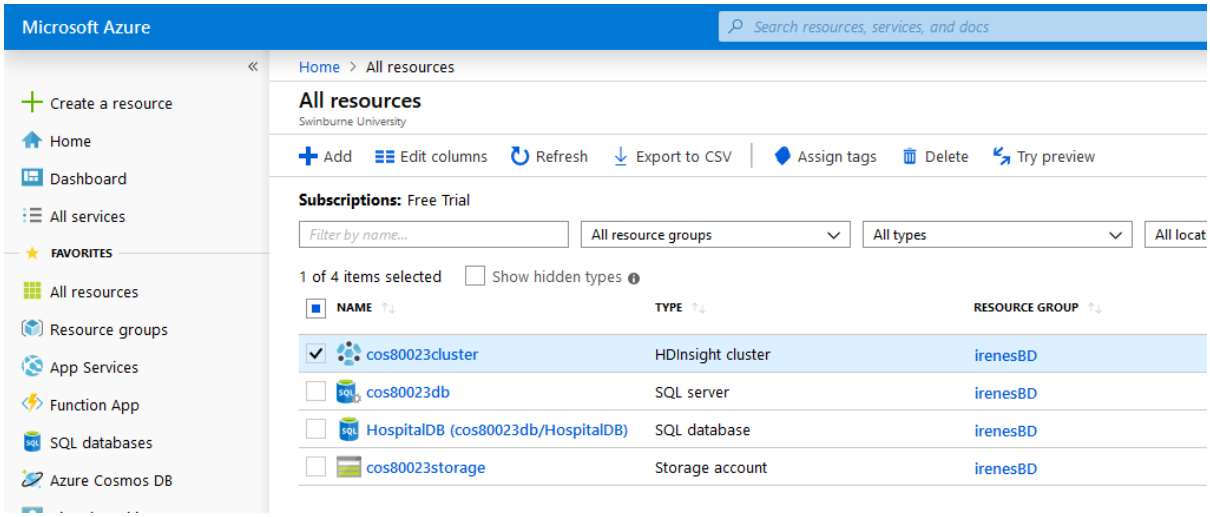
Show the part-r-00000 file on the command line. Use the command:

```
hdfs dfs -cat wasb://<directory-path>/output/part-r-00000
```

Take a screenshot of the command and the beginning of the file and put it into your answer file for Canvas. Example:



Important: You MUST delete the cluster at the end. The cluster will keep running unless you delete it. On your dashboard, go to All resources. Tick the box in front of your cluster and click Delete. You have to confirm in the next step.



Observe the notifications (bell icon). When the cluster has been deleted, refresh and put a screenshot of the resources without the cluster in your answer document.