**COS80023 Big Data – Lab 4**

**Lab 4: Pass Task 4 – Parallelisation with MapReduce**

**Student Name: Arun Ragavendhar Arunachalam Palaniyappan**

**Student ID: 104837257**

**1. Do you think the choice of location matters? Why/why not?**  
Yes, the choice of location matters. The cluster and storage must be in the same region to ensure that resources can connect and work together properly. If they are in different regions, the services may not be able to communicate or it could cause slower performance and higher costs due to cross-region data transfer.

**2. What is the purpose of the ssh protocol? How does it implement security?**  
The purpose of SSH is to allow a secure connection between my local computer and the Hadoop cluster. It provides a way to log in and run commands on the cluster remotely. Security is implemented through encryption and authentication. SSH uses keys or passwords to verify the user and encrypts all data exchanged, so information like login credentials or commands cannot be read by others during transfer.

**3. What does the interface tell you (specifically, the last row that starts with 'Usage')? How do you think you can fix this?**  
The interface shows that the command needs more input to run. The “Usage” line explains the correct way to use the wordcount example, which is wordcount <input> <output>. This means I must provide the input file location in storage and an output directory. The problem can be fixed by adding the file path for the input data and choosing an output directory path when running the command.

</task4.txt> </output>

**4. What does the wasb prefix mean, and how does it relate to HDFS?**  
The “wasb” prefix stands for **Windows Azure Storage Blob**. It is the protocol that allows Hadoop to access files stored in Azure Blob Storage as if they were in HDFS. In other words, wasb provides a connection layer that makes Azure storage behave like the Hadoop Distributed File System, so that Hadoop tools such as MapReduce can read and write data without any changes to the code.



