



UPPSALA
UNIVERSITET

Assignment 2: MapReduce

Junjie Chu

Part 1

Task 1.1

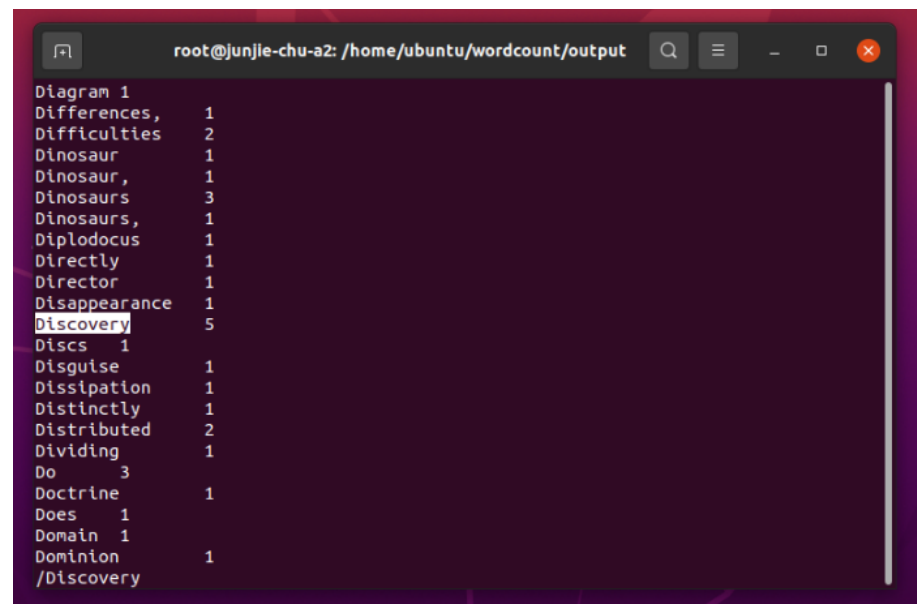
Question a:

There are 2 files in the output folder. `_SUCCESS` is a file which shows whether the program run successfully or not. This would typically be used by job scheduling systems. The output files are by default named `part-x-yyyyy` where:

x is either 'm' or 'r', depending on whether the job was a map only job, or reduce, yyyy is the mapper or reducer task number (zero based).

Question b:

5 times.



```
root@junjie-chu-a2: /home/ubuntu/wordcount/output
Diagram 1
Differences, 1
Difficulties 2
Dinosaur 1
Dinosaur, 1
Dinosaurs 3
Dinosaurs, 1
Diplodocus 1
Directly 1
Director 1
Disappearance 1
Discovery 5
Discs 1
Disguise 1
Dissipation 1
Distinctly 1
Distributed 2
Dividing 1
Do 3
Doctrine 1
Does 1
Domain 1
Dominion 1
/Discovery
```

Question c:

“Local (Standalone) Mode:

- Default mode of Hadoop
- HDFS is not utilized in this mode.



UPPSALA
UNIVERSITET

- Local file system is used for input and output
- Used for debugging purpose
- No Custom Configuration is required in 3 hadoop(mapred-site.xml,core-site.xml, hdfs-site.xml) files.
- Standalone mode is much faster than Pseudo-distributed mode.

Pseudo Distributed Mode (Single Node Cluster):

- Configuration is required in given 3 files for this mode
- Replication factor is one for HDFS.
- Here one node will be used as Master Node / Data Node / Job Tracker / Task Tracker
- Used for Real Code to test in HDFS.
- Pseudo distributed cluster is a cluster where all daemons are running on one node itself.

Task 1.2

Question a:

core-site.xml

1 Specify the location of the name node

2 hadoop.tmp.dir is the basic configuration that the hadoop file system depends on, and many paths depend on it. If the storage location of namenode and datanode is not configured in hdfs-site.xml, they will be placed in this path by default.

hdfs-site.xml:

1 Configure the specific path of namenode and datanode to store files

2 Configure the number of copies

Question b:

```
root@junjie-chu-a2:/usr/local/hadoop# jps
15766 Jps
15067 NameNode
15452 SecondaryNameNode
15245 DataNode
root@junjie-chu-a2:/usr/local/hadoop#
```

NameNode:

Main function: accept the client's read and write requests and distribute them to DataNodes, which are the main storage and processing places for files. The file metadata (metadate) will be stored in the NameNode, including:



UPPSALA
UNIVERSITET

1. File owner, permissions, file name, etc.
2. The block contained in the file
3. In which DataNode these blocks are stored (reported when DataNode starts)

This metadata information is stored as a file "fsimage" on the disk and loaded into the memory after the NameNode starts. That is to say, the metadata is held in the disk and the memory at this time, and they are the same.

After that, since the DataNode will also be started after the NameNode is started, and after the DataNode is started, the block location information will be uploaded to the metadata in the memory instead of the metadata in the disk, so this will cause the metadata in the memory and the disk to be inconsistent.

And the metadata generated after the user uploads the file is the metadata in the memory of the operation, but when the metadata is operated, the operation log is recorded through an edits file. The edits file is stored on the disk.

SecondaryNameNode(SNN):

It is not a backup of the NameNode. Its main function is to help the NameNode merge edits log and reduce the startup time of the NameNode.

When will SNN perform the merger:

1. The time interval `fs.checkpoint.period` set according to the configuration file is 3600 seconds by default
2. Set the edits log size `fs.checkpoint.size` according to the configuration file to specify the maximum value of the edits file, which is 64MB by default.

It means to merge once every once in a while, or merge when the edits file is full.

DataNode (DN):

Main function: storage data block (block)

When the DN starts, it will report the block information to the NN, and then keep in touch by sending a heartbeat to the NN (for example: every 3 seconds). If the NN does not receive the heartbeat of the DN within a period of time, the DN will be considered lost and copy the Block to other DN.

JPS:



UPPSALA
UNIVERSITET

Its function is to display the current system's java process and its id number.

Task 1.3

Successfully running:

```
root@junjie-chu-a2:/home/ubuntu/wordcount# /usr/local/hadoop/bin/hdfs dfs -ls /home/ubuntu/wordcount/output
Found 2 items
-rw-r--r-- 1 root supergroup 0 2021-02-09 11:05 /home/ubuntu/wordcount/output/_SUCCESS
-rw-r--r-- 1 root supergroup 196183 2021-02-09 11:05 /home/ubuntu/wordcount/output/part-000000
root@junjie-chu-a2:/home/ubuntu/wordcount#
```

The results of task1.1 and task 1.2 are the same.

divides 4		root@ju
dividing	4	
diving 3		divided 7
diving, 2		divided. 1
diving-bell	1	divides 4
diving-bell,	1	dividing 4
divisible	1	diving 3
division	9	diving, 2
divisions	1	diving-bell 1
do 96		diving-bell, 1
do, 5		divisible 1
do. 5		division 9
doctrine 1		divisions 1
does 52		do 96
does, 3		do, 5
does. 3		do. 5
does; 1		doctrine 1
dog 14		does 52
dog, 6		does, 3
dog-toothed	1	does. 3
dog; 1		does; 1
dogfish,	1	dog 14
dogma 1		dog, 6
dogs 2		dog-toothed 1
		dog; 1
		/96

Question a:

Text.class: Set the key of job's reduce's output.

IntWritable.class: Set the value of job's reduce's output

Map.class: Set the map function of the job. The map has been re-written in the program.

Reduce.class: Set the reduce function of the job. The reduce has been re-written in the program.

TextInputFormat.class: Set the input format of the map-reduce job.



UPPSALA
UNIVERSITET

TextOutputFormat.class: Set the output format of the map-reduce job.

Question b:

Hadoop distributed file system (HDFS) is a distributed file system designed to run on commodity hardware. HDFS adopts master / slave structure model. An HDFS cluster is composed of a namenode and several datanodes. The namenode is used as the main server to manage the file system's namespace and the client's access to the file; the datanode in the cluster manages the stored data.

HDFS is built on the local file system. HDFS stores data by operating the local file system.

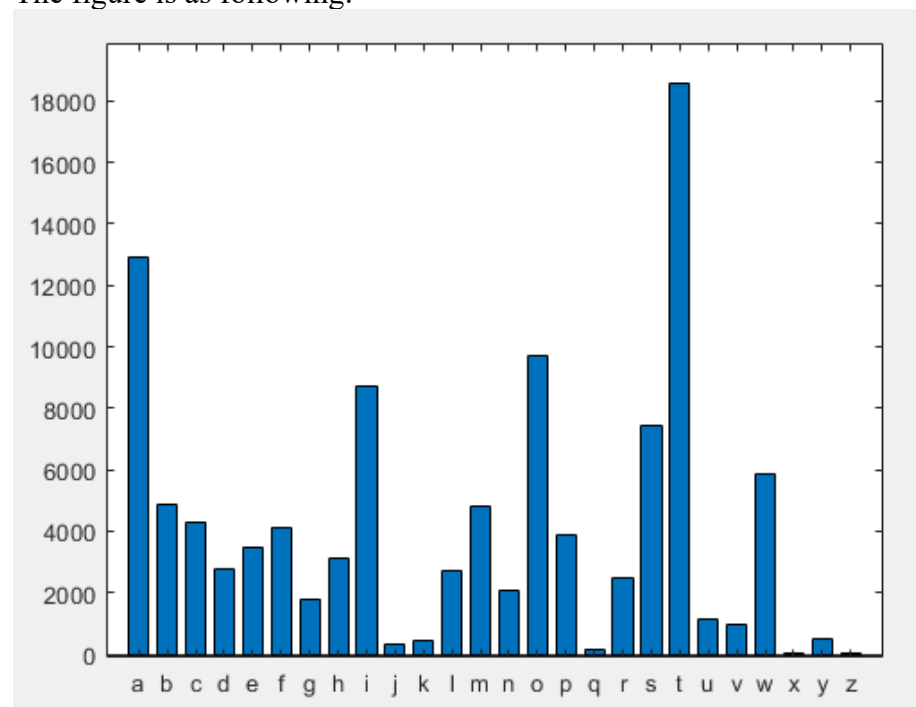
Hadoop abstracts a layer from the existing file system, but not all of them are local file systems.

In order to provide a consistent interface for different data access, Hadoop uses the concept of virtual file system of Linux for reference, introduces Hadoop Abstract file system, and provides a large number of specific implementations on this basis. HDFS is one of them.

HDFS is the abstraction of a higher level file system, which regards the file system composed of multiple machines as a logical whole.

Task 1.4

The figure is as following:





UPPSALA
UNIVERSITET

Task 2.1

1.

Question a:

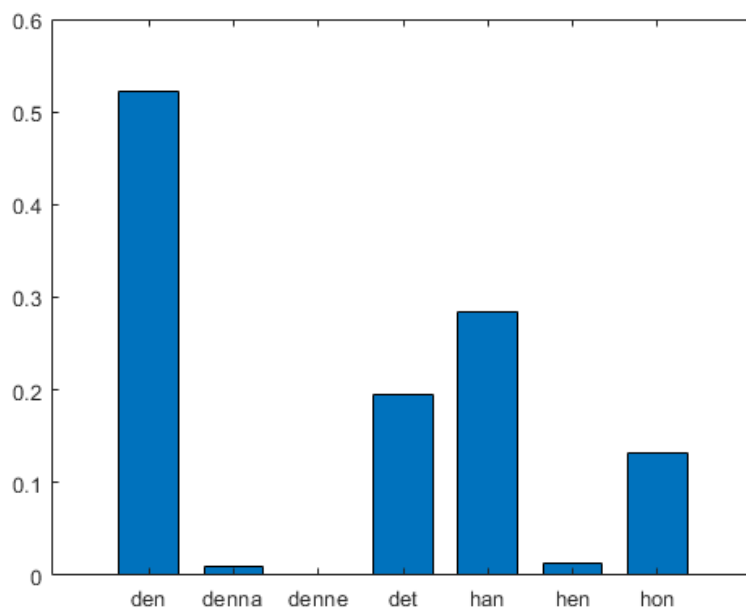
The JSON-formatted tweets is semi-structured data. The data is a bit like structured data, but different objects may have different attributes. Not all the data have the same structure.

Question b:

It is a bit like structured data, but the structure changes a lot. We can't simply create a table to correspond with it. Changes of data usually lead to changes of structural pattern. Relational database can not store this type of data. The composition of semi-structured data is more complex and uncertain, so it has higher flexibility.

2 and 3.

The normalized count:



Part 2

Task 1

1.

SQL:

1) The characteristics of relational database are as follows

-Data relational model is based on relational model, structured storage and integrity constraints.



UPPSALA
UNIVERSITET

- Based on the two-dimensional table and the relationship between them, data operations such as connection, union, intersection, difference and division are needed.
- Structured query language (SQL) is used to read and write data.
- Operations need data consistency, transaction consistency and even strong consistency.

2) Advantages:

- Keeping data consistent (transaction processing)
- It can perform complex queries such as join.
- Universal, mature technology.

3) Disadvantages:

- Data reading and writing must be parsed by SQL, and the reading and writing performance of large amount of data and high concurrency is insufficient.
- When reading and writing data or modifying data structure, lock is needed, which affects concurrent operation.
- Unable to adapt to unstructured storage.
- It is difficult to expand.
- Expensive and complex.

NoSQL database:

1) The characteristics of NoSQL database are as follows:

- Unstructured storage.
- Based on the multidimensional relation model.
- It has a unique use scenario.

2) Advantages:

- High concurrency, strong reading and writing ability in big data.
- Basic support for distributed, easy to expand, scalable.
- Simple, weakly structured storage.

3) Disadvantages:

- The ability of complex operations such as join is weak.
- Transaction support is weak.
- The universality is poor.
- Complex business scenarios without integrity constraints are poorly supported.

Examples:

SQL:

Bank account database

The database of 'Alipay' or 'Paypal'

NOSQL:

LinkedIn's database (for example, a database that stores resumes).

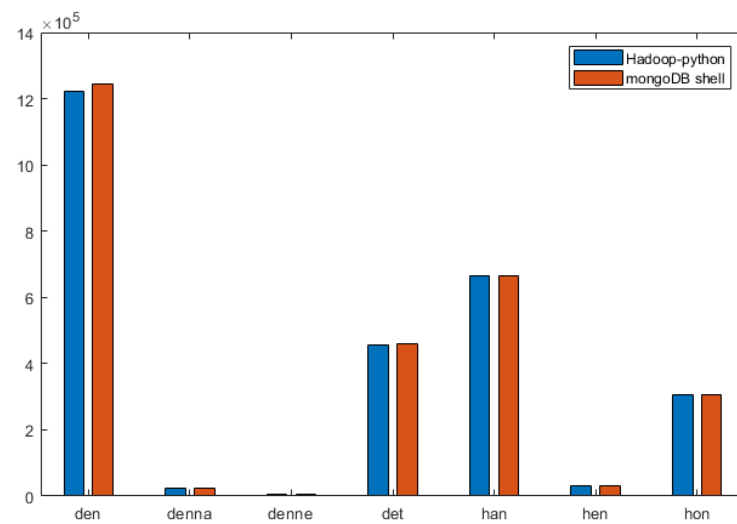
Location Based Services.



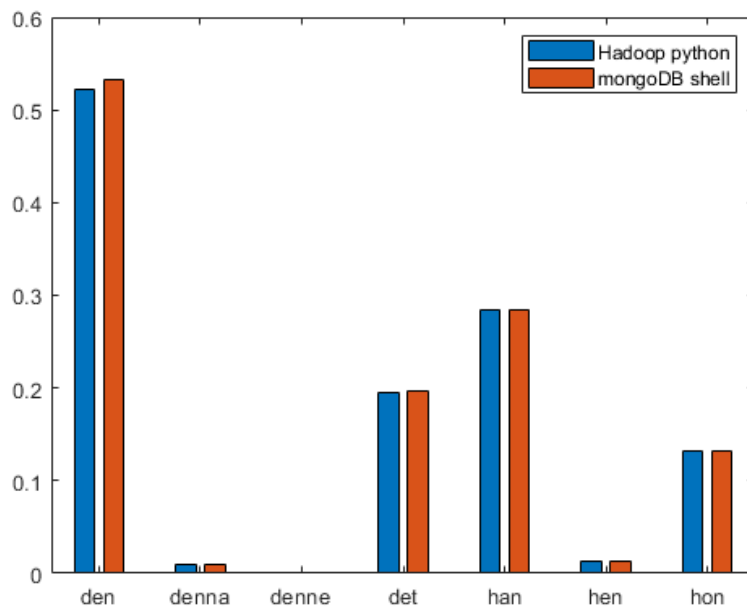
UPPSALA
UNIVERSITET

2.

The count:



The normalized count:



3.

The MongoDB method takes up less hard disk space and run faster than the Hadoop python method. But it almost stores all the data and indexes in the cache and memory. So, it will take up very large memory. I set a large memory for my VM to do the analysis.



UPPSALA
UNIVERSITET

I use MongoDB shell and the built-in MapReduce function. The shell client is more convenient to interact, and the MapReduce function will improve the query speed. Thus, I choose it.

The command `'db.data.find({retweeted_status:null}).count()'` will count all the unique tweets.

The map function is written in Javascript. I split the text of tweets into words array and transform all the words into lower case. Then count the numbers of occurrences of each pronoun in the words array. If the number is larger than 0, use the MongoDB interface `emit(pronoun,1)` to output the key value pair.

The query function is used to choose which data will be the input of map function. `'query:{retweeted_status:null}'` is used to choose unique tweets.

`'out'` is used to set the collection which the output will be stored in.