Uppsala University

Large Datasets For Scientific Applications

1TD268

Assignment A1

Author:
Sudarsan Bhargavan

June 6, 2018

1 Introduction to Hadoop/HDFS WordCount

1.1 Questions

1. Look at the contents of the folder "output" - what are the files place in there? What do they mean?

   Ans:
   There are two files that are created when a job is completed successfully, viz;

   - **_SUCCESS**
   - **part-r-00000**

   The file named _SUCCESS is needed when any application needs to check the completion of the result set by inspecting HDFS.

   The file named part-r-00000 contains the output of the map & reduce jobs as the name has an **r** tag which signifies the type of the job and the **00000** signifies the number of the task. [1]

---

2. In this example we used Hadoop in "Local (Standalone) Mode". What is the difference between this mode and the Pseudo-distributed mode?

   Ans:
   Local Standalone Mode:
   In this mode hadoop is configured in a single node setup as a single java process, with no demons. Local file system is used instead of HDFS. There is a single data-node, single name-node and only one task-tracker.

   Pseudo-Distributed Mode:
   In this configuration hadoop is set to simulate a multi-node cluster in just one VM instance. Unlike local standalone mode, HDFS is used instead of the local file system, hadoop demons run on multiple JVM instances. [2]

1.2 Questions

1. What are the roles of the files core-site.xml and hdfs-site.xml ?

   Ans:
   core-site.xml:
   The core-site.xml configuration file has **Hadoop's core** configurations. Such as the **core I/O settings**, port and IP configurations for the **NameNode daemon**.

   hdfs-site.xml:
   The hdfs-site.xml configuration file has (**Hdfs daemons**, **NameNode**, **SecondaryNameNode**, **DataNodes**)'s configurations. Also **HDFS**'s **Block Replication, Permissions** configurations required to replicate blocks in HDFS and the number of replications required. [3]

---

2. Describe briefly the roles of the different services listed when executing 'jps'.

   Ans:
   NameNode:
   The NameNode has the metadata of the data in the HDFS cluster. It tracks the locations of specific data in the cluster. It returns a list of relevant **DataNode Servers** where the data resides when a client application sends a query for locating a file.[4]

   SecondaryNameNode:
   The SecondaryNameNode creates checkpoints of a particular **Namespace** by keeping track of edits in an **edits file** and merging these edits with the **fsimage file**. It can be hosted on a different machine while running hadoop in cluster mode. It is also known as the **checkpoint node**.[4]

The SecondaryNameNode helps in better functioning of the NameNode by solving the following problems:

- Management of large edit-logs
- Slow restarts of NameNodes due to large amount of edits that have to be merged
- loss of metadata during a crash

DataNode:
The DataNode as the name suggests stores data in HDFS. Usually multiple DataNodes are setup with **data replication** in a functional HDFS. The DataNode responds to the client's queries related to file-system operations.[5]

jps:
JVM Process Status Tool lists various **Java Virtual Machines** in the target system. It needs access permissions to list the target system's JVMs and only lists those it has permissions for.[6]

1.3 Questions

1. Explain the roles of the different classes in the file WordCount.java

   Ans:
   The **WordCount.java** file has two classes, viz;

   - Map Class (Mapper Class)
   - Reduce Class

   The **Mapper Class** - as the name suggests maps input key:value pairs into intermediate key:value pairs by extracting data and transforming them by shuffling and sorting. The input key:value pairs can map zero or more intermediate key:value pairs and the type of the intermediate key:value pairs can be different from that of the input.[7]

   The **Reducer Class** - reduces the intermediate key:value pairs generated by the mapper class into a smaller set of values by aggregating, summarizing, filtering or transforming these intermediate key:value

pairs, and writes them as the result. Here the results of the reducer class is the count of the occurrences of the key.[8]

1.4 Questions

1. Describe the role of Combiners in MapReduce.

   Ans:

   Combiners:
   The combiner class reduces the size of the data output from the mapper class before it is fed into the reducer class. While dealing with large datasets the overhead placed on the reducer function is very high, and transferring large datasets can also cause network latency issues. Combiner summarizes the mapper output with the same key and the output of the combiner is then transferred over the network, which becomes the input for the reducer class. Also a combiner has to implement the reducer interface's reduce() method, since it does not have an interface of its own. It can be implemented by: conf.setCombinerClass(Reduce.class); [9]

   For wordcount of all occurencess of words that start with the same first letter, Please See Figure 1

---

2 Analyzing twitter data using Hadoop streaming and Python

1. Based on the documentation in the above link, how would you classify the JSON-formatted tweets? Structured, semi-structured or unstructured data? What could be the challenges of using traditional row-based RDBMs to store and analyze this dataset (apart from the possibility of very large datasets)?

   Ans:

   Classification:

   The JSON-formatted tweets are semi-structured, it poses various challenges if a traditional row-based RDBMs has to be used to store and analyze this dataset.[10]

   Challenges:

   - Traditional RDBMs are designed to only work with **Structured Data**.
   - Traditional RDBMs are designed around a central data store and have very poor compatibility for distributed data processing.

   For wordcount of all the occurrences of Swedish pronouns, Please See Figure 2

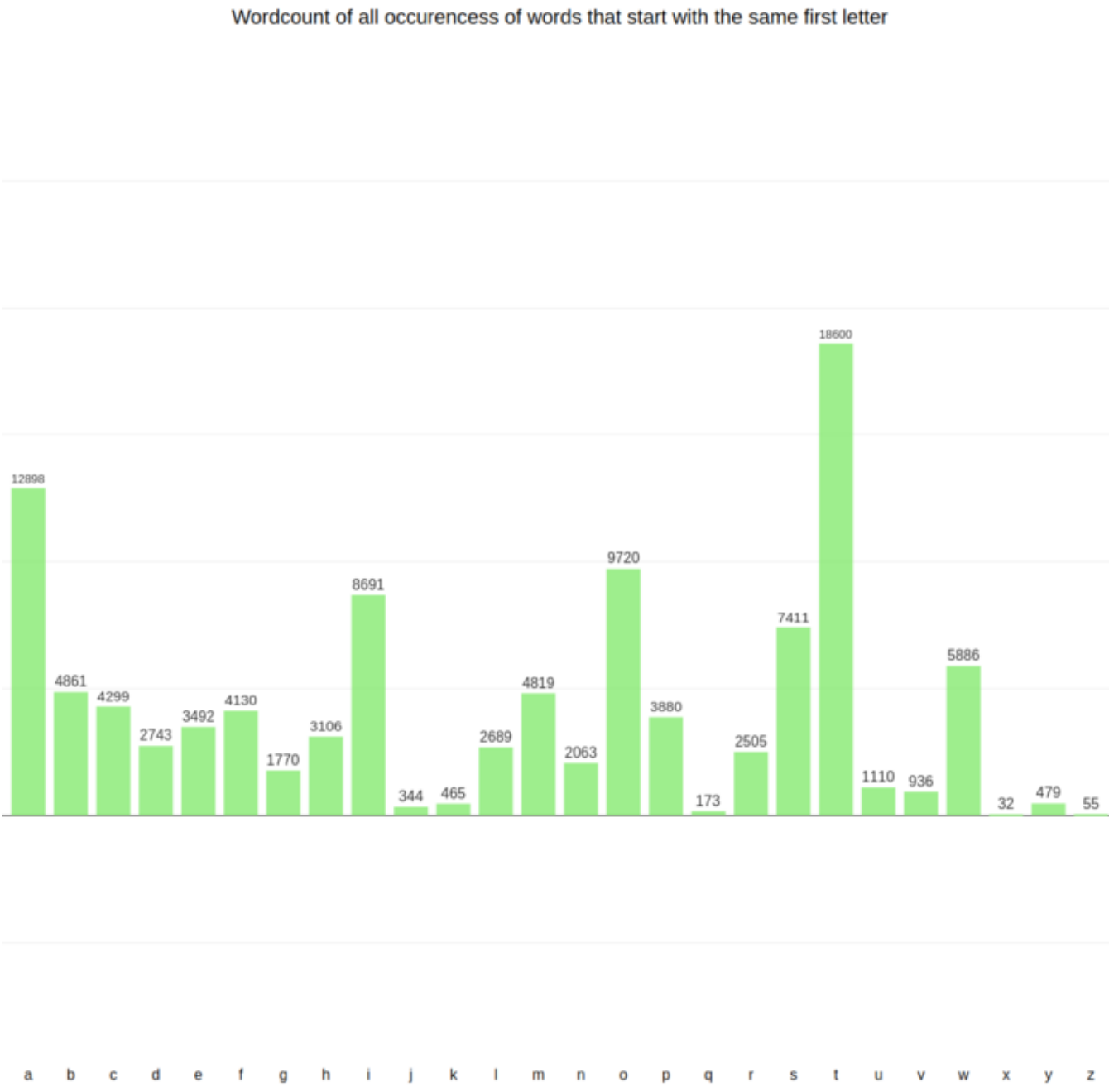

Figure 1: X-axis: Letters, Y-axis: Count, Scale: 1K

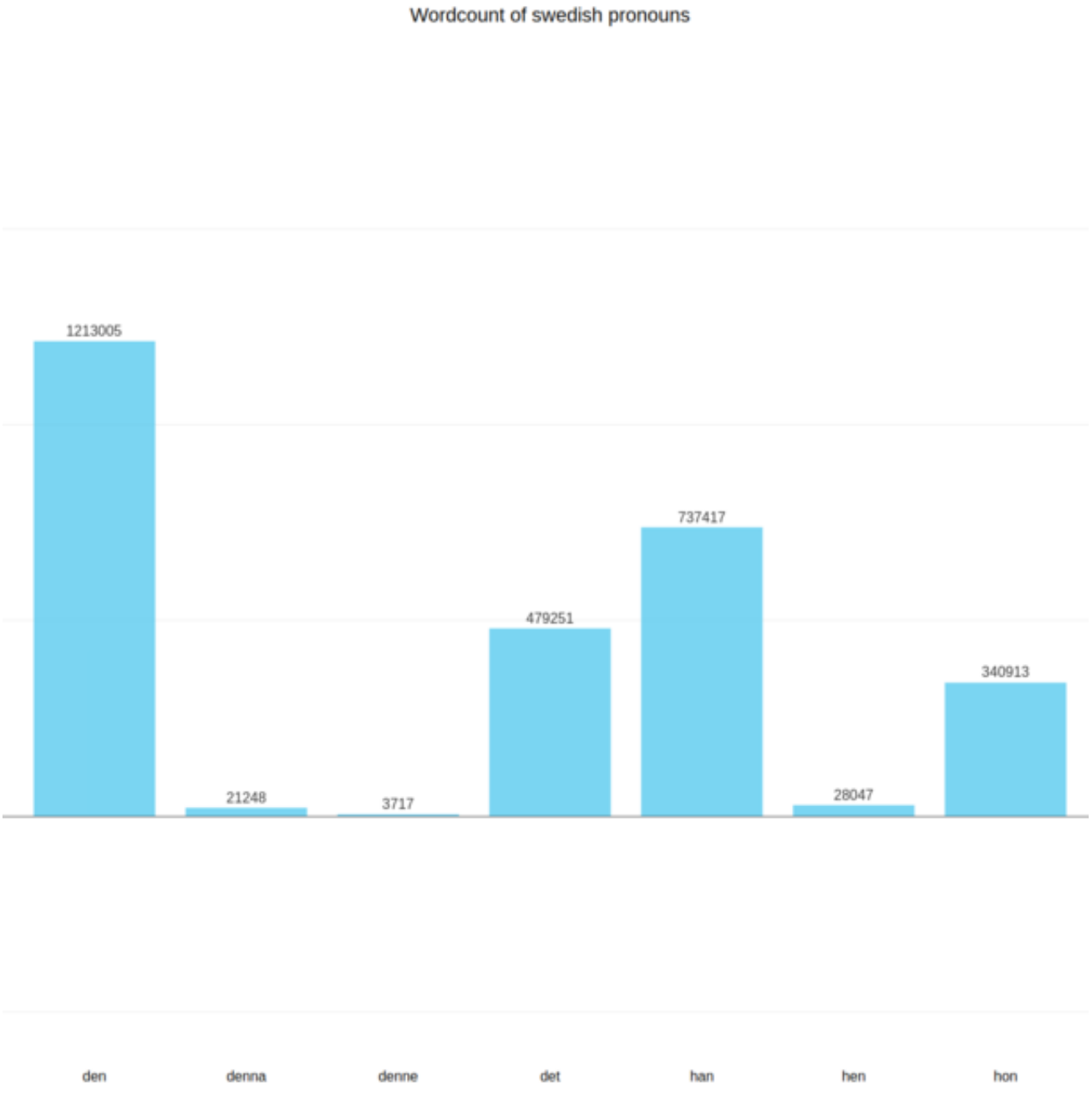

Figure 2: X-axis: Swedish Pronouns, Y-axis: Count, Scale: 1M

References

[1] C. White. (2018) What are success and part-r-00000 files in hadoop. Accessed: 2018-05-01. [Online]. Available: https://stackoverflow.com/questions/10666488/what-are-success-and-part-r-00000-files-in-hadoop#10666874

[2] Tariq. (2018) What is the difference between single node & pseudo-distributed mode in hadoop? Accessed: 2018-05-01. [Online]. Available: https://stackoverflow.com/questions/23435333/what-is-the-difference-between-single-node-pseudo-distributed-mode-in-hadoop#23436266

[3] G. Shankhdhar. (2018) Hadoop cluster configuration files. Accessed: 2018-05-13. [Online]. Available: https://www.edureka.co/blog/hadoop-cluster-configuration-files/

[4] apache.org. (2018) Namenode -hadoop wiki. Accessed: 2018-06-05. [Online]. Available: https://wiki.apache.org/hadoop/NameNode

[5] ——. (2018) Datanode -hadoop wiki. Accessed: 2018-06-05. [Online]. Available: https://wiki.apache.org/hadoop/DataNode

[6] O. J. S. Docs. (2018) jps - java virtual machine process status tool. Accessed: 2018-06-05. [Online]. Available: https://docs.oracle.com/javase/7/docs/technotes/tools/share/jps.html

[7] hadoop.apache.org docs. (2018) Class mapper. Accessed: 2018-06-06. [Online]. Available: https://hadoop.apache.org/docs/current/api/org/apache/hadoop/mapreduce/Mapper.html

[8] ——. (2018) Class reducer. Accessed: 2018-06-06. [Online]. Available: https://hadoop.apache.org/docs/r2.6.2/api/org/apache/hadoop/mapreduce/Reducer.html

[9] T. Point. (2018) Mapreduce - combiners. Accessed: 2018-06-06. [Online]. Available: https://www.tutorialspoint.com/map_reduce/map_reduce_combiners.htm

[10] D. Aggarwal. (2018) Difference between traditional data and big data. Accessed: 2018-06-06. [Online]. Available: https://www.projectguru.in/publications/difference-traditional-data-big-data/