Lecture 2 exercises

Daniel Skjold Toft – datof16@student.sdu.dk

Docker-compose a HDFS cluster

- A namenode and 3 datanodes
- One network – “hadoop”
- 4 volumes – docker’s way of saving data

Docker exec –ti namenode /bin/bash

- Docker exec allows us to execute commands in the docker container
- -ti creates an interactive shell
- “namenode” is the docker container name
- /bin/bash is the shell we want

Basic HDFS shell commands

- hdfs dfs –[command] [path]
- Path is the path in the HDFS folder structure, not the running Linux
- No –cd command, meaning path always have to be specified
- -ls
- -cat
- -put
- -rm, -touch, -mkdir, etc.

Let’s download a book!

- “docker exec –ti namenode /bin/bash”
- “apt update”
- “apt install wget”
- “wget -O alice.txt https://www.gutenberg.org/files/11/11-0.txt”
- “hdfs dfs -put alice.txt /”

Python read-write example

- Dockerfile – Building Python containers
- Run.cmd – Build and run container
- HdfsCli
  - https://hdfscli.readthedocs.io/en/latest/

Python read-write with JSON

- Only example.py changes!
- Extends what we just learned with read-write
- Uses “Counter” to count words
- Dumps a JSON structure

Python read-write with Avro

- AvroWriter from HdfsCli Extensions
  - Example on HdfsCli github page
- Optionally specify a mandatory schema!
- “content” is summary of the remote file
- “reader” can be traversed as a list
- “hdfs dfs –cat /word-count.avro”
  - Is the result expected?

Python read-write with Parquet?

- Write the 10 most common words and read it again using a Parquet file in the HDFS cluster
- Consider looking at
  - pyarrow and pandas: https://arrow.apache.org/docs/python/parquet.html
  - fastparquet: https://fastparquet.readthedocs.io/en/latest/
- I used pyarrow and some extra HdfsCli client methods

Cleanup

- “docker image ls -a”
  - Notice all the “<none>” images? They’re “dangling images”
  - Takes no space on your computer
- Remove by “docker image prune”
  - Will only remove the dangling images
- What about containers?
  - The “run.cmd” contains “-rm”, which removes the container after container termination