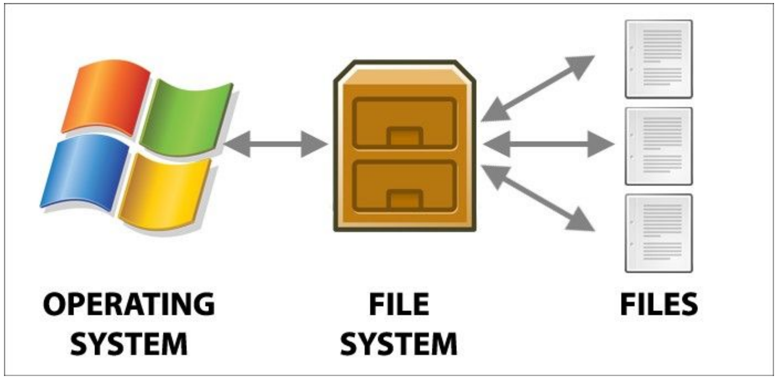
TODAY'S AGENDA

- Introductions
- Introduction to Big Data
- Course Overview
- Container Setup
- Computer Systems & Parallelism
- Guest Lecture
- Exercise
- Assignment

Exercise

- Overview
    - Work in Docker container
    - Copy files to/from HDFS

HADOOP DISTRIBUTED FILE SYSTEM



- File System
  - Data for/from computing is stored in files
  - File system organizes files so data can be stored and retrieved efficiently
- Distributed File System (DFS)
  - For efficient processing of very large data files, data is partitioned across many computer systems
  - DFS manages data that is distributed across many networked systems
  - Each local file system manages its own partition. DFS works on top of local file systems to manage entire file
- Hadoop Distributed File System (HDFS)
  - DFS in Hadoop's ecosystem

EXERCISE STEPS

1. Download data file Shakespeare.txt from Canvas and place it in the same folder as the launch.sh script(working directory)
2. Start a Terminal window in JupyterLab, change to work directory and perform the following steps in the Terminal window.
   a. cd work
3. Run jps to check what daemons are running
   a. jps # Should see NameNode, DataNode, Jps, SecondaryNode
4. Make directory on HDFS
   a. hadoop fs -mkdir /S1
5. Copy data to HDFS and verify file is there
   a. hadoop fs -copyFromLocal Shakespeare.txt /S1
   b. hadoop fs -ls /S1 # Should see Shakespeare.txt
6. See first few lines of Shakespeare.txt
   a. hadoop fs -cat /S1/Shakespeare.txt | head
7. Make copy of data file
   a. hadoop fs -cp /S1/Shakespeare.txt /S1/ShakespeareNew.txt

EXERCISE STEPS

8. List contents in HDFS
   a. `hadoop fs -ls /S1`   # Should see the 2 Shakespeare files
9. Check contents of new file
   a. `hadoop fs -cat /S1/ShakespeareNew.txt`
10. Copy results from HDFS to local file system
    a. `hadoop fs -copyToLocal  /S1/ShakespeareNew.txt  ShakespeareNew.txt`
11. Compare the two Shakespeare files on local system
    a. `diff Shakespeare.txt ShakespeareNew.txt`
12. Check contents of file copied over
    a. `head ShakespeareNew.txt`