

# Apache Hadoop

*Amit Jain*



# Hadoop Running Modes

We can use hadoop in three modes:

- ▶ *Standalone mode*: Everything runs in a single process. Useful for debugging.
- ▶ *Pseudo-distributed mode*: Multiple processes as in distributed mode but they all run on one host. Again useful for debugging distributed mode of operation before unleashing it on a real cluster.
- ▶ *Distributed mode*: “The real thing!” Multiple processes running on multiple machines.
  - ▶ These are typically multi-user Hadoop clusters with appropriate security and access controls.
  - ▶ Users can stage data in and out and run MapReduce commands. They do not have any administrative access to the cluster.

# Hadoop Clusters

- ▶ *Private Production Cluster*: A dedicated set of systems, which could be hosted on-site or in the cloud, but are typically controlled only by your organization. These clusters are typically not visible outside the organization.
  - ▶ The `cscluster` is a dedicated cluster with the head node `cscluster00` and 8 worker nodes (`cscluster01 ... csccluster08`). The `cscluster00` head node isn't visible outside of the Boise State domain.
  - ▶ Typically, the worker nodes are on a private subnet and not accessible directly outside of the head node.
- ▶ *Ad hoc Cluster*: Here we grab random machines that we have access to create a cluster on the fly, use it, and then destroy it.
- ▶ *Clusters in the Clouds*: Amazon Web Services supports Hadoop and Spark clusters that are created on demand and can be scaled automatically. Microsoft Azure has similar capabilities.

# Port Forwarding (aka Tunneling) to Access Hadoop Web Interface

- ▶ Use ssh port forwarding to enable access to Hadoop ports from a browser at home. We will need to forward ports 9870 and 8088. Let us see how to forward one port first.
- ▶ To forward one port, log in to `cscluster00.boisestate.edu` as follows:

```
ssh -Y -L 9870:localhost:9870 csccluster00.boisestate.edu
```

- ▶ Then point browser on your local system to `localhost:9870` and you will have access to the Hadoop web interface without physical presence in the lab or the slow speed of running a browser remotely.
- ▶ To forward multiple ports, use multiple `-L` options as follows:

```
ssh -Y -L 9870:localhost:9870 -L 8088:localhost:8088  
csccluster00.boisestate.edu
```

- ▶ To tunnel via onyx to `cscluster00` from your machine (at home)!

```
ssh -L 9870:localhost:9870 onyx.boisestate.edu ssh -L  
9870:localhost:9870 csccluster00
```

# Fully Distributed Hadoop

Normally, Hadoop runs on a dedicated cluster. In that case, the setup is a bit more complex than for the pseudo-distributed case.

- ▶ Specify hostname or IP address of the master server in the values for `fs.defaultFS` in `core-site.xml` and `mapred.job.tracker` in `mapred-site.xml` file. These are specified as host:port pairs. The default ports are 9000 and 9001.
- ▶ Specify directories for `dfs.name.dir` and `dfs.data.dir` and `dfs.replication` in `conf/hdfs-site.xml`. These are used to hold distributed file system data on the master node and worker nodes respectively. Note that `dfs.data.dir` may contain a space- or comma-separated list of directory names, so that data may be stored on multiple devices.
- ▶ Specify `mapred.system.dir` and `mapred.local.dir` in `conf/hadoop-site.xml`. The system directory must be accessible by server and clients. The local directory determines where temporary MapReduce data is written. It also may be a list of directories.

# Fully Distributed Hadoop (contd.)

- ▶ Specify `mapred.map.tasks` (default value: 2) and `mapred.reduce.tasks` (default value: 1) in `conf/mapred-site.xml`. This is suitable for local or pseudo-distributed mode only. Choosing the right number of map and reduce tasks has significant effects on performance.
- ▶ Default Java memory size is 1000MB. This can be changed in `conf/hadoop-env.sh`. This is related to the parameters discussed above.
- ▶ List all worker host names or IP addresses in your `conf/workers` file, one per line. List name of master nodes (can be more than one) in `conf/masters`.

# Sample Config Files

- ▶ Sample `core-site.xml` file.
- ▶ Sample `hdfs-site.xml` file.
- ▶ Sample `mapred-site.xml` file.

# References

- ▶ *Hadoop: An open source implementation of MapReduce*. The main website: <http://hadoop.apache.org/>.
- ▶ Documentation for Hadoop 3.3.6:  
<https://hadoop.apache.org/docs/r3.3.6/>
- ▶ Documentation for Hadoop Cluster Setup:  
<https://hadoop.apache.org/docs/r3.3.6/hadoop-project-dist/hadoop-common/ClusterSetup.html>