Hadoop Tutorial

Jinlai Xu

Preliminaries

- Familiar with Linux shell: COMMAND --help
  - cd: change current work directory
  - mkdir: make a directory
  - ls: list the directory
  - ln: make a link of a file or a directory
  - cat: print the file content in the shell
  - ssh: secure login
  - scp: secure copy
  - …
- Nano: a text editor on the keyboard
- Trouble shooting:
  - Copy the error message
  - Google it!

Prepare single node- Required Software

- Required Software on client
  - SSH client
    - Cygwin
    - Putty
- Required Software on node
  - Oracle JAVA 8 for Hadoop 3.0.0 (other version is fine)
  - SSH
  - rsync
  - wget (for download the package)

Setup single node – Download Hadoop

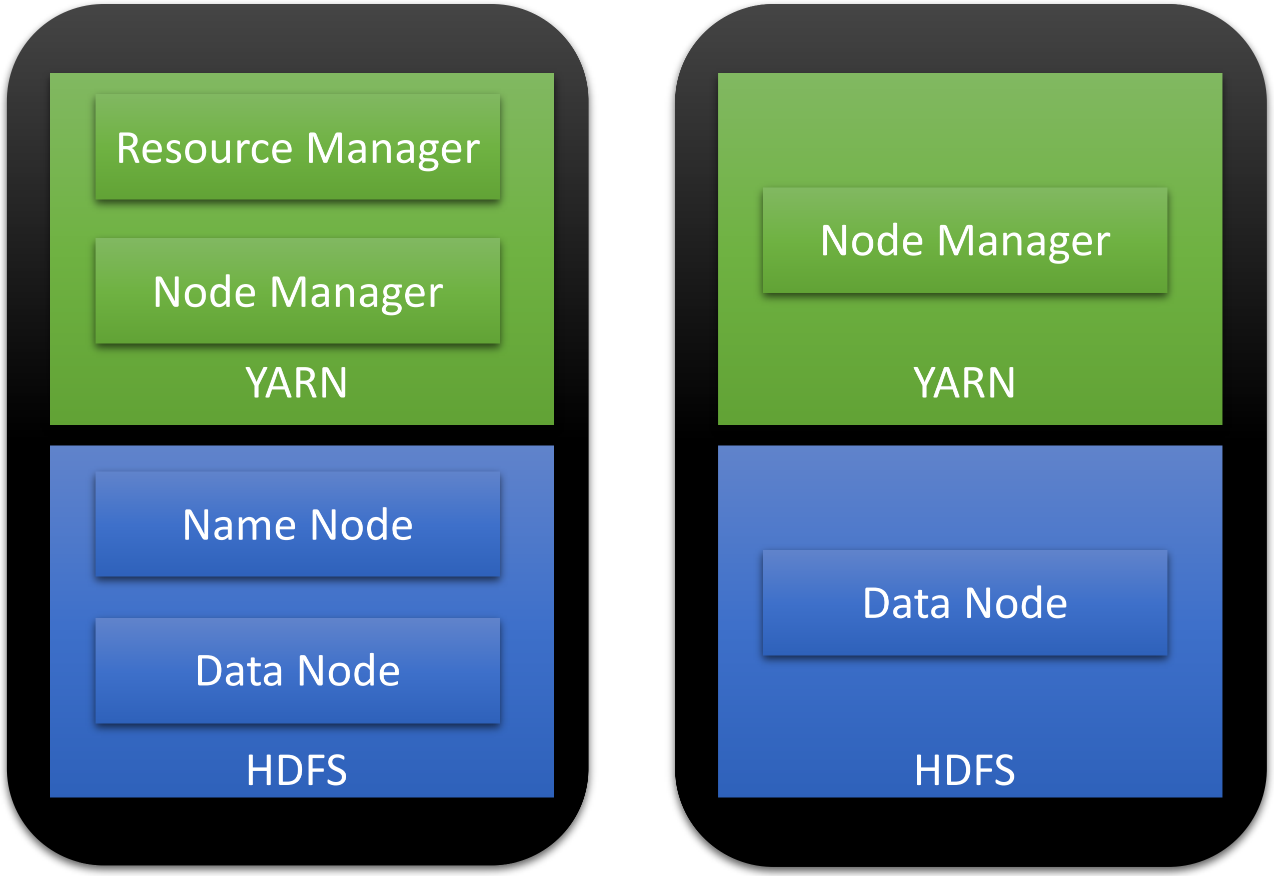
- Download Hadoop and test locally
  - cd /usr/local/
  - wget http://www-us.apache.org/dist/hadoop/common/hadoop-3.0.0/hadoop-3.0.0.tar.gz
  - tar -zxf hadoop-3.0.0.tar.gz
  - ln -s hadoop-3.0.0 hadoop
  - cd hadoop
- Set to the root of your Java installation in hadoop_env.sh
  - export JAVA_HOME=/usr/lib/jvm/java-8-oracle
  - nano etc/hadoop/hadoop_env.sh

Setup single node - Test Hadoop locally

- Hadoop Binary
  - bin/hadoop
- Test with local tasks
  - bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.0.0.jar pi 2 5

  - mkdir input
  - cp etc/hadoop/*.xml input
  - bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.0.0.jar grep input output 'dfs[a-z.]+'
  - cat output/*

Setup Hadoop – Cluster Overview



Resource Manager

Node Manager

YARN

Name Node

Data Node

HDFS

Node Manager

YARN

Data Node

HDFS

Hostname: master
IP: 159.203.71.174

Hostname: slave
IP: 159.203.167.80

Setup Hadoop – Configure the cluster

- SSH interconnection
  - Configure etc/hosts on every node
  - SSH key generate on Master
  - SSH key delivery from Master to every Slaves
- Configure Cluster Environment
  - Set Master node and Slave nodes
  - Configure environment in /etc/environment
  - Configure .bachrc for root
  - Create HDFS directories
    - Name node
    - Data node

Setup Hadoop – Configure Environment

- Configure the cluster setting on every node
  - core-site.xml
  - hdfs-site.xml
  - yarn-site.xml
  - mapred-site.xml
- Format the Name node on master
  - hadoop namenode -format

Start Cluster and Test on master

- Start HDFS
  - sbin/start-dfs.sh
- Start YARN
  - sbin/start-yarn.sh
- Start Job History Server
  - $HADOOP_PREFIX/sbin/mr-jobhistory-daemon.sh --config $HADOOP_CONF_DIR start historyserver
- Monitor the services on master and slave
  - jps
  - ssh slave
  - jps
  - exit
- Test HDFS with Commands
  - hdfs dfs -mkdir input
  - hdfs dfs -put etc/hadoop/ input
  - hdfs dfs -ls input

Monitor and Test

- Monitor the services on the integrated websites:
  - http://159.203.71.174:9870 HDFS
  - http://159.203.71.174:8088 YARN Resource manager
  - http://159.203.71.174:19888 MapReduce History
- Test with some examples
  - bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.0.0.jar pi 2 5
  - bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.0.0.jar wordcount input/ output/
  - hdfs dfs -cat output/*

Thanks!