CS4225/CS5425 Big Data Systems for Data Science

Tutorial 1 Hadoop Introduction

Chengxi Xue, Bingsheng He
School of Computing
National University of Singapore
xuechengxi@u.nus.edu

School of Computing

Hadoop Introduction

- Outline
  - Background about Hadoop
  - Start on different platforms
  - How to setup the Docker for a cluster
  - How to start a Hadoop service
  - How to use HDFS
  - Where you can find the configurations of your Hadoop environments

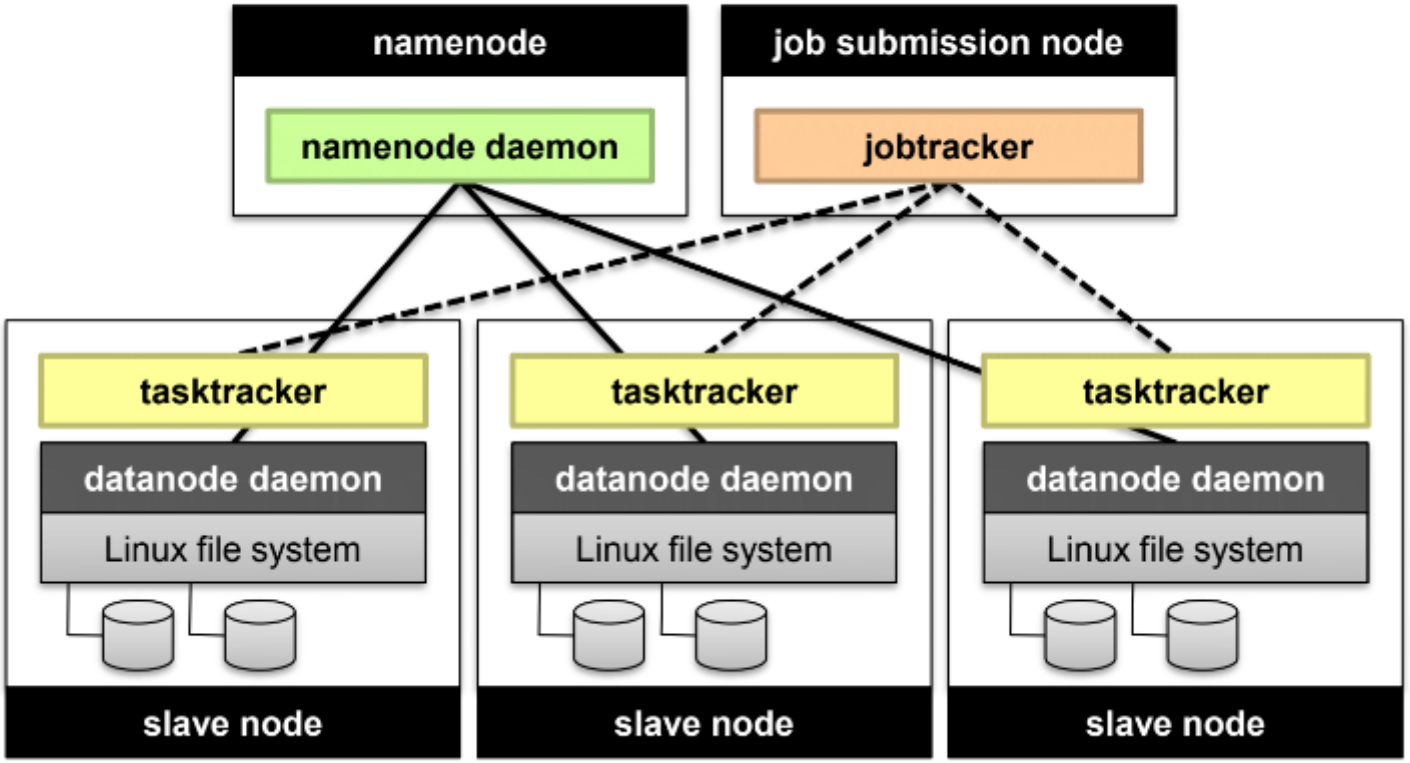
Hadoop

- A collection of open-source software utilities.
- Hadoop provides a software framework for distributed storage and processing of big data using the MapReduce model.
- Hadoop focused on scalability, flexibility, fault tolerance.

Hadoop Modules

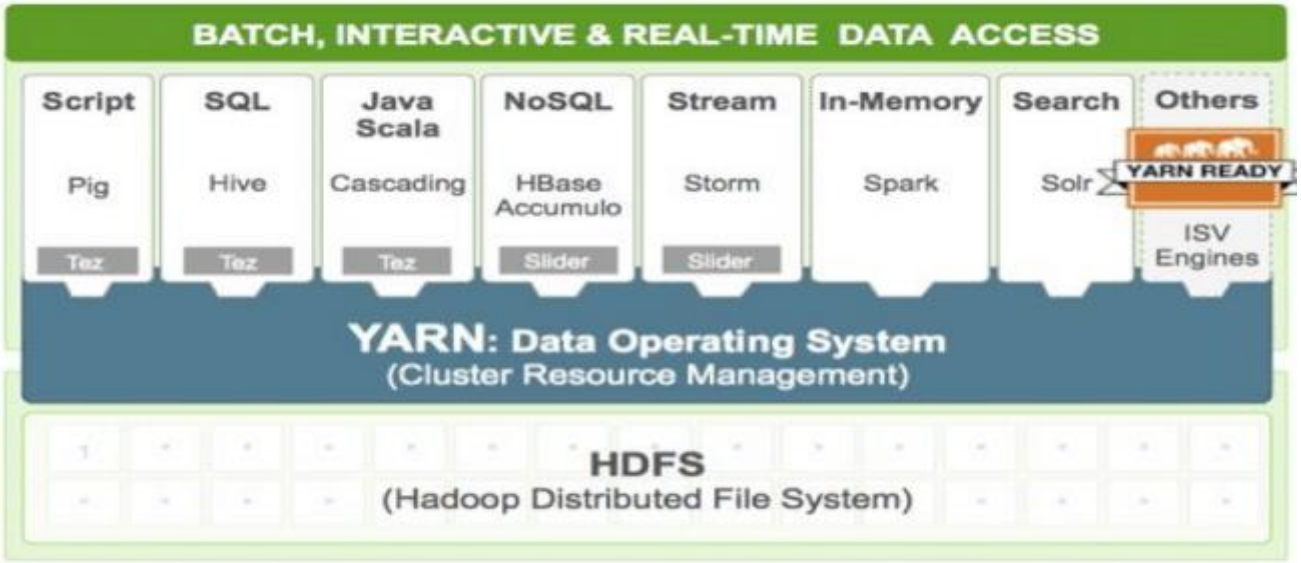
- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

Hadoop Cluster Architecture

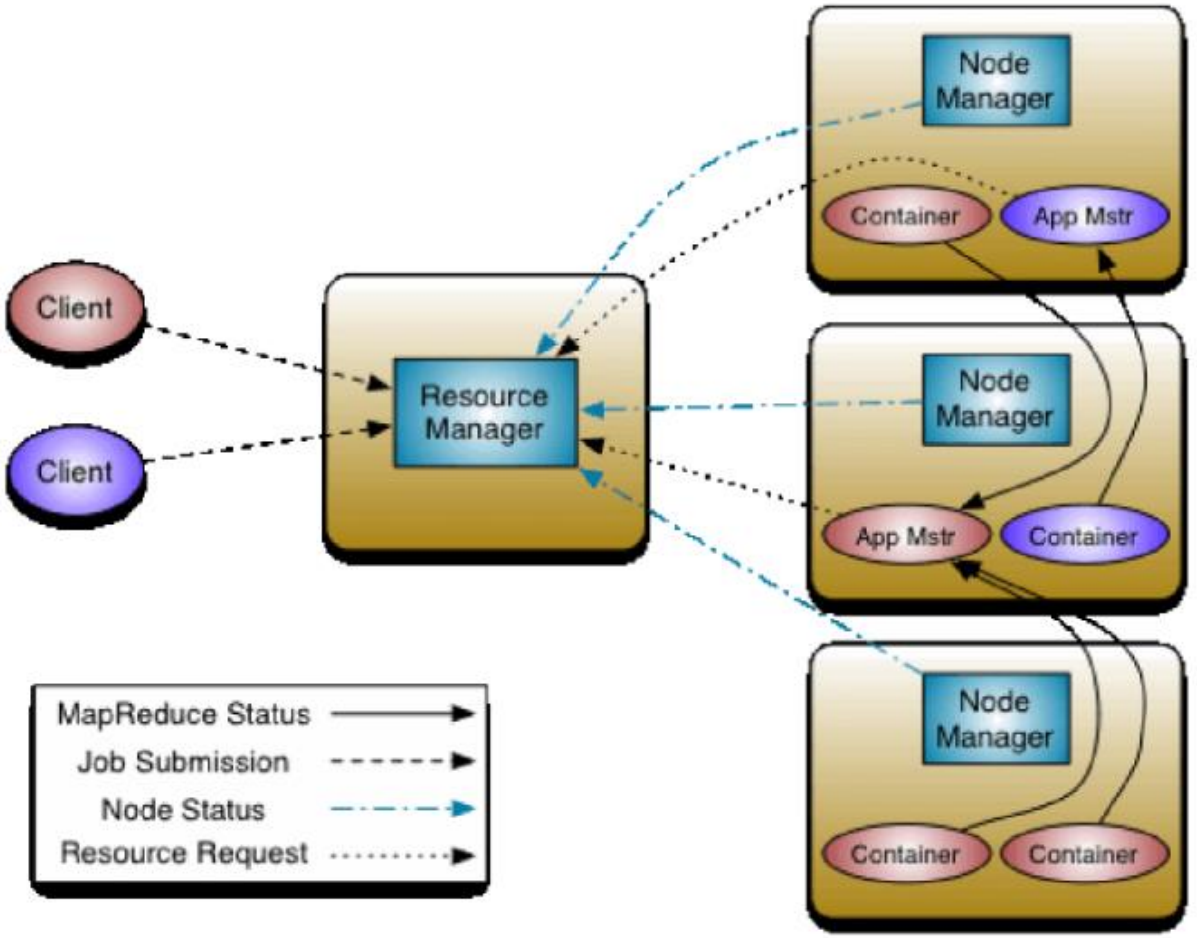


YARN

- Hadoop limitations
  - Can only run MapReduce
  - What if we want to run other distributed framework
- Yarn = Yet-Another-Resource-Negotiator
  - Provides API to develop any generic distribution application
  - Handles scheduling and resource request
  - MapReduce(MR2) is one such application in YARN



YARN: Architecture



About OS

- Ideally, we can build the Hadoop and start the assignments on different platforms.
  - Windows
  - MacOS
  - Linux
- **However, we highly recommend you to use Linux and MacOS.**

Recommended Steps

- Three steps:
  - Step1: Set up the IDE, write the code.
  - Step2: Test your code using single node mode/without Hadoop.
  - Step3: Package your project and submit your task into clusters for final testing.
- Guidelines:
  - Step 1 and Step 2, follow "Run_the_example_win_macos_linux.docx"
  - Step 3, see the following slides and follow "Run_example_using_hadoop.docx"
    - Slide 11: setup a cluster with docker.
    - Slide 12: package the project
    - Slides 13-19: configure Hadoop/HDFS in the docker cluster.

Setup a Docker for a cluster

- Download the Docker.
  - Login “docker app” with your Docker account
  - Run “**docker login**”.
  - Run **“docker pull nusbigdatacs4225/ubuntu-with-Hadoop-spark”**
- Create clusters
  - Run **“docker run -it -h master --name master ubuntu-with-hadoop-spark”, “docker run -it -h slave01 --name slave01 ubuntu-with-hadoop-spark”…….**
  - Check IP address and configure the cluster
- You can refer to the documents “Installation&Configuration” in assignment1

Testing with the cluster

- Package your whole project into a jar file using maven/sbt.
- Copy the jar file to the docker cluster.
- Run the application (jar file) using hadoop as the final testing.

Start a Hadoop service

- Initialize the HDFS
  - You can find Hadoop and Spark in /usr/local/..

```
root@master:/usr/local# ls
bin  etc  games  hadoop  include  lib  man  sbin  share  spark  src
root@master:/usr/local#
```

  - In Hadoop, you can see

```
root@master:/usr/local/hadoop# ls
LICENSE.txt  README.txt  etc   include  libexec  share
NOTICE.txt   bin         hdfs  lib      sbin     tmp
```

  - Structure
    - bin: basic scripts
    - etc: Hadoop configuration files
    - sbin: scripts to start/stop services (HDFS, YARN…)
    - …
  - Set up the master and slaves:
    - e.g. In /etc/Hadoop, vim slaves, add “slave01 slave02”.
  - Initialization: bin/hdfs namenode -format

Start a Hadoop service

- Start the Hadoop
  - Run “sbin./start-all.sh”

```
# Start all hadoop daemons.  Run this on master node.

echo "This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh"

bin=`dirname "${BASH_SOURCE-$0}"`
bin=`cd "$bin"; pwd`

DEFAULT_LIBEXEC_DIR="$bin"/../libexec
HADOOP_LIBEXEC_DIR=${HADOOP_LIBEXEC_DIR:-$DEFAULT_LIBEXEC_DIR}
. $HADOOP_LIBEXEC_DIR/hadoop-config.sh

# start hdfs daemons if hdfs is present
if [ -f "${HADOOP_HDFS_HOME}"/sbin/start-dfs.sh ]; then
  "${HADOOP_HDFS_HOME}"/sbin/start-dfs.sh --config $HADOOP_CONF_DIR
fi

# start yarn daemons if yarn is present
if [ -f "${HADOOP_YARN_HOME}"/sbin/start-yarn.sh ]; then
  "${HADOOP_YARN_HOME}"/sbin/start-yarn.sh --config $HADOOP_CONF_DIR
fi
                                                          38,1          Bot
```

Start a Hadoop service

- You can see following jobs are running in master:

```
root@master:/usr/local/hadoop# jps
3290 DataNode
3147 NameNode
3726 NodeManager
4030 Jps
3455 SecondaryNameNode
3615 ResourceManager
root@master:/usr/local/hadoop#
```

- In slave01 and slave02:

```
root@slave01:/# jps
688 Jps
439 DataNode
555 NodeManager
root@slave01:/#
```

Run The Application

- You can run a jar file as an application using hadoop.

  E.g. bin/hadoop jar <jar> [mainClass] args..

- Notes: Read the APACHE Hadoop website to see all the support materials you want.
  - https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

Operations in HDFS

- Create a file in HDFS:
  - bin./hadoop fs -mkdir ***
- Upload a local file
  - bin/hadoop fs -put *local file* *target path*
- List the file
  - bin/hadoop fs -ls /

.

.

.

- Notes: Read the APACHE Hadoop website to see all the support materials you want.
  - https://hadoop.apache.org/docs/r2.7.3/hadoop-project-dist/hadoop-common/ClusterSetup.html

Configurations

- Some important configuration files you need to read
  - etc/hadoop/core-site.xml
  - etc/hadoop/hdfs-site.xml
  - etc/hadoop/yarn-site.xml
  - etc/hadoop/mapred-site.xml
- You can rewrite these four files to deploy your own system.
- You can control the Hadoop scripts found in the bin/, by setting site-specific values via the etc/hadoop/hadoop-env.sh and etc/hadoop/yarn-env.sh.

Configurations

Configuration Filenames	Description
hadoop-env.sh	Environment variables used in the scripts to run Hadoop
core-site.xml	Configuration settings for Hadoop Core such as I/O settings that are common to HDFS nad MapReduce
hdfs-site.xml	Configuration settings for HDFS daemons, namenode, datanode...
mapred-site.xml	Configuration settings for MapReduce daemons
Yarn-site.xml	Configuration settings for YARN, resource manager, node manager
masters	A lits of machines that each run a secondary namenode
slaves	A list of machines that each run a datanode and task-tracker

Advice

- **Read the APACHE Hadoop website**
  - **https://hadoop.apache.org/docs/r2.7.3/hadoop-project-dist/hadoop-common/ClusterSetup.html**
- **And you will find most of the answers you want!**

Questions?

- Now, make your hands dirty ☺
- Contact me: xuechengxi@u.nus.edu

Img source: http://mpchristianity.com/making-leaders-gets-hands-dirty.html