Hadoop distributed file system

Deadline : Friday, 2019/06/15 23:59

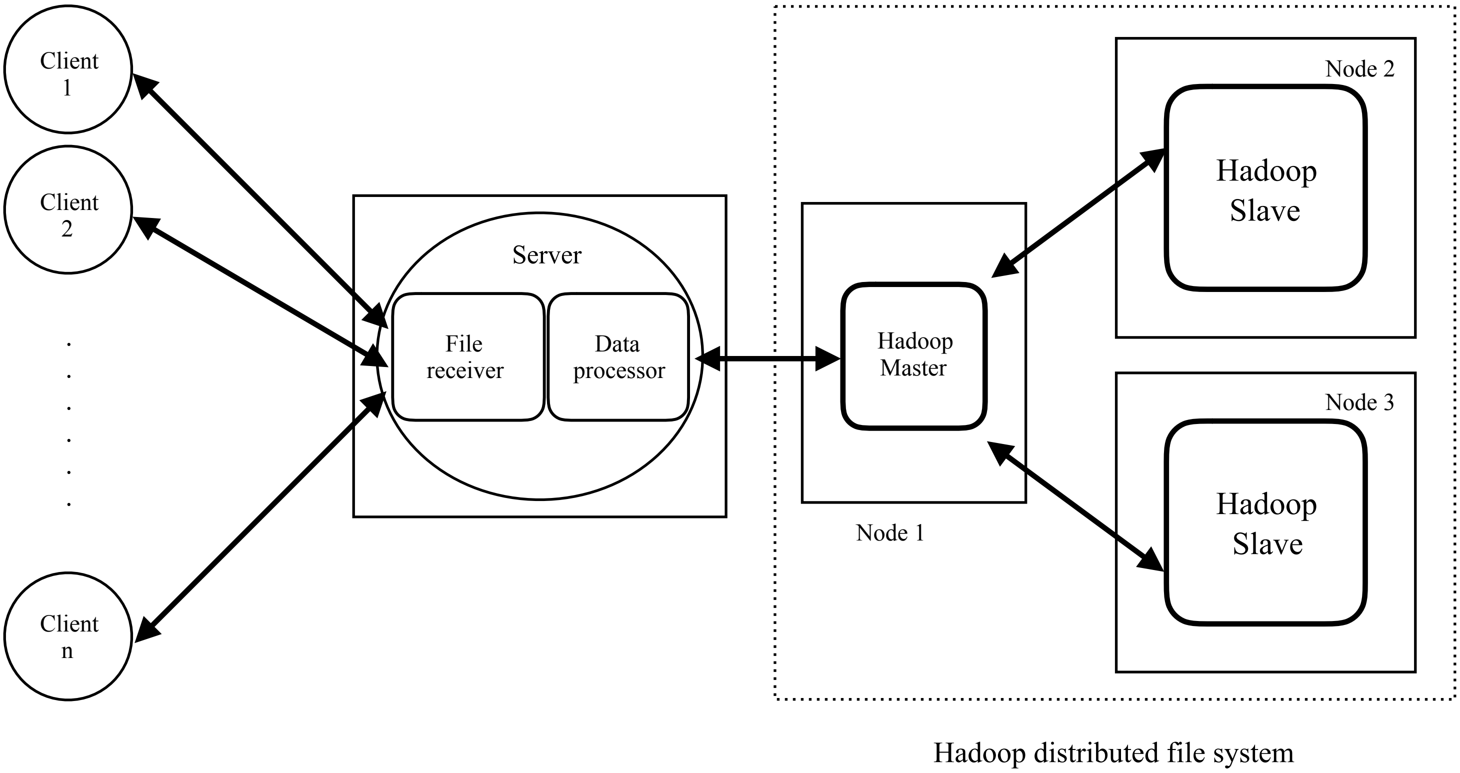
Overview

Apache Hadoop is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation.

In this homework, we are going to setup a Hadoop distributed file system with a real time server to handle the multiple data streaming.

Specification

- Components of server:
  - **File receiver**
    - Receive files from multiple clients at the same time.
    - The maximum number of the clients is 25.
    - The size of the file is about 150MB.
  - **Data processor**
    - Saves the data to the HDFS while file receiver receiving the data.



Note

- We have no limitation on the programming language.
- Server, clients and Hadoop distributed file system should setup on AWS instances.
- The server need to record the duration, from starting to receive the files until server saves all the files.
- The type of the instance is limited. When launch the instances,
  - In step 2, please select the “Free tier eligible” one
    - Family : General purpose
    - Type : t2.micro
    - vCPUs : 1
    - Memory(GiB) : 1
    - Instance storage(GB) : EBS only
  - In step 4, please select
    - Size(GiB) : 30
    - Volume type : General Purpose SSD (gp2)

Grades

During the demo, you need to create 25 clients and send the video files to server.

- If the save all the files safely, you will get **70%**.
- All of the students will be divided to six groups by the duration.
  - The shortest 5 student will get **30%**.
  - 6th to 10th will get **25%**.
  - 11th to 15th will get **20%**.
  - 16th to 20th will get **15%**.
  - 21th to 25% will get **10%**.
  - 26th to 30th will get **5%**.

P.S. if clients do not send the files at the same time, **you will not get any score**.

File submission

Upload you source codes to new E3 platform directly.

TA would validate your source codes by cheating detection. Please finish the assignment by yourself.

Reference

Apache Hadoop

Hadoop cluster setup