Experience in Graph Computation in distributed system

**Group members:** Linna Hu, Hongyu Huang, Jiashuo Sun

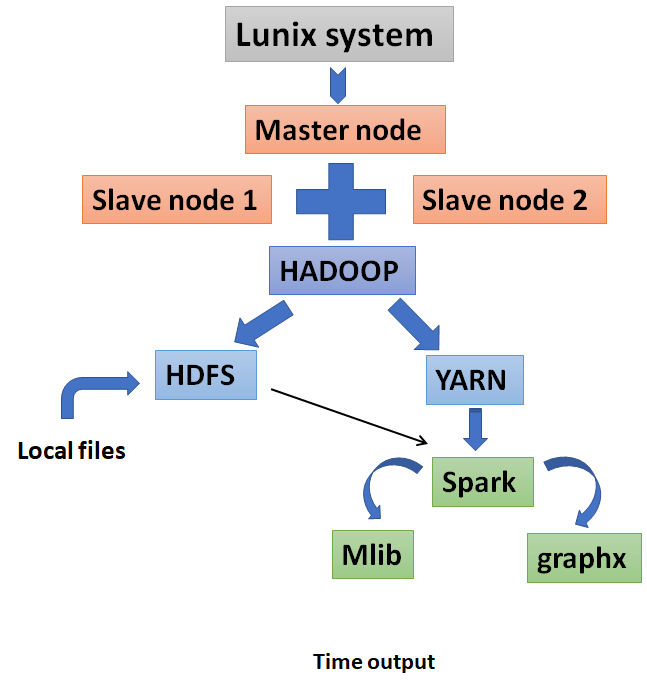
**Background:**

Increasing number of data content and user volume makes it necessary to envelope an optimized and more efficient distributed computation system. Now, machine learning plays an important role in intelligent service. Improved Topic-modeling and click-through prediction, a family of algorithms, make it more convenient for users to enjoy the smart service.

Nowadays, Big Data plays a popular role both in industrial and academic fields. Hadoop is developed as a solution to Big Data. It provides reliable, scalable, fault-tolerance and efficient service for large scale data processing based on HDFS and MapReduce.

**Preparation:**

We set up three nodes in a cluster. Then we set up an environment in the cluster which included Hadoop, Spark.



**Main work:**

**2.1**  We registered a new account and got our SSH key address using linux system. Then we started a new cluster on the Cloudlab. We have to make sure each node could switch to another one without any problem and any password when we doing the work. For the next step, we deploy a java environment.

**2.2** We set up **Hadoop**.

Before setting up it, we should download Yarn 2.7.2 and set up HDFS,Yarn to start the Hadoop.

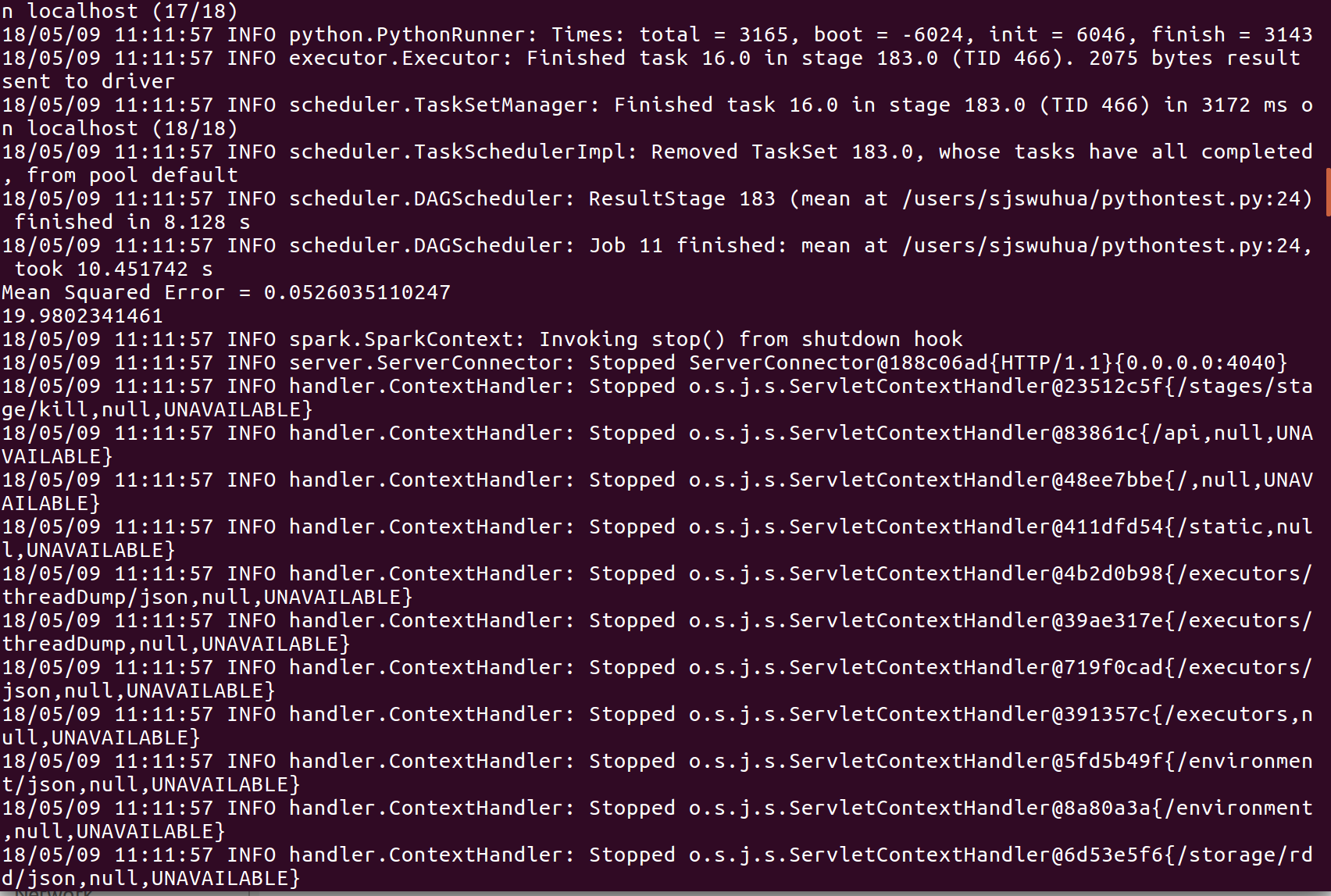
**2.3** To make sure the success of running Hadoop, we log in the 50070 and 8088 website. And we saw that the status of application in Yarn was just accepted, not in running. The website reminded us that “waiting for the AM to register and launch to RM”. To correct this mistake, we did a lot. Finally we noticed that we missed a step : edit hostname and IP in the /etc/hosts.

**2.4** Then we set up **Spark** and Mapreduce on Yarn. We need to run the pyspark in the spark, and import dataset and address dataset on the pyspark.

We first uploaded the dataset to the hdfs and run the script that we edited in the pyspark to address the dataset.

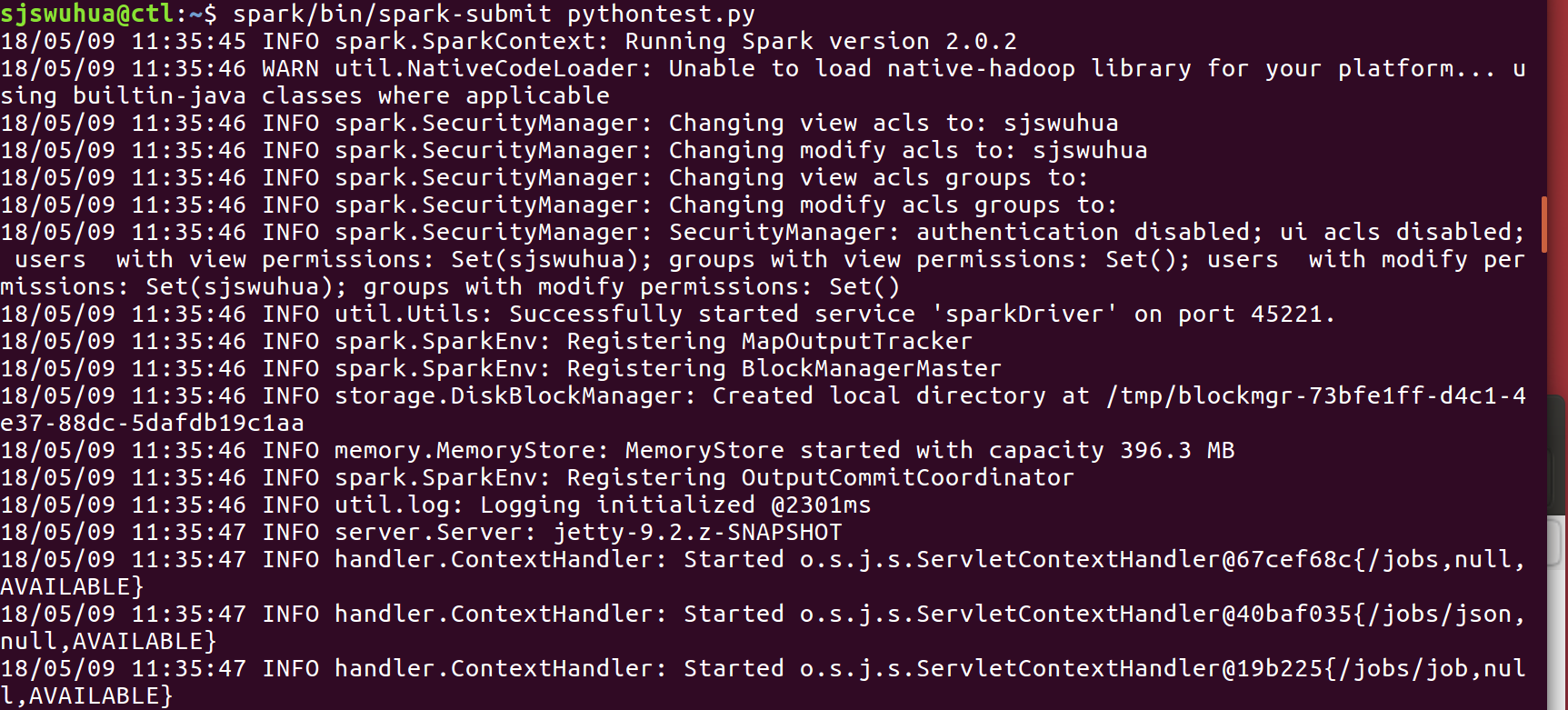
We want to compare the time that running application just in ctl to the time running in all of nodes, 2 slaves and a master.

Run a 1MB dataset.



We can see that the running time is 19.9802341461.

Then we run a 10MB dataset.



In the website of Spark, we can see the working status.



The completed status.

