Setup Yahoo SAMOA on S4

On Virtual Machine with Ubuntu Linux

Li Huang

Computer Engineering and Computer Science
J.B. Speed School of Engineering
University of Louisville
l0huan08@gmail.com

2014/1/29
Edit: 2014/2/5

Statement:

This article only reflects the author's personal opinion, and may contain some views not right. If you find something wrong in this article or have advices please contact the author. This article is for non-commercial purpose, which helps the people who are interested to SAMOA. Some figures are referenced from some other sources.

Contents

1. Setup Environment....2
1. Preparation....2
2. Setup SAMOA....2
3. Test SAMOA....4
   3.1 Simple Test -- Cluster Evaluation....4
   3.2 Forest Cover Type Classification....6
4. Recommend Readings....8

SAMOA(Scalable Advanced Massive Online Analysis) is a new distributed streaming data mining platform which is now being developing by Yahoo, and S4 is one of the support platform for running SAMOA(another is Twitter Storm). Since currently there is very few material can be find to introducing the Setup process of SAMOA and S4, in this article I will introduce my setup process of Yahoo! SAMOA based on S4. This article assumes you have already setup the Ubuntu system and Yahoo S4. For the detail of setup Ubuntu and S4, please see my another article "***Yahoo SAMOA Whitepaper 2-Setup Yahoo S4***".

Comparing with setup S4, setup SAMOA is much **easier**! The whole process takes about **1 hour**.

1.Setup Environment

I use my daily-use laptop to setup SAMOA and S4. Its configuration is below:

Hardware:

Computer: Lenovo Laptop U310

CPU: Intel Core i3-2367M, with 2 cores at 1.40Ghz

Memory(RAM): 2.00 GB

System Type: 64 bit

Software:

Base operating system: Windows 7 Professional Service Pack 1 - 64 bit

Virtual Machine to host S4: VMware Player 5.0.2

Operating system on Virtual Machine: Ubuntu Linux Desktop 12.04 – 64 bit

Stream Platform: Yahoo S4 0.6.0

1.Preparation

Before setup SAMOA, you need setup a Linux and a Stream Process Engine (SPE) in your computer. For operating system, I setup Ubuntu 12.04 desktop 64bit in a virtual machine. Currently SAMOA supported Twitter Storm and Yahoo S4. In this article, I use Yahoo S4 as my SPE, since I failed to setup Storm on my virtual machine (it is hard to configuration).

- Note: Another reason I gave up Storm is that: To test Storm in Cluster environment, we need at least 3 nodes--a nimbus(center server) and 2 computing nodes, but my computer has not so much memory for running 3 VMs at same time.

2.Setup SAMOA

This process is simple and as same as the official document. The only things you need to note are:

Firstly you need install a tool called “maven” to build SAMOA.

SAMOA support three modes: local, S4, Storm. Local mode run on a single machine without any SPE; S4 mode run SAMOA on S4; Storm mode run SAMOA on Storm.

The process of setup SAMOA:

1. Setup maven. (I use maven 3.1.1. It is a tool to build SAMOA)
   `apt-get install maven`
   Or: find “maven” in “**ubuntu software center**”
   Or: download it from the official website
2. Download SAMOA from its website
   Go to the dir “`cd /opt`“.
   Clone SAMOA 0.0.1 from GitHub.

`git clone git://github.com/yahoo/samoa.git`

And set its write permission to all user by "`sudo chmod -R 777 /opt/samoa`"

3. In the "**samoa**" folder,

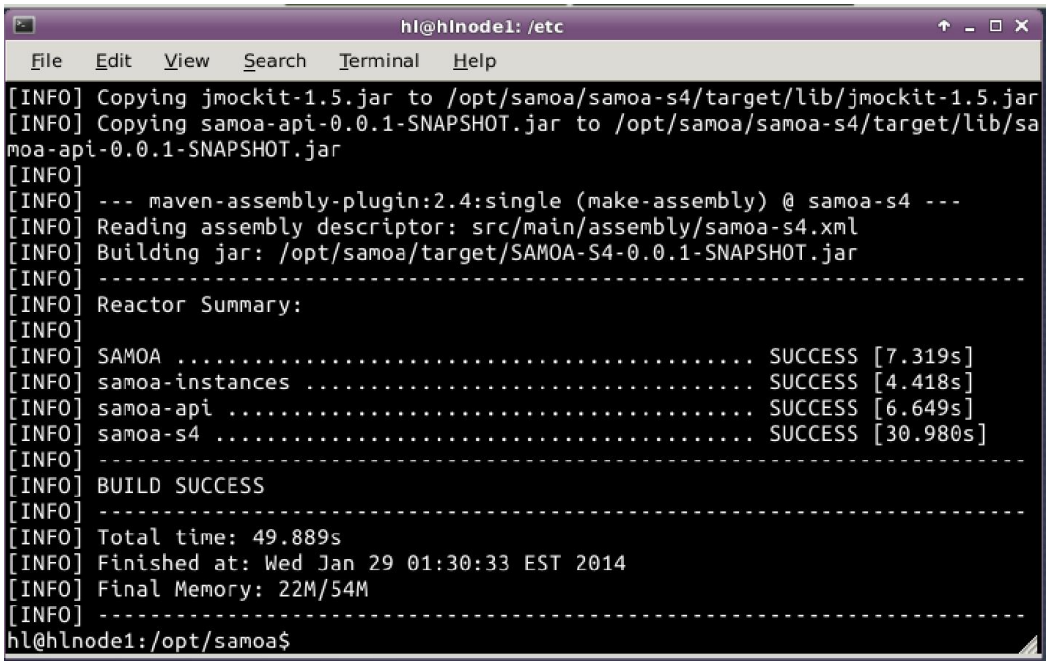
Run "`mvn package`",

to build the software package of "**local mode**" will be build. The result file is "**target/SAMOA-Local-0.0.1-SNAPSHOT.jar**"

Run "`mvn -Ps4 package`"

to build the softwarepackage of "**S4 mode**" will be build. The result file is " **target/SAMOA-S4-0.0.1-SNAPSHOT.jar**"

- ***Note**: SAMOA local mode do not need any SPE, but can only run a single machine rather than a cluster.*



4. Edit configure file, "**<samoa folder>/bin/samoa-s4.properties**".

Same as the official documents, but the document is not clear with this part:

```
# Deployment strategy
samoa.deploy.mode=local
```

- **Note**: When I saw this setting at first time, I did not know what was it mean. Now I know, it means that: If you are run SAMOA in a single (virtual) machine, you should set deploy mode as "**local**". When you want to run SAMOA in several (virtual)machines, you need to change "**local**" into "**cluster**"

5. Add SAMOA to environment file by add these line to the end of "**/etc/profile**":

```
export SAMOA_HOME=/opt/samoa
```

And run "source /etc/profile" to update your environment variables.

3. Test SAMOA

After setup SAMOA, you should test SAMOA. We have two simple test app for SAMOA, “Cluster Evaluation” and “Forest Cover Type”. So far (by 2014/1/29).

Note

SAMOA Bug: However, SAMOA 0.0.1 has a bug that you can only run these two examples in “local” mode, but cannot run them on S4 mode. Currently the developers are trying to solve this problem.

3.1 Simple Test -- Cluster Evaluation

This task will do a cluster and evaluate the performance. This example can be found in the last part of :

https://github.com/yahoo/samoa/wiki/Executing-SAMOA-with-Apache-S4

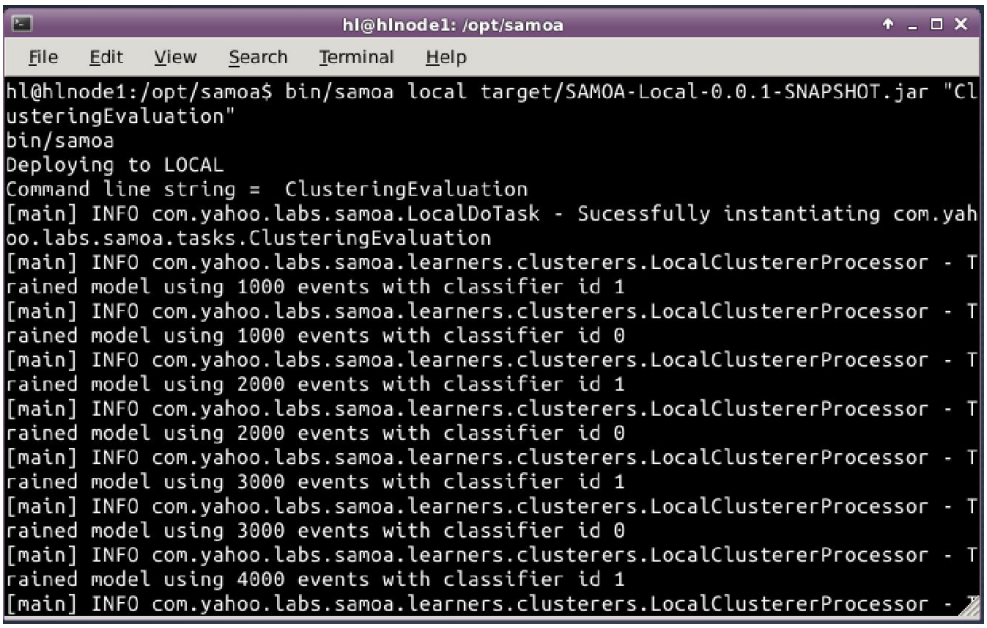
Syntax: `bin/samoa <platform> <jar-location> <task & options>`

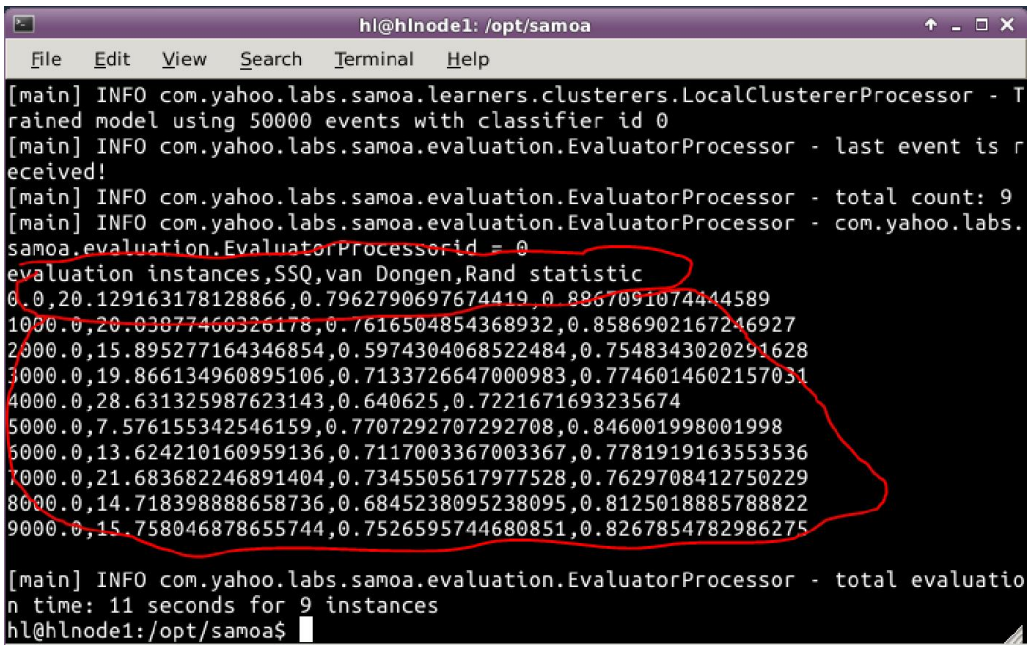
(1) Run in Local Mode

You can run in **local** mode:

```
bin/samoa local target/SAMOA-Local-0.0.1-SNAPSHOT.jar
"ClusteringEvaluation"
```

You will see output result below:





(2) Run in S4 Mode

You can also run it in **S4** mode:

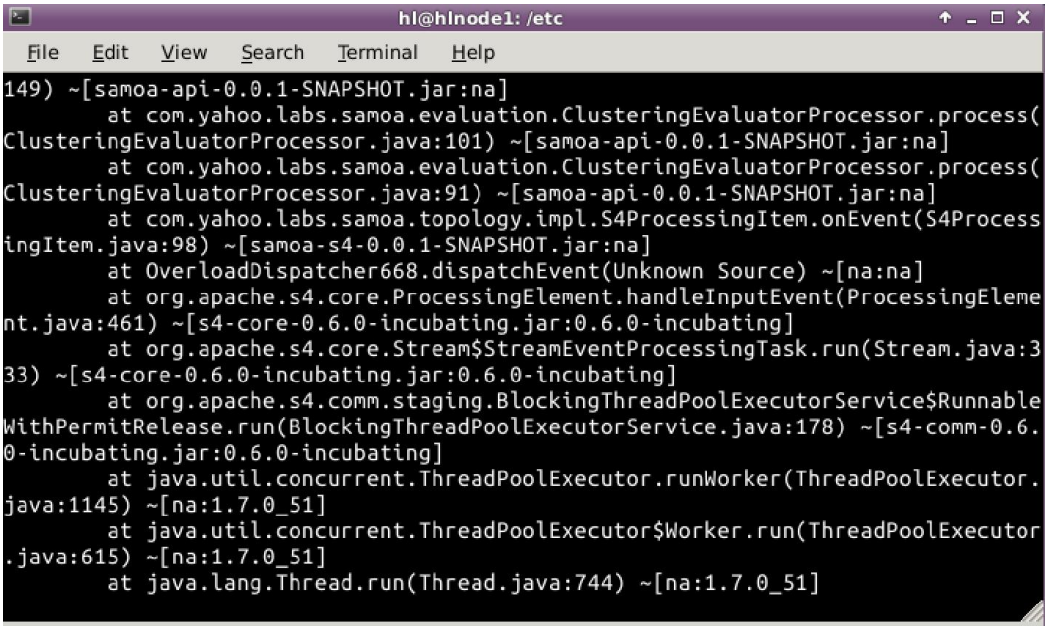
```
bin/samoa S4 target/SAMOA-S4-0.0.1-SNAPSHOT.jar
"ClusteringEvaluation"
```

Important Note:

However, running `ClusteringEvaluation` in **S4 mode** has a bug (by January.29.2014).

Please see https://github.com/yahoo/samoa/issues/33 to find the detail.

In my computer, it shows error message like:



3.2 Forest Cover Type Classification

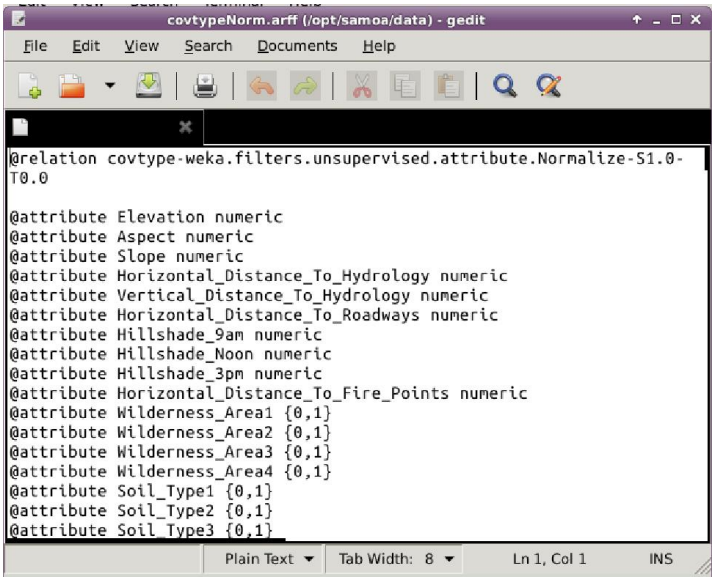
The second example does a classification job on a datasource contains many points of a forest. This example can be found in:

https://github.com/yahoo/samoa/wiki/Getting%20Started

1. Download dataset and unzip it to "**/opt/samoa/data**":

http://downloads.sourceforge.net/project/moa-datastream/Datasets/Classification/covtypeNorm.arff.zip

If we open this file with a text editor, we can see the file like this:



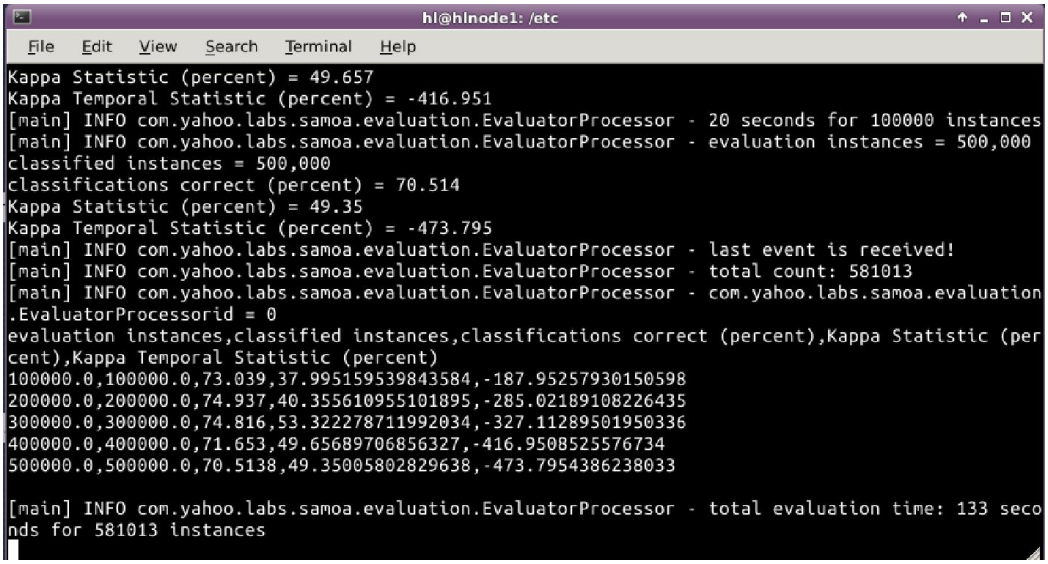
2. classifying the CoverType dataset with the bagging algorithm.

- **Local Mode**

```
bin/samoa local target/SAMOA-Local-0.0.1-SNAPSHOT.jar
"PrequentialEvaluation -l classifiers.ensemble.Bagging -s
(ArffFileStream -f /opt/samoa/data/covtypeNorm.arff) -f 100000“
```

The output will be a list of the evaluation results, plotted each 100,000 instances.

The output will be a list of the evaluation of classification performance results, plotted each 100,000 instances. As the figure below shows:



You can also add "`-d `pwd`/result.csv`" in the command to output the result into a **.csv** file.

- **S4 mode**

You can also try it in **S4 mode**, by typing:

```
bin/samoa S4 target/SAMOA-S4-0.0.1-SNAPSHOT.jar
"PrequentialEvaluation -l classifiers.ensemble.Bagging -s
(ArffFileStream -f /opt/samoa/data/covtypeNorm.arff) -f 100000"
```

Note:

*** The command parameters can be seen at:**

https://github.com/yahoo/samoa/wiki/Prequential%20Evaluation%20Task

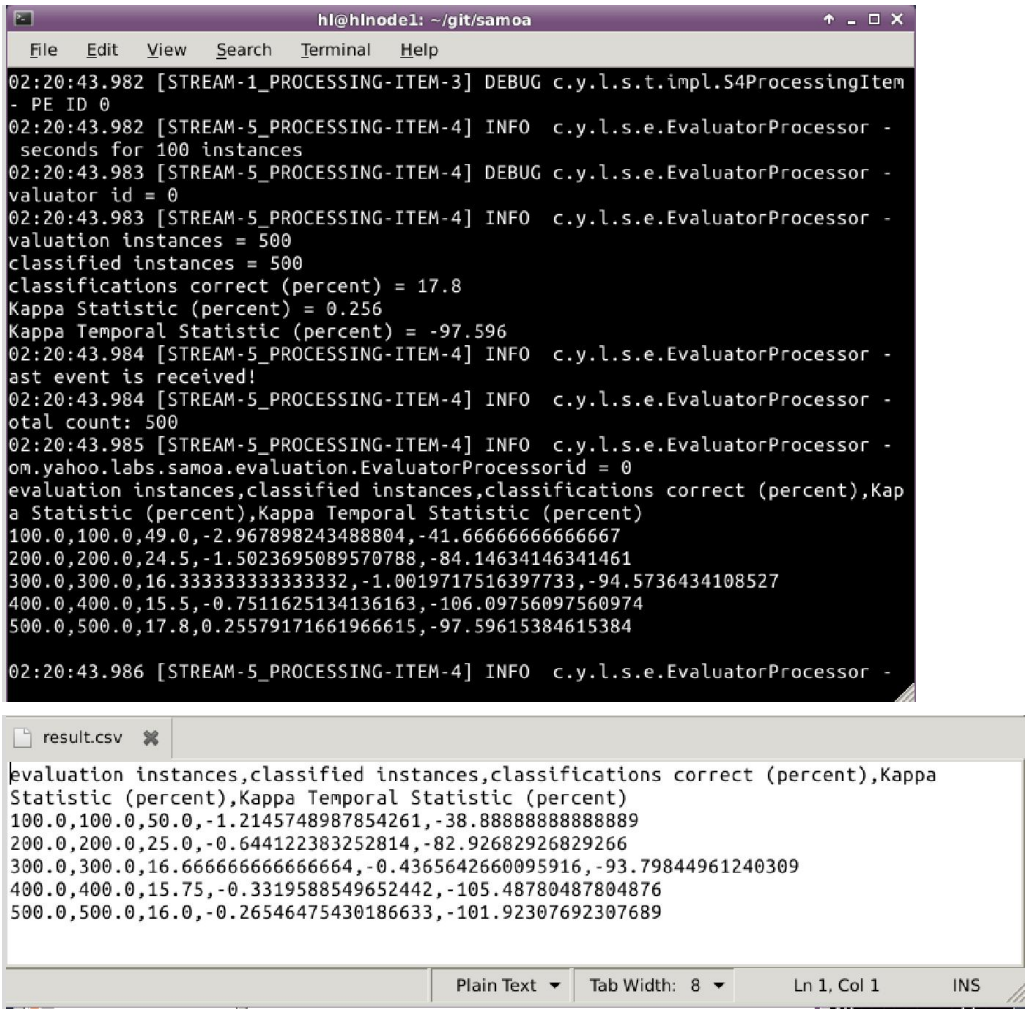
`-f 100000` means the result will be plotted every 100,000 instances

* You can add `-i 200000` to limit the maximum number of instances to test/train on (as 200000)

I recommend use this command below, which only evaluate 500 instances to save your time, and output the result file to "result.csv":

```
bin/samoa S4 target/SAMOA-S4-0.0.1-SNAPSHOT.jar
"PrequentialEvaluation -l classifiers.ensemble.Bagging -s
(ArffFileStream -f `pwd`/covtypeNorm.arff) -f 100 -i 500 -d
`pwd`/result.csv"
```

You will see the result like below, with the result.csv in your SAMOA folder:



This result.csv files shows the **performance** (correct rate, Kappa Statistic) of classification of the first 500 instances.

4. Recommend Readings

- Official website of Yahoo SAMOA, including a document tell how to setup SAMOA on S4:

http://yahoo.github.io/samoa/

https://github.com/yahoo/samoa/wiki/Executing-SAMOA-with-Apache-S4

- Official website of Install Maven

http://maven.apache.org/download.cgi#Installation