Introduction of Yahoo SAMOA

An Open Source Distributed Stream Data Mining Platform

Li Huang

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

2014/1/24

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.SAMOA Introduction

SAMOA(Scalable Advanced Massive Online Analysis) is a framework for **mining big data streams in a cluster/cloud environment**. It is a **distributed streaming machine learning (ML) framework** that contains a programming abstraction for distributed streaming ML algorithms. SAMOA is written in **JAVA**, and must running on **Linux**. SAMOA is a new project which was started at October 2013, and current version (by Jan.4.2014) is 0.0.1, with few developers. This project is belong to Yahoo! and was initially want to work for detecting the spam emails.

There are four questions might be most questioned: Why Yahoo develop SAMOA? What is

SAMOA's architecture? How is SAMOA's ability?

1.1 What is the purpose of SAMOA?

Big Data

The reason is that Yahoo! wants a new platform to deal with "Big Data". Nowadays, "Big Data" becomes a popular word, which means data are exploding and more data is not exploit, because lack of resources (time, people, storage). "Big Data" has three properties called 3V: Volume, Velocity and Variety.

Data Mining

However, just raw data has no meaning. People need to transform, analysis or summary the data. To make the process of data automated and fast, many "machine learning" and "data mining" algorithms and tools have been invented, to let the computers process the data. Currently the popular data mining software includes WEKA, R, SAS, RapidMiner, etc.

One computer cannot process too big data in limited time, so distributed computing platform is used.

Stream

Stream (or called Stream Data) is a kind of Big Data. As the name represent, it is "changing" data. For example, stream data includes the data coming from sensor networks or the Web, online news, micro-blogs, or search queries. They are usually real time data.

Stream arrives at high speed. To process the stream, **streaming algorithms** must process it in one pass under very strict constrains of space and time.

The value of stream is freshness and relatedness to onging events (e.g. online news). Report the news a week before might be useless. So the learning algorithms of stream must be rapid and can deal with evolving data.

Properties of Stream Data:

- Objects arrive continuously.
- Stream sizes can be unbounded and potentially infinite.
- There is no guarantee in arrival order of data objects.
- Data objects are discarded after being processed.
- The data source might be evolving.

Current conventional data mining tools

The list below shows the mostly used data mining tools, most of them are free software. There should be some other commercial software I do not mention here.

1.Non-distributed:

Batch: WEKA, R, SAS …

Batch-(data must be download and stored before analysis. Batch processe need much more space than stream process)

Stream: Aurora, STREAM;

2.Distributed:

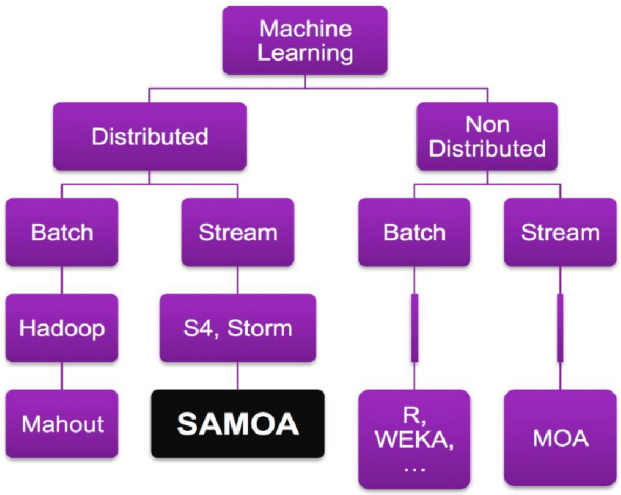
Distributed Batch Process Platforms: Hadoop, MapReduce;

Data mining tool: Mahout (running on Hadoop);
Distributed Stream Process Platforms:
Apache(Yahoo) S4, Twitter Storm;
However, S4 and Storm are generic platforms without data mining tools!
3.Stream Machine Learning Frameworks:
MOA (massive online analysis), VFML(very fast machine learning), RapidMiner.
But they are not distributed!

To utilize the advantages of both distributed stream process platform and machine learning tools, Yahoo! decided to develop SAMOA. We can consider it as:

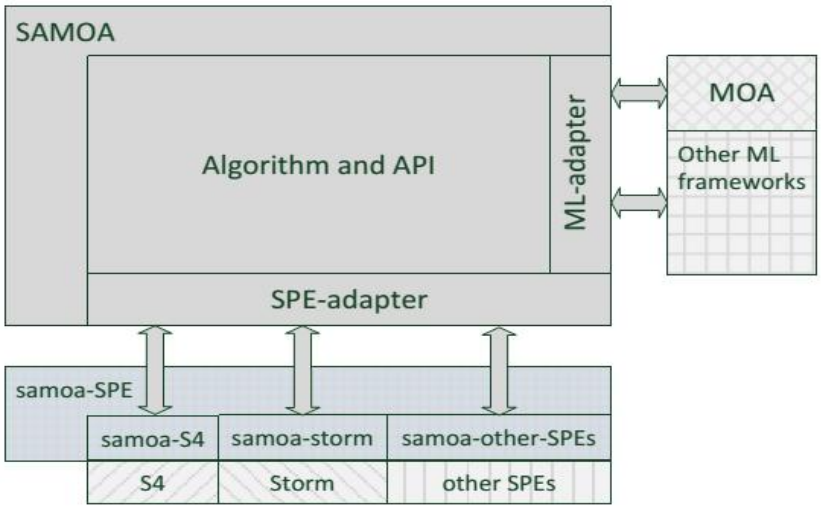
SAMOA = Distributed stream process platform + Machine learning tools

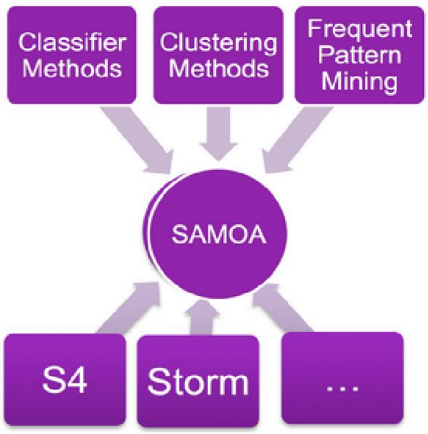
The figure below shows the place of SAMOA in the forest of data mining tools.



1.2 Architecture of SAMOA

SAMOA including two parts: Data Mining Algorithms and Stream Platform. The Algorithms part contains the many data mining or machine learning(ML) algorithms you can use, which should be APIs you can directly called in Java. You also can integrate other machine learning frameworks's algorithms by implementing the ML-adapter classes. The Platform part including a abstract layer called samoa-SPE (SPE=Stream Processing Engines), which Algorithms can run on, and the SPE-adapters to integrate the specific SPEs. Currently Yahoo S4 and Twitter Storm platforms have been supported, and you can also extend the SPE-adapter class to integrated with other SPEs. Once the SPE-adapter for a SPE is implemented, the user does not need to consider the specific API for this SPE, but only need to call the API of ML-algorithms or develop new data mining algorithms based on samoa-SPE abstract layer.
The figures below shows SAMOA's architecture.



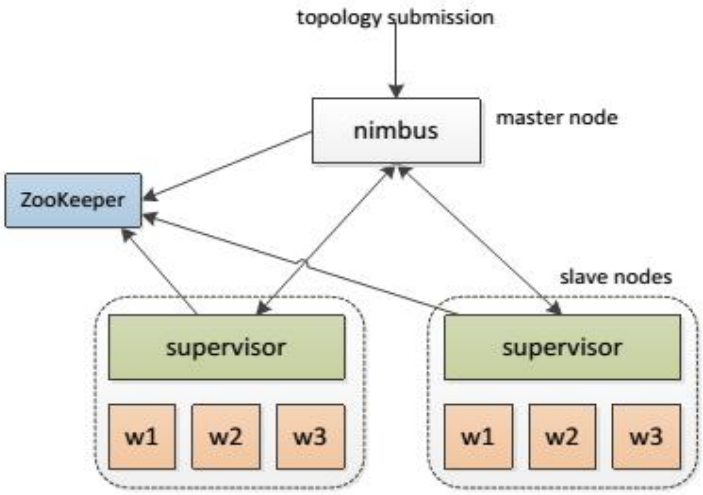


Hardware Architecture and Process Model of SAMOA Cluster

(1) Storm Cluster:

Storm Cluster is a centralized structure, it includes:

Central server (nimbus), Coordinate servers (ZooKeeper cluster), Processing nodes(supervisors). In each supervisor, there are many Workers running (w1,w2,w3,…), and each Worker is a Java virtual machine(JVM) process.



(2) S4 Cluster

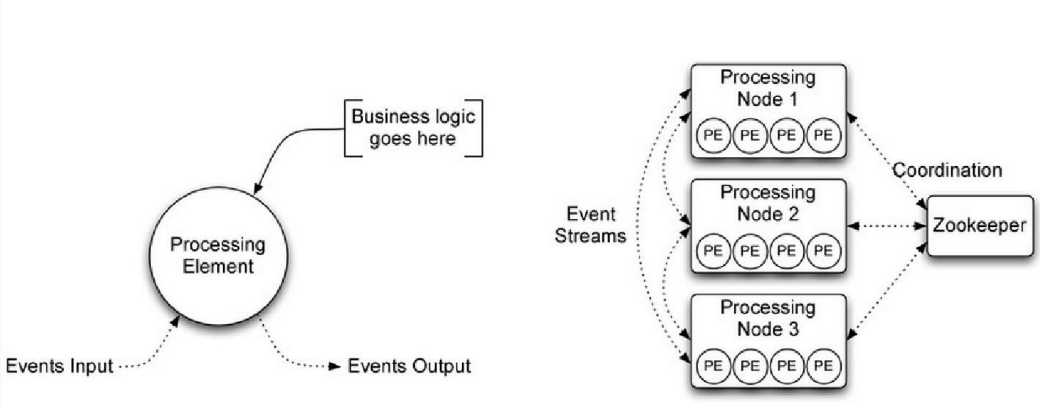
S4 cluster is non-central structure, it includes:

Coordinate servers (ZooKeeper cluster), Processing nodes.

In each Processing ndoe, there are some PE running, and each PE is only a process.

PE: Processing Element; Node; Event

Stream: connects PE instances. A PE emits events in a stream and consumes events from a stream.



Integrate SAMOA to S4

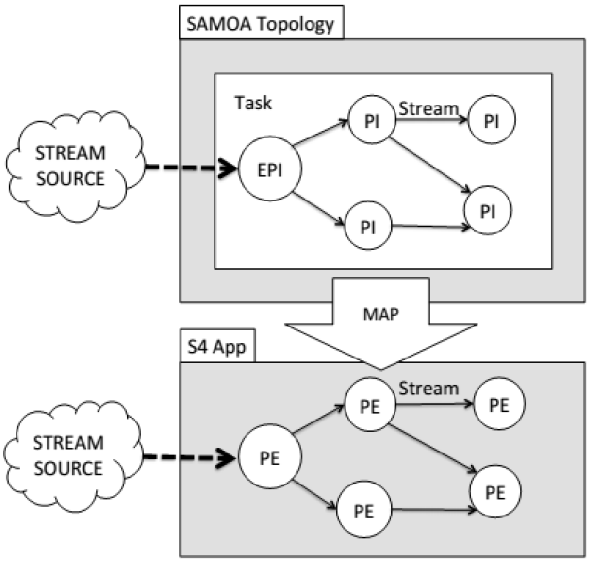
The model of processing on SAMOA is that:

1. Task(the app) is deployed to SAMOA.
2. An entree process item(EPI) read the source data “stream source”
3. The EPI send the “Events” with data to some other Process Items(PI) through stream.
4. The PI process the events received, and send some other event to other PIs.
5. At last, the output PI output the result. (to a file, database, or other place)

The figure below shows the processing model, and its corresponding trigger map to S4 when

SAMOA running on S4.

PI: process items



1.3 Current Abilities of SAMOA

Running Modes:

SAMOA now support 3 running modes:

(1) Local mode: In local mode, SAMOA can run in a single computer, without setup any SPE;

(2) Storm mode: In Storm mode, SAMOA runs on the Twitter Strom Platform.

(3) S4 mode: SAMOA runs on Yahoo S4 Platform.

When running in Storm or S4 mode, SPE must be setup and the cluster must be configured first.

Available machine learning algorithms:

Currently SAMOA have 4 algorithms, and more algorithms will be developed in the future.

(1) **Prequential Evaluation Task**:

Evaluating performance of online classifiers.

(2) **Vertical Hoeffding Tree Classifier**:

Distributed Decision Tree classifier (Vertical parallelised VFDT algorithm)

(3) **Distributed Stream Clustering**

(4) **Distributed Stream Frequent Itemset Mining**:

Base on PARMA algorithm. (The itemsets that were frequent last year may not be frequent this year. To handle this, SAMOA implements Time Biased Sampling approach.)

1.4 Advantages and disadvantages

Advantages:

- User can concentrated on Algorithms rather than the detail of a specific SPE.
- Write algorithm code once, run on all supported stream process engine.

Disadvantages:

- Still at the beginning, few documents and supports.
- Only 4 algorithms implemented so far, lack of common data mining algorithms.

2. Recommend Readings

Online Articles:

- Official SAMOA website: including general documents and a PowerPoint introduction of SAMOA.

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

- SAMOA project code official website:

https://github.com/yahoo/samoa

- A brief blog introducing SAMOA:

http://yahooeng.tumblr.com/post/65453012905/introducing-samoa-an-open-source-platform-for-mining

- Arinto Murdopo's very good blog introducing SAMOA. He is one of SAMOA's developers. His blog also including some other article talk about distributed and stream data mining algorithms:

http://www.otnira.com/2013/10/06/samoa/#more-970

And his slides: http://www.slideshare.net/arinto/final-presentation-34

- Albert Bifet's website. He is one of SAMOA's project leader, and his website includes some slides and study materials about "Big Data" and distributed data mining algorithms.

http://albertbifet.com/

Papers:

- Gianmarco De Francisci Morales. ***SAMOA: A Platform for Mining Big Data Streams***.

http://melmeric.files.wordpress.com/2013/04/samoa-a-platform-for-mining-big-data-streams.pdf

- Arinto Murdopo. ***Distributed Decision Tree Learning for Mining Big Data Streams***.

http://www.slideshare.net/arinto/emdc-thesis

- Antonio Loureiro Severien. ***Scalable Distributed Real-Time Clustering for Big Data Streams.***

http://people.ac.upc.edu/leandro/emdc/antonio-severien-thesis-updated.pdf