Ensemble :: Cloud Army

Quick Start Guide

INSTALL ENSEMBLE CLOUD ARMY AND REQUIRED SUPPORTING SOFTWARE ON LOCAL COMPUTER

Windows Users: Download the latest release from http://sourceforge.net/projects/ica . This zip file contains a windows installer that will offer to install Python and R for you if needed – ECA requires Python to run, and you'll want R installed if you plan to prototype Ensemble :: Cloud Army jobs on your local computer (e.g. as in the section below, "Run Test on Local Computer").

Linux and Mac Users: first read the "Manual Installation" section at the end of this document, then proceed from here.

SETUP WITH AMAZON WEB SERVICES

1) Go to http://aws.amazon.com and create an account with Amazon Web Services (AWS). You will need a credit card.
2) Make note of the AWS *account username* and *password* you chose.
   - These are required to access the various AWS resource and account management portals.
3) Save the AWS *access key* and AWS *secret access key* assigned to your account to a secure place on your local computer.
   - These are required when you dynamically allocate new resources in the Amazon cloud.
   - You can retrieve these again in the future by logging on to the AWS account management portal.
4) Sign up for AWS's Elastic Compute Cloud (EC2) service. Also sign up for AWS's Elastic Map Reduce (EMR) service if you intend to use Hadoop instead of MPI (Hadoop is generally more scalable, so use this if you want higher node counts).
5) Sign up for AWS's Simple Storage Service (S3).
6) Create an S3 bucket with a name of your choosing. NOTE: The S3 namespace is shared across all AWS users, so if the name you want is already taken, you'll have to choose another.
7) From your web browser, logon to the AWS Management Console (http://aws.amazon.com/console/), using your AWS account username and password.
8) Click on the "Amazon S3" tab in the Management Console and confirm your S3 bucket was created.
9) Click on the "Amazon EC2" tab in the Management Console, then the "Key Pairs" link (lower left), then the "Create Key Pair" button to create a new RSA *key pair*. Choose a name for your key pair and save the resulting `<name>.pem` file to a secure place on your computer.
   - The `.pem` file is required for communication between your local computer and any EC2 nodes you launch in the AWS cloud.

CONFIGURE ENSEMBLE :: CLOUD ARMY ON LOCAL COMPUTER

1) Go to `\InsilicosCloudArmy\ECA\src\`, make a copy of `general_config_template.json`, and name it `general_config.json`.
2) Edit general_config.json to insert the values of your personal:
   - AWS access key
   - AWS secret access key
   - RSA key pair name
   - RSA key pair file (`.pem`) location (NOTE: Make sure the file pathname uses forward slash separators ("/"), not backslash ("\"); `simplejson` expects this, even on Windows systems.)
   - S3 bucket name

RUN TEST ON LOCAL COMPUTER

Skip this section if you do not have R installed on your local computer.

This test creates an ensemble of 64 decision trees on your local computer. The compute cluster, being local, has effectively a single worker node. The ensemble is trained on and makes class predictions for the `satimage` dataset. The training and test sets for `satimage` are in directory `\InsilicosCloudArmy\ECA\data\satimage\` in compressed form.

1) Open a Windows DOS console.
2) Navigate to `\InsilicosCloudArmy\ECA\src\`.
3) Enter the following on the command line:

```
eca_launch_rmpi.py general_config.json decision_tree_satimage_job_config.json
--local
```

or

```
eca_launch_mapreduce.py general_config.json
decision_tree_satimage_job_config.json --local
```

or you can use the provided example script which runs the command for you:

```
example_satimage.py
```

4) The test run should finish in about a minute (on a 64-bit machine). Output regarding the accuracy of individual decision trees and the ensemble will appear in two places:
   - The Windows DOS console.
   - A local file in `\InsilicosCloudArmy\ECA\src\` called `decision_tree_satimage_job_config.json_runs\<timestamp>\decision_tree_satimage_job_config.json_results.txt`, where `<timestamp>` gives the date and time the run started.

RUN TEST IN AWS CLOUD

This test creates an ensemble of 64 decision trees on a cluster of 2 worker nodes in the AWS cloud. The ensemble is trained on and makes class predictions for the `satimage` dataset.

1) Open a Windows DOS console.
2) Navigate to `\InsilicosCloudArmy\ECA\src\`.
3) Enter the following on the command line:

```
eca_launch_rmpi.py general_config.json decision_tree_satimage_job_config.json
```

or, for MapReduce use:

```
eca_launch_rmpi.py general_config.json decision_tree_satimage_job_config.json
```

or you can use the provided example script which runs either command for you:

```
example_satimage.py
```

4) Using the default node type (`m1.small`; 32-bit), the test run should finish in about 7 minutes, of which about 2 minutes is the actual ensemble computation. The Windows DOS console will display the logfile from the cluster's head node, together with output regarding the accuracy of individual decision trees and the ensemble. The logfile allows you to follow the setup and shutdown of the cluster, and recognize the occasional failure of cluster setup.
5) In addition to the console, output regarding the accuracy of the individual trees and ensemble will appear in two places:
   - A local file in `\InsilicosCloudArmy\ECA\src\` called `decision_tree_satimage_job_config.json_runs\<`*`timestamp`*`>\ decision_tree_satimage_job_config.json_results.txt`, where <*timestamp*> gives the date and time the run started.
   - A file in your S3 bucket called `rmpi_computation.log`, in directory `decision_tree_satimage_job_config.json_runs\<`*`timestamp`*`>`. Use the AWS Management Console to confirm the file was created, and download to your local computer if you like.
6) ***IMPORTANT!*** Ensemble :: Cloud Army is designed to shut down the cluster after the ensemble computation is complete. Occasionally this will not happen properly, usually due to a problem with creating the cluster in the first place. You should <u>always</u> check that all nodes in the cluster were terminated. Use the AWS Management Console ("Amazon EC2" tab, then "Instances" link) to view the nodes allocated to your account. If the status of any node is other than *shutting down* or *terminated*, select that node, then choose "Terminate" from the menu of "Instance Actions". Remember, you will continue to pay Amazon for any nodes that continue to run after Ensemble :: Cloud Army is finished.

RUN BIGGER TEST IN AWS CLOUD

This test creates an ensemble of 256 decision trees on a cluster of 8 worker nodes in the AWS cloud. The ensemble is trained on and makes class predictions for the `jones` dataset [1, 2]. The training and test sets for `jones` are in directory `\InsilicosCloudArmy\ECA\data\jones\` in compressed form.

1) Open a Windows DOS console.
2) Navigate to `\InsilicosCloudArmy\ECA\src\`.
3) Enter the following on the command line:

`eca_launch_rmpi.py general_config.json decision_tree_jones_job_config.json`
or, for MapReduce, use
`eca_launch_mapreduce.py general_config.json decision_tree_jones_job_config.json`
or you can use the provided example script which runs either command for you:
`example_jones.py`

4) Using the default node type (`m1.small`; 32-bit), the test run should finish in about 15 minutes, of which about 6 minutes is the actual ensemble computation. As before, the console will display the logfile from the cluster's head node, together with output regarding the accuracy of the individual trees and ensemble.
5) In addition to the console, output regarding the accuracy of the individual trees and ensemble will appear in two places:
    a. A local file in `\InsilicosCloudArmy\ECA\src\` called `decision_tree_jones_job_config.json_runs\<timestamp>\decision_tree_jones_job_config.json_results.txt`, where `<timestamp>` gives the date and time the run started.
    b. A file in your S3 bucket called `rmpi_computation.log`, in directory `decision_tree_jones_ job_config.json_runs\<timestamp>`.
6) Follow the instructions under Step 6 in the first AWS cloud test to make sure all cluster nodes are properly terminated.

[1] Jones DT (1999). "Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices." *Journal of Molecular Biology*, **252**, 195-202.

[2] Chawla NV, Moore TE, Hall LO, Bowyer KW, Kegelmeyer WP, Springer C (2003). "Distributed Learning with Bagging-like Performance." *Pattern Recognition Letters*, **24**, 455-471.

Manual Installation

Linux and mac users should generalize the following instructions for their particular operating system. Windows users should just use the installer program as described at the start of this document.

1) Download the latest release file from http://sourceforge.net/projects/ica . This is a zip archive, extract setup_eca.sh then run it:

```
bash setup_eca.sh
```

possibly you will need to run it as

```
sudo bash setup_eca.sh
```

This will create a new directory ~/InsilicosCloudArmy and download ECA files from Sourceforge to it. The script may also install the following packages:

python
python-setuptools
python-devel
python-crypto

2) If you plan to prototype Ensemble :: Cloud Army jobs on your local computer (e.g. as in the section above, “Run Test on Local Computer”), you need to install R on your local computer (highly recommended). Version 2.13.0 (the latest at time of writing) is known to work with this package.