MR-Tandem

Quick Start Guide

for Windows

MR-Tandem is a parallelized version of the X!Tandem search engine for use with Amazon Web Services Elastic Map Reduce or your own Hadoop cluster.

INSTALL MR-TANDEM AND REQUIRED SUPPORTING SOFTWARE ON LOCAL COMPUTER

Download and run the latest windows installer (MR-Tandem_Setup_xxx.exe) from http://sourceforge.net/projects/ica .

Linux users should be able to generalize these instructions to their platform of choice - see the “Manual Installation” section at the end of this document.

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 MapReduce (EMR).
5) Sign up for AWS’s Elastic Compute Cloud (EC2).
6) Sign up for AWS’s Simple Storage Service (S3).
7) 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.
8) From your web browser, logon to the AWS Management Console (http://aws.amazon.com/console/), using your AWS account username and password.
9) Click on the “Amazon S3” tab in the Management Console and confirm your S3 bucket was created.
10) 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 MR-TANDEM ON LOCAL COMPUTER

1) Go to `\InsilicosCloudArmy\MR-Tandem`, 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 A SEARCH ON AWS EMR

You can use MR-Tandem with any parameter file that you already use with regular X!Tandem.

1) Open a Windows DOS console.
2) Enter the following on the command line:

```
mr-tandem <path_to_existing_xtandem_params_file>
```

where *path*`_to_existing_xtandem_params_file` is the name of a normal X!Tandem parameters file. "mr-tandem" is a batch file created by the windows installer that invokes the mr-tandem.py script with the general_config.json config file you set up in the previous step.

3) Your job will run, the results will go where your xtandem parameters file directs them. Logs will be written to a newly created directory named
`<path_to_existing_xtandem_params_file>_runs/<timestamp>`
4) IMPORTANT! MR-Tandem is designed to shut down the cluster after the computation is complete. Occasionally this will not happen properly, usually due to a problem with creating the cluster in the first place. You should always 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 MR-Tandem is finished.

Manual Installation

Linux users should generalize these Windows-oriented instructions for their particular operating system. Windows users should just use the installer program.

1) Download `MR-Tandem.zip` from http://sourceforge.net/projects/ica/ and extract to \InsilicosCloudArmy\MR-Tandem.

2) Install Python. Version 2.7.2 (the latest version of Python 2.x at time of writing) is known to work with this package, but earlier versions may work also. Python 3 is unsupported.
   a. Go to http://www.python.org/download and download the installer for one of the Windows binaries to your desktop (`python-2.7.2.msi` for 32-bit systems or `python-2.7.2-amd64.msi` for 64-bit systems).
   b. Double-click on the downloaded file to run install. We recommend accepting the default executable directory (`C:\Python27\`) and the defaults on customization.
3) Install needed Python modules.
   a. Open a Windows DOS console and navigate to the \InsilicosCloudArmy\MR-Tandem folder containing file `setup.py`.
   b. Enter command "`setup.py install`". You may get a message saying the setup program needs to fetch and install python module `setuptools`; allow it to proceed.
4) Unix/Linux utilities package `UnxUtils` (only needed for Windows).
   a. Go to http://sourceforge.net/projects/unxutils, download `unxutils.zip` to `C:\UnxUtils\` (or other folder of your choosing), and extract the file contents there.
   b. Add `C:\UnxUtils\bin\` and `C:\UnxUtils\usr\local\wbin\` to the Path variable in the system environment variables.