Amazon AWS

Andrew Peterson

Big Data in Finance
Baruch Master in Financial Engineering

March 8, 2016

Outline

- AWS credentials
- Hands-on: Launch instance for Spark 1.6
- Then cover setup, other basics – transferring files from s3, etc.

AWS credentials

- Key ID and Secret Key
- `.pem` file
- you are part of a secruity group with ec2, s3 access
  (If you need something else let me know)

Materials

Materials for this tutorial are at:

`https://github.com/aristotle-tek/cuny-bdif/`

Quick Start

- Start by launching ec2, since this takes a few minutes.
- Launch scripts that automate master-slave setup, etc:
- In spark installation or
  `git clone` <my cuny-bdif files> / AWS / ec2
- Can follow script `setup.sh`

- more info:
  `http://spark.apache.org/docs/latest/ec2-scripts.html`

Launching Spark on ec2

```
./spark-ec2 --key-pair=smallhands \\
--identity-file=/<filepath>/smallhands.pem -t m3.large \\
--ebs-vol-size 60 --region=us-east-1 launch my-spark-cluster
```

- `ebs-vol-size` is an additional drive that can be preserved while shutting down compute
- NB: keeping the ec2 in the same region as s3 prevents data transfer charges

› with that launching, let's back up a little...

AWS

- web interface & command line tools
- ec2 (compute)

```
m3.medium 1CPU,  3.75MB RAM, 1x4 SSD $0.067/hour
m3.large 2CPUs,  7.5MB RAM, 1x32 SSD $0.133/hour
```

- s3 storage (bucket: `s3://bdif-tweets`)
- many other components...

Transferring Files

- Transfer to ec2: Can use scp like the HPC or command line tools
- Transfer to/ from s3 need Amazon command line tools

AWS Command line tools

- `apt-get install awscli` or `pip install awsli`
- `aws configure`
  then enter key ID and secret key.
- `aws s3 cp <fromfile> <tofile>`

AWS Command line tools

```
curl -O https://bootstrap.pypa.io/get-pip.py
sudo python27 get-pip.py
pip install awscli
# Need ~/bin to be in PATH var for the symlink to work:
# See if $PATH contains ~/bin (if not, null output)
echo $PATH | grep ~/bin
export PATH=~/bin:$PATH
```

Mount the ebs volume

```
sudo mkdir ebsvol
sudo mount /dev/xvds ebsvol
```

Get data from s3

```
cd ebsvol
aws s3 cp s3://bdif-tweets/sample/sampletweets.tar sample.tar
tar -xvf sample.tar
```

Bunzip2

For the full original data, to find all files ending in `.json.bz2` and unzip them, can use find & xargs:

```
find . -name "*.json.bz2" -printf '%P\n' |  xargs bunzip2
```

Useful Additions: Tmux

```
sudo yum install tmux
```

- `tmux` – launch tmux
- `tmux a` – attach existing session (e.g. after disconnect)
- `cntl + b, c` – create new window
- `cntl + b, n` – next window (or `cntl + #` for number)
- `cntl + b, [` – scroll mode, then `q` to quit
- `cntl + b, x` – kill new window

Stopping the Instance

Note that when you log off (e.g. `exit`) the instance is still running, and the college is still paying for it!

You can stop the instance, keeping the ebs store active, with:

```
./spark-ec2 --region=us-east-1 stop my-spark-cluster
```

Then you can re-start it with:

```
./spark-ec2 -i <privatekey>.pem --region=us-east-1 start my-spark-cluster
```

Terminating the Instance

To fully terminate the instance and delete the ebs volume, use:

```
./spark-ec2 --region=us-east-1 destroy my-spark-cluster
```