More about AWS cluster manipulation
&
Data/Experiment generation

Gang Luo
Oct. 5, 2010

Hadoop on AWS Cluster

- Run Hadoop instance on the AWS cluster
  - Manual configuration required if you start cluster from web console
  - For convenience, use the harness tools
- Don't copy large files from/to AWS cluster
  - Will be charged for this
  - Use data generator we provided to get datasets *
- Automate the experiments
  - You can run them manually, exactly the same as you did on hadoop21.cs.duke.edu
  - Better generate experiments and run them by one command *

see details at http://www.cs.duke.edu/courses/fall10/cps216/TA_Material/data_expr_instructions.pdf

4. generate datasets
5. generate experiments
6. run experiment in batch
7. ensure you get all your result or runtime information before you shut down your cluster

1. environment setting
2. launch a hadoop cluster on AWS
3. copy the necessary tools (e.g. data generator) to AWS cluster
8. shut down your cluster

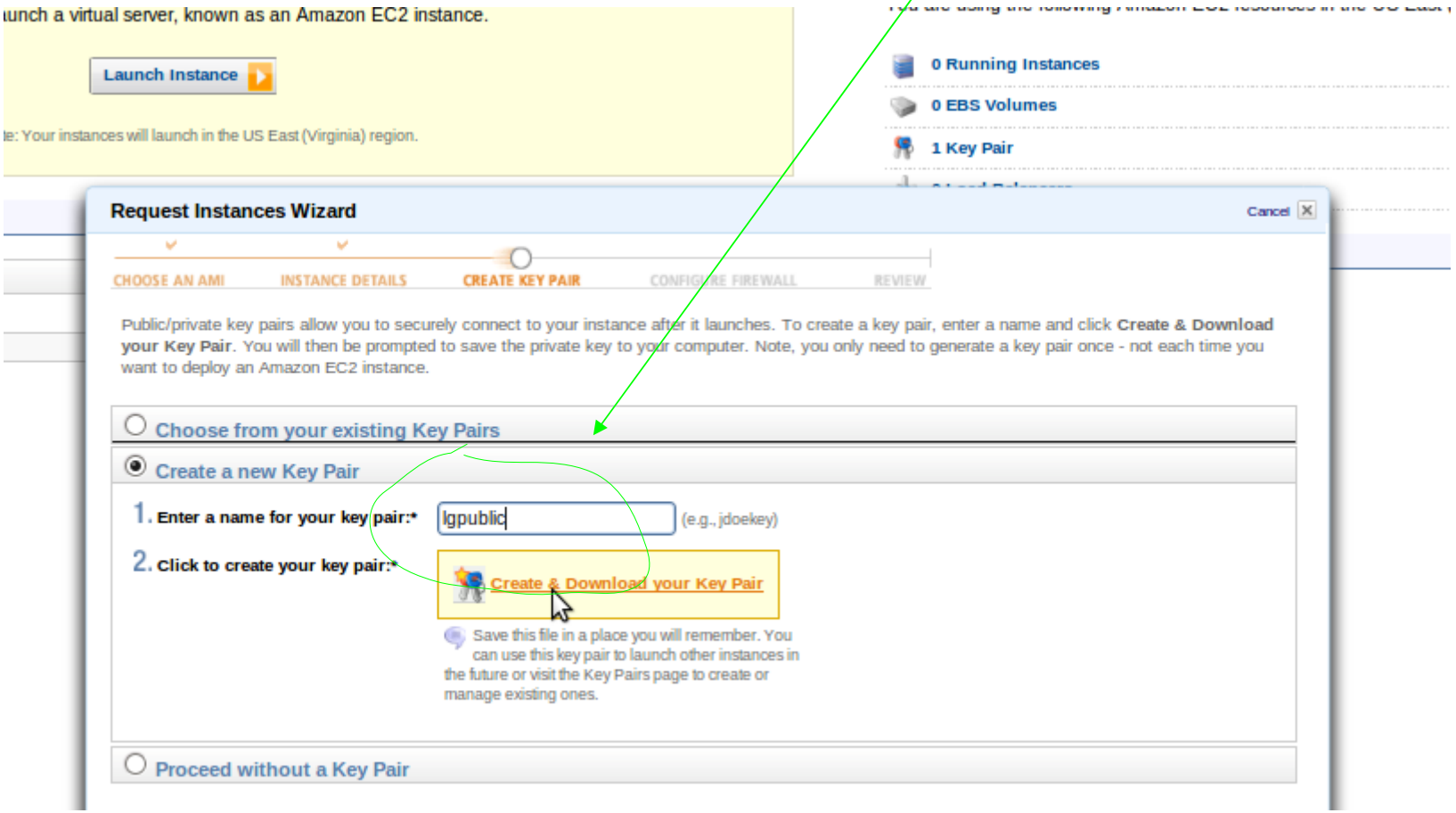
AWS hadoop cluster

linux.cs.duke.edu or your local machine

harness/hadoop_ec2_contrib_bin/hadoop-ec2-env.sh

```
# The EC2 key name used to launch instances. Change it as needed.
KEY_NAME=lgpublic
# Ned's convention:
#KEY_NAME="${EC2_KEYPAIR_NAME}"

# Where your EC2 private key is stored (created, for example, when following the
#    Amazon Getting Started guide).
#PRIVATE_KEY_PATH=`echo "${EC2_KEYDIR}"/"id_rsa-${KEY_NAME}"`
PRIVATE_KEY_PATH="/home/lgpublic/Documents/AWS/keypair/lgpublic.pem"
# Ned's convention:
```



Data Generation

- Follow the instructions at our course website
- Entire dataset contains 8 tables*. lineitem.tbl is the largest one (take 80% of the total size). You can use that.
- Make the number of piece for each table large enough (larger than number of slave nodes) to accelerate the process
  - perl gen_data.pl scale_factor num_files zipf_factor host_list local_dir hdfs_dir
- May takes few minutes to hours depending on the size of your data and cluster

* the schema of all the tables could be found here:
http://www.tpc.org/tpch/spec/tpch2.12.0.doc

Experiment Generation

- Follow the instructions at our course website
- Generate configuration.xml for each experiment specifying the values for some parameters
- Generate script to run this experiment, where the configuration.xml will appear in the command.
- A global run.sh will call each of the scripts and run all the experiments.
- For those parameters assigned a value in the program, they will not change by the external configuration file

Configuration file support

```
public class MyMapReduce extends Configured implements Tool{
    public int run(String[] args) throws Exception {
     JobConf conf = new JobConf(getConf(), MyMapReduce.class);

      ...

     JobClient.runJob(conf);
    }

    public static void main(String[] args) throws Exception{

     int res=ToolRunner.run(new Configuration(), new MyMapReduce(),
    args);
     System.exit(res);
    }

}
```