0. Warm-up

You are supposed to created your own cluster on Amazon AWS here. For more information about environment configuration and cluster manipulation through command lines, please refer to the previous instructions.

After launching your cluster, you need to copy necessary tools/files to the cluster. Follow the steps below.

0.1 Go to the harness home directory by

```
cd ${AWS_HADOOP_HARNESS_HOME}/..
```

0.2 pack the directory which contains necessary tools/information. (e.g. data generator, all the slaves name)

```
tar -zcvf AWS_HADOOP_HARNESS.tar.gz aws_hadoop_harness
```

0.3 Copy this packed file to your cluster

```
${HADOOP_EC2_HOME}/hadoop-ec2 push YOUR_CLUSTER_NAME ${AWS_HADOOP_HARNESS_HOME}/../AWS_HADOOP_HARNESS.tar.gz
```

Remember to replace YOUR_CLUSTER_NAME with the name you are using when launch the cluster. In the example in the previous instructions, the name of the cluster is “test-hadoop-cluster”

0.4 Login to the cluster

```
${HADOOP_EC2_HOME}/hadoop-ec2 login  YOUR_CLUSTER_NAME
```

Remember to replace YOUR_CLUSTER_NAME with your cluster name.

0.5 unpack the file

```
tar -zxvf AWS_HADOOP_HARNESS.tar.gz
```

You should find a directory named “aws_hadoop_harness” generated. All you need are there.

1. Data Generation

We need data before we start doing anything! The AMI we are using contains all the software/tools we may use. But there is no data that we can play with. So, the first step is use the tools to generate datasets we will work on later.

We are using TPC-H datasets. This datasets include 8 tables. When we generate the TPC-H datasets, we will get all the 8 tables. To generate the datasets, we should go to aws_hadoop_harness/data_gen, and use this this command there:

```
perl gen_data.pl scale_factor num_files zipf_factor host_list local_dir hdfs_dir
```

where:

scale_factor = TPCH Scale factor (GB of data to generate)
num_files = The number of files to generate for each table
zipf_factor = Zipfian distribution factor (0-4, 0 means uniform)
host_list = File containing a list of host machines
local_dir = Local directory to use in the host machines
hdfs_dir = HDFS directory to store the generated data

An example for this command would be

```
perl gen_data.pl 20 10 2 SLAVE_NAMES.txt /root/tpch_data /usr/root/dataset/in
```

where I generate the datasets of 20 GB totally, each of the table will be splited into 10 pieces. The skew level is 2. Those datasets will be first generated to local disk at /root/tpch_data, and later copied to HDFS at /usr/root/dataset/in

2. Experiment Generation

After generating the data, you can start run Hadoop jobs (Of course, your jar file should be uploaded first). But for experiments with different configurations, we provide you the tools to generate experiments and run them automatically. All you need to do is specify the parameters you want to in an xml file. You will get all the configurations which are actually the cross products of those parameters, and each job will run on one of these configurations. You can go to "aws_hadoop_harness/sample_configs" directory to see some example configuration files.

Assume you specify the parameters in file at "aws_hadoop_harness/sample_configs/my_conf.xml" , you can go to "aws_hadoop_harness" directory and generate experiments by:

```
perl gen_exper.pl sample_configs/my_conf.xml DIR
```

where the first parameter is the path to the xml file, and the directory DIR is where the generated experiments be placed.

3. Run Experiment in batch

Assume you place the generated experiments in directory "aws_hadoop_harness/expr", run all the experiments sequentially by:

```
./run_exper.sh aws_hadoop_harness/expr
```

Now the jobs will run one by one.