Working with EMR and HBase


1. Create HBase workflow on amazon EC2; Hadoop is added by default; Hive + Pig can be added optionally
2. SSH into the HBase instance with key-pair created earlier (hadoop@ec2...)
3. Hadoop + Hbase are already started;
4. Hbase shell command opens up Hbase interactive shell

   >hbase shell

   hbase(main):001:0> create 't1', 'f1'

   hbase(main):001:0>put 't1', 'r1', 'f1', 'v1'

   hbase(main):001:0>get 't1', 'r1'

   hbase(main):001:0>scan 't1'

5. Insert several values with f1:q1 and scan the table to see versions of same data.

   hbase(main):001:0>put 't1', 'r1', 'f1:q1', 'v1'

   Repeat this about “put” with different values of data (“v1”) field. Then “scan” the table for various versions of data. Scan without specifying versions gives only the latest value.

   scan 't1', { VERSIONS => 2}

   Observe that only column family “schema” needs to be provided at the time of table creations. Columns can be added on the fly during run-time.

6. Using scripts: The Shell is based on JRuby, you can mix Ruby commands with Hbase commands. Add some synthetic data

   hbase(main):064:0> for i in 'a'..'z' do for j in 'a'..'z' do put 't1','r#{i}#{j}','f1:#{j}',"#{j}" end end

   scan 't1'

   hbase(main):073:1> for i in 'a'..'z' do for j in 'a'..'z' do x = rand(6);put 't1','r#{i}#{j}','f1:#{j}',"#{j}" end end

   Remember all these is done programmatically (Java/Python).