HBASE – PIG INTEGRATION

By.Mr.Gopal Krishna
Sr.Hadoop Technical Architect
CCA 175-Spark and Hadoop Developer Certified Consultant

TO WRITE THE DATA FROM PIG TO HBASE

FROM HBASE SIDE

++

VERSIONS USED:
PIG – pig-0.15.0
HBASE – hbase-0.98.4-hadoop2
HADOOP-2.6.0

hbase(main):009:0> create 'hbasepig_emptable','empdetails'
0 row(s) in 0.5520 seconds

=> Hbase::Table - hbasepig_emptable
hbase(main):010:0> scan 'hbasepig_emptable'
ROW   COLUMN+CELL
0 row(s) in 0.0470 seconds

hbase(main):011:0>

FROM PIG SIDE

++
INPUT DATA:

gopalkrishna@ubuntu:~/PRAC/HBASE$ cat HBase_Pig_InputData.log
```
1000    RAVI    VARMA   9       TECHNOLOGY LEAD
1001    KOUSIK  MISTRA  11      ENTRYLEVEL ARCHITECT
1002    SAIED   FAISAL  21      DELIVERY HEAD
1003    SOURAV  JAIN    18      SENIOUR PROJECT MANAGER
1004    CHRIS   WAYNE   15      TEST MANAGER
1005    ANEES   FLEXWALA        19      CONFIGURATION CONTROLLER
1006    GOPAL   KRISHNA         13      SR.BIGDATA ARCHITECT
1007    RAMYA   SRI     8       BIGDATA LEAD
1008    MUNI    KRISHNA         5       SR.HADOOP DEVELOPER
1009    RAMYA   SRI     7       SR.SPARK DEVELOPER
1010    RAM     CHANDAR         16      PROJECT MANAGER
1011    AKHILA  PRIYA   3       HADOOP DEVELOPER
```

```
1012    RAHUL   DEV     6       HADOOP ADMINISTRATOR
1013    AMIT    MISRA   4       BIGDATA TESTER
1014    PUJA    YERUNKAR        11      BIGDATA TEST LEAD
```
gopalkrishna@ubuntu:~/PRAC/HBASE$

gopalkrishna@ubuntu:~/PRAC/HBASE$ cat HBase_Pig_Script.pig

```
REGISTER /home/gopalkrishna/INSTALL/hbase-0.98.4-hadoop2/lib/zookeeper-3.4.6.jar;
REGISTER /home/gopalkrishna/INSTALL/hbase-0.98.4-hadoop2/lib/guava-12.0.1.jar;
REGISTER /home/gopalkrishna/INSTALL/hbase-0.98.4-hadoop2/lib/protobuf-java-2.5.0.jar;
REGISTER /home/gopalkrishna/INSTALL/hbase-0.98.4-hadoop2/lib/hbase-common-0.98.4-hadoop2.jar;
REGISTER /home/gopalkrishna/INSTALL/hbase-0.98.4-hadoop2/lib/hbase-client-0.98.4-hadoop2.jar;
REGISTER /home/gopalkrishna/INSTALL/hbase-0.98.4-hadoop2/lib/hbase-server-0.98.4-hadoop2.jar;

iniData = LOAD '/HBase_Pig_Data/HBase_Pig_InputData.log' USING PigStorage('\t') as
    (empid:int,firstname:chararray,lastname:chararray,yrsofexp:int,desg:chararray);
STORE iniData INTO 'hbase://hbasepig_emptable' USING org.apache.pig.backend.hadoop.hbase.HBaseStorage(
 'empdetails:firstname,empdetails:lastname,empdetails:yrsofexp,empdetails:desg');
```

TO RUN PIG SCRIPT:

gopalkrishna@ubuntu:~/PRAC/HBASE$ pig HBase_Pig_Script.pig

VERIFICATION FROM HBASE SIDE

+++

AFTER RUNNING THE PIG SCRIPT => GO TO HBASE SHELL AND AGAIN RUN THE
SCAN COMMAND ON THE HBASE TABLE LIKE BELOW

hbase(main):011:0> scan 'hbasepig_emptable'

```
ROW             COLUMN+CELL
1000            column=empdetails:desg, timestamp=1505304875777, value=TECHNOLOGY LEAD
1000            column=empdetails:firstname, timestamp=1505304875777, value=RAVI
1000            column=empdetails:lastname, timestamp=1505304875777, value=VARMA
1000            column=empdetails:yrsofexp, timestamp=1505304875777, value=9
1001            column=empdetails:desg, timestamp=1505304875805, value=ENTRYLEVEL ARCHITECT
1001            column=empdetails:firstname, timestamp=1505304875805, value=KOUSIK
1001            column=empdetails:lastname, timestamp=1505304875805, value=MISTRA
1001            column=empdetails:yrsofexp, timestamp=1505304875805, value=11
1002            column=empdetails:desg, timestamp=1505304875806, value=DELIVERY HEAD
1002            column=empdetails:firstname, timestamp=1505304875806, value=SAIED
1002            column=empdetails:lastname, timestamp=1505304875806, value=FAISAL
1002            column=empdetails:yrsofexp, timestamp=1505304875806, value=21
1003            column=empdetails:desg, timestamp=1505304875807, value=SENIOUR PROJECT MANAGER
1003            column=empdetails:firstname, timestamp=1505304875807, value=SOURAV
1003            column=empdetails:lastname, timestamp=1505304875807, value=JAIN
1003            column=empdetails:yrsofexp, timestamp=1505304875807, value=18
1004            column=empdetails:desg, timestamp=1505304875808, value=TEST MANAGER
1004            column=empdetails:firstname, timestamp=1505304875808, value=CHRIS
1004            column=empdetails:lastname, timestamp=1505304875808, value=WAYNE
1004            column=empdetails:yrsofexp, timestamp=1505304875808, value=15
1005            column=empdetails:desg, timestamp=1505304875809, value=CONFIGURATION CONTROLLER
1005            column=empdetails:firstname, timestamp=1505304875809, value=ANEES
1005            column=empdetails:lastname, timestamp=1505304875809, value=FLEXWALA
1005            column=empdetails:yrsofexp, timestamp=1505304875809, value=19
```

```
1006              column=empdetails:desg, timestamp=1505304875809, value=SR.BIGDATA ARCHITECT
1006              column=empdetails:firstname, timestamp=1505304875809, value=GOPAL
1006              column=empdetails:lastname, timestamp=1505304875809, value=KRISHNA
1006              column=empdetails:yrsofexp, timestamp=1505304875809, value=13
1007              column=empdetails:desg, timestamp=1505304875810, value=BIGDATA LEAD
1007              column=empdetails:firstname, timestamp=1505304875810, value=RAMYA
1007              column=empdetails:lastname, timestamp=1505304875810, value=SRI
1007              column=empdetails:yrsofexp, timestamp=1505304875810, value=8
1008              column=empdetails:desg, timestamp=1505304875811, value=SR.HADOOP DEVELOPER
1008              column=empdetails:firstname, timestamp=1505304875811, value=MUNI
1008              column=empdetails:lastname, timestamp=1505304875811, value=KRISHNA
1008              column=empdetails:yrsofexp, timestamp=1505304875811, value=5
1009              column=empdetails:desg, timestamp=1505304875811, value=SR.SPARK DEVELOPER
1009              column=empdetails:firstname, timestamp=1505304875811, value=RAMYA
1009              column=empdetails:lastname, timestamp=1505304875811, value=SRI
1009              column=empdetails:yrsofexp, timestamp=1505304875811, value=7
1010              column=empdetails:desg, timestamp=1505304875812, value=PROJECT MANAGER
1010              column=empdetails:firstname, timestamp=1505304875812, value=RAM
1010              column=empdetails:lastname, timestamp=1505304875812, value=CHANDAR
1010              column=empdetails:yrsofexp, timestamp=1505304875812, value=16
1011              column=empdetails:desg, timestamp=1505304875813, value=HADOOP DEVELOPER
1011              column=empdetails:firstname, timestamp=1505304875813, value=AKHILA
1011              column=empdetails:lastname, timestamp=1505304875813, value=PRIYA
1011              column=empdetails:yrsofexp, timestamp=1505304875813, value=3
1012              column=empdetails:desg, timestamp=1505304875813, value=HADOOP ADMINISTRATOR
1012              column=empdetails:firstname, timestamp=1505304875813, value=RAHUL
1012              column=empdetails:lastname, timestamp=1505304875813, value=DEV
1012              column=empdetails:yrsofexp, timestamp=1505304875813, value=6
1013              column=empdetails:desg, timestamp=1505304875814, value=BIGDATA TESTER
1013              column=empdetails:firstname, timestamp=1505304875814, value=AMIT
1013              column=empdetails:lastname, timestamp=1505304875814, value=MISRA
1013              column=empdetails:yrsofexp, timestamp=1505304875814, value=4
1014              column=empdetails:desg, timestamp=1505304875815, value=BIGDATA TEST LEAD
1014              column=empdetails:firstname, timestamp=1505304875815, value=PUJA
1014              column=empdetails:lastname, timestamp=1505304875815, value=YERUNKAR
1014              column=empdetails:yrsofexp, timestamp=1505304875815, value=11
15 row(s) in 0.6160 seconds
```

··

TO READ THE DATA FROM HBASE TO PIG

FROM PIG SIDE

++

PIG SCRIPT TO READ THE DATA FROM HBASE TABLE

gopalkrishna@ubuntu:~/PRAC/HBASE$ cat HBase_Pig_Script.pig

```
REGISTER /home/gopalkrishna/INSTALL/hbase-0.98.4-hadoop2/lib/zookeeper-3.4.6.jar;
REGISTER /home/gopalkrishna/INSTALL/hbase-0.98.4-hadoop2/lib/guava-12.0.1.jar;
REGISTER /home/gopalkrishna/INSTALL/hbase-0.98.4-hadoop2/lib/protobuf-java-2.5.0.jar;
REGISTER /home/gopalkrishna/INSTALL/hbase-0.98.4-hadoop2/lib/hbase-common-0.98.4-hadoop2.jar;
REGISTER /home/gopalkrishna/INSTALL/hbase-0.98.4-hadoop2/lib/hbase-client-0.98.4-hadoop2.jar;
REGISTER /home/gopalkrishna/INSTALL/hbase-0.98.4-hadoop2/lib/hbase-server-0.98.4-hadoop2.jar;

--iniData = LOAD '/HBase_Pig_Data/HBase_Pig_InputData.log' USING PigStorage('\t') as
--     (empid:int,firstname:chararray,lastname:chararray,yrsofexp:int,desg:chararray);
--STORE iniData INTO 'hbase://hbasepig_emptable' USING org.apache.pig.backend.hadoop.hbase.HBaseStorage(
--  'empdetails:firstname,empdetails:lastname,empdetails:yrsofexp,empdetails:desg');
A = LOAD 'hbase://hbasepig_emptable'
```

using
org.apache.pig.backend.hadoop.hbase.HBaseStorage('empdetails:firstname,empdetails:lastname,empdetails:yrsof
exp,empdetails:desg', '-loadKey true')
  as (rowkey:bytearray, firstname:chararray,lastname:chararray,yrsofexp:int,desg:chararray);
DUMP A;
gopalkrishna@ubuntu:~/PRAC/HBASE$

TO RUN PIG SCRIPT:

gopalkrishna@ubuntu:~/PRAC/HBASE$ pig HBase_Pig_Script.pig

**AFTER RUNNING THE PIG SCRIPT => VERIFY THE RESULT ON THE CONSOLE , ALL
THE RECORDS IN HBASE TABLE(hbasepig_emptable) WILL BE DISPLAYED.**

Input(s):
Successfully read 15 records (337 bytes) from: "hbase://hbasepig_emptable"

Output(s):
Successfully stored 15 records (777 bytes) in: "hdfs://localhost:8020/tmp/temp645148187/tmp-1399112497"

Counters:
Total records written : 15
Total bytes written : 777
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1505297080421_0008

2017-09-13 06:27:24,563 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at
localhost/127.0.0.1:8032
2017-09-13 06:27:24,597 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is
completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-09-13 06:27:24,684 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at
localhost/127.0.0.1:8032
2017-09-13 06:27:24,705 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is
completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-09-13 06:27:24,809 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at
localhost/127.0.0.1:8032
2017-09-13 06:27:24,830 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is
completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-09-13 06:27:24,987 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-09-13 06:27:24,993 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
deprecated. Instead, use fs.defaultFS
2017-09-13 06:27:25,001 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not
set... will not generate code.
2017-09-13 06:27:25,095 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to
process : 1
2017-09-13 06:27:25,107 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total
input paths to process : 1
(1000,RAVI,VARMA,9,TECHNOLOGY LEAD)
(1001,KOUSIK,MISTRA,11,ENTRYLEVEL ARCHITECT)
(1002,SAIED,FAISAL,21,DELIVERY HEAD)
(1003,SOURAV,JAIN,18,SENIOUR PROJECT MANAGER)
(1004,CHRIS,WAYNE,15,TEST MANAGER)
(1005,ANEES,FLEXWALA,19,CONFIGURATION CONTROLLER)
(1006,GOPAL,KRISHNA,13,SR.BIGDATA ARCHITECT)
(1007,RAMYA,SRI,8,BIGDATA LEAD)
(1008,MUNI,KRISHNA,5,SR.HADOOP DEVELOPER)
(1009,RAMYA,SRI,7,SR.SPARK DEVELOPER)
(1010,RAM,CHANDAR,16,PROJECT MANAGER)

(1011,AKHILA,PRIYA,3,HADOOP DEVELOPER)
(1012,RAHUL,DEV,6,HADOOP ADMINISTRATOR)
(1013,AMIT,MISRA,4,BIGDATA TESTER)
(1014,PUJA,YERUNKAR,11,BIGDATA TEST LEAD)
2017-09-13 06:27:25,350 [main] INFO org.apache.pig.Main - Pig script completed in 47 seconds and 647 milliseconds (47647 ms)

gopalkrishna@ubuntu:~/PRAC/HBASE$

ADD 1 MORE RECORD IN HBASE TABLE => AND RUN THE PIG SCRIPT AGAIN , WE WILL FIND THE NEW RECORD IN PIG SIDE

hbase(main):002:0> put 'hbasepig_emptable','1100','empdetails:firstname','RAJESH'
0 row(s) in 0.0900 seconds

hbase(main):003:0> put 'hbasepig_emptable','1100','empdetails:lastname','KANNA'
0 row(s) in 0.0120 seconds

hbase(main):004:0> put 'hbasepig_emptable','1100','empdetails:yrsofexp',14
0 row(s) in 0.0080 seconds

hbase(main):005:0> put 'hbasepig_emptable','1100','empdetails:desg','SOLUTION ARCHITECT'
0 row(s) in 0.0130 seconds

hbase(main):006:0>
+++

FROM PIG SIDE – AFTER RUNNING THE SCRIPT AGAIN, WE WILL GET BELOW RESULTS:

Input(s):
Successfully read 16 records (337 bytes) from: "hbase://hbasepig_emptable"

Output(s):
Successfully stored 16 records (777 bytes) in: "hdfs://localhost:8020/tmp/temp645148187/tmp-1399112497"

Counters:
Total records written : 16
Total bytes written : 777
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1505297080421_0008

2017-09-13 06:27:24,563 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2017-09-13 06:27:24,597 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-09-13 06:27:24,684 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2017-09-13 06:27:24,705 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-09-13 06:27:24,809 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2017-09-13 06:27:24,830 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-09-13 06:27:24,987 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-09-13 06:27:24,993 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-09-13 06:27:25,001 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.

2017-09-13 06:27:25,095 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-09-13 06:27:25,107 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1000,RAVI,VARMA,9,TECHNOLOGY LEAD)
(1001,KOUSIK,MISTRA,11,ENTRYLEVEL ARCHITECT)
(1002,SAIED,FAISAL,21,DELIVERY HEAD)
(1003,SOURAV,JAIN,18,SENIOUR PROJECT MANAGER)
(1004,CHRIS,WAYNE,15,TEST MANAGER)
(1005,ANEES,FLEXWALA,19,CONFIGURATION CONTROLLER)
(1006,GOPAL,KRISHNA,13,SR.BIGDATA ARCHITECT)
(1007,RAMYA,SRI,8,BIGDATA LEAD)
(1008,MUNI,KRISHNA,5,SR.HADOOP DEVELOPER)
(1009,RAMYA,SRI,7,SR.SPARK DEVELOPER)
(1010,RAM,CHANDAR,16,PROJECT MANAGER)
(1011,AKHILA,PRIYA,3,HADOOP DEVELOPER)
(1012,RAHUL,DEV,6,HADOOP ADMINISTRATOR)
(1013,AMIT,MISRA,4,BIGDATA TESTER)
(1014,PUJA,YERUNKAR,11,BIGDATA TEST LEAD)
(1100,RAJESH,KANNA,14,SOLUTION ARCHITECT)
2017-09-13 06:27:25,350 [main] INFO org.apache.pig.Main - Pig script completed in 47 seconds and 647 milliseconds (47647 ms)
gopalkrishna@ubuntu:~/PRAC/HBASE$