Joins in HIVE

1. Inner join
2. Left outer
3. Right outer
4. Full outer All records from bioth tables and matching record will get printed only once

hive (niit)> create table employee(name string, salary float,city string)

> row format delimited

> fields terminated by ',';

OK

Time taken: 0.047 seconds

hive (niit)> load data local inpath '/home/hduser/emp.txt' into table employee;

Loading data to table niit.employee

Table niit.employee stats: [numFiles=1, totalSize=89]

OK

Time taken: 0.081 seconds

hive (niit)> select \* from employee;

OK

employee.name employee.salary employee.city

swetha 250000.0 Chennai

anamika 200000.0 Kanyakumari

tarun 300000.0 Pondi

anita 250000.0 Selam

NULL NULL

NULL NULL

Time taken: 0.061 seconds, Fetched: 6 row(s)

hive (niit)> create table mailid (name string, email string)

> row format delimited

> fields terminated by ',';

OK

Time taken: 0.032 seconds

hive (niit)> load data local inpath '/home/hduser/email.txt' into table mailid;

Loading data to table niit.mailid

Table niit.mailid stats: [numFiles=1, totalSize=97]

OK

Time taken: 0.081 seconds

hive (niit)> select \* from mailid;

OK

mailid.name mailid.email

swetha swetha@gmail.com

tarun tarun@edureka.in

nagesh nagesh@yahoo.com

venkatesh venki@gmail.com

inner join

hive (niit)> select a.name,a.city,a.salary,b.email from

> employee a join mailid b on a.name = b.name;

Query ID = hduser\_20171224145425\_79d9f19c-38ec-4fa0-a9bd-70074b1baf4c

Total jobs = 1

17/12/24 14:54:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Execution log at: /tmp/hduser/hduser\_20171224145425\_79d9f19c-38ec-4fa0-a9bd-70074b1baf4c.log

2017-12-24 14:54:27 Starting to launch local task to process map join; maximum memory = 477626368

2017-12-24 14:54:27 Dump the side-table for tag: 0 with group count: 5 into file: file:/usr/local/hive/iotmp/8c68e5a2-4bec-422a-a153-d6eca81e2e2c/hive\_2017-12-24\_14-54-25\_257\_6391777356404955959-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile30--.hashtable

2017-12-24 14:54:27 Uploaded 1 File to: file:/usr/local/hive/iotmp/8c68e5a2-4bec-422a-a153-d6eca81e2e2c/hive\_2017-12-24\_14-54-25\_257\_6391777356404955959-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile30--.hashtable (431 bytes)

2017-12-24 14:54:27 End of local task; Time Taken: 0.777 sec.

Execution completed successfully

MapredLocal task succeeded

Launching Job 1 out of 1

Number of reduce tasks is set to 0 since there's no reduce operator

Starting Job = job\_1514094803278\_0016, Tracking URL = http://rootuser:8088/proxy/application\_1514094803278\_0016/

Kill Command = /usr/local/hadoop/bin/hadoop job -kill job\_1514094803278\_0016

Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0

2017-12-24 14:54:31,532 Stage-3 map = 0%, reduce = 0%

2017-12-24 14:54:35,692 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 2.19 sec

MapReduce Total cumulative CPU time: 2 seconds 190 msec

Ended Job = job\_1514094803278\_0016

MapReduce Jobs Launched:

Stage-Stage-3: Map: 1 Cumulative CPU: 2.19 sec HDFS Read: 6492 HDFS Write: 79 SUCCESS

Total MapReduce CPU Time Spent: 2 seconds 190 msec

OK

a.name a.city a.salary b.email

swetha Chennai 250000.0 swetha@gmail.com

tarun Pondi 300000.0 tarun@edureka.in

Time taken: 12.509 seconds, Fetched: 2 row(s)

left outer join

hive (niit)> select \* from employee;

OK

employee.name employee.salary employee.city

swetha 250000.0 Chennai

anamika 200000.0 Kanyakumari

tarun 300000.0 Pondi

anita 250000.0 Selam

NULL NULL

NULL NULL

Time taken: 0.036 seconds, Fetched: 6 row(s)

hive (niit)> select \* from mailid;

OK

mailid.name mailid.email

swetha swetha@gmail.com

tarun tarun@edureka.in

nagesh nagesh@yahoo.com

venkatesh venki@gmail.com

Time taken: 0.033 seconds, Fetched: 4 row(s)

hive (niit)> select a.name,a.city,a.salary,b.email from

> employee a left outer join mailid b on a.name = b.name;

Query ID = hduser\_20171224145611\_07878552-4a35-4db1-ac59-d421d4527640

Total jobs = 1

17/12/24 14:56:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Execution log at: /tmp/hduser/hduser\_20171224145611\_07878552-4a35-4db1-ac59-d421d4527640.log

2017-12-24 14:56:12 Starting to launch local task to process map join; maximum memory = 477626368

2017-12-24 14:56:13 Dump the side-table for tag: 1 with group count: 4 into file: file:/usr/local/hive/iotmp/8c68e5a2-4bec-422a-a153-d6eca81e2e2c/hive\_2017-12-24\_14-56-11\_101\_6806483037933427412-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile41--.hashtable

2017-12-24 14:56:13 Uploaded 1 File to: file:/usr/local/hive/iotmp/8c68e5a2-4bec-422a-a153-d6eca81e2e2c/hive\_2017-12-24\_14-56-11\_101\_6806483037933427412-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile41--.hashtable (429 bytes)

2017-12-24 14:56:13 End of local task; Time Taken: 0.579 sec.

Execution completed successfully

MapredLocal task succeeded

Launching Job 1 out of 1

Number of reduce tasks is set to 0 since there's no reduce operator

Starting Job = job\_1514094803278\_0017, Tracking URL = http://rootuser:8088/proxy/application\_1514094803278\_0017/

Kill Command = /usr/local/hadoop/bin/hadoop job -kill job\_1514094803278\_0017

Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0

2017-12-24 14:56:18,421 Stage-3 map = 0%, reduce = 0%

2017-12-24 14:56:22,527 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 1.52 sec

MapReduce Total cumulative CPU time: 1 seconds 520 msec

Ended Job = job\_1514094803278\_0017

MapReduce Jobs Launched:

Stage-Stage-3: Map: 1 Cumulative CPU: 1.52 sec HDFS Read: 6471 HDFS Write: 155 SUCCESS

Total MapReduce CPU Time Spent: 1 seconds 520 msec

OK

a.name a.city a.salary b.email

swetha Chennai 250000.0 swetha@gmail.com

anamika Kanyakumari 200000.0 NULL

tarun Pondi 300000.0 tarun@edureka.in

anita Selam 250000.0 NULL

NULL NULL NULL

NULL NULL NULL

Time taken: 12.498 seconds, Fetched: 6 row(s)

right outer

hive (niit)> select b.name,a.city,a.salary,b.email from

> employee a right outer join mailid b on a.name = b.name;

Query ID = hduser\_20171224145626\_3cc1c040-a7c6-48bd-844f-43fc1f408052

Total jobs = 1

17/12/24 14:56:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Execution log at: /tmp/hduser/hduser\_20171224145626\_3cc1c040-a7c6-48bd-844f-43fc1f408052.log

2017-12-24 14:56:28 Starting to launch local task to process map join; maximum memory = 477626368

2017-12-24 14:56:28 Dump the side-table for tag: 0 with group count: 5 into file: file:/usr/local/hive/iotmp/8c68e5a2-4bec-422a-a153-d6eca81e2e2c/hive\_2017-12-24\_14-56-26\_268\_2427948028839272408-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile50--.hashtable

2017-12-24 14:56:28 Uploaded 1 File to: file:/usr/local/hive/iotmp/8c68e5a2-4bec-422a-a153-d6eca81e2e2c/hive\_2017-12-24\_14-56-26\_268\_2427948028839272408-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile50--.hashtable (431 bytes)

2017-12-24 14:56:28 End of local task; Time Taken: 0.567 sec.

Execution completed successfully

MapredLocal task succeeded

Launching Job 1 out of 1

Number of reduce tasks is set to 0 since there's no reduce operator

Starting Job = job\_1514094803278\_0018, Tracking URL = http://rootuser:8088/proxy/application\_1514094803278\_0018/

Kill Command = /usr/local/hadoop/bin/hadoop job -kill job\_1514094803278\_0018

Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0

2017-12-24 14:56:33,830 Stage-3 map = 0%, reduce = 0%

2017-12-24 14:56:37,951 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 1.31 sec

MapReduce Total cumulative CPU time: 1 seconds 310 msec

Ended Job = job\_1514094803278\_0018

MapReduce Jobs Launched:

Stage-Stage-3: Map: 1 Cumulative CPU: 1.31 sec HDFS Read: 6476 HDFS Write: 141 SUCCESS

Total MapReduce CPU Time Spent: 1 seconds 310 msec

OK

b.name a.city a.salary b.email

swetha Chennai 250000.0 swetha@gmail.com

tarun Pondi 300000.0 tarun@edureka.in

nagesh NULL NULL nagesh@yahoo.com

venkatesh NULL NULL venki@gmail.com

Time taken: 12.752 seconds, Fetched: 4 row(s)

full outer join

hive (niit)> select a.name,a.city,a.salary,b.email from

> employee a full outer join mailid b on a.name = b.name;

Query ID = hduser\_20171224145646\_2458615f-5db5-419f-bddf-736f14999047

Total jobs = 1

Launching Job 1 out of 1

Number of reduce tasks not specified. Estimated from input data size: 1

In order to change the average load for a reducer (in bytes):

set hive.exec.reducers.bytes.per.reducer=<number>

In order to limit the maximum number of reducers:

set hive.exec.reducers.max=<number>

In order to set a constant number of reducers:

set mapreduce.job.reduces=<number>

Starting Job = job\_1514094803278\_0019, Tracking URL = http://rootuser:8088/proxy/application\_1514094803278\_0019/

Kill Command = /usr/local/hadoop/bin/hadoop job -kill job\_1514094803278\_0019

Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1

2017-12-24 14:56:50,269 Stage-1 map = 0%, reduce = 0%

2017-12-24 14:56:54,412 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.59 sec

2017-12-24 14:56:59,552 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.07 sec

MapReduce Total cumulative CPU time: 4 seconds 70 msec

Ended Job = job\_1514094803278\_0019

MapReduce Jobs Launched:

Stage-Stage-1: Map: 2 Reduce: 1 Cumulative CPU: 4.07 sec HDFS Read: 12882 HDFS Write: 206 SUCCESS

Total MapReduce CPU Time Spent: 4 seconds 70 msec

OK

a.name a.city a.salary b.email

NULL NULL NULL

NULL NULL NULL

anamika Kanyakumari 200000.0 NULL

anita Selam 250000.0 NULL

NULL NULL NULL nagesh@yahoo.com

swetha Chennai 250000.0 swetha@gmail.com

tarun Pondi 300000.0 tarun@edureka.in

NULL NULL NULL venki@gmail.com

Time taken: 13.716 seconds, Fetched: 8 row(s)

Setting up local variables and parameters in hive

hive(niit)> set myage=25;

hive (niit)> select \* from customer where age >= ${hiveconf:myage} limit 10;

OK

customer.custno customer.firstname customer.lastname customer.age customer.profession

4000001 Kristina Chung 55 Pilot

4000002 Paige Chen 74 Teacher

4000003 Sherri Melton 34 Firefighter

4000004 Gretchen Hill 66 Computer hardware engineer

4000005 Karen Puckett 74 Lawyer

4000006 Patrick Song 42 Veterinarian

4000007 Elsie Hamilton 43 Pilot

4000008 Hazel Bender 63 Carpenter

4000009 Malcolm Wagner 39 Artist

4000010 Dolores McLaughlin 60 Writer

-------------------------------------------------------------------------------------------------------

hduser@rootuser:~$ gedit professionalcount.sql

set myage=25;

select profession, count(\*) from customer where age >= ${hiveconf:myage} group by profession order by profession;

hduser@rootuser:~$ hive -f professionalcount.sql

Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-common-1.2.1.jar!/hive-log4j.properties

Query ID = hduser\_20171224163003\_c61ab665-3716-42c7-b6b3-599552b0a915

Total jobs = 2

Launching Job 1 out of 2

Number of reduce tasks not specified. Estimated from input data size: 1

In order to change the average load for a reducer (in bytes):

set hive.exec.reducers.bytes.per.reducer=<number>

In order to limit the maximum number of reducers:

set hive.exec.reducers.max=<number>

In order to set a constant number of reducers:

set mapreduce.job.reduces=<number>

Starting Job = job\_1514094803278\_0020, Tracking URL = http://rootuser:8088/proxy/application\_1514094803278\_0020/

Kill Command = /usr/local/hadoop/bin/hadoop job -kill job\_1514094803278\_0020

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2017-12-24 16:30:09,281 Stage-1 map = 0%, reduce = 0%

2017-12-24 16:30:13,413 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.57 sec

2017-12-24 16:30:18,594 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.83 sec

MapReduce Total cumulative CPU time: 3 seconds 830 msec

Ended Job = job\_1514094803278\_0020

Launching Job 2 out of 2

Number of reduce tasks determined at compile time: 1

In order to change the average load for a reducer (in bytes):

set hive.exec.reducers.bytes.per.reducer=<number>

In order to limit the maximum number of reducers:

set hive.exec.reducers.max=<number>

In order to set a constant number of reducers:

set mapreduce.job.reduces=<number>

Starting Job = job\_1514094803278\_0021, Tracking URL = http://rootuser:8088/proxy/application\_1514094803278\_0021/

Kill Command = /usr/local/hadoop/bin/hadoop job -kill job\_1514094803278\_0021

Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1

2017-12-24 16:30:28,206 Stage-2 map = 0%, reduce = 0%

2017-12-24 16:30:31,378 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 0.71 sec

2017-12-24 16:30:35,504 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 1.93 sec

MapReduce Total cumulative CPU time: 1 seconds 930 msec

Ended Job = job\_1514094803278\_0021

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.83 sec HDFS Read: 4817573 HDFS Write: 2260 SUCCESS

Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 1.93 sec HDFS Read: 6778 HDFS Write: 1150 SUCCESS

Total MapReduce CPU Time Spent: 5 seconds 760 msec

OK

profession \_c1

Accountant 191

Actor 197

Agricultural and food scientist 190

Air Sports 872

Architect 195

Artist 168

Athlete 191

Automotive mechanic 185

Carpenter 174

User Define Functions

hive (niit)> create table testing(id string,unixtime string)

> row format delimited

> fields terminated by ',';

OK

Time taken: 0.351 seconds

hive (niit)> load data local inpath '/home/hduser/count.txt' into table testing;

Loading data to table niit.testing

Table niit.testing stats: [numFiles=1, totalSize=86]

OK

Time taken: 0.566 seconds

hive (niit)> select \* from testing;

OK

testing.id testing.unixtime

one 1386023259550 NULL

two 1389523259550 NULL

three 1389523259550 NULL

four 1389523259550 NULL

Time taken: 0.285 seconds, Fetched: 4 row(s)

Unixtime stores date & time as milliseconds. So we have to convert it into proper format.

Word count in hive:

hive (niit)> create table wordcountInput(line string);

OK

Time taken: 0.073 seconds

hive (niit)> desc wordcountInput;

OK

col\_name data\_type comment

line string

Time taken: 0.083 seconds, Fetched: 1 row(s)

hive (niit)> load data local inpath '/home/hduser/filein' overwrite into table wordcountInput;

Loading data to table niit.wordcountinput

Table niit.wordcountinput stats: [numFiles=1, numRows=0, totalSize=45, rawDataSize=0]

OK

Time taken: 0.217 seconds

hive (niit)> select \* from wordcountInput;

OK

wordcountinput.line

Hadoop Pig Hive

Hive Pig Pig

Pig Hive Hadoop

Time taken: 0.071 seconds, Fetched: 3 row(s)

hive (niit)> select split(line, ' ' ) As word from wordcountInput;

OK

word

["Hadoop","Pig","Hive"]

["Hive","Pig","Pig"]

["Pig","Hive","Hadoop"]

Time taken: 0.083 seconds, Fetched: 3 row(s)

hive (niit)> select explode(split(line, ' ')) As word from wordcountInput;

OK

word

Hadoop

Pig

Hive

Hive

Pig

Pig

Pig

Hive

Hadoop

Time taken: 0.058 seconds, Fetched: 9 row(s)

hive (niit)> select word, count(\*) as count from( select explode(split(line, ' ' )) as word from wordcountInput) a group by word;

Query ID = hduser\_20171224164918\_3bbeadcc-b22b-40d7-8a44-299a4aaa711d

Total jobs = 1

Launching Job 1 out of 1

Number of reduce tasks not specified. Estimated from input data size: 1

In order to change the average load for a reducer (in bytes):

set hive.exec.reducers.bytes.per.reducer=<number>

In order to limit the maximum number of reducers:

set hive.exec.reducers.max=<number>

In order to set a constant number of reducers:

set mapreduce.job.reduces=<number>

Starting Job = job\_1514094803278\_0022, Tracking URL = http://rootuser:8088/proxy/application\_1514094803278\_0022/

Kill Command = /usr/local/hadoop/bin/hadoop job -kill job\_1514094803278\_0022

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2017-12-24 16:49:25,946 Stage-1 map = 0%, reduce = 0%

2017-12-24 16:49:31,221 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.84 sec

2017-12-24 16:49:37,484 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 5.23 sec

MapReduce Total cumulative CPU time: 5 seconds 230 msec

Ended Job = job\_1514094803278\_0022

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 5.23 sec HDFS Read: 7554 HDFS Write: 22 SUCCESS

Total MapReduce CPU Time Spent: 5 seconds 230 msec

OK

word count

Hadoop 2

Hive 3

Pig 4

Time taken: 19.905 seconds, Fetched: 3 row(s)