Last login: Thu May 22 16:37:11 on ttys000

(base) pavethraavadiar@Pavethras-MacBook-Air ~ % $HIVE\_HOME/bin/hive

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding in [jar:file:/Users/pavethraavadiar/hive/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/opt/homebrew/Cellar/hadoop/3.4.1/libexec/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: See http://www.slf4j.org/codes.html#multiple\_bindings for an explanation.

SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Hive Session ID = 111ed813-6edf-4ea8-932c-bc598b388a6e

Logging initialized using configuration in jar:file:/Users/pavethraavadiar/hive/lib/hive-common-3.1.3.jar!/hive-log4j2.properties Async: true

Hive Session ID = 1bb0bc12-b570-4367-b93c-6bd17cfe4b3f

Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.

hive> CREATE TABLE test\_table(id INT, name STRING);

OK

Time taken: 1.711 seconds

hive> SHOW TABLES;

OK

test\_table

Time taken: 0.261 seconds, Fetched: 1 row(s)

hive> hdfs dfs -mkdir -p /user/pavethraavadiar/input

> hdfs dfs -put "/Users/pavethraavadiar/Documents/Masters - Sem 2/Data management/project 1 -/insurance.csv" /user/pavethraavadiar/input/

> CREATE EXTERNAL TABLE IF NOT EXISTS insurance (

> age INT,

> sex STRING,

> bmi FLOAT,

> children INT,

> smoker STRING,

> region STRING,

> charges FLOAT

> )

> ROW FORMAT DELIMITED

> FIELDS TERMINATED BY ','

> STORED AS TEXTFILE

> LOCATION '/user/pavethraavadiar/input/';

NoViableAltException(24@[])

at org.apache.hadoop.hive.ql.parse.HiveParser.statement(HiveParser.java:1387)

at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:220)

at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:74)

at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:67)

at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:616)

at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1826)

at org.apache.hadoop.hive.ql.Driver.compileAndRespond(Driver.java:1773)

at org.apache.hadoop.hive.ql.Driver.compileAndRespond(Driver.java:1768)

at org.apache.hadoop.hive.ql.reexec.ReExecDriver.compileAndRespond(ReExecDriver.java:126)

at org.apache.hadoop.hive.ql.reexec.ReExecDriver.run(ReExecDriver.java:214)

at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:239)

at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:188)

at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:402)

at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:821)

at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:759)

at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:683)

at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)

at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)

at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)

at java.lang.reflect.Method.invoke(Method.java:498)

at org.apache.hadoop.util.RunJar.run(RunJar.java:330)

at org.apache.hadoop.util.RunJar.main(RunJar.java:245)

FAILED: ParseException line 1:0 cannot recognize input near 'hdfs' 'dfs' '-'

hive> CREATE EXTERNAL TABLE IF NOT EXISTS insurance (

> age INT,

> sex STRING,

> bmi FLOAT,

> children INT,

> smoker STRING,

> region STRING,

> charges FLOAT

> )

> ROW FORMAT DELIMITED

> FIELDS TERMINATED BY ','

> STORED AS TEXTFILE

> LOCATION '/user/pavethraavadiar/input/';

OK

Time taken: 0.44 seconds

hive> SET hive.exec.skip.header.line.count=1;

Query returned non-zero code: 1, cause: hive configuration hive.exec.skip.header.line.count does not exists.

hive> CREATE EXTERNAL TABLE IF NOT EXISTS insurance (

> age INT,

> sex STRING,

> bmi FLOAT,

> children INT,

> smoker STRING,

> region STRING,

> charges FLOAT

> )

> ROW FORMAT DELIMITED

> FIELDS TERMINATED BY ','

> STORED AS TEXTFILE

> LOCATION '/user/pavethraavadiar/input/'

> TBLPROPERTIES ("skip.header.line.count"="1");

OK

Time taken: 0.382 seconds

hive> SELECT \* FROM insurance LIMIT 5;

OK

NULL sex NULL NULL smoker region NULL

19 female 27.9 0 yes southwest 16884.924

18 male 33.77 1 no southeast 1725.5522

28 male 33.0 3 no southeast 4449.462

33 male 22.705 0 no northwest 21984.47

Time taken: 2.059 seconds, Fetched: 5 row(s)

hive> SELECT COUNT(\*) AS total\_records FROM insurance;

Query ID = pavethraavadiar\_20250522165648\_003d4113-8c90-47f7-8094-c527a04c3c93

Total jobs = 1

Launching Job 1 out of 1

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>

Job running in-process (local Hadoop)

2025-05-22 16:56:52,059 Stage-1 map = 100%, reduce = 100%

Ended Job = job\_local1302600622\_0001

MapReduce Jobs Launched:

Stage-Stage-1: HDFS Read: 116768 HDFS Write: 0 SUCCESS

Total MapReduce CPU Time Spent: 0 msec

OK

1339

Time taken: 3.124 seconds, Fetched: 1 row(s)

hive> SELECT region, COUNT(\*) AS count\_per\_region FROM insurance GROUP BY region;

Query ID = pavethraavadiar\_20250522165720\_f5ac861f-6e96-45e1-8784-4c8f84fe9862

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>

Job running in-process (local Hadoop)

2025-05-22 16:57:23,343 Stage-1 map = 100%, reduce = 100%

Ended Job = job\_local2033692582\_0002

MapReduce Jobs Launched:

Stage-Stage-1: HDFS Read: 225344 HDFS Write: 0 SUCCESS

Total MapReduce CPU Time Spent: 0 msec

OK

northeast 324

northwest 325

region 1

southeast 364

southwest 325

Time taken: 2.387 seconds, Fetched: 5 row(s)

hive> SELECT smoker, COUNT(\*) AS count\_per\_smoker FROM insurance GROUP BY smoker;

Query ID = pavethraavadiar\_20250522165741\_081c7501-c1c7-4a24-9069-a664fcc0327c

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>

Job running in-process (local Hadoop)

2025-05-22 16:57:43,483 Stage-1 map = 100%, reduce = 100%

Ended Job = job\_local730757274\_0003

MapReduce Jobs Launched:

Stage-Stage-1: HDFS Read: 333920 HDFS Write: 0 SUCCESS

Total MapReduce CPU Time Spent: 0 msec

OK

no 1064

smoker 1

yes 274

Time taken: 2.178 seconds, Fetched: 3 row(s)

hive> SELECT \* FROM insurance LIMIT 5;

OK

NULL sex NULL NULL smoker region NULL

19 female 27.9 0 yes southwest 16884.924

18 male 33.77 1 no southeast 1725.5522

28 male 33.0 3 no southeast 4449.462

33 male 22.705 0 no northwest 21984.47

Time taken: 0.325 seconds, Fetched: 5 row(s)

hive> CREATE OR REPLACE VIEW insurance\_clean AS

> SELECT \*

> FROM insurance

> WHERE age != 'age';

OK

Time taken: 0.778 seconds

hive> DROP TABLE IF EXISTS insurance;

OK

Time taken: 1.195 seconds

hive> CREATE EXTERNAL TABLE insurance (

> age INT,

> sex STRING,

> bmi FLOAT,

> children INT,

> smoker STRING,

> region STRING,

> charges FLOAT

> )

> ROW FORMAT DELIMITED

> FIELDS TERMINATED BY ','

> STORED AS TEXTFILE

> LOCATION '/user/pavethraavadiar/input/';

OK

Time taken: 0.24 seconds

hive> SELECT region, COUNT(\*) AS count\_per\_region FROM insurance GROUP BY region;

Query ID = pavethraavadiar\_20250522170409\_ca6be1a2-cf1e-4619-921b-98c420623bb3

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>

Job running in-process (local Hadoop)

2025-05-22 17:04:11,204 Stage-1 map = 100%, reduce = 100%

Ended Job = job\_local576860099\_0004

MapReduce Jobs Launched:

Stage-Stage-1: HDFS Read: 450602 HDFS Write: 0 SUCCESS

Total MapReduce CPU Time Spent: 0 msec

OK

northeast 324

northwest 325

southeast 364

southwest 325

Time taken: 1.969 seconds, Fetched: 4 row(s)

hive> SELECT smoker, COUNT(\*) AS count\_per\_smoker FROM insurance GROUP BY smoker;

Query ID = pavethraavadiar\_20250522170417\_f31b6b1e-6e62-404d-81b1-754a31131a95

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>

Job running in-process (local Hadoop)

2025-05-22 17:04:19,691 Stage-1 map = 100%, reduce = 100%

Ended Job = job\_local126439118\_0005

MapReduce Jobs Launched:

Stage-Stage-1: HDFS Read: 559092 HDFS Write: 0 SUCCESS

Total MapReduce CPU Time Spent: 0 msec

OK

no 1064

yes 274

Time taken: 1.953 seconds, Fetched: 2 row(s)

hive> SELECT \* FROM insurance LIMIT 5;

OK

19 female 27.9 0 yes southwest 16884.924

18 male 33.77 1 no southeast 1725.5522

28 male 33.0 3 no southeast 4449.462

33 male 22.705 0 no northwest 21984.47

32 male 28.88 0 no northwest 3866.8552

Time taken: 0.501 seconds, Fetched: 5 row(s)

hive> SELECT COUNT(\*) AS total\_records FROM insurance;

Query ID = pavethraavadiar\_20250522170502\_9fd30f98-7c8c-49f5-b03f-31280ca314b2

Total jobs = 1

Launching Job 1 out of 1

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>

Job running in-process (local Hadoop)

2025-05-22 17:05:04,722 Stage-1 map = 100%, reduce = 100%

Ended Job = job\_local1037066783\_0006

MapReduce Jobs Launched:

Stage-Stage-1: HDFS Read: 675774 HDFS Write: 0 SUCCESS

Total MapReduce CPU Time Spent: 0 msec

OK

1338

Time taken: 1.936 seconds, Fetched: 1 row(s)

hive> SELECT smoker, AVG(charges) AS avg\_charges FROM insurance GROUP BY smoker;

Query ID = pavethraavadiar\_20250522170511\_d3aa9dfd-4d0d-4b25-a442-e7a1fb546778

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>

Job running in-process (local Hadoop)

2025-05-22 17:05:13,296 Stage-1 map = 100%, reduce = 100%

Ended Job = job\_local221625851\_0007

MapReduce Jobs Launched:

Stage-Stage-1: HDFS Read: 784264 HDFS Write: 0 SUCCESS

Total MapReduce CPU Time Spent: 0 msec

OK

no 8434.268304122123

yes 32050.23178033759

Time taken: 2.098 seconds, Fetched: 2 row(s)

hive> SELECT region, AVG(bmi) AS avg\_bmi FROM insurance GROUP BY region;

Query ID = pavethraavadiar\_20250522170521\_f866ddb5-7e78-4853-bdd7-48e95a980ef9

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>

Job running in-process (local Hadoop)

2025-05-22 17:05:23,467 Stage-1 map = 100%, reduce = 100%

Ended Job = job\_local405501327\_0008

MapReduce Jobs Launched:

Stage-Stage-1: HDFS Read: 892754 HDFS Write: 0 SUCCESS

Total MapReduce CPU Time Spent: 0 msec

OK

northeast 29.173503078060385

northwest 29.19978463979868

southeast 33.355989031739284

southwest 30.59661532475398

Time taken: 2.118 seconds, Fetched: 4 row(s)

hive> SELECT region,

> MAX(charges) AS max\_charge,

> MIN(charges) AS min\_charge,

> AVG(charges) AS avg\_charge

> FROM insurance

> GROUP BY region;

Query ID = pavethraavadiar\_20250522170531\_def10883-f49f-4013-9bfa-7a73aa97824f

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>

Job running in-process (local Hadoop)

2025-05-22 17:05:33,767 Stage-1 map = 100%, reduce = 100%

Ended Job = job\_local1318613592\_0009

MapReduce Jobs Launched:

Stage-Stage-1: HDFS Read: 1001244 HDFS Write: 0 SUCCESS

Total MapReduce CPU Time Spent: 0 msec

OK

northeast 58571.074 1694.7964 13406.38453128014

northwest 60021.4 1621.3402 12417.575369591346

southeast 63770.43 1121.8739 14735.411452073316

southwest 52590.83 1241.565 12346.93732797476

Time taken: 2.067 seconds, Fetched: 4 row(s)

hive> SELECT children, COUNT(\*) AS count FROM insurance GROUP BY children ORDER BY children;

Query ID = pavethraavadiar\_20250522170549\_b219df21-fc28-457b-aaee-f96252548d47

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>

Job running in-process (local Hadoop)

2025-05-22 17:05:51,439 Stage-1 map = 100%, reduce = 100%

Ended Job = job\_local277433313\_0010

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>

Job running in-process (local Hadoop)

2025-05-22 17:05:53,124 Stage-2 map = 100%, reduce = 100%

Ended Job = job\_local2069818482\_0011

MapReduce Jobs Launched:

Stage-Stage-1: HDFS Read: 1109734 HDFS Write: 0 SUCCESS

Stage-Stage-2: HDFS Read: 1109734 HDFS Write: 0 SUCCESS

Total MapReduce CPU Time Spent: 0 msec

OK

0 574

1 324

2 240

3 157

4 25

5 18

Time taken: 3.874 seconds, Fetched: 6 row(s)

hive> SELECT \* FROM insurance ORDER BY charges DESC LIMIT 5;

Query ID = pavethraavadiar\_20250522170559\_76f0ba1e-1236-4ca2-b8d2-d78ffe2cc87d

Total jobs = 1

Launching Job 1 out of 1

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>

Job running in-process (local Hadoop)

2025-05-22 17:06:01,572 Stage-1 map = 100%, reduce = 100%

Ended Job = job\_local1667326854\_0012

MapReduce Jobs Launched:

Stage-Stage-1: HDFS Read: 1218224 HDFS Write: 0 SUCCESS

Total MapReduce CPU Time Spent: 0 msec

OK

54 female 47.41 0 yes southeast 63770.43

45 male 30.36 0 yes southeast 62592.875

52 male 34.485 3 yes northwest 60021.4

31 female 38.095 1 yes northeast 58571.074

33 female 35.53 0 yes northwest 55135.402

Time taken: 1.871 seconds, Fetched: 5 row(s)

hive> SELECT \* FROM insurance WHERE smoker='no' AND charges > 20000;

OK

33 male 22.705 0 no northwest 21984.47

60 female 25.84 0 no northwest 28923.137

55 male 37.3 0 no southwest 20630.283

64 male 24.7 1 no northwest 30166.62

18 female 30.115 0 no northeast 21344.848

60 male 28.595 0 no northeast 30259.996

54 female 31.9 3 no southeast 27322.734

34 male 22.42 2 no northeast 27375.904

24 female 23.21 0 no southeast 25081.768

58 female 41.91 0 no southeast 24227.338

55 female 26.8 1 no southwest 35160.133

54 male 30.02 0 no northwest 24476.479

53 female 38.06 3 no southeast 20462.998

52 male 26.4 3 no southeast 25992.82

29 male 29.64 1 no northeast 20277.807

28 female 27.5 2 no southwest 20177.672

26 female 29.64 4 no northeast 24671.664

46 male 27.6 0 no southwest 24603.049

62 male 31.46 1 no southeast 27000.984

55 male 33.0 0 no southeast 20781.488

50 male 25.365 2 no northwest 30284.643

19 male 33.1 0 no southwest 23082.955

59 female 36.52 1 no southeast 28287.898

59 male 37.4 0 no southwest 21797.0

28 female 24.32 1 no northeast 23288.928

47 female 33.345 0 no northeast 20878.785

61 female 25.08 0 no southeast 24513.092

20 male 35.31 1 no southeast 27724.29

50 female 27.36 0 no northeast 25656.576

19 female 30.59 2 no northwest 24059.68

53 male 31.35 0 no southeast 27346.043

62 female 36.86 1 no northeast 31620.002

52 female 37.525 2 no northwest 33471.973

35 female 38.095 2 no northeast 24915.047

48 female 35.91 1 no northeast 26392.26

57 female 23.98 1 no southeast 22192.438

47 female 24.1 1 no southwest 26236.58

53 female 32.3 2 no northeast 29186.482

61 male 36.1 3 no southwest 27941.287

40 female 41.42 1 no northwest 28476.734

49 female 27.1 1 no southwest 26140.36

50 male 32.11 2 no northeast 25333.332

44 male 29.735 2 no northeast 32108.662

48 male 36.67 1 no northwest 28468.92

52 male 36.765 2 no northwest 26467.098

54 male 25.46 1 no northeast 25517.113

45 female 27.645 1 no northwest 28340.19

48 male 29.6 0 no southwest 21232.182

25 male 24.985 2 no northeast 23241.475

61 female 33.33 4 no southeast 36580.28

21 female 32.68 2 no northwest 26018.951

23 male 18.715 0 no northwest 21595.383

19 male 27.265 2 no northwest 22493.66

52 female 30.875 0 no northeast 23045.566

37 male 29.8 0 no southwest 20420.605

57 male 40.28 0 no northeast 20709.02

52 female 24.86 0 no southeast 27117.994

59 female 34.8 2 no southwest 36910.61

39 male 34.1 2 no southeast 23563.016

55 male 37.715 3 no northwest 30063.58

23 female 24.225 2 no northeast 22395.744

Time taken: 0.693 seconds, Fetched: 61 row(s)

hive> SELECT sex, COUNT(\*) AS count FROM insurance GROUP BY sex;

Query ID = pavethraavadiar\_20250522170630\_44b9cc3c-d84d-4143-9947-e271a9e9fe00

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>

Job running in-process (local Hadoop)

2025-05-22 17:06:31,969 Stage-1 map = 100%, reduce = 100%

Ended Job = job\_local1683718212\_0013

MapReduce Jobs Launched:

Stage-Stage-1: HDFS Read: 1435204 HDFS Write: 0 SUCCESS

Total MapReduce CPU Time Spent: 0 msec

OK

female 662

male 676

Time taken: 1.807 seconds, Fetched: 2 row(s)

hive> SELECT sex, smoker, AVG(charges) AS avg\_charges FROM insurance GROUP BY sex, smoker;

Query ID = pavethraavadiar\_20250522170644\_28a25822-8905-46eb-b6f6-f3f054fe8fcd

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>

Job running in-process (local Hadoop)

2025-05-22 17:06:46,885 Stage-1 map = 100%, reduce = 100%

Ended Job = job\_local681561785\_0014

MapReduce Jobs Launched:

Stage-Stage-1: HDFS Read: 1543694 HDFS Write: 0 SUCCESS

Total MapReduce CPU Time Spent: 0 msec

OK

female no 8762.297322665534

female yes 30678.99621263587

male no 8087.204719705785

male yes 33042.00593307783

Time taken: 2.195 seconds, Fetched: 4 row(s)

hive> CREATE VIEW insurance\_summary AS

> SELECT sex, smoker, region, AVG(charges) AS avg\_charges, COUNT(\*) AS count\_records

> FROM insurance

> GROUP BY sex, smoker, region;

OK

Time taken: 0.383 seconds

hive> xSELECT \* FROM insurance\_summary WHERE region = 'southeast' ORDER BY avg\_charges DESC;

NoViableAltException(24@[])

at org.apache.hadoop.hive.ql.parse.HiveParser.statement(HiveParser.java:1387)

at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:220)

at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:74)

at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:67)

at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:616)

at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1826)

at org.apache.hadoop.hive.ql.Driver.compileAndRespond(Driver.java:1773)

at org.apache.hadoop.hive.ql.Driver.compileAndRespond(Driver.java:1768)

at org.apache.hadoop.hive.ql.reexec.ReExecDriver.compileAndRespond(ReExecDriver.java:126)

at org.apache.hadoop.hive.ql.reexec.ReExecDriver.run(ReExecDriver.java:214)

at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:239)

at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:188)

at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:402)

at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:821)

at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:759)

at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:683)

at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)

at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)

at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)

at java.lang.reflect.Method.invoke(Method.java:498)

at org.apache.hadoop.util.RunJar.run(RunJar.java:330)

at org.apache.hadoop.util.RunJar.main(RunJar.java:245)

FAILED: ParseException line 1:0 cannot recognize input near 'xSELECT' '\*' 'FROM'

hive> CREATE VIEW insurance\_summary AS

> SELECT sex, smoker, region, AVG(charges) AS avg\_charges, COUNT(\*) AS count\_records

> FROM insurance

> GROUP BY sex, smoker, region;

FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. Table already exists: default.insurance\_summary

hive> SELECT \* FROM insurance\_summary WHERE region = 'southeast' ORDER BY avg\_charges DESC;

Query ID = pavethraavadiar\_20250522170754\_02356fee-0531-405e-8ae7-8967b5c551d4

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>

Job running in-process (local Hadoop)

2025-05-22 17:07:56,781 Stage-1 map = 100%, reduce = 100%

Ended Job = job\_local1170129872\_0015

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>

Job running in-process (local Hadoop)

2025-05-22 17:07:58,419 Stage-2 map = 100%, reduce = 100%

Ended Job = job\_local399435209\_0016

MapReduce Jobs Launched:

Stage-Stage-1: HDFS Read: 1652184 HDFS Write: 0 SUCCESS

Stage-Stage-2: HDFS Read: 1652184 HDFS Write: 0 SUCCESS

Total MapReduce CPU Time Spent: 0 msec

OK

male yes southeast 36029.839417613635 55

female yes southeast 33034.82074652778 36

female no southeast 8440.205582213915 139

male no southeast 7609.003565546292 134

Time taken: 3.784 seconds, Fetched: 4 row(s)

hive>

Perfect! The file `insurance\_noheader.csv` is now successfully uploaded into HDFS at `/user/pavethraavadiar/input/`.

### Next steps for Hive:

Since your data file no longer has a header row, you can now create or update your Hive external table \*\*without\*\* the `TBLPROPERTIES ("skip.header.line.count"="1")` property.

---

### 1. Drop existing table (if needed):

```sql

DROP TABLE IF EXISTS insurance;

```

### 2. Create the external table pointing to the HDFS directory, assuming no header row:

```sql

CREATE EXTERNAL TABLE insurance (

age INT,

sex STRING,

bmi FLOAT,

children INT,

smoker STRING,

region STRING,

charges FLOAT

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

LOCATION '/user/pavethraavadiar/input/';

```

---

### 3. Run some queries on the table, e.g.:

\* Count by region:

```sql

SELECT region, COUNT(\*) AS count\_per\_region FROM insurance GROUP BY region;

```

\* Count smokers vs non-smokers:

```sql

SELECT smoker, COUNT(\*) AS count\_per\_smoker FROM insurance GROUP BY smoker;

```

\* View sample data:

```sql

SELECT \* FROM insurance LIMIT 5;

```

---

If you want, I can help you write more queries or prepare some data insights to get full marks! Just say the word.

Great! Here’s a checklist of queries and analysis you can do in Hive to make your assignment stand out and show solid understanding:

---

### 1. \*\*Basic descriptive stats\*\*

\* Count total records:

```sql

SELECT COUNT(\*) AS total\_records FROM insurance;

```

\* Average charge by smoker status:

```sql

SELECT smoker, AVG(charges) AS avg\_charges FROM insurance GROUP BY smoker;

```

\* Average BMI by region:

```sql

SELECT region, AVG(bmi) AS avg\_bmi FROM insurance GROUP BY region;

```

---

### 2. \*\*More insights\*\*

\* Max, Min, Avg charges by region:

```sql

SELECT region,

MAX(charges) AS max\_charge,

MIN(charges) AS min\_charge,

AVG(charges) AS avg\_charge

FROM insurance

GROUP BY region;

```

\* Number of children distribution:

```sql

SELECT children, COUNT(\*) AS count FROM insurance GROUP BY children ORDER BY children;

```

---

### 3. \*\*Filtering and sorting\*\*

\* Top 5 highest charges:

```sql

SELECT \* FROM insurance ORDER BY charges DESC LIMIT 5;

```

\* Non-smokers with charges above 20,000:

```sql

SELECT \* FROM insurance WHERE smoker='no' AND charges > 20000;

```

---

### 4. \*\*Simple aggregations for report\*\*

\* Count of males vs females:

```sql

SELECT sex, COUNT(\*) AS count FROM insurance GROUP BY sex;

```

\* Average charges by sex and smoker (combined grouping):

```sql

SELECT sex, smoker, AVG(charges) AS avg\_charges FROM insurance GROUP BY sex, smoker;

```

---

### 5. \*\*Optional: Create a view for easier queries\*\*

```sql

CREATE VIEW insurance\_summary AS

SELECT sex, smoker, region, AVG(charges) AS avg\_charges, COUNT(\*) AS count\_records

FROM insurance

GROUP BY sex, smoker, region;

```

Then query:

```sql

SELECT \* FROM insurance\_summary WHERE region = 'southeast' ORDER BY avg\_charges DESC;

```

---

### 6. \*\*Document your results\*\*

\* Copy-paste query outputs (or screenshots) in your report.

\* Explain what each query tells you about the dataset.

\* Interpret results (e.g., smokers have higher charges on average, southwest region has the most patients, etc.)

\* Mention how this analysis could help an insurance company.

---

If you want, I can help you draft that explanation or create a simple report outline to maximize your marks. Would you like me to help with that?