

Remember to combine tables before starting these steps

1. Create analytic table

```
INFO : Compiling command(queryId=hive_20201202143333_dcbe85cd-38fb-4dd5-bbab-d86800e1c9eb): create table analytic
as
select
state,
county,
bridges,
residents,
pctmediumtofairbridges as pct_medium_bridges,
pctpoorbridges as pct_poor_bridges,
milesfreightrailroad as miles_freight_railroad,
roadsacceptable as roads_acceptable,
countyarea as county_area,
taxrespondents as tax_respondants,
statelocalincometax as state_local_income_tax,
realestatetax as real_estate_tax,
populationsserved as population_served,
watersystems as water_systems,
pctmediumtofairbridges/pctpoorbridges as ratio_fair_to_poor,
milesfreightrailroad/countyarea as freight_per_sq_mile,
populationsserved/watersystems as water_sys_per_capita,
realestatetax/countyarea as real_estate_tax_per_sq_mile,
residents/countyarea as population_density
from combined
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:state, type:string, comment:null), FieldSchema(n
ame:county, type:string, comment:null), FieldSchema(name:bridges, type:int, comment:null), FieldSchema(name:residents
, type:int, comment:null), FieldSchema(name:pct_medium_bridges, type:double, comment:null), FieldSchema(name:pct_poor
_bridges, type:double, comment:null), FieldSchema(name:miles_freight_railroad, type:double, comment:null), FieldSchem
a(name:roads_acceptable, type:double, comment:null), FieldSchema(name:county_area, type:double, comment:null), FieldS
chema(name:tax_respondants, type:double, comment:null), FieldSchema(name:state_local_income_tax, type:double, comment
:null), FieldSchema(name:real_estate_tax, type:double, comment:null), FieldSchema(name:population_served, type:int, c
omment:null), FieldSchema(name:water_systems, type:int, comment:null), FieldSchema(name:ratio_fair_to_poor, type:doub
le, comment:null), FieldSchema(name:freight_per_sq_mile, type:double, comment:null), FieldSchema(name:water_sys_per_c
apita, type:double, comment:null), FieldSchema(name:real_estate_tax_per_sq_mile, type:double, comment:null), FieldSch
ema(name:population_density, type:double, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20201202143333_dcbe85cd-38fb-4dd5-bbab-d86800e1c9eb); Time taken: 0.
328 seconds
INFO : Executing command(queryId=hive_20201202143333_dcbe85cd-38fb-4dd5-bbab-d86800e1c9eb): create table analytic
as
```

```

INFO : Executing command(queryId=hive_20201202143333_dcbe85cd-38fb-4dd5-bbab-d86800e1c9eb): create table analytic
as
select
state,
county,
bridges,
residents,
pctmediumtofairbridges as pct_medium_bridges,
pctpoorbridges as pct_poor_bridges,
milesfreightrailroad as miles_freight_railroad,
roadsacceptable as roads_acceptable,
countyarea as county_area,
taxrespondents as tax_respondants,
statelocalincometax as state_local_income_tax,
realestatetax as real_estate_tax,
populationserved as population_served,
watersystems as water_systems,
pctmediumtofairbridges/pctpoorbridges as ratio_fair_to_poor,
milesfreightrailroad/countyarea as freight_per_sq_mile,
populationserved/watersystems as water_sys_per_capita,
realestatetax/countyarea as real_estate_tax_per_sq_mile,
residents/countyarea as population_density
from combined
INFO : Query ID = hive_20201202143333_dcbe85cd-38fb-4dd5-bbab-d86800e1c9eb
INFO : Total jobs = 3
INFO : Launching Job 1 out of 3
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Number of reduce tasks is set to 0 since there's no reduce operator
INFO : number of splits:1
INFO : Submitting tokens for job: job_1604346392376_7458
INFO : The url to track the job: http://babar.es.its.nyu.edu:8088/proxy/application_1604346392376_7458/
INFO : Starting Job = job_1604346392376_7458, Tracking URL = http://babar.es.its.nyu.edu:8088/proxy/application_1604
346392376_7458/
INFO : Kill Command = /opt/cloudera/parcels/CDH-5.15.2-1.cd5.15.2.p0.3/lib/hadoop/bin/hadoop job -kill job_1604346
392376_7458
INFO : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
INFO : 2020-12-02 14:33:43,950 Stage-1 map = 0%, reduce = 0%
INFO : 2020-12-02 14:33:51,135 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.59 sec
INFO : MapReduce Total cumulative CPU time: 3 seconds 590 msec
INFO : Ended Job = job_1604346392376_7458
INFO : Starting task [Stage-7:CONDITIONAL] in serial mode
INFO : Stage-4 is selected by condition resolver.
INFO : Stage-3 is filtered out by condition resolver.
INFO : Stage-5 is filtered out by condition resolver.
INFO : Starting task [Stage-4:MOVE] in serial mode
INFO : Moving data to: hdfs://dumbo/user/hive/warehouse/jc8017.db/.hive-staging_hive_2020-12-02_14-33-27_741_8975220
081670741264-1269/-ext-10001 from hdfs://dumbo/user/hive/warehouse/jc8017.db/.hive-staging_hive_2020-12-02_14-33-27_7
41_8975220081670741264-1269/-ext-10003
INFO : Starting task [Stage-0:MOVE] in serial mode
INFO : Moving data to: hdfs://dumbo/user/hive/warehouse/jc8017.db/.hive-staging_hive_2020-12-02_14-33-27_741_8975220081670741264-1269/-ext-10001
INFO : Starting task [Stage-8:DDL] in serial mode
INFO : Starting task [Stage-2:STATS] in serial mode
INFO : Table jc8017.analytic stats: [numFiles=1, numRows=3130, totalSize=583026, rawDataSize=579896]
INFO : MapReduce Jobs Launched:
INFO : Stage-Stage-1: Map: 1 Cumulative CPU: 3.59 sec HDFS Read: 326906 HDFS Write: 583103 SUCCESS
INFO : Total MapReduce CPU Time Spent: 3 seconds 590 msec
INFO : Completed executing command(queryId=hive_20201202143333_dcbe85cd-38fb-4dd5-bbab-d86800e1c9eb); Time taken: 25
.877 seconds
INFO : OK
No rows affected (26.309 seconds)

```

2. Create analytic_mean (Get the means of each column)

```

0: jdbc:hive2://babar.es.its.nyu.edu:10000/> CREATE TABLE analytic_mean
. . . . .> as
. . . . .> select
. . . . .> avg(bridges) as bridges_mean,
. . . . .> avg(residents) as resident_mean,
. . . . .> avg(pct_medium_bridges) as pct_medium_bridges_mean,
. . . . .> avg(pct_poor_bridges) as pct_poor_bridges_mean,
. . . . .> avg(miles_freight_railroad) as miles_freight_railroad_mean,
. . . . .> avg(roads_acceptable) as roads_acceptable_mean,
. . . . .> avg(county_area) as county_area_mean,
. . . . .> avg(tax_respondants) as tax_respondants_mean,
. . . . .> avg(state_local_income_tax) as state_local_income_tax_mean,
. . . . .> avg(real_estate_tax) as real_estate_tax_mean,
. . . . .> avg(population_served) as population_served_mean,
. . . . .> avg(water_systems) as water_systems_mean,
. . . . .> avg(ratio_fair_to_poor) as ratio_fair_to_poor_mean,
. . . . .> avg(freight_per_sq_mile) as freight_per_sq_mile_mean,
. . . . .> avg(water_sys_per_capita) as water_sys_per_cap_mean,
. . . . .> avg(real_estate_tax_per_sq_mile) as estate_tax_per_sq_mean,
. . . . .> avg(population_density) as population_density_mean
. . . . .> FROM analytic;
INFO : Compiling command(queryId=hive_20201202143737_f41f8db9-286c-4fc0-952b-c695c389cc91): CREATE TABLE analytic_me
an
as
select
avg(bridges) as bridges_mean,
avg(residents) as resident_mean,
avg(pct_medium_bridges) as pct_medium_bridges_mean,
avg(pct_poor_bridges) as pct_poor_bridges_mean,
avg(miles_freight_railroad) as miles_freight_railroad_mean,
avg(roads_acceptable) as roads_acceptable_mean,
avg(county_area) as county_area_mean,
avg(tax_respondants) as tax_respondants_mean,
avg(state_local_income_tax) as state_local_income_tax_mean,
avg(real_estate_tax) as real_estate_tax_mean,
avg(population_served) as population_served_mean,
avg(water_systems) as water_systems_mean,
avg(ratio_fair_to_poor) as ratio_fair_to_poor_mean,
avg(freight_per_sq_mile) as freight_per_sq_mile_mean,
avg(water_sys_per_capita) as water_sys_per_cap_mean,
avg(real_estate_tax_per_sq_mile) as estate_tax_per_sq_mean,
avg(population_density) as population_density_mean
FROM analytic

```



```

INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:bridges_mean, type:double, comment:null), FieldSchema(name:resident_mean, type:double, comment:null), FieldSchema(name:pct_medium_bridges_mean, type:double, comment:null), FieldSchema(name:pct_poor_bridges_mean, type:double, comment:null), FieldSchema(name:miles_freight_railroad_mean, type:double, comment:null), FieldSchema(name:roads_acceptable_mean, type:double, comment:null), FieldSchema(name:county_area_mean, type:double, comment:null), FieldSchema(name:tax_respondants_mean, type:double, comment:null), FieldSchema(name:state_local_income_tax_mean, type:double, comment:null), FieldSchema(name:real_estate_tax_mean, type:double, comment:null), FieldSchema(name:population_served_mean, type:double, comment:null), FieldSchema(name:water_systems_mean, type:double, comment:null), FieldSchema(name:ratio_fair_to_poor_mean, type:double, comment:null), FieldSchema(name:freight_per_sq_mile_mean, type:double, comment:null), FieldSchema(name:water_sys_per_cap_mean, type:double, comment:null), FieldSchema(name:estate_tax_per_sq_mean, type:double, comment:null), FieldSchema(name:population_density_mean, type:double, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20201202143737_f41f8db9-286c-4fc0-952b-c695c389cc91); Time taken: 0.33 seconds
INFO : Executing command(queryId=hive_20201202143737_f41f8db9-286c-4fc0-952b-c695c389cc91): CREATE TABLE analytic_mean
as
select
avg(bridges) as bridges_mean,
avg(residents) as resident_mean,
avg(pct_medium_bridges) as pct_medium_bridges_mean,
avg(pct_poor_bridges) as pct_poor_bridges_mean,
avg(miles_freight_railroad) as miles_freight_railroad_mean,
avg(roads_acceptable) as roads_acceptable_mean,
avg(county_area) as county_area_mean,
avg(tax_respondants) as tax_respondants_mean,
avg(state_local_income_tax) as state_local_income_tax_mean,
avg(real_estate_tax) as real_estate_tax_mean,
avg(population_served) as population_served_mean,
avg(water_systems) as water_systems_mean,
avg(ratio_fair_to_poor) as ratio_fair_to_poor_mean,
avg(freight_per_sq_mile) as freight_per_sq_mile_mean,
avg(water_sys_per_capita) as water_sys_per_cap_mean,
avg(real_estate_tax_per_sq_mile) as estate_tax_per_sq_mean,
avg(population_density) as population_density_mean
FROM analytic
INFO : Query ID = hive_20201202143737_f41f8db9-286c-4fc0-952b-c695c389cc91
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Number of reduce tasks determined at compile time: 1
INFO : In order to change the average load for a reducer (in bytes):
INFO :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO : In order to limit the maximum number of reducers:
INFO :   set hive.exec.reducers.max=<number>
INFO : In order to set a constant number of reducers:
INFO :   set mapreduce.job.reduces=<number>
INFO : number of splits:1
INFO : Submitting tokens for job: job_1604346392376_7459
INFO : The url to track the job: http://babar.es.its.nyu.edu:8088/proxy/application_1604346392376_7459/
INFO : Starting Job = job_1604346392376_7459, Tracking URL = http://babar.es.its.nyu.edu:8088/proxy/application_1604346392376_7459/

```

```

346392376_7459/
INFO : Kill Command = /opt/cloudera/parcels/CDH-5.15.2-1.cdh5.15.2.p0.3/lib/hadoop/bin/hadoop job -kill job_1604346392376_7459
INFO : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO : 2020-12-02 14:37:20,803 Stage-1 map = 0%, reduce = 0%
INFO : 2020-12-02 14:37:28,001 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.1 sec
INFO : 2020-12-02 14:37:46,527 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.21 sec
INFO : MapReduce Total cumulative CPU time: 6 seconds 650 msec
INFO : Ended Job = job_1604346392376_7459
INFO : Starting task [Stage-0:MOVE] in serial mode
INFO : Moving data to: hdfs://dumbo/user/hive/warehouse/jc8017.db/analytic_mean from hdfs://dumbo/user/hive/warehouse/jc8017.db/.hive-staging_hive_2020-12-02_14-37-04_557_5305732271166368159-1269/-ext-10001
INFO : Starting task [Stage-3:DDL] in serial mode
INFO : Starting task [Stage-2:STATS] in serial mode
INFO : Table jc8017.analytic_mean stats: [numFiles=1, numRows=1, totalSize=317, rawDataSize=316]
INFO : MapReduce Jobs Launched:
INFO : Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.65 sec HDFS Read: 602596 HDFS Write: 394 SUCCESS
INFO : Total MapReduce CPU Time Spent: 6 seconds 650 msec
INFO : Completed executing command(queryId=hive_20201202143737_f41f8db9-286c-4fc0-952b-c695c389cc91); Time taken: 99.947 seconds
INFO : OK
No rows affected (100.449 seconds)

```

3. Create analytic normalized (column data/ avg(column))

```
0: jdbc:hive2://babar.es.its.nyu.edu:10000/> CREATE TABLE analytic_normalized
. . . . . AS SELECT
. . . . . state as state,
. . . . . county as county,
. . . . . bridges/avg(bridges) over () bridges_normalized,
. . . . . residents/avg(residents) over () residents_normalized,
. . . . . pct_medium_bridges/avg(pct_medium_bridges) over () pct_medium_bridges_no
rmalized,
. . . . . pct_poor_bridges/avg(pct_poor_bridges) over () pct_poor_bridges_normaliz
ed,
. . . . . miles_freight_railroad/avg(miles_freight_railroad) over () miles_freight
_railroad_normalized,
. . . . . roads_acceptable/avg(roads_acceptable) over () roads_acceptable_normaliz
ed,
. . . . . county_area/avg(county_area) over () county_area_normalized,
. . . . . tax_respondants/avg(tax_respondants) over () tax_respondants_normalized,
. . . . . state_local_income_tax/avg(state_local_income_tax) over () state_local_i
ncome_tax_normalized,
. . . . . real_estate_tax/avg(real_estate_tax) over () real_estate_tax_normalized,
. . . . . population_served/avg(population_served) over () population_served_norma
lized,
. . . . . water_systems/avg(water_systems) over () water_systems_normalized,
. . . . . ratio_fair_to_poor/avg(ratio_fair_to_poor) over () ratio_fair_to_poor_no
rmalized,
. . . . . freight_per_sq_mile/avg(freight_per_sq_mile) over () freight_per_sq_mile
_normalized,
. . . . . water_sys_per_capita/avg(water_sys_per_capita) over () water_sys_per_cap
_normalized,
. . . . . real_estate_tax_per_sq_mile/avg(real_estate_tax_per_sq_mile) over () est
ate_tax_per_sq_normalized,
. . . . . population_density/avg(population_density) over () population_density_no
rmalized
. . . . . FROM analytic;
INFO : Compiling command(queryId=hive_20201202144141_2da2a2b0-eb02-4a3c-8850-c6b2c28ee1cc): CREATE TABLE analytic_no
rmalized
AS SELECT
state as state,
county as county,
bridges/avg(bridges) over () bridges_normalized,
residents/avg(residents) over () residents_normalized,
pct_medium_bridges/avg(pct_medium_bridges) over () pct_medium_bridges_normalized,
pct_poor_bridges/avg(pct_poor_bridges) over () pct_poor_bridges_normalized,
miles_freight_railroad/avg(miles_freight_railroad) over () miles_freight_railroad_normalized,
roads_acceptable/avg(roads_acceptable) over () roads_acceptable_normalized,
county_area/avg(county_area) over () county_area_normalized,
tax_respondants/avg(tax_respondants) over () tax_respondants_normalized,
state_local_income_tax/avg(state_local_income_tax) over () state_local_income_tax_normalized,
real_estate_tax/avg(real_estate_tax) over () real_estate_tax_normalized,
population_served/avg(population_served) over () population_served_normalized,
water_systems/avg(water_systems) over () water_systems_normalized,
ratio_fair_to_poor/avg(ratio_fair_to_poor) over () ratio_fair_to_poor_normalized,
freight_per_sq_mile/avg(freight_per_sq_mile) over () freight_per_sq_mile_normalized,
water_sys_per_capita/avg(water_sys_per_capita) over () water_sys_per_cap_normalized,
real_estate_tax_per_sq_mile/avg(real_estate_tax_per_sq_mile) over () estate_tax_per_sq_normalized,
population_density/avg(population_density) over () population_density_normalized
FROM analytic
```



```

FROM analytic
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:state, type:string, comment:null), FieldSchema(name:county, type:string, comment:null), FieldSchema(name:bridges_normalized, type:double, comment:null), FieldSchema(name:residents_normalized, type:double, comment:null), FieldSchema(name:pct_medium_bridges_normalized, type:double, comment:null), FieldSchema(name:pct_poor_bridges_normalized, type:double, comment:null), FieldSchema(name:miles_freight_railroad_normalized, type:double, comment:null), FieldSchema(name:roads_acceptable_normalized, type:double, comment:null), FieldSchema(name:county_area_normalized, type:double, comment:null), FieldSchema(name:tax_respondants_normalized, type:double, comment:null), FieldSchema(name:state_local_income_tax_normalized, type:double, comment:null), FieldSchema(name:real_estate_tax_normalized, type:double, comment:null), FieldSchema(name:population_served_normalized, type:double, comment:null), FieldSchema(name:water_systems_normalized, type:double, comment:null), FieldSchema(name:ratio_fair_to_poor_normalized, type:double, comment:null), FieldSchema(name:freight_per_sq_mile_normalized, type:double, comment:null), FieldSchema(name:water_sys_per_cap_normalized, type:double, comment:null), FieldSchema(name:estate_tax_per_sq_normalized, type:double, comment:null), FieldSchema(name:population_density_normalized, type:double, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20201202144141_2da2a2b0-eb02-4a3c-8850-c6b2c28ee1cc); Time taken: 0.331 seconds
INFO : Executing command(queryId=hive_20201202144141_2da2a2b0-eb02-4a3c-8850-c6b2c28ee1cc): CREATE TABLE analytic_normalized
AS SELECT
state as state,
county as county,
bridges/avg(bridges) over () bridges_normalized,
residents/avg(residents) over () residents_normalized,
pct_medium_bridges/avg(pct_medium_bridges) over () pct_medium_bridges_normalized,
pct_poor_bridges/avg(pct_poor_bridges) over () pct_poor_bridges_normalized,
miles_freight_railroad/avg(miles_freight_railroad) over () miles_freight_railroad_normalized,
roads_acceptable/avg(roads_acceptable) over () roads_acceptable_normalized,
county_area/avg(county_area) over () county_area_normalized,
tax_respondants/avg(tax_respondants) over () tax_respondants_normalized,
state_local_income_tax/avg(state_local_income_tax) over () state_local_income_tax_normalized,
real_estate_tax/avg(real_estate_tax) over () real_estate_tax_normalized,
population_served/avg(population_served) over () population_served_normalized,
water_systems/avg(water_systems) over () water_systems_normalized,
ratio_fair_to_poor/avg(ratio_fair_to_poor) over () ratio_fair_to_poor_normalized,
freight_per_sq_mile/avg(freight_per_sq_mile) over () freight_per_sq_mile_normalized,
water_sys_per_capita/avg(water_sys_per_capita) over () water_sys_per_cap_normalized,
real_estate_tax_per_sq_mile/avg(real_estate_tax_per_sq_mile) over () estate_tax_per_sq_normalized,
population_density/avg(population_density) over () population_density_normalized
FROM analytic
INFO : Query ID = hive_20201202144141_2da2a2b0-eb02-4a3c-8850-c6b2c28ee1cc
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Number of reduce tasks not specified. Estimated from input data size: 1
INFO : In order to change the average load for a reducer (in bytes):
INFO :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO : In order to limit the maximum number of reducers:
INFO :   set hive.exec.reducers.max=<number>
INFO : In order to set a constant number of reducers:
INFO :   set mapreduce.job.reduces=<number>
INFO : number of splits:1
INFO : Submitting tokens for job: job_1604346392376_7461
INFO : The url to track the job: http://babar.es.its.nyu.edu:8088/proxy/application_1604346392376_7461/
INFO : Starting Job = job_1604346392376_7461, Tracking URL = http://babar.es.its.nyu.edu:8088/proxy/application_1604346392376_7461/

```

```

INFO : Kill Command = /opt/cloudera/parcels/CDH-5.15.2-1.cdh5.15.2.p0.3/lib/hadoop/bin/hadoop job -kill job_1604346392376_7461
INFO : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO : 2020-12-02 14:41:58,212 Stage-1 map = 0%, reduce = 0%
INFO : 2020-12-02 14:42:05,386 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.82 sec
INFO : 2020-12-02 14:42:13,611 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.59 sec
INFO : MapReduce Total cumulative CPU time: 8 seconds 590 msec
INFO : Ended Job = job_1604346392376_7461
INFO : Starting task [Stage-0:MOVE] in serial mode
INFO : Moving data to: hdfs://dumbo/user/hive/warehouse/jc8017.db/analytic_normalized from hdfs://dumbo/user/hive/warehouse/jc8017.db/.hive-staging_hive_2020-12-02_14-41-41_180_7452995436156277979-1269/-ext-10001
INFO : Starting task [Stage-3:DDL] in serial mode
INFO : Starting task [Stage-2:STATS] in serial mode
INFO : Table jc8017.analytic_normalized stats: [numFiles=1, numRows=3130, totalSize=1045826, rawDataSize=1042696]
INFO : MapReduce Jobs Launched:
INFO : Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.59 sec HDFS Read: 607244 HDFS Write: 1045915 SUCCESS
INFO : Total MapReduce CPU Time Spent: 8 seconds 590 msec
INFO : Completed executing command(queryId=hive_20201202144141_2da2a2b0-eb02-4a3c-8850-c6b2c28ee1cc); Time taken: 35.429 seconds
INFO : OK
No rows affected (35.913 seconds)

```

