

Data Preparation for Modelling

1. Non-event Dataset : Train and brand devices

1.1 non_event_data_external table creation

create external table if not exists non_event_data_external (device_id string, phone_brand string, device_model string, gender string, age int, group_train string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile;

```
0: jdbc:hive2://localhost:10000/default> create external table if not exists non_event_data_external (device_id string, phone_brand string, device_model string, gender string, age int, group_train string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile;
INFO : Compiling command(queryId=hive_20211014170542_82a17afe-3ba7-44dc-979d-1b60faa915ee): create external table if not exists non_event_data_external (device_id string, phone_brand string, device_model string, gender string, age int, group_train string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO : Completed compiling command(queryId=hive_20211014170542_82a17afe-3ba7-44dc-979d-1b60faa915ee); Time taken: 0.017 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20211014170542_82a17afe-3ba7-44dc-979d-1b60faa915ee): create external table if not exists non_event_data_external (device_id string, phone_brand string, device_model string, gender string, age int, group_train string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20211014170542_82a17afe-3ba7-44dc-979d-1b60faa915ee); Time taken: 0.043 seconds
INFO : OK
No rows affected (0.068 seconds)
0: jdbc:hive2://localhost:10000/default> insert overwrite table non_event_data_external
...
...-> from brand_device_external br
...-> inner join train_external tr
...-> on tr.device_id = br.device_id;
INFO : Compiling command(queryId=hive_20211014170557_ab2eae0-eddf-4011-b5ea-f10d02745987): insert overwrite table non_event_data_external
select tr.device_id, br.phone_brand, br.device_model, tr.gender, tr.age, tr.group_train
from brand_device_external br
inner join train_external tr
on tr.device_id = br.device_id
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tr.device_id, type:string, comment:null), FieldSchema(name:br.phone_brand, type:string, comment:null), FieldSchema(name:br.device_model, type:string, comment:null), FieldSchema(name:tr.gender, type:string, comment:null), FieldSchema(name:tr.age, type:int, comment:null), FieldSchema(name:tr.group_train, type:string, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20211014170557_ab2eae0-eddf-4011-b5ea-f10d02745987); Time taken: 0.133 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20211014170557_ab2eae0-eddf-4011-b5ea-f10d02745987): insert overwrite table non_event_data_external
select tr.device_id, br.phone_brand, br.device_model, tr.gender, tr.age, tr.group_train
from brand_device_external br
inner join train_external tr
on tr.device_id = br.device_id
INFO : Query ID = hive_20211014170557_ab2eae0-eddf-4011-b5ea-f10d02745987
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Session is already open
INFO : Dag name: insert overwrite table non_ev...br.device_id(Stage-1)
INFO : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO : Status: Running (Executing on YARN cluster with App id application_1634228715435_0005)

INFO : Map 1: 0/1      Map 2: 0/1
INFO : Map 1: 0(+1)/1  Map 2: 0(+1)/1
INFO : Map 1: 0(+1)/1  Map 2: 1/1
INFO : Map 1: 1/1      Map 2: 1/1
INFO : Starting task [Stage-2:DEPENDENCY_COLLECTION] in serial mode
INFO : Starting task [Stage-0:MOVE] in serial mode
INFO : Loading data to table mlctest.non_event_data_external from hdfs://ip-172-31-75-0.ec2.internal:8020/user/hive/warehouse/mlctest.db/non_event_data_external/.hive-staging_hive_2021-10-14_17-05-57_9
81.388509210722140907-2/-ext-10000
INFO : Starting task [Stage-3:STATS] in serial mode
INFO : Completed executing command(queryId=hive_20211014170557_ab2eae0-eddf-4011-b5ea-f10d02745987); Time taken: 5.55 seconds
INFO : OK
No rows affected (5.694 seconds)
0: jdbc:hive2://localhost:10000/default>
```

1.2 non_event_data_external table data load

```
insert overwrite table non_event_data_external
select tr.device_id, br.phone_brand, br.device_model, tr.gender, tr.age, tr.group_train
from brand_device_external br
inner join train_external tr
on tr.device_id = br.device_id;
```

```
0: jdbc:hive2://localhost:10000/default> create external table if not exists non_event_data_external (device_id string, phone_brand string, device_model string, gender string, age int, group_train string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile;
INFO : Compiling command(queryId=hive_20211014170542_82a17afe-3ba7-44dc-979d-1b60faa915ee): create external table if not exists non_event_data_external (device_id string, phone_brand string, device_model string, gender string, age int, group_train string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO : Completed compiling command(queryId=hive_20211014170542_82a17afe-3ba7-44dc-979d-1b60faa915ee); Time taken: 0.017 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20211014170542_82a17afe-3ba7-44dc-979d-1b60faa915ee): create external table if not exists non_event_data_external (device_id string, phone_brand string, device_model string, gender string, age int, group_train string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile
INFO : Starting task [Stage-0:DOL] in serial mode
INFO : Completed executing command(queryId=hive_20211014170542_82a17afe-3ba7-44dc-979d-1b60faa915ee); Time taken: 0.043 seconds
INFO : OK
No rows affected (0.068 seconds)
0: jdbc:hive2://localhost:10000/default> insert overwrite table non_event_data_external
. . . . .> select tr.device_id, br.phone_brand, br.device_model, tr.gender, tr.age, tr.group_train
. . . . .> from brand_device_external br
. . . . .> inner join train_external tr
. . . . .> on tr.device_id = br.device_id;
INFO : Compiling command(queryId=hive_20211014170557_ab2eae80-eddf-4011-b5ea-f10d02745987): insert overwrite table non_event_data_external
select tr.device_id, br.phone_brand, br.device_model, tr.gender, tr.age, tr.group_train
from brand_device_external br
inner join train_external tr
on tr.device_id = br.device_id
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tr.device_id, type:string, comment:null), FieldSchema(name:br.phone_brand, type:string, comment:null), FieldSchema(name:br.device_model, type:string, comment:null), FieldSchema(name:tr.gender, type:string, comment:null), FieldSchema(name:tr.age, type:int, comment:null), FieldSchema(name:tr.group_train, type:string, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20211014170557_ab2eae80-eddf-4011-b5ea-f10d02745987); Time taken: 0.133 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20211014170557_ab2eae80-eddf-4011-b5ea-f10d02745987): insert overwrite table non_event_data_external
select tr.device_id, br.phone_brand, br.device_model, tr.gender, tr.age, tr.group_train
from brand_device_external br
inner join train_external tr
on tr.device_id = br.device_id
INFO : Query ID = hive_20211014170557_ab2eae80-eddf-4011-b5ea-f10d02745987
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Session is already open
INFO : Dag name: Insert overwrite table non_ev...br.device_id[Stage-1]
INFO : Setting task.scale.memory-reserve-fraction to 0.30000001192092896
INFO : Status: Running (Executing on YARN cluster with App id application_1634228715435_0005)

INFO : Map 1: 0/1 Map 2: 0/1
INFO : Map 1: 0(1)/1 Map 2: 0(1)/1
INFO : Map 1: 0(1)/1 Map 2: 1/1
INFO : Map 1: 1/1 Map 2: 1/1
INFO : Starting task [Stage-2:DEPENDENCY_COLLECTION] in serial mode
INFO : Starting task [Stage-0:MOVE] in serial mode
INFO : Loading data to table mlctest.non_event_data_external from hdfs://ip-172-31-75-0.ec2.internal:8020/user/hive/warehouse/mlctest.db/non_event_data_external/.hive-staging_hive_2021-10-14_17-05-57_9
81.3853992107223140907-2/-ext-10000
INFO : Starting task [Stage-3:STATS] in serial mode
INFO : Completed executing command(queryId=hive_20211014170557_ab2eae80-eddf-4011-b5ea-f10d02745987); Time taken: 5.55 seconds
INFO : OK
No rows affected (5.694 seconds)
0: jdbc:hive2://localhost:10000/default> 
```

1.3 non_event_data_external table data count - 74840

```
0: jdbc:hive2://localhost:10000/default> select count(*) from non_event_data_external;
INFO : Compiling command(queryId=hive_20211014170645_a6a06ea6-3cdd-422a-a9bf-16b388a30516): select count(*) from non_event_data_external
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:c0, type:bigint, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20211014170645_a6a06ea6-3cdd-422a-a9bf-16b388a30516); Time taken: 0.12 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20211014170645_a6a06ea6-3cdd-422a-a9bf-16b388a30516): select count(*) from non_event_data_external
INFO : Completed executing command(queryId=hive_20211014170645_a6a06ea6-3cdd-422a-a9bf-16b388a30516); Time taken: 0.001 seconds
INFO : OK
+-----+
|_c0|
+-----+
| 74840 |
+-----+
1 row selected (0.134 seconds)
0: jdbc:hive2://localhost:10000/default> 
```

2.Event Dataset : Train and event

2.1 event_train_external table creation

create external table if not exists events_train_external (device_id string, event_id int, event_time timestamp, latitude float, longitude float, gender string, age int, group_train string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile;

```
0: jdbc:hive2://localhost:10000/default> create external table if not exists events_train_external (device_id string, event_id int, event_time timestamp, latitude float, longitude float, gender string, age int, group_train string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile;
INFO : Compiling command(queryId=hive_20211014170715_70dccc77b-1d14-4170-a10a-b2db0640a85d): create external table if not exists events_train_external (device_id string, event_id int, event_time timestamp, latitude float, longitude float, gender string, age int, group_train string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO : Completed compiling command(queryId=hive_20211014170715_70dccc77b-1d14-4170-a10a-b2db0640a85d): create external table if not exists events_train_external (device_id string, event_id int, event_time timestamp, latitude float, longitude float, gender string, age int, group_train string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20211014170715_70dccc77b-1d14-4170-a10a-b2db0640a85d): create external table if not exists events_train_external (device_id string, event_id int, event_time timestamp, latitude float, longitude float, gender string, age int, group_train string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile
INFO : Starting task [Stage=0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20211014170715_70dccc77b-1d14-4170-a10a-b2db0640a85d): create external table if not exists events_train_external (device_id string, event_id int, event_time timestamp, latitude float, longitude float, gender string, age int, group_train string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile
INFO : OK
No rows affected (0.055 seconds)
0: jdbc:hive2://localhost:10000/default> insert overwrite table events_train_external
...> select ev.device_id, ev.event_id, ev.event_time, ev.latitude, ev.longitude,
...> tr.gender, tr.age, tr.group_train
...> from events_external ev
...> inner join train_external tr
...> on ev.device_id = tr.device_id
INFO : Compiling command(queryId=hive_20211014170725_f333b9f4-6b99-4829-8f3d-d324ac03a26e): insert overwrite table events_train_external
select ev.device_id, ev.event_id, ev.event_time, ev.latitude, ev.longitude,
tr.gender, tr.age, tr.group_train
from events_external ev
inner join train_external tr
on ev.device_id = tr.device_id
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name=ev.device_id, type:string, comment:null), FieldSchema(name=ev.event_id, type:int, comment:null), FieldSchema(name=ev.event_time, type=timestamp, comment:null), FieldSchema(name=ev.latitude, type=float, comment:null), FieldSchema(name=ev.longitude, type=float, comment:null), FieldSchema(name=tr.gender, type:string, comment:null), FieldSchema(name=tr.age, type=int, comment:null), FieldSchema(name=tr.group_train, type:string, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20211014170725_f333b9f4-6b99-4829-8f3d-d324ac03a26e): insert overwrite table events_train_external
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20211014170725_f333b9f4-6b99-4829-8f3d-d324ac03a26e): insert overwrite table events_train_external
select ev.device_id, ev.event_id, ev.event_time, ev.latitude, ev.longitude,
tr.gender, tr.age, tr.group_train
from events_external ev
inner join train_external tr
on ev.device_id = tr.device_id
INFO : Query ID = hive_20211014170725_f333b9f4-6b99-4829-8f3d-d324ac03a26e
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage=1:MAPRED] in serial mode
INFO : Session is already open
INFO : Dag Name: Insert overwrite table events...tr.device_id(Stage=1)
INFO : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192892896
INFO : Status: Running (Executing on YARN cluster with App id application_1634228715435_0005)

INFO : Map 1: 0/11 Map 2: 0/1
INFO : Map 1: 0(+1)/11 Map 2: 0/1
INFO : Map 1: 0(+2)/11 Map 2: 0(+1)/1
INFO : Map 1: 0(+3)/11 Map 2: 0(+1)/1
INFO : Map 1: 0(+5)/11 Map 2: 0(+1)/1
INFO : Map 1: 0(+6)/11 Map 2: 0(+1)/1
INFO : Map 1: 0(+7)/11 Map 2: 1/1
INFO : Map 1: 0(+7)/11 Map 2: 1/1
INFO : Map 1: 2(+7)/11 Map 2: 1/1
INFO : Map 1: 3(+7)/11 Map 2: 1/1
INFO : Map 1: 4(+6)/11 Map 2: 1/1
INFO : Map 1: 6(+4)/11 Map 2: 1/1
INFO : Map 1: 8(+3)/11 Map 2: 1/1
INFO : Map 1: 9(+2)/11 Map 2: 1/1
INFO : Map 1: 10(+1)/11 Map 2: 1/1
INFO : Map 1: 11/11 Map 2: 1/1
```

2.2 event_train_external table data load

```
0: jdbc:hive2://localhost:10000/default> create external table if not exists events_train_external (device_id string, event_id int, event_time timestamp, latitude float, longitude float, gender string, age int, group_train string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile;
INFO : Compiling command(queryId=hive_20211014170715_70dcc77b-1d14-4170-a10a-b2db0640a85d): create external table if not exists events_train_external (device_id string, event_id int, event_time timestamp, latitude float, longitude float, gender string, age int, group_train string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO : Completed compiling command(queryId=hive_20211014170715_70dcc77b-1d14-4170-a10a-b2db0640a85d); Time taken: 0.017 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20211014170715_70dcc77b-1d14-4170-a10a-b2db0640a85d): create external table if not exists events_train_external (device_id string, event_id int, event_time timestamp, latitude float, longitude float, gender string, age int, group_train string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20211014170715_70dcc77b-1d14-4170-a10a-b2db0640a85d); Time taken: 0.033 seconds
INFO : OK
No rows affected (0.055 seconds)
0: jdbc:hive2://localhost:10000/default> insert overwrite table events_train_external
... ..> select ev.device_id, ev.event_id, ev.event_time, ev.latitude, ev.longitude,
... ..> tr.gender, tr.age, tr.group_train
... ..> from events_external ev
... ..> inner join train_external tr
... ..> on ev.device_id = tr.device_id;
INFO : Compiling command(queryId=hive_20211014170725_f333b9f4-6b99-4829-8f3d-d324ac03a26e): insert overwrite table events_train_external
select ev.device_id, ev.event_id, ev.event_time, ev.latitude, ev.longitude,
tr.gender, tr.age, tr.group_train
from events_external ev
inner join train_external tr
on ev.device_id = tr.device_id
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name=ev.device_id, type:string, comment:null), FieldSchema(name=ev.event_id, type:int, comment:null), FieldSchema(name=ev.event_time, type:timestamp, comment:null), FieldSchema(name=ev.latitude, type:float, comment:null), FieldSchema(name=ev.longitude, type:float, comment:null), FieldSchema(name=tr.gender, type:string, comment:null), FieldSchema(name=tr.age, type:int, comment:null), FieldSchema(name=tr.group_train, type:string, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20211014170725_f333b9f4-6b99-4829-8f3d-d324ac03a26e); Time taken: 0.134 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20211014170725_f333b9f4-6b99-4829-8f3d-d324ac03a26e): insert overwrite table events_train_external
select ev.device_id, ev.event_id, ev.event_time, ev.latitude, ev.longitude,
tr.gender, tr.age, tr.group_train
from events_external ev
inner join train_external tr
on ev.device_id = tr.device_id
INFO : Query ID = hive_20211014170725_f333b9f4-6b99-4829-8f3d-d324ac03a26e
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Session is already open
INFO : Dag name: insert overwrite table events...tr.device_id(Stage-1)
INFO : Setting io.task.scale.memory.reserve-fraction to 0.3000000192092896
INFO : Status: Running (Executing on YARN cluster with App id application_1634228715435_0005)

INFO : Map 1: 0/11 Map 2: 0/1
INFO : Map 1: 0(+1)/11 Map 2: 0/1
INFO : Map 1: 0(+2)/11 Map 2: 0(+1)/1
INFO : Map 1: 0(+3)/11 Map 2: 0(+1)/1
INFO : Map 1: 0(+5)/11 Map 2: 0(+1)/1
INFO : Map 1: 0(+6)/11 Map 2: 0(+1)/1
INFO : Map 1: 0(+7)/11 Map 2: 1/1
INFO : Map 1: 0(+7)/11 Map 2: 1/1
INFO : Map 1: 2(+7)/11 Map 2: 1/1
INFO : Map 1: 3(+7)/11 Map 2: 1/1
INFO : Map 1: 4(+6)/11 Map 2: 1/1
INFO : Map 1: 4(+4)/11 Map 2: 2/1
INFO : Map 1: 8(+3)/11 Map 2: 1/1
INFO : Map 1: 9(+2)/11 Map 2: 1/1
INFO : Map 1: 10(+1)/11 Map 2: 1/1
INFO : Map 1: 11/11 Map 2: 1/1
```

```
insert overwrite table events_train_external
select ev.device_id, ev.event_id, ev.event_time, ev.latitude, ev.longitude,
tr.gender, tr.age, tr.group_train
from events_external ev
inner join train_external tr
on ev.device_id = tr.device_id;
```

2.3 event_train_external table data count - 1215598

```
0: jdbc:hive2://localhost:10000/default> select count(*) from events_train_external;
INFO : Compiling command(queryId=hive_20211014170818_55b1454c-0d1f-4291-b5ec-b2d28a35e893): select count(*) from events_train_external
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name=c0, type:bigint, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20211014170818_55b1454c-0d1f-4291-b5ec-b2d28a35e893); Time taken: 0.112 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20211014170818_55b1454c-0d1f-4291-b5ec-b2d28a35e893): select count(*) from events_train_external
INFO : Completed executing command(queryId=hive_20211014170818_55b1454c-0d1f-4291-b5ec-b2d28a35e893); Time taken: 0.002 seconds
INFO : OK
+-----+
| _c0 |
+-----+
| 1215598 |
+-----+
1 row selected (0.12 seconds)
0: jdbc:hive2://localhost:10000/default>
```

3. App Data : app_events, app_labels and label_categories

3.1 app_data_external table creation

```
create external table if not exists app_data_external (event_id int, app_id string, is_installed int, is_active
int, label_id int, category string) row format delimited fields terminated by "," lines terminated by "\n"
stored as textfile;
```

[illegible]

3.2 app_data_external table data load

```
insert overwrite table app_data_external
```

```
select app_eve.event_id, app_eve.app_id, app_eve.is_installed, app_eve.is_active, lbl.label_id, lbl.category
from app_events_external app_eve
join app_labels_external app_lbl
on app_eve.app_id = app_lbl.app_id
join label_categories_external lbl
on lbl.label_id = app_lbl.label_id;
```

[illegible]

3.3 app_data_external table data count - 209355710

```
00: jdbc:hive2://localhost:10000/default> select count(*) from app_data_external;
INFO : Compiling command(queryId=hive_20211014171221_e5e862dc-e7e0-4755-8f1a-f647b8de33bc): select count(*) from app_data_external
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name=_c0, type=bigint, comment=null)], properties:null)
INFO : Completed compiling command(queryId=hive_20211014171221_e5e862dc-e7e0-4755-8f1a-f647b8de33bc); Time taken: 0.117 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20211014171221_e5e862dc-e7e0-4755-8f1a-f647b8de33bc): select count(*) from app_data_external
INFO : Completed executing command(queryId=hive_20211014171221_e5e862dc-e7e0-4755-8f1a-f647b8de33bc); Time taken: 0.001 seconds
INFO : OK

+-----+
|      _c0      |
+-----+
| 209355710     |
+-----+

1 row selected (0.13 seconds)
00: jdbc:hive2://localhost:10000/default>
```

4. CSV file creation from external tables

4.1 Commands

```
hive -e 'set hive.cli.print.header=true; select * from mlctest.non_event_data_external' | sed 's/[\\t],/g' > /home/hadoop/non_events.csv;
hive -e 'set hive.cli.print.header=true; select * from mlctest.events_train_external' | sed 's/[\\t],/g' > /home/hadoop/events.csv;
hive -e 'set hive.cli.print.header=true; select * from mlctest.app_data_external' | sed 's/[\\t],/g' > /home/hadoop/appdata.csv;
```

```
[hadoop@ip-172-31-75-0 ~]$ hive -e 'set hive.cli.print.header=true; select * from mlctest.non_event_data_external' | sed 's/[\\t],/g' > /home/hadoop/non_events.csv;
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
OK
Time taken: 2.471 seconds, Fetched: 74840 row(s)
[hadoop@ip-172-31-75-0 ~]$ hive -e 'set hive.cli.print.header=true; select * from mlctest.events_train_external' | sed 's/[\\t],/g' > /home/hadoop/events.csv;
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
OK
Time taken: 2.245 seconds, Fetched: 1215598 row(s)
[hadoop@ip-172-31-75-0 ~]$ hive -e 'set hive.cli.print.header=true; select * from mlctest.app_data_external' | sed 's/[\\t],/g' > /home/hadoop/appdata.csv;
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
OK
^[[3]+ Stopped
[hadoop@ip-172-31-75-0 ~]$ hive -e 'set hive.cli.print.header=true; select * from mlctest.app_data_external' | sed 's/[\\t],/g' > /home/hadoop/appdata.csv;
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
OK
sed: couldn't write 43 items to stdout: No space left on device
Time taken: 2.24 seconds, Fetched: 55616200 row(s)
[hadoop@ip-172-31-75-0 ~]$
```

5. Transfer CSV Data Files to S3 Bucket

5.1 Commands

```
aws s3 cp non_events.csv s3://capstone-sanjaykarthik/non_events.csv;
aws s3 cp events.csv s3://capstone-sanjaykarthik/events.csv;
aws s3 cp appdata.csv s3://capstone-sanjaykarthik/appdata.csv;
```

```
[hadoop@ip-172-31-75-0 ~]$ aws s3 cp non_events.csv s3://capstone-sanjanameghna/non_events.csv;
upload: ./non_events.csv to s3://capstone-sanjanameghna/non_events.csv
[hadoop@ip-172-31-75-0 ~]$ aws s3 cp events.csv s3://capstone-sanjanameghna/events.csv;
upload: ./events.csv to s3://capstone-sanjanameghna/events.csv
[hadoop@ip-172-31-75-0 ~]$ aws s3 cp appdata.csv s3://capstone-sanjanameghna/appdata.csv;
upload: ./appdata.csv to s3://capstone-sanjanameghna/appdata.csv
[hadoop@ip-172-31-75-0 ~]$
```

capstone-sanjanameghna [Info](#)

[Objects](#) | [Properties](#) | [Permissions](#) | [Metrics](#) | [Management](#) | [Access Points](#)

Objects (3)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#)

[Upload](#)

< 1 > [Settings](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	appdata.csv	csv	October 14, 2021, 23:02:09 (UTC+05:30)	2.5 GB	Standard
<input type="checkbox"/>	events.csv	csv	October 14, 2021, 23:02:00 (UTC+05:30)	81.5 MB	Standard
<input type="checkbox"/>	non_events.csv	csv	October 14, 2021, 23:01:47 (UTC+05:30)	3.4 MB	Standard

