

Hive

Create table for partitioning , bucketing & partitioning with bucketing for project

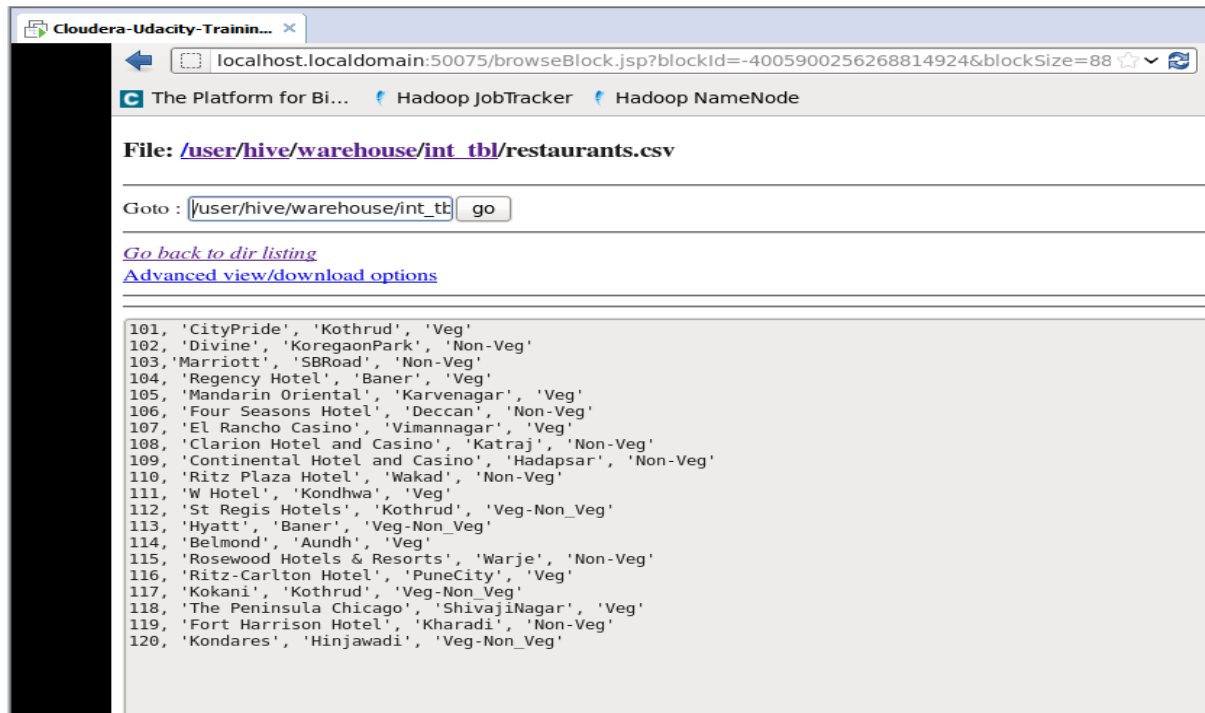
1.create table insert data in operating system

```
[root@localhost ~]# cat restaurants.csv
101, 'CityPride', 'Kothrud', 'Veg'
102, 'Divine', 'KoregaonPark', 'Non-Veg'
103, 'Marriott', 'SBRoad', 'Non-Veg'
104, 'Regency Hotel', 'Baner', 'Veg'
105, 'Mandarin Oriental', 'Karvenagar', 'Veg'
106, 'Four Seasons Hotel', 'Deccan', 'Non-Veg'
107, 'El Rancho Casino', 'Vimannagar', 'Veg'
108, 'Clarion Hotel and Casino', 'Katraj', 'Non-Veg'
109, 'Continental Hotel and Casino', 'Hadapsar', 'Non-Veg'
110, 'Ritz Plaza Hotel', 'Wakad', 'Non-Veg'
111, 'W Hotel', 'Kondhwa', 'Veg'
112, 'St Regis Hotels', 'Kothrud', 'Veg-Non_Veg'
113, 'Hyatt', 'Baner', 'Veg-Non_Veg'
114, 'Belmond', 'Aundh', 'Veg'
115, 'Rosewood Hotels & Resorts', 'Warje', 'Non-Veg'
116, 'Ritz-Carlton Hotel', 'PuneCity', 'Veg'
117, 'Kokani', 'Kothrud', 'Veg-Non_Veg'
118, 'The Peninsula Chicago', 'ShivajiNagar', 'Veg'
119, 'Fort Harrison Hotel', 'Kharadi', 'Non-Veg'
120, 'Kondares', 'Hinjawadi', 'Veg-Non_Veg'
[root@localhost ~]#
```

Create internal data:-

```
Cloudera-Udacity-Train...
hive> drop table parte_tbl;
OK
Time taken: 0.435 seconds
hive> create table int_tbl(id int,rest_name string,city string,rest_type string)row format delimited fields terminated by ' ',
'lines terminated by '\n';
OK
Time taken: 0.071 seconds
hive> load data local inpath '/root/restaurants.csv' into table int_tbl;
Copying data from file:/root/restaurants.csv
Copying file: file:/root/restaurants.csv
Loading data to table default.int_tbl
OK
Time taken: 0.226 seconds
hive> select * from int_tbl;
OK
101      'CityPride'      'Kothrud'      'Veg'
102      'Divine'       'KoregaonPark' 'Non-Veg'
103      'Marriott'     'SBRoad'       'Non-Veg'
104      'Regency Hotel'   'Baner'        'Veg'
105      'Mandarin Oriental' 'Karvenagar'   'Veg'
106      'Four Seasons Hotel' 'Deccan'       'Non-Veg'
107      'El Rancho Casino' 'Vimannagar'   'Veg'
108      'Clarion Hotel and Casino' 'Katraj'       'Non-Veg'
109      'Continental Hotel and Casino' 'Hadapsar'     'Non-Veg'
110      'Ritz Plaza Hotel' 'Wakad'        'Non-Veg'
111      'W Hotel'         'Kondhwa'      'Veg'
112      'St Regis Hotels' 'Kothrud'      'Veg-Non_Veg'
113      'Hyatt'           'Baner'        'Veg-Non_Veg'
114      'Belmond'         'Aundh'        'Veg'
115      'Rosewood Hotels & Resorts' 'Warje'        'Non-Veg'
116      'Ritz-Carlton Hotel' 'PuneCity'     'Veg'
117      'Kokani'          'Kothrud'      'Veg-Non_Veg'
118      'The Peninsula Chicago' 'ShivajiNagar' 'Veg'
119      'Fort Harrison Hotel' 'Kharadi'      'Non-Veg'
120      'Kondares'       'Hinjawadi'    'Veg-Non_Veg'
Time taken: 0.23 seconds
```

Then create table on hive (internal table):- File Location



Create external table on HDFS

Create directory on hdfs on operating system

```
File Edit View Search Terminal Help
[root@localhost ~]# hdfs dfs -rmkdir /project
rmkdir: `/project': Directory is not empty
[root@localhost ~]# hdfs dfs -rmkdir /project/restaurants.csv
rmkdir: `/project/restaurants.csv': Is not a directory
[root@localhost ~]# ls
Address.java    derby.log      Menu.java      practice
Customers.java  Drivers.java   Orders.java    Rating.java
DA1             file.csv       Payment.java   restaurants.csv
DANames_file.csv hello.csv      pral.java      Restaurants.java
[root@localhost ~]# cd /root
[root@localhost ~]# ls
Address.java    derby.log      Menu.java      practice
Customers.java  Drivers.java   Orders.java    Rating.java
DA1             file.csv       Payment.java   restaurants.csv
DANames_file.csv hello.csv      pral.java      Restaurants.java
[root@localhost ~]# cd /
[root@localhost /]# ls
bin  data  etc  lib  lost+found  mnt  practice  root  selinux  sys  usr
boot  dev  home  logs  media      opt  proc      sbin  srv      var
[root@localhost /]# cd
[root@localhost ~]# vi restaurants.csv
[root@localhost ~]# hdfs dfs -mkdir /prj
[root@localhost ~]# hdfs dfs -put /root/restaurants.csv /prj
[root@localhost ~]#
```

Data transfer from restaurants table to prj directory

For external table -File Location

Contents of directory /prj

Goto :

[Go to parent directory](#)

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
restaurants.csv	file	0.86 KB	1	64 MB	2023-08-30 07:44	rw-r--r--	root	supergroup

[Go back to DFS home](#)

Cloudera-Udacity-Trainin...

The Platform for Bi...Hadoop JobTrackerHadoop NameNode

File: [/prj/restaurants.csv](#)

Goto :

[Go back to dir listing](#)
[Advanced view/download options](#)

```
101, 'CityPride', 'Kothrud', 'Veg'
102, 'Divine', 'KoregaonPark', 'Non-Veg'
103, 'Marriott', 'SBRoad', 'Non-Veg'
104, 'Regency Hotel', 'Baner', 'Veg'
105, 'Mandarin Oriental', 'Karvenagar', 'Veg'
106, 'Four Seasons Hotel', 'Deccan', 'Non-Veg'
107, 'El Rancho Casino', 'Vimannagar', 'Veg'
108, 'Clarion Hotel and Casino', 'Katraj', 'Non-Veg'
109, 'Continental Hotel and Casino', 'Hadapsar', 'Non-Veg'
110, 'Ritz Plaza Hotel', 'Wakad', 'Non-Veg'
111, 'W Hotel', 'Kondhwa', 'Veg'
112, 'St Regis Hotels', 'Kothrud', 'Veg-Non_Veg'
113, 'Hyatt', 'Baner', 'Veg-Non_Veg'
114, 'Belmond', 'Aundh', 'Veg'
115, 'Rosewood Hotels & Resorts', 'Warje', 'Non-Veg'
116, 'Ritz-Carlton Hotel', 'PuneCity', 'Veg'
117, 'Kokani', 'Kothrud', 'Veg-Non_Veg'
118, 'The Peninsula Chicago', 'ShivajiNagar', 'Veg'
119, 'Fort Harrison Hotel', 'Kharadi', 'Non-Veg'
120, 'Kondares', 'Hinjawadi', 'Veg-Non_Veg'
```

Data partitioning

First we create normal table.

```
hive> create table npr_tbl(id int,rest_name string,city string,rest_type string)row format delimited fields terminated by ','
lines terminated by '\n';
OK
Time taken: 0.088 seconds
hive> select * from npr_tbl;
OK
Time taken: 0.183 seconds
hive> load data local inpath '/root/restaurants.csv'into table npr_tbl;
Copying data from file:/root/restaurants.csv
Copying file: file:/root/restaurants.csv
Loading data to table default.npr_tbl
OK
Time taken: 0.231 seconds
```

Then partitioning the table with help of normal table.

```

hive> create table parte_tbl(id int,rest_name string,city string)partitioned by(rest_type string)row format delimited fields
terminated by ','lines terminated by '\n';
OK
Time taken: 0.066 seconds
hive> insert into table parte_tbl partition(rest_type) select * from npr_tbl;
Total MapReduce jobs = 2
Launching Job 1 out of 2
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1693333757154_0005, Tracking URL = http://localhost.localdomain:8088/proxy/application_1693333757154_0005/
Kill Command = /usr/lib/hadoop/bin/hadoop job -Dmapred.job.tracker=localhost:9101 -kill job_1693333757154_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2023-08-30 07:54:16,900 Stage-1 map = 0%, reduce = 0%
2023-08-30 07:54:24,843 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.07 sec
MapReduce Total cumulative CPU time: 2 seconds 70 msec
Ended Job = job_1693333757154_0005
Ended Job = 836649352, job is filtered out (removed at runtime).
Moving data to: hdfs://localhost:8020/tmp/hive-root/hive_2023-08-30_07-54-00_036_8201568629472411001/-ext-10000
Loading data to table default.parte_tbl partition (rest_type=null)
Loading partition {rest_type= 'Veg' }
Loading partition {rest_type= 'Veg-Non-Veg' }
Loading partition {rest_type= 'Non-Veg' }
Loading partition {rest_type= 'Veg' }
Loading partition {rest_type= 'Non-Veg' }
20 Rows loaded to parte_tbl
MapReduce Jobs Launched:
Job 0: Map: 1 Cumulative CPU: 2.07 sec HDFS Read: 1079 HDFS Write: 672 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 70 msec
OK
Time taken: 26.147 seconds
hive>

```

Partitioning file location:-

Contents of directory [/user/hive/warehouse/parte_tbl](#)

Goto:

[Go to parent directory](#)

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
rest_type= %27Non-Veg%27	dir				2023-08-30 07:54	rwxr-xr-x	root	supergroup
rest_type= %27Veg%27	dir				2023-08-30 07:54	rwxr-xr-x	root	supergroup
rest_type= %27Veg-Non Veg%27	dir				2023-08-30 07:54	rwxr-xr-x	root	supergroup

[Go back to DFS home](#)

Nonveg hotels:-

File: [/user/hive/warehouse/parte_tbl/rest_type= %27Non-Veg%27/000000_0_copy_1](#)

Goto:

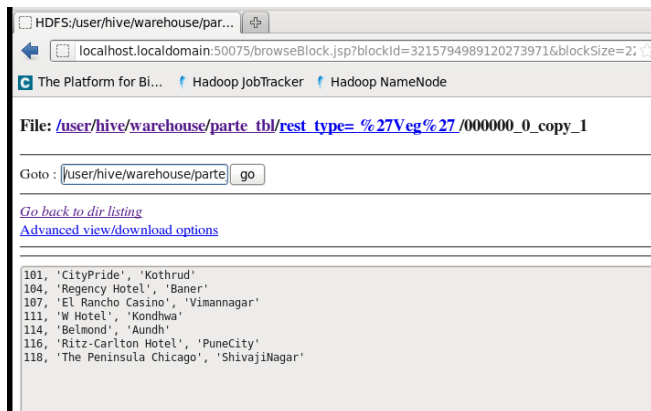
[Go back to dir listing](#)
[Advanced view/download options](#)

```

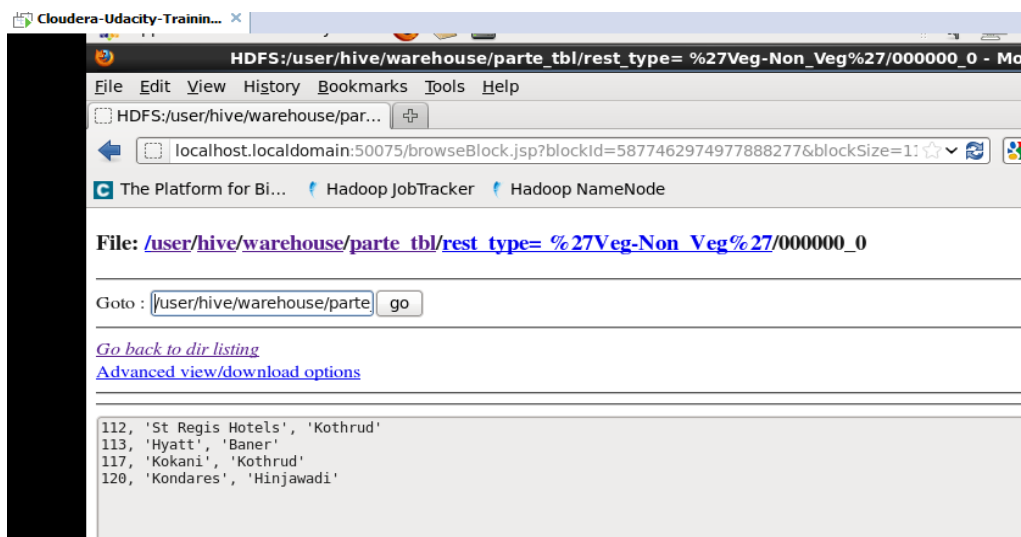
102, 'Divine', 'KoregaonPark'
109, 'Continental Hotel and Casino', 'Hadapsar'
115, 'Rosewood Hotels & Resorts', 'Warje'
119, 'Fort Harrison Hotel', 'Kharadi'

```

Veg hotels:-



Veg & Non-veg hotels:-



Bucketing the data:-

Bucketing the data on customers data

1st we create data operating system

```
[root@localhost ~]# cat customers.csv
1, 'Aditya Gawade', 'adityagawade@gmail.com'
2, 'Ashish Kondare' , 'ashishkondare@gmail.com'
3, 'Viraj Walanj' , 'virajwalanj@gmail.com'
4, 'Akash Chavan' , 'akashchavan@gmail.com'
5, 'Sourabh Gawade', 'sourabhgawade@gmail.com'
6, 'Manasi Zagade', 'manasizagade@gmail.com'
7, 'Pooja Vaddepalli', 'poojawadepalli@gmail.com'
8, 'Ashwini Khade', 'ashwinikhade@gmail.com'
9, 'Ashwini Patil', 'ashwinipatil@gmail.com'
10, 'Jitesh Deore', 'jiteshdeore@gmail.com'
[root@localhost ~]#
```

Then we transfer the data from operating system to hive

```
hive> create table npr_tbl1(id int,name string,email string)row format delimited fields terminated by ','lines terminated by '\n';
OK
Time taken: 0.089 seconds
hive> load data local inpath '/root/customers.csv'into table npr_tbl1;
Copying data from file:/root/customers.csv
Copying file: file:/root/customers.csv
Loading data to table default.npr_tbl1
OK
Time taken: 0.244 seconds
```

```
Cloudera-Udacity-Trainin...
Time taken: 22.538 seconds
hive> set hive.enforce.bucketing=true;
hive> drop table buck_tbl;
OK
Time taken: 0.157 seconds
hive> create table buck_tbl(id int,name string,email string)clustered by(id)into 2 buckets row format delimited fields termin
ated by ',';
OK
Time taken: 0.091 seconds
hive> insert into table buck_tbl select * from npr_tbl1;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 2
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 mapred.reduce.tasks=<number>
Starting Job = job_1693333757154_0009, Tracking URL = http://localhost.localdomain:8088/proxy/application_1693333757154_0009/
Kill Command = /usr/lib/hadoop/bin/hadoop job -Dmapred.job.tracker=localhost:9101 -kill job_1693333757154_0009
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 2
2023-08-30 08:47:54,028 Stage-1 map = 0%, reduce = 0%
2023-08-30 08:48:22,320 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.97 sec
2023-08-30 08:48:23,744 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 5.79 sec
2023-08-30 08:48:25,356 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 5.79 sec
2023-08-30 08:48:26,960 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 7.71 sec
2023-08-30 08:48:28,645 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 7.71 sec
2023-08-30 08:48:30,004 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 9.39 sec
MapReduce Total cumulative CPU time: 9 seconds 390 msec
Ended Job = job_1693333757154_0009
Loading data to table default.buck_tbl
10 Rows loaded to buck_tbl
MapReduce Jobs Launched:
Job 0: Map: 1 Reduce: 2 Cumulative CPU: 9.39 sec HDFS Read: 654 HDFS Write: 457 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 390 msec
```

HDFS location:-

Cloudera-Udacity-Trainin... HDFS:/user/hive/warehouse/buck_tbl - Mozilla Firefox

File Edit View History Bookmarks Tools Help

HDFS:/user/hive/warehouse/buc... localhost.localdomain:50075/browseDirectory.jsp?dir=%2Fuser%2Fhive%2Fwarehouse%2Fbuc... Google

The Platform for Bi... Hadoop JobTracker Hadoop NameNode

Contents of directory /user/hive/warehouse/buck_tbl

Goto: /user/hive/warehouse/buck go

[Go to parent directory](#)

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
000000_0	file	0.22 KB	1	64 MB	2023-08-30 08:48	rw-r--r--	root	supergroup
000001_0	file	0.23 KB	1	64 MB	2023-08-30 08:48	rw-r--r--	root	supergroup

[Go back to DFS home](#)

Bucket 1

Cloudera-Udacity-Trainin... x

THE Platform for BI... | Hadoop JobTracker | Hadoop NameNode

File: [/user/hive/warehouse/buck_tbl/000001_0](#)

Goto :

[Go back to dir listing](#)

[Advanced view/download options](#)

```
9, 'Ashwini Patil', 'ashwinipatil@gmail.com'
7, 'Pooja Vaddepalli', 'poojawadepalli@gmail.com'
5, 'Sourabh Gawade', 'sourabhgawade@gmail.com'
3, 'Viraj Walanj', 'virajwalanj@gmail.com'
1, 'Aditya Gawade', 'adityagawade@gmail.com'
```

Bucket 2:-

Cloudera-Udacity-Trainin... x

THE Platform for BI... | Hadoop JobTracker | Hadoop NameNode

File: [/user/hive/warehouse/buck_tbl/000000_0](#)

Goto :

[Go back to dir listing](#)

[Advanced view/download options](#)

```
10, 'Jitesh Deore', 'jiteshdeore@gmail.com'
8, 'Ashwini Khade', 'ashwinikhade@gmail.com'
6, 'Manasi Zagade', 'manasizagade@gmail.com'
4, 'Akash Chavan', 'akashchavan@gmail.com'
2, 'Ashish Kondare', 'ashishkondare@gmail.com'
```

Then we do both at a time partitioning & bucketing


```
root@localhost:/home/training
File Edit View Search Terminal Help
hive> create table parte_buck_tbl(id int,rest_name string,city string)partitioned by (rest_type string) clustered by(id)
o 2 buckets row format delimited fields terminated by ',';
OK
Time taken: 0.103 seconds
hive> insert into table parte_buck_tbl select * from npr_tbl;
FAILED: SemanticException 1:18 Need to specify partition columns because the destination table is partitioned. Error encou
red near token 'parte_buck_tbl'
hive> insert into table parte_buck_tbl partition (rest_type)select * from npr_tbl;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 2
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 mapred.reduce.tasks=<number>
Starting Job = job_1693333757154_0010, Tracking URL = http://localhost.localdomain:8088/proxy/application_1693333757154_00
Kill Command = /usr/lib/hadoop/bin/hadoop job -Dmapred.job.tracker=localhost:9101 -kill job_1693333757154_0010
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 2
2023-08-30 09:03:06,011 Stage-1 map = 0%, reduce = 0%
2023-08-30 09:03:28,884 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.24 sec
2023-08-30 09:03:30,101 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.24 sec
2023-08-30 09:03:31,705 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.24 sec
```

partitioning & bucketing location:- hive location

partitioning:-

Cloudera-Udacity-Trainin...

HDFS:/user/hive/warehouse/par...

localhost.localdomain:50075/browseDirectory.jsp?dir=/user/hive/warehouse/parte_buck_tbl&r

The Platform for Bi... Hadoop JobTracker Hadoop NameNode

Contents of directory /user/hive/warehouse/parte_buck_tbl

Goto : /user/hive/warehouse/parte go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
rest_type= %27Non-Veg%27	dir				2023-08-30 09:03	rwxf-r-x	root	supergroup
rest_type= %27Veg%27	dir				2023-08-30 09:03	rwxf-r-x	root	supergroup
rest_type= %27Veg-Non_Veg%27	dir				2023-08-30 09:03	rwxf-r-x	root	supergroup

Go back to DFS home

Bucketing:-

dera-Udacity-Trainin...

HDFS:/user/hive/warehouse/par...

localhost.localdomain:50075/browseDirectory.jsp?dir=%2Fuser%2Fhive%2Fwarehouse%2Fpe

The Platform for Bi... Hadoop JobTracker Hadoop NameNode

Contents of directory /user/hive/warehouse/parte_buck_tbl/rest_type= %27Non-Veg%27

Goto : /user/hive/warehouse/parte go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
000000_0	file	0.03 KB	1	64 MB	2023-08-30 09:03	rw-r--r--	root	supergroup
000000_0_copy_1	file	0.11 KB	1	64 MB	2023-08-30 09:03	rw-r--r--	root	supergroup
000001_0	file	0.12 KB	1	64 MB	2023-08-30 09:03	rw-r--r--	root	supergroup
000001_0_copy_1	file	0.02 KB	1	64 MB	2023-08-30 09:03	rw-r--r--	root	supergroup

Go back to DFS home