Hive Bucketing Concepts:

Hive will improve its performance based on the below Techniques:

- Partitioning & Bucketing
- Window Functions
- Join Optimization

Bucketing:

Bucketing splits the data into smaller subsets and make it more manageable parts. It is due to the hash function based on the column value.

Each bucket gets stored as a file under table directory it is not partitioned and if partitioned is made on the table, then bucketed files will get stored under partitioned directory in HDFS location.

Key Features:

- Buckets are stores as files in HDFS where as partitions are basically stored in as a directory
- Bucketing is the place where knowing which file each record is getting stored and it certainly makes the operation faster.
- Bucketing can be done on the column like ex:product_id since data is getting stored in terms files which is not a overhead for NameNode.
- Ex: if the buckets is determined as “3” for the bucketed column ex:product_id ,it split the data based on the operation 1%3 or 2%3 or 3%3 and it goes on **[Modulo operator on Integer column hash function]**
- Buckets are almost same size and it works based on the hash value of the column
- Fixed number of buckets and easy to know which record stored get stored in which file.
- Hive tables both partitioned and bucketed can be done
- Faster query response
- Since the column Is bucketed and due to this join operations will be too faster when you perform a join between a normal table vs bucketed table

Sample scenario on Bucketing concepts – Cloudera:

Step 1: Create a non-bucketed tables and load the data to the table with holds product information

create table products(

id int,

name string,

cost double,

category string

)

row format delimited fields terminated by ',';

```
hive> create table products(
    > id int,
    > name string,
    > cost double,
    > category string
    > )
    > row format delimited fields terminated by ',';
OK
Time taken: 6.21 seconds
```

Step 2 :

Load the data from local file system to Hive:

load data local inpath '/home/cloudera/projects/newproducts.csv' into table products;

```
hive> load data local inpath '/home/cloudera/projects/newproducts.csv' into table products;
Loading data to table default.products
Table default.products stats: [numFiles=1, totalSize=213]
OK
Time taken: 43.957 seconds
```

Step 3:

See the results once the data gets loaded

```
hive> select * from default.products;
OK
1       iPhone  379.99  mobiles
2       doll    8.99    toys
3       Galaxy X        100.0   mobile
5       Nokia Y 39.99   mobile
6       truck   7.99    toys
7       makeup  100.0   fashion
8       earings 69.0    fashion
9       chair   129.0   furniture
10      table   269.0   furniture
11      waterpistol     9.0     toys
Time taken: 0.43 seconds, Fetched: 10 row(s)
hive>
```

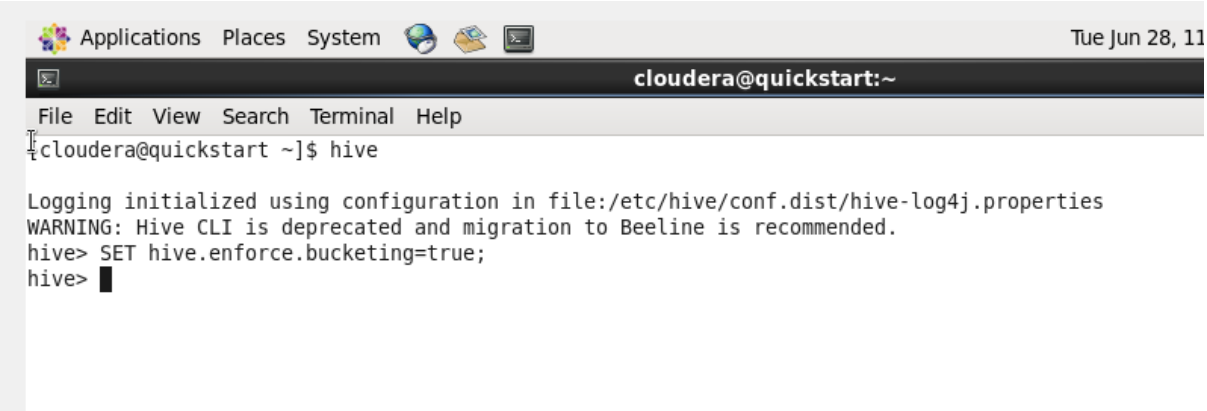
Cloudera Live : Welco... | cloudera@quickstart:~

Step 4:

Create a bucketed table with column “id” as bucketed column and load the data from non bucketed table to bucketed table.

Ensure you enable the below property for Bucketing:

SET hive.enforce.bucketing=true;




```
create table products_bucketed_info(

id int,

name string,

cost double,

category string

)

CLUSTERED BY (id) INTO 4 BUCKETS;
```

Step 5 :

Load the data from Non bucketed table which is created above with bucketed table in Hive

```
from products

insert into table products_bucketed_info select id,name,cost,category;
```

```
hive> SET hive.enforce.bucketing=true;
hive> create table products_bucketed_info(
    > id int,
    > name string,
    > cost double,
    > category string
    > )
    > CLUSTERED BY (id) INTO 4 BUCKETS;
OK
Time taken: 2.052 seconds
hive>
    > from products
    > insert into table products_bucketed_info select id,name,cost,category;
Query ID = cloudera_20220628112525_7e197555-0a0d-42dc-96ac-2fc89a908898
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 4
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1656438724556_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1656438724556_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1656438724556_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 4
2022-06-28 11:26:26,531 Stage-1 map = 0%,  reduce = 0%
2022-06-28 11:26:43,622 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.02 sec
2022-06-28 11:27:11,122 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 15.56 sec
MapReduce Total cumulative CPU time: 15 seconds 560 msec
Ended Job = job_1656438724556_0002
Loading data to table default.products_bucketed_info
Table default.products_bucketed_info stats: [numFiles=4, numRows=10, totalSize=225, rawDataSize=215]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 4   Cumulative CPU: 15.56 sec   HDFS Read: 19584 HDFS Write: 569 SUCCESS
Total MapReduce CPU Time Spent: 15 seconds 560 msec
OK
Time taken: 82.319 seconds
hive>
```

Cloudera Live : Welco... | cloudera@quickstart:~

Bucketed data in HDFS Location with 4 Buckets:

Applications Places System Tue Jun 28, 11:29 AM cloudera

cloudera@quickstart:~

File Edit View Search Terminal Help

```
[cloudera@quickstart ~]$ hadoop fs -ls /user/hive/warehouse/products_bucketed_info
Found 4 items
-rwxrwxrwx   1 cloudera supergroup         23 2022-06-28 11:27 /user/hive/warehouse/products_bucketed_info/000000_0
-rwxrwxrwx   1 cloudera supergroup         71 2022-06-28 11:27 /user/hive/warehouse/products_bucketed_info/000001_0
-rwxrwxrwx   1 cloudera supergroup         60 2022-06-28 11:27 /user/hive/warehouse/products_bucketed_info/000002_0
-rwxrwxrwx   1 cloudera supergroup         71 2022-06-28 11:27 /user/hive/warehouse/products_bucketed_info/000003_0
[cloudera@quickstart ~]$
```

To pull few sample data from Buckets :

```
hive> select * from default.products_bucketed_info TABLESAMPLE(bucket 3 out of 4);
OK
10      table   269.0   furniture
6       truck   7.99    toys
2       doll    8.99    toys
Time taken: 4.942 seconds, Fetched: 3 row(s)
hive> select * from default.products_bucketed_info TABLESAMPLE(bucket 2 out of 4);
OK
9       chair   129.0   furniture
5       Nokia Y 39.99   mobile
1       iPhone  379.99  mobiles
Time taken: 0.127 seconds, Fetched: 3 row(s)
hive>
```

Cloudera Live : Welco... | cloudera@quickstart:~

```
[cloudera@quickstart ~]$ hadoop fs -cat  /user/hive/warehouse/products_bucketed_info/000000_0
8earings69.0fashion
[cloudera@quickstart ~]$ hadoop fs -cat  /user/hive/warehouse/products_bucketed_info/000000_1
cat: `/user/hive/warehouse/products_bucketed_info/000000_1': No such file or directory
[cloudera@quickstart ~]$ hadoop fs -cat  /user/hive/warehouse/products_bucketed_info/000001_0
9chair129.0furniture
5Nokia Y39.99mobile
1iPhone379.99mobiles
[cloudera@quickstart ~]$ hadoop fs -cat  /user/hive/warehouse/products_bucketed_info/000002_0
10table269.0furniture
6truck7.99toys
2doll8.99toys
[cloudera@quickstart ~]$ hadoop fs -cat  /user/hive/warehouse/products_bucketed_info/000003_0
11waterpistol9.0toys
7makeup100.0fashion
3Galaxy X100.0mobile
[cloudera@quickstart ~]$
```

Bucketed Files in UI :