Caleb Post
Lab 6 Questions

Question 1: Hive supports the MSCK repair command. What does it do? Explain with an example.

What this command does is it rebuilds the metastore for the table adding metadata for partitions that don’t already exist. This could be used in place of the ADD PARTITION command. Take for example our manually added partition in task 3 for employees. Instead of using the commands

```
ALTER TABLE table_name ADD PARTITION ‘csse’
```

We could have run

```
MSCK REPAIR TABLE table_name
```

And instead of us running the multiple alter tables we could have done this with just the MSCK repair command.

Question 2: Explain the difference between order by, sort by commands with an example.

The difference between order by and sort by is the number of reducers that is used. Sort by uses the default number of reducers and sorts based on each reducer. Order by guarantees that there will only be one reducer so the output is completely sorted.

An example of this is running is say that there is this data that has been run through sort by, in 2 reducers,

```
    0   5
    0   3
    3   9

    0   4
    1   2
```

And we can see that it is sorted by the first column, but it is only sorted within each reducer. If we ran this through order by it would go through one reducer and give the output

```
0       5
0       3
0       4
1       2
3       9
```

Question 3: Explain the purpose of the distribute by command. What happens when you add a sort by to the output of a distribute by? Is there an equivalent command that replaces the distribute by and sort by? What is it?

The purpose of the distribute by command is to specify what column determines what elements go to which reducer. Each value in a distribute by column is guaranteed to end up at the same reducer as others with that same value. If you add a sort by to a distribute by then you can end up with a complete sorting if you are willing to piece together the results yourself. There is an equivalent command that combines the distribute by and sort by, it is called cluster by.

Question 4: Explain the following commands/concepts with an example.

a. Bucketing

   Bucketing is much like partitioning except that there are a fixed number of buckets available at table creation and no more are added. Instead, elements that are in the bucking column are hashed into the specific bucket they are to appear in. This functions like a hashtable would for elements in the bucket.

   If we had a partition therefore on department name and we add 'csse' that has not been added before we would see a new partition csse be created.

   Instead, if we had bucketed on department name, no new buckets would be created, instead, all rows with csse would be added into the proper bucket for csse.

b. UNION ALL

   Union all is the bag union that hive uses to combine different select statements into the same result set. It does not eliminate duplicates and requires the that number and names of columns returned must be the same.

   Using the example from the hive confluence pages, if we assume that there are two separate tables containing information, one in which the user has published a video and one in which a user has made a comment, if we want to determine therefore the actions taken by this individual we could run this query which combines the result on video and comment into a single bag.

```
SELECT u.id, actions.date
FROM (
    SELECT av.uid AS uid
    FROM action_video av
    WHERE av.date = '2008-06-03'
    UNION ALL
    SELECT ac.uid AS uid
    FROM action_comment ac
    WHERE ac.date = '2008-06-03'
 ) actions JOIN users u ON (u.id = actions.uid)
```