Big Data Technologies - CSP 554
Assignment 4
Prabhu Avula | A20522815
Illinois Institute of Technology, Chicago

Summary of the Article

The article addresses the crucial challenge of optimizing data organization in Big Data Warehousing (BDW) systems. With the exponential growth of data in terms of volume, variety, and velocity, traditional Data Warehouses (DWs) face significant difficulties in handling and processing this vast amount of information efficiently. Hive, a data warehousing solution built on top of Hadoop, offers a promising approach to managing large-scale datasets using SQL-like languages in distributed environments, making it a focal point for this study.

The research focuses on Hive's primary data organization strategies: partitioning and bucketing. Partitioning involves dividing a table into smaller segments based on certain column values, which can significantly reduce query processing times by limiting the data scanned during queries. Bucketing, however, splits data into more manageable parts within partitions, which can be particularly useful for optimizing join operations. The study aims to evaluate the practical benefits of these strategies and provide guidance on their optimal use.

Initial motivations for the study stem from the observation that, while partitioning and bucketing are theoretically advantageous, their real-world applications and benefits were poorly understood. The research seeks to fill this gap by conducting a series of experiments to measure the impact of these strategies on query performance in Hive-based BDWs. The ultimate goal is to offer actionable insights and best practices for data practitioners working with Hive.

The experimental setup involved various datasets and query workloads to simulate typical Big Data processing scenarios. The researchers measured query execution times and system performance with and without the use of partitioning and bucketing. By aligning partitions with frequently queried attributes, they aimed to demonstrate how strategic data organization can lead to significant performance improvements.

Results from the experiments highlighted that partitioning can substantially decrease query processing time. Specifically, when partitions were aligned with query attributes, processing times for intensive workloads decreased by up to 40%. This finding underscores the importance of understanding query patterns and aligning data organization strategies accordingly to maximize efficiency in Hive-based systems.

Bucketing, while theoretically beneficial, showed mixed results in the experiments. The study found that bucketing did not significantly improve performance across all scenarios. However, it proved useful in specific cases, particularly in optimizing join operations between large tables. This suggests that bucketing should be used selectively based on the specific requirements and query patterns of the BDW system.

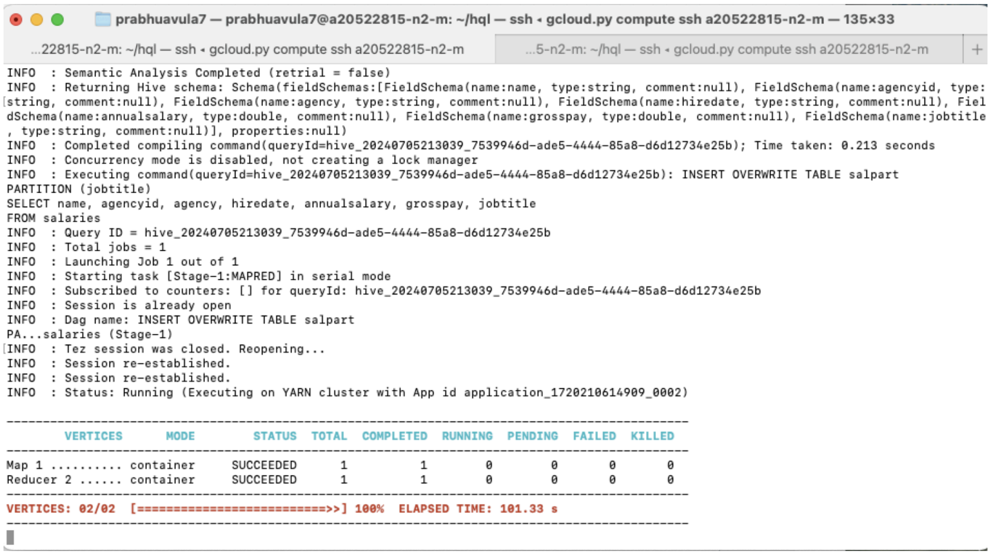
Based on these findings, the article offers practitioners a set of best practices. It recommends focusing on partitioning strategies that align with frequently used query attributes to achieve significant performance gains. It advises a more cautious and selective approach for bucketing, utilizing it primarily to enhance joint operations rather than as a general data organization strategy.

The study concludes by emphasizing the critical role of proper data organization in Big Data environments. Effective partitioning strategies can lead to marked improvements in system performance, making them a key consideration for data practitioners. While not universally beneficial, bucketing can provide targeted performance enhancements when used appropriately.

The article provides valuable insights into optimizing Hive-based BDWs through strategic data organization. By highlighting the practical benefits of partitioning and offering guidance on the selective use of bucketing, the study equips data practitioners with the knowledge to enhance the performance of their Big Data systems.

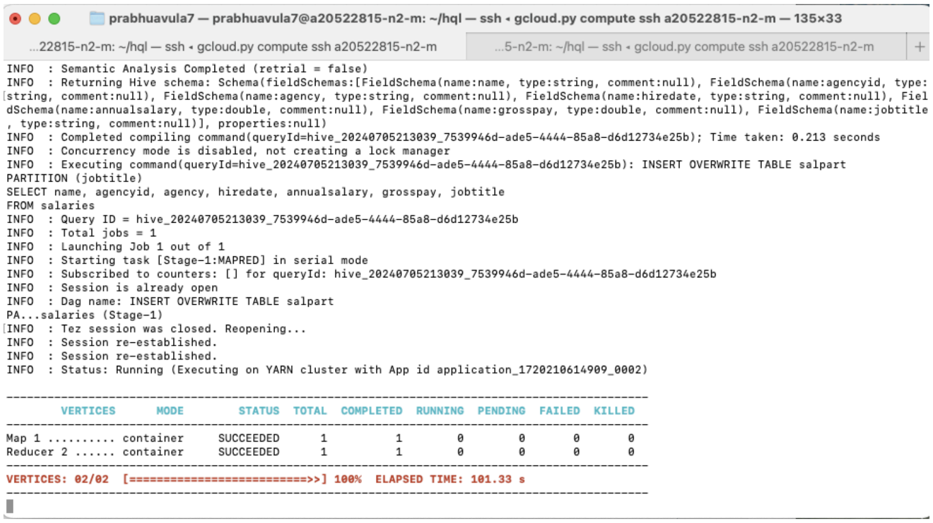
HIVE Based Coding Exercises

1.



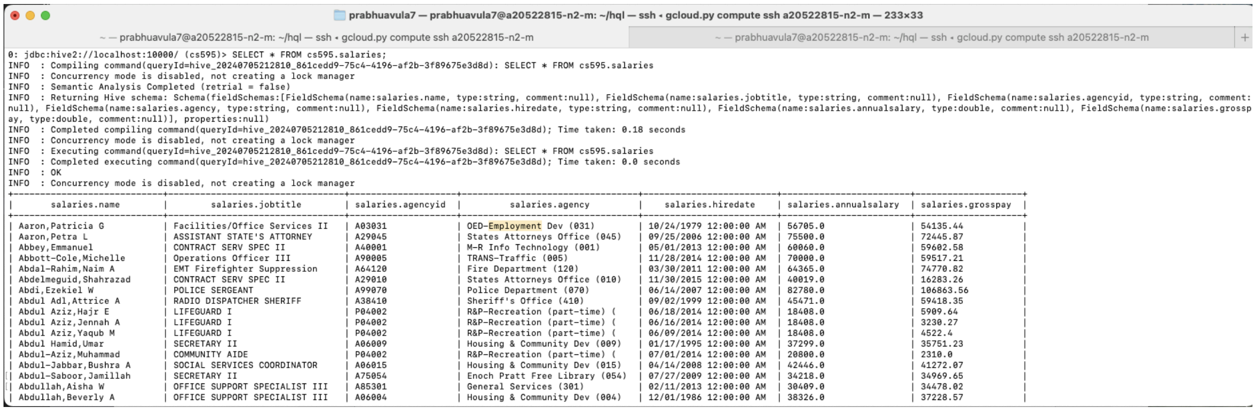
```
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:name, type:string, comment:null), FieldSchema(name:agencyid, type:
string, comment:null), FieldSchema(name:agency, type:string, comment:null), FieldSchema(name:hiredate, type:string, comment:null), Fiel
dSchema(name:annualsalary, type:double, comment:null), FieldSchema(name:grosspay, type:double, comment:null), FieldSchema(name:jobtitle
, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240705213039_7539946d-ade5-4444-85a8-d6d12734e25b); Time taken: 0.213 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240705213039_7539946d-ade5-4444-85a8-d6d12734e25b): INSERT OVERWRITE TABLE salpart
PARTITION (jobtitle)
SELECT name, agencyid, agency, hiredate, annualsalary, grosspay, jobtitle
FROM salaries
INFO  : Query ID = hive_20240705213039_7539946d-ade5-4444-85a8-d6d12734e25b
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20240705213039_7539946d-ade5-4444-85a8-d6d12734e25b
INFO  : Session is already open
INFO  : Dag name: INSERT OVERWRITE TABLE salpart
PA...salaries (Stage-1)
INFO  : Tez session was closed. Reopening...
INFO  : Session re-established.
INFO  : Session re-established.
INFO  : Status: Running (Executing on YARN cluster with App id application_1720210614909_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 101.33 s
----------------------------------------------------------------------------------------------
```


prabhuavula7 — prabhuavula7@a20522815-n2-m: ~/hql — ssh ◂ gcloud.py compute ssh a20522815-n2-m — 135×33
...22815-n2-m: ~/hql — ssh ◂ gcloud.py compute ssh a20522815-n2-m
...5-n2-m: ~/hql — ssh ◂ gcloud.py compute ssh a20522815-n2-m
+

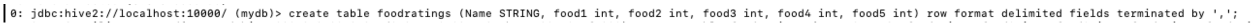
```
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:name, type:string, comment:null), FieldSchema(name:agencyid, type:
string, comment:null), FieldSchema(name:agency, type:string, comment:null), FieldSchema(name:hiredate, type:string, comment:null), Fiel
dSchema(name:annualsalary, type:double, comment:null), FieldSchema(name:grosspay, type:double, comment:null), FieldSchema(name:jobtitle
, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240705213039_7539946d-ade5-4444-85a8-d6d12734e25b); Time taken: 0.213 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240705213039_7539946d-ade5-4444-85a8-d6d12734e25b): INSERT OVERWRITE TABLE salpart
PARTITION (jobtitle)
SELECT name, agencyid, agency, hiredate, annualsalary, grosspay, jobtitle
FROM salaries
INFO  : Query ID = hive_20240705213039_7539946d-ade5-4444-85a8-d6d12734e25b
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20240705213039_7539946d-ade5-4444-85a8-d6d12734e25b
INFO  : Session is already open
INFO  : Dag name: INSERT OVERWRITE TABLE salpart
PA...salaries (Stage-1)
INFO  : Tez session was closed. Reopening...
INFO  : Session re-established.
INFO  : Session re-established.
INFO  : Status: Running (Executing on YARN cluster with App id application_1720210614909_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 101.33 s
----------------------------------------------------------------------------------------------
```

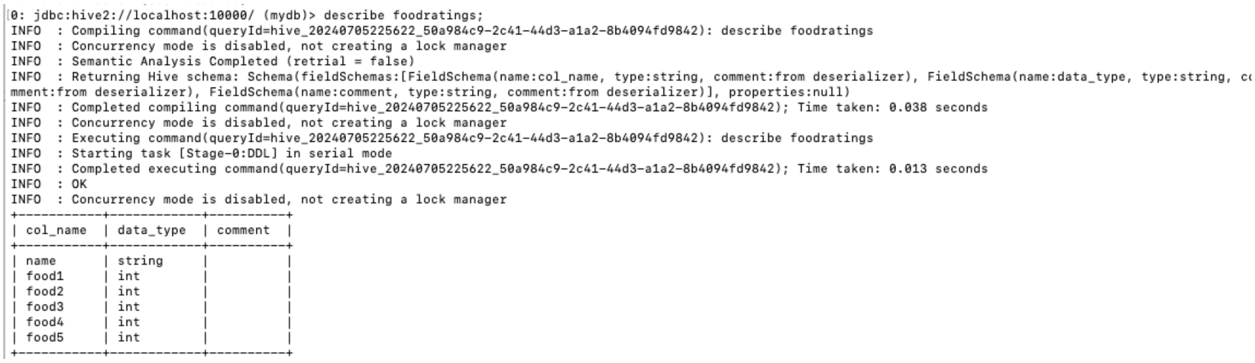

prabhuavula7 — prabhuavula7@a20522815-n2-m: ~/hql — ssh ◂ gcloud.py compute ssh a20522815-n2-m — 233×33
~ — prabhuavula7@a20522815-n2-m: ~/hql — ssh ◂ gcloud.py compute ssh a20522815-n2-m
~ — prabhuavula7@a20522815-n2-m: ~/hql — ssh ◂ gcloud.py compute ssh a20522815-n2-m
+

```
0: jdbc:hive2://localhost:10000/ (cs595)> SELECT * FROM cs595.salaries;
INFO  : Compiling command(queryId=hive_20240705212810_861cedd9-75c4-4196-af2b-3f89675e3d8d): SELECT * FROM cs595.salaries
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:salaries.name, type:string, comment:null), FieldSchema(name:salaries.jobtitle, type:string, comment:null), FieldSchema(name:salaries.agencyid, type:string, comment:
null), FieldSchema(name:salaries.agency, type:string, comment:null), FieldSchema(name:salaries.hiredate, type:string, comment:null), FieldSchema(name:salaries.annualsalary, type:double, comment:null), FieldSchema(name:salaries.grossp
ay, type:double, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240705212810_861cedd9-75c4-4196-af2b-3f89675e3d8d); Time taken: 0.18 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240705212810_861cedd9-75c4-4196-af2b-3f89675e3d8d): SELECT * FROM cs595.salaries
INFO  : Completed executing command(queryId=hive_20240705212810_861cedd9-75c4-4196-af2b-3f89675e3d8d); Time taken: 0.0 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------------------------+----------------------------------+--------------------+---------------------------------+-------------------------+-------------------------+---------------------+
|       salaries.name         |        salaries.jobtitle         | salaries.agencyid  |         salaries.agency         |    salaries.hiredate    | salaries.annualsalary   | salaries.grosspay   |
+-----------------------------+----------------------------------+--------------------+---------------------------------+-------------------------+-------------------------+---------------------+
| Aaron,Patricia G            | Facilities/Office Services II    | A03031             | OED-Employment Dev (031)        | 10/24/1979 12:00:00 AM  | 56705.0                 | 54135.44            |
| Aaron,Petra L               | ASSISTANT STATE'S ATTORNEY       | A29045             | States Attorneys Office (045)   | 09/25/2006 12:00:00 AM  | 75500.0                 | 72445.87            |
| Abbey,Emmanuel              | CONTRACT SERV SPEC II            | A40001             | M-R Info Technology (001)       | 05/01/2013 12:00:00 AM  | 60060.0                 | 59602.58            |
| Abbott-Cole,Michelle        | Operations Officer III           | A90005             | TRANS-Traffic (005)             | 11/28/2014 12:00:00 AM  | 70000.0                 | 59517.21            |
| Abdal-Rahim,Naim A          | EMT Firefighter Suppression      | A64120             | Fire Department (120)           | 03/30/2011 12:00:00 AM  | 64365.0                 | 74770.82            |
| Abdelmeguid,Shahrazad       | CONTRACT SERV SPEC II            | A29010             | States Attorneys Office (010)   | 11/30/2015 12:00:00 AM  | 40019.0                 | 16283.26            |
| Abdi,Ezekiel W              | POLICE SERGEANT                  | A99070             | Police Department (070)         | 06/14/2007 12:00:00 AM  | 82780.0                 | 106863.56           |
| Abdul Adl,Attrice A         | RADIO DISPATCHER SHERIFF         | A38410             | Sheriff's Office (410)          | 09/02/1999 12:00:00 AM  | 45471.0                 | 59418.35            |
| Abdul Aziz,Hajr E           | LIFEGUARD I                      | P04002             | R&P-Recreation (part-time) (    | 06/18/2014 12:00:00 AM  | 18408.0                 | 5909.64             |
| Abdul Aziz,Jennah A         | LIFEGUARD I                      | P04002             | R&P-Recreation (part-time) (    | 06/16/2014 12:00:00 AM  | 18408.0                 | 3230.27             |
| Abdul Aziz,Yaqub M          | LIFEGUARD I                      | P04002             | R&P-Recreation (part-time) (    | 06/09/2014 12:00:00 AM  | 18408.0                 | 4522.4              |
| Abdul Hamid,Umar            | SECRETARY II                     | A06009             | Housing & Community Dev (009)   | 01/17/1995 12:00:00 AM  | 37299.0                 | 35751.23            |
| Abdul-Aziz,Muhammad         | COMMUNITY AIDE                   | P04002             | R&P-Recreation (part-time) (    | 07/01/2014 12:00:00 AM  | 20800.0                 | 2310.0              |
| Abdul-Jabbar,Bushra A       | SOCIAL SERVICES COORDINATOR      | A06015             | Housing & Community Dev (015)   | 04/14/2008 12:00:00 AM  | 42446.0                 | 41272.07            |
| Abdul-Saboor,Jamillah       | SECRETARY II                     | A75054             | Enoch Pratt Free Library (054)  | 07/27/2009 12:00:00 AM  | 34218.0                 | 34969.65            |
| Abdullah,Aisha W            | OFFICE SUPPORT SPECIALIST III    | A85301             | General Services (301)          | 02/11/2013 12:00:00 AM  | 30409.0                 | 34478.02            |
| Abdullah,Beverly A          | OFFICE SUPPORT SPECIALIST III    | A06004             | Housing & Community Dev (004)   | 12/01/1986 12:00:00 AM  | 38326.0                 | 37228.57            |
```

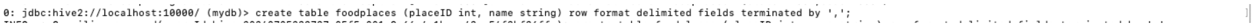
2.



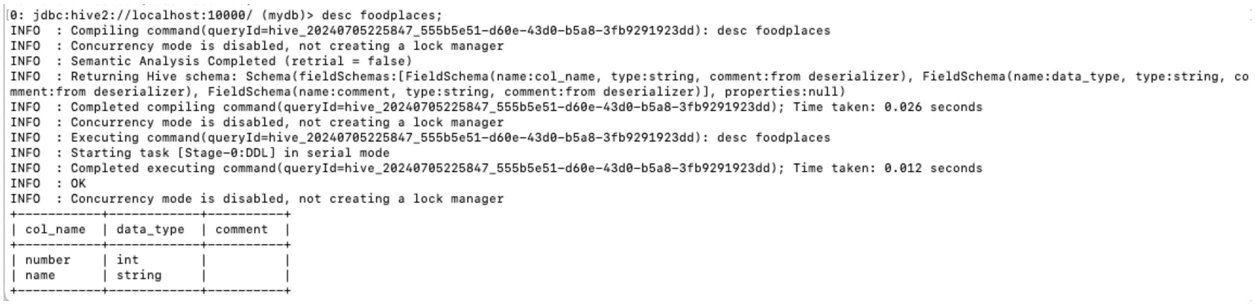
```
0: jdbc:hive2://localhost:10000/ (mydb)> create table foodratings (Name STRING, food1 int, food2 int, food3 int, food4 int, food5 int) row format delimited fields terminated by ',';
```



```
0: jdbc:hive2://localhost:10000/ (mydb)> describe foodratings;
INFO  : Compiling command(queryId=hive_20240705225622_50a984c9-2c41-44d3-a1a2-8b4094fd9842): describe foodratings
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:col_name, type:string, comment:from deserializer), FieldSchema(name:data_type, type:string, co
mment:from deserializer), FieldSchema(name:comment, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240705225622_50a984c9-2c41-44d3-a1a2-8b4094fd9842); Time taken: 0.038 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240705225622_50a984c9-2c41-44d3-a1a2-8b4094fd9842): describe foodratings
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20240705225622_50a984c9-2c41-44d3-a1a2-8b4094fd9842); Time taken: 0.013 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------+------------+----------+
| col_name  | data_type  | comment  |
+-----------+------------+----------+
| name      | string     |          |
| food1     | int        |          |
| food2     | int        |          |
| food3     | int        |          |
| food4     | int        |          |
| food5     | int        |          |
+-----------+------------+----------+
```

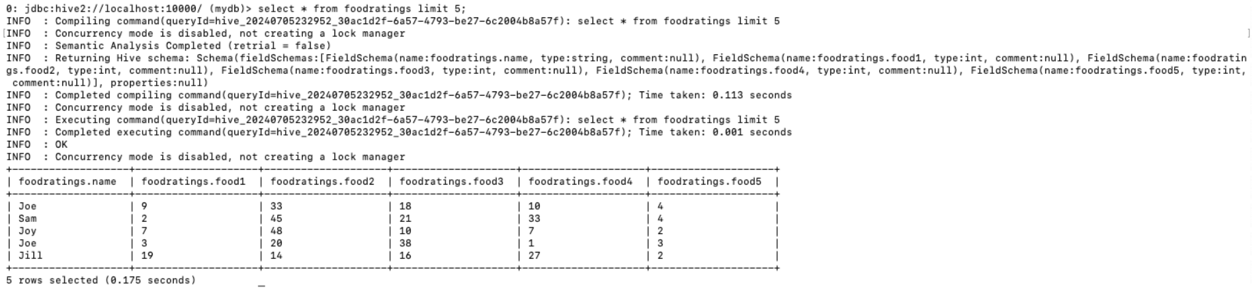


```
0: jdbc:hive2://localhost:10000/ (mydb)> create table foodplaces (placeID int, name string) row format delimited fields terminated by ',';
```

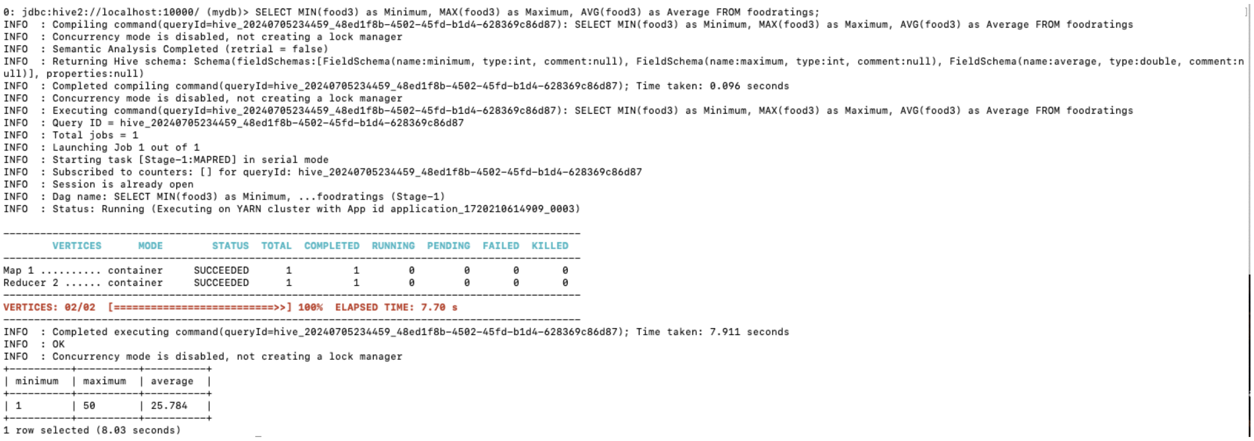


```
0: jdbc:hive2://localhost:10000/ (mydb)> desc foodplaces;
INFO  : Compiling command(queryId=hive_20240705225847_555b5e51-d60e-43d0-b5a8-3fb9291923dd): desc foodplaces
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:col_name, type:string, comment:from deserializer), FieldSchema(name:data_type, type:string, co
mment:from deserializer), FieldSchema(name:comment, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240705225847_555b5e51-d60e-43d0-b5a8-3fb9291923dd); Time taken: 0.026 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240705225847_555b5e51-d60e-43d0-b5a8-3fb9291923dd): desc foodplaces
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20240705225847_555b5e51-d60e-43d0-b5a8-3fb9291923dd); Time taken: 0.012 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------+------------+----------+
| col_name  | data_type  | comment  |
+-----------+------------+----------+
| number    | int        |          |
| name      | string     |          |
+-----------+------------+----------+
```

3.



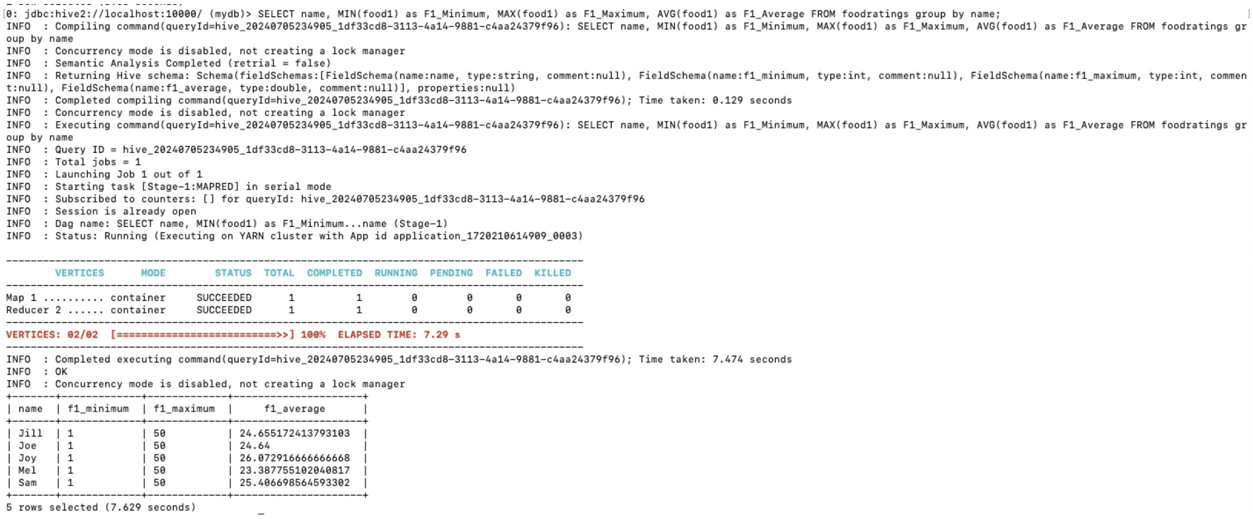
```
0: jdbc:hive2://localhost:10000/ (mydb)> select * from foodratings limit 5;
INFO  : Compiling command(queryId=hive_20240705232952_30ac1d2f-6a57-4793-be27-6c2004b8a57f): select * from foodratings limit 5
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:foodratings.name, type:string, comment:null), FieldSchema(name:foodratings.food1, type:int, comment:null), FieldSchema(name:foodratin
gs.food2, type:int, comment:null), FieldSchema(name:foodratings.food3, type:int, comment:null), FieldSchema(name:foodratings.food4, type:int, comment:null), FieldSchema(name:foodratings.food5, type:int,
 comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240705232952_30ac1d2f-6a57-4793-be27-6c2004b8a57f); Time taken: 0.113 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240705232952_30ac1d2f-6a57-4793-be27-6c2004b8a57f): select * from foodratings limit 5
INFO  : Completed executing command(queryId=hive_20240705232952_30ac1d2f-6a57-4793-be27-6c2004b8a57f); Time taken: 0.001 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------------------+--------------------+--------------------+--------------------+--------------------+--------------------+
| foodratings.name  | foodratings.food1  | foodratings.food2  | foodratings.food3  | foodratings.food4  | foodratings.food5  |
+-------------------+--------------------+--------------------+--------------------+--------------------+--------------------+
| Joe               | 9                  | 33                 | 18                 | 10                 | 4                  |
| Sam               | 2                  | 45                 | 21                 | 33                 | 4                  |
| Joy               | 7                  | 48                 | 10                 | 7                  | 2                  |
| Joe               | 3                  | 20                 | 38                 | 1                  | 3                  |
| Jill              | 19                 | 14                 | 16                 | 27                 | 2                  |
+-------------------+--------------------+--------------------+--------------------+--------------------+--------------------+
5 rows selected (0.175 seconds)
```



```
0: jdbc:hive2://localhost:10000/ (mydb)> SELECT MIN(food3) as Minimum, MAX(food3) as Maximum, AVG(food3) as Average FROM foodratings;
INFO  : Compiling command(queryId=hive_20240705234459_48ed1f8b-4502-45fd-b1d4-628369c86d87): SELECT MIN(food3) as Minimum, MAX(food3) as Maximum, AVG(food3) as Average FROM foodratings
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:minimum, type:int, comment:null), FieldSchema(name:maximum, type:int, comment:null), FieldSchema(name:average, type:double, comment:n
ull)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240705234459_48ed1f8b-4502-45fd-b1d4-628369c86d87); Time taken: 0.096 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240705234459_48ed1f8b-4502-45fd-b1d4-628369c86d87): SELECT MIN(food3) as Minimum, MAX(food3) as Maximum, AVG(food3) as Average FROM foodratings
INFO  : Query ID = hive_20240705234459_48ed1f8b-4502-45fd-b1d4-628369c86d87
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20240705234459_48ed1f8b-4502-45fd-b1d4-628369c86d87
INFO  : Session is already open
INFO  : Dag name: SELECT MIN(food3) as Minimum, ...foodratings (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1720210614909_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.70 s
----------------------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20240705234459_48ed1f8b-4502-45fd-b1d4-628369c86d87); Time taken: 7.911 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+----------+----------+----------+
| minimum  | maximum  | average  |
+----------+----------+----------+
| 1        | 50       | 25.784   |
+----------+----------+----------+
1 row selected (8.03 seconds)
```

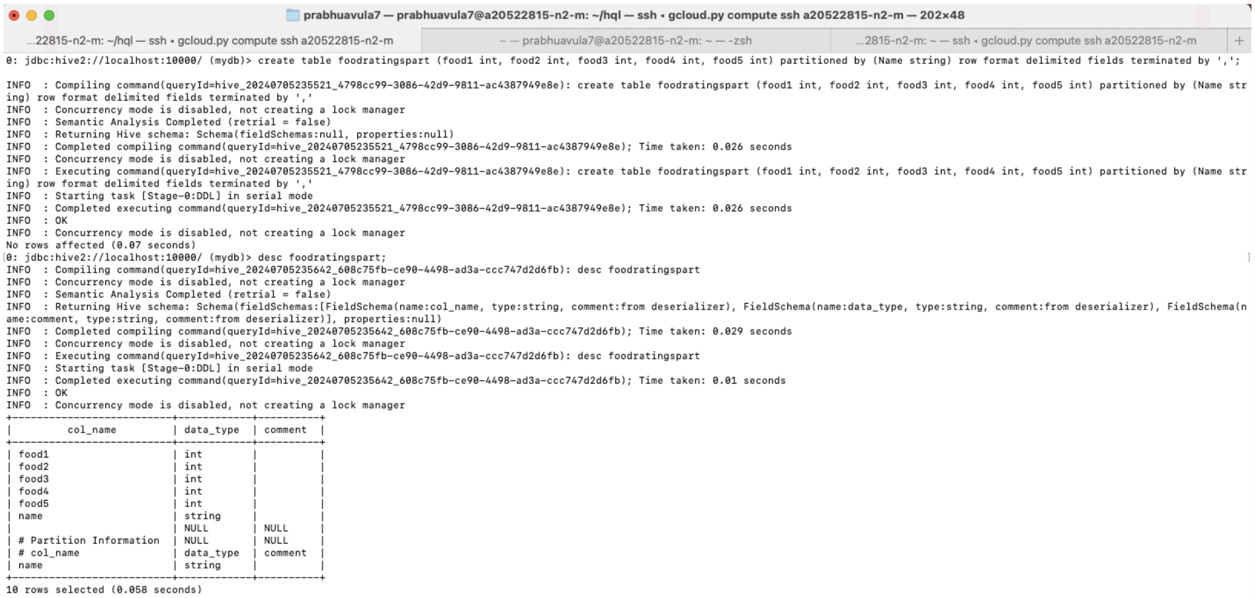
4.



```
0: jdbc:hive2://localhost:10000/ (mydb)> SELECT name, MIN(food1) as F1_Minimum, MAX(food1) as F1_Maximum, AVG(food1) as F1_Average FROM foodratings group by name;
INFO  : Compiling command(queryId=hive_20240705234905_1df33cd8-3113-4a14-9881-c4aa24379f96): SELECT name, MIN(food1) as F1_Minimum, MAX(food1) as F1_Maximum, AVG(food1) as F1_Average FROM foodratings gr
oup by name
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:name, type:string, comment:null), FieldSchema(name:f1_minimum, type:int, comment:null), FieldSchema(name:f1_maximum, type:int, commen
t:null), FieldSchema(name:f1_average, type:double, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240705234905_1df33cd8-3113-4a14-9881-c4aa24379f96); Time taken: 0.129 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240705234905_1df33cd8-3113-4a14-9881-c4aa24379f96): SELECT name, MIN(food1) as F1_Minimum, MAX(food1) as F1_Maximum, AVG(food1) as F1_Average FROM foodratings gr
oup by name
INFO  : Query ID = hive_20240705234905_1df33cd8-3113-4a14-9881-c4aa24379f96
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20240705234905_1df33cd8-3113-4a14-9881-c4aa24379f96
INFO  : Session is already open
INFO  : Dag name: SELECT name, MIN(food1) as F1_Minimum...name (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1720210614909_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.29 s
----------------------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20240705234905_1df33cd8-3113-4a14-9881-c4aa24379f96); Time taken: 7.474 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------+-------------+-------------+---------------------+
| name  | f1_minimum  | f1_maximum  |     f1_average      |
+-------+-------------+-------------+---------------------+
| Jill  | 1           | 50          | 24.655172413793103  |
| Joe   | 1           | 50          | 24.64               |
| Joy   | 1           | 50          | 26.072916666666668  |
| Mel   | 1           | 50          | 23.387755102040817  |
| Sam   | 1           | 50          | 25.406698564593302  |
+-------+-------------+-------------+---------------------+
5 rows selected (7.629 seconds)
```

5.



prabhuavula7 — prabhuavula7@a20522815-n2-m: ~/hql — ssh ◂ gcloud.py compute ssh a20522815-n2-m — 202×48
...22815-n2-m: ~/hql — ssh ◂ gcloud.py compute ssh a20522815-n2-m
~ — prabhuavula7@a20522815-n2-m: ~ — -zsh
...2815-n2-m: ~ — ssh ◂ gcloud.py compute ssh a20522815-n2-m
+

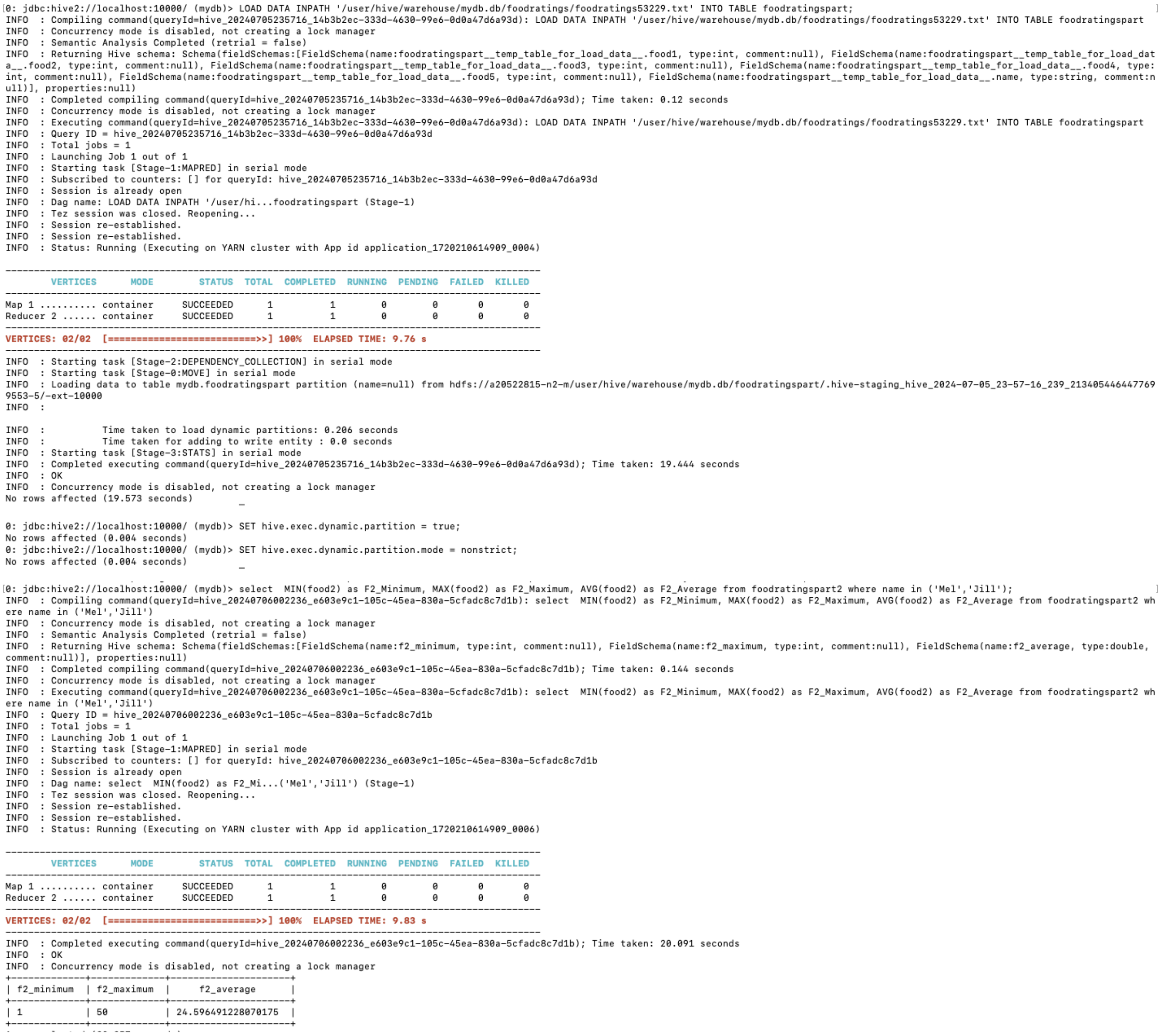
```
0: jdbc:hive2://localhost:10000/ (mydb)> create table foodratingspart (food1 int, food2 int, food3 int, food4 int, food5 int) partitioned by (Name string) row format delimited fields terminated by ',';

INFO  : Compiling command(queryId=hive_20240705235521_4798cc99-3086-42d9-9811-ac4387949e8e): create table foodratingspart (food1 int, food2 int, food3 int, food4 int, food5 int) partitioned by (Name str
ing) row format delimited fields terminated by ','
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20240705235521_4798cc99-3086-42d9-9811-ac4387949e8e); Time taken: 0.026 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240705235521_4798cc99-3086-42d9-9811-ac4387949e8e): create table foodratingspart (food1 int, food2 int, food3 int, food4 int, food5 int) partitioned by (Name str
ing) row format delimited fields terminated by ','
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20240705235521_4798cc99-3086-42d9-9811-ac4387949e8e); Time taken: 0.026 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.07 seconds)
0: jdbc:hive2://localhost:10000/ (mydb)> desc foodratingspart;
INFO  : Compiling command(queryId=hive_20240705235642_608c75fb-ce90-4498-ad3a-ccc747d2d6fb): desc foodratingspart
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:col_name, type:string, comment:from deserializer), FieldSchema(name:data_type, type:string, comment:from deserializer), FieldSchema(n
ame:comment, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240705235642_608c75fb-ce90-4498-ad3a-ccc747d2d6fb); Time taken: 0.029 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240705235642_608c75fb-ce90-4498-ad3a-ccc747d2d6fb): desc foodratingspart
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20240705235642_608c75fb-ce90-4498-ad3a-ccc747d2d6fb); Time taken: 0.01 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+--------------------------+------------+----------+
|         col_name         | data_type  | comment  |
+--------------------------+------------+----------+
| food1                    | int        |          |
| food2                    | int        |          |
| food3                    | int        |          |
| food4                    | int        |          |
| food5                    | int        |          |
| name                     | string     |          |
|                          | NULL       | NULL     |
| # Partition Information  | NULL       | NULL     |
| # col_name               | data_type  | comment  |
| name                     | string     |          |
+--------------------------+------------+----------+
10 rows selected (0.058 seconds)
```

6.

Partitioning the records by the critic's name is a better approach as we have a defined number of critics (5). The places ID, though defined separately in another document, may appear several times for each user in our main document. Hence, partitioning the records by the place ID would result in over-partitioning as there would be too many files with small sizes, cluttering the space.

7.



```
0: jdbc:hive2://localhost:10000/ (mydb)> LOAD DATA INPATH '/user/hive/warehouse/mydb.db/foodratings/foodratings53229.txt' INTO TABLE foodratingspart;
INFO  : Compiling command(queryId=hive_20240705235716_14b3b2ec-333d-4630-99e6-0d0a47d6a93d): LOAD DATA INPATH '/user/hive/warehouse/mydb.db/foodratings/foodratings53229.txt' INTO TABLE foodratingspart
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:foodratingspart__temp_table_for_load_data__.food1, type:int, comment:null), FieldSchema(name:foodratingspart__temp_table_for_load_dat
a__.food2, type:int, comment:null), FieldSchema(name:foodratingspart__temp_table_for_load_data__.food3, type:int, comment:null), FieldSchema(name:foodratingspart__temp_table_for_load_data__.food4, type:
int, comment:null), FieldSchema(name:foodratingspart__temp_table_for_load_data__.food5, type:int, comment:null), FieldSchema(name:foodratingspart__temp_table_for_load_data__.name, type:string, comment:n
ull)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240705235716_14b3b2ec-333d-4630-99e6-0d0a47d6a93d); Time taken: 0.12 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240705235716_14b3b2ec-333d-4630-99e6-0d0a47d6a93d): LOAD DATA INPATH '/user/hive/warehouse/mydb.db/foodratings/foodratings53229.txt' INTO TABLE foodratingspart
INFO  : Query ID = hive_20240705235716_14b3b2ec-333d-4630-99e6-0d0a47d6a93d
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20240705235716_14b3b2ec-333d-4630-99e6-0d0a47d6a93d
INFO  : Session is already open
INFO  : Dag name: LOAD DATA INPATH '/user/hi...foodratingspart (Stage-1)
INFO  : Tez session was closed. Reopening...
INFO  : Session re-established.
INFO  : Session re-established.
INFO  : Status: Running (Executing on YARN cluster with App id application_1720210614909_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 9.76 s
----------------------------------------------------------------------------------------------
INFO  : Starting task [Stage-2:DEPENDENCY_COLLECTION] in serial mode
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table mydb.foodratingspart partition (name=null) from hdfs://a20522815-n2-m/user/hive/warehouse/mydb.db/foodratingspart/.hive-staging_hive_2024-07-05_23-57-16_239_213405446447769
9553-5/-ext-10000
INFO  :

INFO  :          Time taken to load dynamic partitions: 0.206 seconds
INFO  :          Time taken for adding to write entity : 0.0 seconds
INFO  : Starting task [Stage-3:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20240705235716_14b3b2ec-333d-4630-99e6-0d0a47d6a93d); Time taken: 19.444 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (19.573 seconds)

0: jdbc:hive2://localhost:10000/ (mydb)> SET hive.exec.dynamic.partition = true;
No rows affected (0.004 seconds)
0: jdbc:hive2://localhost:10000/ (mydb)> SET hive.exec.dynamic.partition.mode = nonstrict;
No rows affected (0.004 seconds)

0: jdbc:hive2://localhost:10000/ (mydb)> select  MIN(food2) as F2_Minimum, MAX(food2) as F2_Maximum, AVG(food2) as F2_Average from foodratingspart2 where name in ('Mel','Jill');
INFO  : Compiling command(queryId=hive_20240706002236_e603e9c1-105c-45ea-830a-5cfadc8c7d1b): select  MIN(food2) as F2_Minimum, MAX(food2) as F2_Maximum, AVG(food2) as F2_Average from foodratingspart2 wh
ere name in ('Mel','Jill')
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:f2_minimum, type:int, comment:null), FieldSchema(name:f2_maximum, type:int, comment:null), FieldSchema(name:f2_average, type:double,
comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240706002236_e603e9c1-105c-45ea-830a-5cfadc8c7d1b); Time taken: 0.144 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240706002236_e603e9c1-105c-45ea-830a-5cfadc8c7d1b): select  MIN(food2) as F2_Minimum, MAX(food2) as F2_Maximum, AVG(food2) as F2_Average from foodratingspart2 wh
ere name in ('Mel','Jill')
INFO  : Query ID = hive_20240706002236_e603e9c1-105c-45ea-830a-5cfadc8c7d1b
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20240706002236_e603e9c1-105c-45ea-830a-5cfadc8c7d1b
INFO  : Session is already open
INFO  : Dag name: select  MIN(food2) as F2_Mi...('Mel','Jill') (Stage-1)
INFO  : Tez session was closed. Reopening...
INFO  : Session re-established.
INFO  : Session re-established.
INFO  : Status: Running (Executing on YARN cluster with App id application_1720210614909_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 9.83 s
----------------------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20240706002236_e603e9c1-105c-45ea-830a-5cfadc8c7d1b); Time taken: 20.091 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------------+-------------+---------------------+
| f2_minimum  | f2_maximum  |     f2_average      |
+-------------+-------------+---------------------+
| 1           | 50          | 24.596491228070175  |
+-------------+-------------+---------------------+
```

Short Answers based on the article “An Introduction to Big Data Formats”

1. The most important consideration is the nature of your queries. Row-based storage is ideal for accessing all or most columns for each row, such as in web log files or structured databases. Column-based storage is more efficient for analytical queries that only require a subset of columns across large datasets

2. Splittability refers to breaking a file into smaller, independently processable parts. It is crucial for large data sets as it allows for parallel processing, significantly improving performance by distributing the workload across multiple processors. Columnar formats are often more splittable because they store data in a way that aligns with processing by columns

3. Files that store similar data types together, such as all dates or all numerical values, achieve better compression in column format. This is because similar data can be more efficiently compressed when stored sequentially compared to the mixed data types typically found in row-based storage

4. Parquet best suits read-heavy workloads and analytical queries on wide datasets with many columns. It offers high compression and is splittable, which makes it efficient for large-scale data processing, particularly with Hadoop ecosystems like Apache Impala for low latency and high concurrency queries