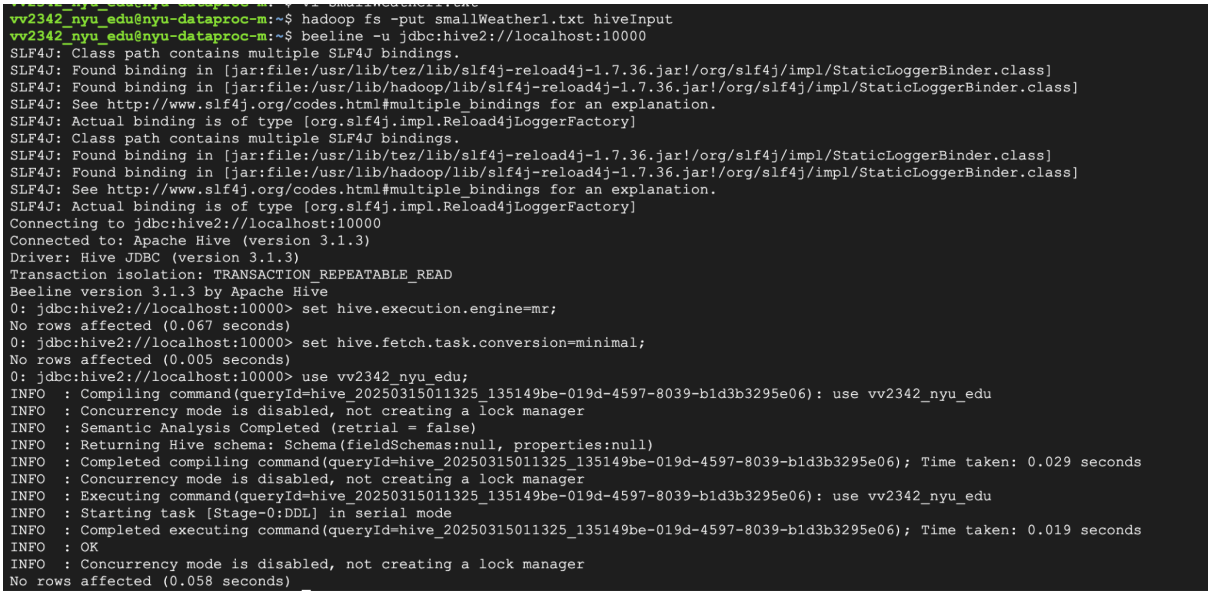
LAB 4

Vaishnavi Venkateswaran

vv2342

Create a directory in HDFS and put input file into it :

Connect to Hive shell and set execution engine to MapReduce . Also selected the database created for me:



Create external table and display it

```
No rows selected (0.07 seconds)
0: jdbc:hive2://localhost:10000> create external table w1 (data1 string, year int, data2 string, temperature int, quality tinyint, data3 string)row format delimited fields terminated by ','
location '/user/vv2342_nyu_edu/hiveInput/';
INFO  : Compiling command(queryId=hive_20250315014916_3aa0f095-6689-4866-b2a9-3677d47a6a51): create external table w1 (data1 string, year int, data2 string, temperature int, quality tinyint,
 data3 string)row format delimited fields terminated by ',' location '/user/vv2342_nyu_edu/hiveInput/'
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_20250315014916_3aa0f095-6689-4866-b2a9-3677d47a6a51); Time taken: 0.029 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250315014916_3aa0f095-6689-4866-b2a9-3677d47a6a51): create external table w1 (data1 string, year int, data2 string, temperature int, quality tinyint,
 data3 string)row format delimited fields terminated by ',' location '/user/vv2342_nyu_edu/hiveInput/'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250315014916_3aa0f095-6689-4866-b2a9-3677d47a6a51); Time taken: 0.069 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.105 seconds)
0: jdbc:hive2://localhost:10000> show tables;
INFO  : Compiling command(queryId=hive_20250315014935_c6bdd8f9-03c7-4b6d-b7df-f3a010717012): show tables
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250315014935_c6bdd8f9-03c7-4b6d-b7df-f3a010717012); Time taken: 0.032 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250315014935_c6bdd8f9-03c7-4b6d-b7df-f3a010717012): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250315014935_c6bdd8f9-03c7-4b6d-b7df-f3a010717012); Time taken: 0.017 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------+
| tab_name  |
+-----------+
| w1        |
+-----------+
1 row selected (0.063 seconds)
0: jdbc:hive2://localhost:10000>
```

```
INFO  : Executing command(queryId=hive_20250315015042_1b9e581d-dcdf-435f-97ce-c9a310a09546): select * from w1
INFO  : Completed executing command(queryId=hive_20250315015042_1b9e581d-dcdf-435f-97ce-c9a310a09546); Time taken: 0.0 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------------+----------+-----------------------------------------------------+-----------------+-------------+---------------+
|     w1.data1    | w1.year  |                      w1.data2                       | w1.temperature  | w1.quality  |   w1.data3    |
+-----------------+----------+-----------------------------------------------------+-----------------+-------------+---------------+
| 006701199099999 | 1950     | 051507004+68750+023550FM-12+038299999V0203301N00671220001CN9999999N9 | 0               | 1           | +99999999999  |
| 004301199099999 | 1950     | 051512004+68750+023550FM-12+038299999V0203201N00671220001CN9999999N9 | 22              | 1           | +99999999999  |
| 004301199099999 | 1950     | 051518004+68750+023550FM-12+038299999V0203201N00261220001CN9999999N9 | -11             | 1           | +99999999999  |
| 004301265099999 | 1949     | 032412004+62300+010750FM-12+048599999V0202701N00461220001CN0500001N9 | 111             | 1           | +99999999999  |
| 004301265099999 | 1949     | 032418004+62300+010750FM-12+048599999V0202701N00461220001CN0500001N9 | 78              | 1           | +99999999999  |
+-----------------+----------+-----------------------------------------------------+-----------------+-------------+---------------+
5 rows selected (0.089 seconds)
```

Displaying all the rows of the table w1

Running the queries which are fast :

```
0: jdbc:hive2://localhost:10000> select * from w1 limit 2;
INFO  : Compiling command(queryId=hive_20250315015049_30125f53-1093-47b1-a9bf-47bd051dcfe9): select * from w1 limit 2
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:w1.data1, type:string, comment:null), FieldSchema(name:w1.year, type:int, comment:null), FieldSchema(name:w1.data2, type
string, comment:null), FieldSchema(name:w1.temperature, type:int, comment:null), FieldSchema(name:w1.quality, type:tinyint, comment:null), FieldSchema(name:w1.data3, type:string, comment:nu
ll)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250315015049_30125f53-1093-47b1-a9bf-47bd051dcfe9); Time taken: 0.06 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250315015049_30125f53-1093-47b1-a9bf-47bd051dcfe9): select * from w1 limit 2
INFO  : Completed executing command(queryId=hive_20250315015049_30125f53-1093-47b1-a9bf-47bd051dcfe9); Time taken: 0.0 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------------+----------+-----------------------------------------------------+-----------------+-------------+---------------+
|     w1.data1    | w1.year  |                      w1.data2                       | w1.temperature  | w1.quality  |   w1.data3    |
+-----------------+----------+-----------------------------------------------------+-----------------+-------------+---------------+
| 006701199099999 | 1950     | 051507004+68750+023550FM-12+038299999V0203301N00671220001CN9999999N9 | 0               | 1           | +99999999999  |
| 004301199099999 | 1950     | 051512004+68750+023550FM-12+038299999V0203201N00671220001CN9999999N9 | 22              | 1           | +99999999999  |
+-----------------+----------+-----------------------------------------------------+-----------------+-------------+---------------+
2 rows selected (0.079 seconds)
0: jdbc:hive2://localhost:10000> select year from w1;
INFO  : Compiling command(queryId=hive_20250315015054_0affbd30-121d-4ae5-b9ed-7e0bbe6c95e0): select year from w1
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:year, type:int, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250315015054_0affbd30-121d-4ae5-b9ed-7e0bbe6c95e0); Time taken: 0.061 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250315015054_0affbd30-121d-4ae5-b9ed-7e0bbe6c95e0): select year from w1
INFO  : Completed executing command(queryId=hive_20250315015054_0affbd30-121d-4ae5-b9ed-7e0bbe6c95e0); Time taken: 0.0 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------+
| year  |
+-------+
| 1950  |
| 1950  |
| 1950  |
| 1949  |
| 1949  |
+-------+
5 rows selected (0.077 seconds)
```

The following queries are slower because MapReduce also runs this time .

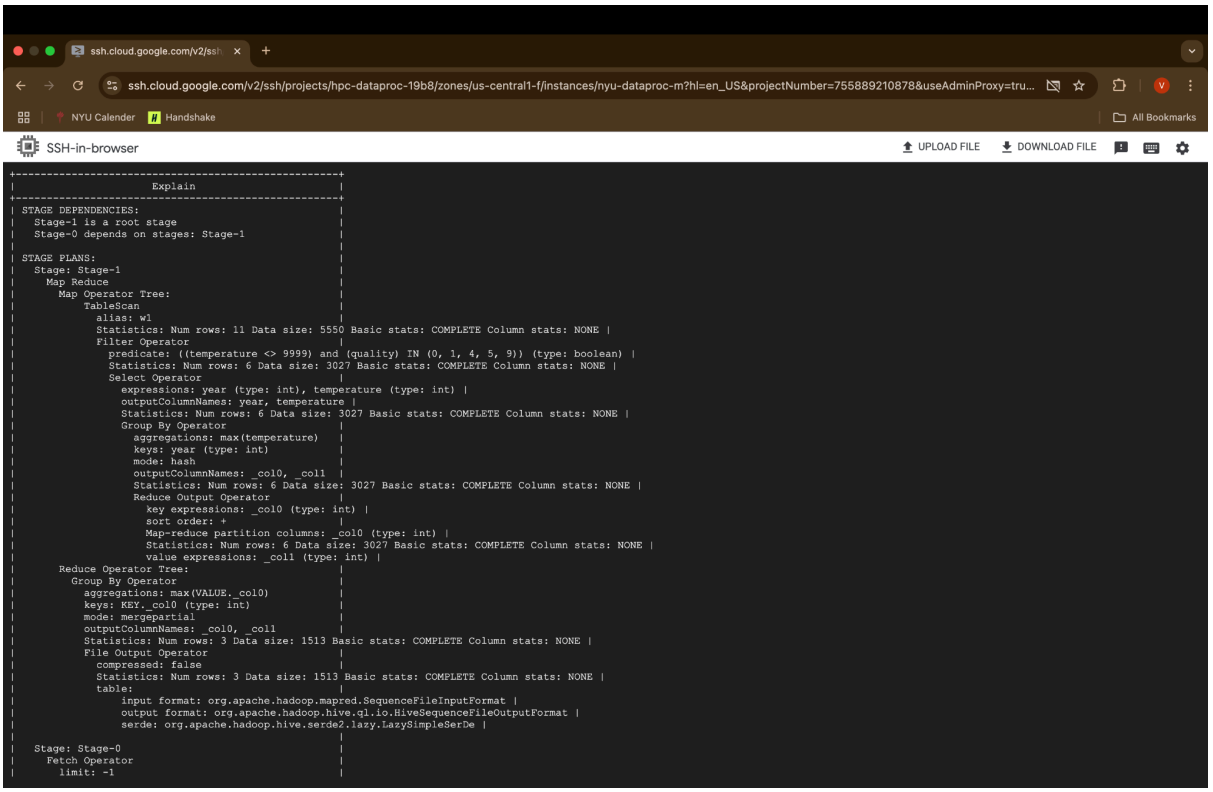
```
0: jdbc:hive2://localhost:10000> select * from w1 where year > 1949;
INFO  : Compiling command(queryId=hive_20250315015517_a8039f22-e449-465a-b1d8-d76938148057): select * from w1 where year > 1949
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:w1.data1, type:string, comment:null), FieldSchema(name:w1.year, type:int, comment:null), FieldSchema(name:w1.data2, type:
string, comment:null), FieldSchema(name:w1.temperature, type:int, comment:null), FieldSchema(name:w1.quality, type:tinyint, comment:null), FieldSchema(name:w1.data3, type:string, comment:nul
l)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250315015517_a8039f22-e449-465a-b1d8-d76938148057); Time taken: 0.133 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250315015517_a8039f22-e449-465a-b1d8-d76938148057): select * from w1 where year > 1949
WARN  : Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
INFO  : Query ID = hive_20250315015517_a8039f22-e449-465a-b1d8-d76938148057
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Number of reduce tasks is set to 0 since there's no reduce operator
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1740271921940_5063
INFO  : Executing with tokens: []
INFO  : The url to track the job: http://nyu-dataproc-m.local.:8088/proxy/application_1740271921940_5063/
INFO  : Starting Job = job_1740271921940_5063, Tracking URL = http://nyu-dataproc-m.local.:8088/proxy/application_1740271921940_5063/
INFO  : Kill Command = /usr/lib/hadoop/bin/mapred job  -kill job_1740271921940_5063
INFO  : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
INFO  : 2025-03-15 01:55:27,741 Stage-1 map = 0%,  reduce = 0%
INFO  : 2025-03-15 01:55:37,015 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.8 sec
INFO  : MapReduce Total cumulative CPU time: 2 seconds 800 msec
INFO  : Ended Job = job_1740271921940_5063
INFO  : MapReduce Jobs Launched:
INFO  : Stage-Stage-1: Map: 1   Cumulative CPU: 2.8 sec   HDFS Read: 7224 HDFS Write: 447 SUCCESS
INFO  : Total MapReduce CPU Time Spent: 2 seconds 800 msec
INFO  : Completed executing command(queryId=hive_20250315015517_a8039f22-e449-465a-b1d8-d76938148057); Time taken: 21.156 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------------+----------+-----------------------------------------------------------+-----------------+-------------+---------------+
|    w1.data1     | w1.year  |                         w1.data2                          | w1.temperature  | w1.quality  |   w1.data3    |
+-----------------+----------+-----------------------------------------------------------+-----------------+-------------+---------------+
| 006701199099999 | 1950     | 051507004+68750+023550FM-12+038299999V0203301N00671220001CN9999999N9 | 0               | 1           | +99999999999  |
| 004301199099999 | 1950     | 051512004+68750+023550FM-12+038299999V0203201N00671220001CN9999999N9 | 22              | 1           | +99999999999  |
| 004301199099999 | 1950     | 051518004+68750+023550FM-12+038299999V0203201N00261220001CN9999999N9 | -11             | 1           | +99999999999  |
+-----------------+----------+-----------------------------------------------------------+-----------------+-------------+---------------+
3 rows selected (21.306 seconds)
```

```
0: jdbc:hive2://localhost:10000> select distinct year from w1;
INFO  : Compiling command(queryId=hive_20250315015541_25f3ec0e-ba98-45e4-af5e-e25f2ec6199d): select distinct year from w1
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:year, type:int, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250315015541_25f3ec0e-ba98-45e4-af5e-e25f2ec6199d); Time taken: 0.269 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250315015541_25f3ec0e-ba98-45e4-af5e-e25f2ec6199d): select distinct year from w1
WARN  : Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or usi
INFO  : Query ID = hive_20250315015541_25f3ec0e-ba98-45e4-af5e-e25f2ec6199d
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Number of reduce tasks not specified. Estimated from input data size: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1740271921940_5065
INFO  : Executing with tokens: []
INFO  : The url to track the job: http://nyu-dataproc-m.local.:8088/proxy/application_1740271921940_5065/
INFO  : Starting Job = job_1740271921940_5065, Tracking URL = http://nyu-dataproc-m.local.:8088/proxy/application_1740271921940_5065/
INFO  : Kill Command = /usr/lib/hadoop/bin/mapred job  -kill job_1740271921940_5065
INFO  : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO  : 2025-03-15 01:55:52,986 Stage-1 map = 0%,  reduce = 0%
INFO  : 2025-03-15 01:56:01,218 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.57 sec
INFO  : 2025-03-15 01:56:10,466 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.82 sec
INFO  : MapReduce Total cumulative CPU time: 5 seconds 820 msec
INFO  : Ended Job = job_1740271921940_5065
INFO  : MapReduce Jobs Launched:
INFO  : Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.82 sec   HDFS Read: 13088 HDFS Write: 121 SUCCESS
INFO  : Total MapReduce CPU Time Spent: 5 seconds 820 msec
INFO  : Completed executing command(queryId=hive_20250315015541_25f3ec0e-ba98-45e4-af5e-e25f2ec6199d); Time taken: 31.174 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------+
| year  |
+-------+
| 1949  |
| 1950  |
+-------+
2 rows selected (31.457 seconds)
```

```
0: jdbc:hive2://localhost:10000> select year, max(temperature) as maxTemp from w1 where temperature != 9999 and quality in (0, 1, 4, 5, 9) group by year;
INFO  : Compiling command(queryId=hive_20250315015237_b2a7b74e-9d8a-4e67-9dea-6a1404f6c174): select year, max(temperature) as maxTemp from w1 where temperature != 9999 and quality in (0, 1,
4, 5, 9) group by year
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:year, type:int, comment:null), FieldSchema(name:maxtemp, type:int, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250315015237_b2a7b74e-9d8a-4e67-9dea-6a1404f6c174); Time taken: 0.118 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250315015237_b2a7b74e-9d8a-4e67-9dea-6a1404f6c174): select year, max(temperature) as maxTemp from w1 where temperature != 9999 and quality in (0, 1,
4, 5, 9) group by year
WARN  : Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
INFO  : Query ID = hive_20250315015237_b2a7b74e-9d8a-4e67-9dea-6a1404f6c174
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Number of reduce tasks not specified. Estimated from input data size: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1740271921940_5061
INFO  : Executing with tokens: []
INFO  : The url to track the job: http://nyu-dataproc-m.local.:8088/proxy/application_1740271921940_5061/
INFO  : Starting Job = job_1740271921940_5061, Tracking URL = http://nyu-dataproc-m.local.:8088/proxy/application_1740271921940_5061/
INFO  : Kill Command = /usr/lib/hadoop/bin/mapred job  -kill job_1740271921940_5061
INFO  : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO  : 2025-03-15 01:52:50,150 Stage-1 map = 0%,  reduce = 0%
INFO  : 2025-03-15 01:52:59,453 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.55 sec
INFO  : 2025-03-15 01:53:07,726 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.9 sec
INFO  : MapReduce Total cumulative CPU time: 5 seconds 900 msec
INFO  : Ended Job = job_1740271921940_5061
INFO  : MapReduce Jobs Launched:
INFO  : Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.9 sec   HDFS Read: 11044 HDFS Write: 128 SUCCESS
INFO  : Total MapReduce CPU Time Spent: 5 seconds 900 msec
INFO  : Completed executing command(queryId=hive_20250315015237_b2a7b74e-9d8a-4e67-9dea-6a1404f6c174); Time taken: 32.754 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------+----------+
| year  | maxtemp  |
+-------+----------+
| 1949  | 111      |
| 1950  | 22       |
+-------+----------+
2 rows selected (32.89 seconds)
```

Results of running the query : select year, max(temperature) as maxTemp from w1 where temperature != 9999 and quality in (0, 1, 4, 5, 9) group by year; Same as what we did in Lab 2 but now we use one line of code

Now we prefix the above query with ‘explain’ to see how Hive converts it into a MapReduce job :



```
+----------------------------------------------------+
|                      Explain                       |
+----------------------------------------------------+
| STAGE DEPENDENCIES:                                |
|   Stage-1 is a root stage                          |
|   Stage-0 depends on stages: Stage-1               |
|                                                    |
| STAGE PLANS:                                       |
|   Stage: Stage-1                                   |
|     Map Reduce                                     |
|       Map Operator Tree:                           |
|           TableScan                                |
|             alias: w1                              |
|             Statistics: Num rows: 11 Data size: 5550 Basic stats: COMPLETE Column stats: NONE |
|             Filter Operator                        |
|               predicate: ((temperature <> 9999) and (quality) IN (0, 1, 4, 5, 9)) (type: boolean) |
|               Statistics: Num rows: 6 Data size: 3027 Basic stats: COMPLETE Column stats: NONE |
|               Select Operator                      |
|                 expressions: year (type: int), temperature (type: int) |
|                 outputColumnNames: year, temperature |
|                 Statistics: Num rows: 6 Data size: 3027 Basic stats: COMPLETE Column stats: NONE |
|                 Group By Operator                  |
|                   aggregations: max(temperature)   |
|                   keys: year (type: int)           |
|                   mode: hash                       |
|                   outputColumnNames: _col0, _col1  |
|                   Statistics: Num rows: 6 Data size: 3027 Basic stats: COMPLETE Column stats: NONE |
|                   Reduce Output Operator           |
|                     key expressions: _col0 (type: int) |
|                     sort order: +                  |
|                     Map-reduce partition columns: _col0 (type: int) |
|                     Statistics: Num rows: 6 Data size: 3027 Basic stats: COMPLETE Column stats: NONE |
|                     value expressions: _col1 (type: int) |
|       Reduce Operator Tree:                        |
|         Group By Operator                          |
|           aggregations: max(VALUE._col0)           |
|           keys: KEY._col0 (type: int)              |
|           mode: mergepartial                       |
|           outputColumnNames: _col0, _col1          |
|           Statistics: Num rows: 3 Data size: 1513 Basic stats: COMPLETE Column stats: NONE |
|           File Output Operator                     |
|             compressed: false                      |
|             Statistics: Num rows: 3 Data size: 1513 Basic stats: COMPLETE Column stats: NONE |
|             table:                                |
|                 input format: org.apache.hadoop.mapred.SequenceFileInputFormat |
|                 output format: org.apache.hadoop.hive.ql.io.HiveSequenceFileOutputFormat |
|                 serde: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe |
|                                                    |
|   Stage: Stage-0                                   |
|     Fetch Operator                                 |
|       limit: -1                                    |
```

The second table (Table w2)

```
+------------------+----------+-----------------------------------------------------------------+-----------------+-------------+-----------+
|     w2.data1     | w2.year  |                           w2.data2                              | w2.temperature  | w2.quality  | w2.nines  |
+------------------+----------+-----------------------------------------------------------------+-----------------+-------------+-----------+
| 0067011990999999 | 1950     | 051507004+68750+023550FM-12+038299999V0203301N00671220001CN9999999N9 | 0               | 1           | NULL      |
| 004301199099999  | 1950     | 051512004+68750+023550FM-12+038299999V0203201N00671220001CN9999999N9 | 22              | 1           | NULL      |
| 004301199099999  | 1950     | 051518004+68750+023550FM-12+038299999V0203201N00261220001CN9999999N9 | -11             | 1           | NULL      |
| 004301265099999  | 1949     | 032412004+62300+010750FM-12+048599999V0202701N00461220001CN0500001N9 | 111             | 1           | NULL      |
| 004301265099999  | 1949     | 032418004+62300+010750FM-12+048599999V0202701N00461220001CN0500001N9 | 78              | 1           | NULL      |
+------------------+----------+-----------------------------------------------------------------+-----------------+-------------+-----------+
5 rows selected (0.068 seconds)
0: jdbc:hive2://localhost:10000>
```

```
+------------------+----------+-----------------------------------------------------------------+-----------------+-------------+-------------+
|     w3.data1     | w3.year  |                           w3.data2                              | w3.temperature  | w3.quality  |  w3.nines   |
+------------------+----------+-----------------------------------------------------------------+-----------------+-------------+-------------+
| 0067011990999999 | 1950     | 051507004+68750+023550FM-12+038299999V0203301N00671220001CN9999999N9 | 0               | 1           | 99999999999 |
| 004301199099999  | 1950     | 051512004+68750+023550FM-12+038299999V0203201N00671220001CN9999999N9 | 22              | 1           | 99999999999 |
| 004301199099999  | 1950     | 051518004+68750+023550FM-12+038299999V0203201N00261220001CN9999999N9 | -11             | 1           | 99999999999 |
| 004301265099999  | 1949     | 032412004+62300+010750FM-12+048599999V0202701N00461220001CN0500001N9 | 111             | 1           | 99999999999 |
| 004301265099999  | 1949     | 032418004+62300+010750FM-12+048599999V0202701N00461220001CN0500001N9 | 78              | 1           | 99999999999 |
+------------------+----------+-----------------------------------------------------------------+-----------------+-------------+-------------+
5 rows selected (0.079 seconds)
```

The table w3

What Might Be Happening?

- When you create multiple external tables (w2, w3) over the same file but define **different schemas**, Hive doesn’t enforce structure matching.
- Instead, it reads the raw file and interprets it based on the schema you provide.
- If the schema has too few or too many fields, data may get assigned incorrectly, leading to unexpected NULLs or incorrect values.

Dropping the table w2

Now , we create multiple copies of the same input file

```
vv2342_nyu_edu@nyu-dataproc-m:~$ hadoop fs -cp hiveInput/smallWeather1.txt hiveInput/smallWeather2.txt
vv2342_nyu_edu@nyu-dataproc-m:~$ hadoop fs -cp hiveInput/smallWeather1.txt hiveInput/smallWeather3.txt
vv2342_nyu_edu@nyu-dataproc-m:~$ hadoop fs -ls hiveInput
Found 3 items
-rw-r--r--   1 vv2342_nyu_edu vv2342_nyu_edu        555 2025-03-15 01:12 hiveInput/smallWeather1.txt
-rw-r--r--   1 vv2342_nyu_edu vv2342_nyu_edu        555 2025-03-15 02:07 hiveInput/smallWeather2.txt
-rw-r--r--   1 vv2342_nyu_edu vv2342_nyu_edu        555 2025-03-15 02:07 hiveInput/smallWeather3.txt
vv2342_nyu_edu@nyu-dataproc-m:~$
```

Now, when we query out the table we see this :

```
0: jdbc:hive2://localhost:10000> select * from w1;
INFO  : Compiling command(queryId=hive_20250315020813_cdbea765-f6b1-4aa1-bcb7-7aae137cc263): select * from w1
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:w1.data1, type:string, comment:null), FieldSchema(name:w1.year, type:int, comment:null), FieldSchema(name:w1.data2, type:string, comment:null), FieldSchema(name:w1.temperature, type:int, comment:null), FieldSchema(name:w1.quality, type:tinyint, comment:null), FieldSchema(name:w1.data3, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250315020813_cdbea765-f6b1-4aa1-bcb7-7aae137cc263); Time taken: 0.061 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250315020813_cdbea765-f6b1-4aa1-bcb7-7aae137cc263): select * from w1
INFO  : Completed executing command(queryId=hive_20250315020813_cdbea765-f6b1-4aa1-bcb7-7aae137cc263); Time taken: 0.0 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------------+----------+----------------------------------------------------------+-----------------+-------------+--------------+
|    w1.data1     | w1.year  |                         w1.data2                         | w1.temperature  | w1.quality  |   w1.data3   |
+-----------------+----------+----------------------------------------------------------+-----------------+-------------+--------------+
| 006701199099999 | 1950     | 051507004+68750+023550FM-12+038299999V0203301N00671220001CN9999999N9 | 0               | 1           | +99999999999  |
| 004301199099999 | 1950     | 051512004+68750+023550FM-12+038299999V0203201N00671220001CN9999999N9 | 22              | 1           | +99999999999  |
| 004301199099999 | 1950     | 051518004+68750+023550FM-12+038299999V0203201N00261220001CN9999999N9 | -11             | 1           | +99999999999  |
| 004301265099999 | 1949     | 032412004+62300+010750FM-12+048599999V0202701N00461220001CN0500001N9 | 111             | 1           | +99999999999  |
| 004301265099999 | 1949     | 032418004+62300+010750FM-12+048599999V0202701N00461220001CN0500001N9 | 78              | 1           | +99999999999  |
| 006701199099999 | 1950     | 051507004+68750+023550FM-12+038299999V0203301N00671220001CN9999999N9 | 0               | 1           | +99999999999  |
| 004301199099999 | 1950     | 051512004+68750+023550FM-12+038299999V0203201N00671220001CN9999999N9 | 22              | 1           | +99999999999  |
| 004301199099999 | 1950     | 051518004+68750+023550FM-12+038299999V0203201N00261220001CN9999999N9 | -11             | 1           | +99999999999  |
| 004301265099999 | 1949     | 032412004+62300+010750FM-12+048599999V0202701N00461220001CN0500001N9 | 111             | 1           | +99999999999  |
| 004301265099999 | 1949     | 032418004+62300+010750FM-12+048599999V0202701N00461220001CN0500001N9 | 78              | 1           | +99999999999  |
| 006701199099999 | 1950     | 051507004+68750+023550FM-12+038299999V0203301N00671220001CN9999999N9 | 0               | 1           | +99999999999  |
| 004301199099999 | 1950     | 051512004+68750+023550FM-12+038299999V0203201N00671220001CN9999999N9 | 22              | 1           | +99999999999  |
| 004301199099999 | 1950     | 051518004+68750+023550FM-12+038299999V0203201N00261220001CN9999999N9 | -11             | 1           | +99999999999  |
| 004301265099999 | 1949     | 032412004+62300+010750FM-12+048599999V0202701N00461220001CN0500001N9 | 111             | 1           | +99999999999  |
| 004301265099999 | 1949     | 032418004+62300+010750FM-12+048599999V0202701N00461220001CN0500001N9 | 78              | 1           | +99999999999  |
+-----------------+----------+----------------------------------------------------------+-----------------+-------------+--------------+
15 rows selected (0.087 seconds)
```

```
0: jdbc:hive2://localhost:10000> closing: 0: jdbc:hive2://localhost:10000
vv2342_nyu_edu@nyu-dataproc-m:~$ trino
trino> show catalogs;
       Catalog
----------------------
 bigquery
 bigquery_public_data
 hive
 mastersql
 memory
 system
 tpcds
 tpch
(8 rows)

Query 20250315_021016_00025_weaaj, FINISHED, 2 nodes
Splits: 20 total, 20 done (100.00%)
0.03 [0 rows, 0B] [0 rows/s, 0B/s]
```

Running Trino

We see that running any query is way faster in Trino than in Hive

```
trino:vv2342_nyu_edu> select year, max(temperature) as maxTemp from w1 where temperature != 9999 and quality in (0, 1, 4, 5, 9) group by year;
 year | maxTemp
------+---------
 1949 |     111
 1950 |      22
(2 rows)

Query 20250315_021134_00029_weaaj, FINISHED, 2 nodes
Splits: 35 total, 35 done (100.00%)
0.65 [15 rows, 1.63KB] [22 rows/s, 2.49KB/s]

trino:vv2342_nyu_edu>
```