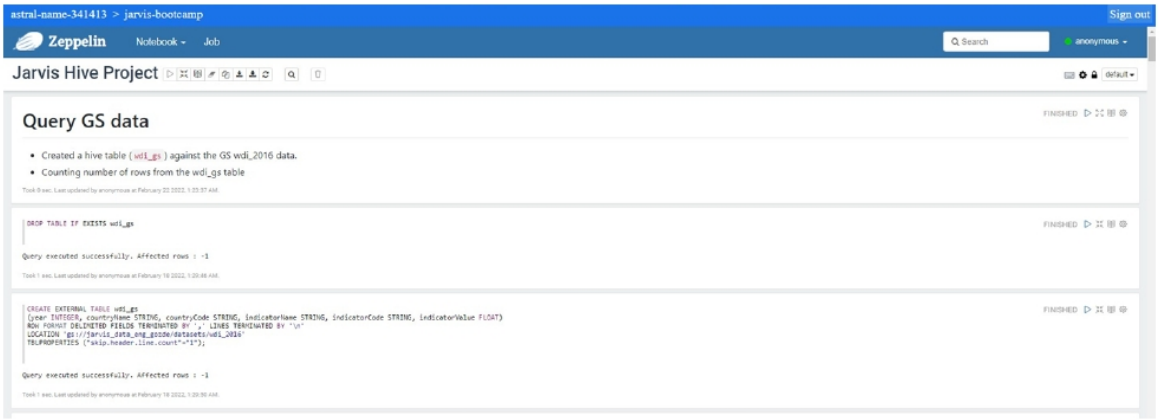
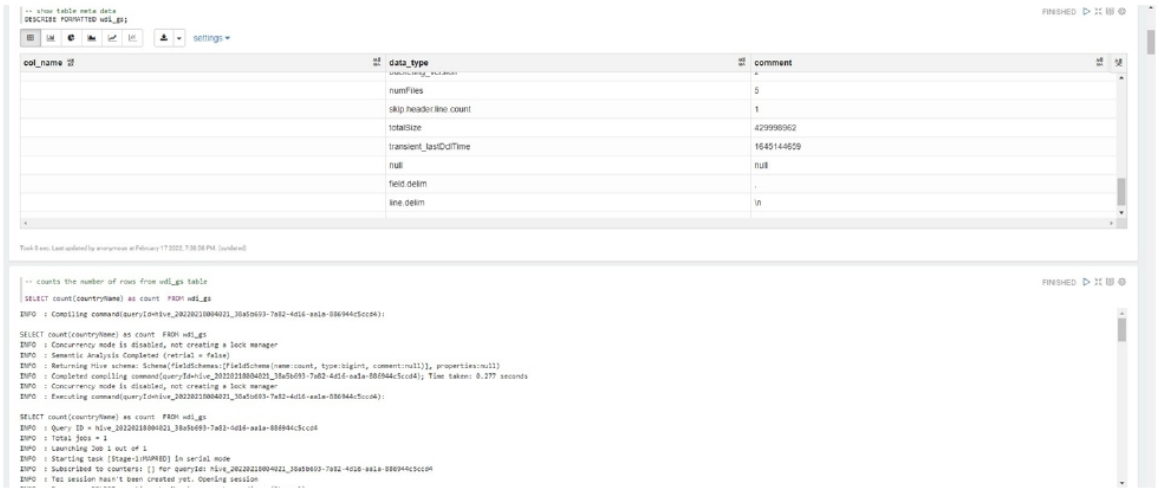

astral-name-341413 > jarvis-bootcamp
Sign out
Zeppelin
Notebook
Job
Search
anonymous
Jarvis Hive Project
default
Query GS data
FINISHED
Created a hive table (wdi_gs) against the GS wdi_2016 data.
Counting number of rows from the wdi_gs table
DROP TABLE IF EXISTS wdi_gs
Query executed successfully. Affected rows : -1
CREATE EXTERNAL TABLE wdi_gs
(year INTEGER, countryName STRING, countryCode STRING, indicatorName STRING, indicatorCode STRING, indicatorValue FLOAT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n'
LOCATION 'gs://jarvis_data_eng_gcs2m/datasets/wdi_2016'
TBLPROPERTIES ("skip.header.line.count"="1");
Query executed successfully. Affected rows : -1


-- show table meta data
DESCRIBE FORMATTED wdi_gs;
settings
col_name
data_type
comment
numFiles
5
skip.header.line.count
1
totalSize
429998962
transient_lastDdlTime
1645144659
null
null
field.delim
,
line.delim
\n

-- counts the number of rows from wdi_gs table
SELECT count(countryName) as count FROM wdi_gs
INFO : Compiling command(queryId=hive_20220218004021_38a5b693-7a82-4d16-aa1a-886944c5ccd4):
SELECT count(countryName) as count FROM wdi_gs
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:count, type:bigint, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20220218004021_38a5b693-7a82-4d16-aa1a-886944c5ccd4); Time taken: 0.277 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20220218004021_38a5b693-7a82-4d16-aa1a-886944c5ccd4):
SELECT count(countryName) as count FROM wdi_gs
INFO : Query ID = hive_20220218004021_38a5b693-7a82-4d16-aa1a-886944c5ccd4
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_20220218004021_38a5b693-7a82-4d16-aa1a-886944c5ccd4
INFO : Tez session hasn't been created yet. Opening session


-- counts the number of rows from wdi_gs table
SELECT count(countryName) as count FROM wdi_gs
INFO : Compiling command(queryId=hive_20220218004021_38a5b693-7a82-4d16-aa1a-886944c5ccd4):
SELECT count(countryName) as count FROM wdi_gs
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:count, type:bigint, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20220218004021_38a5b693-7a82-4d16-aa1a-886944c5ccd4); Time taken: 0.277 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20220218004021_38a5b693-7a82-4d16-aa1a-886944c5ccd4):
SELECT count(countryName) as count FROM wdi_gs
INFO : Query ID = hive_20220218004021_38a5b693-7a82-4d16-aa1a-886944c5ccd4
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_20220218004021_38a5b693-7a82-4d16-aa1a-886944c5ccd4
INFO : Tez session hasn't been created yet. Opening session
settings
count
21759408

Load GS data to HDFS
FINISHED

```
DROP TABLE IF EXISTS wdi_csv_text
```

FINISHED

```
Query executed successfully. Affected rows : -1
```

Took 1 sec. Last updated by anonymous at February 18 2022, 1:32:26 AM.

```
CREATE EXTERNAL TABLE wdi_csv_text
(year INTEGER, countryName STRING, countryCode STRING, indicatorName STRING, indicatorCode STRING, indicatorValue FLOAT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n'
LOCATION 'hdfs:///user/garslan/hive/wdi/wdi_csv_text';
```

FINISHED

```
Query executed successfully. Affected rows : -1
```

Took 1 sec. Last updated by anonymous at February 18 2022, 1:32:30 AM.

- Query that loads data from wdi_gs table to wdi_csv_text table.

FINISHED

Took 0 sec. Last updated by anonymous at February 22 2022, 1:33:22 AM.

```
INSERT OVERWRITE TABLE wdi_csv_text
SELECT * FROM wdi_gs
```

FINISHED

```
INFO  : Compiling command(queryId=hive_20220218063235_f188442a-f46a-4fc2-9b82-d0bda8531508): INSERT OVERWRITE TABLE wdi_csv_text
SELECT * FROM wdi_gs

INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:wdi_gs.year, type:int, comment:null), FieldSchema(name:wdi_gs.countryname, type:string, comment:null), FieldSchema(name:wdi_gs.countrycode, type:string, comment:null), FieldSchema(name:wdi_gs.indicatorname,
type:string, comment:null), FieldSchema(name:wdi_gs.indicatorcode, type:string, comment:null), FieldSchema(name:wdi_gs.indicatorvalue, type:float, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20220218063235_f188442a-f46a-4fc2-9b82-d0bda8531508); Time taken: 0.36 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220218063235_f188442a-f46a-4fc2-9b82-d0bda8531508): INSERT OVERWRITE TABLE wdi_csv_text
SELECT * FROM wdi_gs

INFO  : Query ID = hive_20220218063235_f188442a-f46a-4fc2-9b82-d0bda8531508
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_20220218063235_f188442a-f46a-4fc2-9b82-d0bda8531508

Query executed successfully. Affected rows : INFO  : Star-ttl
ng task [Stage-2:DEPENDENCY_COLLECTION] in serial mode
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table default.wdi_csv_text from hdfs:///jarvis-bootcamp-m/user/garslan/hive/wdi/wdi_csv_text/.hive-staging_hive_2022-02-18_06-32-35_240_7119954303282732200-8/-ext-10000
INFO  : Starting task [Stage-3:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20220218063235_f188442a-f46a-4fc2-9b82-d0bda8531508); Time taken: 111.988 seconds
```

Took 1 min 52 sec. Last updated by anonymous at February 18 2022, 1:34:27 AM. (outdated)

- Check HDFS file size for wdi_csv_text file

FINISHED

Took 0 sec. Last updated by anonymous at February 22 2022, 1:33:17 AM.

```
%sh
 hdfs dfs -ls -h /user/garslan/hive/wdi/wdi_csv_text
```

FINISHED

```
Found 5 items
-rw-r--r--   2 gpadmin hadoop    386.0 M 2022-02-18 06:34 /user/garslan/hive/wdi/wdi_csv_text/000000_0
-rw-r--r--   2 gpadmin hadoop    385.8 M 2022-02-18 06:33 /user/garslan/hive/wdi/wdi_csv_text/000001_0
-rw-r--r--   2 gpadmin hadoop    386.0 M 2022-02-18 06:34 /user/garslan/hive/wdi/wdi_csv_text/000002_0
-rw-r--r--   2 gpadmin hadoop    385.6 M 2022-02-18 06:34 /user/garslan/hive/wdi/wdi_csv_text/000003_0
-rwxr-xr-x   2 gpadmin hadoop    210.4 M 2022-02-18 06:34 /user/garslan/hive/wdi/wdi_csv_text/000004_0
```

Took 3 sec. Last updated by anonymous at February 18 2022, 1:37:36 AM.

```
%sh
 hdfs dfs -du -s -h /user/garslan/hive/wdi/wdi_csv_text

1.7 G  5.4 G  /user/garslan/hive/wdi/wdi_csv_text
```

FINISHED

Took 3 sec. Last updated by anonymous at February 18 2022, 1:37:59 AM.

```
SELECT count(countryName) as count  FROM wdi_csv_text
```

FINISHED

```
INFO  : Compiling command(queryId=hive_20220218063912_6d5ac6c9-90f0-4048-b2e9-479dc4be6731): SELECT count(countryName) as count  FROM wdi_csv_text

INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:count, type:bigint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20220218063912_6d5ac6c9-90f0-4048-b2e9-479dc4be6731); Time taken: 0.202 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220218063912_6d5ac6c9-90f0-4048-b2e9-479dc4be6731): SELECT count(countryName) as count  FROM wdi_csv_text

INFO  : Query ID = hive_20220218063912_6d5ac6c9-90f0-4048-b2e9-479dc4be6731
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_20220218063912_6d5ac6c9-90f0-4048-b2e9-479dc4be6731
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: SELECT count(countryName) as ...wdi_csv_text (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1645140844819_0013)
```

settings

count
21759408

Took 30 sec. Last updated by anonymous at February 18 2022, 1:39:42 AM. (outdated)

- Clear filesystem cache and execute the count query again.

FINISHED

Took 0 sec. Last updated by anonymous at February 22 2022, 1:34:01 AM.

```
%sh
echo 3 | sudo tee /proc/sys/vm/drop_caches
```

ERROR

```
We trust you have received the usual lecture from the local System
Administrator. It usually boils down to these three things:

    #1) Respect the privacy of others.
    #2) Think before you type.
    #3) With great power comes great responsibility.

sudo: no tty present and no askpass program specified

ExitValue: 1
```

Took 0 sec. Last updated by anonymous at February 18 2022, 1:40:55 AM.

Hive vs Bash

FINISHED

Bash took 3 seconds to execute, while Hive took 20 seconds. The total time of the Hive technique was longer. Because of the cost of parsing queries, developing an implementation of execution plan, and performing Hadoop Map Reduce task.

Took 0 sec. Last updated by anonymous at February 22 2022, 1:34:07 AM.

```
%sh
#SSH to master node
cd ~
hdfs  dfs -get  hdfs:///user/garslan/hive/wdi/wdi_csv_text .
cd wdi_csv_text
#calculate current directory size
du -ch .
#1.8G  total

#clear fs cache
echo 3 | sudo tee /proc/sys/vm/drop_caches
#hash row count
date +%s && cat * | wc && date +%s
```

```
get: `wdi_csv_text/000000_0': File exists
get: `wdi_csv_text/000001_0': File exists
get: `wdi_csv_text/000002_0': File exists
get: `wdi_csv_text/000003_0': File exists
get: `wdi_csv_text/000004_0': File exists
1.8G    .
1.8G    total

We trust you have received the usual lecture from the local System
Administrator. It usually boils down to these three things:

    #1) Respect the privacy of others.
    #2) Think before you type.
    #3) With great power comes great responsibility.

sudo: no tty present and no askpass program specified
1645151268
```

Took 28 sec. Last updated by anonymous at February 17 2022, 9:28:03 PM. (outdated)

```
%sh
cd ~
hdfs  dfs -get  hdfs:///user/garslan/hive/wdi/wdi_csv_text .
```

```
get: `wdi_csv_text/000000_0': File exists
get: `wdi_csv_text/000001_0': File exists
get: `wdi_csv_text/000002_0': File exists
get: `wdi_csv_text/000003_0': File exists
get: `wdi_csv_text/000004_0': File exists

ExitValue: 1
```

Took 3 sec. Last updated by anonymous at February 18 2022, 1:42:58 AM.

Parsing Issue solved

The indicatorName column may contain commas that causes problem when parsing data by using LazysimpleSerDe. In this case opencsvSerDe should be used to parse data into columsn in the Hive table.

Create an external table wdi_gs_debug with one column without any SerDe parsing

Took 0 sec. Last updated by anonymous at February 22 2022, 1:24:13 AM.

```
%hive
SELECT distinct(indicatorcode)
FROM wdi_csv_text
ORDER BY indicatorcode
LIMIT 20
```

```
INFO  : Compiling command(queryId=hive_20220218064124_06483e85-e8b4-47dd-ac94-17fedc093a01):
SELECT distinct(indicatorcode)
FROM wdi_csv_text
ORDER BY indicatorcode
LIMIT 20

INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:indicatorcode, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20220218064124_06483e85-e8b4-47dd-ac94-17fedc093a01); Time taken: 0.306 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220218064124_06483e85-e8b4-47dd-ac94-17fedc093a01):
SELECT distinct(indicatorcode)
FROM wdi_csv_text
ORDER BY indicatorcode
LIMIT 20
```

settings

indicatorcode
% of exports of goods
%)"
(% of urban population)"
15+
Atlas method (current US$)"
Australia (current US$)"
Austria (current US$)"
Belgium (current US$)"

Took 24 sec. Last updated by anonymous at February 18 2022, 1:41:38 AM.

1. Identify issue by creating a debug table

Took 0 sec. Last updated by anonymous at February 22 2022, 1:24:37 AM.

```
%hive
-- 1a. Create an external table wdi_gs_debug with one column without any SerDe parsing

CREATE EXTERNAL TABLE wdi_gs_debug
(line STRING)
ROW FORMAT DELIMITED LINES TERMINATED BY '\n'
STORED AS TEXTFILE
LOCATION 'hdfs:///user/garslan/hive/wdi/wdi_csv_text';
```

```
Error while processing statement: FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. AlreadyExistsException(message:Table hive.default.wdi_gs_debug already exists)
```

Took 0 sec. Last updated by anonymous at February 18 2022, 1:42:44 AM.

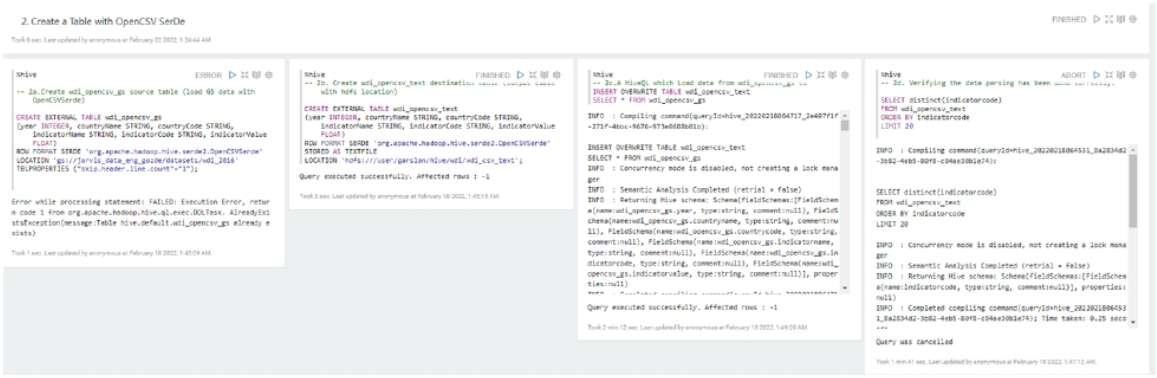
```
%hive
-- 1b.Query the line that have parsing issue

SELECT line FROM wdi_gs_debug
WHERE line like "%\(%of urban population\)\"%"
```

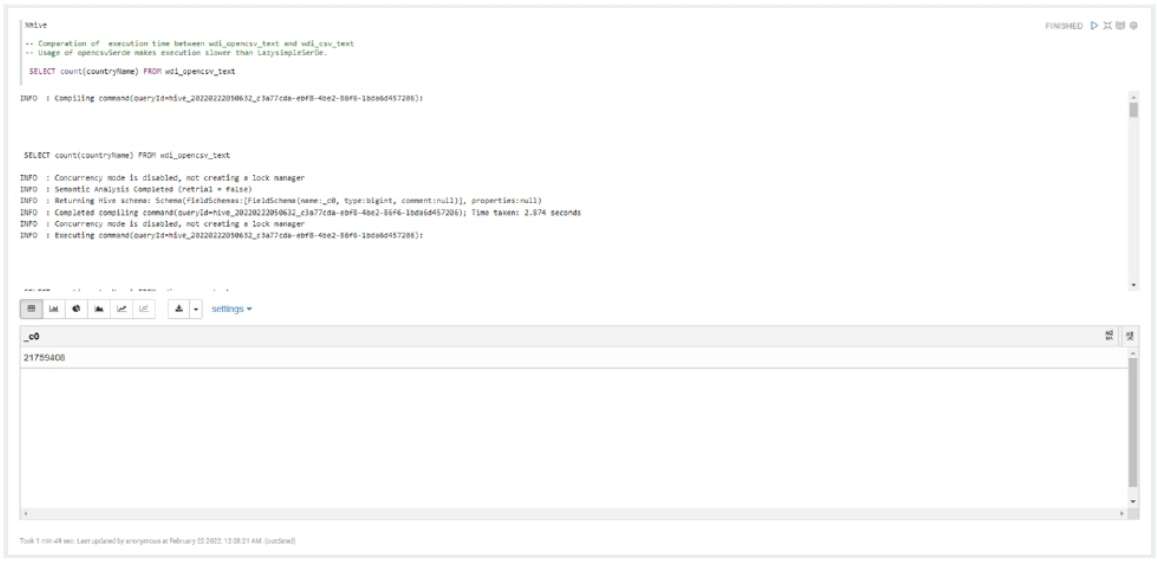
```
VERTICES: 00/01  [==================>>---------] 71%   ELAPSED TIME: 33.67 s
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ......    container     RUNNING      7          5        1        0        0       0
--------------------------------------------------------------------------------
VERTICES: 00/01  [==================>>---------] 71%   ELAPSED TIME: 36.17 s
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..... container     SUCCEEDED      7          7        0        0       0       0
--------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20220218064304_8dffc2cb-87c3-4096-9c8f-782c5303bfd6); Time taken: 47.556 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
```

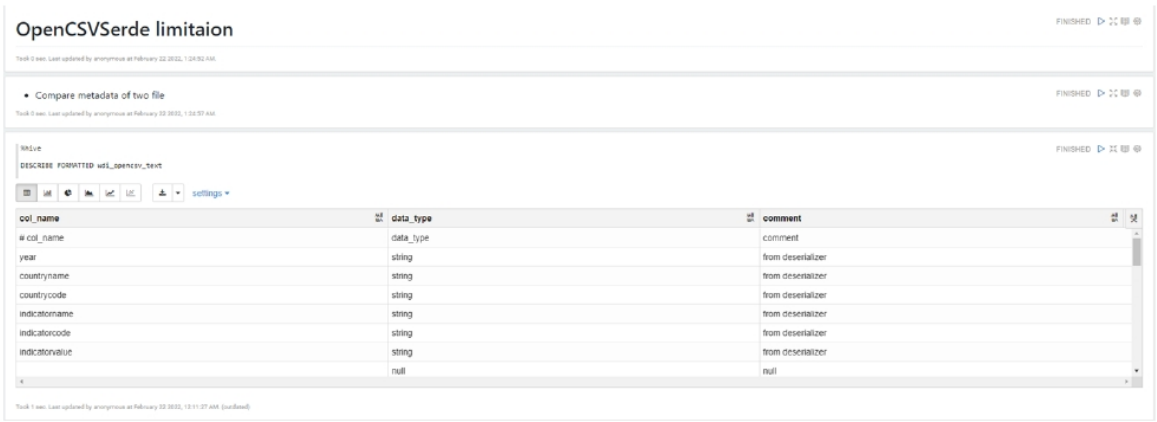
settings

line
2007,Hungary,HUN,"Population living in slums, (% of urban population)",\N
2014,Hungary,HUN,"Access to electricity, urban (% of urban population)",\N
1962,Iceland,ISL,"Access to non-solid fuel, urban (% of urban population)",\N
1974,Iceland,ISL,"Population living in slums, (% of urban population)",\N
1975,Iceland,ISL,"Population living in slums, (% of urban population)",\N
1982,Iceland,ISL,"People practicing open defecation, urban (% of urban population)",\N
1992,Iceland,ISL,"Access to non-solid fuel, urban (% of urban population)",\N
1996,Iceland,ISL,"Population living in slums, (% of urban population)",\N

Took 46 sec. Last updated by anonymous at February 18 2022, 1:43:52 AM.


2. Create a Table with OpenCSV SerDe
FINISHED
%hive
ERROR
-- 2a.Create wdi_opencsv_gs source table (load GS data with
OpenCSVSerde)
CREATE EXTERNAL TABLE wdi_opencsv_gs
(year INTEGER, countryName STRING, countryCode STRING,
IndicatorName STRING, IndicatorCode STRING, IndicatorValue
FLOAT)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
LOCATION 'gs://jarvis_data_eng_gozde/datasets/wdi_2016'
TBLPROPERTIES ("skip.header.line.count"="1");
Error while processing statement: FAILED: Execution Error, retur
n code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. AlreadyExi
stsException(message:Table hive.default.wdi_opencsv_gs already e
xists)
Took 1 sec. Last updated by anonymous at February 18 2022, 1:45:09 AM.
%hive
FINISHED
-- 2b. Create wdi_opencsv_text destinatio
with hdfs location)
CREATE EXTERNAL TABLE wdi_opencsv_text
(year INTEGER, countryName STRING, countryCode STRING,
IndicatorName STRING, IndicatorCode STRING, IndicatorValue
FLOAT)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
STORED AS TEXTFILE
LOCATION 'hdfs:///user/garslan/hive/wdi/wdi_csv_text';
Query executed successfully. Affected rows : -1
Took 0 sec. Last updated by anonymous at February 18 2022, 1:49:13 AM.
%hive
FINISHED
-- 2c.A HiveQL which Load date from wdi
INSERT OVERWRITE TABLE wdi_opencsv_text
SELECT * FROM wdi_opencsv_gs
INFO : Compiling command(queryId=hive_20220218064717_1e407f1f
-271f-4bbc-9674-973e0d83b81b):
INSERT OVERWRITE TABLE wdi_opencsv_text
SELECT * FROM wdi_opencsv_gs
INFO : Concurrency mode is disabled, not creating a lock mana
ger
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchem
a(name:wdi_opencsv_gs.year, type:string, comment:null), FieldS
chema(name:wdi_opencsv_gs.countryname, type:string, comment:nu
ll), FieldSchema(name:wdi_opencsv_gs.countrycode, type:string,
comment:null), FieldSchema(name:wdi_opencsv_gs.indicatorname,
type:string, comment:null), FieldSchema(name:wdi_opencsv_gs.in
dicatorcode, type:string, comment:null), FieldSchema(name:wdi_
opencsv_gs.indicatorvalue, type:string, comment:null)], proper
ties:null)
Query executed successfully. Affected rows : -1
Took 2 min 12 sec. Last updated by anonymous at February 18 2022, 1:49:28 AM.
%hive
ABORT
-- 2c. Verifying the date parsing has been
SELECT distinct(indicatorcode)
FROM wdi_opencsv_text
ORDER BY indicatorcode
LIMIT 20
INFO : Compiling command(queryId=hive_20220218064531_8a2834d2
-3b82-4eb5-80f8-c04ae30b1e74):
SELECT distinct(indicatorcode)
FROM wdi_opencsv_text
ORDER BY indicatorcode
LIMIT 20
INFO : Concurrency mode is disabled, not creating a lock mana
ger
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchem
a(name:indicatorcode, type:string, comment:null)], properties:
null)
INFO : Completed compiling command(queryId=hive_20220218064531
_8a2834d2-3b82-4eb5-80f8-c04ae30b1e74); Time taken: 0.25 seco
Query was cancelled
Took 1 min 41 sec. Last updated by anonymous at February 18 2022, 1:47:12 AM.


%hive
FINISHED
-- Comparation of execution time between wdi_opencsv_text and wdi_csv_text
-- Usage of opencsvSerde makes execution slower than LazySimpleSerDe.
SELECT count(countryName) FROM wdi_opencsv_text
INFO : Compiling command(queryId=hive_20220222050632_c3a77cda-ebf8-4be2-86f6-1bda6d457206):
SELECT count(countryName) FROM wdi_opencsv_text
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:_c0, type:bigint, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20220222050632_c3a77cda-ebf8-4be2-86f6-1bda6d457206); Time taken: 2.874 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20220222050632_c3a77cda-ebf8-4be2-86f6-1bda6d457206):
settings
_c0
21759408
Took 1 min 49 sec. Last updated by anonymous at February 22 2022, 12:08:21 AM. (outdated)


OpenCSVSerde limitaion
FINISHED
Took 0 sec. Last updated by anonymous at February 22 2022, 1:24:52 AM.
• Compare metadata of two file
FINISHED
Took 0 sec. Last updated by anonymous at February 22 2022, 1:24:57 AM.
%hive
FINISHED
DESCRIBE FORMATTED wdi_opencsv_text
settings
col_name
data_type
comment
col_name
data_type
comment
year
string
from deserializer
countryname
string
from deserializer
countrycode
string
from deserializer
indicatorname
string
from deserializer
indicatorcode
string
from deserializer
indicatorvalue
string
from deserializer
null
null
Took 1 sec. Last updated by anonymous at February 22 2022, 12:11:27 AM. (outdated)

```
%hive
DESCRIBE FORMATTED wdi_csv_text
```

FINISHED

settings

col_name	data_type	comment
# col_name	data_type	comment
year	int	
countryname	string	
countrycode	string	
indicatorname	string	
indicatorcode	string	
indicatorvalue	float	
	null	null

Took 1 sec. Last updated by anonymous at February 22 2022, 12:11:48 AM. (outdated)

```
%hive
-- Create a view on top of wdi_opencsv_text to cast specific columns to correct data type.

DROP VIEW IF EXISTS wdi_opencsv_text_view;

CREATE VIEW IF NOT EXISTS wdi_opencsv_text_view
AS
SELECT cast(year as INTEGER), countryName, countryCode,indicatorName, indicatorCode,cast(indicatorValue as Float) from wdi_opencsv_text
```

```
Query executed successfully. Affected rows : -1


Query executed successfully. Affected rows : -1
```

Took 1 sec. Last updated by anonymous at February 22 2022, 12:12:36 AM.

FINISHED

```
%hive
-- Check if the data types casted properly
DESCRIBE FORMATTED wdi_opencsv_text_view
```

FINISHED

settings

col_name	data_type	comment
# col_name	data_type	comment
year	int	
countryname	string	
countrycode	string	
indicatorname	string	
indicatorcode	string	
indicatorvalue	float	
	null	null

Took 0 sec. Last updated by anonymous at February 22 2022, 12:12:40 AM. (outdated)

2015 Canada GDP Growth HQL

FINISHED

Took 0 sec. Last updated by anonymous at February 22 2022, 1:23:06 AM.

- Find 2015 GDP growth (annual %) for Canada.

FINISHED

Took 0 sec. Last updated by anonymous at February 22 2022, 1:23:13 AM.

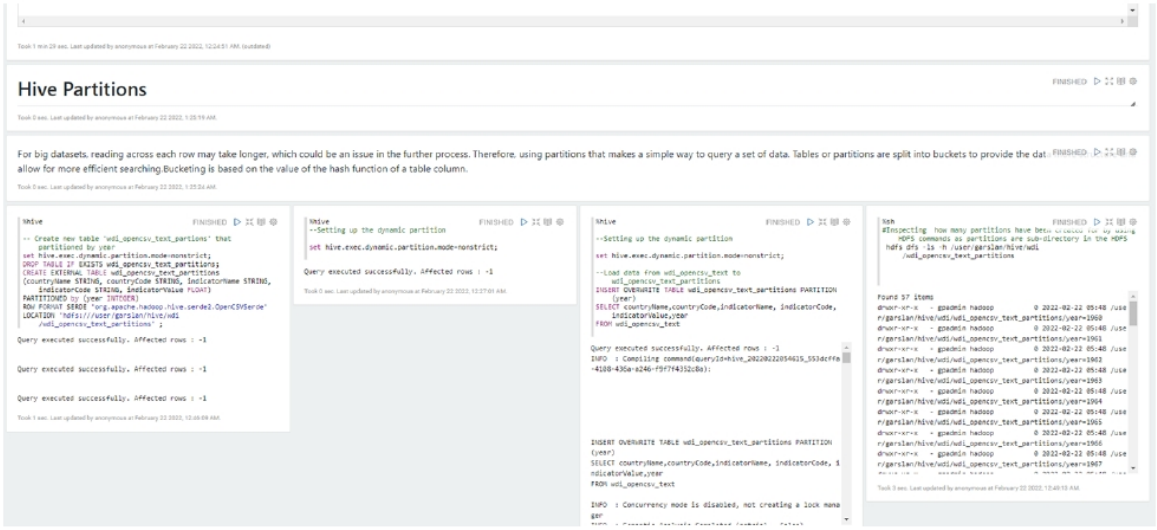
```
%hive
SELECT year ,countryName , indicatorValue AS GDP_growth_value
FROM wdi_opencsv_text_view
WHERE indicatorName LIKE "%GDP growth%" AND year =2015 AND countryName ='Canada'
```

FINISHED

```
INFO  : Compiling command(queryId=hive_20220223052322_bb87a695-c07f-474e-8f30-5337e9d30cf7):
SELECT year ,countryName , indicatorValue AS GDP_growth_value
FROM wdi_opencsv_text_view
WHERE indicatorName LIKE "%GDP growth%" AND year =2015 AND countryName ='Canada'

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:countryname, type:string, comment:null), FieldSchema(name:gdp_growth_value, type:float, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20220223052322_bb87a695-c07f-474e-8f30-5337e9d30cf7); Time taken: 0.375 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220223052322_bb87a695-c07f-474e-8f30-5337e9d30cf7):
SELECT year ,countryName , indicatorValue AS GDP_growth_value
FROM wdi_opencsv_text_view
WHERE indicatorName LIKE "%GDP growth%" AND year =2015 AND countryName ='Canada'

INFO  : Query ID = hive_20220223052322_bb87a695-c07f-474e-8f30-5337e9d30cf7
INFO  : Total jobs = 1
```

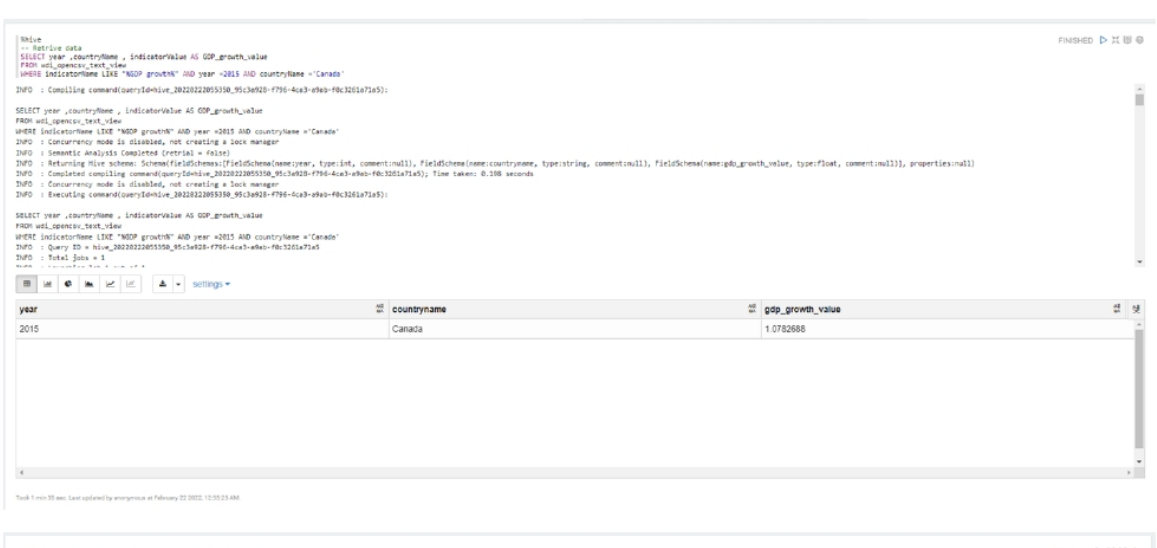
settings

year	countryname	gdp_growth_value
2015	Canada	1.0782688

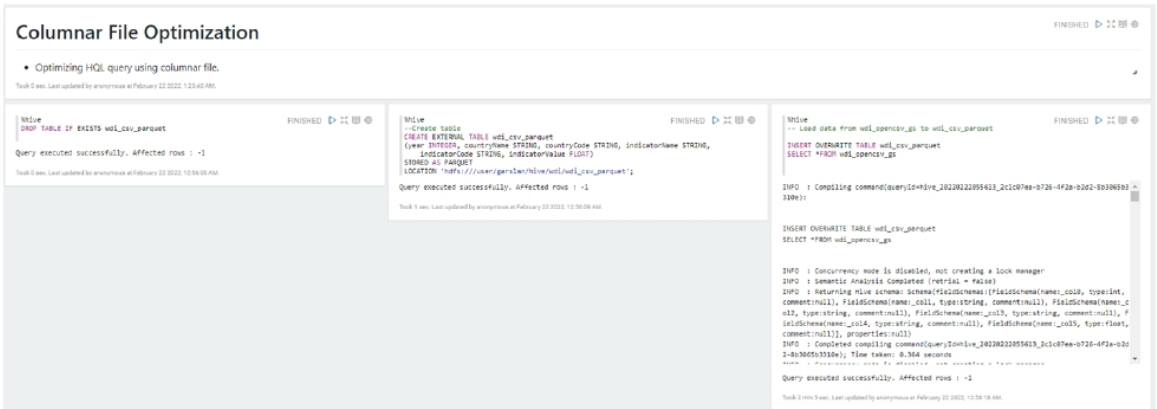


Hive Partitions

For big datasets, reading across each row may take longer, which could be an issue in the further process. Therefore, using partitions that makes a simple way to query a set of data. Tables or partitions are split into buckets to provide the dat... allow for more efficient searching.Bucketing is based on the value of the hash function of a table column.



year	countryname	gdp_growth_value
2015	Canada	1.0782688



Columnar File Optimization

- Optimizing HQL query using columnar file.

```
%sh
#Inspecting  how many parquet have been created for by using HDFS commands as partitions are sub-directory in the HDFS
 hdfs dfs -ls -h /user/garslan/hive/wdi/wdi_csv_parquet

Found 5 items
-rw-r--r--   2 gpadmin hadoop     20.9 M 2022-02-22 05:57 /user/garslan/hive/wdi/wdi_csv_parquet/000000_0
-rw-r--r--   2 gpadmin hadoop     30.0 M 2022-02-22 05:57 /user/garslan/hive/wdi/wdi_csv_parquet/000001_0
-rw-r--r--   2 gpadmin hadoop     30.0 M 2022-02-22 05:57 /user/garslan/hive/wdi/wdi_csv_parquet/000002_0
-rw-r--r--   2 gpadmin hadoop     29.9 M 2022-02-22 05:58 /user/garslan/hive/wdi/wdi_csv_parquet/000003_0
-rw-r--r--   2 gpadmin hadoop     17.4 M 2022-02-22 05:58 /user/garslan/hive/wdi/wdi_csv_parquet/000004_0
```

Took 2 sec. Last updated by anonymous at February 22 2022, 1:2:59:17 AM.

FINISHED

```
%sh
 hdfs dfs -ls -h /user/garslan/hive/wdi/wdi_opencsv_text

ls: `/user/garslan/hive/wdi/wdi_opencsv_text': No such file or directory

ExitValue: 1
```

Took 2 sec. Last updated by anonymous at February 22 2022, 1:2:59:27 AM.

ERROR

Runtime comparison

Took 0 sec. Last updated by anonymous at February 22 2022, 1:23:45 AM.

FINISHED

```
%hive
SELECT count(countryName) FROM wdi_csv_parquet;

INFO  : Compiling command(queryId=hive_20220222060816_c6fc5e93-e9a0-41da-8d8f-a27d0a69f75a):

SELECT count(countryName) FROM wdi_csv_parquet
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:_c0, type:bigint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20220222060816_c6fc5e93-e9a0-41da-8d8f-a27d0a69f75a); Time taken: 0.287 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220222060816_c6fc5e93-e9a0-41da-8d8f-a27d0a69f75a):

SELECT count(countryName) FROM wdi_csv_parquet
INFO  : Query ID = hive_20220222060816_c6fc5e93-e9a0-41da-8d8f-a27d0a69f75a
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_20220222060816_c6fc5e93-e9a0-41da-8d8f-a27d0a69f75a
INFO  : Tez session hasn't been created yet. Opening session
```

settings

_c0
21759408

Took 28 sec. Last updated by anonymous at February 22 2022, 1:01:42 AM. (outdated)

FINISHED

```
%hive
SELECT count(countryName) FROM wdi_opencsv_text;

INFO  : Compiling command(queryId=hive_20220222060849_8dc062e6-ee5a-4788-88bc-ab569af8824c):
SELECT count(countryName) FROM wdi_opencsv_text
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:_c0, type:bigint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20220222060849_8dc062e6-ee5a-4788-88bc-ab569af8824c); Time taken: 0.242 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220222060849_8dc062e6-ee5a-4788-88bc-ab569af8824c):
SELECT count(countryName) FROM wdi_opencsv_text
INFO  : Query ID = hive_20220222060849_8dc062e6-ee5a-4788-88bc-ab569af8824c
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_20220222060849_8dc062e6-ee5a-4788-88bc-ab569af8824c
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: SELECT count(countryName)...wdi_opencsv_text (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1645500859648_0012)
```

settings

_c0
21759408

Took 1 min 33 sec. Last updated by anonymous at February 22 2022, 1:03:38 AM. (outdated)

FINISHED

```
%hive
-- Retrive data from wdi_csv_parquet
SELECT year ,countryname , indicatorvalue AS GDP_growth_value
FROM wdi_csv_parquet
WHERE indicatorname LIKE "%GDP growth%" AND year =2015 AND countryname ='Canada'


INFO  : Compiling command(queryId=hive_20220222060906_e8319c27-81d6-4796-9dc4-e833dc93cd0c):

SELECT year ,countryname , indicatorvalue AS GDP_growth_value
FROM wdi_csv_parquet
WHERE indicatorname LIKE "%GDP growth%" AND year =2015 AND countryname ='Canada'


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:countryname, type:string, comment:null), FieldSchema(name:gdp_growth_value, type:float, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20220222060906_e8319c27-81d6-4796-9dc4-e833dc93cd0c); Time taken: 0.182 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220222060906_e8319c27-81d6-4796-9dc4-e833dc93cd0c):

SELECT year ,countryname , indicatorvalue AS GDP_growth_value
FROM wdi_csv_parquet
```

settings

year	countryname	gdp_growth_value
2015	Canada	1.0782688

Took 21 sec. Last updated by anonymous at February 22 2022, 1:05:07 AM. (outdated)

FINISHED

```
%hive
-- Retrive data from wdi_opencsv_text
SELECT year ,countryname , indicatorvalue AS GDP_growth_value
FROM wdi_opencsv_text
WHERE indicatorname LIKE "%GDP growth%" AND year =2015 AND countryname ='Canada'

INFO  : Compiling command(queryId=hive_20220222060549_9c1d1f67-0b6a-4266-8c2e-332795711b9c):

SELECT year ,countryname , indicatorvalue AS GDP_growth_value
FROM wdi_opencsv_text
WHERE indicatorname LIKE "%GDP growth%" AND year =2015 AND countryname ='Canada'
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:string, comment:null), FieldSchema(name:countryname, type:string, comment:null), FieldSchema(name:gdp_growth_value, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20220222060549_9c1d1f67-0b6a-4266-8c2e-332795711b9c); Time taken: 0.266 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220222060549_9c1d1f67-0b6a-4266-8c2e-332795711b9c):

SELECT year ,countryname , indicatorvalue AS GDP_growth_value
FROM wdi_opencsv_text
WHERE indicatorname LIKE "%GDP growth%" AND year =2015 AND countryname ='Canada'
INFO  : Query ID = hive_20220222060549_9c1d1f67-0b6a-4266-8c2e-332795711b9c
```

settings

year	countryname	gdp_growth_value
2015	Canada	1.0782687507638 1

Took 1 min 39 sec. Last updated by anonymous at February 22 2022, 1:07:28 AM. (outdated)

while wdi_csv_parquet took 31 seconds. wdi_opencsv_text took 1 min 39 seconds. Columnar file optimization drastically reduced the runtime.

Took 0 sec. Last updated by anonymous at February 22 2022, 1:10:52 AM.

FINISHED

Highest GDP Growth

FINISHED

Took 0 sec. Last updated by anonymous at February 22 2022, 1:05:58 AM.

FINISHED

```
%hive

--Finding  the highest GDP growth (NY.GDP.MKTP.KD.ZG) year for each country.
SELECT w.year ,w.countryName , w.indicatorValue AS GDP_growth_value
FROM
(
   SELECT max(indicatorValue) as value, countryName
   FROM wdi_csv_parquet
   WHERE indicatorCode = 'NY.GDP.MKTP.KD.ZG' AND indicatorValue!=0
   GROUP BY countryName )t
 INNER JOIN wdi_csv_parquet w ON t.value = w.indicatorValue AND t.countryname=w.countryName
 ORDER BY GDP_growth_value DESC;
```

```
Reducer 4       container       INITED     1          0        0        1       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/05  [=====================>>----] 84%   ELAPSED TIME: 75.66 s
----------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Map 5 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 05/05  [==========================>>] 100%  ELAPSED TIME: 75.64 s
INFO  : Completed executing command(queryId=hive_20220222062037_0ce03c9c-a6b0-41aa-989c-d7ca5a095752); Time taken: 93.985 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
```

settings

year	countryname	gdp_growth_value
1965	Oman	169.82993
1997	Equatorial Guinea	149.97296
1997	Liberia	106.27981
2012	Libya	104.486786
1996	Bosnia and Herzegovina	88.957664
1990	Iraq	57.81783
1974	Kiribati	45.302753
1974	Gabon	39.487095

Took 1 min 35 sec. Last updated by anonymous at February 22 2022, 1:22:12 AM. (outdated)

SPARK JOB FINISHED

```
%spark.sql

--Executing the query using Spark to compare execution time

SELECT w.year ,w.countryName , w.indicatorValue AS GDP_growth_value
FROM
( SELECT max(indicatorValue) as value, countryName
   FROM wdi_csv_parquet
   WHERE indicatorCode = 'NY.GDP.MKTP.KD.ZG' AND indicatorValue!=0
   GROUP BY countryName
   )m
 INNER JOIN wdi_csv_parquet w ON m.value = w.indicatorValue AND w.countryname =m.countryname
 ORDER BY GDP_growth_value DESC
```

settings

SPARK JOB FINISHED


```
%spark.sql

--Executing the query using Spark to compare execution time

SELECT w.year ,w.countryName , w.indicatorValue AS GDP_growth_value
FROM
( SELECT max(indicatorValue) as value, countryName
   FROM wdi_csv_parquet
   WHERE indicatorCode = 'NY.GDP.MKTP.KD.ZG' AND indicatorValue!=0
   GROUP BY countryName
   )m
 INNER JOIN wdi_csv_parquet w ON m.value = w.indicatorValue AND w.countryname =m.countryname
 ORDER BY GDP_growth_value DESC
```


settings

year	countryName	GDP_growth_value
1965	Oman	169.82993
1997	Equatorial Guinea	149.97296
1997	Liberia	106.27981
2012	Libya	104.486786
1996	Bosnia and Herzegovina	88.957664
1990	Iraq	57.81783
1974	Kiribati	45.302753
1974	Gabon	39.487095

Took 1 min 42 sec. Last updated by anonymous at February 22 2022, 1:32:51 AM. (outdated)

Sort GDP by country and year

FINISHED

Took 0 sec. Last updated by anonymous at February 22 2022, 1:33:42 AM.

ERROR

```
-- Retrieve GDP growth for all countries sorted by country name and year.

SELECT countryName, year,indicatorCode, indicatorValue
FROM wdi_csv_parquet
WHERE  indicatorCode ='NY.GDP.MKTP.KD.ZG'
ORDER BY countryName , year;
```

```
INFO  : Compiling command(queryId=hive_20220222064221_17f54920-cff0-4a68-bc2a-da843cd4e57d):


SELECT countryName, year,indicatorCode, indicatorValue
FROM wdi_csv_parquet
WHERE  indicatorCode ='NY.GDP.MKTP.KD.ZG'
ORDER BY countryName , 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:countryname, type:string, comment:null), FieldSchema(name:year, type:int, comment:null), FieldSchema(name:indicatorcode, type:string, comment:null), FieldSchema(name:indicatorvalue, type:float, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20220222064221_17f54920-cff0-4a68-bc2a-da843cd4e57d); Time taken: 0.132 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
```

settings

countryname	year	indicatorcode
Afghanistan	1993	NY.GDP.MKTP.KD.ZG
Afghanistan	1994	NY.GDP.MKTP.KD.ZG
Afghanistan	1995	NY.GDP.MKTP.KD.ZG
Afghanistan	1996	NY.GDP.MKTP.KD.ZG
Afghanistan	1997	NY.GDP.MKTP.KD.ZG
Afghanistan	1998	NY.GDP.MKTP.KD.ZG
Afghanistan	1999	NY.GDP.MKTP.KD.ZG
Afghanistan	2000	NY.GDP.MKTP.KD.ZG