Jarvis Hive Project

default

```
%md
### Create the 'wdi_gs' table through Zeppelin this time
- Drop table if it exists statement
- DDL statement for 'wdi_gs'
- Show metadata of table
- Run simple DML statement to test
```

FINISHED

Create the 'wdi_gs' table through Zeppelin this time

- Drop table if it exists statement
- DDL statement for 'wdi_gs'
- Show metadata of table
- Run simple DML statement to test

Took 1 sec. Last updated by anonymous at March 20 2025, 12:23:45 PM. (outdated)

```
-- Drop table if exists, remember semicolons don't work in Zeppelin
DROP TABLE IF EXISTS wdi_gs
```

FINISHED

```
INFO

Query executed successfully. Affected rows : -1
 : Compiling command(queryId=hive_20250320160523_fb409379-2671-471e-85a6-d859c459094d):
DROP TABLE IF EXISTS wdi_gs
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_20250320160523_fb409379-2671-471e-85a6-d859c459094d); Time taken: 1.464 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250320160523_fb409379-2671-471e-85a6-d859c459094d):
DROP TABLE IF EXISTS wdi_gs
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250320160523_fb409379-2671-471e-85a6-d859c459094d); Time taken: 1.936 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
```

Took 9 sec. Last updated by anonymous at March 20 2025, 12:05:28 PM. (outdated)

```
-- Create Table DDL Statement
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_ahmedamer/datasets/wdi_2016'
TBLPROPERTIES ("skip.header.line.count"="1")
```

FINISHED

```
INFO  : Compiling command(queryId=hive_20250320160529_864aefad-f905-4ca3-bc94-ece45dbf6868):
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_ahmedamer/datasets/wdi_2016'
TBLPROPERTIES ("skip.header.line.count"="1")
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_20250320160529_864aefad-f905-4ca3-bc94-ece45dbf6868); Time taken: 0.72 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250320160529_864aefad-f905-4ca3-bc94-ece45dbf6868):
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_ahmedamer/datasets/wdi_2016'
TBLPROPERTIES ("skip.header.line.count"="1")

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

Took 4 sec. Last updated by anonymous at March 20 2025, 12:05:32 PM. (outdated)

```
-- Describe the table and show metadata
DESCRIBE FORMATTED wdi_gs
```

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 March 20 2025, 12:05:33 PM. (outdated)

```
-- SELECT COUNT for testing
SELECT count(countryName) from wdi_gs
```

HIVE JOB FINISHED

```
INFO  : Compiling command(queryId=hive_20250320160534_c18d7319-c18a-4acb-b518-60927cb385ad):
SELECT count(countryName) 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:_c0, type:bigint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250320160534_c18d7319-c18a-4acb-b518-60927cb385ad); Time taken: 2.868 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250320160534_c18d7319-c18a-4acb-b518-60927cb385ad):
SELECT count(countryName) from wdi_gs
INFO  : Query ID = hive_20250320160534_c18d7319-c18a-4acb-b518-60927cb385ad
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_20250320160534_c18d7319-c18a-4acb-b518-60927cb385ad
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: SELECT count(countryName) from wdi_gs (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1742485905867_0001)
```

settings

_c0
21759408

Took 51 sec. Last updated by anonymous at March 20 2025, 12:06:24 PM. (outdated)

```
%md
### Create External Table 'wdi_csv_text' for loading data into HDFS
```

FINISHED

Create External Table 'wdi_csv_text' for loading data into HDFS

Took 0 sec. Last updated by anonymous at March 20 2025, 12:06:24 PM. (outdated)

```
-- Drop if exists
DROP TABLE IF EXISTS wdi_csv_text
```

FINISHED

```
INFO  : Compiling command(queryId=hive_20250320160625_f8e377a7-8430-4efa-aba4-718324dbe42b):
DROP TABLE IF EXISTS 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:null, properties:null)


Query exINFO  : Completed compiling command(queryId=hive_20250320160625_f8e377a7-8430-4efa-aba4-718324dbe42b); Time taken: 0.069 seconds
INFO  : Concurrencyc modutee isd  dsucicseassfbleuld,ly no. tAf fcercetaetind g ra owlso ck:  -m1an
ager
INFO  : Executing command(queryId=hive_20250320160625_f8e377a7-8430-4efa-aba4-718324dbe42b):
DROP TABLE IF EXISTS wdi_csv_text
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250320160625_f8e377a7-8430-4efa-aba4-718324dbe42b); Time taken: 0.907 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
```

Took 2 sec. Last updated by anonymous at March 20 2025, 12:06:26 PM. (outdated)

```
-- Create External Table 'wdi_csv_text'
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/ahmedamer/hive/wdi/wdi_csv_text'
```

FINISHED

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

Took 1 sec. Last updated by anonymous at March 20 2025, 12:06:27 PM. (outdated)

```
-- We need to transfer data from wdi_gs to wdi_csv_text
-- This is so that our data will now reside on the HDFS and not the GCP storage
INSERT OVERWRITE TABLE wdi_csv_text
SELECT * FROM wdi_gs
```

HIVE JOB FINISHED

```
INFO  : Compiling command(queryId=hive_20250320160627_f9041d23-3756-4aa9-bbac-02ca4be02251):

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), FieldSche
ma(name:wdi_gs.countrycode, type:string, comment:null), FieldSchema(name:wdi_gs.indicatorname, type:string, comment:null), FieldSchema(name:wdi_gs.indicatorcode, type:string, commen
t:null), FieldSchema(name:wdi_gs.indicatorvalue, type:float, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250320160627_f9041d23-3756-4aa9-bbac-02ca4be02251); Time taken: 0.531 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250320160627_f9041d23-3756-4aa9-bbac-02ca4be02251):

INSERT OVERWRITE TABLE wdi_csv_text
SELECT * FROM wdi_gs
INFO  : Query ID = hive_20250320160627_f9041d23-3756-4aa9-bbac-02ca4be02251
INFO  : Total jobs = 1
```

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

Took 1 min 34 sec. Last updated by anonymous at March 20 2025, 12:08:01 PM. (outdated)

```
-- Check if the data was successfully transferred, this has to be done by SSHing to the master node
-- Run the bash cmds 'hdfs dfs -ls -h /user/ahmedamer/hive/wdi/wdi_csv_text' : we see a total of 1.7GB data
```

FINISHED

Took 0 sec. Last updated by anonymous at March 20 2025, 12:08:01 PM. (outdated)

```
-- Now we will demonstrate filesystem cache benefits by running a SELECT statement twice
-- To clear cache of worker nodes use bash cmd : sudo rm -rf /var/cache/*
-- Needs to be used in the nodes themselves

SELECT count(countryName) FROM wdi_csv_text
```

HIVE JOB FINISHED

```
INFO  : Compiling command(queryId=hive_20250320160801_95d3a0dc-f27e-4e53-9566-e3dffa053a06):



SELECT count(countryName) 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:_c0, type:bigint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250320160801_95d3a0dc-f27e-4e53-9566-e3dffa053a06); Time taken: 0.241 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250320160801_95d3a0dc-f27e-4e53-9566-e3dffa053a06):



SELECT count(countryName) FROM wdi_csv_text
INFO  : Query ID = hive_20250320160801_95d3a0dc-f27e-4e53-9566-e3dffa053a06
INFO  : Total jobs = 1
```

settings

_c0
21759408

Took 30 sec. Last updated by anonymous at March 20 2025, 12:08:31 PM. (outdated)

```
-- Second time : note execution times 40.346 for the first and 23.64 for the second

SELECT count(countryName) FROM wdi_csv_text
```

HIVE JOB FINISHED

```
INFO  : Compiling command(queryId=hive_20250320160831_8706bafa-0042-42b5-a121-9a220f5db09e):

SELECT count(countryName) 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:_c0, type:bigint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250320160831_8706bafa-0042-42b5-a121-9a220f5db09e); Time taken: 0.23 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250320160831_8706bafa-0042-42b5-a121-9a220f5db09e):

SELECT count(countryName) FROM wdi_csv_text
INFO  : Query ID = hive_20250320160831_8706bafa-0042-42b5-a121-9a220f5db09e
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_20250320160831_8706bafa-0042-42b5-a121-9a220f5db09e
INFO  : Tez session hasn't been created yet. Opening session
```

settings

_c0
21759408

Took 33 sec. Last updated by anonymous at March 20 2025, 12:09:04 PM. (outdated)

FINISHED

```
%sh

# Let's compare the task of finding the count of rows through bash commands as well

cd ~

hdfs dfs -get hdfs:///user/ahmedamer/hive/wdi/wdi_csv_text .
cd wdi_csv_text

# Display the total size of the data
du -ch .

# Clear cache
echo 3 | sudo tee /proc/sys/vm/drop_caches

# Row count functions
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
get: `wdi_csv_text/000005_0': File exists
get: `wdi_csv_text/000006_0': File exists
get: `wdi_csv_text/000007_0': File exists
get: `wdi_csv_text/000008_0': File exists
get: `wdi_csv_text/000009_0': File exists
get: `wdi_csv_text/000010_0': File exists
1.8G    .
1.8G    total
sudo: a terminal is required to read the password; either use the -S option to read from standard input or configure an askpass helper
sudo: a password is required
1742486952
21759408 179709942 1838962843
```

Took 29 sec. Last updated by anonymous at March 20 2025, 12:09:33 PM. (outdated)

FINISHED

```
%md
### Hive vs Bash
You can see that using bash compared to just HiveQL is more complicated and the process takes longer (47 seconds):
- Pulling the file from HDFS consumes bandwith, memory and CPU
- The computation is only executed on the master node

HiveQL transforms the SQL-like code into MapReduce jobs that span and utilize the entire cluster for efficiency
```

Hive vs Bash

You can see that using bash compared to just HiveQL is more complicated and the process takes longer (47 seconds):

- Pulling the file from HDFS consumes bandwith, memory and CPU
- The computation is only executed on the master node

HiveQL transforms the SQL-like code into MapReduce jobs that span and utilize the entire cluster for efficiency

Took 0 sec. Last updated by anonymous at March 20 2025, 12:09:34 PM. (outdated)

HIVE JOB FINISHED

```
-- Let's run more queries using HiveQL

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

```
INFO  : Compiling command(queryId=hive_20250320160934_caafe0c5-3d56-46fc-b5bb-a6d0e9139949):

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_20250320160934_caafe0c5-3d56-46fc-b5bb-a6d0e9139949); Time taken: 0.331 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250320160934_caafe0c5-3d56-46fc-b5bb-a6d0e9139949):

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 36 sec. Last updated by anonymous at March 20 2025, 12:10:10 PM. (outdated)

```
%md
### Parsing Problem
It appears we have found a problem with which the column 'indicatorcode' is not being parsed correctly by SerDe. Let's make a debug table to investigate whole rows instead of columns.
```

FINISHED

Parsing Problem

It appears we have found a problem with which the column ‘indicatorcode’ is not being parsed correctly by SerDe. Let’s make a debug table to investigate whole rows instead of columns.

Took 0 sec. Last updated by anonymous at March 20 2025, 12:10:11 PM. (outdated)

```
DROP TABLE IF EXISTS wdi_gs_debug
```

FINISHED

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

Took 0 sec. Last updated by anonymous at March 20 2025, 12:10:11 PM. (outdated)

```
CREATE EXTERNAL TABLE wdi_gs_debug (debug_col STRING)
    STORED AS TEXTFILE
    LOCATION 'gs://jarvis_data_eng_ahmedamer/datasets/wdi_2016'
    TBLPROPERTIES ("skip.header.line.count" = "1")
```

FINISHED

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

Took 1 sec. Last updated by anonymous at March 20 2025, 12:10:12 PM. (outdated)

```
SELECT debug_col FROM wdi_gs_debug
    WHERE debug_col LIKE "%\(\% of urban population\)\"%"
```

HIVE JOB FINISHED

```
INFO  : Compiling command(queryId=hive_20250320161013_4b21a523-a356-43f7-86fc-8cbb67864017): SELECT debug_col FROM wdi_gs_debug
    WHERE debug_col LIKE "%\(\% of urban population\)\"%"
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:debug_col, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250320161013_4b21a523-a356-43f7-86fc-8cbb67864017); Time taken: 0.325 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250320161013_4b21a523-a356-43f7-86fc-8cbb67864017): SELECT debug_col FROM wdi_gs_debug
    WHERE debug_col LIKE "%\(\% of urban population\)\"%"
INFO  : Query ID = hive_20250320161013_4b21a523-a356-43f7-86fc-8cbb67864017
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_20250320161013_4b21a523-a356-43f7-86fc-8cbb67864017
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: SELECT debug_col FROM wdi...population\)\"%" (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1742485905867_0006)
```

settings

debug_col
1960,Argentina,ARG,"Access to non-solid fuel, urban (% of urban population)",EG.NSF.ACCS.UR.ZS,0
1960,American Samoa,ASM,"Access to electricity, urban (% of urban population)",EG.ELC.ACCS.UR.ZS,0
1960,Belgium,BEL,"People practicing open defecation, urban (% of urban population)",SH.STA.ODFC.UR.ZS,0
1960,Bangladesh,BGD,"Access to non-solid fuel, urban (% of urban population)",EG.NSF.ACCS.UR.ZS,0
1960,Barbados,BRB,"People practicing open defecation, urban (% of urban population)",SH.STA.ODFC.UR.ZS,0

```
1960,Brunei Darussalam,BRN,"Access to non-solid fuel, urban (% of urban population)",EG.NSF.ACCS.UR.ZS,0
1960,Central Europe and the Baltics,CEB,"Access to non-solid fuel, urban (% of urban population)",EG.NSF.ACCS.UR.ZS,0
1960,China,CHN,"People practicing open defecation, urban (% of urban population)",SH.STA.ODFC.UR.ZS,0
```

Took 1 min 5 sec. Last updated by anonymous at March 20 2025, 12:11:18 PM. (outdated)

```
%md
### Debugging
Through querying the debug table, we are able to identify the reason there are parsing problems (Quotations in indicatorName column are not being recognized, so the comma in the middle is being used as a delimiter).
```

FINISHED

Debugging

Through querying the debug table, we are able to identify the reason there are parsing problems (Quotations in indicatorName column are not being recognized, so the comma in the middle is being used as a delimiter).

Took 0 sec. Last updated by anonymous at March 20 2025, 12:21:47 PM. (outdated)

```
%md
Let's use OPENCSV SerDe to create a solution for this.
```

FINISHED

Let's use OPENCSV SerDe to create a solution for this.

Took 0 sec. Last updated by anonymous at March 20 2025, 12:11:18 PM. (outdated)

```
DROP TABLE IF EXISTS wdi_opencsv_gs
```

FINISHED

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

Took 0 sec. Last updated by anonymous at March 20 2025, 12:11:19 PM. (outdated)

```
-- With OpenCSVSerDe create new table
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'
    WITH SERDEPROPERTIES (
        "separatorChar" = ",",
        "quoteChar" = "\"",
        "escapeChar" = "\\"
    )
    STORED AS TEXTFILE
    LOCATION 'gs://jarvis_data_eng_ahmedamer/datasets/wdi_2016'
    TBLPROPERTIES ("skip.header.line.count"="1")
```

FINISHED

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

Took 1 sec. Last updated by anonymous at March 20 2025, 12:11:20 PM. (outdated)

```
DROP TABLE IF EXISTS wdi_opencsv_text
```

FINISHED

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

Took 1 sec. Last updated by anonymous at March 20 2025, 12:11:21 PM. (outdated)

```
-- Let's make a destination table for the data in HDFS
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'
    WITH SERDEPROPERTIES (
        "separatorChar" = ",",
        "quoteChar" = "\"",
        "escapeChar" = "\\"
    )
LOCATION 'hdfs:///user/ahmedamer/hive/wdi/wdi_opencsv_text'
```

FINISHED

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

Took 0 sec. Last updated by anonymous at March 20 2025, 12:11:21 PM. (outdated)

```
-- Transfer statement from wdi_opencsv_gs to wdi_opencsv_text
INSERT OVERWRITE TABLE wdi_opencsv_text
SELECT * FROM wdi_opencsv_gs
```

HIVE JOB FINISHED

```
INFO  : Compiling command(queryId=hive_20250320161122_e07d07f5-783b-43a4-82dc-a0e6d189cadc):
INSERT OVERWRITE TABLE wdi_opencsv_text
SELECT * FROM wdi_opencsv_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_opencsv_gs.year, type:string, comment:null), FieldSchema(name:wdi_opencsv_gs.countryname, type:string, comment:null), 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.indicatorcode, type:string, comment:null), FieldSchema(name:wdi_opencsv_gs.indicatorvalue, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250320161122_e07d07f5-783b-43a4-82dc-a0e6d189cadc); Time taken: 0.443 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250320161122_e07d07f5-783b-43a4-82dc-a0e6d189cadc):
INSERT OVERWRITE TABLE wdi_opencsv_text
SELECT * FROM wdi_opencsv_gs
INFO  : Query ID = hive_20250320161122_e07d07f5-783b-43a4-82dc-a0e6d189cadc
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
```

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

Took 1 min 42 sec. Last updated by anonymous at March 20 2025, 12:13:03 PM. (outdated)

```
-- Let's verify that the parsing worked this time!
SELECT distinct(indicatorcode) FROM wdi_opencsv_text
    ORDER BY indicatorcode
        LIMIT 20
```

HIVE JOB FINISHED

```
INFO  : Compiling command(queryId=hive_20250320161303_728fa5fe-4953-45dd-965a-e9d901f33d78):
SELECT distinct(indicatorcode) FROM wdi_opencsv_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_20250320161303_728fa5fe-4953-45dd-965a-e9d901f33d78); Time taken: 0.221 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250320161303_728fa5fe-4953-45dd-965a-e9d901f33d78):
SELECT distinct(indicatorcode) FROM wdi_opencsv_text
    ORDER BY indicatorcode
        LIMIT 20
INFO  : Query ID = hive_20250320161303_728fa5fe-4953-45dd-965a-e9d901f33d78
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
```

settings

indicatorcode
AG.AGR.TRAC.NO
AG.CON.FERT.PT.ZS
AG.CON.FERT.ZS
AG.LND.AGRI.K2
AG.LND.AGRI.ZS
AG.LND.ARBL.HA
AG.LND.ARBL.HA.PC
AG.LND.ARBL.ZS

Took 1 min 19 sec. Last updated by anonymous at March 20 2025, 12:14:22 PM. (outdated)

```
%md
### OpenCSVSerde Limitations
Now that we can parse the columns correctly, we ask ourselves about the limitations of OpenCSVSerde and we see that it treats all columns as String types as seen if we describe the
'wdi_opencsv_text' table.
```

FINISHED

OpenCSVSerde Limitations

Now that we can parse the columns correctly, we ask ourselves about the limitations of OpenCSVSerde and we see that it treats all columns as String types as seen if we describe the 'wdi_opencsv_text' table.

Took 0 sec. Last updated by anonymous at March 20 2025, 12:14:22 PM. (outdated)

```
DESCRIBE FORMATTED wdi_opencsv_text
```

FINISHED

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 0 sec. Last updated by anonymous at March 20 2025, 12:14:22 PM. (outdated)

```
%md
We need to create a view on top of this table in order to correct this limitation. We need to have indicatorValue as FLOAT.
```

FINISHED

We need to create a view on top of this table in order to correct this limitation. We need to have indicatorValue as FLOAT.

Took 0 sec. Last updated by anonymous at March 20 2025, 12:14:23 PM. (outdated)

```
CREATE VIEW IF NOT EXISTS wdi_opencsv_text_view
AS
SELECT year, countryName, countryCode, indicatorName, indicatorCode, CAST(indicatorValue AS FLOAT) AS indicatorValue
FROM wdi_opencsv_text
```

FINISHED

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

Took 0 sec. Last updated by anonymous at March 20 2025, 12:14:23 PM. (outdated)

```
%md
### HiveQL Optimizations
Let's compute the 2015 GDP Growth for Canada and then let's discuss optimization stategies to make queries run faster.
```

FINISHED

HiveQL Optimizations

Let's compute the 2015 GDP Growth for Canada and then let's discuss optimization stategies to make queries run faster.

Took 0 sec. Last updated by anonymous at March 20 2025, 12:14:24 PM. (outdated)

```
-- Here we find the correct indicatorCode to use
SELECT DISTINCT indicatorCode as indicatorCode, countryName FROM wdi_opencsv_text_view
WHERE countryName = 'Canada' AND indicatorName LIKE 'GDP growth (annual %)'
```

HIVE JOB FINISHED

```
INFO  : Compiling command(queryId=hive_20250320161424_5a01f3f1-4def-4a35-94bb-2755aa893386):
SELECT DISTINCT indicatorCode as indicatorCode, countryName FROM wdi_opencsv_text_view
WHERE countryName = 'Canada' AND indicatorName LIKE 'GDP growth (annual %)'

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), FieldSchema(name:countryname, type:string, comment:null)], propertie
s:null)
INFO  : Completed compiling command(queryId=hive_20250320161424_5a01f3f1-4def-4a35-94bb-2755aa893386); Time taken: 0.408 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250320161424_5a01f3f1-4def-4a35-94bb-2755aa893386):
SELECT DISTINCT indicatorCode as indicatorCode, countryName FROM wdi_opencsv_text_view
WHERE countryName = 'Canada' AND indicatorName LIKE 'GDP growth (annual %)'

INFO  : Query ID = hive_20250320161424_5a01f3f1-4def-4a35-94bb-2755aa893386
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
```

settings

indicatorcode	countryname
NY.GDP.MKTP.KD.ZG	Canada

Took 1 min 12 sec. Last updated by anonymous at March 20 2025, 12:15:36 PM. (outdated)

```
-- Compute 2015 Canada GDP growth
SELECT indicatorValue AS GDP_growth_value, year, countryName FROM wdi_opencsv_text_view
WHERE year = '2015' AND indicatorCode = 'NY.GDP.MKTP.KD.ZG' AND countryName = 'Canada'
```

HIVE JOB FINISHED

```
INFO  : Compiling command(queryId=hive_20250320161536_c6b7666e-65a3-4783-9fbb-0bc16426d4a1):
SELECT indicatorValue AS GDP_growth_value, year, countryName FROM wdi_opencsv_text_view
WHERE year = '2015' AND indicatorCode = 'NY.GDP.MKTP.KD.ZG' 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:gdp_growth_value, type:float, comment:null), FieldSchema(name:year, type:string, comment:null), FieldSchema(nam
e:countryname, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250320161536_c6b7666e-65a3-4783-9fbb-0bc16426d4a1); Time taken: 0.237 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250320161536_c6b7666e-65a3-4783-9fbb-0bc16426d4a1):
SELECT indicatorValue AS GDP_growth_value, year, countryName FROM wdi_opencsv_text_view
WHERE year = '2015' AND indicatorCode = 'NY.GDP.MKTP.KD.ZG' AND countryName = 'Canada'
INFO  : Query ID = hive_20250320161536_c6b7666e-65a3-4783-9fbb-0bc16426d4a1
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_20250320161536_c6b7666e-65a3-4783-9fbb-0bc16426d4a1
```

settings

gdp_growth_value	year	countryname
1.0782688	2015	Canada

Took 1 min 14 sec. Last updated by anonymous at March 20 2025, 12:16:50 PM. (outdated)

```
%md
This query takes a long time (~50secs) to run, so how can we optimize it? Let's partition the data
```

FINISHED

This query takes a long time (~50secs) to run, so how can we optimize it? Let's partition the data

Took 0 sec. Last updated by anonymous at March 20 2025, 12:16:50 PM. (outdated)

```
DROP TABLE IF EXISTS wdi_opencsv_text_partitions
```

FINISHED

```
INFO  : Compiling command(queryId=hive_20250320161650_d7154a5a-adb9-45b1-a4ed-e20e9dcbde65): DROP TABLE IF EXISTS wdi_opencsv_text_partitions
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas

Query executed successfully. Affected rows : -1
:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20250320161650_d7154a5a-adb9-45b1-a4ed-e20e9dcbde65); Time taken: 0.044 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250320161650_d7154a5a-adb9-45b1-a4ed-e20e9dcbde65): DROP TABLE IF EXISTS wdi_opencsv_text_partitions
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250320161650_d7154a5a-adb9-45b1-a4ed-e20e9dcbde65); Time taken: 1.778 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
```

Took 3 sec. Last updated by anonymous at March 20 2025, 12:16:53 PM. (outdated)

```
-- Create a partition table
CREATE EXTERNAL TABLE wdi_opencsv_text_partitions
(countryName STRING, countryCode STRING, indicatorName STRING, indicatorCode STRING, indicatorValue FLOAT)
PARTITIONED BY (year STRING)
STORED AS TEXTFILE
LOCATION 'hdfs:///user/ahmedamer/hive/wdi/wdi_opencsv_text_partitions'
```

FINISHED

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

Took 0 sec. Last updated by anonymous at March 20 2025, 12:16:53 PM. (outdated)

```
-- Enable Dynamic Partitioning without static partitions, increase max partitions
SET hive.exec.dynamic.partition.mode=nonstrict;
SET hive.exec.dynamic.partition=TRUE;
-- SET hive.exec.max.dynamic.partitions=1000;
-- SET hive.exec.max.dynamic.partitions.pernode=500;


-- Load partitioned data
-- After experiencing errors where there were too many partitions being created, I decided to preaggregate the data
INSERT OVERWRITE TABLE wdi_opencsv_text_partitions
PARTITION (year)
SELECT countryName, countryCode, indicatorName, indicatorCode, indicatorValue, year
FROM wdi_opencsv_text_view
```

HIVE JOB FINISHED

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


Query executed successfully. Affected rows : -1
INFO  : Compiling command(queryId=hive_20250320161654_8d2b0580-9b69-4d03-8de2-9faaf4f5d60e):

INSERT OVERWRITE TABLE wdi_opencsv_text_partitions
PARTITION (year)
SELECT countryName, countryCode, indicatorName, indicatorCode, indicatorValue, year
FROM wdi_opencsv_text_view
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:countrycode, type:string, comment:null), FieldSchema(n

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

Took 2 min 18 sec. Last updated by anonymous at March 20 2025, 12:19:12 PM. (outdated)

```
-- Show Partitions and whether it was successful
SHOW PARTITIONS wdi_opencsv_text_partitions;
```

FINISHED

settings

partition

```
year=1960
year=1961
year=1962
year=1963
year=1964
year=1965
year=1966
year=1967
```

Took 1 sec. Last updated by anonymous at March 20 2025, 12:19:13 PM. (outdated)

```
%md
Let's rerun the same computation and find Canada's GDP growth in 2015 to compare execution times after partitioning. We will see it takes a considerably less time to fulfill!
```

FINISHED

Let's rerun the same computation and find Canada's GDP growth in 2015 to compare execution times after partitioning. We will see it takes a considerably less time to fulfill!

Took 0 sec. Last updated by anonymous at March 20 2025, 12:19:13 PM. (outdated)

```
SELECT indicatorValue AS GDP_growth_value, year, countryName FROM wdi_opencsv_text_partitions
WHERE year = '2015' AND indicatorCode = 'NY.GDP.MKTP.KD.ZG' AND countryName = 'Canada'
```

FINISHED

```
INFO  : Compiling command(queryId=hive_20250320161913_631b58a2-891d-40de-9621-c9d37d7f44d4): SELECT indicatorValue AS GDP_growth_value, year, countryName FROM wdi_opencsv_text_partiti
ons
WHERE year = '2015' AND indicatorCode = 'NY.GDP.MKTP.KD.ZG' AND countryName =%table gdp_growth_value    year    countryname
1.0782688       2015    Canada


 '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:gdp_growth_value, type:float, comment:null), FieldSchema(name:year, type:string, comment:null), FieldSchema(name:c
ountryname, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250320161913_631b58a2-891d-40de-9621-c9d37d7f44d4); Time taken: 0.744 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250320161913_631b58a2-891d-40de-9621-c9d37d7f44d4): SELECT indicatorValue AS GDP_growth_value, year, countryName FROM wdi_opencsv_text_partiti
ons
WHERE year = '2015' AND indicatorCode = 'NY.GDP.MKTP.KD.ZG' AND countryName = 'Canada'
INFO  : Completed executing command(queryId=hive_20250320161913_631b58a2-891d-40de-9621-c9d37d7f44d4); Time taken: 0.004 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
```

Took 2 sec. Last updated by anonymous at March 20 2025, 12:19:15 PM. (outdated)

```
%md
### Columnar File Optimization
Now we will explore storing the data as a Parquet file
```

FINISHED

Columnar File Optimization

Now we will explore storing the data as a Parquet file

Took 0 sec. Last updated by anonymous at March 20 2025, 12:19:15 PM. (outdated)

```
DROP TABLE IF EXISTS wdi_csv_parquet
```

FINISHED

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

Took 0 sec. Last updated by anonymous at March 20 2025, 12:19:15 PM. (outdated)

```
-- Create Parquet Table
CREATE EXTERNAL TABLE wdi_csv_parquet
(year STRING, countryName STRING, countryCode STRING, indicatorName STRING, indicatorCode STRING, indicatorValue STRING)
    -- ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    -- WITH SERDEPROPERTIES (
    --     "separatorChar" = ",",
    --     "quoteChar" = "\"",
    --     "escapeChar" = "\\"
    -- )
    STORED AS PARQUET
    LOCATION 'hdfs:///user/ahmedamer/hive/wdi/wdi_csv_parquet'
```

FINISHED

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

Took 1 sec. Last updated by anonymous at March 20 2025, 12:19:16 PM. (outdated)

```
-- Load data from wdi_opencsv_gs
INSERT OVERWRITE TABLE wdi_csv_parquet
SELECT *
FROM wdi_opencsv_gs
```

HIVE JOB FINISHED

```
INFO  : Compiling command(queryId=hive_20250320161916_75794d3e-3c76-4c7d-9794-c4acd5b92e0d):
INSERT OVERWRITE TABLE wdi_csv_parquet
SELECT *
FROM wdi_opencsv_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_opencsv_gs.year, type:string, comment:null), FieldSchema(name:wdi_opencsv_gs.countryname, type:string, comment:null), 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.indicatorcode, type:string, comment:null), FieldSchema(name:wdi_opencsv_gs.indicatorvalue, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250320161916_75794d3e-3c76-4c7d-9794-c4acd5b92e0d); Time taken: 0.294 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250320161916_75794d3e-3c76-4c7d-9794-c4acd5b92e0d):
INSERT OVERWRITE TABLE wdi_csv_parquet
SELECT *
FROM wdi_opencsv_gs
INFO  : Query ID = hive_20250320161916_75794d3e-3c76-4c7d-9794-c4acd5b92e0d
INFO  : Total jobs = 1
```

Query executed successfully. Affected rows : -1

Took 1 min 48 sec. Last updated by anonymous at March 20 2025, 12:21:04 PM. (outdated)

```
%md
Let's compare the sizes of the data stored as Parquet vs Textfile format
```

FINISHED

Let's compare the sizes of the data stored as Parquet vs Textfile format

Took 0 sec. Last updated by anonymous at March 20 2025, 12:21:04 PM. (outdated)

```
%sh
#SSH to master node
cd ~
hdfs  dfs -get  hdfs:///user/ahmedamer/hive/wdi/wdi_csv_parquet .
cd wdi_csv_parquet
#calculate current directory size
du -ch .
#217MB, much less than 1.8G of the textfile format
```

FINISHED

```
get: `wdi_csv_parquet/000000_0': File exists
get: `wdi_csv_parquet/000001_0': File exists
get: `wdi_csv_parquet/000002_0': File exists
get: `wdi_csv_parquet/000003_0': File exists
get: `wdi_csv_parquet/000004_0': File exists
get: `wdi_csv_parquet/000005_0': File exists
get: `wdi_csv_parquet/000006_0': File exists
get: `wdi_csv_parquet/000007_0': File exists
get: `wdi_csv_parquet/000008_0': File exists
get: `wdi_csv_parquet/000009_0': File exists
get: `wdi_csv_parquet/000010_0': File exists
217M    .
217M    total
```

Took 5 sec. Last updated by anonymous at March 20 2025, 12:21:09 PM. (outdated)

```
%md
Let's run the count query and compare execution times of that, too
```

FINISHED

Let's run the count query and compare execution times of that, too

Took 0 sec. Last updated by anonymous at March 20 2025, 12:21:09 PM. (outdated)

```
SELECT COUNT(countryName) FROM wdi_csv_parquet
```

HIVE JOB FINISHED

```
INFO  : Compiling command(queryId=hive_20250320162110_bb51565b-135b-497d-80a4-6d4a02137054): 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_20250320162110_bb51565b-135b-497d-80a4-6d4a02137054); Time taken: 0.158 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250320162110_bb51565b-135b-497d-80a4-6d4a02137054): SELECT COUNT(countryName) FROM wdi_csv_parquet

INFO  : Query ID = hive_20250320162110_bb51565b-135b-497d-80a4-6d4a02137054
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_20250320162110_bb51565b-135b-497d-80a4-6d4a02137054
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: SELECT COUNT(countryName) ...wdi_csv_parquet (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1742485905867_0013)
```

settings

_c0
21759408

Took 31 sec. Last updated by anonymous at March 20 2025, 12:21:41 PM. (outdated)

```
SELECT COUNT(countryName) FROM wdi_opencsv_text
```

HIVE JOB FINISHED

```
INFO  : Compiling command(queryId=hive_20250320162141_3c810a8a-c90c-41ab-8edd-2057598a3347): 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_20250320162141_3c810a8a-c90c-41ab-8edd-2057598a3347); Time taken: 0.161 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250320162141_3c810a8a-c90c-41ab-8edd-2057598a3347): SELECT COUNT(countryName) FROM wdi_opencsv_text
INFO  : Query ID = hive_20250320162141_3c810a8a-c90c-41ab-8edd-2057598a3347
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_20250320162141_3c810a8a-c90c-41ab-8edd-2057598a3347
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_1742485905867_0014)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
```

settings

_c0
21759408

Took 1 min 1 sec. Last updated by anonymous at March 20 2025, 12:22:42 PM. (outdated)

```
%md
We see ~25secs for the Parquet file and ~59secs for Textfile formatting. In this way the benefits of Columnar File formats can be seen:
- Vectorized Processing; batches of columnar data used instead of reading each row
- Perfect for aggregating large datasets
```

FINISHED

We see ~25secs for the Parquet file and ~59secs for Textfile formatting. In this way the benefits of Columnar File formats can be seen:

- Vectorized Processing; batches of columnar data used instead of reading each row
- Perfect for aggregating large datasets

Took 0 sec. Last updated by anonymous at March 20 2025, 12:22:42 PM. (outdated)

```
%md
### Spark vs Hive
Let's answer the question: 'Highest GDP growth year for each country'
```

FINISHED

Spark vs Hive

Let's answer the question: 'Highest GDP growth year for each country'

Took 0 sec. Last updated by anonymous at March 20 2025, 12:22:42 PM. (outdated)

```
-- This query uses a subquery with a window function that calculates the highest GDP year
-- Then we select that row via the outer query
SELECT indicatorValue as GDP_growth_value, year, countryName
FROM (
    SELECT countryName, year, indicatorValue,
        ROW_NUMBER() OVER (
            Partition By countryName ORDER BY indicatorValue DESC
        ) as rank
    FROM wdi_opencsv_text_partitions
    WHERE indicatorCode = 'NY.GDP.MKTP.KD.ZG'
) inner_select
WHERE rank = 1
ORDER BY countryName
```

HIVE JOB FINISHED

```
INFO  : Compiling command(queryId=hive_20250320162243_412d6ad2-f87a-4d23-b3dd-a661ff904668):

SELECT indicatorValue as GDP_growth_value, year, countryName
FROM (
    SELECT countryName, year, indicatorValue,
        ROW_NUMBER() OVER (
            Partition By countryName ORDER BY indicatorValue DESC
        ) as rank
    FROM wdi_opencsv_text_partitions
    WHERE indicatorCode = 'NY.GDP.MKTP.KD.ZG'
) inner_select
WHERE rank = 1
ORDER BY countryName
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
```

```
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:gdp_growth_value, type:float, comment:null), FieldSchema(name:year, type:string, comment:null), FieldSchema(name:countryname, type:string, comment:null)], properties:null)
```

settings

gdp_growth_value	year	countryname
21.020649	2009	Afghanistan
13.501173	1999	Albania
34.31373	1963	Algeria
0.0	1993	American Samoa
12.02392	2003	Andorra
22.593054	2007	Angola
13.3764	2006	Antigua and Barbuda
12.896334	1976	Arab World

Took 41 sec. Last updated by anonymous at March 20 2025, 12:23:23 PM. (outdated)

```
%md
Since we get an execution time of ~35secs in Hive, let's compare that to Spark (we see a time of ~45secs)
```

FINISHED

Since we get an execution time of ~35secs in Hive, let's compare that to Spark (we see a time of ~45secs)

Took 0 sec. Last updated by anonymous at March 20 2025, 12:23:24 PM. (outdated)

```
%spark.sql
SELECT indicatorValue as GDP_growth_value, year, countryName
FROM (
    SELECT countryName, year, indicatorValue,
        ROW_NUMBER() OVER (
            Partition By countryName ORDER BY indicatorValue DESC
        ) as rank
    FROM wdi_opencsv_text_partitions
    WHERE indicatorCode = 'NY.GDP.MKTP.KD.ZG'
) inner_select
WHERE rank = 1
ORDER BY countryName
```

SPARK JOB FINISHED

settings

GDP_growth_value	year	countryName
21.020649	2009	Afghanistan
13.501173	1999	Albania
34.31373	1963	Algeria
0.0	1994	American Samoa
12.02392	2003	Andorra
22.593054	2007	Angola
13.3764	2006	Antigua and Barbuda
12.896334	1976	Arab World

Took 1 min 27 sec. Last updated by anonymous at March 20 2025, 12:24:51 PM. (outdated)

```
%md
In most circumstances we would expect a performance advantage with Spark due to:
- Spark uses in-memory processing and doesn't write to the disk as much
- Optimized queries and better handling of complex queries

However, this is not what we see here and maybe this is due to:
- The data is transformed and partitioned already
- It's not that great of a dataset where using Spark would matter more
- The query is not too complex
```

FINISHED

In most circumstances we would expect a performance advantage with Spark due to:

- Spark uses in-memory processing and doesn't write to the disk as much
- Optimized queries and better handling of complex queries

However, this is not what we see here and maybe this is due to:

- The data is transformed and partitioned already
- It's not that great of a dataset where using Spark would matter more
- The query is not too complex

Took 0 sec. Last updated by anonymous at March 20 2025, 12:24:52 PM. (outdated)

```
%md
### Sorted GDP by country and year
```

FINISHED

Sorted GDP by country and year

Took 0 sec. Last updated by anonymous at March 20 2025, 12:24:52 PM. (outdated)

HIVE JOB FINISHED

```
-- Final query
SELECT countryName as Country, year, indicatorCode as indicator, indicatorValue as GDP_growth
FROM wdi_csv_parquet
WHERE indicatorCode = 'NY.GDP.MKTP.KD.ZG'
ORDER BY countryName DESC, year
```

```
INFO  : Compiling command(queryId=hive_20250320162452_a9c4ee7b-2035-479b-b139-78f5f87a1b68):
SELECT countryName as Country, year, indicatorCode as indicator, indicatorValue as GDP_growth
FROM wdi_csv_parquet
WHERE indicatorCode = 'NY.GDP.MKTP.KD.ZG'
ORDER BY countryName DESC, 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:country, type:string, comment:null), FieldSchema(name:year, type:string, comment:null), FieldSchema(name:indicat
or, type:string, comment:null), FieldSchema(name:gdp_growth, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250320162452_a9c4ee7b-2035-479b-b139-78f5f87a1b68); Time taken: 0.17 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250320162452_a9c4ee7b-2035-479b-b139-78f5f87a1b68):
SELECT countryName as Country, year, indicatorCode as indicator, indicatorValue as GDP_growth
FROM wdi_csv_parquet
WHERE indicatorCode = 'NY.GDP.MKTP.KD.ZG'
ORDER BY countryName DESC, year
INFO  : Query ID = hive_20250320162452_a9c4ee7b-2035-479b-b139-78f5f87a1b68
```

settings

country	year	indicator	gdp_growth
Zimbabwe	1960	NY.GDP.MKTP.KD.ZG	0
Zimbabwe	1961	NY.GDP.MKTP.KD.ZG	6.31615726938115
Zimbabwe	1962	NY.GDP.MKTP.KD.ZG	1.43447088725841
Zimbabwe	1963	NY.GDP.MKTP.KD.ZG	6.24434450666239
Zimbabwe	1964	NY.GDP.MKTP.KD.ZG	-1.1061718588287
Zimbabwe	1965	NY.GDP.MKTP.KD.ZG	4.91057058670184
Zimbabwe	1966	NY.GDP.MKTP.KD.ZG	1.52313002640238
Zimbabwe	1967	NY.GDP.MKTP.KD.ZG	8.36700893022744

Took 31 sec. Last updated by anonymous at March 20 2025, 12:25:23 PM. (outdated)