```
# Check the size of wdi_csv_parquet
hdfs dfs -du -s -h hdfs://user/homaalmasieh/hive/wdi/wdi_opencsv_text
2.0 G  6.0 G  hdfs://user/homaalmasieh/hive/wdi/wdi_opencsv_text
```

Took 3 sec. Last updated by anonymous at November 25 2021, 2:34:43 PM

```
-- Compare runtime
SELECT count(countryname) FROM wdi_csv_parquet

INFO : Compiling command(queryId=hive_20211125193626_396487cc-155b-4178-bda4-2e8320fe0054): 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_20211125193626_396487cc-155b-4178-bda4-2e8320fe0054); Time taken: 0.184 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20211125193626_396487cc-155b-4178-bda4-2e8320fe0054):
SELECT count(countryname) FROM wdi_csv_parquet

INFO : Query ID = hive_20211125193626_396487cc-155b-4178-bda4-2e8320fe0054
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_20211125193626_396487cc-155b-4178-bda4-2e8320fe0054
INFO : Tez session hasn't been created yet. Opening session
```

_c0
21759408

Took 26 sec. Last updated by anonymous at November 25 2021, 2:36:52 PM. (outdated)

```
SELECT count(countryname) FROM wdi_opencsv_text

INFO : Compiling command(queryId=hive_20211125193703_a321dfb0-9d2d-4d25-a5a7-5c0fid39334b): 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_20211125193703_a321dfb0-9d2d-4d25-a5a7-5c0fid39334b); Time taken: 0.15 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20211125193703_a321dfb0-9d2d-4d25-a5a7-5c0fid39334b): SELECT count(countryname) FROM wdi_opencsv_text
INFO : Query ID = hive_20211125193703_a321dfb0-9d2d-4d25-a5a7-5c0fid39334b
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_20211125193703_a321dfb0-9d2d-4d25-a5a7-5c0fid39334b
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_1637855588469_0050)

----------------------------------------------------------------------------------------------
```

_c0
21759408

Took 28 sec. Last updated by anonymous at November 25 2021, 2:37:30 PM. (outdated)

```
-- Execute 2015 GDP Growth HQL against wdi_csv_parquet and wdi_opencsv_text tables
SELECT indicatorvalue, year, countryname
FROM wdi_csv_parquet
WHERE countryname = 'Canada' AND indicatorcode = 'NY.GDP.MKTP.KD.ZG' AND year = '2015'

INFO : Compiling command(queryId=hive_20211125194032_a5dcedcc-c0bf-4bc2-afc7-5ed981cd0899):
SELECT indicatorvalue, year, countryname
FROM wdi_csv_parquet
WHERE countryname = 'Canada' AND indicatorcode = 'NY.GDP.MKTP.KD.ZG' AND year = '2015'
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:indicatorvalue, type:float, comment:null), FieldSchema(name:year, type:int, comment:null), FieldSchema(name:countryname, type:string, comment:nul
ll), properties:null)
INFO : Completed compiling command(queryId=hive_20211125194032_a5dcedcc-c0bf-4bc2-afc7-5ed981cd0899); Time taken: 0.238 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20211125194032_a5dcedcc-c0bf-4bc2-afc7-5ed981cd0899):
SELECT indicatorvalue, year, countryname
FROM wdi_csv_parquet
WHERE countryname = 'Canada' AND indicatorcode = 'NY.GDP.MKTP.KD.ZG' AND year = '2015'
INFO : Query ID = hive_20211125194032_a5dcedcc-c0bf-4bc2-afc7-5ed981cd0899
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
```

indicatorvalue	year	countryname
1.0782688	2015	Canada

Took 29 sec. Last updated by anonymous at November 25 2021, 2:41:01 PM. (outdated)

```
SELECT indicatorvalue, year, countryname
FROM wdi_opencsv_text
WHERE countryname = 'Canada' AND indicatorcode = 'NY.GDP.MKTP.KD.ZG' AND year = '2015'

INFO : Compiling command(queryId=hive_20211125194140_92a0be61-afa6-4767-9ede-4b0047a520af): SELECT indicatorvalue, year, countryname
FROM wdi_opencsv_text
WHERE countryname = 'Canada' AND indicatorcode = 'NY.GDP.MKTP.KD.ZG' AND year = '2015'
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:indicatorvalue, type:float, comment:null), FieldSchema(name:year, type:int, comment:null), FieldSchema(name:countryname, type:string, comment:nul
ll), properties:null)
INFO : Completed compiling command(queryId=hive_20211125194140_92a0be61-afa6-4767-9ede-4b0047a520af); Time taken: 0.173 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20211125194140_92a0be61-afa6-4767-9ede-4b0047a520af): SELECT indicatorvalue, year, countryname
FROM wdi_opencsv_text
WHERE countryname = 'Canada' AND indicatorcode = 'NY.GDP.MKTP.KD.ZG' AND year = '2015'
INFO : Query ID = hive_20211125194140_92a0be61-afa6-4767-9ede-4b0047a520af
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_20211125194140_92a0be61-afa6-4767-9ede-4b0047a520af
```

indicatorvalue	year	countryname
1.0782688	2015	Canada

Took 48 sec. Last updated by anonymous at November 25 2021, 2:42:34 PM. (outdated)

The dataset using wdi_csv_parquet is significiently faster than when using wdi_opencsv_text.

Took 0 sec. Last updated by anonymous at November 25 2021, 2:43:24 PM.

Highest GDP Growth

Took 0 sec. Last updated by anonymous at November 25 2021, 2:44:42 PM.

```
-- Find the highest GDP growth (NY.GDP.MKTP.KD.ZG) year for each country
SELECT wdi.indicatorValue As GDP_growth_value, wdi.year, wdi.countryName
FROM (
    SELECT max(indicatorValue) AS max_ind, countryname
    FROM wdi_csv_parquet
    WHERE indicatorCode = 'NY.GDP.MKTP.KD.ZG' AND indicatorValue !=0
    GROUP BY countryname
) maxValue
INNER JOIN wdi_csv_parquet AS wdi ON maxValue.max_ind = wdi.indicatorValue AND maxValue.countryname = wdi.countryname
ORDER BY GDP_growth_value DESC

INFO : Compiling command(queryId=hive_20211126015624_eb9f0a1b-3db7-4a9e-9bbd-296c68a441e):
SELECT wdi.indicatorValue As GDP_growth_value, wdi.year, wdi.countryName
FROM (
    SELECT max(indicatorValue) AS max_ind, countryname
    FROM wdi_csv_parquet
    WHERE indicatorCode = 'NY.GDP.MKTP.KD.ZG' AND indicatorValue !=0
    GROUP BY countryname
) maxValue
INNER JOIN wdi_csv_parquet AS wdi ON maxValue.max_ind = wdi.indicatorValue AND maxValue.countryname = wdi.countryname
ORDER BY GDP_growth_value DESC

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:wdi.year, type:int, comment:null), FieldSchema(name:wdi.countryname, type:string, com
ment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20211126015624_eb9f0a1b-3db7-4a9e-9bbd-296c68a441e); Time taken: 3.468 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
```

gdp_growth_value	year	countryname
189.82993	1965	Oman
149.97296	1997	Equatorial Guinea
106.27981	1997	Liberia
104.486786	2012	Libya
88.957664	1996	Bosnia and Herzegovina
57.81783	1990	Iraq
45.302753	1974	Kiribati
39.487095	1974	Gabon

Took 1 min 46 sec. Last updated by anonymous at November 25 2021, 6:58:10 PM. (outdated)

```
%spark.sql
-- Use %spark.sql to switch to Spark interpreter

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

GDP_growth_value	year	countryName
189.82993	1965	Oman
149.97296	1997	Equatorial Guinea
106.27981	1997	Liberia
104.486786	2012	Libya
88.957664	1996	Bosnia and Herzegovina
57.81783	1990	Iraq
45.302753	1974	Kiribati
39.487095	1974	Gabon

Took 53 sec. Last updated by anonymous at November 25 2021, 9:03:08 PM. (outdated)

```
%md
### Compare the execution time with the Hive Tez query and spark.sql

The Hive interpreter took 1.5 minutes which is significiently slower than using spark interpreter.
```

Compare the execution time with the Hive Tez query and spark.sql

The Hive interpreter took 1.5 minutes which is significiently slower than using spark interpreter.

Took 3 sec. Last updated by anonymous at November 25 2021, 10:18:06 PM.

Sort GDP by country and year

Took 0 sec. Last updated by anonymous at November 25 2021, 10:20:31 PM.

```
-- Write a query that returns GDP Growth for all countries. Sort by countryname and year.
SELECT countryname, year, indicatorcode, indicatorvalue AS GDP_growth_value
FROM wdi_csv_parquet
WHERE indicatorcode = 'NY.GDP.MKTP.KD.ZG'
DISTRIBUTE BY countryname
SORT BY countryname ASC, year ASC

INFO : Compiling command(queryId=hive_20211126032927_ddaiba28-e21d-4d2b-9da7-95cf52fc14db):
SELECT countryname, year, indicatorcode, indicatorvalue AS GDP_growth_value
FROM wdi_csv_parquet
WHERE indicatorcode = 'NY.GDP.MKTP.KD.ZG'
DISTRIBUTE BY countryname
SORT BY countryname ASC, year ASC

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:nul
ll), FieldSchema(name:gdp_growth_value, type:float, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20211126032927_ddaiba28-e21d-4d2b-9da7-95cf52fc14db); Time taken: 0.199 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20211126032927_ddaiba28-e21d-4d2b-9da7-95cf52fc14db):
SELECT countryname, year, indicatorcode, indicatorvalue AS GDP_growth_value
FROM wdi_csv_parquet
```

countryname	year	indicatorcode	gdp_growth_value
Afghanistan	1960	NY.GDP.MKTP.KD.ZG	0.0
Afghanistan	1961	NY.GDP.MKTP.KD.ZG	0.0
Afghanistan	1962	NY.GDP.MKTP.KD.ZG	0.0
Afghanistan	1963	NY.GDP.MKTP.KD.ZG	0.0
Afghanistan	1964	NY.GDP.MKTP.KD.ZG	0.0
Afghanistan	1965	NY.GDP.MKTP.KD.ZG	0.0
Afghanistan	1966	NY.GDP.MKTP.KD.ZG	0.0
Afghanistan	1967	NY.GDP.MKTP.KD.ZG	0.0

Took 26 sec. Last updated by anonymous at November 25 2021, 10:29:03 PM. (outdated)