Big_data_project

```
%spark2.pyspark
td = spark.read.csv('/user/Bigdata_project/electricity-exports-and-imports-monthly (1).csv',header=True)
```

FINISHED

Took 1 sec. Last updated by anonymous at December 12 2022, 8:07:49 PM.

```
%spark2
td.show(5)
```

FINISHED

```
+----------+--------+----------------+-----------+-------------+------------------+-----------------+
|    Period|Activity|          Source|Destination|Energy (MW.h)|Total Value (CAN$)|Price (CAN$/MW.h)|
+----------+--------+----------------+-----------+-------------+------------------+-----------------+
|01/01/1990| Exports|British Columbia|     Alaska|       75.196|            4536.0|            60.32|
|01/01/1990| Exports|British Columbia| California|     345904.0|     1.114459603E7|            32.22|
|01/01/1990| Exports|British Columbia|     Oregon|     148800.0|         4770498.0|            32.06|
|01/01/1990| Exports|British Columbia|      Total|   496761.596|     1.600059003E7|            32.21|
|01/01/1990| Exports|British Columbia| Washington|       1982.4|           80960.0|            40.84|
+----------+--------+----------------+-----------+-------------+------------------+-----------------+
only showing top 5 rows
```

Took 0 sec. Last updated by anonymous at December 12 2022, 8:07:25 PM.

```
%spark2.pyspark
import pyspark.sql.functions as f
td.select("Activity", f.translate(f.col("Activity"),'"',"").alias("replacedActivity")).show(5)
```

FINISHED

```
+--------+----------------+
|Activity|replacedActivity|
+--------+----------------+
| Exports|         Exports|
| Exports|         Exports|
| Exports|         Exports|
| Exports|         Exports|
| Exports|         Exports|
+--------+----------------+
only showing top 5 rows
```

Took 0 sec. Last updated by anonymous at December 12 2022, 8:40:43 PM.

```
%spark2.pyspark
from pyspark.sql.functions import substring
td.select("Period", substring("Period",7,4).alias("replacedPeriod")).show(5)
```

FINISHED

```
+----------+--------------+
|    Period|replacedPeriod|
+----------+--------------+
|01/01/1990|          1990|
|01/01/1990|          1990|
|01/01/1990|          1990|
|01/01/1990|          1990|
|01/01/1990|          1990|
+----------+--------------+
only showing top 5 rows
```

Took 0 sec. Last updated by anonymous at December 12 2022, 8:35:25 PM.

```
%spark2.sql
select * from project.imp_exp
WHERE region
```

FINISHED

region	source	year	generated_value	unit
Region	Source	null	null	Unit

region	source	year	generated_value	unit
Canada	Hydro	2005	358386	GW.h
Canada	Wind	2005	1453	GW.h
Canada	Biomass	2005	7687	GW.h
Canada	Solar	2005	0	GW.h
Canada	Nuclear	2005	86668	GW.h
Canada	Coal	2005	96750	GW.h

Output is truncated to 1000 rows. Learn more about **zeppelin.spark.maxResult**

Took 0 sec. Last updated by anonymous at December 13 2022, 6:24:37 PM. (outdated)

```
%spark2.sql
create OR Replace View total_capacity
AS
SELECT region,source,year,capacity_value*5.3 as capacity_value, unit
FROM project.total_capacity
WHERE source != 'Source'
```

FINISHED

Took 0 sec. Last updated by anonymous at December 13 2022, 8:33:38 PM.

```
%spark2.sql
select * from total_capacity
```

FINISHED

region	source	year	capacity_value	unit
Canada	Hydro	2005	386311.7	MW
Canada	Wind	2005	2952.1	MW
Canada	Biomass	2005	9561.2	MW
Canada	Solar	2005	84.8	MW
Canada	Nuclear	2005	67866.5	MW
Canada	Coal	2005	84810.6	MW
Canada	Natural Gas	2005	69907.0	MW
Canada	Oil and Diesel	2005	25408.2	MW
Canada	Hydro	2006	385728.7	MW

Output is truncated to 1000 rows. Learn more about **zeppelin.spark.maxResult**

Took 1 sec. Last updated by anonymous at December 13 2022, 8:33:48 PM.

```
%spark2.sql
create or replace view capacity_Refined_view
AS
select year, region ,sum(capacity_value) as total_capacity, unit
from total_capacity
where region=='ON'
group by year, region, unit
order by year
```

FINISHED

Took 1 sec. Last updated by anonymous at December 13 2022, 8:36:38 PM.

```
%spark2.sql
select * from capacity_Refined_View
```

FINISHED

year	region	total_capacity	unit
2005	ON	167140.8	MW
2006	ON	171359.6	MW
2007	ON	171836.6	MW
2008	ON	179542.8	MW
2009	ON	180470.3	MW
2010	ON	191229.3	MW
2011	ON	193995.9	MW
2012	ON	192527.8	MW
2013	ON	200663.3	MW

Took 2 sec. Last updated by anonymous at December 13 2022, 8:37:24 PM.

```
%spark2.sql
create OR Replace View generation
AS
SELECT * FROM project.generation
WHERE region != 'Region'
```

FINISHED

Took 0 sec. Last updated by anonymous at December 13 2022, 6:32:54 PM.

```
%spark2.sql
select * from generation
```

FINISHED

region	source	year	generated_value	unit
Canada	Hydro	2005	358386	GW.h
Canada	Wind	2005	1453	GW.h
Canada	Biomass	2005	7687	GW.h
Canada	Solar	2005	0	GW.h
Canada	Nuclear	2005	86668	GW.h
Canada	Coal	2005	96750	GW.h
Canada	Natural Gas	2005	40874	GW.h
Canada	Oil and Diesel	2005	10608	GW.h
Canada	Hydro	2006	349481	GW.h

Output is truncated to 1000 rows. Learn more about **zeppelin.spark.maxResult**

Took 0 sec. Last updated by anonymous at December 14 2022, 9:21:40 AM.

```
%spark2.sql
create OR Replace View imp_exp
AS
SELECT time_period,activity,source,destination,activity_value*0.001 as activity_value,total_value,price
FROM project.imp_exp
WHERE source != 'Source'
```

FINISHED

Took 1 sec. Last updated by anonymous at December 13 2022, 8:42:26 PM.

```
%spark2.sql
select * from imp_exp
```

FINISHED

time_period	activity	source	destination	activity_value	total_value	price
1990	Exports	British Columbia	Alaska	0.075	4536	60
1990	Exports	British Columbia	California	345.904	11144596	32
1990	Exports	British Columbia	Oregon	148.800	4770498	32
1990	Exports	British Columbia	Total	496.761	16000590	32
1990	Exports	British Columbia	Washington	1.982	80960	40
1990	Exports	Manitoba	Minnesota	5.100	129051	25
1990	Exports	Manitoba	North Dakota	0.085	3922	46
1990	Exports	Manitoba	Total	5.185	132974	25
1990	Exports	New Brunswick	Maine	181.005	9087736	49

Output is truncated to 1000 rows. Learn more about **zeppelin.spark.maxResult**

Took 0 sec. Last updated by anonymous at December 14 2022, 12:09:00 PM.

```
%spark2.sql
create OR Replace View imp_exp_Refined_view
AS
select time_period, activity,source, sum(activity_value) as total_activity_value
from imp_exp
where (activity=='Exports' and source == 'Ontario')
group by time_period, activity, source
order by time_period
```

FINISHED

Took 1 sec. Last updated by anonymous at December 13 2022, 8:43:44 PM.

```
%spark2.sql
SELECT * FROM imp_exp_Refined_view
```

FINISHED

time_period	activity	source	total_activity_value
1990	Exports	Ontario	1476.212
1991	Exports	Ontario	4628.416
1992	Exports	Ontario	4111.909
1993	Exports	Ontario	9928.159
1994	Exports	Ontario	25326.924
1995	Exports	Ontario	18390.268
1996	Exports	Ontario	11060.911
1997	Exports	Ontario	11801.212
1998	Exports	Ontario	5206.489

Took 2 sec. Last updated by anonymous at December 13 2022, 8:44:06 PM.

READY

```
%spark2.sql
create or replace view generation_Refined_view
AS
select year, region ,sum(generated_value) as total_energy_generated, unit
from generation
where region=='ON'
group by year, region, unit
order by year
```

FINISHED

Took 1 sec. Last updated by anonymous at December 13 2022, 8:19:28 PM.

```
%spark2.sql
select * from generation_Refined_view
```

FINISHED

year	region	total_energy_generated	unit
2005	ON	157278	GW.h
2006	ON	157287	GW.h
2007	ON	157462	GW.h
2008	ON	162641	GW.h
2009	ON	146786	GW.h
2010	ON	149727	GW.h
2011	ON	154116	GW.h
2012	ON	153907	GW.h
2013	ON	161048	GW.h

Took 2 sec. Last updated by anonymous at December 13 2022, 8:20:20 PM.

FINISHED

```
%spark2.sql
select year, region, sum(capacity_value) as total_energy_generated, unit
from total_capacity
where region=='ON'
group by year, region, unit
order by year
```

year	region	total_energy_generated	unit
2005	ON	31536	MW
2006	ON	32332	MW
2007	ON	32422	MW
2008	ON	33876	MW
2009	ON	34051	MW
2010	ON	36081	MW
2011	ON	36603	MW
2012	ON	36326	MW
2013	ON	37861	MW

Took 1 sec. Last updated by anonymous at December 13 2022, 7:11:35 PM.

FINISHED

```
%spark2.sql
create or replace view imp_exp_Refined_view
AS
select time_period, activity,source, (sum(activity_value) as Total_value
from imp_exp
where (activity=='Exports' and source == 'Ontario')
group by time_period, activity, source
order by time_period
```

Took 1 sec. Last updated by anonymous at December 13 2022, 7:50:52 PM. (outdated)

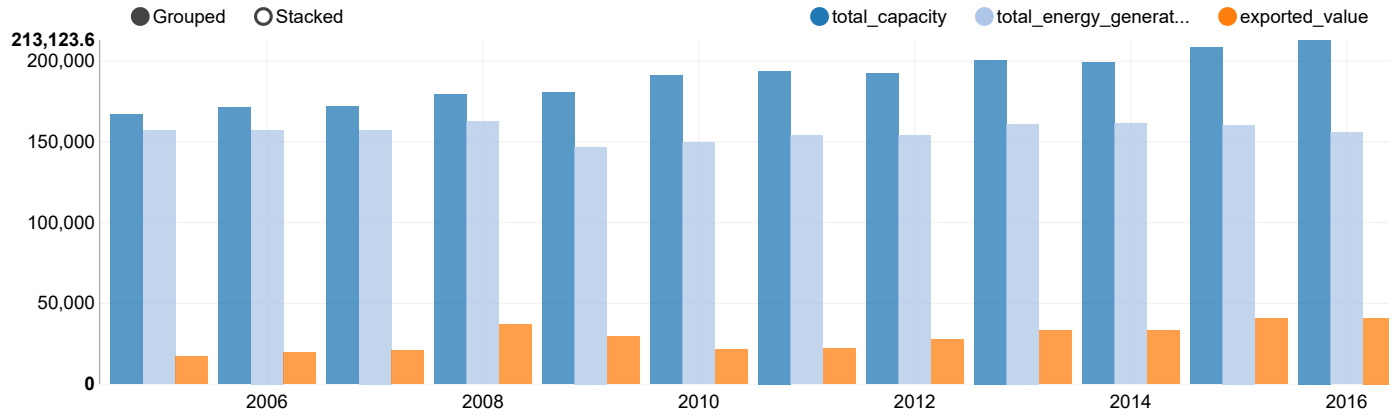
```
%md
For Ontario
```

READY

```
%spark2.sql
SELECT c.year,i.total_activity_value as exported_value,g.total_energy_generated,c.total_capacity
FROM capacity_Refined_view c
JOIN imp_exp_Refined_view i
ON c.year=i.time_period
JOIN generation_Refined_view g
ON c.year=g.year
```

FINISHED

settings

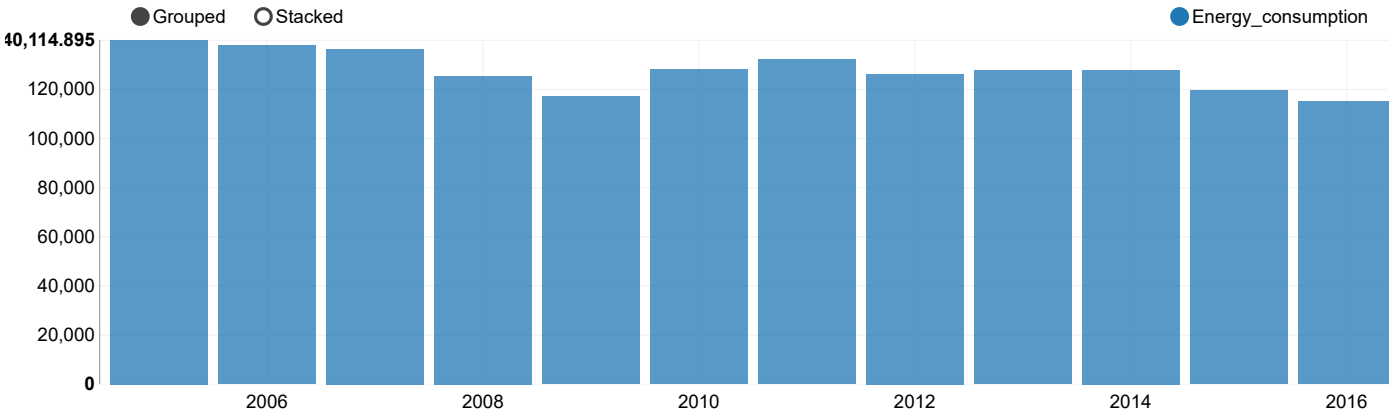


Took 7 sec. Last updated by anonymous at December 13 2022, 10:30:56 PM. (outdated)

```
%spark2.sql
SELECT c.year,(g.total_energy_generated-i.total_activity_value) as Energy_consumption
FROM capacity_Refined_view c
JOIN imp_exp_Refined_view i
ON c.year=i.time_period
JOIN generation_Refined_view g
ON c.year=g.year
```

FINISHED

settings

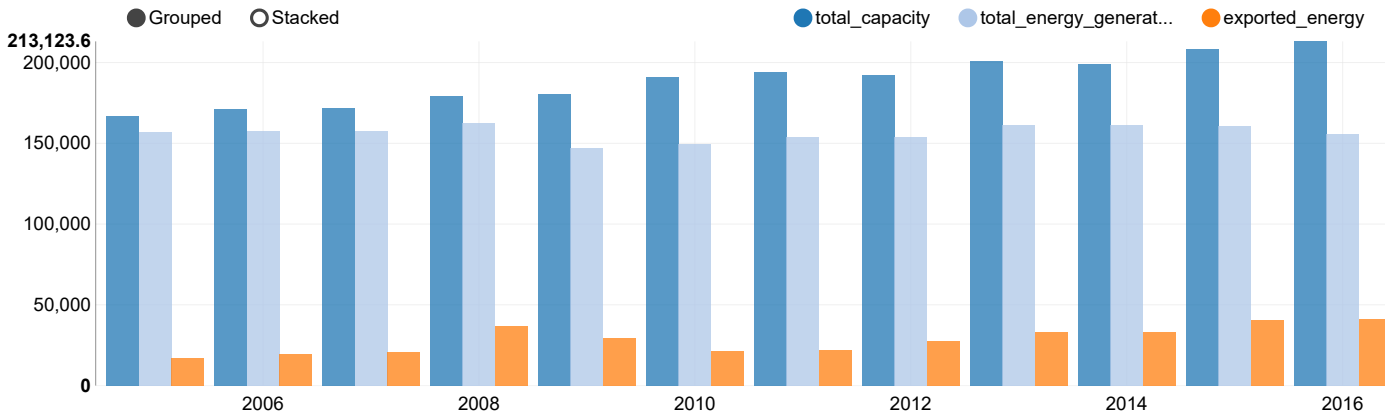


Took 7 sec. Last updated by anonymous at December 13 2022, 10:47:29 PM. (outdated)

```
%spark2.sql
SELECT c.year,c.total_capacity,g.total_energy_generated,i.total_activity_value as exported_energy,(g.total_energy_generated
    -total_activity_value) as Relative_Energy_consumption
FROM capacity_Refined_view c
JOIN imp_exp_Refined_view i
ON c.year=i.time_period
JOIN generation_Refined_view g
ON c.year=g.year
```

FINISHED

settings



Took 6 sec. Last updated by anonymous at December 13 2022, 11:18:04 PM. (outdated)

```
%spark2.sql
```

READY