Assignment Final Report

Student Name: Aruna Bellgutte Ramesh

Cloud Web App: https://energy-statistics-application.herokuapp.com/site.html

GitHub Link for code:
https://github.com/arunabellgutteramesh/EnergyConsumptionBigDataAnalysisWithBigquery

Video Walkthrough: https://youtu.be/plPzZUzcWUk

The Data

The dataset chosen to analyse was based on Energy Statistics on Kaggle from the following URL. The dataset includes over 1 million rows of data, inclusive of 7 different columns. These columns include: **Countries**, **Quantity** of Energy Consumed, **Category**, and **Type** of Energy to name a few.

Data Processing

To clean the data we used *Apache PIG* through *Google Data Proc*. The queries to clean the data are as follows:

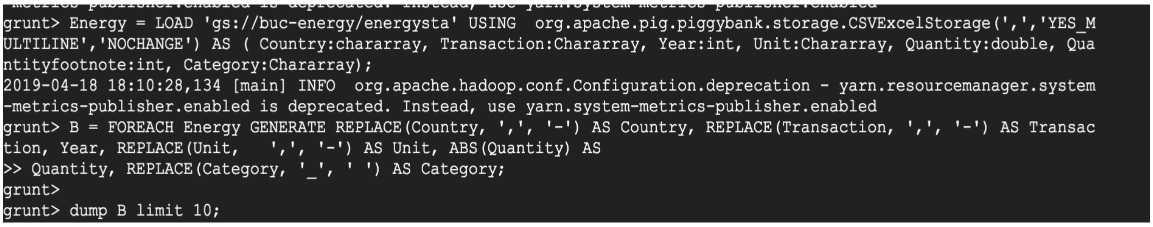
- Energy = LOAD 'gs://buc-uthera/Cloud2' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE','NOCHANGE','SKIP_INPUT_HEADER') AS (Country:chararray, Transaction:Chararray, Year:int, Unit:Chararray, Quantity:double, Quantityfootnote:int, Category:Chararray);
- B = FOREACH Energy GENERATE Country, REPLACE(Transaction, ',', '-') AS Transaction, Year, REPLACE(Unit, ',', '-') AS Unit, ABS(Quantity) AS Quantity, REPLACE(Category,'_', ' ') AS Category;

We stored the data in a bucket on Google Cloud. To process and query the data we used Google BigQuery. Here we were able to store the data in a JSON object and query it, outputting the result on our front-end. To store the data in the bucket, we used the following line:

- STORE B INTO 'gs://buc-energy/E1' USING PigStorage(',');

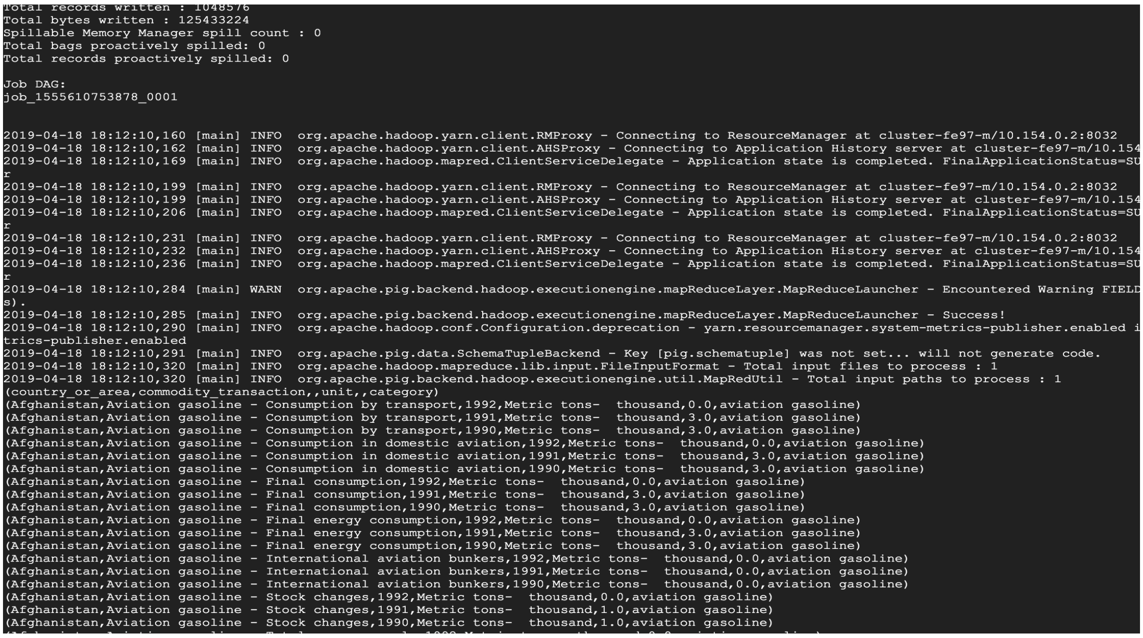
Screenshots for the *Apache PIG* Scripts execution through *Google Data Proc:*

- Queries Execution



```
-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
grunt> Energy = LOAD 'gs://buc-energy/energysta' USING  org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_M
ULTILINE','NOCHANGE') AS ( Country:chararray, Transaction:Chararray, Year:int, Unit:Chararray, Quantity:double, Qua
ntityfootnote:int, Category:Chararray);
2019-04-18 18:10:28,134 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system
-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
grunt> B = FOREACH Energy GENERATE REPLACE(Country, ',', '-') AS Country, REPLACE(Transaction, ',', '-') AS Transac
tion, Year, REPLACE(Unit,   ',', '-') AS Unit, ABS(Quantity) AS
>> Quantity, REPLACE(Category, '_', ' ') AS Category;
grunt>
grunt> dump B limit 10;
```

- Results

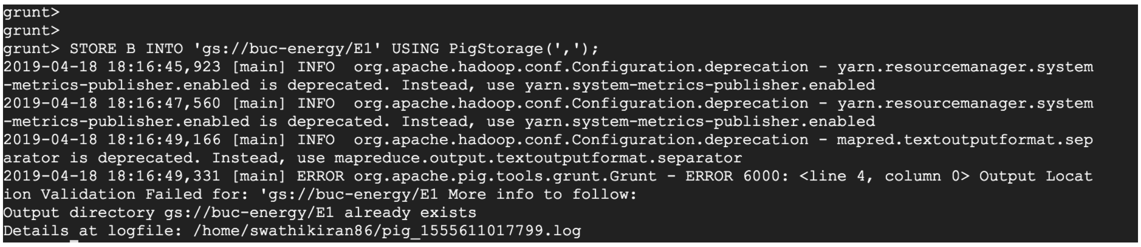


```
Total records written : 1048576
Total bytes written : 125433224
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1555610753878_0001


2019-04-18 18:12:10,160 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster-fe97-m/10.154.0.2:8032
2019-04-18 18:12:10,162 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster-fe97-m/10.154
2019-04-18 18:12:10,169 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SU
r
2019-04-18 18:12:10,199 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster-fe97-m/10.154.0.2:8032
2019-04-18 18:12:10,199 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster-fe97-m/10.154
2019-04-18 18:12:10,206 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SU
r
2019-04-18 18:12:10,231 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster-fe97-m/10.154.0.2:8032
2019-04-18 18:12:10,232 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster-fe97-m/10.154
2019-04-18 18:12:10,236 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SU
r
2019-04-18 18:12:10,284 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD
s).
2019-04-18 18:12:10,285 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2019-04-18 18:12:10,290 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled i
trics-publisher.enabled
2019-04-18 18:12:10,291 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2019-04-18 18:12:10,320 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2019-04-18 18:12:10,320 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(country_or_area,commodity_transaction,,unit,,category)
(Afghanistan,Aviation gasoline - Consumption by transport,1992,Metric tons-  thousand,0.0,aviation gasoline)
(Afghanistan,Aviation gasoline - Consumption by transport,1991,Metric tons-  thousand,3.0,aviation gasoline)
(Afghanistan,Aviation gasoline - Consumption by transport,1990,Metric tons-  thousand,3.0,aviation gasoline)
(Afghanistan,Aviation gasoline - Consumption in domestic aviation,1992,Metric tons-  thousand,0.0,aviation gasoline)
(Afghanistan,Aviation gasoline - Consumption in domestic aviation,1991,Metric tons-  thousand,3.0,aviation gasoline)
(Afghanistan,Aviation gasoline - Consumption in domestic aviation,1990,Metric tons-  thousand,3.0,aviation gasoline)
(Afghanistan,Aviation gasoline - Final consumption,1992,Metric tons-  thousand,0.0,aviation gasoline)
(Afghanistan,Aviation gasoline - Final consumption,1991,Metric tons-  thousand,3.0,aviation gasoline)
(Afghanistan,Aviation gasoline - Final consumption,1990,Metric tons-  thousand,3.0,aviation gasoline)
(Afghanistan,Aviation gasoline - Final energy consumption,1992,Metric tons-  thousand,0.0,aviation gasoline)
(Afghanistan,Aviation gasoline - Final energy consumption,1991,Metric tons-  thousand,3.0,aviation gasoline)
(Afghanistan,Aviation gasoline - Final energy consumption,1990,Metric tons-  thousand,3.0,aviation gasoline)
(Afghanistan,Aviation gasoline - International aviation bunkers,1992,Metric tons-  thousand,0.0,aviation gasoline)
(Afghanistan,Aviation gasoline - International aviation bunkers,1991,Metric tons-  thousand,0.0,aviation gasoline)
(Afghanistan,Aviation gasoline - International aviation bunkers,1990,Metric tons-  thousand,0.0,aviation gasoline)
(Afghanistan,Aviation gasoline - Stock changes,1992,Metric tons-  thousand,0.0,aviation gasoline)
(Afghanistan,Aviation gasoline - Stock changes,1991,Metric tons-  thousand,1.0,aviation gasoline)
(Afghanistan,Aviation gasoline - Stock changes,1990,Metric tons-  thousand,1.0,aviation gasoline)
```

- Loading the processed data



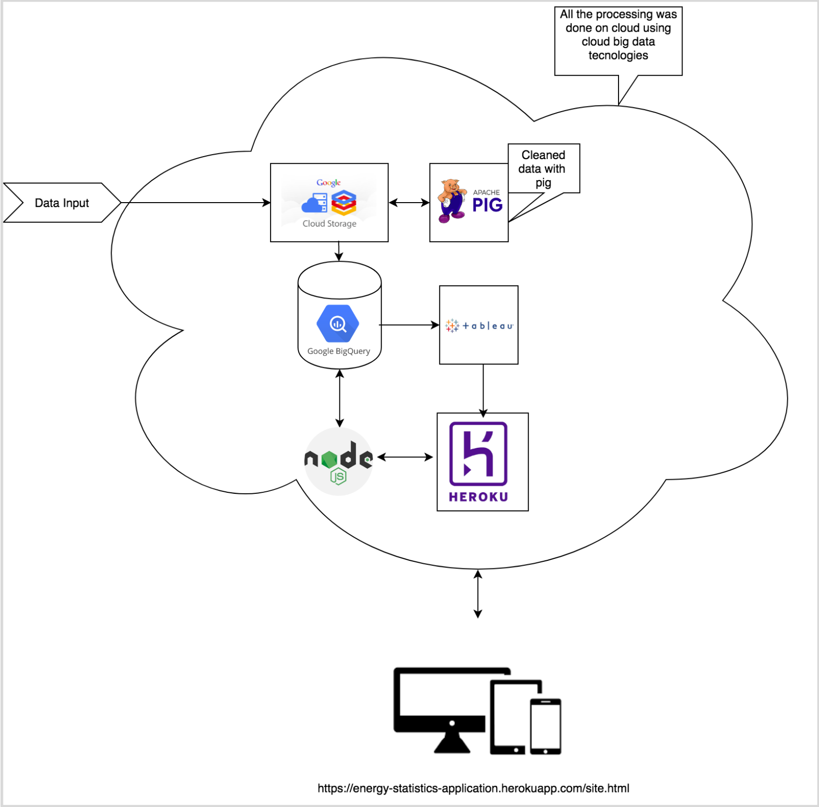
```
grunt>
grunt>
grunt> STORE B INTO 'gs://buc-energy/E1' USING PigStorage(',');
2019-04-18 18:16:45,923 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system
-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
2019-04-18 18:16:47,560 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system
-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
2019-04-18 18:16:49,166 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.sep
arator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2019-04-18 18:16:49,331 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 6000: <line 4, column 0> Output Locat
ion Validation Failed for: 'gs://buc-energy/E1 More info to follow:
Output directory gs://buc-energy/E1 already exists
Details at logfile: /home/swathikiran86/pig_1555611017799.log
```

Front End

We used HTML, CSS, AngularJS and Bootstrap to build the front end. To visualise the data we used Tableau and Bootstrap Interactive widgets. We used Tableau to graph the countries and allow the user to filter by year, quantity, type of consumption etc. This allows the user to interact with a host of filters and functions through the tableau features. We also used the Bootstrap Widget to show real time integration with the cloud technology server, the drop down menus allow the user to choose a category and a year, the widgets will return the top 4 results from the category and year chosen.

System Architecture

To the right, a data flow diagram to represent how our application works.


All the processing was done on cloud using cloud big data tecnologies
Data Input
Google Cloud Storage
Cleaned data with pig
APACHE PIG
Google BigQuery
+ableau
node JS
HEROKU
https://energy-statistics-application.herokuapp.com/site.html

The Web App

We chose to complete almost all of the tasks on the cloud, we cleaned our data on the cloud, we processed and query the data on the cloud and we hosted our final application on the cloud. This application performs a cloud to cloud interaction whereby the query at the top of the page returns the data from our Google BigTable using BigQuery. We wanted to incorporate a cloud to cloud interaction rather than just deploying a static webapp.

The technologies used throughout our project consists of Apache PIG, Google Cloud, Data Proc, BigTable & BigQuery, Tableau, HTML, CSS, Angular, Node and Bootstrap.

Challenges & Lessons Learned

We learnt a lot from this assignment. We found that using hive to query and call real time information from the bucket proved more difficult than expected. We learnt to adapt where required in this case and integrate google BigQuery instead of using Hive to perform this real time interaction. We integrated this cloud to cloud interaction instead of statistical predictions, which had proven to be too difficult to incorporate on the site and we also felt it didn't relate to cloud technologies as we were using machine learning technologies instead.

Related Work

There are a variety of mapping tools currently out there. We were unable to find something to directly match our project. We found a mapping of New York City which shows the energy consumption of each building. See the URL for more details and visualisation of this map. Also, ESRI have an application called ArcGIS which show a multitude of industry data mapped with potential for future developments in some locations. Here is a link to how they can map electric and gas utilities. The ESRI also map and predict different industries based on yearly statistics building coherent spatial maps with objects, statistics and future predictions.

References

➢ https://www.kaggle.com/unitednations/international-energy-statistics

➢ https://pig.apache.org/docs/latest/basic.html

➢ https://cloud.google.com/bigquery/?utm_source=google&utm_medium=cpc&utm_campaign=emea-emea-all-en-dr-bkws-all-all-trial-e-gcp-1003963&utm_content=text-ad-lpquickdataemeactr-any-DEV_c-CRE_167357408135-ADGP_Hybrid+%7C+AW+SEM+%7C+BKWS+~+EXA_1:1_EMEA_EN_Data_BigQuery_TOP_google+bigquery-KWID_43700016288510698-kwd-63326440124-userloc_1007850&utm_term=KW_google%20bigquery-ST_google+bigquery&ds_rl=1242853&ds_rl=1245734&ds_rl=1245734&gclid=EAIaIQobChMIz7yH_4Dc4QIV4bDtCh0hrQsvEAAYASAAEgK3NvD_BwE

➢ https://cloud.google.com/bigquery/docs/quickstarts/quickstart-web-ui

➢ https://www.measurementlab.net/data/docs/bq/examples/

➢ https://devcenter.heroku.com/start

➢ https://devcenter.heroku.com/categories/nodejs-support