Slaughter-House Analysis

An ETL Project

By Rowena Sagaria

Table of Contents

- Dataset used
- Extraction
- Transformation
- Load
- Project Inference

Dataset Used

- The dataset used was taken from open.Canada.ca , downloaded as a csv file.

https://open.canada.ca/data/en/dataset/3c981dfe-30ac-44cb-b9a3-0fb450913d1b

- It was chosen from federally inspected slaughterhouses to analyze the number of heads for cattle, calves, hogs, sheep, and lambs.

- The dataset has 4 columns and 14868 rows. The fields include date, livestock type, livestock category and headcount aggregated weekly from 1997 to 2024.

- The dataset was imported into MySQL database using the terminal to connect to Timberlea via SSH.

Data Extraction

- The database connection was established using the command :

  mysql -u$DBUSER -p$DBPWD -hdb.cs.dal.ca $DBNAME

- A table called ”slaughterdata” was created using the command:

CREATE TABLE slaughterhouse (

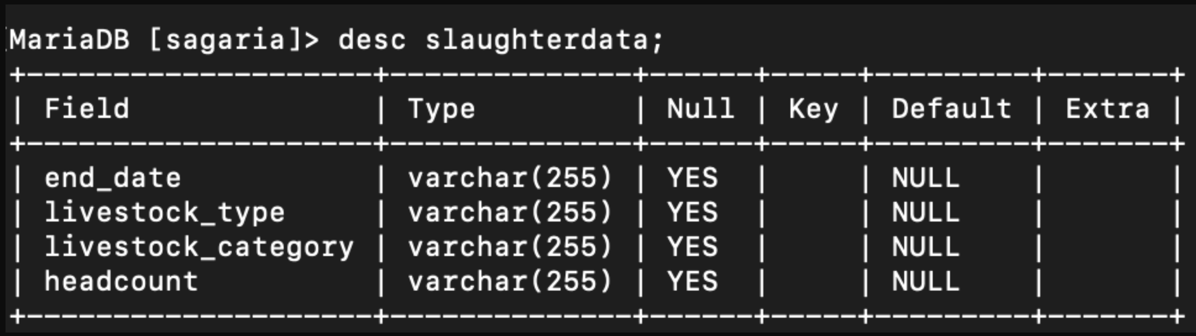
  end_date VARCHAR(225) PRIMARY KEY,

  livestock_type VARCHAR(225),

  livestock_category VARCHAR(225),

  headcount VARCHAR(225));

- The below query was executed:
  “load data infile 'ETL_Prj.csv' into TABLE slaughterdata FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '/n' IGNORE 1 ROWS; “



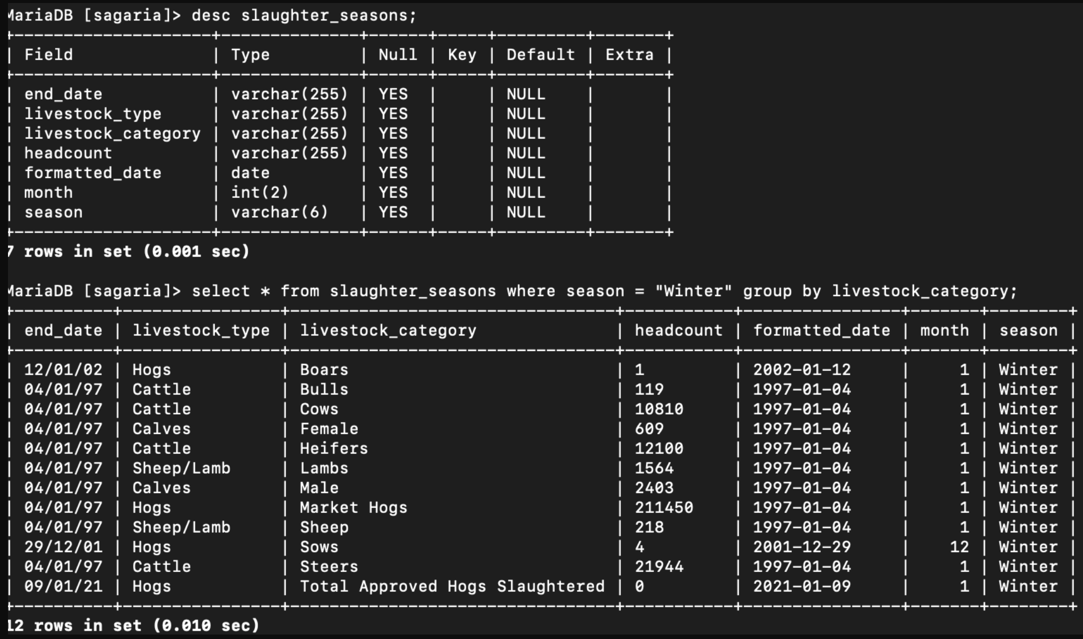
```
MariaDB [sagaria]> desc slaughterdata;
+--------------------+--------------+------+-----+---------+-------+
| Field              | Type         | Null | Key | Default | Extra |
+--------------------+--------------+------+-----+---------+-------+
| end_date           | varchar(255) | YES  |     | NULL    |       |
| livestock_type     | varchar(255) | YES  |     | NULL    |       |
| livestock_category | varchar(255) | YES  |     | NULL    |       |
| headcount          | varchar(255) | YES  |     | NULL    |       |
+--------------------+--------------+------+-----+---------+-------+
```

Data Transformation

- A new table was created to analyse the data based on seasons which was done by extracting the month from the end_date column and mapping the months to a specific season.
- This is done using a case statement as shown below:

```
CASE
WHEN MONTH(STR_TO_DATE(end_date, '%d/%m/%y')) IN (12, 1, 2) THEN 'Winter’
WHEN MONTH(STR_TO_DATE(end_date, '%d/%m/%y')) IN (3, 4, 5) THEN 'Spring'
WHEN MONTH(STR_TO_DATE(end_date, '%d/%m/%y')) IN (6, 7, 8) THEN    'Summer'
WHEN MONTH(STR_TO_DATE(end_date, '%d/%m/%y')) IN (9, 10, 11) THEN 'Fall'
ELSE NULL
END AS season
FROM slaughterdata;
```

- Now to retrive only the data for the number of slaughters in winter, use : select * from slaughter_seasons where season = "Winter" group by livestock_category;



```
MariaDB [sagaria]> desc slaughter_seasons;
+--------------------+--------------+------+-----+---------+-------+
| Field              | Type         | Null | Key | Default | Extra |
+--------------------+--------------+------+-----+---------+-------+
| end_date           | varchar(255) | YES  |     | NULL    |       |
| livestock_type     | varchar(255) | YES  |     | NULL    |       |
| livestock_category | varchar(255) | YES  |     | NULL    |       |
| headcount          | varchar(255) | YES  |     | NULL    |       |
| formatted_date     | date         | YES  |     | NULL    |       |
| month              | int(2)       | YES  |     | NULL    |       |
| season             | varchar(6)   | YES  |     | NULL    |       |
+--------------------+--------------+------+-----+---------+-------+
7 rows in set (0.001 sec)

MariaDB [sagaria]> select * from slaughter_seasons where season = "Winter" group by livestock_category;
+----------+----------------+---------------------------------+-----------+----------------+-------+--------+
| end_date | livestock_type | livestock_category              | headcount | formatted_date | month | season |
+----------+----------------+---------------------------------+-----------+----------------+-------+--------+
| 12/01/02 | Hogs           | Boars                           | 1         | 2002-01-12     |     1 | Winter |
| 04/01/97 | Cattle         | Bulls                           | 119       | 1997-01-04     |     1 | Winter |
| 04/01/97 | Cattle         | Cows                            | 10810     | 1997-01-04     |     1 | Winter |
| 04/01/97 | Calves         | Female                          | 609       | 1997-01-04     |     1 | Winter |
| 04/01/97 | Cattle         | Heifers                         | 12100     | 1997-01-04     |     1 | Winter |
| 04/01/97 | Sheep/Lamb     | Lambs                           | 1564      | 1997-01-04     |     1 | Winter |
| 04/01/97 | Calves         | Male                            | 2403      | 1997-01-04     |     1 | Winter |
| 04/01/97 | Hogs           | Market Hogs                     | 211450    | 1997-01-04     |     1 | Winter |
| 04/01/97 | Sheep/Lamb     | Sheep                           | 218       | 1997-01-04     |     1 | Winter |
| 29/12/01 | Hogs           | Sows                            | 4         | 2001-12-29     |    12 | Winter |
| 04/01/97 | Cattle         | Steers                          | 21944     | 1997-01-04     |     1 | Winter |
| 09/01/21 | Hogs           | Total Approved Hogs Slaughtered | 0         | 2021-01-09     |     1 | Winter |
+----------+----------------+---------------------------------+-----------+----------------+-------+--------+
12 rows in set (0.010 sec)
```

Data Load

1. This step is the final step where only the required cleaned and transformed data is extracted for further analysis via Data Visualization.

2. To extract the cleansed data into a csv file using the terminal:

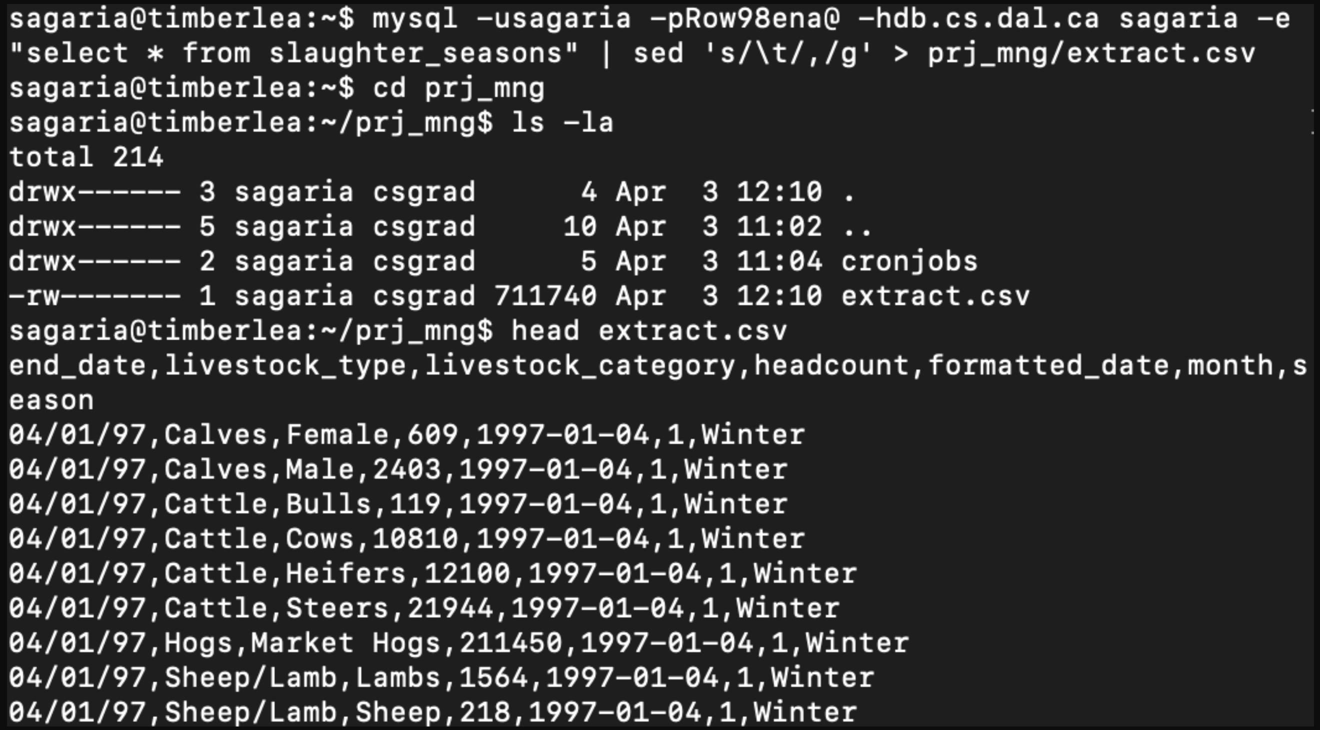
mysql -usagaria –p$DBPWD -hdb.cs.dal.ca sagaria -e "select livestock_type, livestock_category, headcount, season from slaughter_seasons" | sed 's/\t/,/g' > extract.csv

3. Display the extracted file in linux shell :

i. cd prj_mng

ii. ls -la extract.csv

iii. head extract.csv or cat extract.csv



```
sagaria@timberlea:~$ mysql -usagaria -pRow98ena@ -hdb.cs.dal.ca sagaria -e
"select * from slaughter_seasons" | sed 's/\t/,/g' > prj_mng/extract.csv
sagaria@timberlea:~$ cd prj_mng
sagaria@timberlea:~/prj_mng$ ls -la
total 214
drwx------ 3 sagaria csgrad      4 Apr  3 12:10 .
drwx------ 5 sagaria csgrad     10 Apr  3 11:02 ..
drwx------ 2 sagaria csgrad      5 Apr  3 11:04 cronjobs
-rw------- 1 sagaria csgrad 711740 Apr  3 12:10 extract.csv
sagaria@timberlea:~/prj_mng$ head extract.csv
end_date,livestock_type,livestock_category,headcount,formatted_date,month,s
eason
04/01/97,Calves,Female,609,1997-01-04,1,Winter
04/01/97,Calves,Male,2403,1997-01-04,1,Winter
04/01/97,Cattle,Bulls,119,1997-01-04,1,Winter
04/01/97,Cattle,Cows,10810,1997-01-04,1,Winter
04/01/97,Cattle,Heifers,12100,1997-01-04,1,Winter
04/01/97,Cattle,Steers,21944,1997-01-04,1,Winter
04/01/97,Hogs,Market Hogs,211450,1997-01-04,1,Winter
04/01/97,Sheep/Lamb,Lambs,1564,1997-01-04,1,Winter
04/01/97,Sheep/Lamb,Sheep,218,1997-01-04,1,Winter
```

Job Automation - Cron Jobs

1. ssh to timberlea

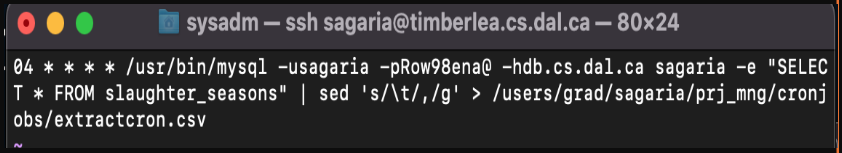
2. Cron job file path: /users/grad/sagaria/prj_mng/cronjobs/cron2

3. Contents of cron2: where 38 is the minute of execution:

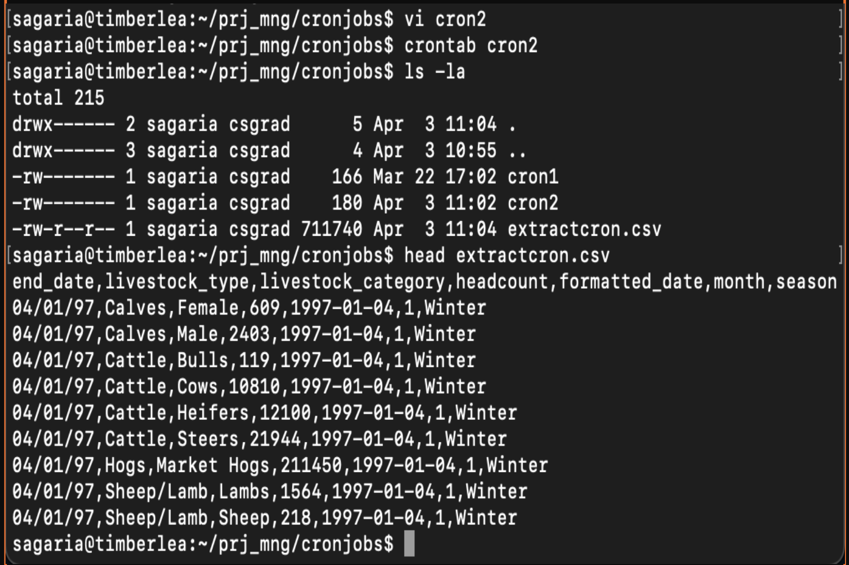
```
38 * * * * /usr/bin/mysql -u$DBUSER -p$DBPWD-hdb.cs.dal.ca
$DBNAME -e "SELECT * FROM slaughter_seasons" | sed 's/\t/,/g' >
/users/grad/sagaria/cronjobs/extractcron.csv
```

4. Running the Cron file:
In the command line type → crontab cron2

5. Check in the path given in the command if the job is complete by cross-verifying the time displayed to the time given in the cron job file.



```
04 * * * * /usr/bin/mysql -usagaria -pRow98ena@ -hdb.cs.dal.ca sagaria -e "SELEC
T * FROM slaughter_seasons" | sed 's/\t/,/g' > /users/grad/sagaria/prj_mng/cronj
obs/extractcron.csv
~
```



```
sagaria@timberlea:~/prj_mng/cronjobs$ vi cron2
sagaria@timberlea:~/prj_mng/cronjobs$ crontab cron2
sagaria@timberlea:~/prj_mng/cronjobs$ ls -la
total 215
drwx------ 2 sagaria csgrad      5 Apr  3 11:04 .
drwx------ 3 sagaria csgrad      4 Apr  3 10:55 ..
-rw------- 1 sagaria csgrad    166 Mar 22 17:02 cron1
-rw------- 1 sagaria csgrad    180 Apr  3 11:02 cron2
-rw-r--r-- 1 sagaria csgrad 711740 Apr  3 11:04 extractcron.csv
sagaria@timberlea:~/prj_mng/cronjobs$ head extractcron.csv
end_date,livestock_type,livestock_category,headcount,formatted_date,month,season
04/01/97,Calves,Female,609,1997-01-04,1,Winter
04/01/97,Calves,Male,2403,1997-01-04,1,Winter
04/01/97,Cattle,Bulls,119,1997-01-04,1,Winter
04/01/97,Cattle,Cows,10810,1997-01-04,1,Winter
04/01/97,Cattle,Heifers,12100,1997-01-04,1,Winter
04/01/97,Cattle,Steers,21944,1997-01-04,1,Winter
04/01/97,Hogs,Market Hogs,211450,1997-01-04,1,Winter
04/01/97,Sheep/Lamb,Lambs,1564,1997-01-04,1,Winter
04/01/97,Sheep/Lamb,Sheep,218,1997-01-04,1,Winter
sagaria@timberlea:~/prj_mng/cronjobs$
```

Project Inference

1. Gain insights into source data structure, quality, and dependencies for effective integration.
2. Address inconsistencies and inaccuracies; enhance data quality through cleansing and enrichment.
3. Align transformation rules with business objectives; develop a deep understanding of business processes.
4. Improve proficiency in SQL, data processing techniques, and optimization strategies.
5. Embrace agile practices for iterative refinement; incorporate stakeholder feedback for improvements.
6. Overcome challenges associated with disparate data sources; develop expertise in integration technologies.
7. Document processes and metadata for reproducibility; implement robust metadata management practices.
8. Establish monitoring mechanisms for ongoing enhancement; define metrics for data quality and alignment with objectives.

Thank You !!