Analysing Real and Fraudulent Job Postings with Hive and Pig

Name: Edward Bolger
Student Number: 20364133
[Git Repo]

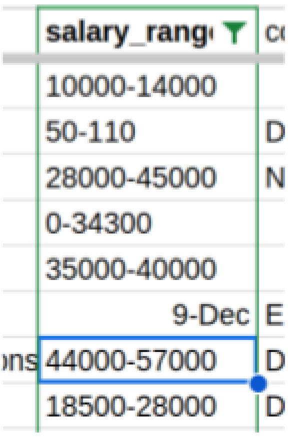
1. Introduction

Apache Hadoop is a framework tailored for processing vast amounts of data across clusters of computers. It employs a MapReduce model to divide workloads across these clusters, enabling parallel processing of data. There are many software projects built on top of Hadoop, including Hive and Pig. Hive is a data warehousing project which allows for SQL-like queries to be run on datasets stored in Hadoop's distributed file system (HDFS). Meanwhile, Pig is a high-level scripting language with extensive support for transforming and cleaning data. This report outlines the process of using Pig and Hive to analyse a dataset describing real and fraudulent job postings. The job posting dataset can be found here, while the code used is contained in this repository.

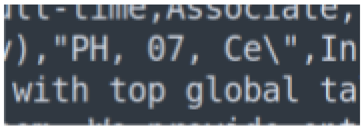
2. Cleaning the data with Pig

Cleaning the job postings dataset was more difficult than expected due to the presence of escape characters and incorrect data types within some of the fields. Here is the process that the cleaning_job_postings.pig script goes through to improve the quality of the dataset.

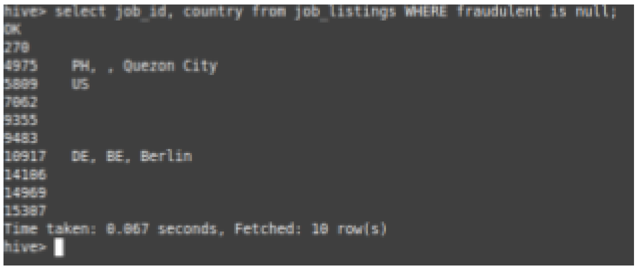
1. The job postings are loaded with CSVExcelStorage, as it ignores quoted commas letting it correctly parse the fields in the CSV. The first row of the CSV is also skipped as it contains the header names.
2. When loading the data into hive the some rows were incorrectly parsed due to the presence of \ in entries of the location field which was acting as an escape character on the quotes causing data loss. To resolve this occurrences of \ in the location are replaced with / using the REPLACE() function.
3. The STRSPLIT function is used to extract the Country from the location field.
4. For some reason the salary range field incorrectly contains some dates. To fix this, and to extract the salary figures, first the data is filtered based on salary columns that match the format: 20000-40000. STRSPLIT is then used to extract the lower and upper bounds of the salaries. Some of the salaries are measured in thousands e.g 20-40 so they are multiplied by 1000 to match the others. Then the midpoint of the salary ranges is calculated.
5. As the filter created a separate table, these salary figures are joined back with the full job listings dataset. This renames the fields based on their origin table e.g. extract_country::job_id.
6. The fields are renamed removing the table of origin, and unnecessary fields are dropped.
7. The data is then saved as a CSV with headers using CSVExcelStorage().



An example of date in the salary_range field



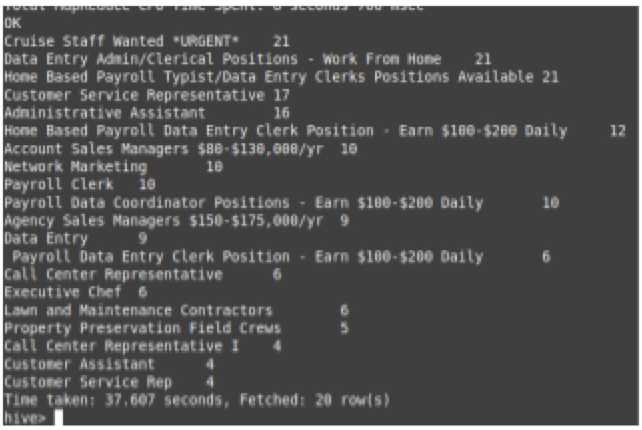
An example of a \ escaping " in the location field



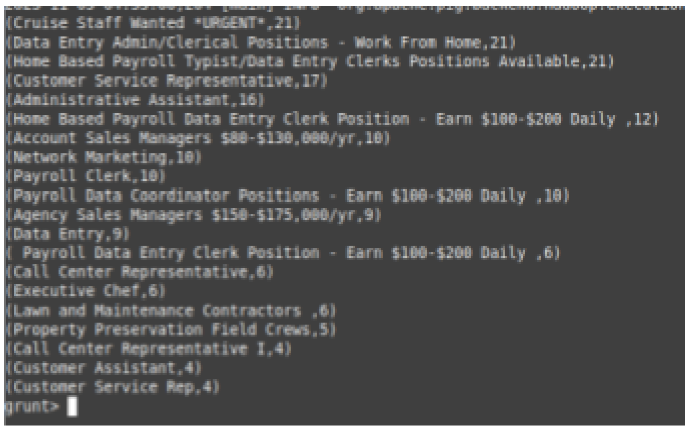
The rows in Hive that were initially corrupted by these errors

3. Analysing the Data with Hive and Pig

Query 1: Firstly, two simple queries were carried out with both Hive and Pig. The first of these examines the most common titles of fraudulent jobs in descending order. Both Hive and Pig yielded the same results.

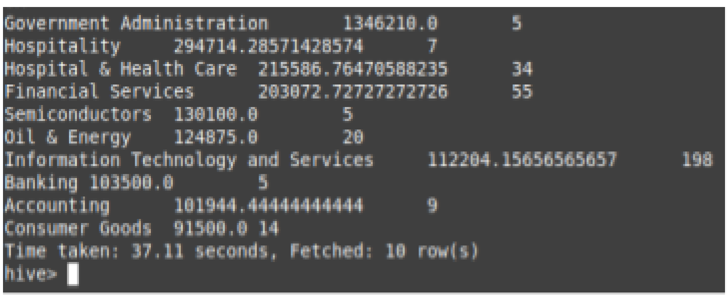


Query 1 In Hive

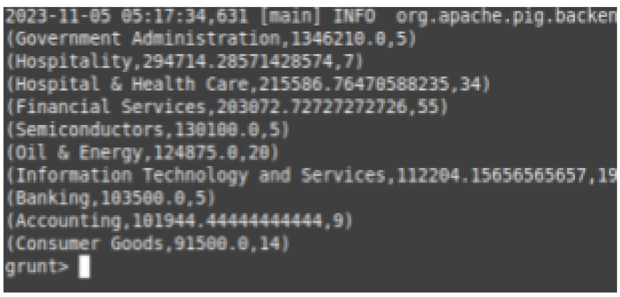


Query 1 In Pig

Query 2: The second query shows the top 10 average salaries by industry in the US. I chose one country in this example to remove issues related to currency rates. Additionally, to reduce the influence of outliers, only industries with more than 5 salaries were included. Once again Hive and Pig reached the same conclusion.

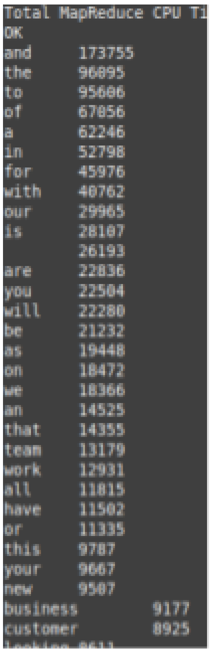


Query 2 in Hive



Query 2 in Pig

Query 3: The next three queries were done solely in Hive using some more advanced functions. The first of these gets a count of the words in the description giving an overview of the topics the jobs in the dataset tend to be related to. The SPLIT(' ') function in hive separates the description strings into an array of words. Then LATERAL VIEW explode turns this into a table which is then grouped by words and ordered by their count. The words are made lowercase to measure their frequency more accurately.



```
Total MapReduce CPU Ti
OK
and         173755
the         96095
to          95606
of          67056
a           62246
in          52798
for         45976
with        40762
our         29965
is          28107
            26193
are         22836
you         22504
will        22280
be          21232
as          19448
on          18472
we          18366
an          14525
that        14355
team        13179
work        12931
all         11815
have        11502
or          11335
this        9787
your        9667
new         9507
business         9177
customer         8925
```

Top word counts in the description field

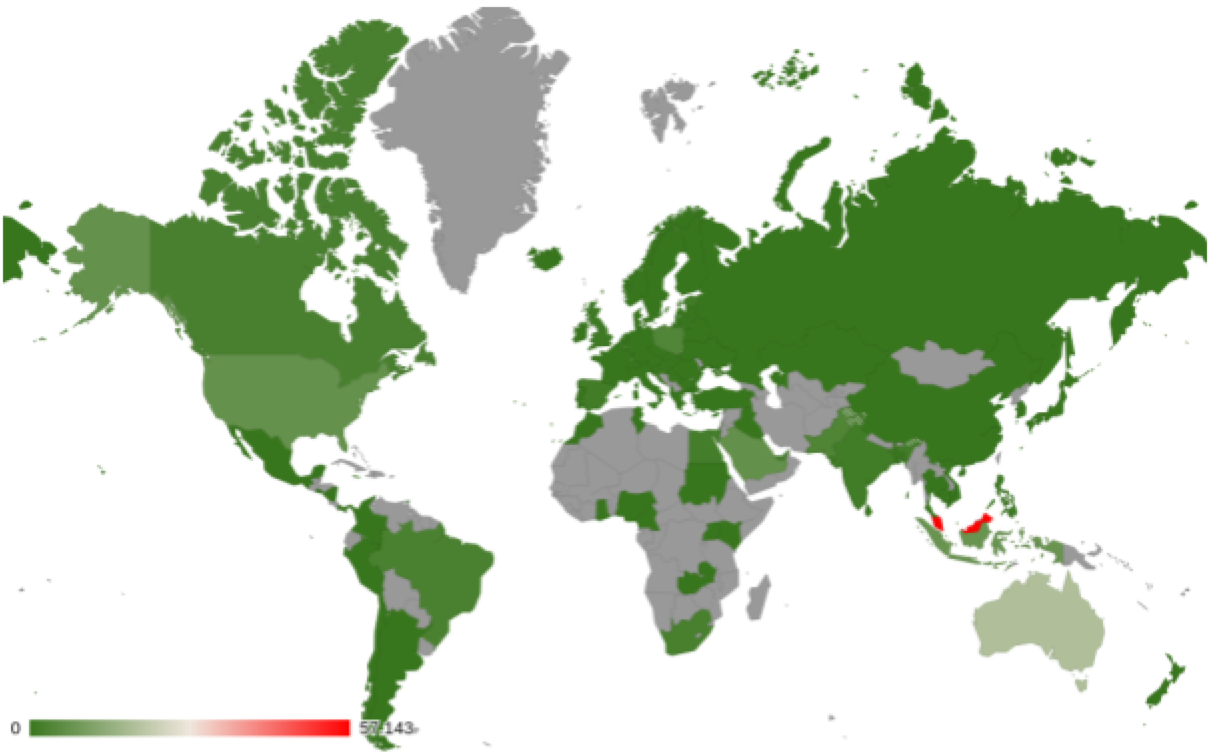
Top word counts after stop words are removed

As you might have expected, stop words such as 'and' or 'the' occur the most frequently, with more business and technology focused words starting to appear as you go down the list.

Query 4: The next query explores the percentage of fraudulent job postings by country. As the two letter country codes are quite hard to understand on their own I downloaded another table which elaborates on each of these codes from here. The job_listings table was joined with the coutry_codes table using the 2 letter representation as a common key. From here the country name, sub-region and region are included in the output alongside the total job count and the percentage of fraudulent jobs.

MY	Malaysia	South-eastern Asia	Asia	21	57.14285714285714
BH	Bahrain	Western Asia	Asia	9	55.55555555555556
TW	Taiwan, Province of China	Eastern Asia	Asia	4	50.0
QA	Qatar	Western Asia	Asia	21	28.57142857142857
AU	Australia	Australia and New Zealand	Oceania	214	18.69158878504673
ID	Indonesia	South-eastern Asia	Asia	13	7.6923076923076925
US	United States of America	Northern America	Americas	10656	6.850600600600601
SA	Saudi Arabia	Western Asia	Asia	15	6.666666666666667
PL	Poland	Eastern Europe	Europe	76	3.9473684210526314
PK	Pakistan	Southern Asia	Asia	27	3.7037037037037033
BR	Brazil	Latin America and the Caribbean	Americas	36	2.7777777777777777
CA	Canada	Northern America	Americas	457	2.62582056892779
ZA	South Africa	Sub-Saharan Africa	Africa	40	2.5

The countries with the largest percentage of fraudulent job postings



Countries coloured by the percentage of fraudulent job postings

Query 5: The final query makes use of the sampling features in Hive to examine 10% of the dataset. Then the sampled jobs were grouped by level of experience required, showing the total number of jobs per experience level, and the percentage of the sample that these jobs make up. In order to calculate the percentage of the sample the sum and count functions had to be combined using SUM(COUNT(required experience)) then the OVER() function calculates this over the whole sample.

```
        720     38.897893030794165
Mid-Senior level        410     22.15018908698001
Entry level     263     14.208535926526203
Associate       245     13.236088600756348
Not Applicable  131     7.0772555375472725
Director        39      2.106969205834684
Internship      34      1.8368449486763911
Executive       9       0.48622366288492713
Time taken: 59.291 seconds, Fetched: 8 row(s)
```

A 10% sample of the dataset broken down by experience required

4. Conclusion

In conclusion, this report demonstrates the level of transformation that can be required before a dataset can be properly loaded and queried. Pig proved to be effective at executing these transformations before the data was loaded to Hive. Both Hive and Pig were also able to perform in-depth analysis on the data unveiling new insights from the job listings. Running these tools in a distributed setting would give them large advantages over traditional querying and scripting languages, especially when working with large datasets.