Integrated CA2

HDip in Science in Artificial Intelligence Applications

Ciaran Quinlan(sbs23098@student.cct.ie)

Word count:

Lecturer: James Garza, Muhammad Iqbal

Table of Contents

Table of Contents 3
1. Introduction 4
2. SSBD Assessment Tasks 5
1. Question 1: Can you define Big Data? 5
1. Question 2: What is HDFS 5
Question 3: MySQL 7
Question 4:Flink 9
Question 5: Storm 10
7. Data Visualization Assessment 12
Irish Dataset 12
USA Dataset 12
Types of Data Visualizations 13
Choropleth Map. 13
Stock Heatmap aka Treemap 13
8. Conclusion 14
9. References 15
Appendix 1 15
Filenames and directories 15

1. Introduction

This is the integrated assessment for Muhammad and James.

- See the Appendix for list of files
- See Git @ https://github.com/ciaranquinlan/ssbd-ca2.git, you are both invited to this repo, everything i did is in there, images, code, markdown.
- All my assessment was run in a PC with the following ubuntu version

```
hduser@ciaran-ubun:~/ssbdca/ssbd-ca2/hadoop$ lsb_release -a
Distributor ID:Ubuntu
Description:   Ubuntu 22.04.3 LTS
Release:       22.04
Codename:      jammy
```

2. SSBD Assessment Tasks

1. Question 1: Can you define Big Data?

I would define Big Data as very large (terrabytes+) and complex data sets that cannot be handled or stored for fast retrieval and analysis by traditional data processing software. Big data is characterised in 3V's
Volume: Massive amounts of data stored on solid state disks with petebytes of data collected from webpages, IOT devices,cameras, sensors,vehicles etc
Velocity: Real time, instantaneous and changing data.
Variety: Structured (transactions, databases,data feeds), semi structured (XML, SGML, JSON), and unstructured (text, images, voice, video) data in multiple data formats.
and some add a further 2 V's Variability (data in numerous formats) and veracity (reliable and unreliable data).
Banks use big data to combat internal and external fraud, spot customer opportunities, better server regulators and manage risk in the money markets

I have come across these organisations who use big data:
Dunsink Observatory use big data to analyse the dark sky above Ireland to find new stars, monitor meteors and observe astrological threats.
Tesla use big data for analytics in their self-driving cars. This helps improve performance and safety and can identify and predict potential hazards on irish roads, such as other vehicles, pedestrians, and weather challenges.
Irish Revenue use big data to improve tax management, one example is capturing sharing economy income data from Airbnb. Also they use big data to manage collection of vat & customs fees from Amazon sales. The revenue also use big data to model taxpayer's behavior and risk profiles.

1. Question 2: What is HDFS

Hadoop Distributed File System (HDFS) is part of Apache Hadoop. HDFS is designed for storing and processing large data sets that can be distributed across many servers.
HDFS is different because it is:

- Distributed, meaning it can use many servers in multiple locations that operate together and Scalable, meaning that you can continue to add more servers as your data grows
- Fault Tolerant because it replicates data across multiple nodes so if nodes fail it can still operate..
- Optimised for massive datasets in the terabytes or petabytes in size.
- Write-Once, Read-Many simplifying data consistency with high performance.
- Data Locality: HDFS strives to place data on the same or nearby node where

The Hadoop framework consists of several layers, each serving a specific purpose:
HDFS storing and managing large data sets
YARN manages and allocates resources (
MapReduce does data processing tasks.
Hive is for data warehousing with SQL-like query language

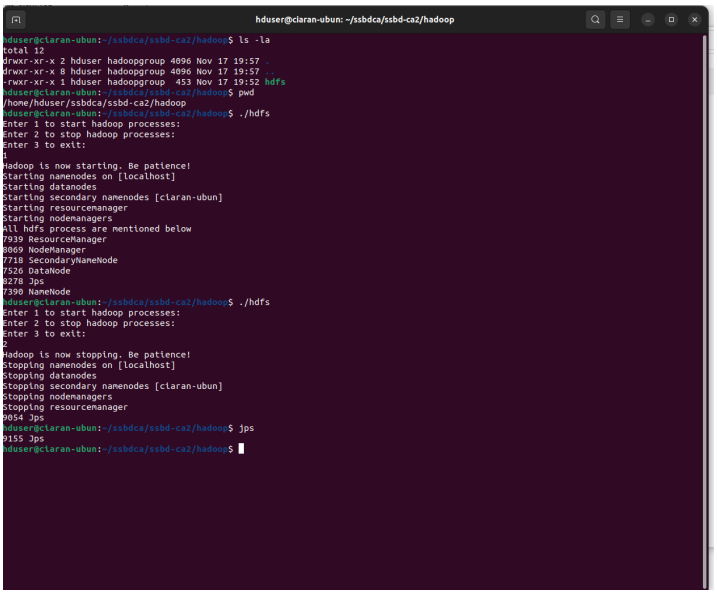
Pig a high-level data processing language
Spark not part of the Hadoop core but supports batch processing, real-time streaming, machine learning, and graph processing.

Characteristics of the Hadoop Framework:

Distributed enabling scalability and fault tolerance.
Fault Tolerant ensuring data availability if node fail.
Parallel Processing by splitting data into small chunks for speed processing.
Scalable accommodating growing data and processing needs.
Hadoop has a rich ecosystem of tools and libraries.

My deployment of HDFS is visible @
https://github.com/ciaranquinlan/ssbd-ca2/tree/main/hadoop

images @
https://github.com/ciaranquinlan/ssbd-ca2/blob/main/hadoop/Hadoop%20Screenshot%20from%202023-11-17%2020-02-10.png. It is running on a acer pc with ubuntu 18

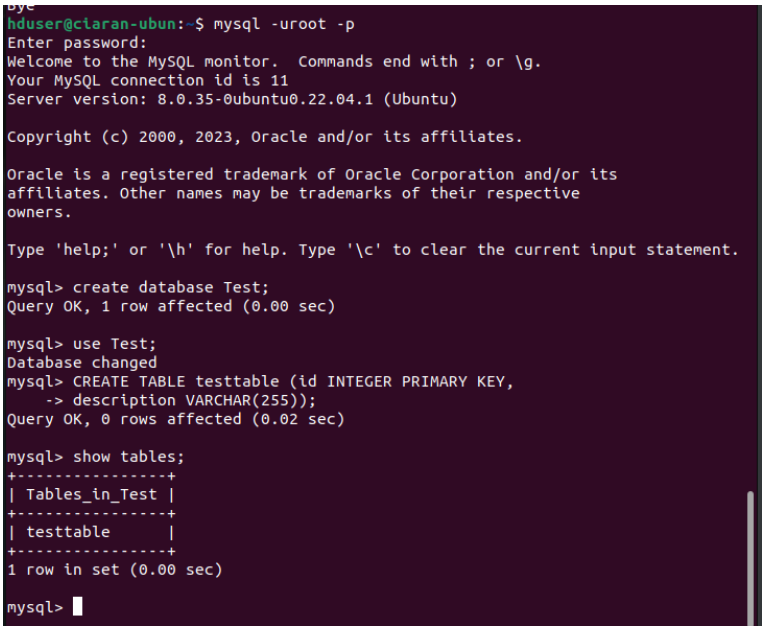


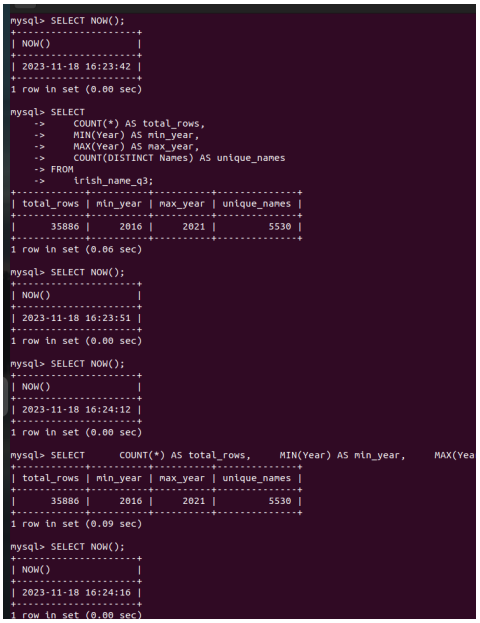
Question 3: MySQL

My Demonstration of MySQL and Apache Hive can be see in the folder
https://github.com/ciaranquinlan/ssbd-ca2/tree/main/mysql_hive

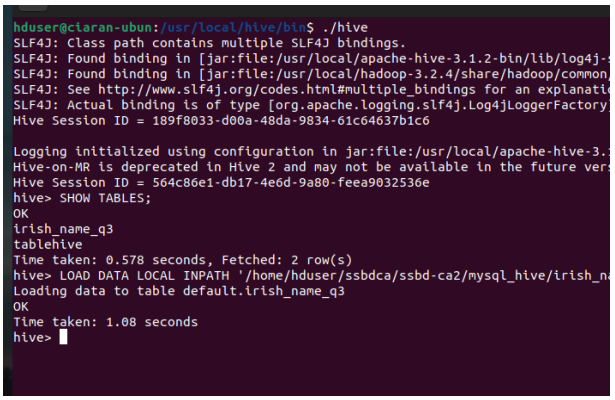
I used a Dataset : Irish names from 2015-2021, 35886 rows x 5 columns
Query: calculate total_rows, min_year & max_year, number of unique_names
Mysql: I created a database and imported the dataset of 35886 rows. I ran a query and the output in 0.9 sec was , see screenshot below:

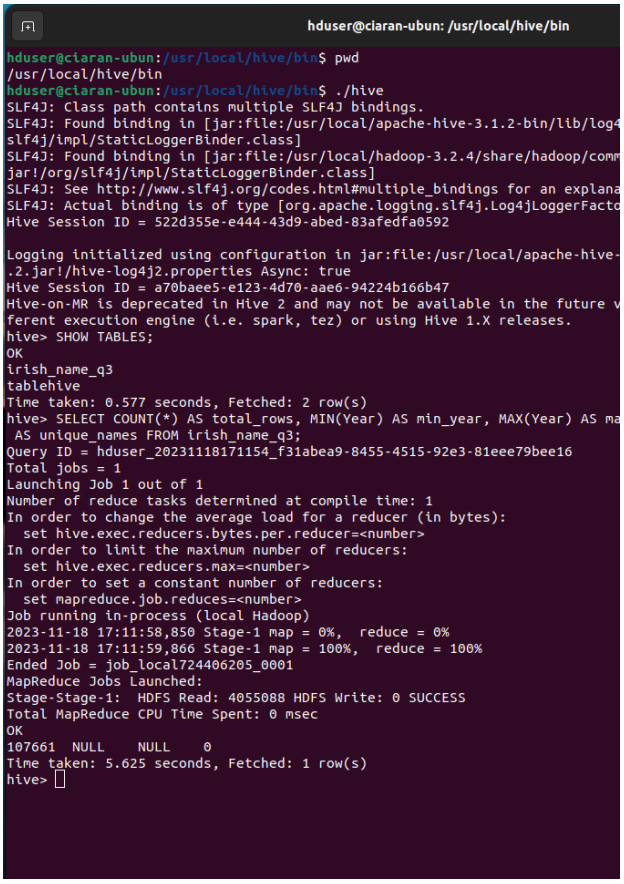
```
+------------+----------+----------+--------------+
| total_rows | min_year | max_year | unique_names |
+------------+----------+----------+--------------+
|      35886 |     2016 |     2021 |         5530 |
+------------+----------+----------+--------------+
```





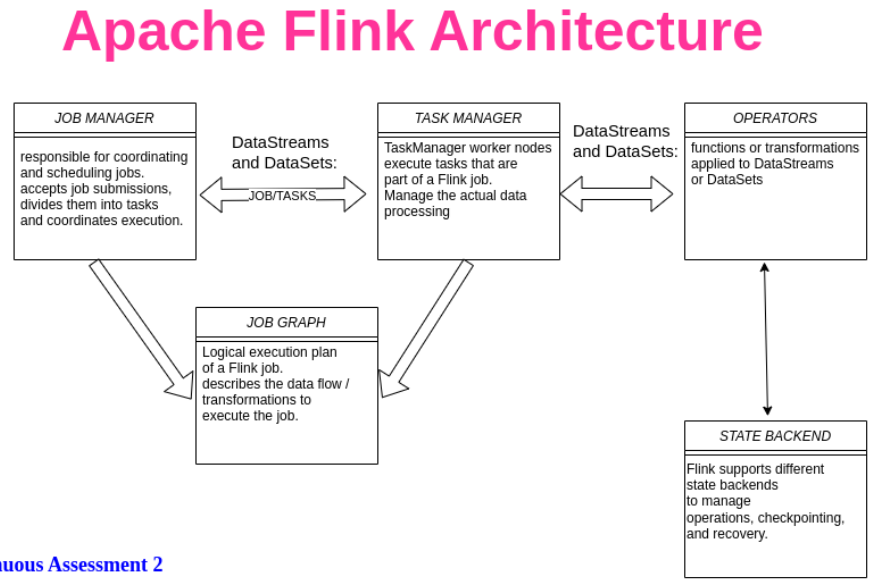
Hive: https://github.com/ciaranquinlan/ssbd-ca2/tree/main/mysql_hive
I started hadoop, ran the hive and imported the dataset, i ran the same query in the hive and it took longer to process than the mysql query. here is the hive screenshots, also in the mysql-hive directory.





Question 4:Flink

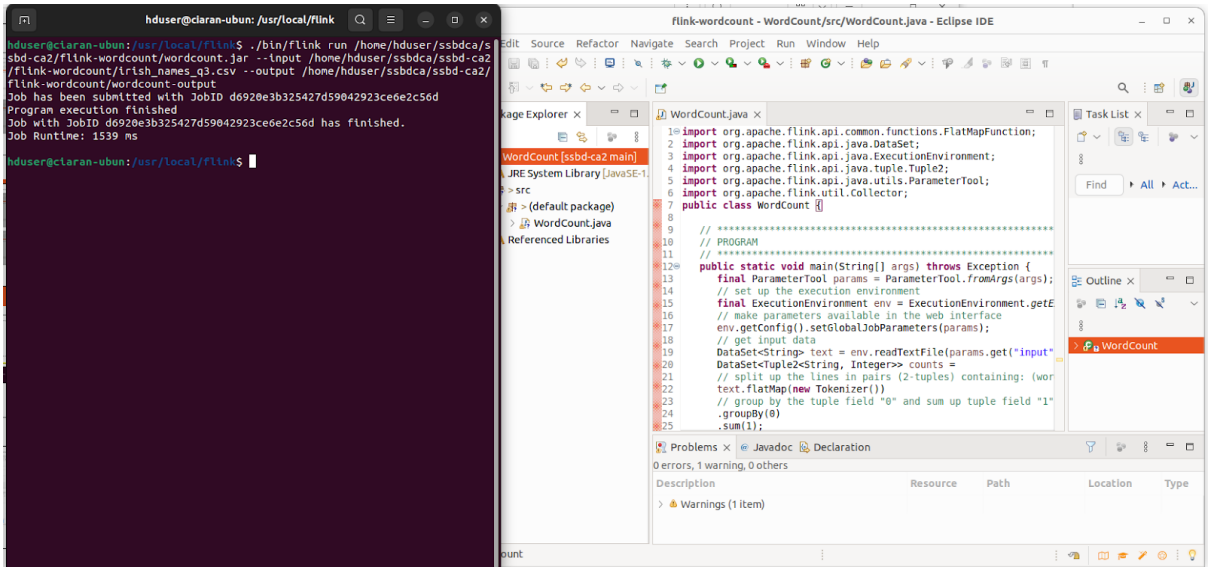
Apache Flink is a distributed stream processing framework and consists of the following components:
JobManager, TaskManagers, JobGraph, Operators, State Backend.
This is my Apache Flink Architecture:



Integrated Continuous Assessment 2
Lecturer Name:
Muhammad Iqbal, James Garza

Ciaran Quinlan (sbs23098@student.cct.ie)

My flink was installed and up and running see
https://github.com/ciaranquinlan/ssbd-ca2/tree/main/flink-wordcount
I used a same Dataset : Irish names from 2015-2021, 35886 rows x 5 columns. The Files are in this directory and I successfully ran the wordcount java and got the output file created.



Question 5: Storm

Why is Apache Storm useful for Stream processing specifically? Distinguish the characteristics of Apache storm as compared to Hadoop.
What is the role of Apache Zookeeper in Apache Storm deployment.

Why Apache Storm for Stream Processing:

Apache Storm is for real-time stream processing and it can process data streams that require low-latency and continuous processing.
Apache Storm is useful for stream processing because it has:

- Low Latency, that is Storm can run quickly when data processing and deliver real-time responses to applications such as fraud detection, live analytics and location recommendations.
- Event Processing: Storm handles events and messages in real time.
- Scalability: being part of the Apache hadoop ecosystem, storm can scale quickly.
- Fault Tolerant, Storm ensures data processing continues even in the event of node failures.
- Ease of Use, Storm makes it easier for developers to build real-time applications without dealing with event and error handling.

Here are some key distinctions between Apache Storm and Hadoop:

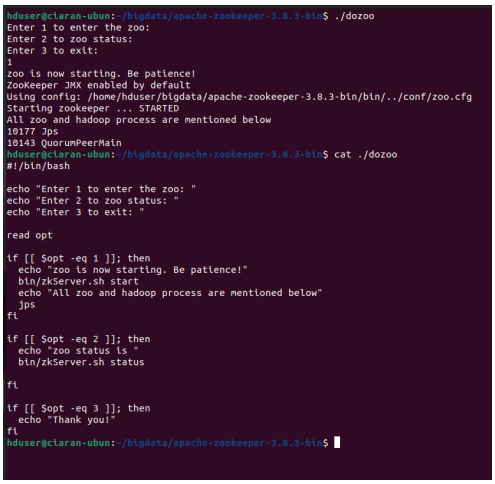
Storm is designed for real-time stream processing, Hadoop is primarily designed for batch processing. Storm uses a micro-batch processing model as events or messages arrive.
Hadoop uses a batch processing model

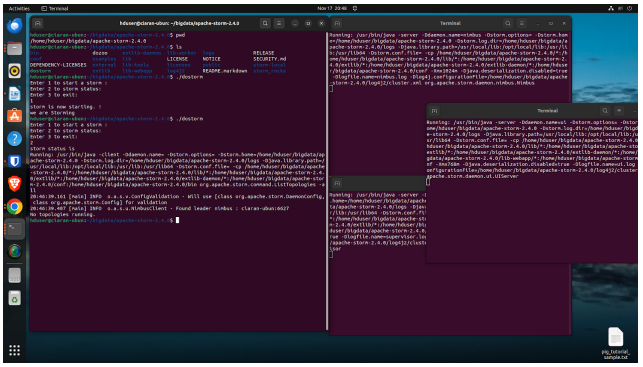
Storm is best suited for real-time analytics, event-driven processing, and immediate responses to data streams. Hadoop is ideal for batch-oriented tasks, offline data analysis, and processing of historical data.

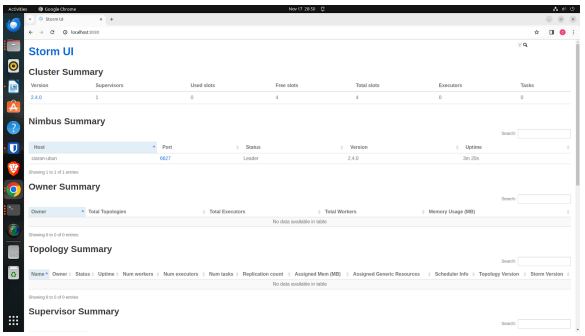
Apache ZooKeeper plays a critical role in Apache Storm deployments by providing coordination, configuration management, and distributed synchronisation among Storm components. Distributed State Management: Storm uses ZooKeeper for managing distributed states, such as tracking the progress of topologies and worker nodes. It helps maintain the state of tasks and components in the cluster. In summary, Apache ZooKeeper acts as the central nervous system of an Apache Storm cluster, providing coordination and management services to ensure the cluster's proper functioning and fault tolerance.

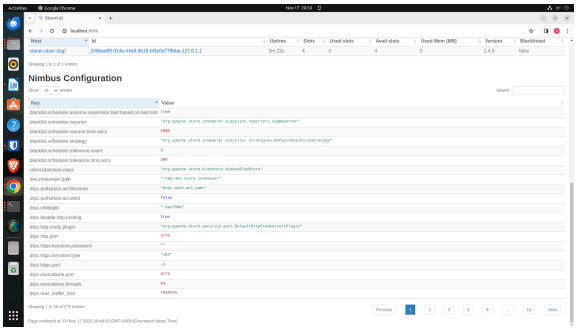
Screenshot of your VM to show working of Storm UI including Cluster, Nimbus and Owner summary.
All the files are in my storm directory in the repo :
https://github.com/ciaranquinlan/ssbd-ca2/tree/main/storm

I started Apache zoo with my dozoo script - 1, then started all 3 of the storm processes - 2 and went to localhost:8080 to view the storm process running on my ubuntu 3 and 4. All picture in the repo storm directory.









7. Data Visualization Assessment

This is the integrated assessment for James.

- See Git @ https://github.com/ciaranquinlan/ssbd-ca2.git, and https://github.com/ciaranquinlan/ssbd-ca2/tree/main/dataviz
  For all the dataviz file.

- All my assessment was run in a PC with ubuntu 21

Project Theme

The theme of my data visualisation project is to compare Baby names in Ireland and Baby names in the USA.

I hoped to build a dashboard where you enter a name and it will tell you how many times it is used in Ireland from 1964 to 2021 and if the name has been registered in any American state between 1910 and 2021. With such a large Irish American diaspora I wondered if they used a lot of the Irish names and if they avoided the ones that are hard to pronounce. We all know how Americans can't say or get their head around names like , Siobhan, Ailbhe, Caoilfhionn, Meadhbh, Saoirse, Seoirse, Niamh or even Caoimhín ?

The genesis behind this was researching my daughter's name Sine, (pronounced like Sheena) is registered not so frequently in Ireland and not at all in the USA, I could not believe it.
Some Irish names like Niamh do get entries in the USA but my data viz revealed it is confined mostly to the NY area with a few in Texas and CA. So this was an interesting dataset to work with. I did have some problems with special characters and I hope to work further on this issue.

One proviso is that any names registered in a year that are less that 3 in ireland or 5 in a US state are recorded as a zero in the dataset to protect privacy. So if you name does not appear for a year, my data visualisation will consider you special.
Another reason I dug a bit deeper into the dataset was my daughter who was born in 2004 was only one of possible 3 children born that year with that name as the name is registered as zero in the Irish dataset..

Irish Dataset

The dataset I used was found at https://www.kaggle.com/datasets/megan3/irish-baby-names, this dataset is based on the CSO data found here https://www.cso.ie/en/releasesandpublications/ep/p-ibn/irishbabiesnames2022/data/

USA Dataset

The datasets for US Baby Names at https://www.kaggle.com/datasets/robikscube/us-baby-name-popularity.

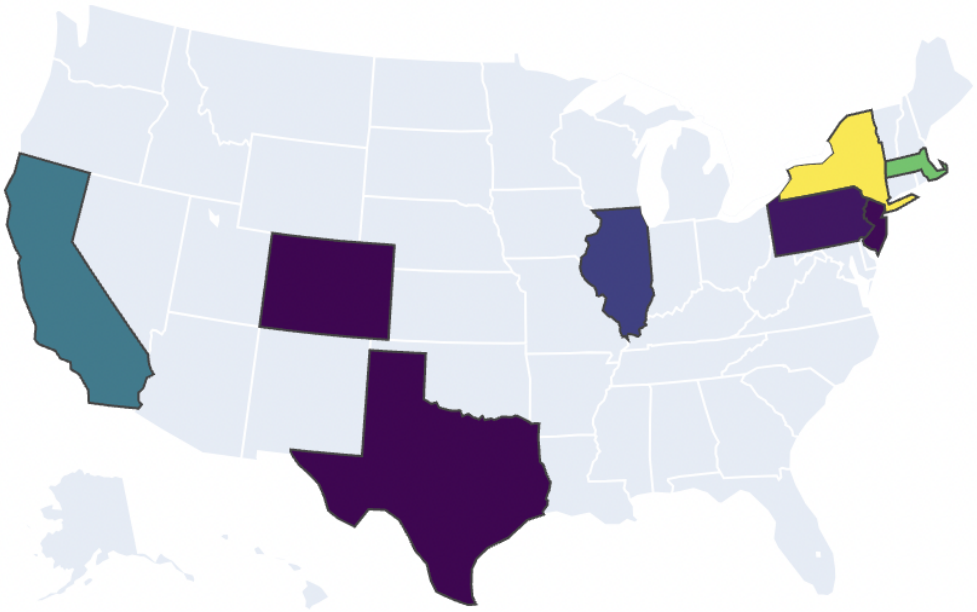
I did a bunch of changes to the datasets and merged both datasets. I am able to use the full datasets locally but since the file is over 350mb, for the project I removed any year before 1985 so that I could get the dataset small enough to fit in the git repository @ 99mb. The notebook for preparing the data is CIARANQ-IRE-USA-PrepData_forDataviz

Types of Data Visualizations

I spent a good bit of time working in draw.io I found it very clunky but I got the design done. I wanted the following plots

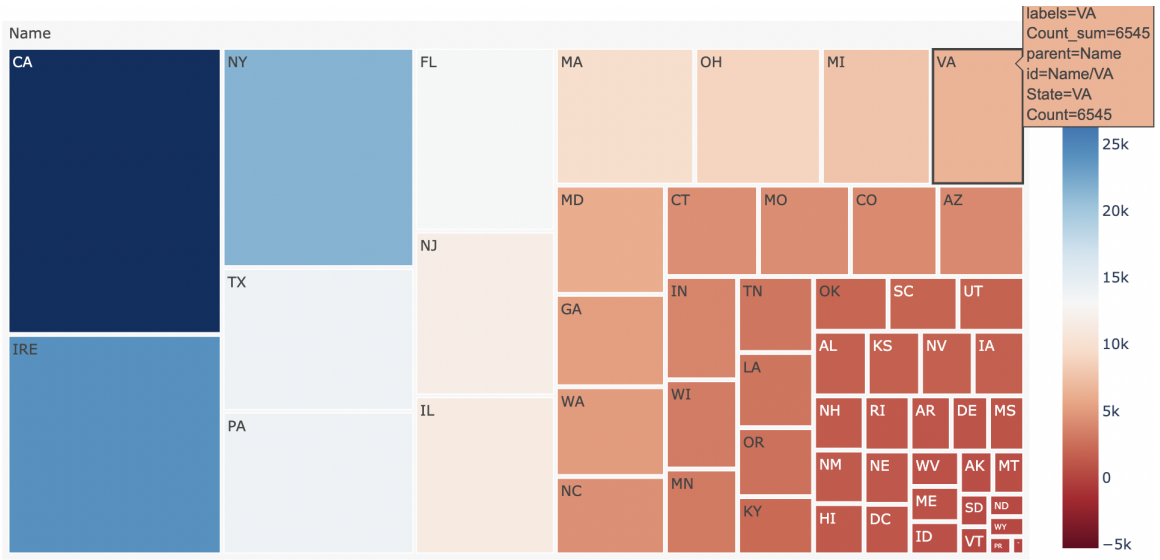
Choropleth Map.

With the USA data by state it was easy to map the count of names in each state. I could not get Ireland displayed side by side.



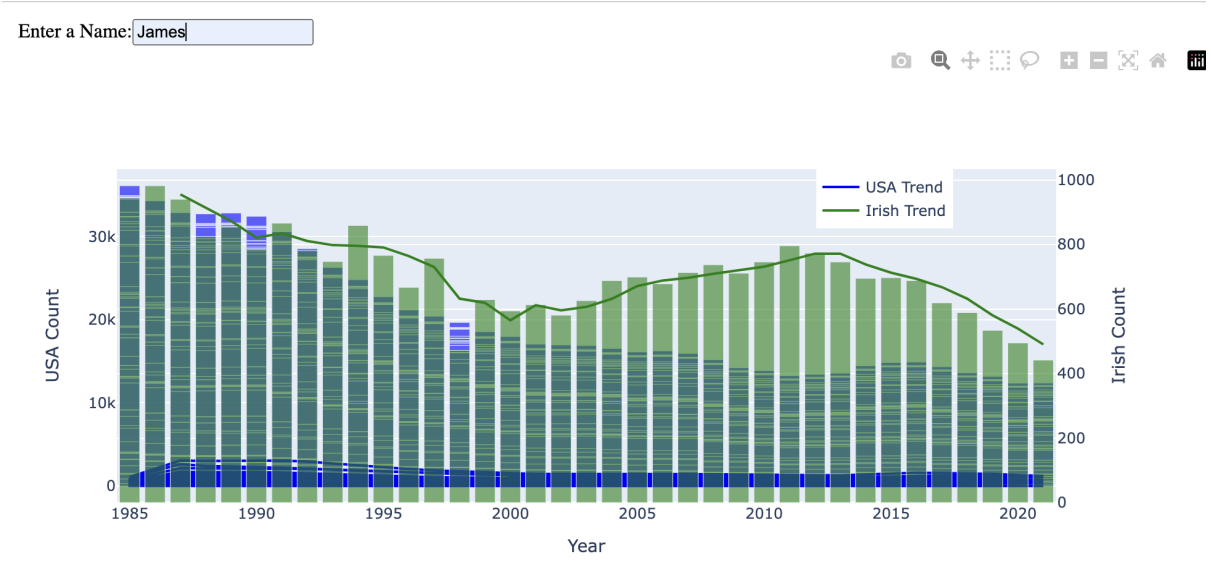
Stock Heatmap aka Treemap

I wanted to show the names by state with Ireland also represented. I got this to work in the test notebook but had issues when I ran it in dash. Here is a treemap for Sean showing just under California in the count for all years. CA, Ireland, NY,TX,PA,FL



2 Axis Bar Chart

I wanted to display a barchart which would show the trend over the years. I managed to do this and change the colours to show the bars and the trend.



8. Conclusion

I really enjoyed both projects and it was very rewarding to do the assignments. I ran out of time at the end to do some final touches but I learnt a lot about the subjects.

Many thanks for all the help and guidance throughout the year.

Ciaran Quinan

9. References

Appendix 1

Filenames and directories

Appendix A

Test Notebooks

Bar charts - I did a 2 axis bar chart, this worked well and showed the trend in the names in both countries. Getting the same effect with Dash was a little harder and I could not get the effect desired in this notebook.

Stock Heatmap
I spend ages trying to get a stock heat map like they use in financial examples of stocks. Finally I found out it's actually called a treemap and got the code on plotly.com that did a treemap that displayed my count of names as a set of nested rectangles. This was exactly what I wanted but crashed sometimes if the given name was zero for some states.

GivennameLine chart
This was my test to get a trend line of a name, I was able to use this within my bar chart to show the trend and the bar on a 2 axis graph.

Plot choropleth.
This was my test of choropleth, I hoped to get the USA and Ireland side by side on the same plot but ran into difficulties.