

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 |
| Project Theme | 12 |
| Irish Dataset | 12 |
| USA Dataset | 13 |
| Types of Data Visualizations | 13 |
| Choropleth Map. | 13 |
| Stock Heatmap aka Treemap | 13 |
| 2 Axis Bar Chart | 14 |
| 8. Conclusion | 15 |
| 9. References | 16 |
| Appendix 1 | 16 |
| Filenames and directories | 16 |
| Appendix 2 | 17 |
| Test Notebooks | 17 |

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 (terabytes+) 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 petabytes 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 serve 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.

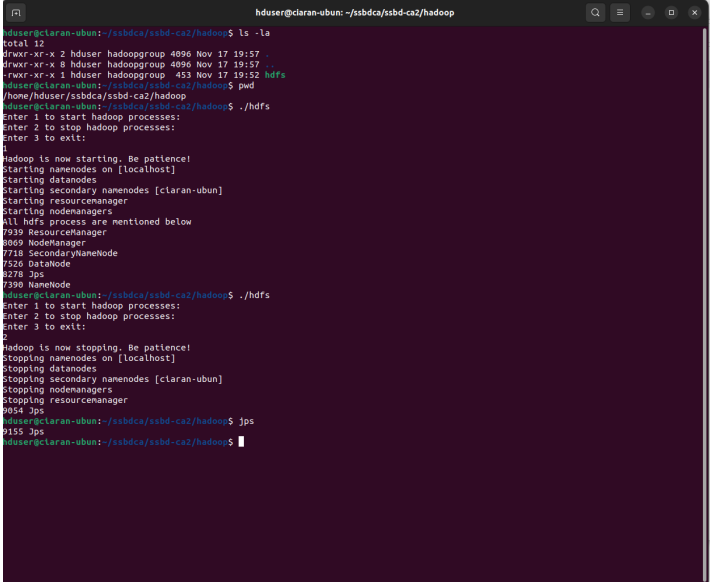
My deployment of HDFS is visible @

<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%2002-10.png>. It is running on a acer pc with ubuntu

18



```
hduser@ciaran-ubuntu: ~/ssbdca/ssbd-ca2/hadoop
hduser@ciaran-ubuntu:~/ssbdca/ssbd-ca2/hadoop$ ls -la
total 12
drwxr-xr-x 2 hduser hadoopgroup 4096 Nov 17 19:57
drwxr-xr-x 8 hduser hadoopgroup 4096 Nov 17 19:57
-rwxr-xr-x 1 hduser hadoopgroup 453 Nov 17 19:52 hdfs
hduser@ciaran-ubuntu:~/ssbdca/ssbd-ca2/hadoop$ pwd
/home/hduser/ssbdca/ssbd-ca2/hadoop
hduser@ciaran-ubuntu:~/ssbdca/ssbd-ca2/hadoop$ ./hdfs
Enter 1 to start hadoop processes:
Enter 2 to stop hadoop processes:
Enter 3 to exit:
1
Hadoop is now starting. Be patience!
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [ciaran-ubun]
Starting resourcemanager
Starting nodemanagers
All hdfs process are mentioned below
7930 ResourceManager
8869 NodeManager
7718 SecondaryNameNode
7526 DataNode
8278 Jps
7390 NameNode
hduser@ciaran-ubuntu:~/ssbdca/ssbd-ca2/hadoop$ ./hdfs
Enter 1 to start hadoop processes:
Enter 2 to stop hadoop processes:
Enter 3 to exit:
2
Hadoop is now stopping. Be patience!
Stopping namenodes on [localhost]
Stopping datanodes
Stopping secondary namenodes [ciaran-ubun]
Stopping nodemanagers
Stopping resourcemanager
8054 Jps
9155 Jps
hduser@ciaran-ubuntu:~/ssbdca/ssbd-ca2/hadoop$
```

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 |
+-----+-----+-----+-----+
```

```
hduser@ciaran-ubun:~/ssbdca/ssbd-ca2/mysql$ mysql -uroot -p
Enter password:
Welcome to the MySQL monitor. Commands end with ; or \g.
Your MySQL connection id is 14
Server version: 8.0.35-0ubuntu0.22.04.1 (Ubuntu)

Copyright (c) 2000, 2023, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> create database q3;
ERROR 1007 (HY000): Can't create database 'q3'; database exists
mysql> create database question3;
Query OK, 1 row affected (0.01 sec)

mysql> use question3
Database changed
mysql> CREATE TABLE Irish_name_q3 (
  -> Year INT,
  -> Names VARCHAR(255),
  -> Sex VARCHAR(255),
  -> Count INT,
  -> Ranks VARCHAR(255)
  -> );
Query OK, 0 rows affected (0.02 sec)

mysql>
mysql> SHOW TABLES;
+-----+
| Tables_in_question3 |
+-----+
| Irish_name_q3       |
+-----+
1 row in set (0.00 sec)

mysql> LOAD DATA INFILE '/var/lib/mysql-files/Irish_names_q3.csv'
  -> INTO TABLE Irish_name_q3
  -> FIELDS TERMINATED BY ','
  -> ENCLOSED BY '"'
  -> LINES TERMINATED BY '\n'
  -> IGNORE 1 ROWS
  -> (Year, Names, Sex, @varCount, Ranks)
  -> SET Count = NULLIF(@varCount, '');
Query OK, 35886 rows affected (0.35 sec)
Records: 35886 Deleted: 0 Skipped: 0 Warnings: 0

mysql>
```

```
hduser@ciaran-ubun:~$ mysql -uroot -p
Enter password:
Welcome to the MySQL monitor. Commands end with ; or \g.
Your MySQL connection id is 11
Server version: 8.0.35-0ubuntu0.22.04.1 (Ubuntu)

Copyright (c) 2000, 2023, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> create database Test;
Query OK, 1 row affected (0.00 sec)

mysql> use Test;
Database changed
mysql> CREATE TABLE testtable (id INTEGER PRIMARY KEY,
  -> description VARCHAR(255));
Query OK, 0 rows affected (0.02 sec)

mysql> show tables;
+-----+
| Tables_in_Test |
+-----+
| testtable      |
+-----+
1 row in set (0.00 sec)

mysql>
```

```
mysql> SELECT NOW();
+-----+
| NOW() |
+-----+
| 2023-11-18 16:23:42 |
+-----+
1 row in set (0.00 sec)

mysql> SELECT
  -> COUNT(*) AS total_rows,
  -> MIN(Year) AS min_year,
  -> MAX(Year) AS max_year,
  -> COUNT(DISTINCT Names) AS unique_names
  -> FROM
  -> Irish_name_q3;
+-----+-----+-----+-----+
| total_rows | min_year | max_year | unique_names |
+-----+-----+-----+-----+
| 35886 | 2016 | 2021 | 5530 |
+-----+-----+-----+-----+
1 row in set (0.06 sec)

mysql> SELECT NOW();
+-----+
| NOW() |
+-----+
| 2023-11-18 16:23:51 |
+-----+
1 row in set (0.00 sec)

mysql> SELECT NOW();
+-----+
| NOW() |
+-----+
| 2023-11-18 16:24:12 |
+-----+
1 row in set (0.00 sec)

mysql> SELECT COUNT(*) AS total_rows, MIN(Year) AS min_year, MAX(Yea
  -> total_rows | min_year | max_year | unique_names |
  -> 35886 | 2016 | 2021 | 5530 |
  -> 1 row in set (0.09 sec)

mysql> SELECT NOW();
+-----+
| NOW() |
+-----+
| 2023-11-18 16:24:16 |
+-----+
1 row in set (0.00 sec)
```

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.

```
hduser@ciaran-ubun: /usr/local/hive/bin$ ./hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.2-bin/lib/log4j-
SLF4J: Found binding in [jar:file:/usr/local/hadoop-3.2.4/share/hadoop/common
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanati
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory
Hive Session ID = 189f8033-d00a-48da-9834-61c64637b1c6

Logging initialized using configuration in jar:file:/usr/local/apache-hive-3.
Hive-on-MR is deprecated in Hive 2 and may not be available in the future ver
Hive Session ID = 564c86e1-db17-4e6d-9a80-fee9032536e
hive> SHOW TABLES;
OK
irish_name_q3
tablehive
Time taken: 0.578 seconds, Fetched: 2 row(s)
hive> LOAD DATA LOCAL INPATH '/home/hduser/ssbdca/ssbd-ca2/mysql_hive/irish_n
Loading data to table default.irish_name_q3
OK
Time taken: 1.08 seconds
hive>
```

```
hduser@ciaran-ubun: /usr/local/hive/bin$ pwd
/usr/local/hive/bin
hduser@ciaran-ubun: /usr/local/hive/bin$ ./hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.2-bin/lib/log4j-
slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-3.2.4/share/hadoop/comm
jar/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explan
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFacto
Hive Session ID = 522d355e-e444-43d9-abad-83afedfa0592

Logging initialized using configuration in jar:file:/usr/local/apache-hive-
.2.jar!/hive-log4j2.properties Async: true
Hive Session ID = a70baee5-e123-4d70-aae6-94224b166b47
Hive-on-MR is deprecated in Hive 2 and may not be available in the future v
ferent execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> SHOW TABLES;
OK
irish_name_q3
tablehive
Time taken: 0.577 seconds, Fetched: 2 row(s)
hive> SELECT COUNT(*) AS total_rows, MIN(Year) AS min_year, MAX(Year) AS ma
AS unique_names FROM irish_name_q3;
Query ID = hduser_20231118171154_f31abea9-8455-4515-92e3-81eee79bee16
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2023-11-18 17:11:58,850 Stage-1 map = 0%, reduce = 0%
2023-11-18 17:11:59,866 Stage-1 map = 100%, reduce = 100%
Ended Job = job_local72406205_0001
MapReduce Jobs Launched:
Stage-Stage-1: HDFS Read: 4055088 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
107661 NULL NULL 0
Time taken: 5.625 seconds, Fetched: 1 row(s)
hive>
```

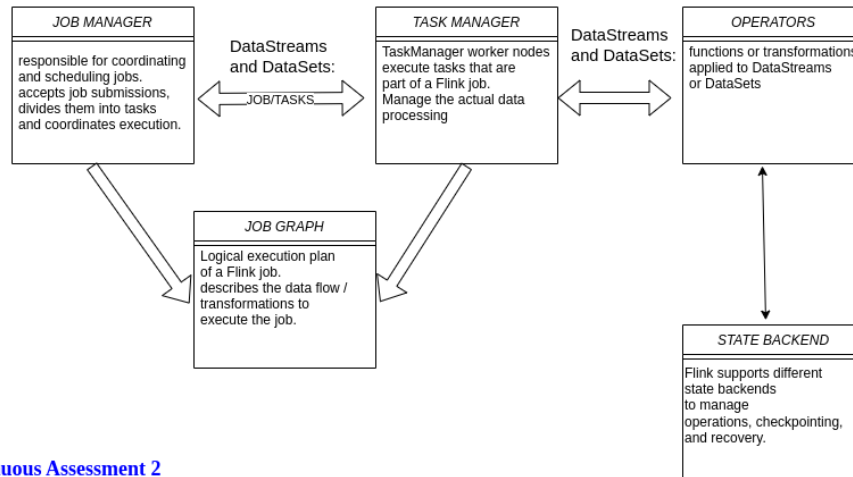

Question 4:Flink

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:

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.

The screenshot shows a terminal window on the left and an Eclipse IDE on the right. The terminal window displays the command to run the Flink wordcount job and the output, including the job ID and runtime. The Eclipse IDE shows the source code of the WordCount.java file, which uses the Flink API to read a CSV file, tokenize the text, and calculate the word counts.

```
hduser@ciaran-ubuntu: /usr/local/flink$ ./bin/flink run /home/hduser/ssbdca/ssbd-ca2/flink-wordcount/wordcount.jar --input /home/hduser/ssbdca/ssbd-ca2/flink-wordcount/irish_names_q3.csv --output /home/hduser/ssbdca/ssbd-ca2/flink-wordcount/wordcount-output
Job has been submitted with JobID d6920e3b325427d59042923ce6e2c56d
Program execution finished
Job with JobID d6920e3b325427d59042923ce6e2c56d has finished.
Job Runtime: 1539 ms

hduser@ciaran-ubuntu: /usr/local/flink$
```

```
1 import org.apache.flink.api.common.functions.FlatMapFunction;
2 import org.apache.flink.api.java.DataSet;
3 import org.apache.flink.api.java.ExecutionEnvironment;
4 import org.apache.flink.api.java.tuple.Tuple2;
5 import org.apache.flink.api.java.util.ParameterTool;
6 import org.apache.flink.util.Collector;
7
8 public class WordCount {
9
10 // PROGRAM
11 // *****
12 public static void main(String[] args) throws Exception {
13     final ParameterTool params = ParameterTool.fromArgs(args);
14     // set up the execution environment
15     final ExecutionEnvironment env = ExecutionEnvironment.getExecutionEnvironment();
16     // make parameters available in the web interface
17     env.getConfig().setGlobalJobParameters(params);
18     // get input data
19     DataSet<String> text = env.readTextFile(params.get("input"));
20     DataSet<Tuple2<String, Integer>> counts =
21         text.flatMap(new Tokenizer())
22             .groupBy(0)
23             .sum(1);
24 }
25
```

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>

```
hdsuser@claran-ubuntu:~/bigdata$ ./dozoo
Enter 1 to enter the zoo:
Enter 2 to zoo status:
Enter 3 to exit:
1
zoo is now starting. Be patience!
ZooKeeper JMX enabled by default
Using config: /home/hdsuser/bigdata/apache-zookeeper-3.8.3-bin/bin/./conf/zoo.cfg
Starting zookeeper... STARTED
All zoo and hadoop process are mentioned below
107177 jps
18143 QuorumPeerMain
hdsuser@claran-ubuntu:~/bigdata/apache-zookeeper-3.8.3-bin$ cat ./dozoo
#!/bin/bash

echo "Enter 1 to enter the zoo: "
echo "Enter 2 to zoo status: "
echo "Enter 3 to exit: "

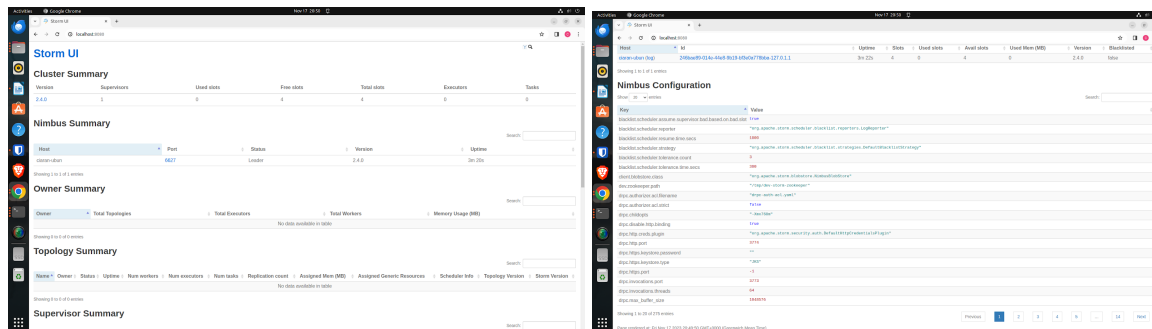
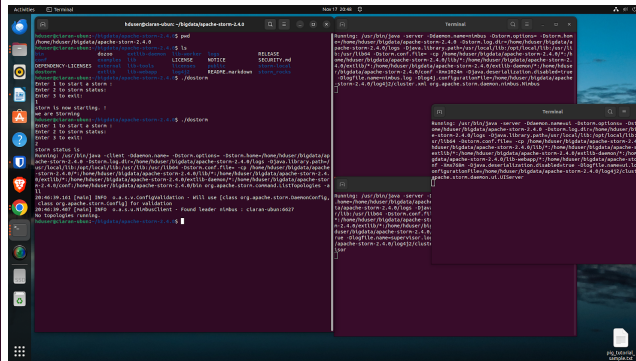
read opt

if [[ $opt -eq 1 ]]; then
    echo "zoo is now starting. Be patience!"
    bin/zkServer.sh start
    echo "All zoo and hadoop process are mentioned below"
    jps
fi

if [[ $opt -eq 2 ]]; then
    echo "zoo status is "
    bin/zkServer.sh status
fi

if [[ $opt -eq 3 ]]; then
    echo "Thank you!"
fi

hdsuser@claran-ubuntu:~/bigdata/apache-zookeeper-3.8.3-bin$
```



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 than 3 in Ireland or 5 in a US state are recorded as a zero in the dataset to protect privacy. So if your 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>.

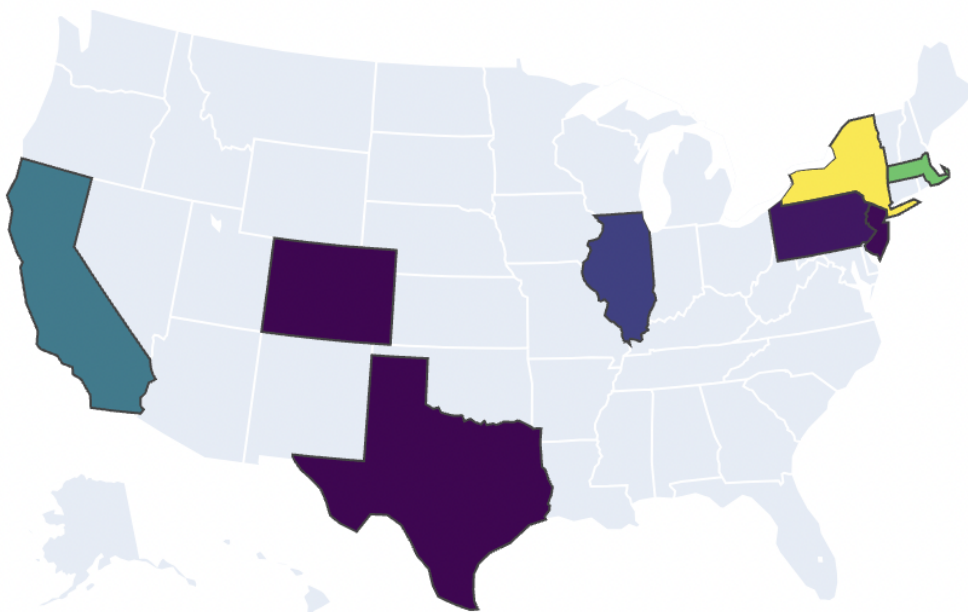
I did a bunch of changes to the datasets and merged both datasets. I am able to use the full datasets locally but since the file is over 350mb, for the project I removed any year before 1985 so that I could get the dataset small enough to fit in the git repository @ 99mb. The notebook for preparing the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

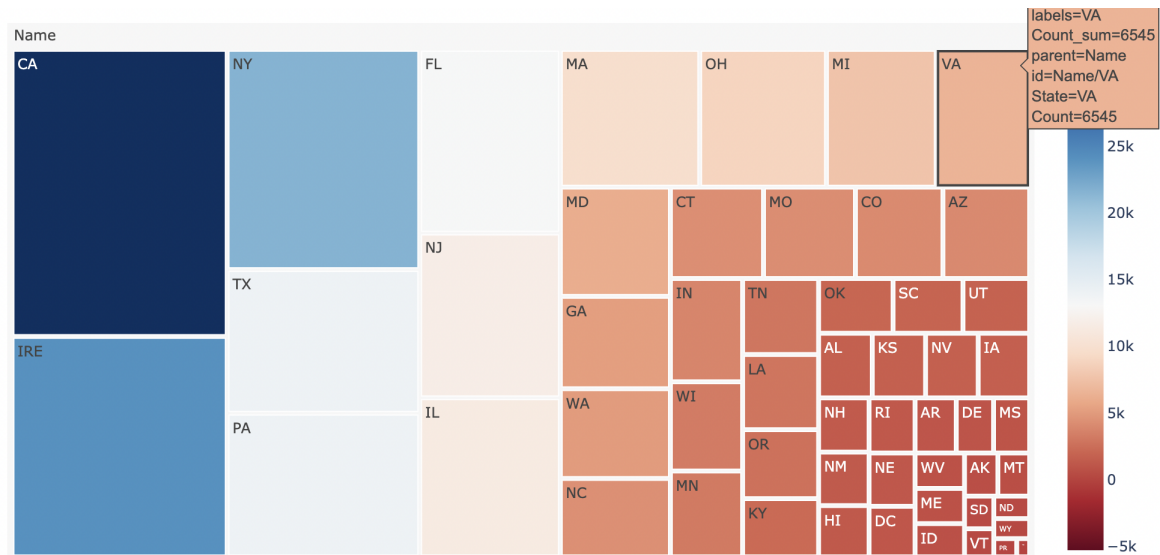
Choropleth Map.

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

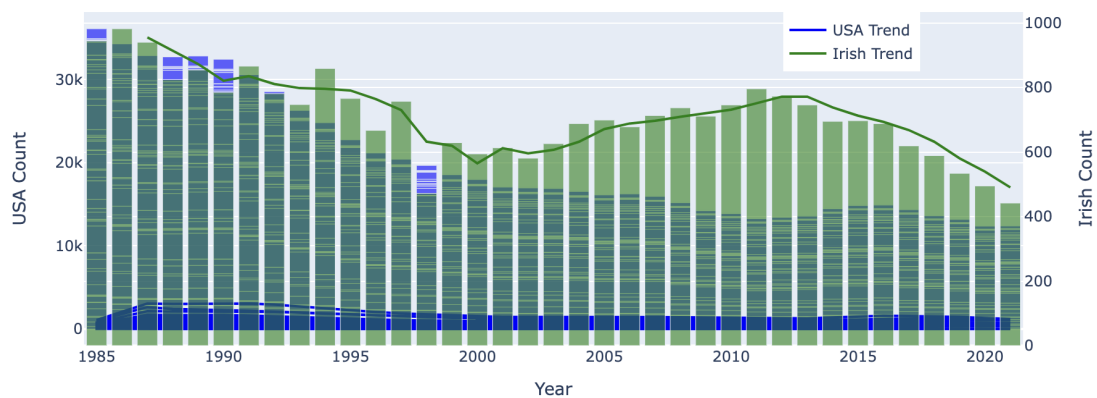
I wanted to show the names by state with Ireland also represented. I got this to work in the test notebook but had issues when I ran it in dash. Here is a treemap for Sean showing just under California in the count for all years. CA, Ireland, NY,TX,PA,FL



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.

Enter a Name:



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

| | |
|--|--|
| 1.CIARANQ-IRE-USA-PrepData_forDataviz.ipynb | Main data prep file |
| 2.CIARANQ-IRE-USA-6panel-name.ipynb | Dashboard |
| bar-charts.ipynb Given-name-linechart.ipynb plot-choropleth.ipynb treemap.ipynb | Test notebooks also see cass notebooks |
| *.csv | datasets |
| Wireframe.drawio wireframe.drawio.png | Draw.io files and wireframes |
| | |
| | |
| | |

Appendix 2

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.