ASSIGNMENT 3

Name – Viraj Joshi

Dal ID – B00924759

Email – viraj.joshi@dal.ca

GIT Problem 1 – https://git.cs.dal.ca/vjoshi/assignment-3-part-1.git

GIT Problem 2 – https://git.cs.dal.ca/vjoshi/assignment-3-part-2.git

PART 1

Data Extraction || Transformation || Store (related to ETL in BigData)

The following algorithm was implemented to extract the news and apply cleaning/transformation on the same.

Step 1: Read the HTTP response from the API for each keyword.

Step2: Save the HTTP response for each keyword response in a separate .txt file.

Step3: Read the contents of each file in a string (one loop execution per keyword.txt file)

Step4: Save the string into a JSON object and retrieve every “article” value into a JSON array.

Step5: Iterate over the array and read contents of each article of each keyword.

Step6: Save unfiltered article content in a string.

Step7: Use string from step 6 and replace all URL and UrlToImage links, replace all null values, carriage return and new link (/r /n) characters, HTML tags (
, , <td> , <th>) tags, Unicode values (\u….) and special characters (+[], emoticons) with empty string.

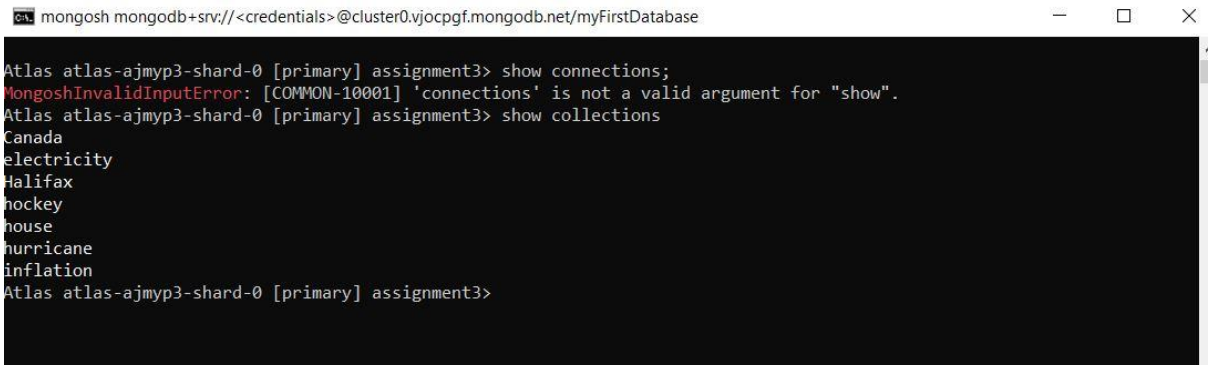
Step8: Save the filtered article content into a BSON Document.

Step9: Create a collection for each keyword with name = keyword.

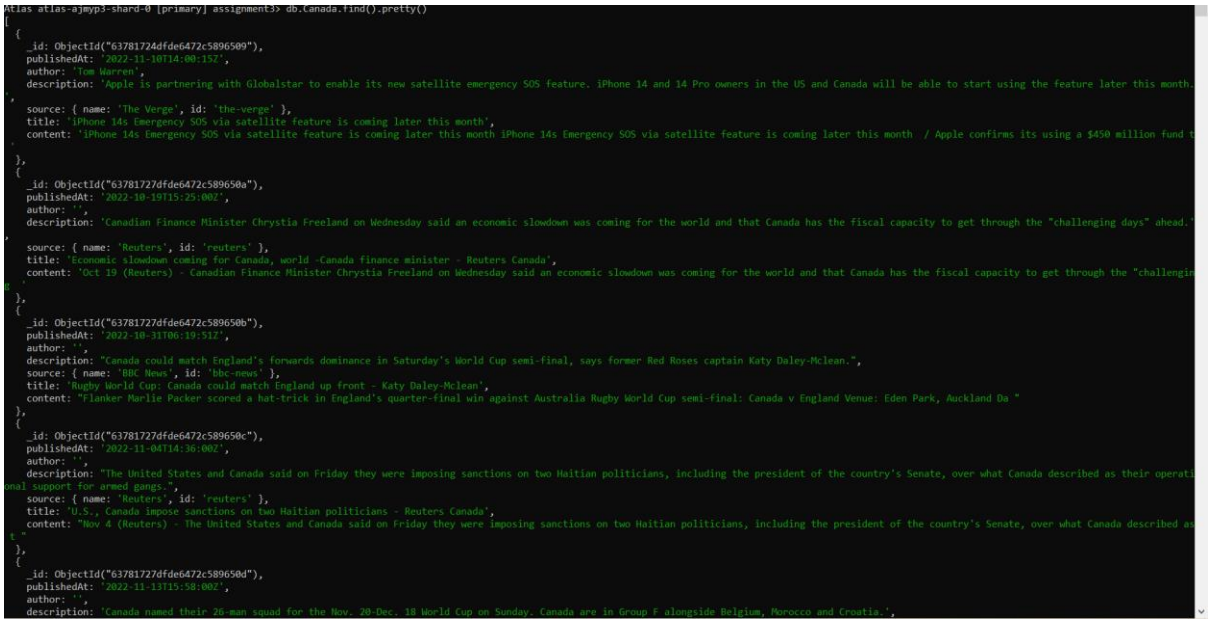
Step10: Insert BSON document from step 8 into the collection.

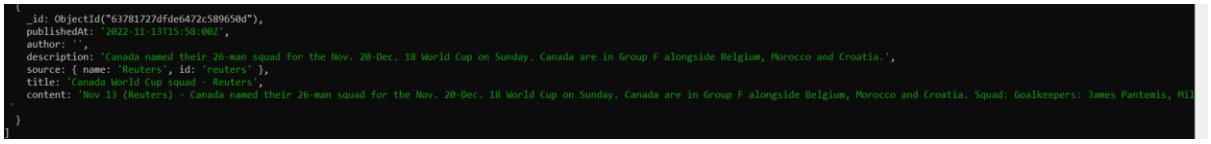
MongoDB screenshots after execution of algorithm.

Collection created for each keyword.

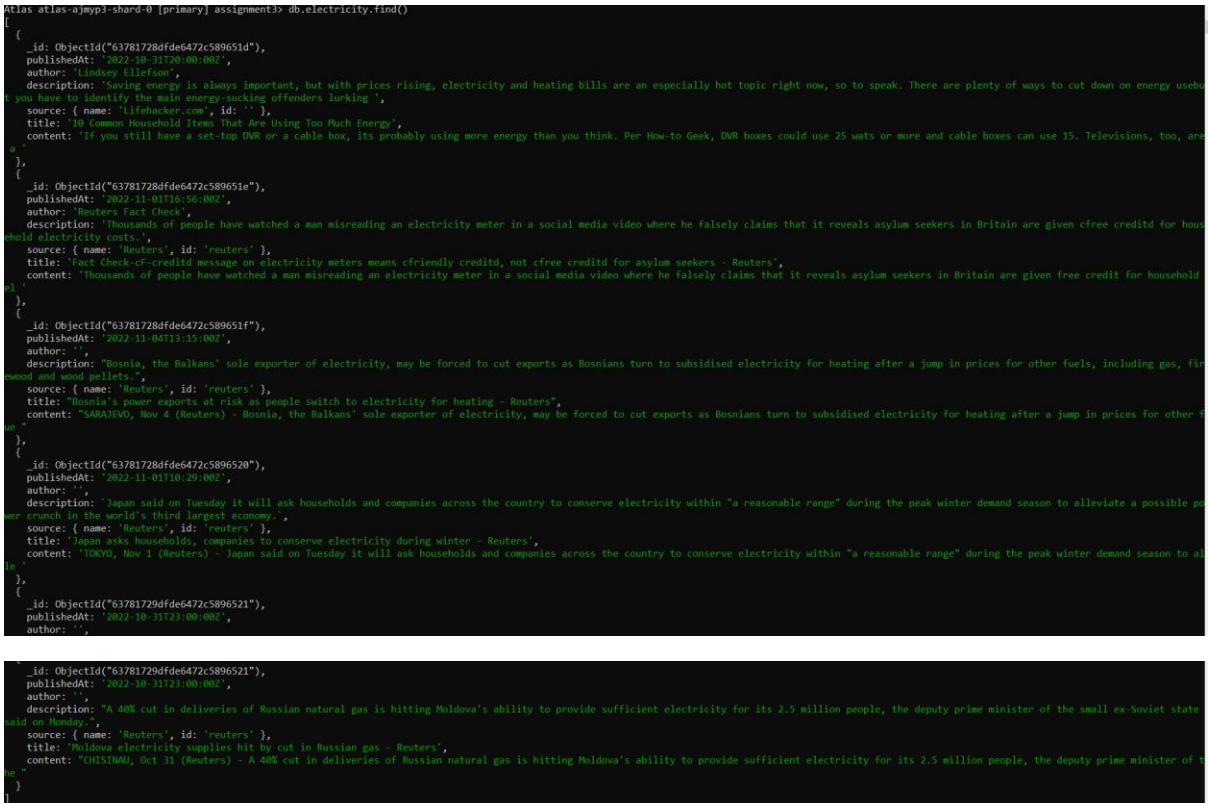


Collection ‘Canada’ with articles as ‘Document’

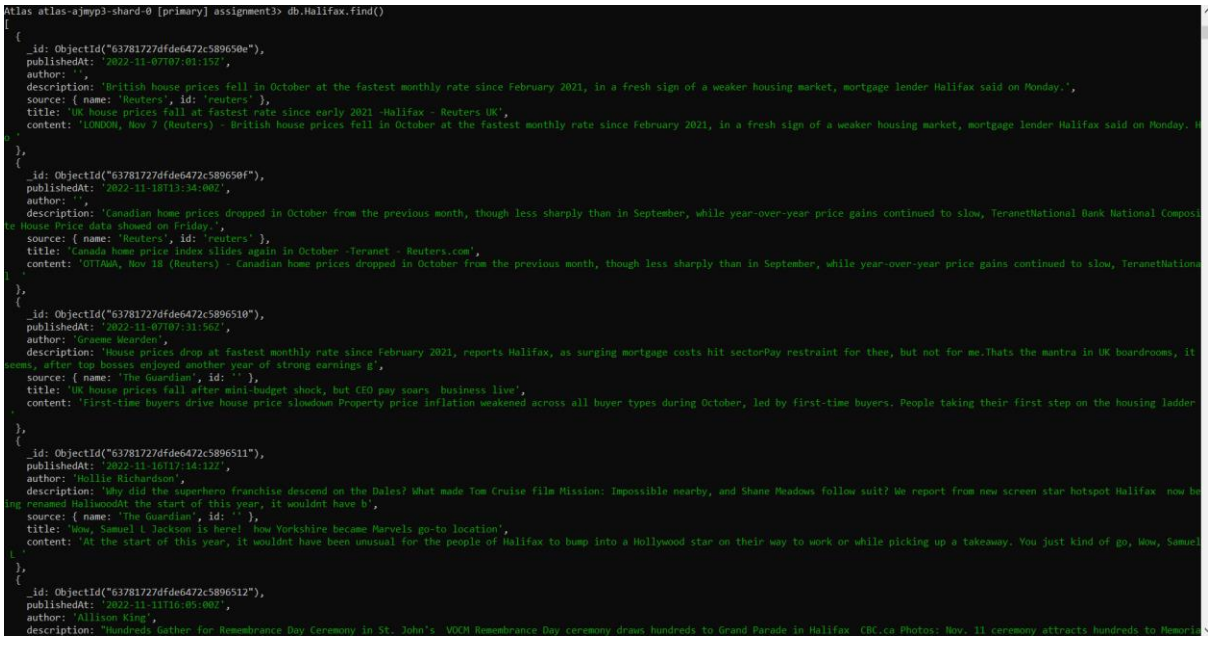


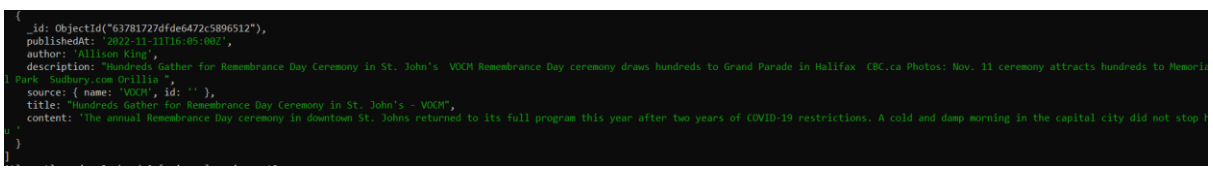


Collection ‘electricity’ with articles as ‘Document’

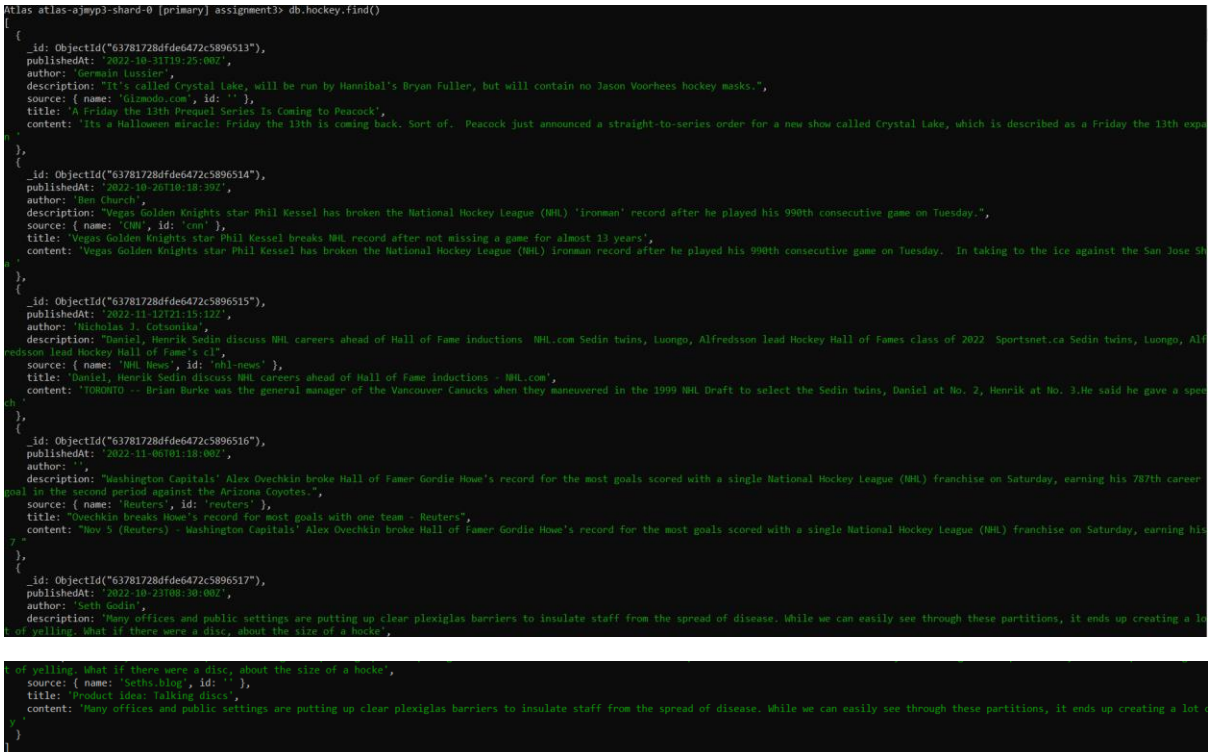


Collection ‘Halifax’ with articles as ‘Document’





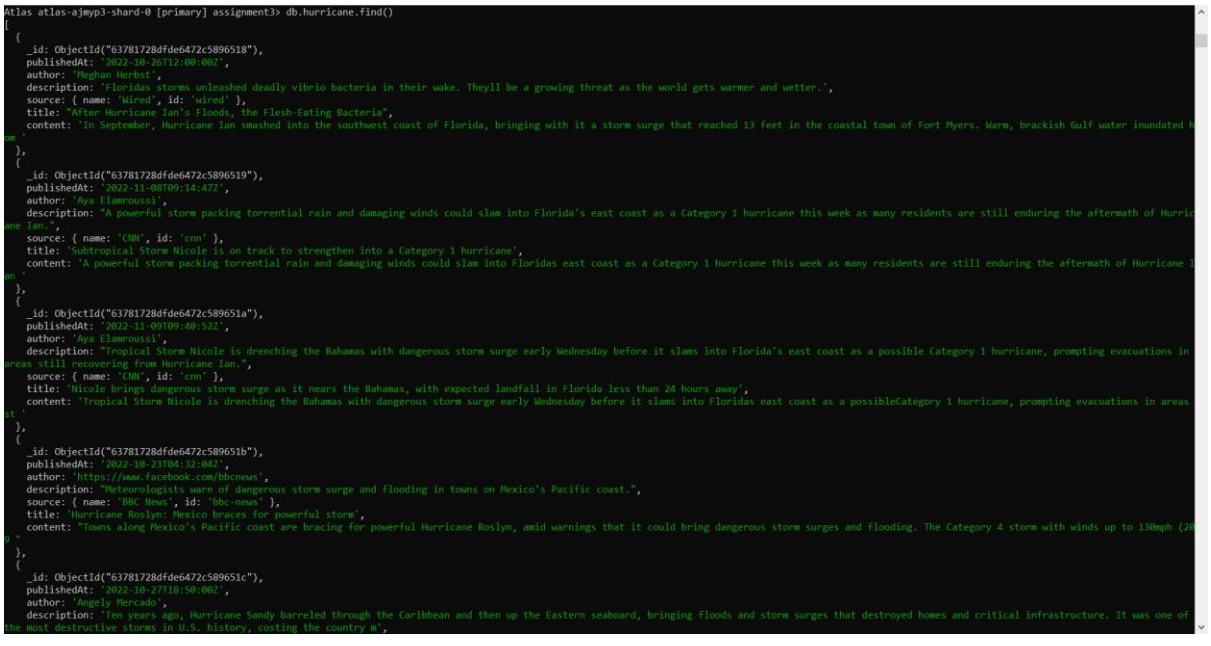
Collection ‘hockey’ with articles as ‘Document’



```
Atlas atlas-ajmyp3-shard-0 [primary] assignment3> db.hockey.find()
[
  {
    _id: ObjectId("63781728dfde6472c5896513"),
    publishedAt: '2022-10-31T19:25:00Z',
    author: 'Germain Lussier',
    description: "It's called Crystal Lake, will be run by Hannibal's Bryan Fuller, but will contain no Jason Voorhees hockey masks.",
    source: { name: 'Gizmodo.com', id: '' },
    title: 'A Friday the 13th Prequel Series Is Coming to Peacock',
    content: 'Its a Halloween miracle: Friday the 13th is coming back. Sort of.  Peacock just announced a straight-to-series order for a new show called Crystal Lake, which is described as a Friday the 13th expa
n '
  },
  {
    _id: ObjectId("63781728dfde6472c5896514"),
    publishedAt: '2022-10-26T10:18:39Z',
    author: 'Ben Church',
    description: 'Vegas Golden Knights star Phil Kessel has broken the National Hockey League (NHL) \'ironman\' record after he played his 990th consecutive game on Tuesday.',
    source: { name: 'CNN', id: 'cnn' },
    title: 'Vegas Golden Knights star Phil Kessel breaks NHL record after not missing a game for almost 13 years',
    content: 'Vegas Golden Knights star Phil Kessel has broken the National Hockey League (NHL) ironman record after he played his 990th consecutive game on Tuesday.  In taking to the ice against the San Jose Sh
a '
  },
  {
    _id: ObjectId("63781728dfde6472c5896515"),
    publishedAt: '2022-11-12T21:15:12Z',
    author: 'Nicholas J. Cotsonika',
    description: "Daniel, Henrik Sedin discuss NHL careers ahead of Hall of Fame inductions  NHL.com Sedin twins, Luongo, Alfredsson lead Hockey Hall of Fames class of 2022  Sportsnet.ca Sedin twins, Luongo, Alf
redsson lead Hockey Hall of Fame's cl",
    source: { name: 'NHL News', id: 'nhl-news' },
    title: 'Daniel, Henrik Sedin discuss NHL careers ahead of Hall of Fame inductions - NHL.com',
    content: 'TORONTO -- Brian Burke was the general manager of the Vancouver Canucks when they maneuvered in the 1999 NHL Draft to select the Sedin twins, Daniel at No. 2, Henrik at No. 3.He said he gave a spee
ch '
  },
  {
    _id: ObjectId("63781728dfde6472c5896516"),
    publishedAt: '2022-11-06T01:18:00Z',
    author: '',
    description: "Washington Capitals' Alex Ovechkin broke Hall of Famer Gordie Howe's record for the most goals scored with a single National Hockey League (NHL) franchise on Saturday, earning his 787th career
goal in the second period against the Arizona Coyotes.",
    source: { name: 'Reuters', id: 'reuters' },
    title: "Ovechkin breaks Howe's record for most goals with one team - Reuters",
    content: "Nov 5 (Reuters) - Washington Capitals' Alex Ovechkin broke Hall of Famer Gordie Howe's record for the most goals scored with a single National Hockey League (NHL) franchise on Saturday, earning his
7 "
  },
  {
    _id: ObjectId("63781728dfde6472c5896517"),
    publishedAt: '2022-10-23T08:30:00Z',
    author: 'Seth Godin',
    description: 'Many offices and public settings are putting up clear plexiglas barriers to insulate staff from the spread of disease. While we can easily see through these partitions, it ends up creating a lo
t of yelling. What if there were a disc, about the size of a hocke',
```

```
t of yelling. What if there were a disc, about the size of a hocke',
    source: { name: 'Seths.blog', id: '' },
    title: 'Product idea: Talking discs',
    content: 'Many offices and public settings are putting up clear plexiglas barriers to insulate staff from the spread of disease. While we can easily see through these partitions, it ends up creating a lot o
y '
  }
]
```

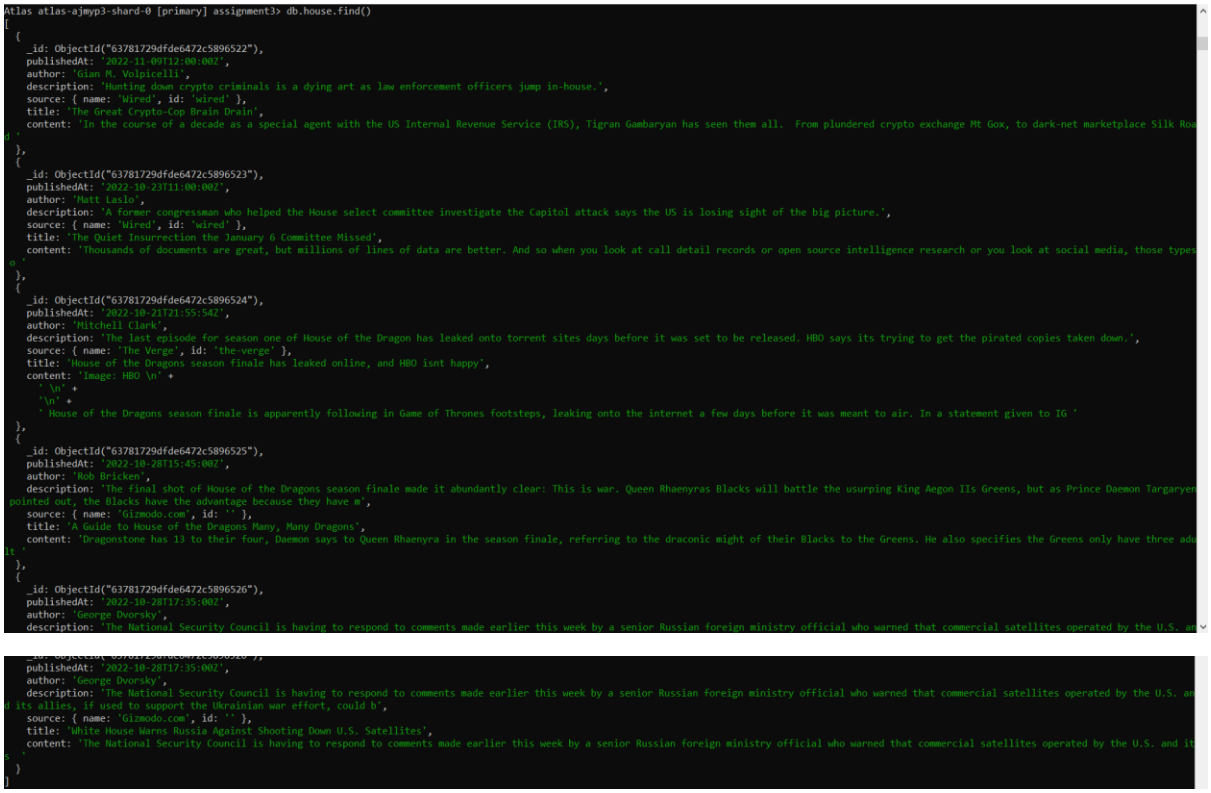
Collection ‘hurricane’ with articles as ‘Document’



```
Atlas atlas-ajmyp3-shard-0 [primary] assignment3> db.hurricane.find()
[
  {
    _id: ObjectId("63781728dfde6472c5896518"),
    publishedAt: '2022-10-26T12:00:00Z',
    author: 'Meghan Herbst',
    description: 'Floridas storms unleashed deadly vibrio bacteria in their wake. Theyll be a growing threat as the world gets warmer and wetter.',
    source: { name: 'Wired', id: 'wired' },
    title: "After Hurricane Ian's Floods, the Flesh-Eating Bacteria",
    content: 'In September, Hurricane Ian smashed into the southwest coast of Florida, bringing with it a storm surge that reached 13 feet in the coastal town of Fort Myers. Warm, brackish Gulf water inundated H
ms '
  },
  {
    _id: ObjectId("63781728dfde6472c5896519"),
    publishedAt: '2022-11-08T09:14:47Z',
    author: 'Aya Elamroussi',
    description: "A powerful storm packing torrential rain and damaging winds could slam into Florida's east coast as a Category 1 hurricane this week as many residents are still enduring the aftermath of Hurric
ane Ian.",
    source: { name: 'CNN', id: 'cnn' },
    title: 'Subtropical Storm Nicole is on track to strengthen into a Category 1 hurricane',
    content: 'A powerful storm packing torrential rain and damaging winds could slam into Floridas east coast as a Category 1 hurricane this week as many residents are still enduring the aftermath of Hurricane I
an '
  },
  {
    _id: ObjectId("63781728dfde6472c589651a"),
    publishedAt: '2022-11-09T09:40:52Z',
    author: 'Aya Elamroussi',
    description: "Tropical Storm Nicole is drenching the Bahamas with dangerous storm surge early Wednesday before it slams into Florida's east coast as a possible Category 1 hurricane, prompting evacuations in
areas still recovering from Hurricane Ian.",
    source: { name: 'CNN', id: 'cnn' },
    title: 'Nicole brings dangerous storm surge as it nears the Bahamas, with expected landfall in Florida less than 24 hours away',
    content: 'Tropical Storm Nicole is drenching the Bahamas with dangerous storm surge early Wednesday before it slams into Floridas east coast as a possibleCategory 1 hurricane, prompting evacuations in areas
st '
  },
  {
    _id: ObjectId("63781728dfde6472c589651b"),
    publishedAt: '2022-10-23T04:32:04Z',
    author: 'https://www.facebook.com/bbcnews',
    description: "Meteorologists warn of dangerous storm surge and flooding in towns on Mexico's Pacific coast.",
    source: { name: 'BBC News', id: 'bbc-news' },
    title: 'Hurricane Roslyn: Mexico braces for powerful storm',
    content: "Towns along Mexico's Pacific coast are bracing for powerful Hurricane Roslyn, amid warnings that it could bring dangerous storm surges and flooding. The Category 4 storm with winds up to 130mph (20
9 "
  },
  {
    _id: ObjectId("63781728dfde6472c589651c"),
    publishedAt: '2022-10-27T18:50:00Z',
    author: 'Angely Mercado',
    description: 'Ten years ago, Hurricane Sandy barreled through the Caribbean and then up the Eastern seaboard, bringing floods and storm surges that destroyed homes and critical infrastructure. It was one of
the most destructive storms in U.S. history, costing the country m',
```

```
  {
    _id: ObjectId("63781728dfde6472c589651c"),
    publishedAt: '2022-10-27T18:50:00Z',
    author: 'Angely Mercado',
    description: 'Ten years ago, Hurricane Sandy barreled through the Caribbean and then up the Eastern seaboard, bringing floods and storm surges that destroyed homes and critical infrastructure. It was one of
the most destructive storms in U.S. history, costing the country m',
    source: { name: 'Gizmodo.com', id: '' },
    title: 'Shocking Images From the Aftermath of Hurricane Sandy',
    content: 'Ten years ago, Hurricane Sandy barreled through the Caribbean and then up the Eastern seaboard, bringing floods and storm surges that destroyed homes and critical infrastructure. It was one of the
mo '
  }
]
```

Collection ‘house’ with articles as ‘Document’

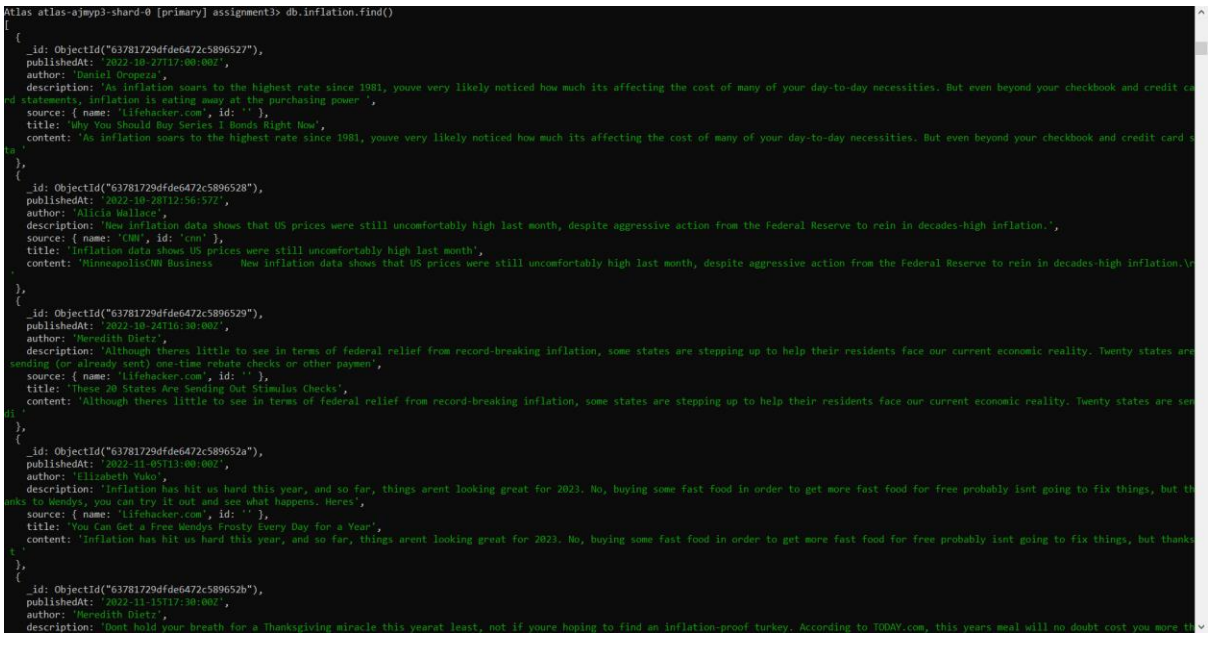


```
Atlas atlas-ajmyp3-shard-0 [primary] assignment3> db.house.find()
[
  {
    _id: ObjectId("63781729dfde6472c5896522"),
    publishedAt: '2022-11-09T12:00:00Z',
    author: 'Gian M. Volpicelli',
    description: 'Hunting down crypto criminals is a dying art as law enforcement officers jump in-house.',
    source: { name: 'Wired', id: 'wired' },
    title: 'The Great Crypto-Cop Brain Drain',
    content: 'In the course of a decade as a special agent with the US Internal Revenue Service (IRS), Tigran Gambaryan has seen them all.  From plundered crypto exchange Mt Gox, to dark-net marketplace Silk Roa
d '
  },
  {
    _id: ObjectId("63781729dfde6472c5896523"),
    publishedAt: '2022-10-23T11:00:00Z',
    author: 'Matt Laslo',
    description: 'A former congressman who helped the House select committee investigate the Capitol attack says the US is losing sight of the big picture.',
    source: { name: 'Wired', id: 'wired' },
    title: 'The Quiet Insurrection the January 6 Committee Missed',
    content: 'Thousands of documents are great, but millions of lines of data are better. And so when you look at call detail records or open source intelligence research or you look at social media, those types
o '
  },
  {
    _id: ObjectId("63781729dfde6472c5896524"),
    publishedAt: '2022-10-21T21:55:54Z',
    author: 'Mitchell Clark',
    description: 'The last episode for season one of House of the Dragon has leaked onto torrent sites days before it was set to be released. HBO says its trying to get the pirated copies taken down.',
    source: { name: 'The Verge', id: 'the-verge' },
    title: 'House of the Dragons season finale has leaked online, and HBO isnt happy',
    content: 'Image: HBO \n' +
      ' \n' +
      '\n' +
      ' House of the Dragons season finale is apparently following in Game of Thrones footsteps, leaking onto the internet a few days before it was meant to air. In a statement given to IG '
  },
  {
    _id: ObjectId("63781729dfde6472c5896525"),
    publishedAt: '2022-10-28T15:45:00Z',
    author: 'Rob Bricken',
    description: 'The final shot of House of the Dragons season finale made it abundantly clear: This is war. Queen Rhaenyras Blacks will battle the usurping King Aegon IIs Greens, but as Prince Daemon Targaryen
pointed out, the Blacks have the advantage because they have m',
    source: { name: 'Gizmodo.com', id: '' },
    title: 'A Guide to House of the Dragons Many, Many Dragons',
    content: 'Dragonstone has 13 to their four, Daemon says to Queen Rhaenyra in the season finale, referring to the draconic might of their Blacks to the Greens. He also specifies the Greens only have three adu
lt '
  },
  {
    _id: ObjectId("63781729dfde6472c5896526"),
    publishedAt: '2022-10-28T17:35:00Z',
    author: 'George Dvorsky',
    description: 'The National Security Council is having to respond to comments made earlier this week by a senior Russian foreign ministry official who warned that commercial satellites operated by the U.S. an
```

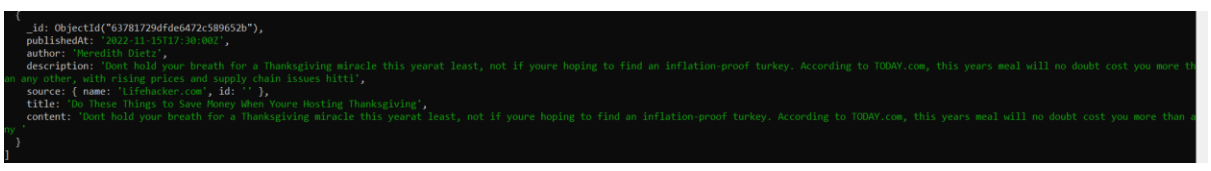


```
    publishedAt: '2022-10-28T17:35:00Z',
    author: 'George Dvorsky',
    description: 'The National Security Council is having to respond to comments made earlier this week by a senior Russian foreign ministry official who warned that commercial satellites operated by the U.S. an
d its allies, if used to support the Ukrainian war effort, could b',
    source: { name: 'Gizmodo.com', id: '' },
    title: 'White House Warns Russia Against Shooting Down U.S. Satellites',
    content: 'The National Security Council is having to respond to comments made earlier this week by a senior Russian foreign ministry official who warned that commercial satellites operated by the U.S. and it
s '
  }
]
```

Collection ‘inflation’ with articles as ‘Document’



```
Atlas atlas-ajmyp3-shard-0 [primary] assignment3> db.inflation.find()
[
  {
    _id: ObjectId("63781729dfde6472c5896527"),
    publishedAt: '2022-10-27T17:00:00Z',
    author: 'Daniel Oropeza',
    description: 'As inflation soars to the highest rate since 1981, youve very likely noticed how much its affecting the cost of many of your day-to-day necessities. But even beyond your checkbook and credit ca
rd statements, inflation is eating away at the purchasing power ',
    source: { name: 'Lifehacker.com', id: '' },
    title: 'Why You Should Buy Series I Bonds Right Now',
    content: 'As inflation soars to the highest rate since 1981, youve very likely noticed how much its affecting the cost of many of your day-to-day necessities. But even beyond your checkbook and credit card s
ta '
  },
  {
    _id: ObjectId("63781729dfde6472c5896528"),
    publishedAt: '2022-10-28T12:56:57Z',
    author: 'Alicia Wallace',
    description: 'New inflation data shows that US prices were still uncomfortably high last month, despite aggressive action from the Federal Reserve to rein in decades-high inflation.',
    source: { name: 'CNN', id: 'cnn' },
    title: 'Inflation data shows US prices were still uncomfortably high last month',
    content: 'MinneapolisCNN Business     New inflation data shows that US prices were still uncomfortably high last month, despite aggressive action from the Federal Reserve to rein in decades-high inflation.\r
'
  },
  {
    _id: ObjectId("63781729dfde6472c5896529"),
    publishedAt: '2022-10-24T16:30:00Z',
    author: 'Meredith Dietz',
    description: 'Although theres little to see in terms of federal relief from record-breaking inflation, some states are stepping up to help their residents face our current economic reality. Twenty states are
 sending (or already sent) one-time rebate checks or other paymen',
    source: { name: 'Lifehacker.com', id: '' },
    title: 'These 20 States Are Sending Out Stimulus Checks',
    content: 'Although theres little to see in terms of federal relief from record-breaking inflation, some states are stepping up to help their residents face our current economic reality. Twenty states are sen
di '
  },
  {
    _id: ObjectId("63781729dfde6472c589652a"),
    publishedAt: '2022-11-05T13:00:00Z',
    author: 'Elizabeth Yuko',
    description: 'Inflation has hit us hard this year, and so far, things arent looking great for 2023. No, buying some fast food in order to get more fast food for free probably isnt going to fix things, but th
anks to Wendy’s, you can try it out and see what happens. Heres',
    source: { name: 'Lifehacker.com', id: '' },
    title: 'You Can Get a Free Wendys Frosty Every Day for a Year',
    content: 'Inflation has hit us hard this year, and so far, things arent looking great for 2023. No, buying some fast food in order to get more fast food for free probably isnt going to fix things, but thanks
t '
  },
  {
    _id: ObjectId("63781729dfde6472c589652b"),
    publishedAt: '2022-11-15T17:30:00Z',
    author: 'Meredith Dietz',
    description: 'Dont hold your breath for a Thanksgiving miracle this yearat least, not if youre hoping to find an inflation-proof turkey. According to TODAY.com, this years meal will no doubt cost you more th
```

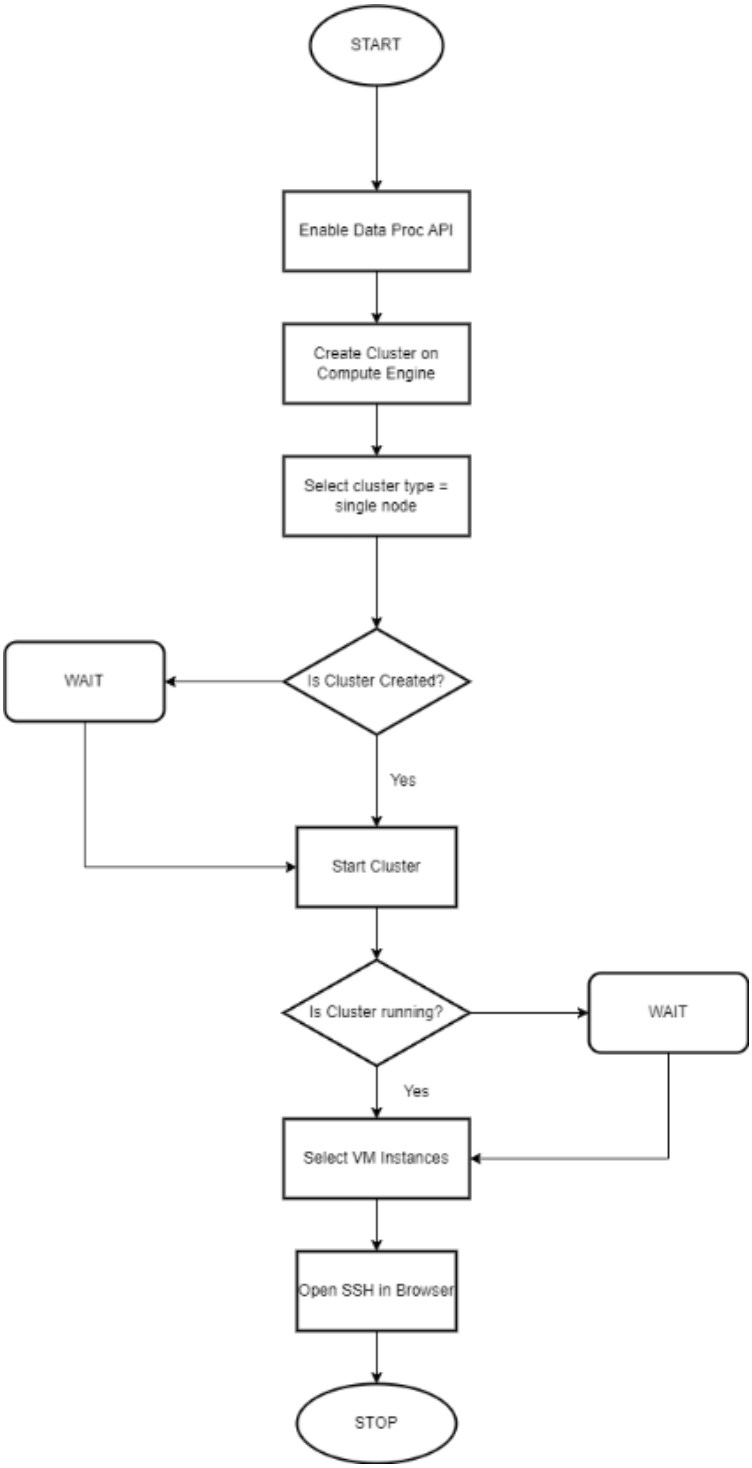


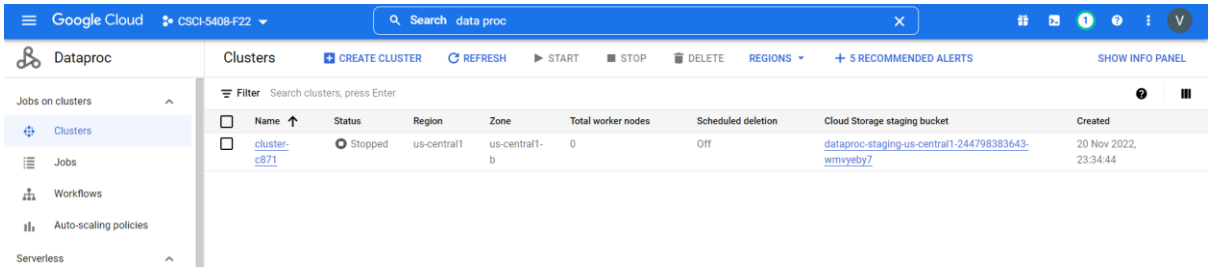
```
    _id: ObjectId("63781729dfde6472c589652b"),
    publishedAt: '2022-11-15T17:30:00Z',
    author: 'Meredith Dietz',
    description: 'Dont hold your breath for a Thanksgiving miracle this yearat least, not if youre hoping to find an inflation-proof turkey. According to TODAY.com, this years meal will no doubt cost you more th
an any other, with rising prices and supply chain issues hitti',
    source: { name: 'Lifehacker.com', id: '' },
    title: 'Do These Things to Save Money When Youre Hosting Thanksgiving',
    content: 'Dont hold your breath for a Thanksgiving miracle this yearat least, not if youre hoping to find an inflation-proof turkey. According to TODAY.com, this years meal will no doubt cost you more than a
ny '
  }
]
```

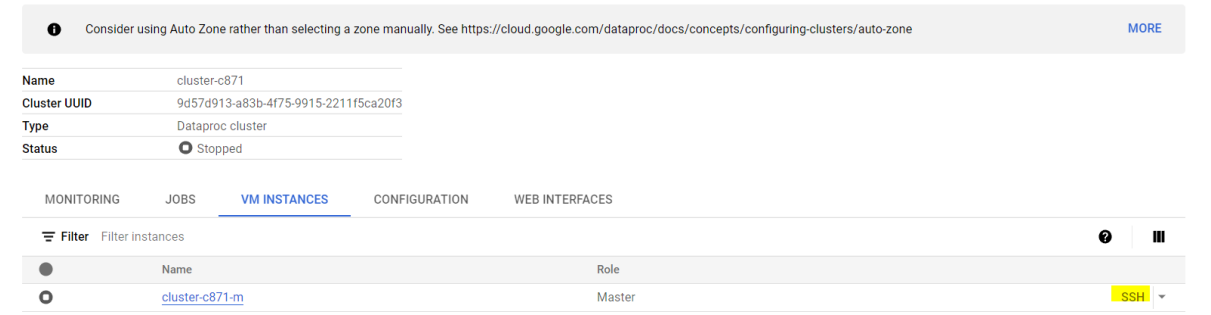
Data Processing and Popularity Detection

Created a cluster on Google Cloud Platform.

Flowchart for creating cluster

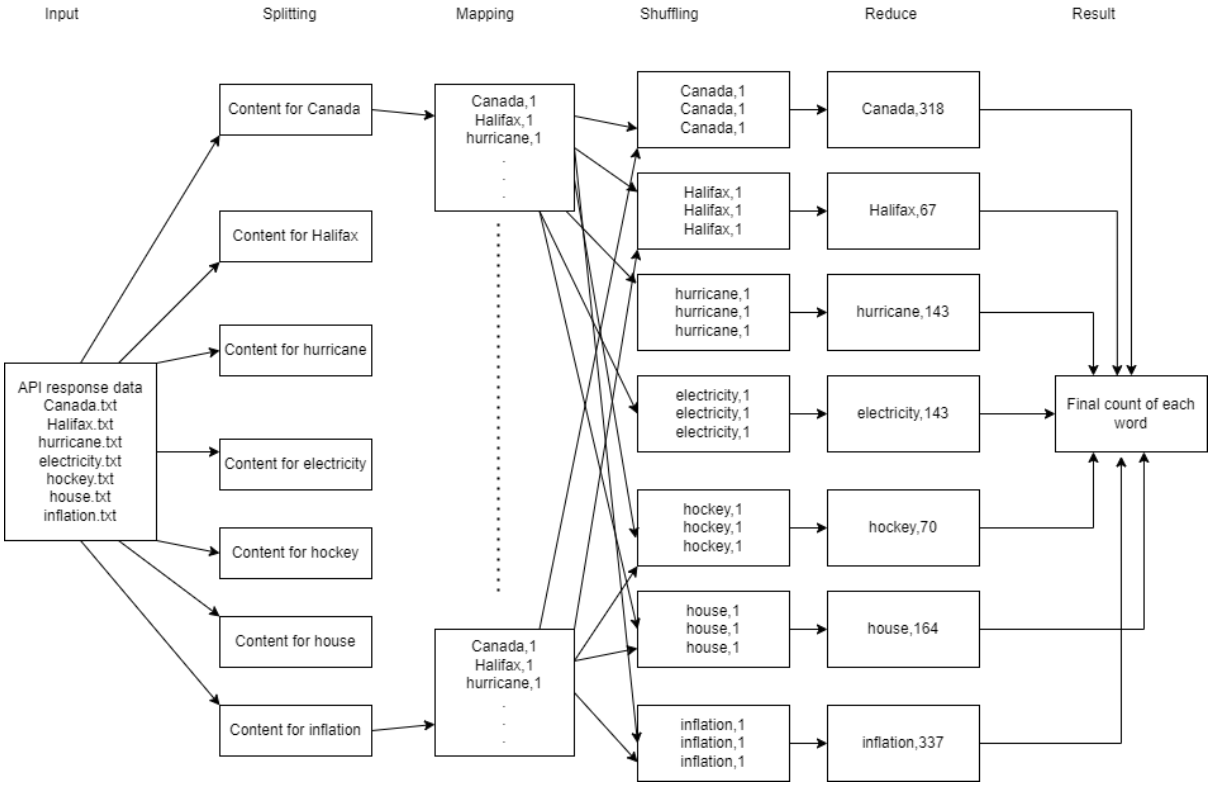






MapReduce program in JAVA to count words

MapReduce logic for word count



Algorithm for MapReduce framework used to count frequency of the following words: - “Canada”, “Halifax”, “hockey”, “hurricane”, “electricity”, “house”, “inflation”.

Input – Canada.txt, Halifax.txt, hockey.txt, hurricane.txt, electricity.txt, house.txt, inflation.txt

Output – word count of each keyword.

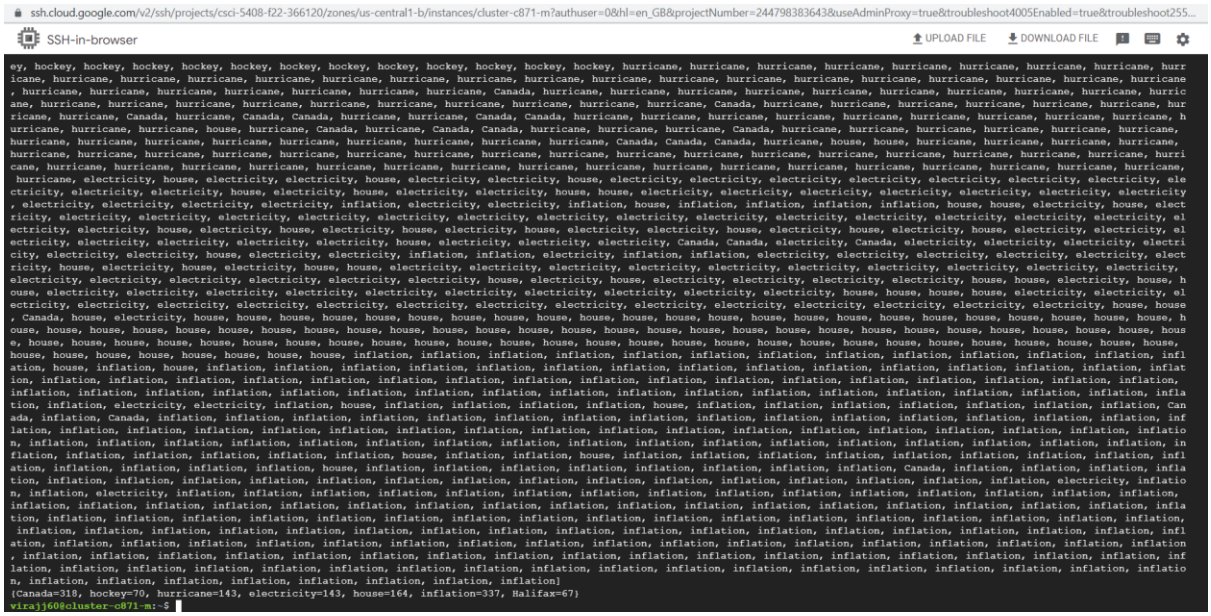
```
Mapper.class
//splitting phase
For each file {
        BufferedReader buffer = read(filename.txt);
        stringBuffer = buffer.readLine();
        array[] array = stringBuffer.toString().split(“ ”);

        //mapping phase
        for each arrayElement {
        for each keyword {
        if (arrayElement contains keyword){
                //list is ordered. So shuffling phase is implicit
                list.add(keyword);
                                                }
                        }
         }
}

Reducer.class
//reduce phase
for each keyword{
        counter=0;
        for each array element {
          if(array element equals keyword){
                counter ++;
        }
        }
        //result phase
        reducerMap.add(keyword,counter);
}
```

Uploaded the java code .jar file to the spark cluster created on GCP.

Executed using **spark-submit –class package.className jarName.jar** command



Final word count -

Canada = 318

hockey=70

hurricane=143

electricity=133

house=164

Halifax=67

inflation=337

Highest frequency is for keyword ‘inflation’ and lowest for keyword ‘Halifax’

PART 2

Neo4j – Summary

Unlike other databases like relational, key-valued or document database, Neo4j has a native graph approach at the core. Data is saved in nodes and each node is connected to other nodes through relationships. Thus, information about the next node in the sequence is available in the node itself making traversal and operations on database much faster than other databases.

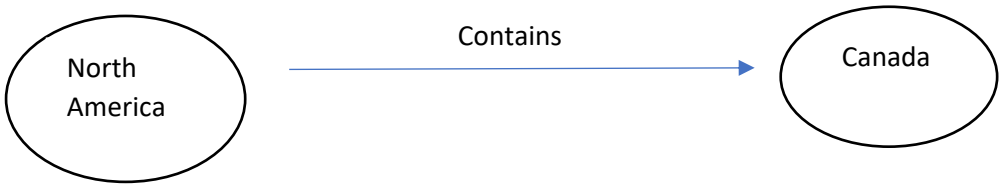
Neo4j organizes the information as nodes, relationships, and properties.

- Node - Nodes are entities in graph like tables in relational databases. Nodes can hold multiple properties and a label can be tagged with labels like table name in relational database.

- Relationships – It represents the connection between two nodes. It gives a real-world/natural connection between the related nodes. Relationships are directed i.e. they have a start node and an end node. Even though the relationships are directed, they can be efficiently traversed in any direction. Relationships have properties just like node.

Cypher! Cypher is Neo4js graph query language like SQL for relational databases. It is a declarative language with syntax that provides a logical and visual way to match patterns of node and relationships between them.

Cypher to create nodes and establish a relationship between them –



CREATE (northAmerica: Continent {name='North America'}) – [rel:Contains] -> (canada: Country {name='Canada'})

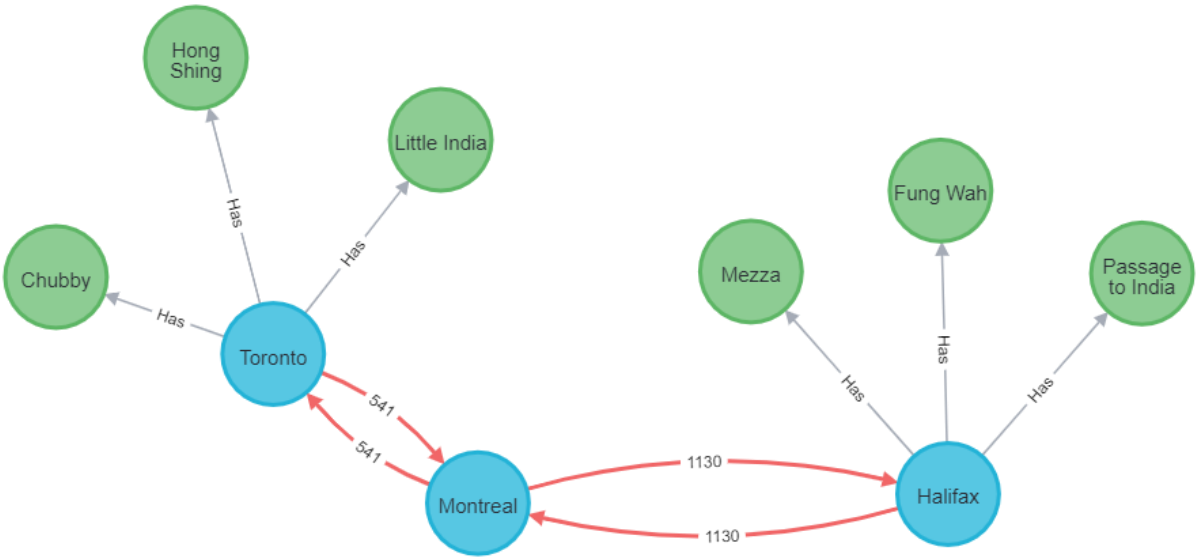
Here, 'northAmerica' is the variable Name, Continent is the node label and name is the node property. - -> depicts the relationship between both nodes. 'rel' is the relationship variable and 'Contains' is the relationship type.

Neo4j provides what is expected of a database system – ACID transactions, runtime failovers and cluster support. It is stable and flexible hence used by various enterprises in production environments.

I would use Neo4j for any future applications which requires a tremendous amount of data to be traversed or processed as Neo4j will provide faster information retrieval and processing. For e.g. – If I intend to build an algorithm to recommend restaurants in a city based on the type of cuisine a user prefers, I will store all the data in a graphical structure using Neo4j with cities, restaurant names as nodes, cuisine offered as properties for restaurants and relationships between cities – cities and cities – restaurant. Then using Cypher, I would fetch restaurants that offer the type of cuisine a user prefers.

Created the following graph using Neo4j.

```
create (halifax:City {name:'Halifax'})
create (montreal:City {name:'Montreal'})
create (toronto:City {name:'Toronto'})
create (halifax) -[:Neighbour {distance:1130}] -> (montreal)
create (montreal) -[:Neighbour {distance:1130}] -> (halifax)
create (montreal) - [:Neighbour {distance:541}] -> (toronto)
create (toronto) - [:Neighbour {distance:541}] -> (montreal)
create (passage:Restaurant {name:'Passage to India',cuisine:'Indian'})
create (fung:Restaurant {name:'Fung Wah',cuisine:'Chinese'})
create (mezza:Restaurant {name:'Mezza',cuisine:'Lebanese'})
create (litInd:Restaurant {name:'Little India',cuisine:'Indian'})
create (hong:Restaurant {name:'Hong Shing',cuisine:'Chinese'})
create (chub:Restaurant {name:'Chubby',cuisine:'Caribbean'})
create (halifax) -[:Has] -> (passage)
create (halifax) -[:Has] -> (fung)
create (halifax) -[:Has] -> (mezza)
create (toronto) -[:Has] -> (litInd)
create (toronto) -[:Has] -> (hong)
create (toronto) -[:Has] -> (chub)
```



In the above graph ‘Toronto’ , ‘Montreal’ and ‘Halifax’ are nodes labels with toronto, montreal and Halifax as the node variables. Node labels Toronto, Montreal and Halifax are connected through a relationship with relationship type = ‘Neighbour’ and ‘distance’ between them as relationship property. Nodes Toronto and Halifax are connected through a relation ‘Has’ to node labels ‘Little India’, ’Chubby’, ’Hong Shing’ and ‘Mezza’, ’Passage to India’ and ‘Fung Wah’ respectively. These nodes have ‘name’ and ‘cuisine’ as node properties.

REFERENCES

[1] “News API – search news and blog articles on the web,” News API Search News and Blog Articles on the Web. [Online]. Available: https://newsapi.org/. [Accessed: 22-Nov-2022].

[2] “MongoDB cloud,” MongoDB. [Online]. Available: https://cloud.mongodb.com/. [Accessed: 22-Nov-2022].

[3] “Dataproc,” Google cloud platform. [Online]. Available: https://console.cloud.google.com/. [Accessed: 22-Nov-2022].

[4] “Getting started guide for neo4j version 5,” Neo4j Graph Data Platform. [Online]. Available: https://neo4j.com/docs/getting-started/current/. [Accessed: 22-Nov-2022].

[5] S. Roussy, “Neo4j: A graph project story.” .