

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 UriToImage links, replace all null values, carriage return and new line (/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.

```
mongosh mongodb+srv://<credentials>@cluster0.vjocpgf.mongodb.net/myFirstDatabase
Atlas atlas-ajmvp3-shard-0 [primary] assignment3> show connections;
MongoshInvalidInputError: [COMMON-10001] 'connections' is not a valid argument for "show".
Atlas atlas-ajmvp3-shard-0 [primary] assignment3> show collections
Canada
electricity
Halifax
hockey
house
hurricane
inflation
Atlas atlas-ajmvp3-shard-0 [primary] assignment3>
```

Collection 'Canada' with articles as 'Document'

```
Atlas atlas-ajmvp3-shard-0 [primary] assignment3> db.canada.find().pretty()
{
  "_id": ObjectId("63781724dfe6472c5896509"),
  "publishedAt": "2022-11-10T14:00:15Z",
  "author": "Tom Warren",
  "description": "Apple is partnering with Globalstar to enable its new satellite emergency SOS feature. iPhone 14 and 14 Pro owners in the US and Canada will be able to start using the feature later this month.",
  "source": { "name": "The Verge", "id": "the-verge" },
  "title": "iPhone 14s Emergency SOS via satellite feature is coming later this month",
  "content": "iPhone 14s Emergency SOS via satellite feature is coming later this month iPhone 14s Emergency SOS via satellite feature is coming later this month / Apple confirms its using a $450 million fund t",
},
{
  "_id": ObjectId("63781724dfe6472c589650a"),
  "publishedAt": "2022-10-19T15:25:00Z",
  "author": "",
  "description": "Canadian Finance Minister Chrystia Freeland on Wednesday said an economic slowdown was coming for the world and that Canada has the fiscal capacity to get through the "challenging days" ahead.",
  "source": { "name": "Reuters", "id": "reuters" },
  "title": "Economic slowdown coming for Canada, world -Canada finance minister - Reuters Canada",
  "content": "Oct 19 (Reuters) - Canadian Finance Minister Chrystia Freeland on Wednesday said an economic slowdown was coming for the world and that Canada has the fiscal capacity to get through the "challengin",
},
{
  "_id": ObjectId("63781724dfe6472c589650b"),
  "publishedAt": "2022-10-31T00:19:31Z",
  "author": "",
  "description": "Canada could match England's forwards dominance in Saturday's World Cup semi-final, says former Red Roses captain Katy Daley-McLean.",
  "source": { "name": "BBC News", "id": "bbc-news" },
  "title": "Rugby World Cup: Canada could match England up front - Katy Daley-McLean",
  "content": "Flanker Hurlie Packer scored a hat-trick in England's quarter-final win against Australia Rugby World Cup semi-final: Canada v England Venue: Eden Park, Auckland Da ",
},
{
  "_id": ObjectId("63781724dfe6472c589650c"),
  "publishedAt": "2022-11-04T14:30:00Z",
  "author": "",
  "description": "The United States and Canada said on Friday they were imposing sanctions on two Haitian politicians, including the president of the country's Senate, over what Canada described as their operati",
  "source": { "name": "Reuters", "id": "reuters" },
  "title": "U.S., Canada impose sanctions on two Haitian politicians - Reuters Canada",
  "content": "Nov 4 (Reuters) - The United States and Canada said on Friday they were imposing sanctions on two Haitian politicians, including the president of the country's Senate, over what Canada described as",
},
{
  "_id": ObjectId("63781724dfe6472c589650d"),
  "publishedAt": "2022-11-13T19:58:00Z",
  "author": "",
  "description": "Canada named their 26-man squad for the Nov. 20-Dec. 18 World Cup on Sunday. Canada are in Group F alongside Belgium, Morocco and Croatia.",
},
{
  "_id": ObjectId("63781724dfe6472c589650e"),
  "publishedAt": "2022-11-13T19:58:00Z",
  "author": "",
  "description": "Canada named their 26-man squad for the Nov. 20-Dec. 18 World Cup on Sunday. Canada are in Group F alongside Belgium, Morocco and Croatia.",
  "source": { "name": "Reuters", "id": "reuters" },
  "title": "Canada world Cup squad - Reuters",
  "content": "Nov 13 (Reuters) - Canada named their 26-man squad for the Nov. 20-Dec. 18 World Cup on Sunday. Canada are in Group F alongside Belgium, Morocco and Croatia. Squad: Goalkeepers: James Pantemis, Mil",
}
```

Collection 'electricity' with articles as 'Document'

```
Atlas atlas-ajmvp3-shard-0 [primary] assignment>> db.electricity.find()
{
  "_id": ObjectId("63781728dfde6472c589651d"),
  "publishedAt": "2022-10-31T20:00:00Z",
  "author": "Lindsey Ellerson",
  "description": "Saving energy is always important, but with prices rising, electricity and heating bills are an especially hot topic right now, so to speak. There are plenty of ways to cut down on energy use, but you have to identify the main energy-sucking offenders lurking in your home.",
  "source": { "name": "Lifehacker.com", "id": "" },
  "title": "10 Common Household Items That Are Using Too Much Energy",
  "content": "If you still have a set-top DVR or a cable box, its probably using more energy than you think. Per How-to Geek, DVR boxes could use 25 wats or more and cable boxes can use 15. Televisions, too, are energy hogs.",
},
{
  "_id": ObjectId("63781728dfde6472c589651e"),
  "publishedAt": "2022-11-01T16:56:00Z",
  "author": "Reuters Fact Check",
  "description": "Thousands of people have watched a man misreading an electricity meter in a social media video where he falsely claims that it reveals asylum seekers in Britain are given free credit for household electricity costs.",
  "source": { "name": "Reuters", "id": "reuters" },
  "title": "Fact Check: Free credit message on electricity meters means free credit, not free credit for asylum seekers - Reuters",
  "content": "Thousands of people have watched a man misreading an electricity meter in a social media video where he falsely claims that it reveals asylum seekers in Britain are given free credit for household electricity costs.",
},
{
  "_id": ObjectId("63781728dfde6472c589651f"),
  "publishedAt": "2022-11-04T13:15:00Z",
  "author": "Reuters",
  "description": "Bosnia, the Balkans' sole exporter of electricity, may be forced to cut exports as Bosnians turn to subsidised electricity for heating after a jump in prices for other fuels, including gas, firewood and wood pellets.",
  "source": { "name": "Reuters", "id": "reuters" },
  "title": "Bosnia's power exports at risk as people switch to electricity for heating - Reuters",
  "content": "SARAJEVO, Nov 4 (Reuters) - Bosnia, the Balkans' sole exporter of electricity, may be forced to cut exports as Bosnians turn to subsidised electricity for heating after a jump in prices for other fuels.",
},
{
  "_id": ObjectId("63781728dfde6472c5896520"),
  "publishedAt": "2022-11-01T10:29:00Z",
  "author": "Reuters",
  "description": "Japan said on Tuesday it will ask households and companies across the country to conserve electricity within "a reasonable range" during the peak winter demand season to alleviate a possible power crunch in the world's third largest economy.",
  "source": { "name": "Reuters", "id": "reuters" },
  "title": "Japan asks households, companies to conserve electricity during winter - Reuters",
  "content": "TOKYO, Nov 1 (Reuters) - Japan said on Tuesday it will ask households and companies across the country to conserve electricity within "a reasonable range" during the peak winter demand season to alleviate a possible power crunch in the world's third largest economy.",
},
{
  "_id": ObjectId("63781729dfde6472c5896521"),
  "publishedAt": "2022-10-31T23:00:00Z",
  "author": "Reuters",
  "description": "A 40% cut in deliveries of Russian natural gas is hitting Moldova's ability to provide sufficient electricity for its 2.5 million people, the deputy prime minister of the small ex-Soviet state said on Monday.",
  "source": { "name": "Reuters", "id": "reuters" },
  "title": "Moldova electricity supplies hit by cut in Russian gas - Reuters",
  "content": "CHISINAU, Oct 31 (Reuters) - A 40% cut in deliveries of Russian natural gas is hitting Moldova's ability to provide sufficient electricity for its 2.5 million people, the deputy prime minister of the small ex-Soviet state said on Monday.",
}
```

Collection 'Halifax' with articles as 'Document'

```
Atlas atlas-ajmvp3-shard-0 [primary] assignment>> db.Halifax.find()
{
  "_id": ObjectId("63781727dfde6472c589650e"),
  "publishedAt": "2022-11-07T07:01:15Z",
  "author": "Reuters",
  "description": "British house prices fell in October at the fastest monthly rate since February 2021, in a fresh sign of a weaker housing market, mortgage lender Halifax said on Monday.",
  "source": { "name": "Reuters", "id": "reuters" },
  "title": "UK house prices fall at fastest rate since early 2021 - Halifax - Reuters UK",
  "content": "LONDON, Nov 7 (Reuters) - British house prices fell in October at the fastest monthly rate since February 2021, in a fresh sign of a weaker housing market, mortgage lender Halifax said on Monday.",
},
{
  "_id": ObjectId("63781727dfde6472c589650f"),
  "publishedAt": "2022-11-18T13:34:00Z",
  "author": "Reuters",
  "description": "Canadian home prices dropped in October from the previous month, though less sharply than in September, while year-over-year price gains continued to slow, TeranetNational Bank National Composite House Price data showed on Friday.",
  "source": { "name": "Reuters", "id": "reuters" },
  "title": "Canada home price index slides again in October - Teranet - Reuters.com",
  "content": "OTTAWA, Nov 18 (Reuters) - Canadian home prices dropped in October from the previous month, though less sharply than in September, while year-over-year price gains continued to slow, TeranetNational Bank National Composite House Price data showed on Friday.",
},
{
  "_id": ObjectId("63781727dfde6472c5896510"),
  "publishedAt": "2022-11-07T07:31:56Z",
  "author": "Graeme Meardon",
  "description": "House prices drop at fastest monthly rate since February 2021, reports mortgage costs hit sectorPay restraint for thee, but not for me.Thats the mantra in UK boardrooms, it seems, after top bosses enjoyed another year of strong earnings.",
  "source": { "name": "The Guardian", "id": "" },
  "title": "UK house prices fall after mini-budget shock, but CEO pay soars - business live",
  "content": "First-time buyers drive house price slowdown Property price inflation weakened across all buyer types during October, led by first-time buyers. People taking their first step on the housing ladder.",
},
{
  "_id": ObjectId("63781727dfde6472c5896511"),
  "publishedAt": "2022-11-10T17:14:12Z",
  "author": "Hollie Richardson",
  "description": "Why did the superhero franchise descend on the Dales? What made Tom Cruise film Mission: Impossible nearby, and Shane Meadows follow suit? We report from new screen star hotspot Halifax - now being renamed HalliwoodAt the start of this year, it wouldnt have been.",
  "source": { "name": "The Guardian", "id": "" },
  "title": "Wow, Samuel L Jackson is here! How Yorkshire became Marvels go-to location",
  "content": "At the start of this year, it wouldnt have been unusual for the people of Halifax to bump into a Hollywood star on their way to work or while picking up a takeaway. You just kind of go, Wow, Samuel L Jackson is here!",
},
{
  "_id": ObjectId("63781727dfde6472c5896512"),
  "publishedAt": "2022-11-11T10:05:00Z",
  "author": "Allison King",
  "description": "Hundreds Gather for Remembrance Day Ceremony in St. John's - VOON Remembrance Day ceremony draws hundreds to Grand Parade in Halifax CBC.ca Photos: Nov. 11 ceremony attracts hundreds to Memorial Park Sudbury.com Orillia",
  "source": { "name": "VOON", "id": "" },
  "title": "Hundreds Gather for Remembrance Day Ceremony in St. John's - VOON",
  "content": "The annual Remembrance Day ceremony in downtown St. Johns returned to its full program this year after two years of COVID-19 restrictions. A cold and damp morning in the capital city did not stop the ceremony.",
}
```

Collection 'hockey' with articles as 'Document'

```

class atlas-ajpy3-shard-0 [primary] assignment3> db.hockey.find(
{
  _id: ObjectId("63781728dfde6472c5896513"),
  publishedAt: "2022-10-31T19:25:00Z",
  author: "Gowain Lusier",
  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: "It's 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 ex...
"},
{
  _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: "CNM", id: "cnm" },
  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...
"},
{
  _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 Fame class of 2022 Sportsnet.ca Sedin twins, Luongo, Alfr...
  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...
"},
{
  _id: ObjectId("63781728dfde6472c5896516"),
  publishedAt: "2022-11-06T01:18:00Z",
  author: "Reuters",
  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...
  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...
"},
{
  _id: ObjectId("63781728dfde6472c5896517"),
  publishedAt: "2022-10-23T08:30:00Z",
  author: "Seth Golin",
  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 lot...
  source: { name: "Seth Golin", 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...
"}
}

```

Collection 'hurricane' with articles as 'Document'

```

Atlas atlas-ajmvp3-shard-0 [primary] assignment> db.hurricane.find()
{
  _id: ObjectId("63781728dfde6472c5896518"),
  publishedAt: "2022-10-26T12:00:00Z",
  author: 'Peghan Herbat',
  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
',
},
{
  _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
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
',
},
{
  _id: ObjectId("63781728dfde6472c589651c"),
  publishedAt: "2022-10-23T04:32:00Z",
  author: 'https://www.facebook.com/bhcnnews',
  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
)',
},
{
  _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 a',
},
{
  _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 a',
  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
most d
',
},

```


Collection 'house' with articles as 'Document'

```

Atlas atlas-ajmp3-shard-0 [query] assignment> db.house.find(
{
  _id: ObjectId("63781729dfde6472c5896522"),
  publishedAt: '2022-10-09T12:00:00Z',
  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-God Brain Drain',
  content: 'In the course of a decade as a special agent with the US Internal Revenue Service (IRS), Tigran Gombaryan has seen them all. From plundered crypto exchange Mt Gox, to dark-net marketplace Silk Roa
},
{
  _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
},
{
  _id: ObjectId("63781729dfde6472c5896524"),
  publishedAt: '2022-10-23T21:55:54Z',
  author: 'Rutledge 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 isn't happy',
  content: 'House: 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 16'
},
{
  _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 Rhaenyra Blacks will battle the usurping King Aegon II's Greens, but as Prince Daemon Targaryen pointed out, the Blacks have the advantage because they have a',
  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 ad
},
{
  _id: ObjectId("63781729dfde6472c5896526"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. in
},
{
  _id: ObjectId("63781729dfde6472c5896527"),
  publishedAt: '2022-10-20T17:35:00Z',
  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
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
},
{
  _id: ObjectId("63781729dfde6472c5896528"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896529"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896530"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896531"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896532"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896533"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896534"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896535"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896536"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896537"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896538"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896539"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896540"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896541"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896542"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896543"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896544"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896545"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896546"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896547"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896548"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896549"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896550"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896551"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896552"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896553"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it
},
{
  _id: ObjectId("63781729dfde6472c5896554"),
  publishedAt: '2022-10-20T17:35:00Z',
  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. and it

```

Collection 'inflation' with articles as 'Document'

```

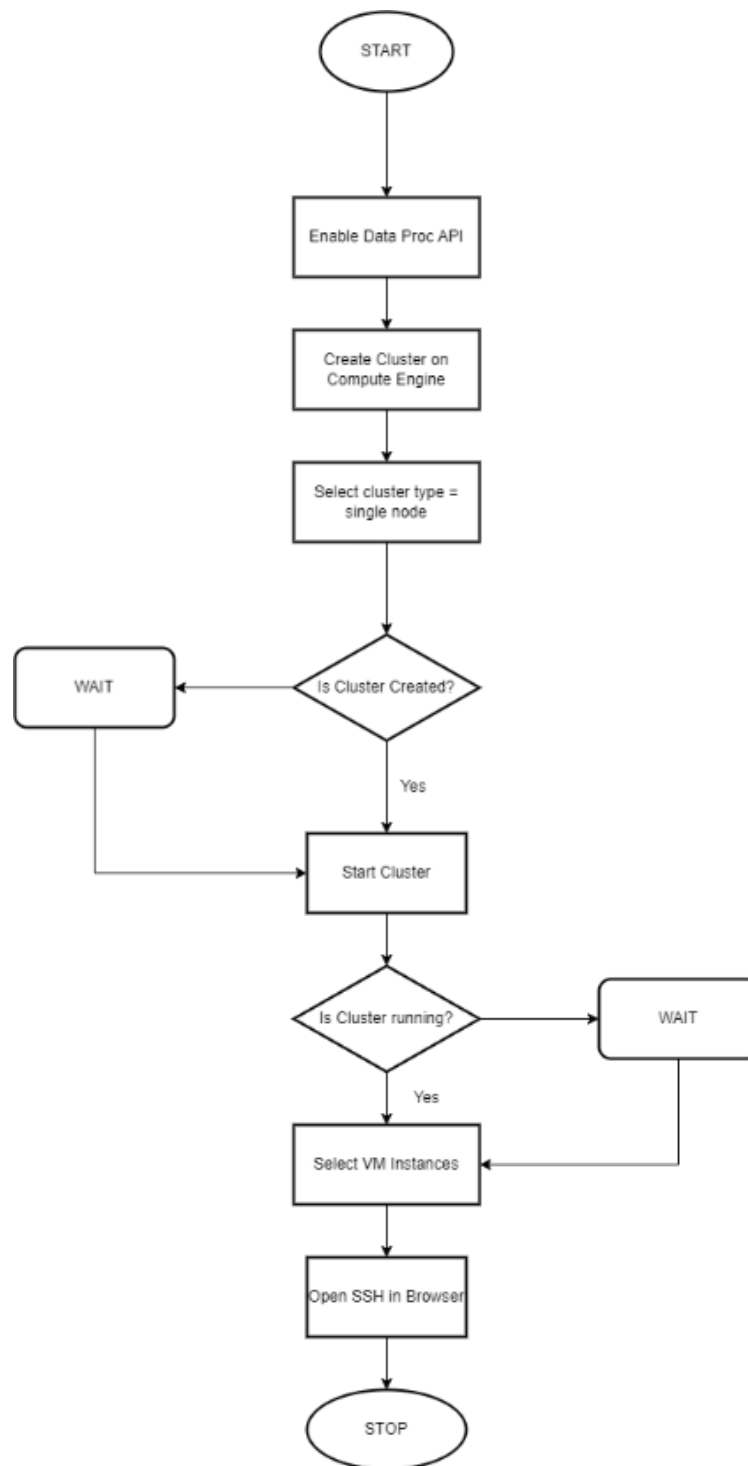
[{"_id": "63781729dfde6472c5896527",
  publishedAt: "2022-10-27 17:00:00Z",
  author: "Daniel Oprea",
  description: "As inflation soars to the highest rate since 1981, youve very likely noticed how much its affecting the cost of many of your day-to-day necessities. But even beyond your checkbook and credit ca
  statements, inflation is eating away at the purchasing power.",
  source: { name: "lifehacker.com", id: "" },
  title: "Why You Should Buy Series Z 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
  },
  {
    _id: "63781729dfde6472c5896528",
    publishedAt: "2022-10-28 12:56:52Z",
    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: "lifehacker.com", id: "" },
    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.Y
  },
  {
    _id: "63781729dfde6472c5896529",
    publishedAt: "2022-10-24 16:30:00Z",
    author: "Meredit 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 see
  },
  {
    _id: "63781729dfde6472c589652a",
    publishedAt: "2022-11-05 13:00:00Z",
    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
    nks to Wendys, 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 thank
  },
  {
    _id: "63781729dfde6472c589652b",
    publishedAt: "2022-11-15 17:30:00Z",
    author: "Meredit 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: "63781729dfde6472c589652b",
    publishedAt: "2022-11-15 17:30:00Z",
    author: "Meredit 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
    any other, with rising prices and supply chain issues hitting",
    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 s
  }
]

```

Data Processing and Popularity Detection

Created a cluster on Google Cloud Platform.

Flowchart for creating cluster



Google Cloud CSCI-5408-F22 Search data proc

Dataproc Clusters CREATE CLUSTER REFRESH START STOP DELETE REGIONS + 5 RECOMMENDED ALERTS SHOW INFO PANEL

Jobs on clusters Clusters Jobs Workflows Auto-scaling policies Serverless

Filter Search clusters, press Enter

Name	Status	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created
cluster-c871	Stopped	us-central1	us-central1-b	0	Off	dataproc-staging-us-central1-244798383643-wmyyby7	20 Nov 2022, 23:34:44

Consider using Auto Zone rather than selecting a zone manually. See <https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/auto-zone>

Name	cluster-c871
Cluster UUID	9d57d913-a83b-4f75-9915-2211f5ca20f3
Type	Dataproc cluster
Status	Stopped

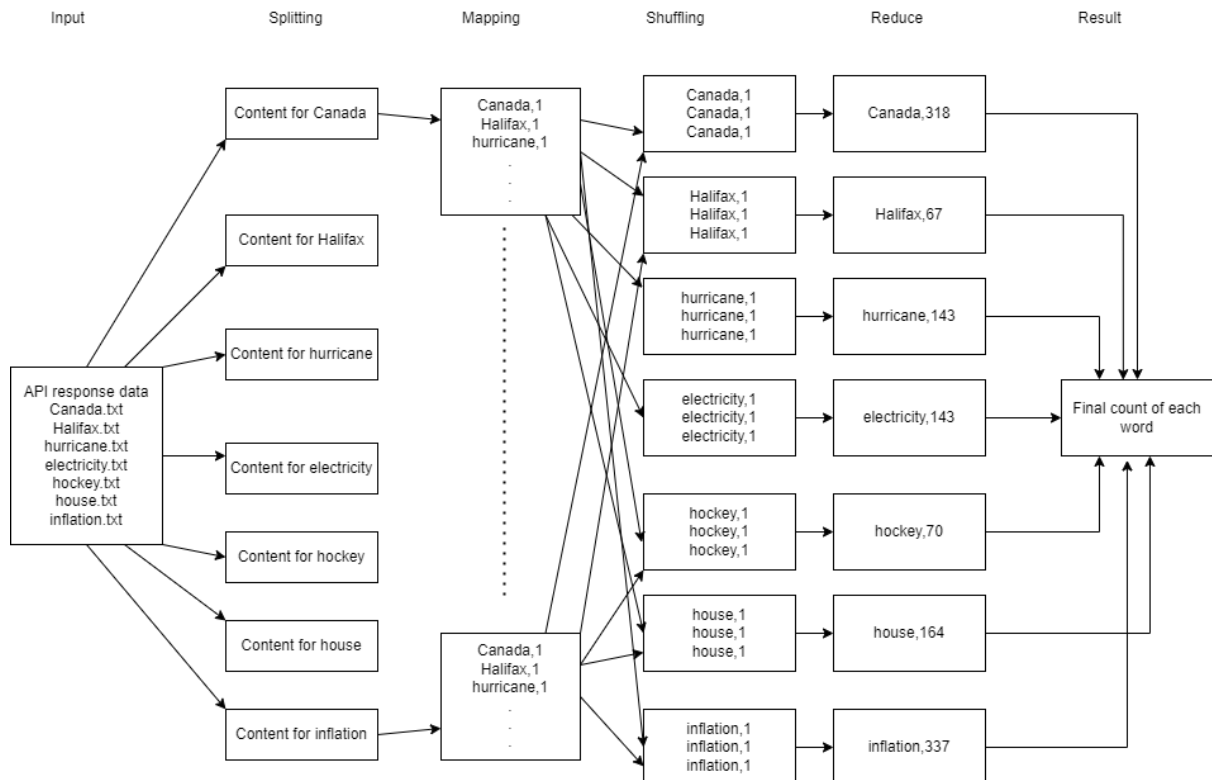
MONITORING JOBS **VM INSTANCES** CONFIGURATION WEB INTERFACES

Filter Filter instances

Name	Role
cluster-c871-m	Master

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

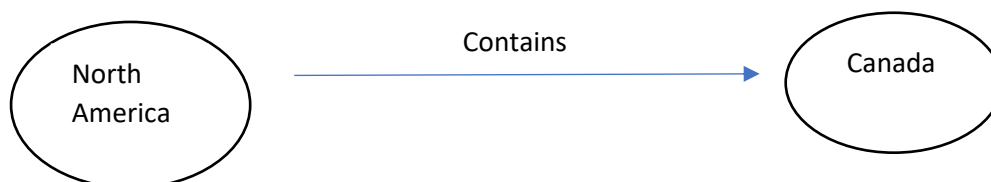
Unlike other databases like relational, key-valued or document database, Neo4j has a native graph approach at the core. Data is saved in nodes and each node is connected to other nodes through relationships. Thus, information about the next node in the sequence is available in the node itself making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

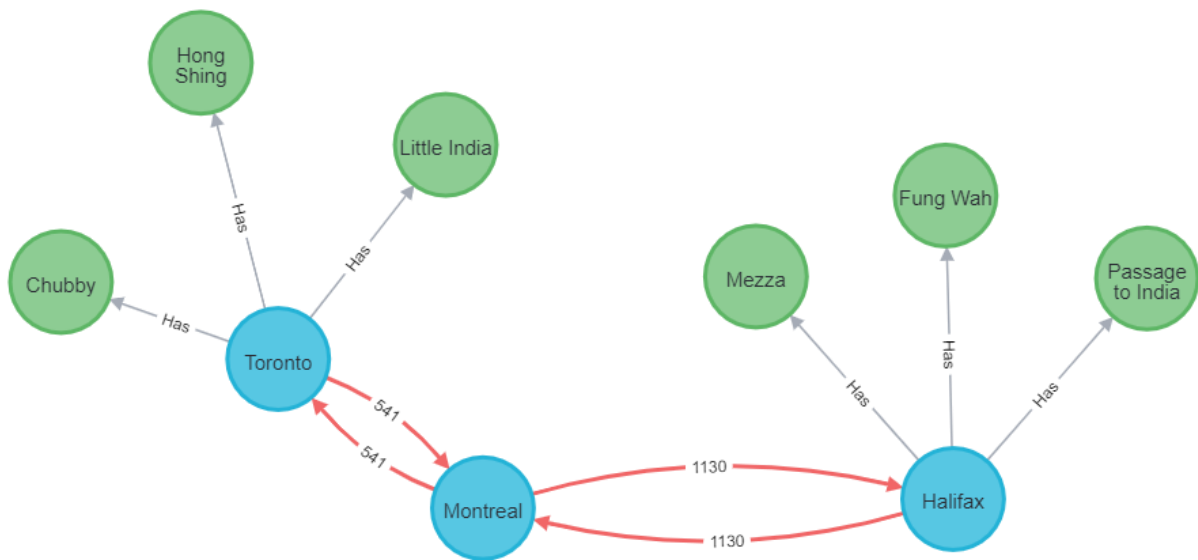
Here, 'northAmerica' is the variable Name, Continent is the node label and name is the node property. - -> depicts the relationship between both nodes. 'rel' is the relationship variable and 'Contains' is the relationship type.

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." .