

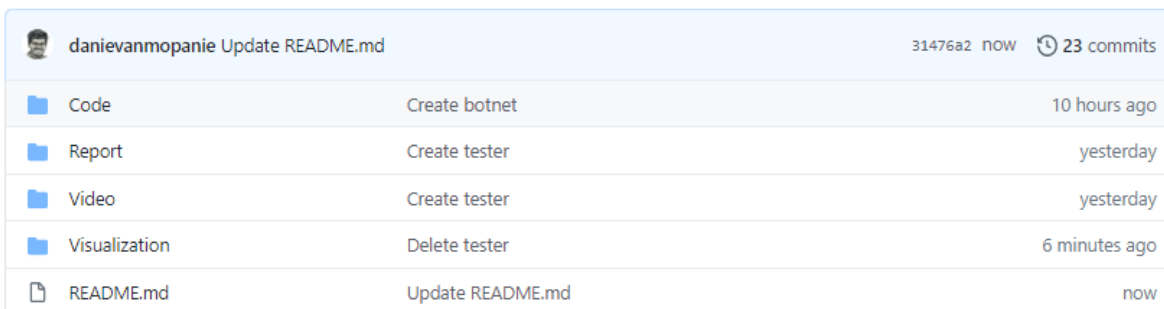
*"The University of Pretoria commits itself to produce academic work of integrity. I affirm that I am aware of and have read the Rules and Policies of the University, more specifically the Disciplinary Procedure and the Tests and Examinations Rules, which prohibit any unethical, dishonest or improper conduct during tests, assignments, examinations and/or any other forms of assessment. I am aware that no student or any other person may assist or attempt to assist another student, or obtain help, or attempt to obtain help from another student or any other person during tests, assessments, assignments, examinations and/or any other forms of assessment."*







Surname: UNGERER - Initials: I.D. - StudentNo: 29129398 - Signature: 

## General

- Please follow the link to my Github here:

[https://github.com/danievanmopanie/MIT805\\_Exam](https://github.com/danievanmopanie/MIT805_Exam)



	danievanmopanie Update README.md		31476a2 now 23 commits
	Code	Create botnet	10 hours ago
	Report	Create tester	yesterday
	Video	Create tester	yesterday
	Visualization	Delete tester	6 minutes ago
	README.md	Update README.md	now

**Figure 1: Github File Structure**

- Please follow the link to my GDrive folder for the MP4 video locations:

[https://drive.google.com/drive/folders/11LcBTyUBXr3hRu\\_hUHETy2h7gsMuJY0E?usp=sharing](https://drive.google.com/drive/folders/11LcBTyUBXr3hRu_hUHETy2h7gsMuJY0E?usp=sharing)

- Please follow the link to the Question 1's video on Youtube here:

<https://www.youtube.com/watch?v=z3K-NiKTR1M>

- Please follow this link to the Tableau Interactive dashboard:

(it has guest permission for external connection):

[http://kizasis075/views/Assignment1Part2/GestationalWeightDashboard?:showAppBanner=false&:display\\_count=n&:showVizHome=n&:origin=viz\\_share\\_link](http://kizasis075/views/Assignment1Part2/GestationalWeightDashboard?:showAppBanner=false&:display_count=n&:showVizHome=n&:origin=viz_share_link)

- Please follow the link to Question 2b's video here:

<https://drive.google.com/file/d/1Pf6ExC70GcmDnQ8DurRtRuTWkevPQarV/view?usp=sharing>

### Question 1 [50 marks]

1. During the semester, you worked on an individual Project i.e., the Project of both assignment 1 and 2 - MIT 805 module. For 50% of the practical questions, you are required to use the same data set (minimum data set 1GB). Show a fully-working system with code and video demo with use cases, core requirements and business value of the project data set using the three defining properties of big data. Then analyse the architecture of the Big Data requirements for all data set using the following questions as guide:

*(For the video answer to questions 1a – 1f, please see GDrive video submission [here](#))*

**Question 2 [30 marks]**

2. During one of our lectures, we discussed the concept of BotNet and how MapReduce could be applied in setting up a BotNet architecture. For this question, present a simulation of cyber-attack where the nodes and the master interact to achieve a cyberattack network of robots used to commit an identified kind of cybercrime.

(a) Describe in detail how MapReduce can be applied to Botnet. In your description, focus on merits, demerits (10) and propose any other big data technique(s) or tool(s) that could best address the BotNet architecture. *Note: For question (2a) submit a PDF document of your answer*

**(a) Description:**

The botnet approach breaches a network of devices by exploiting a weak security layer in a the network of devices. Once the layer is penetrated, command and control software enable a master node (central command) to carry out malicious activities on a large scale due to the distributed network of connected nodes (worker nodes). The level of infiltration by the BotNet attack depends on the ability and extent to which unprotected nodes in the connected network can be reached.

The MapReduce (Hadoop) architecture solution to a BotNet attack follows the “client-server” approach, where a command and control centre sends out the communications protocol to deliver the malicious payload to the worker nodes. Malicious activities are controlled from a central point in this approach.

There are four key elements of Hadoop that are all used in an intertwined way to deliver BotNet attacks. These four elements of Hadoop include :

- i. Distributed File System (HDFS)
- ii. MapReduce
- iii. Hadoop Common
- iv. YARN

MapReduce is the module in the Hadoop architecture that performs the analytics on the datasets. MapReduce takes a large volume of data and reduces it to smaller datasets. There are three major functions executed to deliver the MapReduce component, namely; *Map*, *Reduce* and *Driver* functions. The MapReduce solution uses the *Map* function to crawl through the data and transform it into key-value pairs (smaller chunks of data). The next function, *Reduce*, collects the results from the *Mapping* function. The *Map* function is the

input to the Reduce function. The *Driver* function then delivers (transports) these data packages to a network of connected nodes.

In a BotNet attack, the *Map* function will split the malicious software into smaller chunks. The *Reduce* function processes the data and readies it for distribution to the HDFS. The output of the *Reduce* function is passed to the major element of MapReduce, namely, the *Driver* function. This function sets up the job to be run in Hadoop on the network. It is the actions part of the BotNet distribution.

***Merits:***

- The main advantage of Hadoop – and precisely the point of exploitation for a BotNet attack - is the fact that the technology solution allows reaction time to be as close as possible to real-time. Hadoop enables very fast reaction time, so this means once the BotNet penetrates a network, it is within a heartbeat that the cyber attack's malicious software can be distributed to the nodes in the network.
- The master-slave architecture of Hadoop allows a coordinated attack that spreads from a single node to the connected slave nodes. Data centre-based servers are ideal since they have a lot more bandwidth than IoT devices. Just a small collection of servers can be employed to deliver a very powerful attack.
- Hadoop is extremely scalable – it offers very high bandwidth to allow large workloads to be completed in a blink of an eye. This is perfect to launch a quick and devastating BotNet attack that has already been able to do the damage before being detected.
- It is a very cost-effective architecture solution.
- There is a lot of redundancy in the solution because of its distributed nature. Every node gets to handle a part of the data to process so that no single node can get overburdened.

***Demerits:***

- A Hadoop is complex to manage. It is not a simple technology to use in order to deliver a BotNet attack.

In the MapReduce framework, the paradigm is that the processing units are taken to the data. This is in contrast to a traditional approach where the data is moved to the processing unit. However, this exposed the MapReduce solution to the following vulnerabilities (Sinha, 2016):

- i. Apache Hadoop and MapReduce supports batch processing only.
- ii. It is not very effective with complex algorithms

- iii. It does not support an iterative data processing requirement. Spark technology can assist in this regard when an iterative approach is required.
- iv. When the *Map* phase generates too many keys, it becomes an issue for the *Reduce* function to sorting the keys – this will slow down the performance drastically.

***Other big data technique(s) or tool(s):***

A BotNet attack can also be launched using the YARN module of the Hadoop solution. There is a vulnerability in Hadoop YARN (Yet Another Resource Negotiator) that allows a malicious attack to install malware on as many devices as possible. In Hadoop, YARN is deployed as a resource management and job scheduling system. It is a cluster management tool that allows a central platform for resource management across the Hadoop infrastructure. YARN is responsible for allocating system resources to various applications running in a Hadoop cluster. YARN schedules tasks for execution on the nodes of a connected Hadoop network. A Botnet attack, therefore, uses YARN (as the Hadoop resource scheduler) to deliver malware across a connection of nodes as YARN goes about its task to manage the node network.

Outside the Hadoop infrastructure, there is also a P2P (peer-to-peer) approach to BotNet. This is a more decentralised approach compared to the Hadoop/MapReduce approach. Once a device or collection of devices are infected, these devices scan for malicious websites and even other infected devices in the net of bots. The bots can then deliver the attack based on the latest versions of the BotNet malware from a completely decentralised point of attack (Rouse, 2019).

(b) Setup a Botnet of a minimum of 5-nodes using Hadoop MapReduce consisting of three main parts, i.e., Driver, Mapper and Reducer. *Note: Showcase your simulation using a 5 to 10mins video, then upload the video demo with a voice over as an MP4 format - filename:StudentNumber.Surname-Q2b.MP4 to this link - <https://forms.gle/tWTbSjKoixdnp2PXA> - NO VideoDemo equals '0' mark* (15)

(b) (For the video answer to question 2b, please see GDrive video submission [here](#))

(c) In what business/industries case scenario can BotNet be useful? Provide some detailed example cases. (5)  
*Note: For question (2c) submit a PDF document of your answer*

(c) Here follow examples of businesses/industries where BotNet can be useful:

i. **Advertising (Spam at scale)**

In October of 2007, the internet was bombarded with spam e-mails promoting Ron Paul as U.S. presidential candidate (Stewart, 2007). The reach of the audience using the MapReduce heuristic and distribution of e-mail nodes can be useful in disseminating an advertisement.

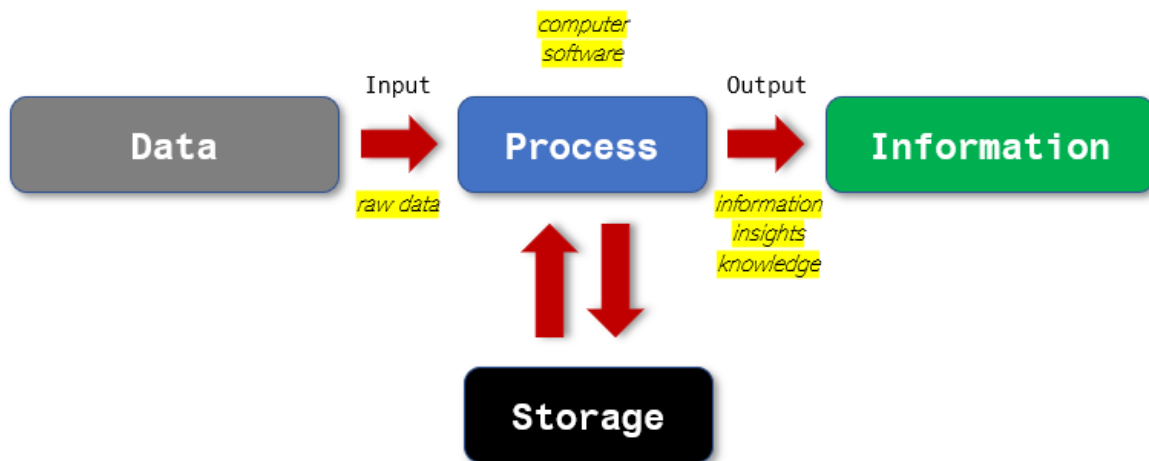
ii. **Security Monitoring**

The same BotNet architecture can be used to monitor malicious BotNet attaches. A distributed solution that monitors bot can acts as detectors for when benign bots launch an attach. An example is SolarWinds solution. This security software gathers logs from multiple sources and synthesises (*map* and *reduce*) them to improve situational awareness across a network.

**Question 3 [20 marks]**

3. (a) Discuss what data processing is? Present your discussion in line with the functions of data processing, that highlights how your data set for the **Semester Project Assignment** were explored? *For question (3a) submit only PDF document of your answer* (5)

(a) Corporations cannot use data in its raw format. Similar to mining that sends ore through a plant beneficiation process to sell a value-add product, data needs to undergo a value-add process to transform unusable, raw data into usable information, insight and knowledge. This value-add process is what is referred to a 'data processing'. Figure 1 represents a visual representation of the data processing life cycle - adapted from *DAMA-DMBOK Data Management Book of Knowledge* (Data Management Association, 2017).



**Figure 2: Data Processing Cycle** (Data Management Association, 2017)

The life cycle for my **Semester Project Assignment** is depicted in Figure 3. I employed various APIs on the Google Cloud Platform to ensure the processing life cycle is properly managed for my business case. This included the data (Natality BigQuery® Public Dataset on Cloud Storage®), sending (Cloud Dataflow®) the data to BigQuery® where the raw data gets cleaned, transformed and processed for machine learning. Throughout the processing process in BigQuery®, the data tables and machine learning predictions get stored in Cloud Storage®. This allows for further analysis and ensures a pragmatic approach to adding value to the raw data. At the end of the process, the results are visualised using bespoke technology, Tableau. This is another beneficiation process since the reports and visuals are convey knowledge and insight to the end-user (Google, 2020).

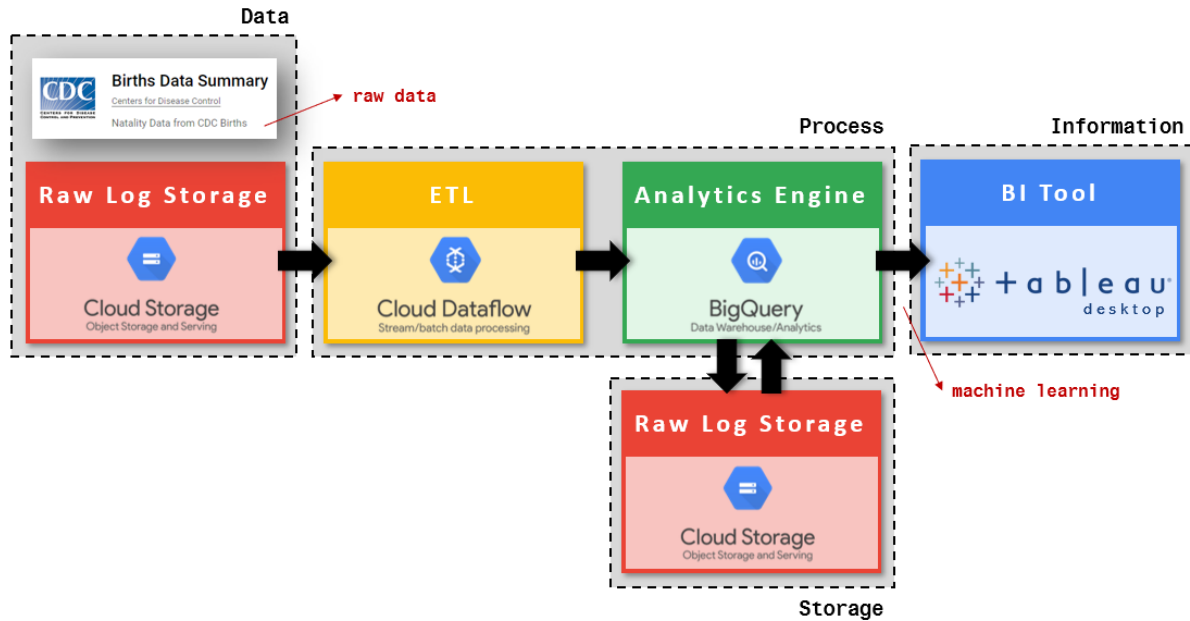


Figure 3: Semester Project Assignment Life Cycle

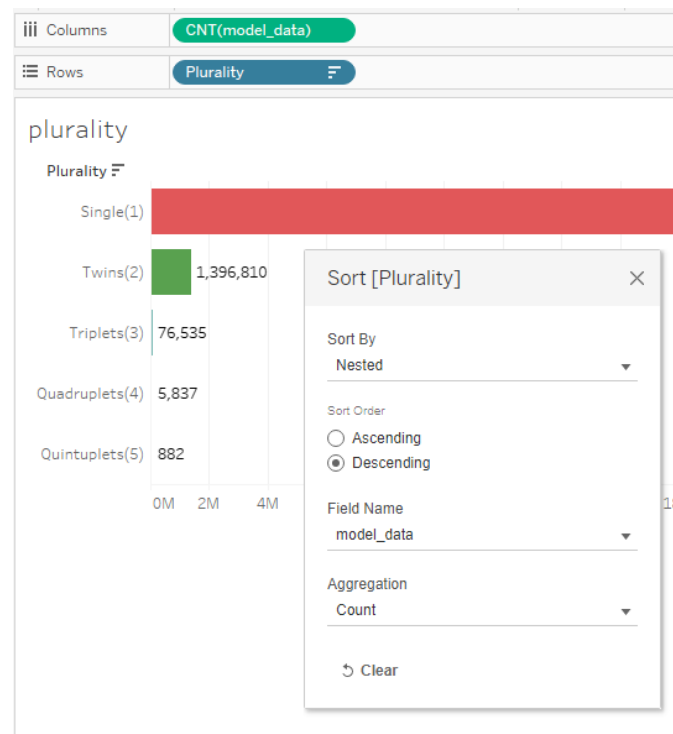
The functions of data processing for the **Semester Project Assignment** is captured in Table 1.

Table 1: Functions of Data Processing

FUNCTION	DESCRIPTION
Classification	The classification function separates the data into various categories. The process was completed in with the SQL queries to ensure that categorical data, that was represented as ordinal data in the raw data, gets converted to factor values with the appropriate level numbers. For example, the [Race] dimension is a numerical set of data (1 to 78). This was changed to a string to represent it as categorical values (i.e. white, black, Asian, et cetera).
Aggregation	The data was batch processes as a single unit from the public dataset gallery in Google. However, there was a requirement to executed a SQL inner join after the linear regression model was evaluated. The 'model feature' and 'weights' table were joined together to inspect the weights of the continuous and categorical regression weight as a unit. The aggregation was safed as a table in the project database for future reference.
Sorting	Tableau, as a visualisation tool, has a potent sorting function (see Figure 4). This sorting function was regularly used to present the the data (bar charts, line charts) in a logical, understandable fashion.
Summarisation	Various levels of aggregation (over time periods, like avarages for a day, week, month of year) very executed. For example, the average weight per year for boys that we born between 1990 and 2020.



FUNCTION	DESCRIPTION
	Those that seek to use the reports, are interested in information that is summarised to the correct level. Too much detail, but also too little detail can lead to miscommunication to the stakeholders.
Analysis	<p>The k-means clustering algorithms is a perfect example. This machine-learning algorithm assisted in summarising the relationships between subgroups of the data. It was the first step in identifying the strong impact that the age of a parent play in gestational weight.</p> <p>The visualisation from the specific k-means clustering result, was extremely effective in summarising the results in a very simplistic way (see Figure 5).</p>
Reporting	<p>The aim of reporting is to summarise the information and analysis. This was done through a detail report that referred to all the assumptions, technical model explanations and resulting tables.</p> <p>Reporting data should effectively capture the essence of the message and be accurate in transferring the required knowledge to the reader.</p>
Validation	<p>The main validation was done trough assessing the accuracy of the two machine learning models.</p> <p>The k-means clustering results were validated against the Davies-Bouldin index. The linear regressing model results were validated through the mean absolute error when fitting a model to the data.</p>



**Figure 4: Tableau Sorting Function used to Sort the Plurality Category Count as Descending for the Bar Chart**

## Metrics

Davies-Bouldin index	1.7656
Mean squared distance	3.8252

## Numeric features

This table shows the centroid value for each feature. Use the select menu to view more numeric features.

Select features (6/6) ▾











































Centroid Id	Count	weight_kilograms	mother_age	father_age	gestation_weeks	weight_gain_kilograms	parent_age
1	3649145	 3.2059	 26.9670	 29.8804	 38.9007	 13.2658	 56.8474
2	1852232	 1.7944	 28.4070	 30.7938	 31.4015	 14.5618	 59.2008
3	3216709	 3.3593	 27.3690	 29.9819	 39.0174	 43.9885	 57.3509
4	13123414	 3.4744	 34.1247	 37.5116	 39.0858	 13.8994	 71.6363
5	14144397	 3.3718	 27.5518	 29.6001	 39.0781	 13.4945	 57.1519
6	6711343	 3.0590	 20.0155	 21.8670	 38.4061	 13.8556	 41.8825
7	6348775	 3.8097	 22.6230	 24.6764	 40.6840	 16.6871	 47.2995

Figure 5: K-Means Clustering Result Output

- (b) Differentiate clearly between traditional streaming data and stateful streaming data [2marks]. *Note: To obtain a full mark to this question, Students are expected to use some real-world examples to obtain [4marks], present merits [2marks] and demerits [2marks] in discussing the question For question (3b) submit only PDF document of your answer* (10)

### (b) Traditional Streaming versus Stateful Streaming

TRADITIONAL STREAMING	STATEFUL STREAMING
<p>Also referred to as stateless streaming (Shwartz, 2016).</p> <p>Each event is handled on its own and has no dependency on a preceding event. The stream processor will, therefore treat every event in the exact same way – regardless of the data that arrived beforehand. It is like a vending machine – there is a single request and a single response.</p> <p>Stateless streaming does not require memory or history to perform an action.</p> <p>Stateless processes can be understood in isolation. Each transaction is created from scratch – as if for the first time with no context required.</p> <p>When the service is interrupted or closed accidentally, it is as easy as just starting up a new stateless streaming operation. There is very little lost from the disruption.</p>	<p>The “state” in the title implies that there is a dependency between events – i.e. past events have an influence on the way current events are processed.</p> <p>A stateful solution keeps track of user-profiles, preferences and user actions. Information is regarded as the “status” of the system.</p> <p>Stateful functions need the ability to return to results over and over again.</p> <p>Previous context (history) is essential here. The current transactions are affected by what happened in the past (like an online shopping cart that is updated continuously). The same server will need to be used for each time a user request is processed.</p> <p>Stateful transactions are like an ongoing conversation over a period with the same person.</p>

TRADITIONAL STREAMING	STATEFUL STREAMING
<p><b>Examples:</b></p> <ul style="list-style-type: none"> <li>i. Keep track of online user sessions (aggregated output is only per session, typically communicated when the monitoring session ends).</li> <li>ii. A very practical example is like an internet search. A once-off hit search without the need to keep a history.</li> <li>iii. Transformations of one record to another record.</li> <li>iv. A filter that filters out data that does not fall in line with the business rule definitions of the corporation.</li> <li>v. Processing of short-term requests (web servers, print servers to process short-term requests).</li> <li>vi. HTTP, UDP (User Datagram Protocol) and DNS (Domain Name System)</li> </ul>	<p><b>Examples:</b></p> <ul style="list-style-type: none"> <li>i. Online banking</li> <li>ii. E-mail</li> <li>iii. Shopping cart when doing online shopping – always aware of the content.</li> <li>iv. FTP (File Transfer Protocol), Telnet</li> <li>v. Tracking in apps of the following: window location, setting preferences and recent activity.</li> <li>vi. Apache Flink is a popular stateful streaming technology solution.</li> <li>vii. Streaming sensor measurements</li> </ul>
<p><b>Merits:</b></p> <ul style="list-style-type: none"> <li>i. Very stable and reliable.</li> <li>ii. Server and client are loosely coupled and can act independently</li> <li>iii. Server design is simple to implement</li> </ul>	<p><b>Merits:</b></p> <ul style="list-style-type: none"> <li>i. Ability to scale to harvest enormous data volumes.</li> <li>ii. Creates very good user experiences when the app tracking data is used to improve the user's experience.</li> </ul>
<p><b>Demerits:</b></p> <ul style="list-style-type: none"> <li>i. Does not keep transactional history. I.e. no context. It is only concerned with processing the task at hand.</li> <li>ii. Changing the schema of a table or scaling a traditional streaming solution requires careful planning and a lot of effort. It is not as flexible to handle changes on-the-go.</li> </ul>	<p><b>Demerits:</b></p> <ul style="list-style-type: none"> <li>i. Harder to scale up due to various "workers" sharing the state.</li> <li>ii. It has complicated hardware and software set-up.</li> <li>iii. Highly dependent on technology to execute desires.</li> <li>iv. State is lost if the node dies.</li> <li>v. Challenging when it comes to fault tolerance.</li> <li>vi. Not easy to recover from errors</li> <li>vii. Security becomes a challenge here. All data is required to be encrypted in-flight.</li> </ul>

- (c) Discuss how Hadoop addresses the needs of data in both the traditional and stateful data streaming. (5)  
 Elaborate using your **Semester Project Assignment**. *For question (3c) submit only PDF document of your answer*

- (c) For my **Semester Project Assignment**, I utilised a technology called BigQuery on the Google Cloud Platform (GCP). BigQuery is a service that manages both the storage of my data (SQL tables and views), and also scales appropriately for calling up large datasets and booking out parallel slots for machine learning training. The scaling is automatically done by the BigQuery API. This is in contrast to Hadoop, where scale is achieved by adding more nodes to a network. With GCP and BigQuery, there are no nodes.

However, the dynamic nature of utilising traditional (stateless) streaming and stateful streaming still applies in this regard.

**Traditional Streaming:** I utilise this when I run my SQL queries from the source (21.9GB Natality Dataset) to pre-process (clean, aggregate and transform my data). I write a lot of filtering statements (exclude nulls or only show data between 2000 and 2005 with a WHERE clause). These filtering clauses are all stateless streaming operations.

**Stateful Streaming:** My business case is inherently not a streaming solution – it is a batch processing solution from the source data that is curated on a quarterly basis. However, the opportuning for a stateful streaming architecture is still there. Due to the fact that the dashboard is an HTTP hosted solution, it could be very interesting and useful to monitor the following in a stateful streaming operation:

- i. Traffic on the host site (metadata like location, duration and hit-patterns are of concern here)
- ii. Search history of users. An aggregation for various sessions per user will be useful stateful streaming operations.

These website-monitoring solutions need access to the right data at the right time and to host this data for whatever visualisation tool will connect to the streaming results. At the very least, such a solution will require 24 hours of website monitoring data to keep the service running and for the aggregations per user to take place. This monitoring activity can build out the business case for the educational solution I built. One is able to see the reach of the dashboards. This will assist in an educational support approach to identify geolocations where

there is little traffic (might require an awareness campaign in the town) or identify active areas where supporting functions might have to be increased (doctor, antenatal care facilities and clinics).

## Bibliography

Data Management Association (2017) *DAMA-DMBOK Data Management Body of Knowledge*. 2nd edn. Edited by D. Henderson. New Jersey: Technics Publications.

Google (2020) *BigQuery public datasets | Google Cloud, Google Cloud*. Available at: <https://cloud.google.com/bigquery/public-data> (Accessed: 14 September 2020).

Rouse, M. (2019) *What is a Botnet and How Does it Work?*, *TechTarget*. Available at: <https://searchsecurity.techtarget.com/definition/botnet> (Accessed: 26 November 2020).

Shwartz, D. (2016) *4 Valuable Resources on Stream processing*, *alooma*. Available at: <https://www.alooma.com/blog/stream-processing-101>.

Sinha, S. (2016) *Fundamentals of MapReduce with MapReduce Example | by Shubham Sinha | Edureka | Medium, Medium*. Available at: <https://medium.com/edureka/mapreduce-tutorial-3d9535ddbe7c> (Accessed: 26 November 2020).

Stewart, J. (2007) *Inside the 'Ron Paul' Spam Botnet | , Secureworks*. Available at: <https://www.secureworks.com/research/srizbi> (Accessed: 27 November 2020).