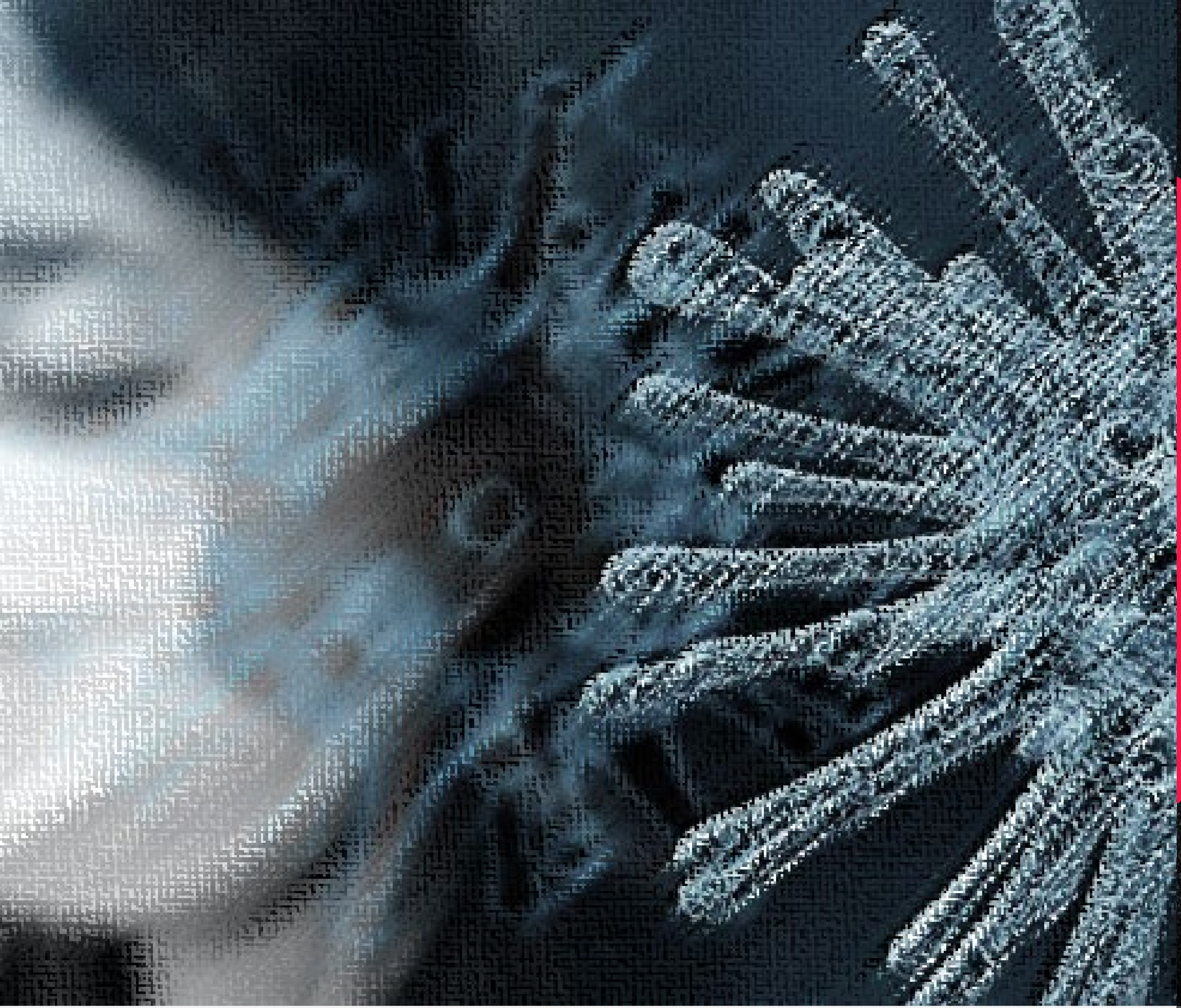


Combating COVID-19 Outbreaks

WQD7005 Data Mining
Master of Data Science | University of Malaya

Part B: Management of Data Assignment

Group Members:
Azwa bin Kamaruddin (WQD170089)
Kok Hon Loong (WQD170086)

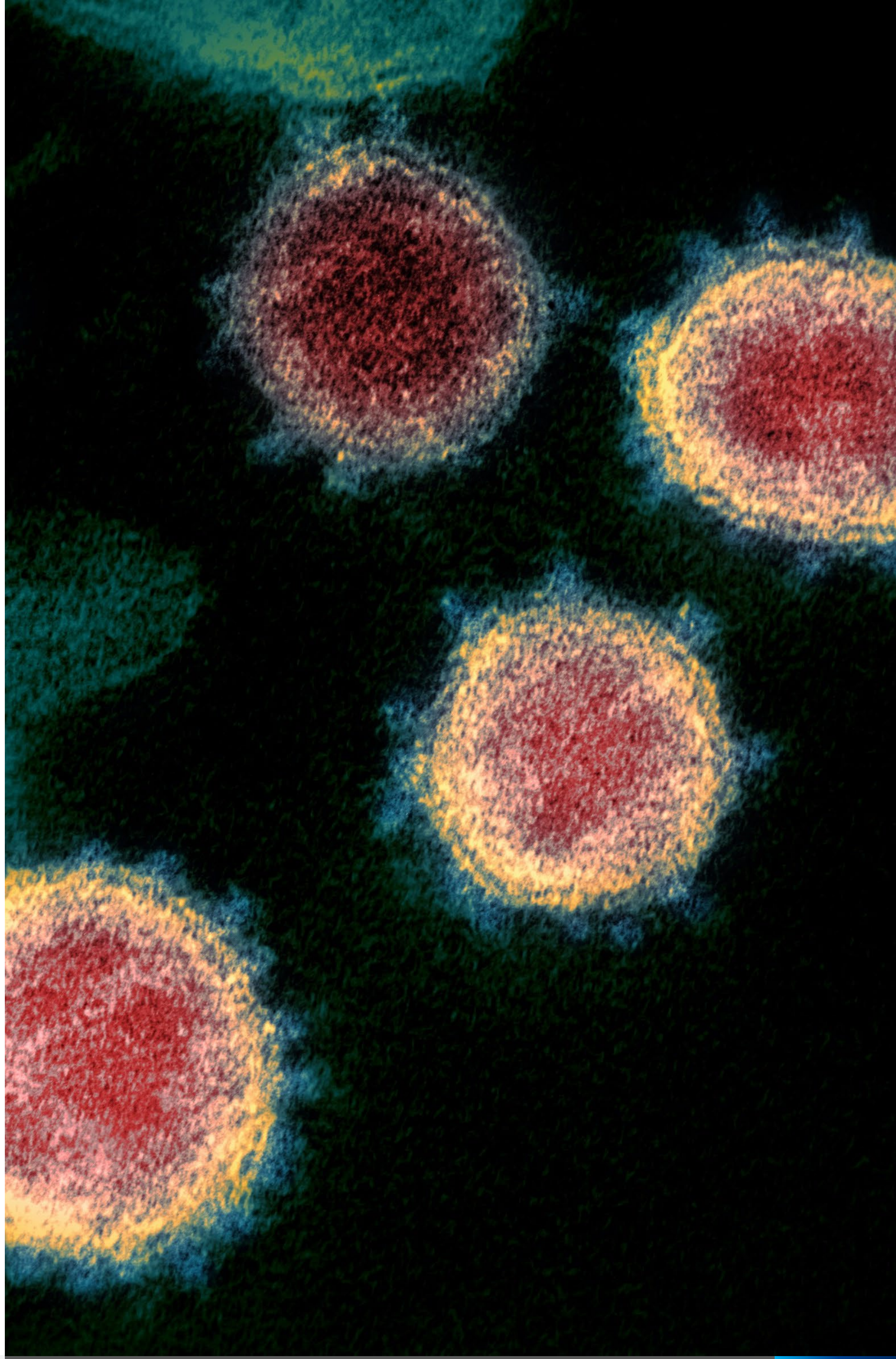
UNIVERSITY OF MALAYA

Assignment Milestones

WQD7005 Data Mining | Semester 2 Session 2019/2020

UNIVERSITY OF MALAYA

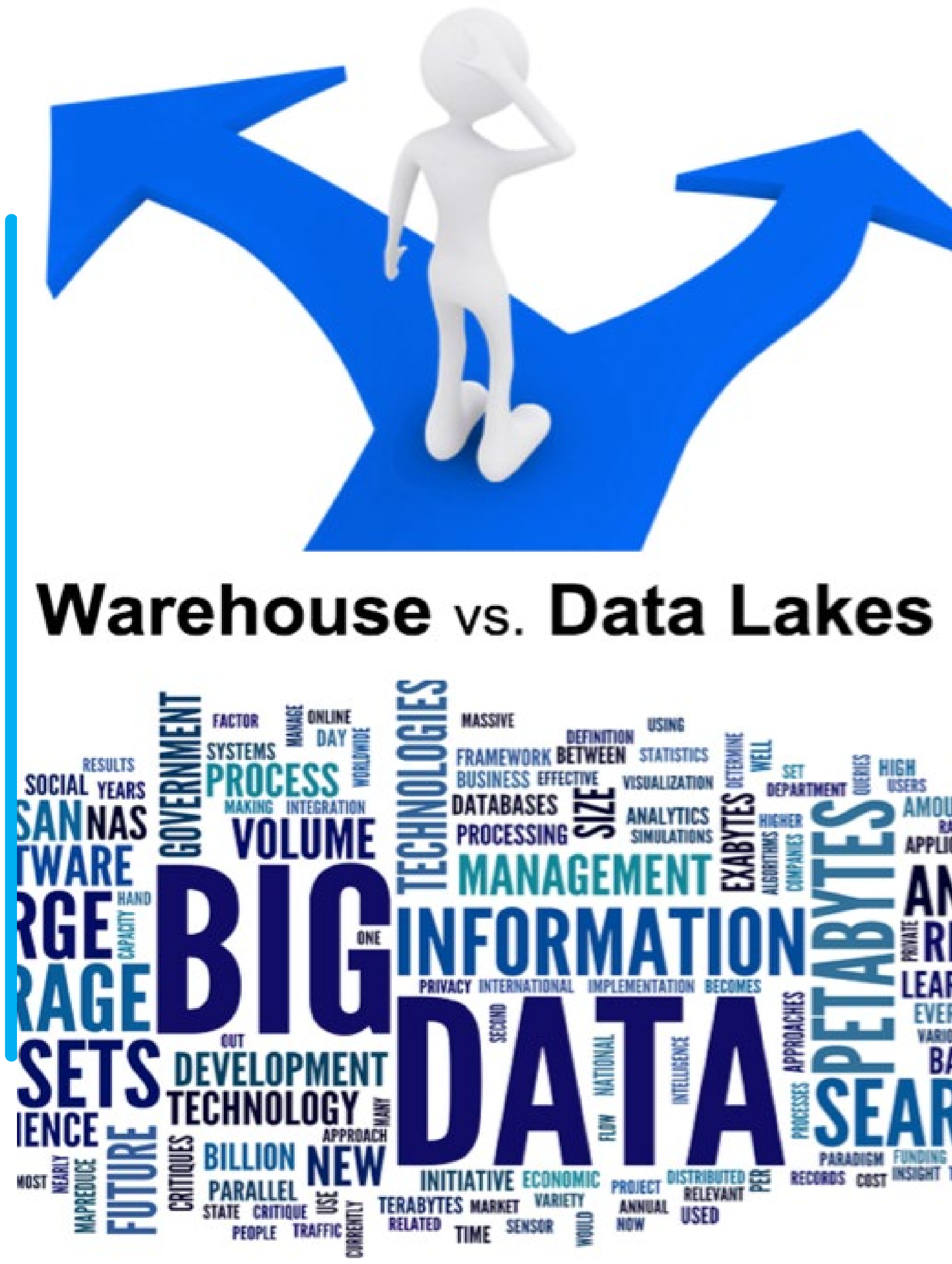
- Part A: (Group)
  - Web Crawling of Real-time Data
- **Part B: (Group)**
  - **Management of Data using Hadoop Data Warehouse or Data Lake**
- Part C: (Group)
  - Accessing and Processing of Data from Hadoop Data Warehouse or Data Lake
- Part D: (Individual)
  - Interpretation and Communication of Data Insights
- Part E: (Group)
  - Deployment of the Data Mining Results on Web (Flask) and Mobile Application (Kivy)



Assignment Background

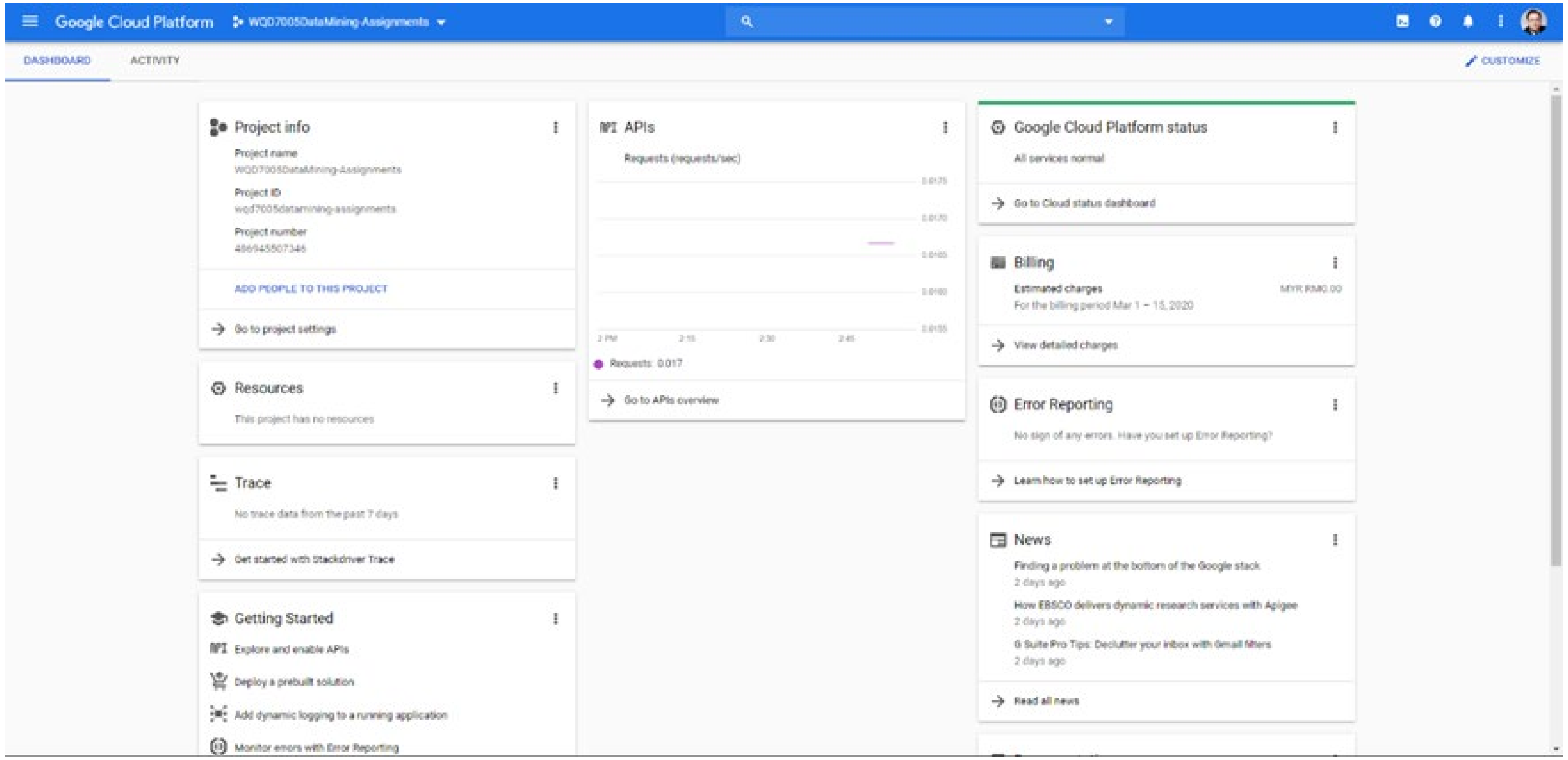
After we have acquired the data from the websites using the web crawling method as explained in Part A of the assignment, our group will be setting up a storage repository so that we can proceed in processing the data over and over to extract more data as we learn more about the contents.

- In this assignment, our team have decided to proceed in setting up the Hadoop Data Warehouse leveraging on the Google Cloud Platform (GCP).
- We will be using Hive as the SQL-like scripting language interface to query and analyze the data acquired from the websites in Hadoop HDFS, and the scripting code is also uploaded in our group assignment GitHub at the link below:
    - ○ https://github.com/scholarazwa/wqd7005-assignment



Configuring of the Hadoop & Hive Data Warehouse Cluster

using Google Cloud Platform (GCP)



Project info

Project name
WQD7005DataMining-Assignments

Project ID
wqd7005datamining-assignments

Project number
486945507346

ADD PEOPLE TO THIS PROJECT

→ Go to project settings

Resources

This project has no resources

Trace

No trace data from the past 7 days

→ Get started with Stackdriver Trace

Getting Started

Explore and enable APIs

Deploy a prebuilt solution

Add dynamic logging to a running application

Monitor errors with Error Reporting

APIs

Requests (requests/sec)

● Requests: 0.017

→ Go to APIs overview

Google Cloud Platform status

All services normal

→ Go to Cloud status dashboard

Billing

Estimated charges MYR RM0.00
For the billing period Mar 1 – 15, 2020

→ View detailed charges

Error Reporting

No sign of any errors. Have you set up Error Reporting?

→ Learn how to set up Error Reporting

News

Finding a problem at the bottom of the Google stack
2 days ago

How EBSCO delivers dynamic research services with Apigee
2 days ago

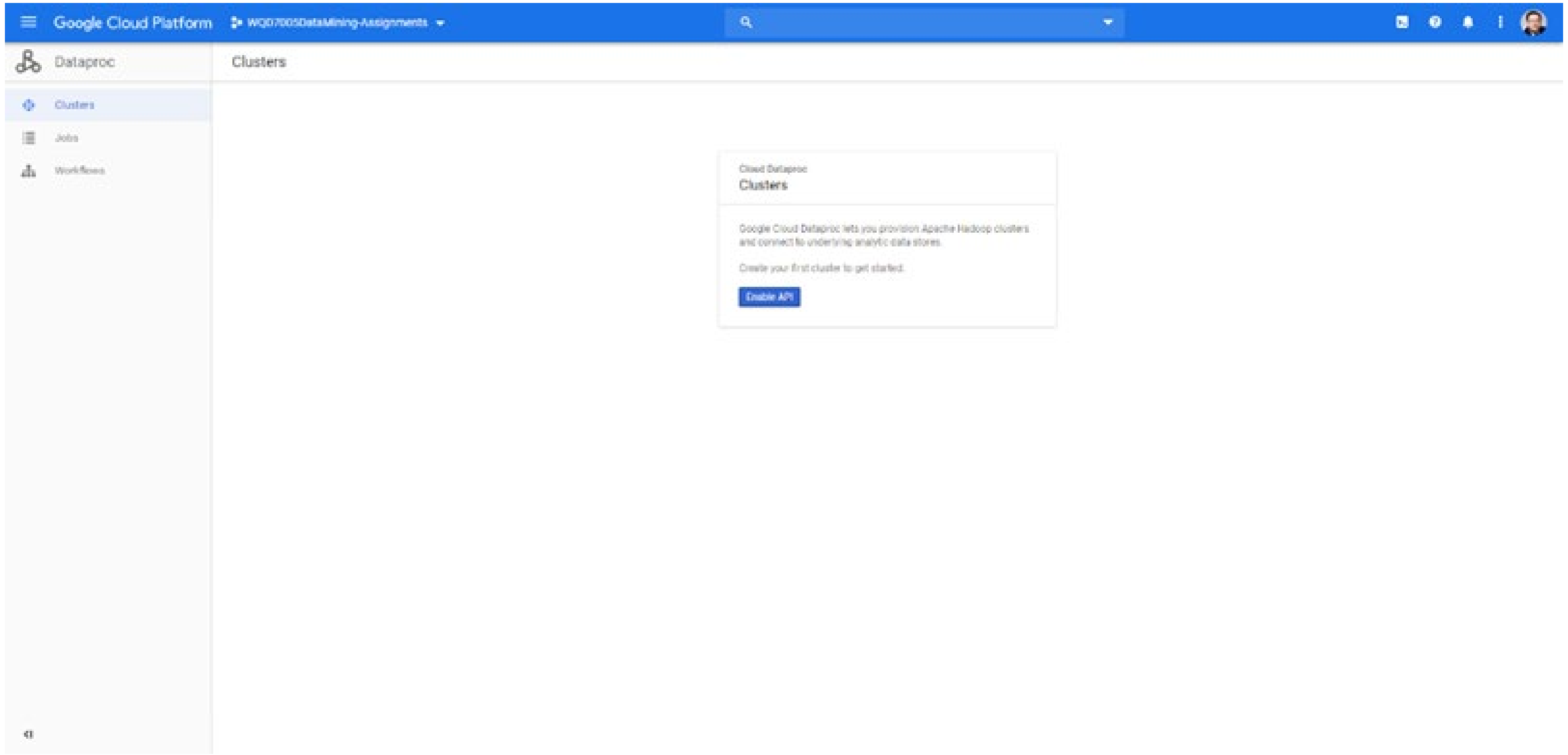
G Suite Pro Tips: Declutter your inbox with Gmail filters
2 days ago

→ Read all news

- For this course assignment, we have created a new project for our team to work on using GCP.
- The screen on the left illustrate a brand new project dashboard that our team have created.

Configuring of the Hadoop & Hive Data Warehouse Cluster

using Google Cloud Platform (GCP)



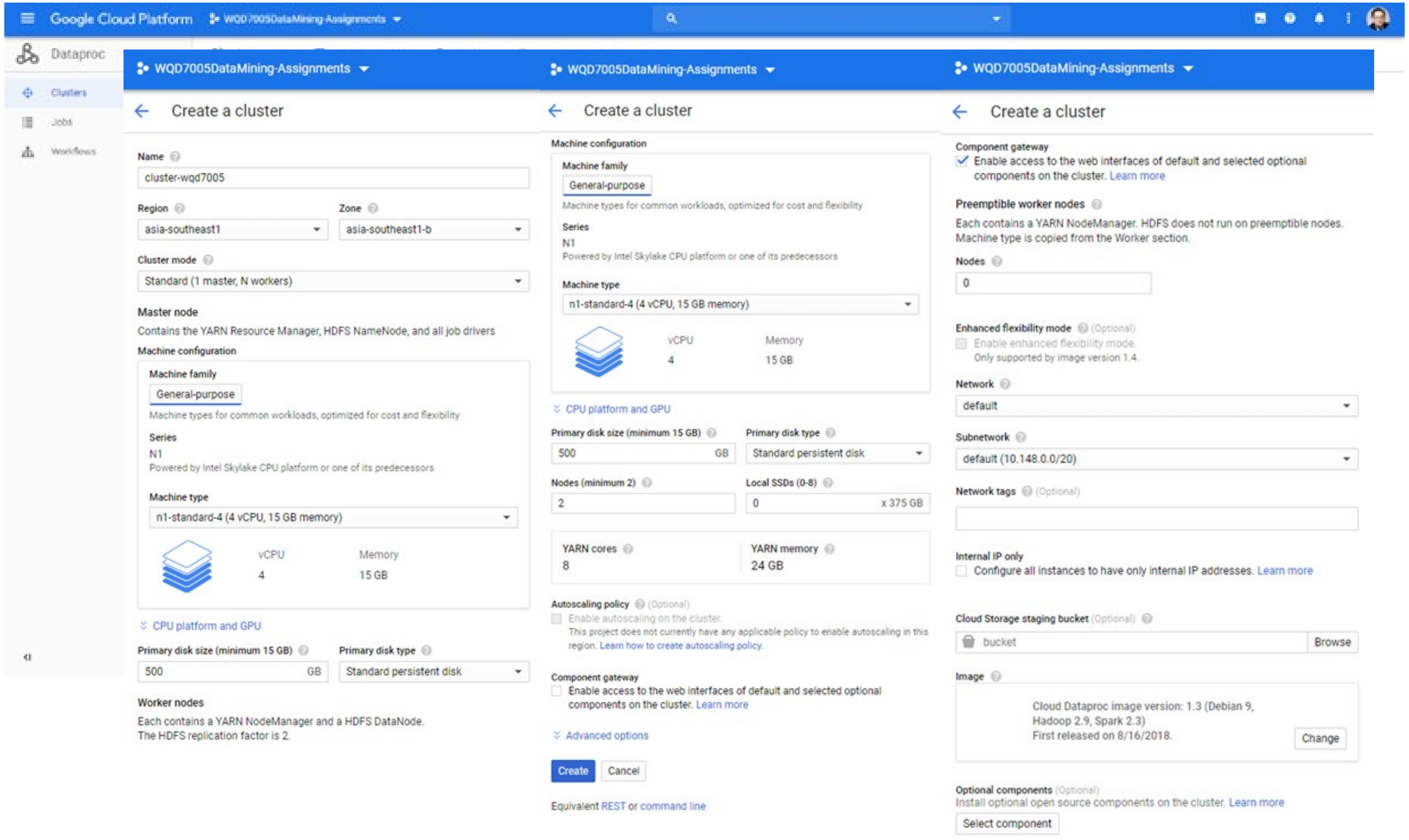
Google Cloud Dataproc lets you provision Apache Hadoop clusters and connect to underlying analytic data stores.

Create your first cluster to get started.

- To configure Hadoop and Hive Data Warehouse in GCP, click on the ***Navigation Menu*** at the top left corner.
- Scroll all the way down until you reach to the ***Big Data*** section, and then click on the ***Dataproc*** to create the cluster.
- Next, click on ***Enable API*** button to proceed.

Configuring of the Hadoop & Hive Data Warehouse Cluster

using Google Cloud Platform (GCP)



WQD7005DataMining-Assignments

Create a cluster

Name: cluster-wqd7005

Region: asia-southeast1

Zone: asia-southeast1-b

Cluster mode: Standard (1 master, N workers)

Master node

Contains the YARN Resource Manager, HDFS NameNode, and all job drivers

Machine configuration

Machine family: General-purpose

Machine types for common workloads, optimized for cost and flexibility

Series: N1

Powered by Intel Skylake CPU platform or one of its predecessors

Machine type: n1-standard-4 (4 vCPU, 15 GB memory)

vCPU: 4

Memory: 15 GB

CPU platform and GPU

Primary disk size (minimum 15 GB): 500 GB

Primary disk type: Standard persistent disk

Worker nodes

Each contains a YARN NodeManager and a HDFS DataNode.
The HDFS replication factor is 2.

WQD7005DataMining-Assignments

Create a cluster

Machine configuration

Machine family: General-purpose

Machine types for common workloads, optimized for cost and flexibility

Series: N1

Powered by Intel Skylake CPU platform or one of its predecessors

Machine type: n1-standard-4 (4 vCPU, 15 GB memory)

vCPU: 4

Memory: 15 GB

CPU platform and GPU

Primary disk size (minimum 15 GB): 500 GB

Primary disk type: Standard persistent disk

Nodes (minimum 2): 2

Local SSDs (0-8): 0 x 375 GB

YARN cores: 8

YARN memory: 24 GB

Autoscaling policy (Optional)

☐ Enable autoscaling on the cluster.
This project does not currently have any applicable policy to enable autoscaling in this region. Learn how to create autoscaling policy.

Component gateway

☐ Enable access to the web interfaces of default and selected optional components on the cluster. Learn more

Advanced options

Create | Cancel

Equivalent REST or command line

WQD7005DataMining-Assignments

Create a cluster

Component gateway

☑ Enable access to the web interfaces of default and selected optional components on the cluster. Learn more

Preemptible worker nodes

Each contains a YARN NodeManager. HDFS does not run on preemptible nodes.
Machine type is copied from the Worker section.

Nodes: 0

Enhanced flexibility mode (Optional)

☐ Enable enhanced flexibility mode.
Only supported by image version 1.4.

Network: default

Subnetwork: default (10.148.0.0/20)

Network tags (Optional)

Internal IP only

☐ Configure all instances to have only internal IP addresses. Learn more

Cloud Storage staging bucket (Optional): bucket | Browse

Image

Cloud Dataproc image version: 1.3 (Debian 9, Hadoop 2.9, Spark 2.3)
First released on 8/16/2018. | Change

Optional components (Optional)

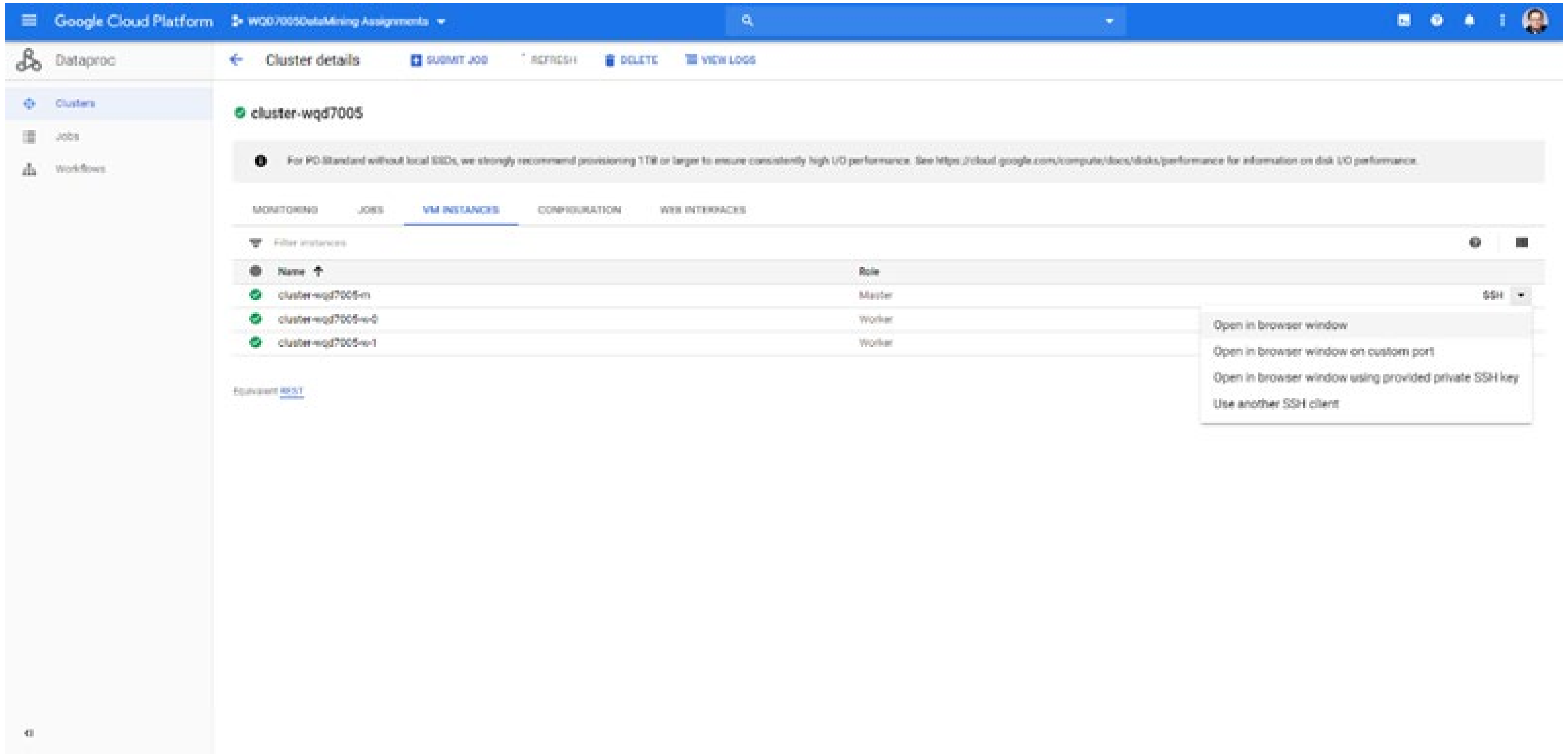
Install optional open source components on the cluster. Learn more

Select component

- Once provisioning the Apache Hadoop is initiated in GCP, click on the ***Create cluster*** button.
- Our team has named our assignment cluster as ***cluster-wqd7005*** and change the ***region*** and ***zone*** align to our current location.
- Expand the ***Advanced options*** and click on the ***Enable access to web interface*** checkbox. Leave the rest of the setting as default.
- Next, click on ***Create*** to proceed.

Configuring of the Hadoop & Hive Data Warehouse Cluster

using Google Cloud Platform (GCP)



cluster-wqd7005

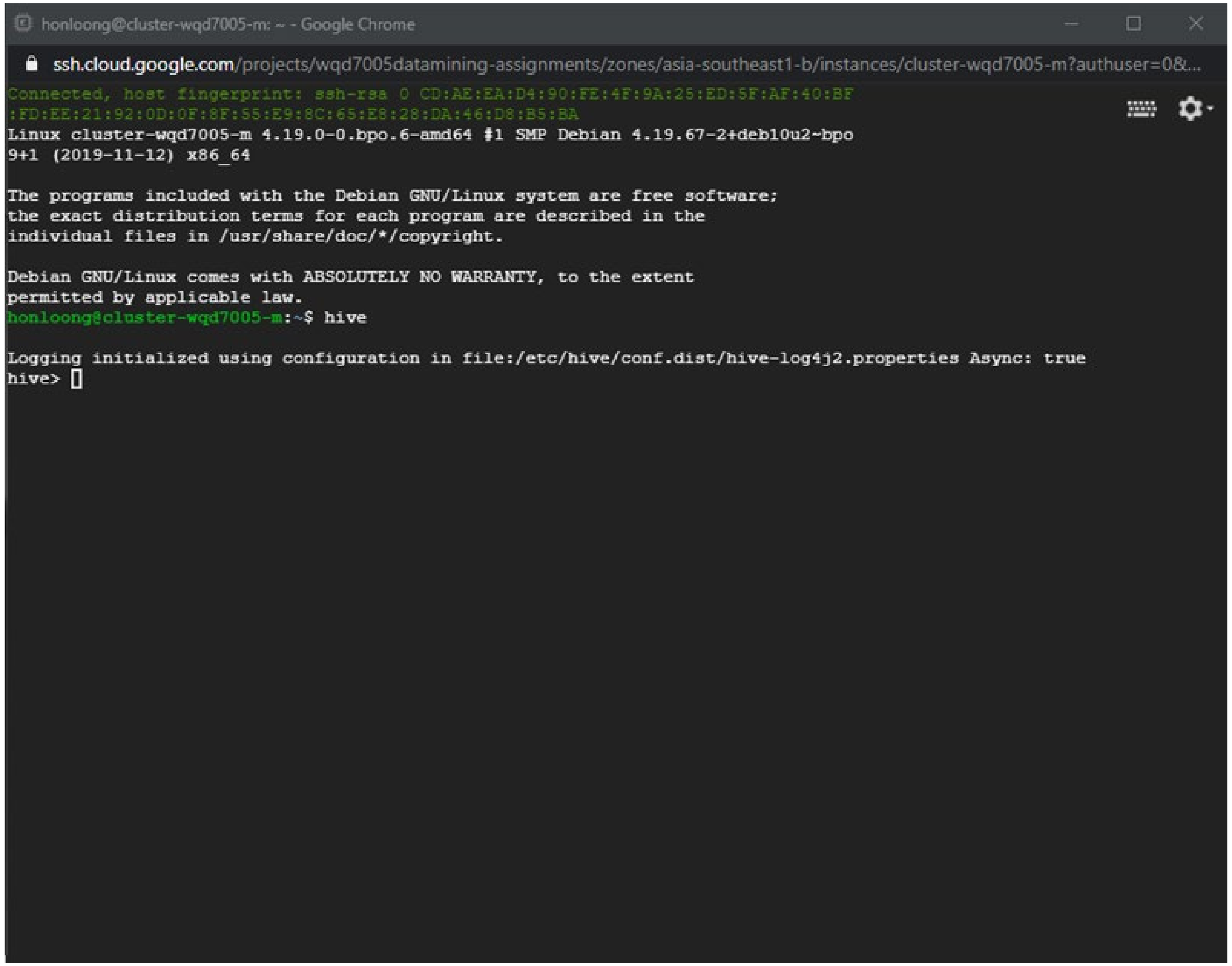
For PD-Standard without local SSDs, we strongly recommend provisioning 1TB or larger to ensure consistently high I/O performance. See https://cloud.google.com/compute/docs/disks/performance for information on disk I/O performance.

Name	Role	
cluster-wqd7005-m	Master	SSH
cluster-wqd7005-w-0	Worker	
cluster-wqd7005-w-1	Worker	

- It will takes a while for clusters to be created and once it's completed, the ***green checkbox*** as shown will appear.
- To verify on the cluster provisioned, click on the ***cluster name*** and it will show the Hadoop services are started.
- Click on the ***Cluster*** on the left, the Master and Workers nodes should be active as well.
- To test out the service, click on the ***VM Instances*** tab and from the ***SSH*** pull-down menu on the far right, select ***Open in browser window***.

Configuring of the Hadoop & Hive Data Warehouse Cluster

using Google Cloud Platform (GCP)



honloong@cluster-wqd7005-m: ~ - Google Chrome

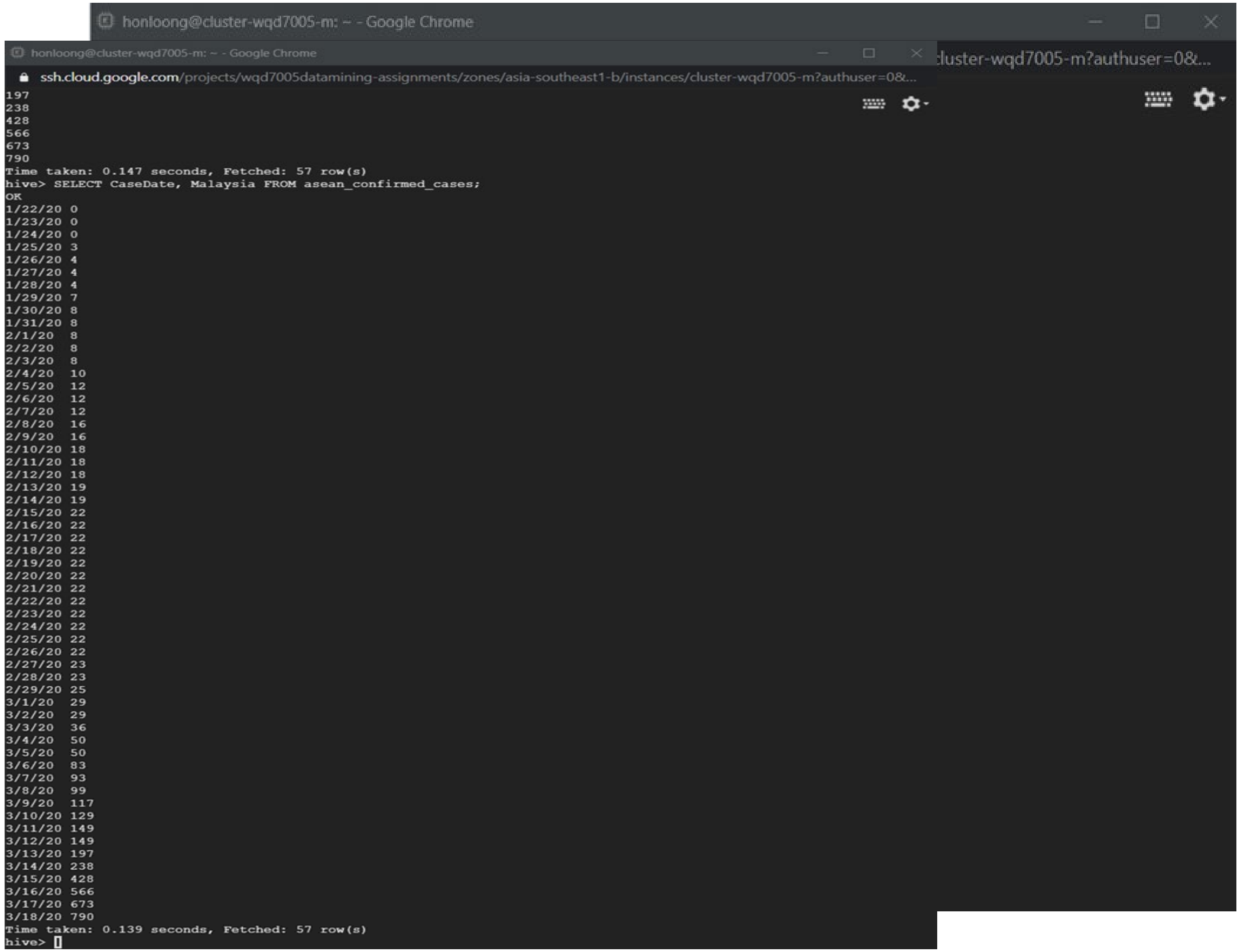
ssh.cloud.google.com/projects/wqd7005datamining-assignments/zones/asia-southeast1-b/instances/cluster-wqd7005-m?authuser=0&...

```
Connected, host fingerprint: ssh-rsa 0 CD:AE:EA:D4:90:FE:4F:9A:25:ED:5F:AF:40:BF
:FD:EE:21:92:0D:0F:8F:55:E9:8C:65:E8:28:DA:46:D8:B5:BA
Linux cluster-wqd7005-m 4.19.0-0.bpo.6-amd64 #1 SMP Debian 4.19.67-2+deb10u2~bpo
9+1 (2019-11-12) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
honloong@cluster-wqd7005-m:~$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
hive> 
```

- The ***connection to the VM instances*** processes kicks in.
- Once the connection is done, the ***SSH command line*** screen will appear.
- To test if the Hive Data Warehouse is created, type ***Hive*** from the command line.
- From here, we can start the ***Hive SQL-like script*** to load the data we have acquired from web crawling.

Using Hive to Query Hadoop Data Warehouse Cluster

using Google Cloud Platform (GCP)


honloong@cluster-wqd7005-m: ~ - Google Chrome
honloong@cluster-wqd7005-m: ~ - Google Chrome
ssh.cloud.google.com/projects/wqd7005datamining-assignments/zones/asia-southeast1-b/instances/cluster-wqd7005-m?authuser=0&...
cluster-wqd7005-m?authuser=0&...

```
197
238
428
566
673
790
Time taken: 0.147 seconds, Fetched: 57 row(s)
hive> SELECT CaseDate, Malaysia FROM asean_confirmed_cases;
OK
1/22/20 0
1/23/20 0
1/24/20 0
1/25/20 3
1/26/20 4
1/27/20 4
1/28/20 4
1/29/20 7
1/30/20 8
1/31/20 8
2/1/20  8
2/2/20  8
2/3/20  8
2/4/20  10
2/5/20  12
2/6/20  12
2/7/20  12
2/8/20  16
2/9/20  16
2/10/20 18
2/11/20 18
2/12/20 18
2/13/20 19
2/14/20 19
2/15/20 22
2/16/20 22
2/17/20 22
2/18/20 22
2/19/20 22
2/20/20 22
2/21/20 22
2/22/20 22
2/23/20 22
2/24/20 22
2/25/20 22
2/26/20 22
2/27/20 23
2/28/20 23
2/29/20 25
3/1/20  29
3/2/20  29
3/3/20  36
3/4/20  50
3/5/20  50
3/6/20  83
3/7/20  93
3/8/20  99
3/9/20  117
3/10/20 129
3/11/20 149
3/12/20 149
3/13/20 197
3/14/20 238
3/15/20 428
3/16/20 566
3/17/20 673
3/18/20 790
Time taken: 0.139 seconds, Fetched: 57 row(s)
hive> 
```

- After the crawl datasets are stored in Hadoop Data Warehouse, we can start querying the dataset by creating a table using the Hive SQL-like script in Hive shell.

```
CREATE TABLE asean_confirmed_cases
(CaseDate STRING,
Thailand INT, Singapore INT, Malaysia INT, Cambodia INT,
Philippines INT, Indonesia INT, Brunei INT, Vietnam INT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE
TBLPROPERTIES("skip.header.line.count"="1");
```

- Next, we load the dataset from the Hadoop Data Warehouse to the Hive Table

```
LOAD DATA INPATH '/user/honloong/t_confirmed_cases_asean.csv'
OVERWRITE INTO TABLE asean_confirmed_cases;
```

- To test out the query from the Hive Table, we can use the following SQL-like scripts:

```
SELECT * FROM asean_confirmed_cases;
Or
SELECT CaseDate, Malaysia FROM asean_confirmed_cases;
```

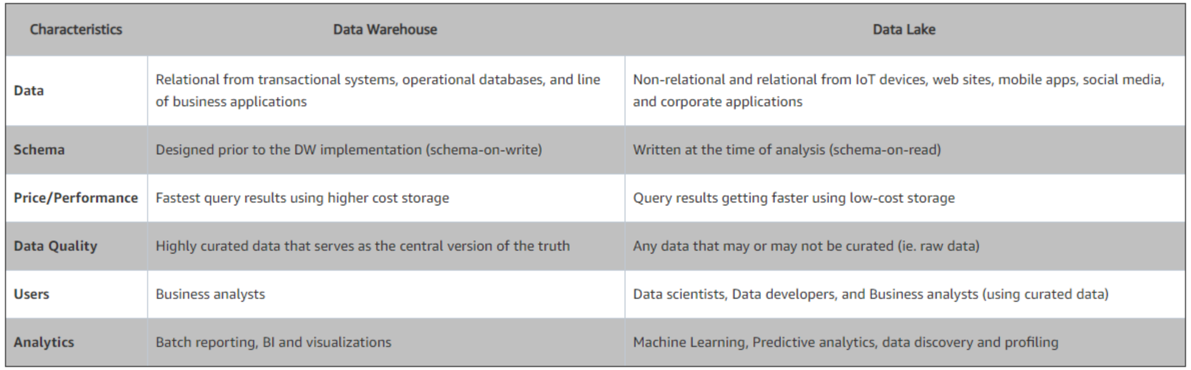
Summary

- To acquire content from a large number of data sources, we need to also prepare the data acquisition and ingestion tools.

- While data scraping can happen in any data array and can be done manually, web scraping or crawling takes place only on the web pages and is performed by special robots i.e. crawlers/scrapers.

- It’s important to start with an agile, flexible and adaptable data repository and can rapidly adapt to changes with the application stacks of choices.

- However, there are different characteristics to consider when choosing the repository to manage the data acquired.



Characteristics	Data Warehouse	Data Lake
Data	Relational from transactional systems, operational databases, and line of business applications	Non-relational and relational from IoT devices, web sites, mobile apps, social media, and corporate applications
Schema	Designed prior to the DW implementation (schema-on-write)	Written at the time of analysis (schema-on-read)
Price/Performance	Fastest query results using higher cost storage	Query results getting faster using low-cost storage
Data Quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (ie. raw data)
Users	Business analysts	Data scientists, Data developers, and Business analysts (using curated data)
Analytics	Batch reporting, BI and visualizations	Machine Learning, Predictive analytics, data discovery and profiling

- Depending on the requirements, typically it will require both data warehouse and data lake to serve different needs and use cases as illustrated in the table above.

Thank You

Azwa Kamaruddin & Hon-Loong Kok (HL)

UNIVERSITY
OF MALAYA