Data Management Ng Jin Bin 1171303325

Q1

Date No.

1. 4V - Volume, Velocity, Variety, Veracity

Volume

- Volume refers to the amount of data (Size of data). Nowadays, with the decreasing storage costs and better storage solution like Hadoop, massive volume of data is getting generated, in range of tera bytes to zeta bytes. In Covid-19 pandemic case, as the people with covid-19 case getting increased, the volume/size of data will increase also. The data generated not only include the number of case, it also contains some medical and personal information of the patients. These data will make the volume of data become massive.

Velocity

- Velocity is the speed at which data is created, stored, analyzed and visualize. In the past, when batch processing was common practice, it was normal to receive an update from the database every ~~pandemic~~ ~~OR~~ night or every week. In big data era, data is created in real-time or near real-time. In Covid-19 pandemic case, Malaysia has some SOP that every people required to follow. People that went to any shop will need to scan the QR code and submit their personal information. The speed of data being created

Variety

- Variety refers to the many sources and types of data. In the past, all data is created in structured data, it neatly fitted in columns and row but those day are over. Nowadays, unstructured data is getting more popular. In Covid-19 pandemic case, there will always have news updating on current situation of Covid-19 globally. The news can be in the form of website, newspaper, article, image, video, dataframe etc. All these form of data is veffering to variety

Veracity

- Veracity refers to the biases, noise, and abnormalities, ambiguities latency in the data. In short, it is related to the correctness of the data. In Covid-19 case, there might be people that purpasely key in false data/fake data to the form after QR code scanning. The correctness of data will be affected.

Q2

Date No.

```
<?xml version="1.0" encoding="UTF-8"?>
    <key id="d0" for="edge" attr.name="weight" attr.type="int">
        <default>100</default>
    </key>
    <graph id="G" edgedefault="undirected">
        <node id="n0"/>
        <node id="n1"/>
        <node id="n2"/>
        <node id="n3"/>
        <edge id="e0" directed="false" source="n0" target="n2"/>
        <edge id="e1" directed="true" source="n2" target="n3">
            <data key="d0">50</data>
        </edge>
        <edge id="e2" directed="true" source="n3" target="n1">
            <data key="d0">100</data>
        </edge>
        <edge id="e3" directed="true" source="n3" target="n0">
            <data key="d0">50</data>
        </edge>
        <edge id="e4" directed="true" source="n3" target="n0">
            <data key="d0">150</data>
        </edge>
        <edge id="e5" directed="false" source="n1" target="n2">
            <data key="d0">150</data>
        </edge>

    </graph>
</graphml>
```

Q3

Default number of replications = 3 → meaning each data nodes need to distribute 3 block.

	Datanode 1	Datanode 2	Datanode 3	Datanode 4	
	A, B, C, D	A, C, D, E	A, B, C, E	B, D, E	
A	✓	✓	✓		3
B	✓		✓	✓	3
C	✓	✓	✓		3
D	✓	✓		✓	3
E		✓	✓	✓	3

Q4

student_info_for_secret_purposes

time stamp 2020/09/17 21:31:25

ID	Name	Gender	Cur-add	home_add	ic-num	passport num
1171303325	Ng Jin Bin	Male	B-26-09, the arc	40, Jln Indah 17/3, Johor Bahru, Johor	990706-01-5163	-