BINUS University

Academic Career: ***Undergraduate / ~~Master~~ / ~~Doctoral~~ *)***	**Class Program:** ***~~International~~/Regular/~~Smart Program/Global Class~~*)***
□ Mid Exam **☑ Final Exam** **□ Short Term Exam** **□ Others Exam : _____________**	**Term : ~~Odd~~/Even/~~Short~~ *)**
☑ Kemanggisan **☑ Alam Sutera** **□ Bekasi** **□ Senayan** **□ Bandung** **□ Malang**	**Academic Year :** **2019 / 2020**

Faculty / Dept.	:	School of Computer Science	Deadline	Day / Date	:	Thursday / Jul 09th, 2020
				Time	:	17:00
Code - Course	:	COMP6579 – Big Data Processing	Class		:	All Classes
Lecturer	:	Team	Exam Type		:	Online

*) *Strikethrough the unnecessary items*

The penalty for CHEATING is DROP OUT!!!

I. Case (100%)

You are a team member of covid19 Task Force in IT Big Data Division and Data Scientist. Your team currently have a Big Data and gather all data about covid19 in Indonesia. When you check you team Big Data repository, you see there are 5 main datasets that are hosted in your repository. This dataset has huge data and it size more than 2TB for each dataset.

The following is the basic information of the datasets:

Covid 19 Suspected Movement Case

This dataset describes about Covid19 Case. This dataset is from surveyor that gather data from covid19 suspected person in movement/mobility before they are quarantined.

Data Source : Apache Hive

Name	Descriptions	Data Type	More Information
Name	Name of suspected person	String	
Gender	Gender	String (Enum)	Male, Female
Age	Age of patient	Int	
PersonStatus	Covid19 suspected category	String(Enum)	*PDP*, *ODP*
PlaceName	Name of place	String	
PlaceType	Place descriptions	String (Enum)	Restaurant, Office, Public, Market, Private
Duration	Duration suspected person take a break in this place (minutes)	Int	
Latitude	Coordinate of place	Double	
Longitude	Coordinate of place	Double	
VisitedOn	Suspected person make visited in date	DateTime	

Example:

- Fepri Putra, Male, 30, ODP, Ps.TanahAbang, Market, 120, 106.8097425, -6.1890043, 10-05-2020 10:10:10
- Ani, Female, 52, ODP, Ps.TanahAbang, Market, 120, 106.8097425, -6.1890043, 10-05-2020 10:10:10

Verified by,

Fidelson Tanzil (D5542) and sent to Department on May 29, 2020

Hospital Covid19 Case

This dataset describes about the status of covid19 patient in care at hospital. This data is updated and inserted daily.
Data Source : csv (HDFS)

Name	Descriptions	Data Type	More Information
Name	Name of patient	String	
Gender	Gender	String(Enum)	Male, Female
Age	Age of patient	Int	
HospitalID	Hospital ID based on master data	Int	
HospitalName	Hospital Name	String	
IsPositive	Covid19 positive status	Boolean	
PatientCovidStatus	Covid19 suspected category	Enum (string)	*ODP*, *PDP*
IsTested	Patient is tested for swab test	Boolean	
TestedDate	Patient swab test date	DateTime	
PatientStatus	Patient status in hospital	String(Enum)	PassedAway, InCare, Healed
CreatedOn	Created record	DateTime	

Example:

- Fepri Putra, Male, 30, 1, RS.Sulianti Saroso, True, PDP, True, 10-05-2020 10:10:10, InCare, 10-05-2020 10:10:10

CCTV Data

This dataset store video cctv from the place where there is a potential of social crowd (traditional market, social district, business district). This dataset is useful for monitoring social distancing program at crowded place.
Data Source: Stream

Name	Descriptions	Data Type	More Information
ID	ID cctv	Int	
Latitude	Coordinate of place	Double	
Longitude	Coordinate of place	Double	
VideoData	CCTV video data	Binary Video Format	
CreatedOn	Crated record	DateTime	

Hospital Medical Item Needs

The dataset describes about medical item that hospital needs.
Data Source : csv (HDFS)

Name	Descriptions	Data Type	More Information
HospitalID	Hospital ID	Int	
HospitalName	Hospital Name	String	
MedicalItemName	Medical item that hospital need	String	Surgical Mask, Hazmat Suite, Disinfectant
Amount	Amount item that hospital needs	Double	
CreatedOn	Covid19 positive status	Boolean	

Example:

- 1, RS.Sulianti Saroso, Surgical Mask, 1000, 10-05-2020 10:10:10
- 1, RS.Sulianti Saroso, Hazmat Suite, 1000, 10-05-2020 10:10:10

Verified by,

Fidelson Tanzil (D5542) and sent to Department on May 29, 2020

Social Assistance Distributions

The dataset describes about social assistance distribution in Indonesia.
Data Source: Apache Hive

Name	Descriptions	Data Type	More Information
SubDiscrictID	SubDistrict ID from master data	String	
SubDisctrictName	SubDistrict Name	String	
DistrictID	District ID from master data.	String	
DistrictName	District Name	String	
ProvinceID	Province ID from master data	String	
ProvinceName	Province Name	String	
TotalAllocation	Amount that distribution in Rupiah	Decimal (12,2)	
DateAllocation	Date of allocation	DateTime	
Institution	Institution that distribute aid	String	

Example:

- DKI1, Keb.Baru, DKI1, JakartaSelatan, 1, DKI Jakarta, 10000000, 10-05-2020 10:10:10, ABC.PT
- DKI1, Keb.Baru, DKI1, JakartaSelatan, 1, DKI Jakarta, 10000000, 10-05-2020 10:10:10, HambaTuhan

You as Data Scientist in covid19 task force has some tasks:

1. You should choose **min. 2 from 5** datasets (you can choose all datasets) and make some analytics (**at least 3**) from dataset that you have chosen. Please **write** the sample data that you use for each dataset and **describe** what analysis you will create based on the dataset you choose! (30%)
2. What type analytics per analysis that you will create (descriptive, diagnostic, predictive, or prescriptive) and explain why! (20%)
3. Please **explain** analytics flow of your analysis from the data source until you visualize it! (20%)
4. What software technology that will help you to process the data in your analytics flow, **explain** it why you choose that! (10%)
5. Please **explain** about diagram/chart per analysis that you will choose (scatter, bar, etc.) and how does it fit with your analysis! (20%)

Notes for this task:

1. You could add more another dataset or field that support your analysis but in your answers you should using this 5 dataset as your main analysis.
2. If you add some features (dataset/field) you should explain about how to get that datasource/field, what is data type and some basic information about that features.
3. You could create output schema from this dataset to accommodate your analysis and please describe some basic information about your output schema. Don't forget you should describe it too in analytics flow.
4. You could add more assumption, but please give basic information about you assumptions.
5. ODP (Orang Dalam Pengawasan) => "People Under Surveillance"
6. PDP (Pasien Dalam Pegawasan) => "Patient Under Suveillance"
7. Date time format "dd-MM-yyyy hh:mm:ss"

-- Good Luck--

Verified by,

Fidelson Tanzil (D5542) and sent to Department on May 29, 2020