TARUMT

TUNKU ABDUL RAHMAN UNIVERSITY OF
MANAGEMENT AND TECHNOLOGY

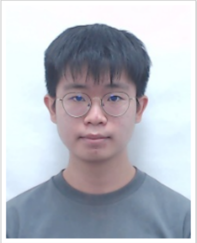
BMDS2013 Data Engineering

CLO2 ASSIGNMENT 202505

Team Reference : **S1G2-1**

Programme / Group : **RDS2 S1 / G2**

Student Name	Student ID	Contribution	Signature
LIM FENG ZHI	24WMR08003	28%	
NG CHIAO HAN	24WMR08010	25%	
THO HUI YEE	24WMR08035	31%	
THAM ZHEN HERN	24WMR08034	16%	
		%	
	TOTAL (must sum to 100%)	100%	

LIM FENG ZHI	NG CHIAO HAN	THO HUI YEE	THAM ZHEN HERN
			

1.0 Introduction

1.1 Brief overview of the selected SDG use case

This project investigates how carbon dioxide (CO_2) emissions are a driving factor behind the increasing intensity and frequency of extreme weather events in China, France, Germany, India, Japan, Russia, the United Kingdom, and the United States. It directly aligns with Sustainable Development Goal (SDG) 13: Climate Action, which emphasizes the urgent need to combat climate change and its far-reaching effects. Rising CO_2 emissions intensify the greenhouse effect, leading to global temperature increases that disrupt natural climate systems. These disruptions manifest as more frequent and severe heatwaves, floods, storms, and unpredictable seasonal shifts. Both developed and developing nations are experiencing these impacts, with vulnerable populations, economies, and infrastructures suffering the most as CO_2 emissions continue to accelerate climate instability.

1.2 Motivation and real-world relevance

The motivation for this project comes from the urgent challenge that rising CO_2 emissions are directly driving extreme weather events around the world. The selected countries—China, France, Germany, India, Japan, Russia, the United Kingdom, and the United States—are among the largest emitters of CO_2, and at the same time, they are increasingly affected by severe climate impacts. High emission levels intensify global warming, which disrupts natural climate systems and results in stronger heatwaves, heavier rainfall, destructive floods, and powerful storms. This problem is highly relevant in the real world because it affects multiple aspects of society:

- **Health impacts**: Extreme heat and flooding endanger vulnerable groups, including the elderly, children, and outdoor workers.
- **Agricultural and economic disruption**: Farming and industrial activities are destabilized by unpredictable and damaging weather.
- **Urban challenges**: Cities face greater risks of flooding, storm damage, and heat stress, putting pressure on infrastructure and resources.

By focusing on the direct role of CO_2 emissions in causing extreme weather, this project highlights an urgent global issue with immediate consequences for public health, food security, economic stability, and the resilience of both rural and urban communities.

1.3 Objectives of the pipeline

a) To collect and process weather records and CO_2 emission data for China, France, Germany, India, Japan, Russia, United Kingdom, and the United States.

b) To analyze extreme weather patterns and explore their relationship with emission levels in each country.

c) To identify high-risk countries where high emissions and frequent severe weather converge, increasing vulnerability to climate impacts.

d) To develop a scalable data pipeline that integrates emissions and weather datasets, enabling real-time monitoring, trend analysis, and early-warning indicators.
e) To generate visualizations and reports that support policymakers, disaster response agencies, and climate planners in making informed, data-driven decisions.

2.0 Task 1: Raw Data Streaming

Use Case:	Weather Data
Raw Data Source:	https://www.kaggle.com/datasets/guillemservera/global-daily-climate-data/data https://carbonmonitor.org.cn/
Kafka Topic:	weather_data co2_data
Output Path:	hdfs://localhost:9000/user/student/data_store/processed_data/cleanned_data/cleaned_weather_parquet hdfs://localhost:9000/user/student/data_store/processed_data/cleanned_data/cleaned_co2_parquet

List of Python (`.py`) files:

Python file(s)	Author
producer.py	THO HUI YEE
carbon_producer.py	THO HUI YEE
weather_producer.py	THO HUI YEE
consumer.py	THO HUI YEE
consumer_config.py	THO HUI YEE
carbon_consumer.py	THO HUI YEE
weather_consumer.py	THO HUI YEE
run_carbon_consumer.py	THO HUI YEE
run_weather_consumer.py	THO HUI YEE
run_carbon_producer.py	LIM FENG ZHI
run_weather_producer.py	LIM FENG ZHI

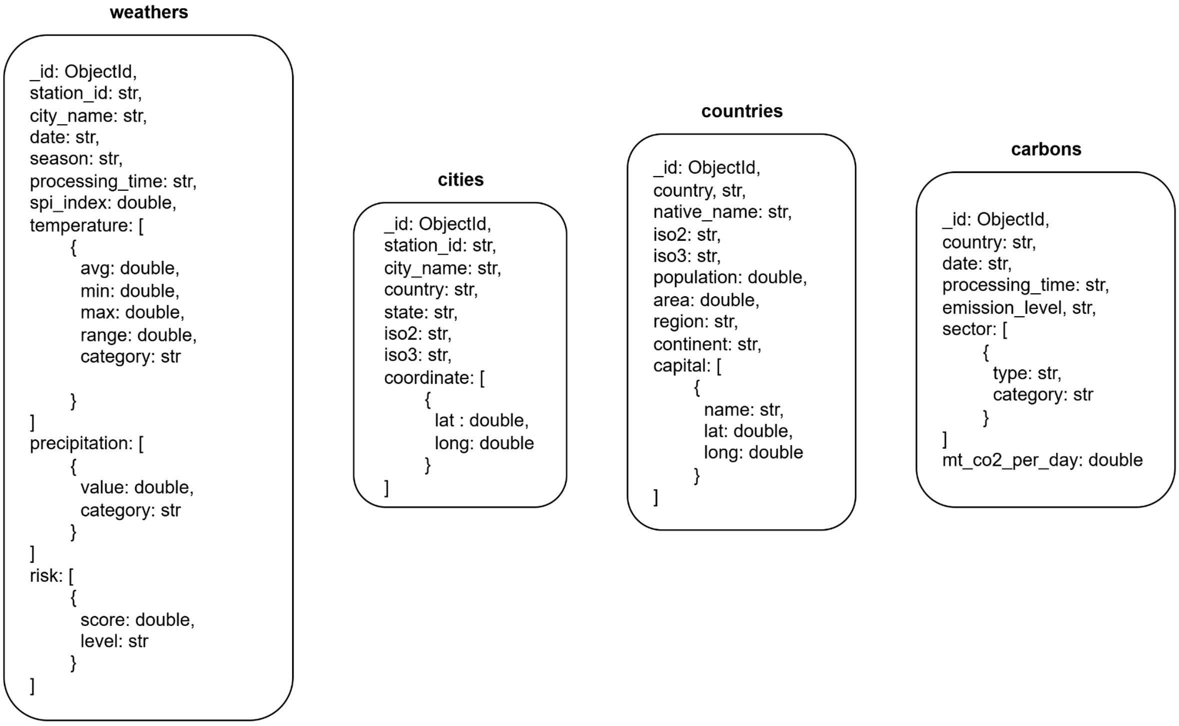
3.0 Task 2: Data Processing

List of Python (`.py`) files:

Python file(s)	Author
preprocessor.py	LIM FENG ZHI
data_preprocessing_config.py	THAM ZHEN HERN
carbon_preprocessor.py	THAM ZHEN HERN
weather_preprocessor.py	THAM ZHEN HERN, LIM FENG ZHI

4.0 Task 3: Document Database

4.1 MongoDB Data Model Design



weathers

```
_id: ObjectId,
station_id: str,
city_name: str,
date: str,
season: str,
processing_time: str,
spi_index: double,
temperature: [
        {
          avg: double,
          min: double,
          max: double,
          range: double,
          category: str

        }
]
precipitation: [
        {
          value: double,
          category: str
        }
]
risk: [
        {
          score: double,
          level: str
        }
]
```

cities

```
_id: ObjectId,
station_id: str,
city_name: str,
country: str,
state: str,
iso2: str,
iso3: str,
coordinate: [
        {
          lat : double,
          long: double
        }
]
```

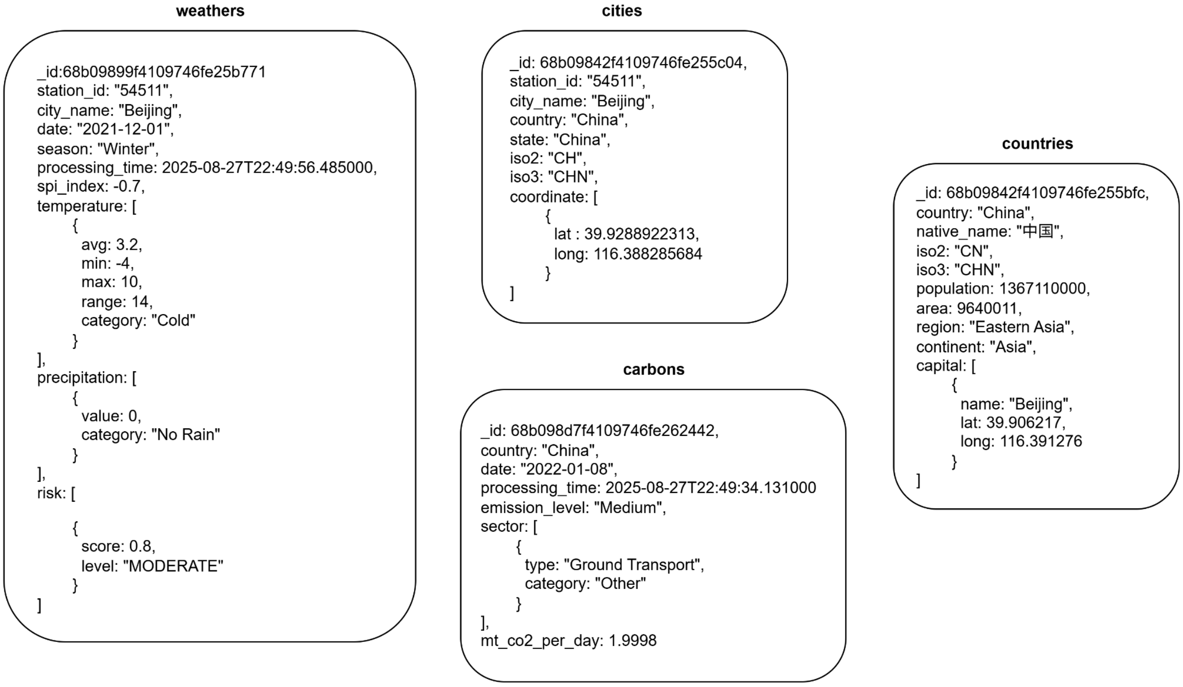
countries

```
_id: ObjectId,
country, str,
native_name: str,
iso2: str,
iso3: str,
population: double,
area: double,
region: str,
continent: str,
capital: [
        {
          name: str,
          lat: double,
          long: double
        }
]
```

carbons

```
_id: ObjectId,
country: str,
date: str,
processing_time: str,
emission_level, str,
sector: [
        {
          type: str,
          category: str
        }
]
mt_co2_per_day: double
```

4.2Diagram with Example of Values in MongoDB



weathers

```
_id:68b09899f4109746fe25b771
station_id: "54511",
city_name: "Beijing",
date: "2021-12-01",
season: "Winter",
processing_time: 2025-08-27T22:49:56.485000,
spi_index: -0.7,
temperature: [
      {
        avg: 3.2,
        min: -4,
        max: 10,
        range: 14,
        category: "Cold"
      }
],
precipitation: [
      {
        value: 0,
        category: "No Rain"
      }
],
risk: [

      {
        score: 0.8,
        level: "MODERATE"
      }
]
```

cities

```
_id: 68b09842f4109746fe255c04,
station_id: "54511",
city_name: "Beijing",
country: "China",
state: "China",
iso2: "CH",
iso3: "CHN",
coordinate: [
      {
        lat : 39.9288922313,
        long: 116.388285684
      }
]
```

countries

```
_id: 68b09842f4109746fe255bfc,
country: "China",
native_name: "中国",
iso2: "CN",
iso3: "CHN",
population: 1367110000,
area: 9640011,
region: "Eastern Asia",
continent: "Asia",
capital: [
      {
        name: "Beijing",
        lat: 39.906217,
        long: 116.391276
      }
]
```

carbons

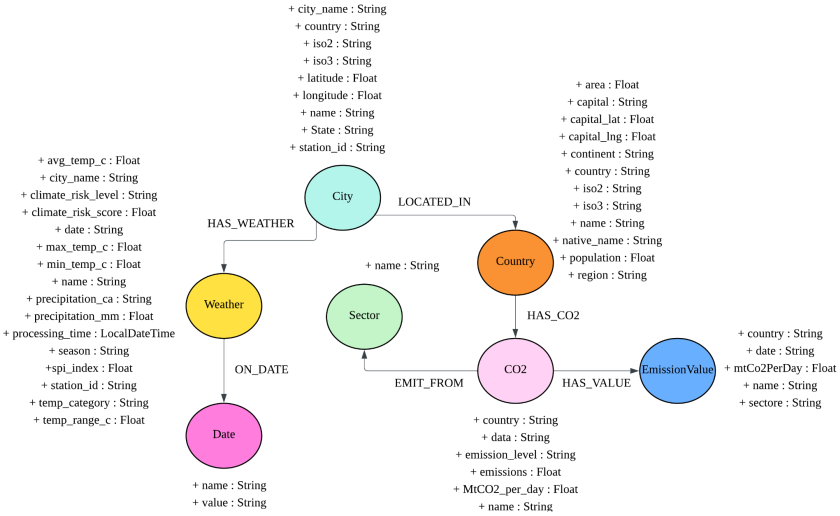
```
_id: 68b098d7f4109746fe262442,
country: "China",
date: "2022-01-08",
processing_time: 2025-08-27T22:49:34.131000
emission_level: "Medium",
sector: [
      {
        type: "Ground Transport",
        category: "Other"
      }
],
mt_co2_per_day: 1.9998
```

4.3 List of Python (`.py`) files:

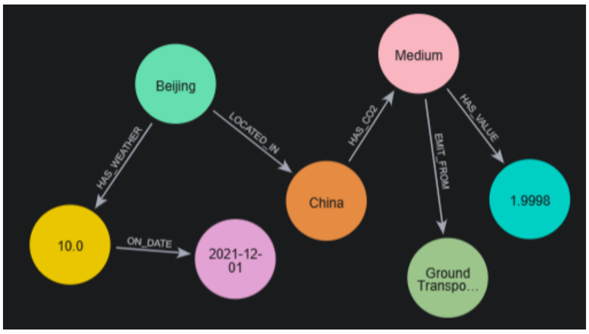
Python file(s)	Author
utilis_mongo.py	THO HUI YEE
mongo_query.py	THO HUI YEE

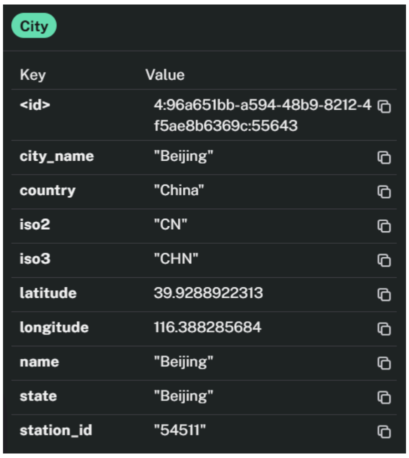
5.0 Graph Database

5.1 Graph Model Design

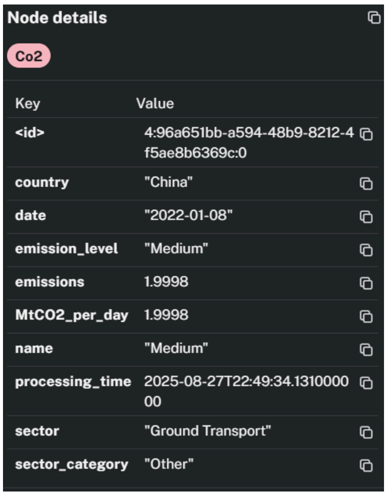


5.2 Diagram with Example of Values in Neo4j



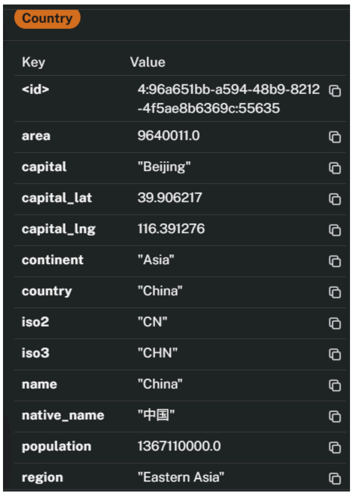

City

Key	Value
<id>	4:96a651bb-a594-48b9-8212-4f5ae8b6369c:55643
city_name	"Beijing"
country	"China"
iso2	"CN"
iso3	"CHN"
latitude	39.9288922313
longitude	116.388285684
name	"Beijing"
state	"Beijing"
station_id	"54511"

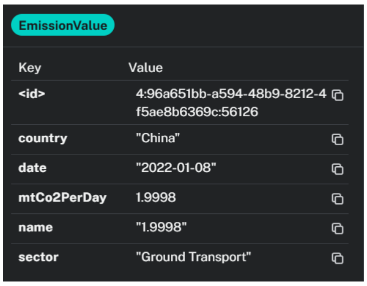

Node details

Co2

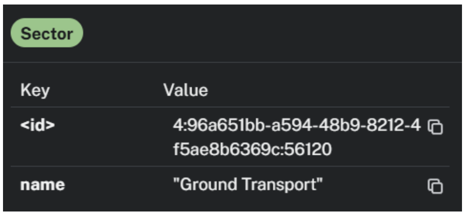
Key	Value
<id>	4:96a651bb-a594-48b9-8212-4f5ae8b6369c:0
country	"China"
date	"2022-01-08"
emission_level	"Medium"
emissions	1.9998
MtCO2_per_day	1.9998
name	"Medium"
processing_time	2025-08-27T22:49:34.131000000
sector	"Ground Transport"
sector_category	"Other"


Country

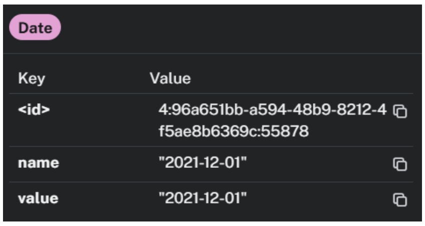
Key	Value
<id>	4:96a651bb-a594-48b9-8212-4f5ae8b6369c:55635
area	9640011.0
capital	"Beijing"
capital_lat	39.906217
capital_lng	116.391276
continent	"Asia"
country	"China"
iso2	"CN"
iso3	"CHN"
name	"China"
native_name	"中国"
population	1367110000.0
region	"Eastern Asia"


EmissionValue

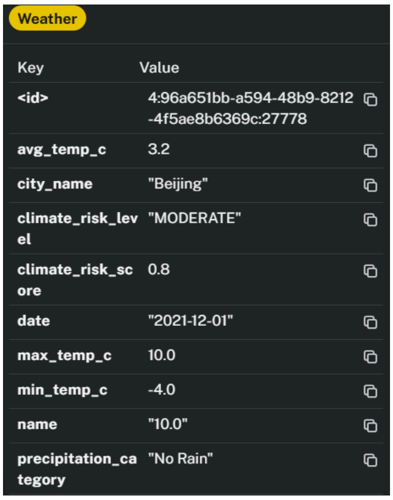
Key	Value
<id>	4:96a651bb-a594-48b9-8212-4f5ae8b6369c:56126
country	"China"
date	"2022-01-08"
mtCo2PerDay	1.9998
name	"1.9998"
sector	"Ground Transport"


Sector

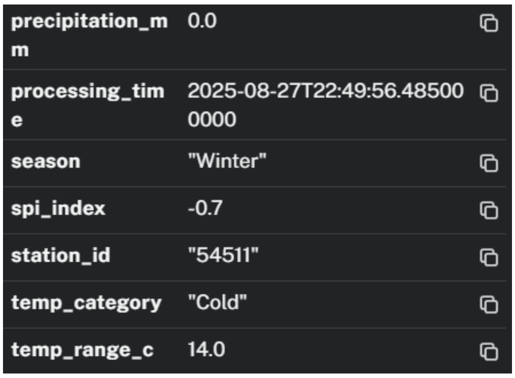
Key	Value
<id>	4:96a651bb-a594-48b9-8212-4f5ae8b6369c:56120
name	"Ground Transport"


Date

Key	Value
<id>	4:96a651bb-a594-48b9-8212-4f5ae8b6369c:55878
name	"2021-12-01"
value	"2021-12-01"


Weather

Key	Value
<id>	4:96a651bb-a594-48b9-8212-4f5ae8b6369c:27778
avg_temp_c	3.2
city_name	"Beijing"
climate_risk_level	"MODERATE"
climate_risk_score	0.8
date	"2021-12-01"
max_temp_c	10.0
min_temp_c	-4.0
name	"10.0"
precipitation_category	"No Rain"



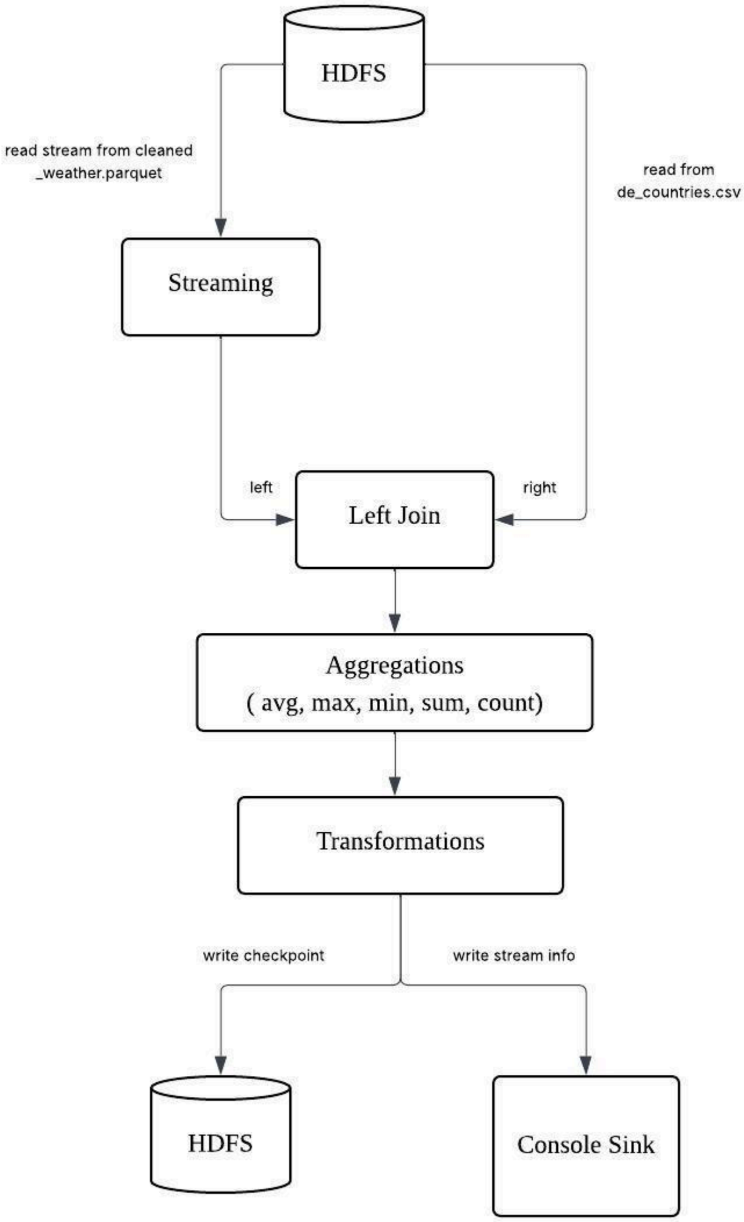
precipitation_mm	0.0
processing_time	2025-08-27T22:49:56.485000000
season	"Winter"
spi_index	-0.7
station_id	"54511"
temp_category	"Cold"
temp_range_c	14.0

5.3 List of Python (`.py`) files:

Python file(s)	Author
utilis_neo4j.py	NG CHIAO HAN
neo4j_query.py	NG CHIAO HAN

6.0 Spark Structured Streaming

6.1 Structured Streaming Workflow Diagram


HDFS
read stream from cleaned _weather.parquet
read from de_countries.csv
Streaming
left
Left Join
right
Aggregations (avg, max, min, sum, count)
Transformations
write checkpoint
write stream info
HDFS
Console Sink

6.2 List of Python (`.py`) files:

Python file(s)	Author
streamer.py	LIM FENG ZHI
climate_risk_streamer.py	LIM FENG ZHI

7.0 Utility Class

7.1 List of Python(.py)files:

Python file(s)	Author
spark_manager.py	LIM FENG ZHI
input_country_manager.py	LIM FENG ZHI

7.2 Task Complete By

	LIMFENGZHI	THO HUI YEE	NG CHIAO HAN	THAM ZHEN HERN
Choose Topic Identify Objective (5%)	1%	1%	2%	1%
Task 1 (18%)	2% change producer	11% consumer	4% find raw data	1% find raw data
Task 2 (18%)	5% do preprocessor class			13%
Task 3 (18%)		18%		
Task 4 (18%)			18%	
Task 5 (18%)	18%			
Combined Task (5%)	2% Combined all the code	1 % test run	1 % test run	1% test run
Total (100%)	28%	31%	25%	16%