DATABASE 415 Final Project – Phase1

(1) Team member

- Da Chen (dchen78)
- Xiaoxiao Liu (xliu91)

(2) Domain:

China Regional Air Pollution and Potential Major Causes

- Real-time
- Historical
- (Forecast)
- Geographical comparisons
- Influencing Factors (Major)
    - Wind Dispersion
    - Precipitation
    - Topography
    - Number of automobile ownership
    - Number of local factories
    - Holidays

We hope through this database, we can provide real-time data of air quality for different cities in China. China's air quality has been a long-discussing topic yet without many definitive conclusions. While collecting and displaying visualized rea-time data on air quality, our website also has functionalities such as historical comparison, city comparison, and searching for peak value. Following the mainstream method, we assume the level of pollutant PM2.5 can stand for haze level.

In addition to solely focusing on air quality, we will also provide some basic analytical relationship between air pollution levels and the commonly-believed factors: natural environment, source of emission and human behavior.

For the natural environment aspect, we collect and display data related to wind condition, humidity, precipitation, temperature, air pressure and landform for each city. For emission source, we take two indexes to briefly represent the emission level of pollutant: the number of car ownership and number of factories in different cities. Human behavior is a more controversial area in terms of effect on air pollution such as haze. We plan to only tentatively compare air pollution variation between workdays and holidays, during which usually more automobiles are used, and more emissions exists.

(3) Answer what questions (15 min)

- How's the air quality for now in different cities
- How's air quality changed in the past given time
- The hourly forecasting of air quality for each city
- In the given time, the cities with the highest and the lowest air pollution
- Find the cities whose air qualities is in the level of a given range (ex 'mid-level')

- The proportion of days over the year a city is polluted in a certain level
- Which city has the most number of hazy days throughout the year
- In the history of existing data, which city has the highest haze index
- For each haze level, which cities has the most number of hazy days of that level
- For each hour (or one certain hour of a day), the air quality index and wind speed
- For each hour (or one certain hour of a day), the air quality index and humidity
- For each hour (or one certain hour of a day), the air quality index and precipitation
- For each hour (or one certain hour of a day), the air quality index and air pressure
- For each hour (or one certain hour of a day), the air quality index and temperature
- In which landform(s) the air pollution is more serious
- What is the relationship between factory number and average AQI
- What is the relationship between number of automobiles and average AQI
- In which holiday(s), the air pollution is more serious

(4) Relation data model

(relationships see attachment: Relational Table)

Schemas

- City

City_code	City_name	Province	Longitude	Latitude

- Pollution

City_name	Date	Time	Quality	Pm2_5	Pm2_5_24	AQI	Primary_pollutant

- Pm2_5_level

Level	StartNum

- Wind

City_name	Date	Time	Speed

- Humidity

City_name	Date	Time	Humidity

- Precipitation

City_name	Date	Time	precipitation

- Air pressure

City_name	Date	Time	Air_Pressure

- Temperature

City_name	Date	Time	Temperature

- City_topography

City_name	Landform

- NumberOfCarOwnership

City_name	Year	CarNumber

- NumberOfFactories

City_name	Year	Factories

- Festival

Date	Festival_Name

(5) SQL

(see in attachment: phase1-sql)

(6) Data Source

API

- City Air Quality API (Real-time synchronization): http://pm25.in/
- City Weather API (Real-time synchronization): http://openweathermap.org
- Festival API: https://www.timeanddate.com/holidays/

Dataset

- City -landform: http://data.stats.gov.cn/search.htm?s=%E5%9C%B0%E5%BD%A2
- City-automobile ownership http://data.stats.gov.cn/easyquery.htm?cn=C01&zb=A0G0I&sj=2014
- City-factory amount http://data.stats.gov.cn/search.htm?s=%E5%B7%A5%E5%8E%82%E6%95%B0%E9%87%8F

(7) Reports

We will display visualized data with diagrams and graphs

- Data visualization: E-CHARTS (tool)
  - Bar graphs/Histograms
  - Pie charts
  - Line graphs
  - Maps
  - Thermodynamic diagrams

(8) Advanced topics

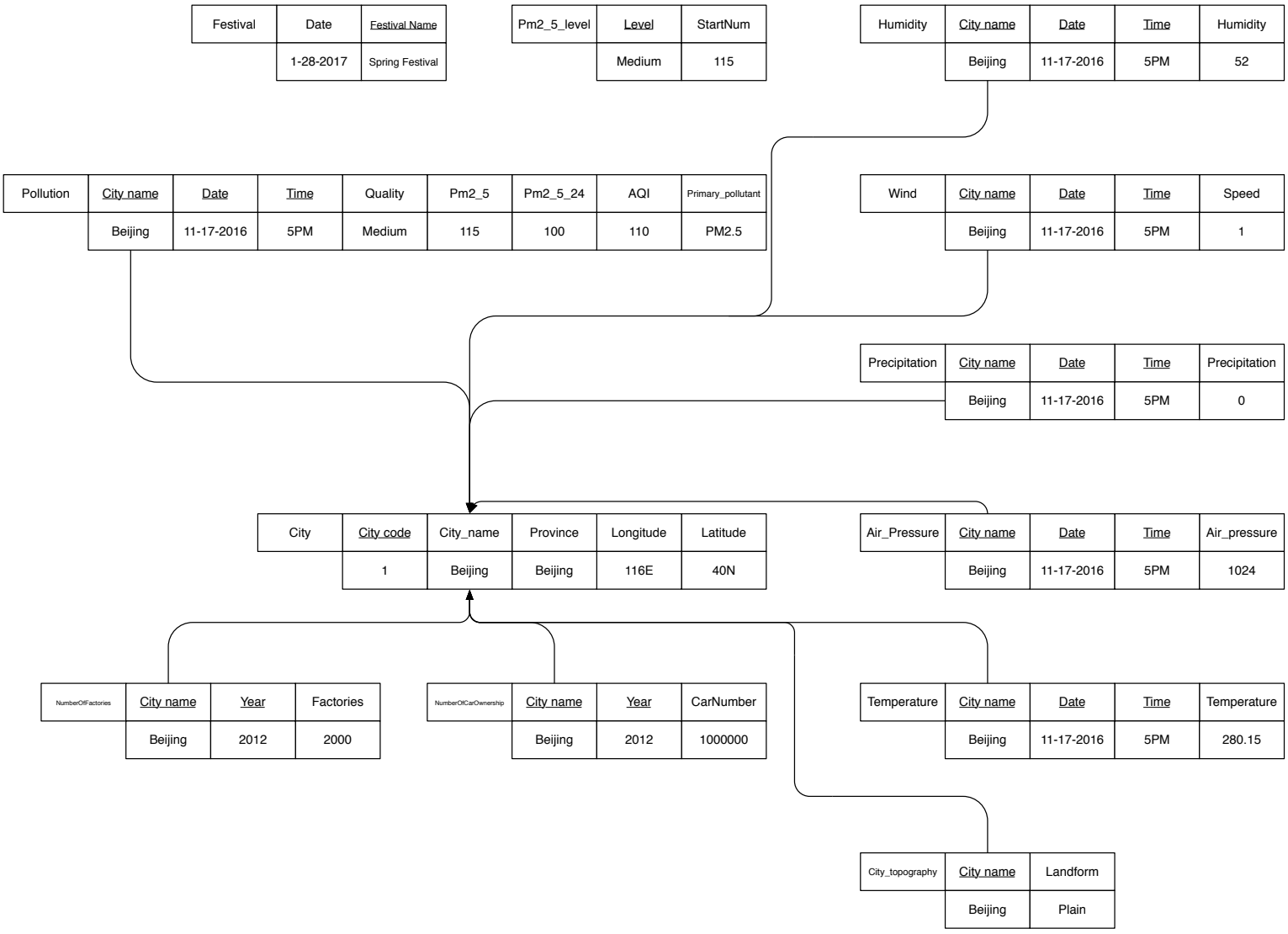
We plan to focus on the following advanced topics:

- Advanced GUI using E-charts
- Real-time data synchronization through online APIs
- (JDBC/Distributed System…)

The data visualization and real-time data synchronization will be or main features.

(9) Database platform

Platform: 5.7 MySQL Community Server (GPL) on gradx server
(Intel(R) Xeon(R) CPU E5-2420 0 @ 1.90GHz; 32G memory)


Festival
Date
Festival Name
1-28-2017
Spring Festival
Pm2_5_level
Level
StartNum
Medium
115
Humidity
City name
Date
Time
Humidity
Beijing
11-17-2016
5PM
52
Pollution
City name
Date
Time
Quality
Pm2_5
Pm2_5_24
AQI
Primary_pollutant
Beijing
11-17-2016
5PM
Medium
115
100
110
PM2.5
Wind
City name
Date
Time
Speed
Beijing
11-17-2016
5PM
1
Precipitation
City name
Date
Time
Precipitation
Beijing
11-17-2016
5PM
0
City
City code
City_name
Province
Longitude
Latitude
1
Beijing
Beijing
116E
40N
Air_Pressure
City name
Date
Time
Air_pressure
Beijing
11-17-2016
5PM
1024
NumberOfFactories
City name
Year
Factories
Beijing
2012
2000
NumberOfCarOwnership
City name
Year
CarNumber
Beijing
2012
1000000
Temperature
City name
Date
Time
Temperature
Beijing
11-17-2016
5PM
280.15
City_topography
City name
Landform
Beijing
Plain

In the given time, the cities with the highest and the lowest air pollution

```
select City_name, Pm2_5
from Poluttion
where Pm2_5 = (
    select max(Pm2_5)
    From Pollution
    where Date between 2016-01-01 and 2016-10-31
)
```

For each haze level, which cities has the most number of hazy days of that level

```
select Base2.*
from
    (select City_name, Level, count(distinct Date) cnt
    from (select p.*, (case
                            when
                                p.Pm2_5 >=(select startNum from Pm2_5_level where lev
el = 'light polluted')
                                and p.Pm2_5 <(select startNum from Pm2_5_level where
level = 'mid polluted')
                            then
                                'light polluted'
                            when
                                p.Pm2_5 >=(select startNum from Pm2_5_level where lev
el = 'mid polluted')
                                and p.Pm2_5 <(select startNum from Pm2_5_level where
level = 'heavy polluted')
                            then
                                'mid polluted'
                            when
                                p.Pm2_5 >=(select startNum from Pm2_5_level where lev
el = 'heavy polluted')
                            then
                                'heavy polluted'
                        end case) as Level
          from Pollution p
          where p.Pm2_5 >= (
              select
              from Pm2_5_level
              where level = 'light polluted')) Base3
```

```
    gorup by City_name, Level) Base4,
    (select Level, max(cnt) max
    from( select City_name, Level, count(distinct Date) cnt
            from (select p.*, (case
                                    when
                                        p.Pm2_5 >=(select startNum from Pm2_5_level w
here level = 'light polluted')
                                        and p.Pm2_5 <(select startNum from Pm2_5_leve
l where level = 'mid polluted')
                                    then
                                        'light polluted'
                                    when
                                        p.Pm2_5 >=(select startNum from Pm2_5_level w
here level = 'mid polluted')
                                        and p.Pm2_5 <(select startNum from Pm2_5_leve
l where level = 'heavy polluted')
                                    then
                                        'mid polluted'
                                    when
                                        p.Pm2_5 >=(select startNum from Pm2_5_level w
here level = 'heavy polluted')
                                    then
                                        'heavy polluted'
                                end case) as Level
                  from Pollution p
                  where p.Pm2_5 >= (
                      select
                      from Pm2_5_level
                      where level = 'light polluted')) Base
            gorup by City_name, Level) Base1
    group by Level) Base2
where Base4.Level = Base2.Level
  and Base4.cnt = Base2.max
```