Data:

4 Pollutants - [*Chloroform, Benzene, Lead PM2.5 LC, Arsenic PM2.5 LC*]

All the cities in [***Texas***] State

Period of Data: Years [*2009 to 2014*]

Data Acquisition:

Converted the data frames from R into CSV files, and then loaded into neo4j:

#Load the City Data into the node: "City"

```
load csv with headers from "file:C:/Users/Suman/Documents/city-data.csv" as city create (a:City {cityname:city.city, statename: city.state })
```

#Load the Pollutant Data into the node: "Pollutant"

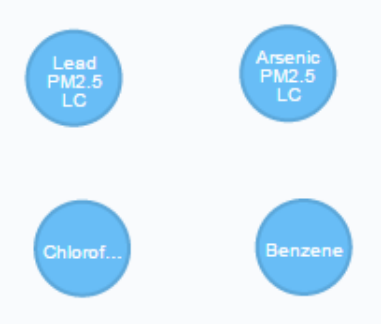
```
load csv with headers from "file:C:/Users/Suman/Documents/pollutant-data.csv" as pollutant create (b:Pollutant {code:pollutant.code, name: pollutant.name })
```

#Establish relationship between City <--- Observations ---> Pollutant

```
load csv with headers from "file:C:/Users/Suman/Documents/observation-data.csv" as observation match (a: City {cityname: observation.city, statename: observation.state}), (b: Pollutant {code: observation.code}) create (a) - [r:Observations {year: observation.year, measurement: observation.measurement}] -> (b)
```

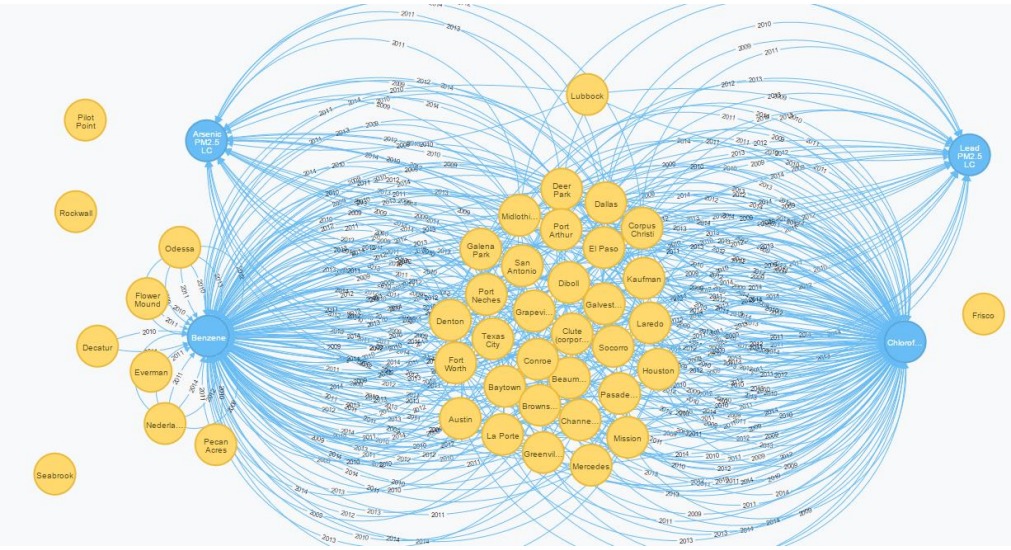
match (a:**City**) return (a)

match (b:**Pollutant**) return (b)



(a:City) - [:**Observations**] -> (b: Pollutant) ➔ Observations has '**year**' and '**measurement**'

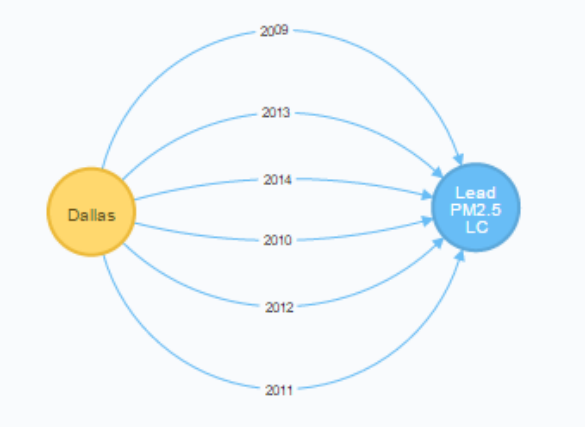
match (n) return (n)



Data Analysis:

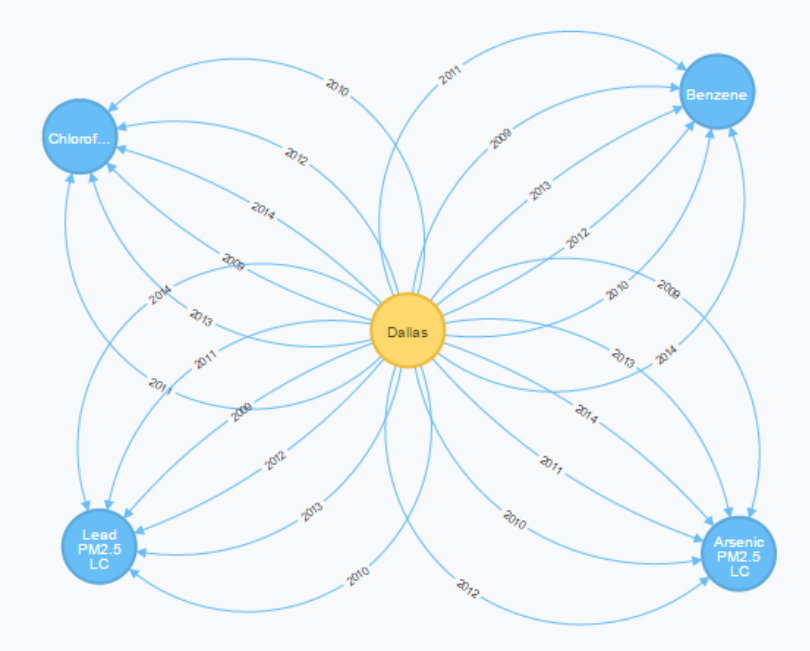
#Dispaly the pollutant data 'Lead PM2.5 LC' in Dallas:

```
match p = (a:City) - [r1:Observations] -> (b:Pollutant) where a.cityname='Dallas' and a.statename='Texas'
and b.name='Lead PM2.5 LC'  return p
```



*#Explore **All** of Dallas Polllution Data*

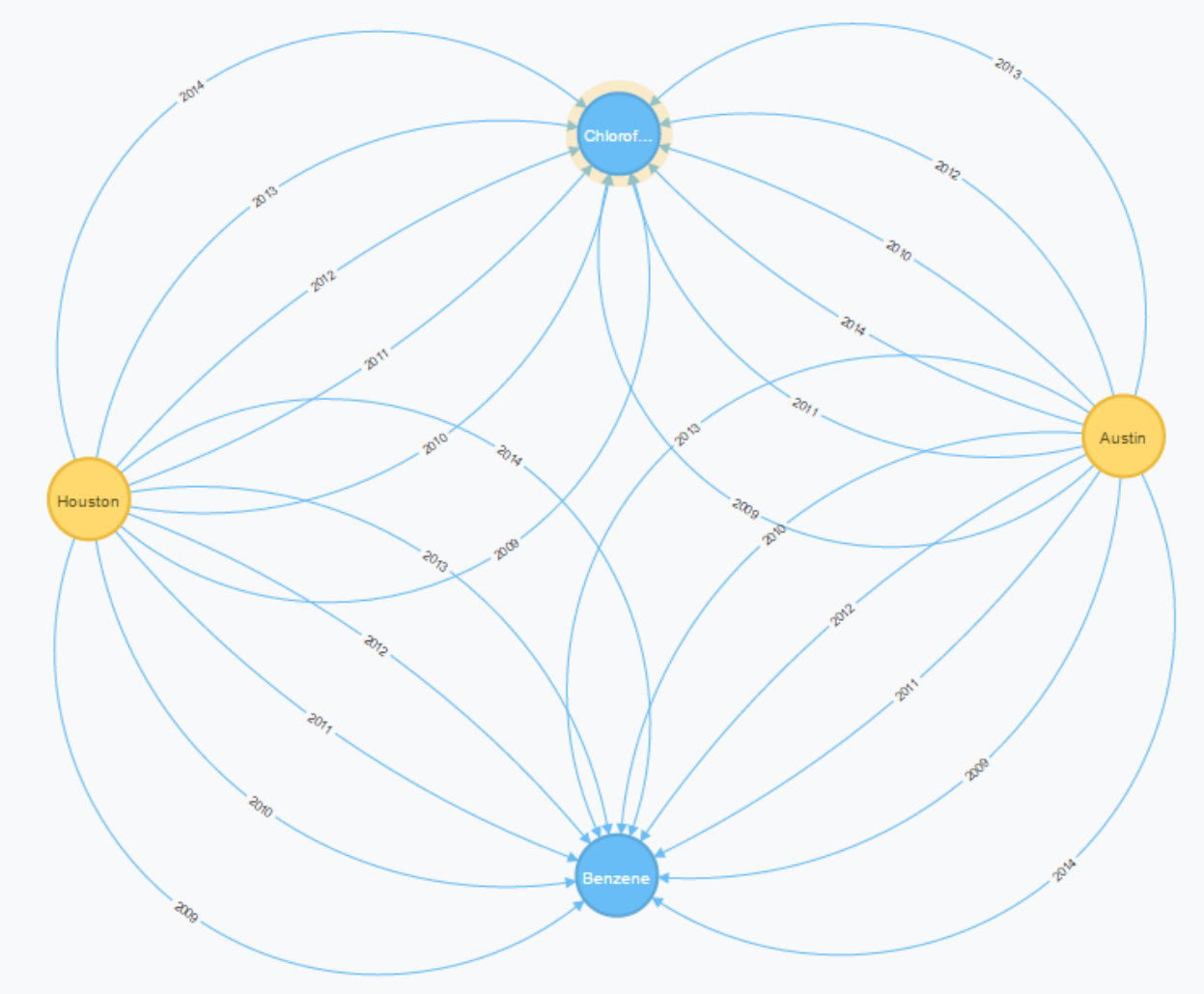
```
match p = (a:City) - [r1:Observations ] -> (b:Pollutant) where a.cityname = 'Dallas'  and
a.statename='Texas'   return a,b
```



Dallas shows the presence of all **4** pollutants for all 6 years

#Explore all of the Austin, Houston Polllution Data

```
match p = (a:City) - [r1:Observations ] -> (b:Pollutant) where a.cityname IN ['Austin','Houston']  and
a.statename='Texas'   return a,b
```



Unlike Dallas, the cities Austin and Houston shows the presence of **2** pollutants for all 6 years.

#Dispaly top 5 Cities in Texas with highest measurement for the pollutant data - 'Lead PM2.5 LC' :

```
match p = (a:City) - [r1:Observations] -> (b:Pollutant) where a.statename='Texas' and b.name='Lead PM2.5 LC' return  a.cityname as City, max(r1.measurement) as Lead_PM2_5_Annual_Measure order by Lead_PM2_5_Annual_Measure desc limit 5
```

City	Lead_PM2_5_Annual_Measure
Corpus Christi	0.004531
El Paso	0.0044775
Dallas	0.003573
Midlothian	0.003503
Lubbock	0.002942

Returned 5 rows in 66 ms.

*#Display ALL the Cities with the presence of **all 4 hazardous pollutants** in the year **2014***

```
match (a:City) - [r1:Observations  { year: '2014'} ] -> (b:Pollutant) with a.cityname as City, count(r1) as PollutantCount  where PollutantCount = 4 return  City, PollutantCount;
```

City	PollutantCount
Corpus Christi	4
Midlothian	4
Dallas	4
Deer Park	4
El Paso	4

Returned 5 rows in 74 ms.

*#Display the Cities in Texas with **absolutely NO presence** of these 4 hazardous pollutants in the last 6 years [2009-14]*

```
match (a:City) where NOT (a)-[:Observations]->() return a
```

--Returns 24 Cities out of 62 ==> which is **38.7% of Texas**.