Group B - Assginment 3

Problem Statememt

Visualize the data using Python libraries matplotlib, seaborn by plotting the graphs for assignment number 1 and 2

In [1]:

```
import pandas as pd
import matplotlib.pyplot as plt
```

In [ ]:

```

```

Read data from CSV file

In [2]:

```
A  =  pd.read_csv("Airquality.csv")
A.head(12)
```

Out[2]:

	Unnamed: 0	Ozone	Solar.R	Wind	Temp	Month	Day	Humidity
0	1	41.0	190.0	7.4	67	5	1	High
1	2	36.0	118.0	8.0	72	5	2	High
2	3	12.0	149.0	12.6	74	5	3	Low
3	4	18.0	313.0	11.5	62	5	4	NaN
4	5	NaN	NaN	14.3	56	5	5	High
5	6	28.0	NaN	14.9	66	5	6	High
6	7	23.0	299.0	8.6	65	5	7	High
7	8	19.0	99.0	13.8	59	5	8	Low
8	9	8.0	19.0	20.1	61	5	9	NaN
9	10	NaN	194.0	8.6	69	5	10	Medium
10	11	7.0	NaN	6.9	74	5	11	Medium
11	12	16.0	256.0	9.7	69	5	12	High

In [3]:

```
A.isnull().sum()
```

Out[3]:

```
Unnamed: 0     0
Ozone         37
Solar.R        7
Wind           2
Temp           0
Month          0
Day            0
Humidity       8
dtype: int64
```

Data Cleaning

In [4]:

```
df = A.drop("Unnamed: 0",axis=1)
df.head(6)
```

Out[4]:

	Ozone	Solar.R	Wind	Temp	Month	Day	Humidity
0	41.0	190.0	7.4	67	5	1	High
1	36.0	118.0	8.0	72	5	2	High
2	12.0	149.0	12.6	74	5	3	Low
3	18.0	313.0	11.5	62	5	4	NaN
4	NaN	NaN	14.3	56	5	5	High
5	28.0	NaN	14.9	66	5	6	High

Replacing null values with mean

In [5]:

```
df['Ozone']=df['Ozone'].fillna(df['Ozone'].mean())
df['Solar.R']=df['Solar.R'].fillna(df['Solar.R'].mean())
df["Wind"]  =  df["Wind"].fillna(df["Wind"].mean())
```

In [ ]:

Replacing null values with mode

In [6]:

```
df['Humidity']=df['Humidity'].fillna(df['Humidity'].mode()[0])
df.isnull().sum()
```

Out[6]:

```
Ozone       0
Solar.R     0
Wind        0
Temp        0
Month       0
Day         0
Humidity    0
dtype: int64
```

In [7]:

```
df.dtypes
```

Out[7]:

```
Ozone       float64
Solar.R     float64
Wind        float64
Temp          int64
Month         int64
Day           int64
Humidity     object
dtype: object
```

In [ ]:

Data Transformation

Converting Continuous to Categorical Values

In [8]:

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
df['Humidity']  =  le.fit_transform(df['Humidity'])
df['Humidity'].unique()
```

Out[8]:

```
array([0, 1, 2])
```

In [9]:

```
df.dtypes
```

Out[9]:

```
Ozone        float64
Solar.R      float64
Wind         float64
Temp           int64
Month          int64
Day            int64
Humidity       int32
dtype: object
```

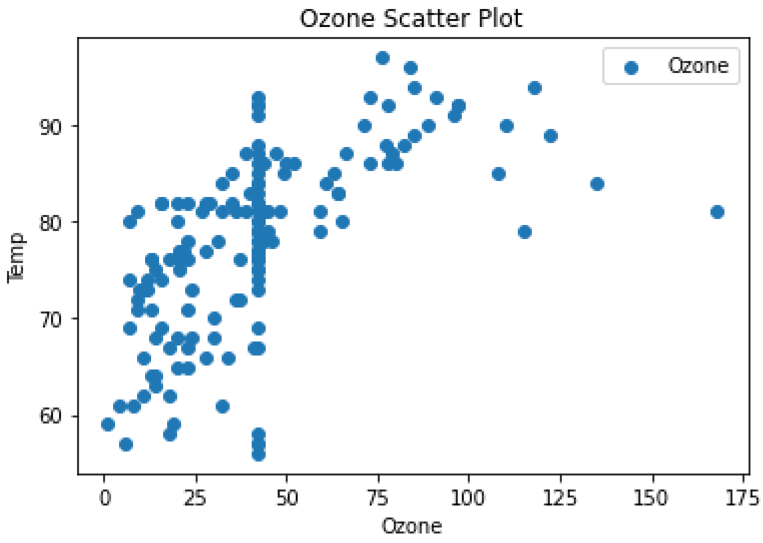
In [ ]:

Visualising the Data

1. Scatter Plot

In [10]:

```
plt.scatter(x = df["Ozone"],y = df["Temp"])
plt.legend(["Ozone"])
plt.xlabel("Ozone")
plt.ylabel("Temp")
plt.title("Ozone  Scatter  Plot")
plt.show()
```



2. Bar Plot

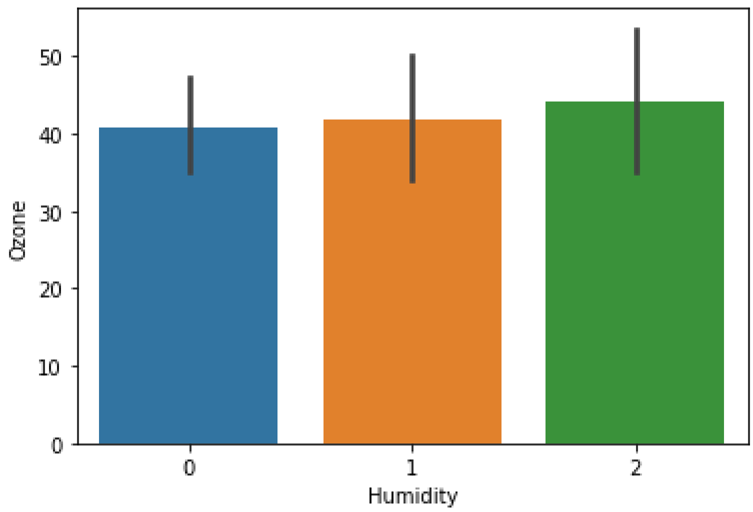
In [37]:

```
import seaborn as sns
sns.barplot(df["Humidity"],df["Ozone"])
```

```
c:\users\hp\appdata\local\programs\python\python39\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

Out[37]:

```
<AxesSubplot:xlabel='Humidity', ylabel='Ozone'>
```



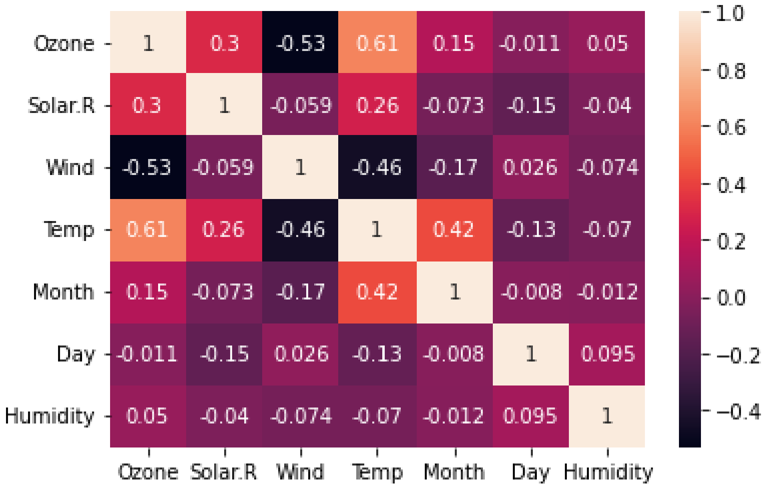
3. Heatmap

In [12]:

```
corr = df.corr()
sns.heatmap(corr, annot = True)
```

Out[12]:

```
<AxesSubplot:>
```



In [ ]:

In [ ]:

In [ ]:

In [ ]:

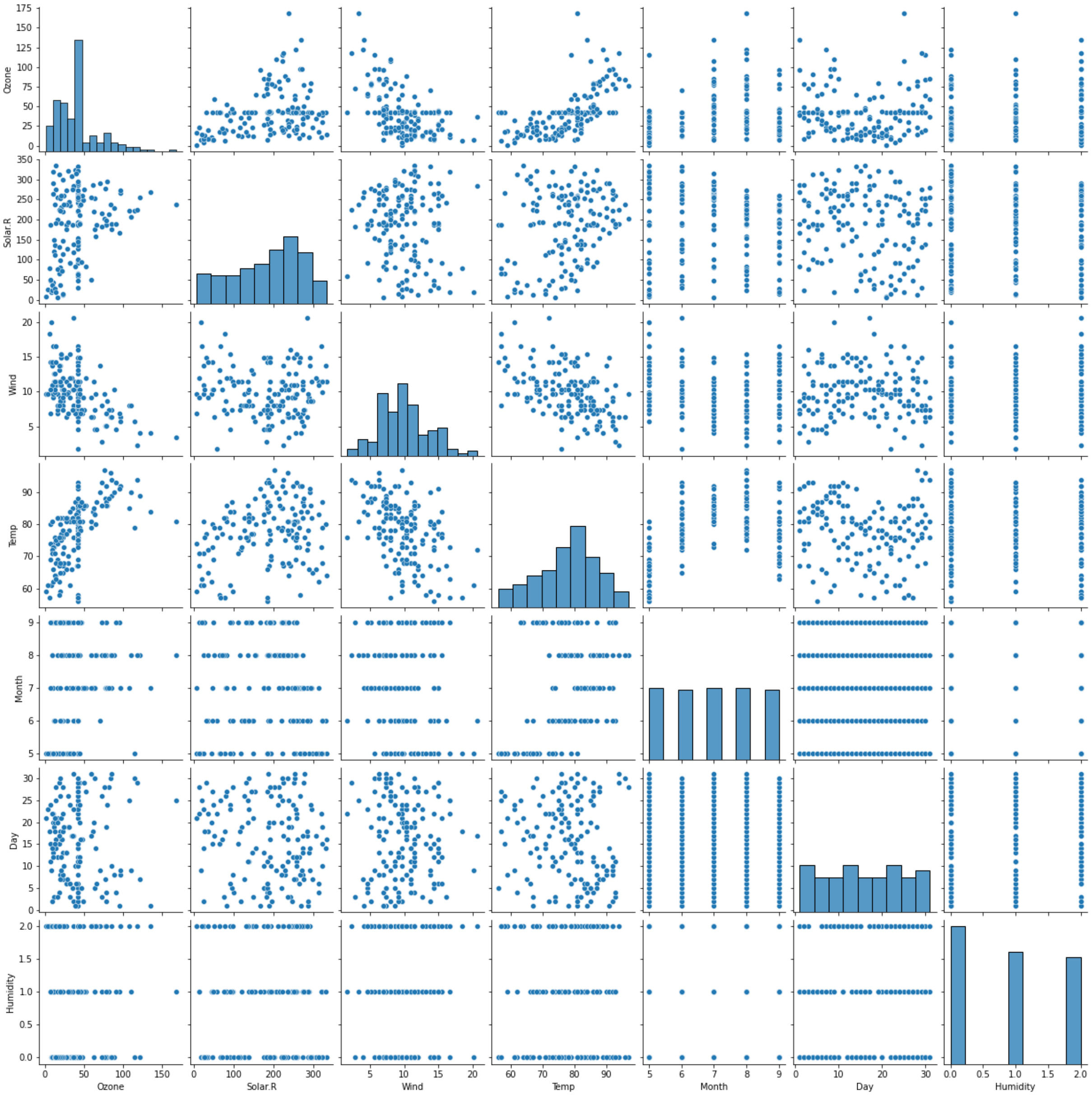
4. Pairplot

In [ ]:

```
sns.pairplot(df)
```

Out[ ]:

```
<seaborn.axisgrid.PairGrid at 0x1525c4861f0>
```



In [ ]:

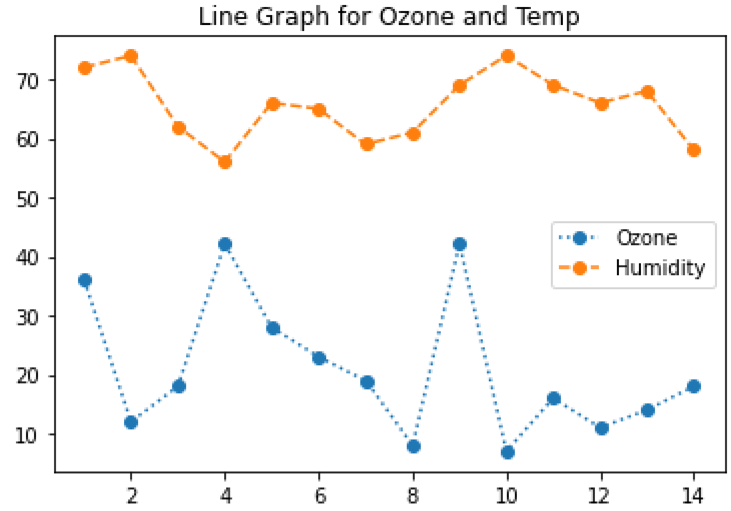
In [ ]:

In [ ]:

5. Line Graph

In [34]:

```
h = df.iloc[1:15, 0]
v = df.iloc[1:15, 3]
plt.plot(h, label="Ozone", marker="o", linestyle="dotted")
plt.plot(v, label="Humidity", marker="o", linestyle="dashed")
plt.title("Line Graph for Ozone and Temp")
plt.legend()
plt.show()
```



In [ ]:

In [ ]:

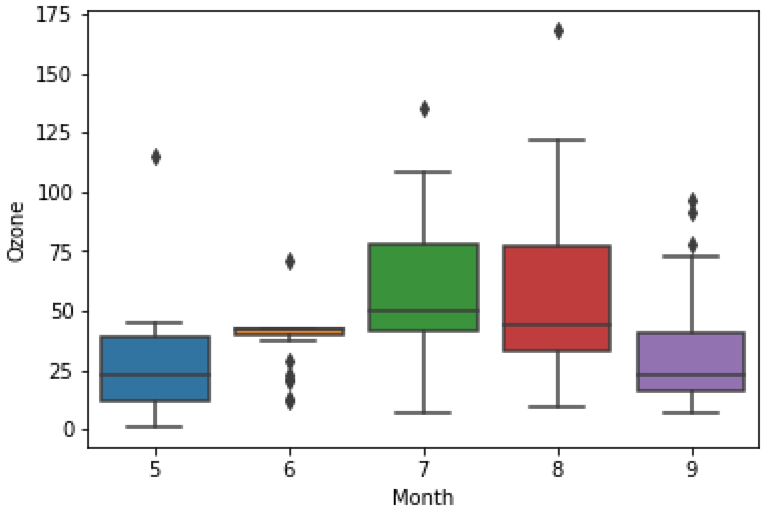
6. Box Plot

In [33]:

```
sns.boxplot(x = df["Month"],y = df["Ozone"])
```

Out[33]:

```
<AxesSubplot:xlabel='Month', ylabel='Ozone'>
```

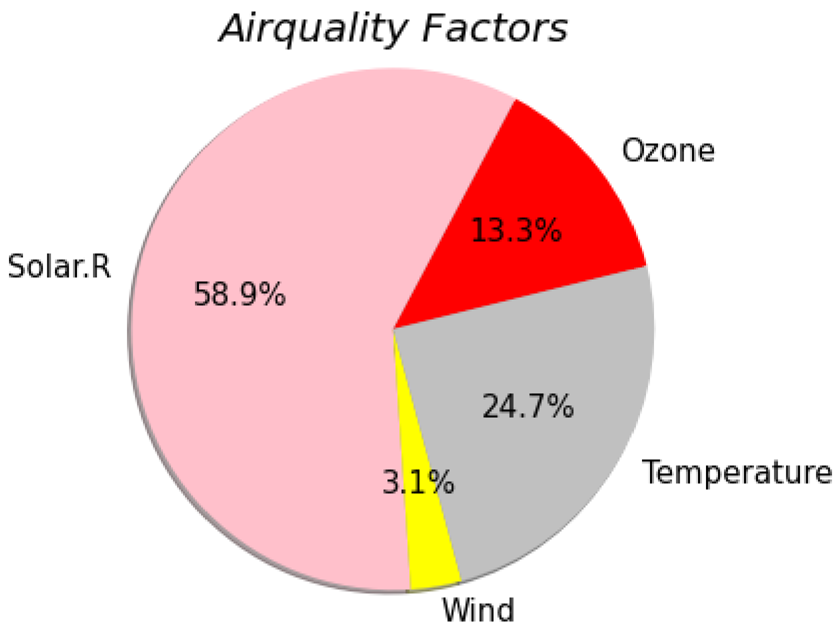


In [ ]:

7. Pie-Chart

```
In [ ]:
labels=  ['Ozone','Solar.R','Wind','Temperature']
sizes=[df['Ozone'].mean(),df["Solar.R"].mean(),df['Wind'].mean(),df["Temp"].mean()]
colors=['red','pink','yellow','silver']
textprops =  {"fontsize":15}
plt.pie(sizes, labels=labels, colors=colors, autopct='%1.1f%%', shadow=True, startan
plt.title("Airquality Factors", fontsize=20, style="italic", pad=35)
```

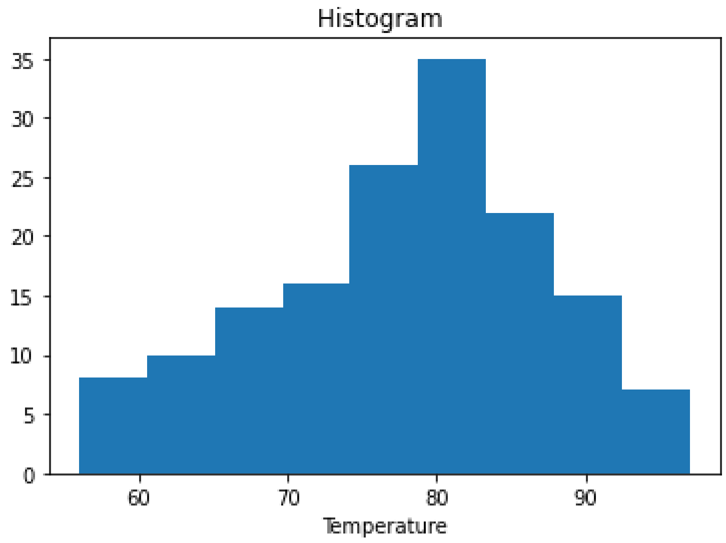
```
Out[ ]: Text(0.5, 1.0, 'Airquality Factors')
```



8. Histogram

```
In [21]:
h=df.iloc[:,-4]
plt.hist(h,bins='auto')
plt.title('Histogram ')
plt.xlabel("Temperature")
```

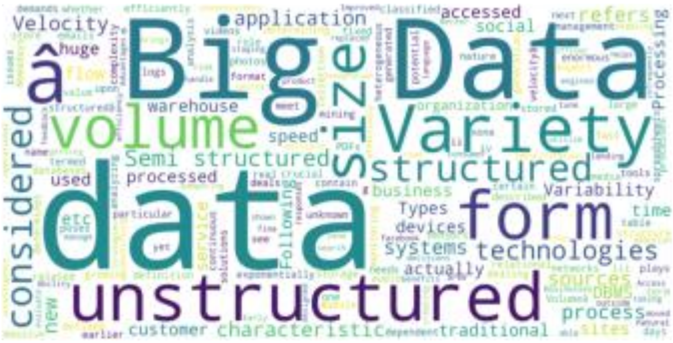
```
Out[21]: Text(0.5, 0, 'Temperature')
```



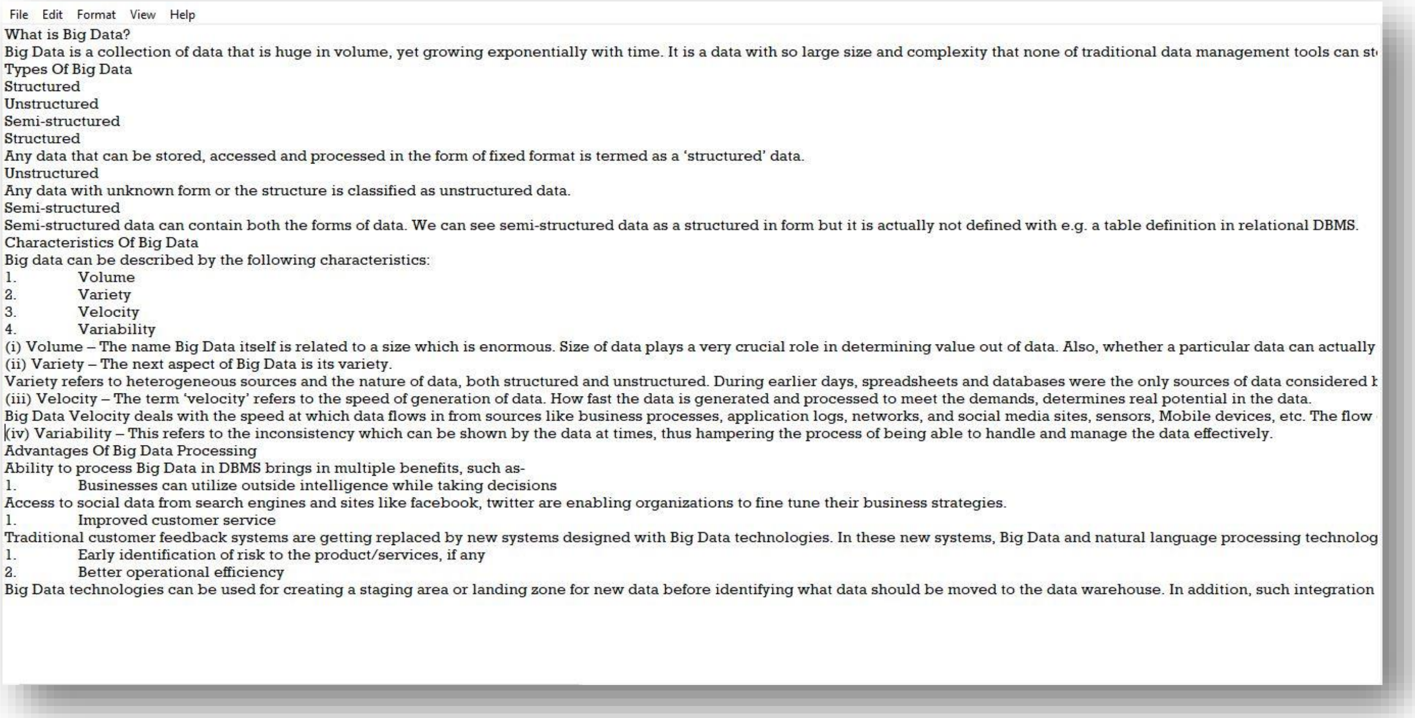
9. Word Cloud

In [18]:

```
from wordcloud import WordCloud, STOPWORDS
text = open("word Cloud.txt").read()
wrd_cld  =  WordCloud(background_color="white", height=2225,  width=4450).generate(text
plt.imshow(wrd_cld)
plt.axis("off")
plt.show()
```



Word-Cloud txt file:



What is Big Data?
Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can st
Types Of Big Data
Structured
Unstructured
Semi-structured
Structured
Any data that can be stored, accessed and processed in the form of fixed format is termed as a ‘structured’ data.
Unstructured
Any data with unknown form or the structure is classified as unstructured data.
Semi-structured
Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS.
Characteristics Of Big Data
Big data can be described by the following characteristics:
1. Volume
2. Variety
3. Velocity
4. Variability
(i) Volume – The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually
(ii) Variety – The next aspect of Big Data is its variety.
Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered
(iii) Velocity – The term ‘velocity’ refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.
Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc. The flow
(iv) Variability – This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.
Advantages Of Big Data Processing
Ability to process Big Data in DBMS brings in multiple benefits, such as-
1. Businesses can utilize outside intelligence while taking decisions
Access to social data from search engines and sites like facebook, twitter are enabling organizations to fine tune their business strategies.
1. Improved customer service
Traditional customer feedback systems are getting replaced by new systems designed with Big Data technologies. In these new systems, Big Data and natural language processing technolog
1. Early identification of risk to the product/services, if any
2. Better operational efficiency
Big Data technologies can be used for creating a staging area or landing zone for new data before identifying what data should be moved to the data warehouse. In addition, such integration