This is the GitHub link: https://github.com/ProtivaArafin/BigDataProject

Explanation of 4Vs of Big Data:

1. Volume:

It refers to the large amount of data processed in big data systems.

Our code uses PySpark, and it reads csv file ('live.csv') into pandas data frame. PySpark splits the dataset across multiple (in our case four) workers so that this cluster-based architecture can scale this easily.

2. Value:

We use two algorithms - Elbow Method and K-Means Clustering. These two provide insights into the data by grouping it into clusters.

Example: using the Elbow method to determine the optimal number of clusters.

3. Variety:

```
from sklearn.datasets import make_blobs
```

```
X,_=make_blobs(n_samples=300,centers=4,cluster_std=0.6,random_state=42)
data = spark.createDataFrame([(float(x[0]), float(x[1])) for x in X], ["x", "y"])
```

'make_blobs' - this function generates a randomized dataset for clustering. This creates a synthetic dataset with specified characteristics.

```
import pandas as pd
pandas_df = pd.read_csv("C:/Users/user/Downloads/Live.csv")
```

This loads a real-life dataset (Live.csv) from the master machine.

4. Visualization:

It refers to the process of presenting data in a visual format, such as graphs, charts, or plots. In our project, we have shown many graphs to visually represent the data and clustering results. Here is one of the graphs: