SOURCES

- Network Hacking Data across various Zip Codes in USA – Source 1
- Internet Usage Data across various Zip Codes in USA – Source 2
- Monthly Living expenditure Data across Various Zip Codes in USA – Source 3

Data Extraction

1. Choose Set of Zip Codes that are available in all 3 Sources such that Data size is around 25 GB each or 50 - 100 GB in total. **Set – ZIP CODES**

2. **Mapper 1** Parses through **SOURCE 1** such that the Zip Codes in **ZIP CODES** and this is output to the **Reducer 1.** The **Reducer 1** takes this and parses out the data required to calculate Network Hacking Statistics. - **OUTPUT 1**
3. **Mapper 2** Parses through **SOURCE 2** such that the Zip Codes in **ZIP CODES** and this is output to the **Reducer 2.** The **Reducer 2** takes this and parses out the data on how much users in that area use the Internet. - **OUTPUT 2**
4. **Mapper 3** Parses through **SOURCE 3** such that the Zip Codes in **ZIP CODES** and this is output to the **Reducer 3.** The **Reducer 3** takes this and parses out the data on how much users in that area spend on living per month. - **OUTPUT 3**

Data Combination

5. **Mapper 4** and **Reducer 4** take **OUTPUT 3** and finds find out the value for three Terms. – **Low Income/Zip Code, Medium Income/Zip Code** and **High Income/Zip Code.** We use these values to compare against end results.
6. **Mapper 5** and **Reduce 5** combine **OUTPUT 2** and the terms **Low Income/Zip Code, Medium Income/Zip Code** and **High Income/Zip Code** to produce the left hand side of what we are trying to prove. – **OUTPUT 4**

End Result

7. **Mapper 5** and **Reduce 5** use **Output 1** and **OUTPUT 4** to produce the final mapping. – **OUTPUT 5**

8. **Mapper 6** and **Reduce 6** combine similar values in **OUTPUT 5** based on locality. This produces the end analytic.
9. **Pictorially Show-off results** to make it easier to explain ☺