**HW1:**

I’m using Popular Baby Names to see which name is the most popular. The data is gathered from 2011 to 2014 and in different ethnicity.

The objective of this project is to explore the relationship between the year of birth and Ethnicity. The hypothesis posits that certain baby names may experience fluctuations in popularity corresponding to cultural shifts, such as the influence of popular media or societal events. By analyzing historical baby names data, I aim to validate or disprove this hypothesis.

1. I have selected the "Popular Baby Names" dataset available on data.gov for this project. I will download this dataset using curl and save it as a CSV file.
2. After downloading the dataset, I will use the head command to display the columns such as "Year," "Gender” ,"Ethnicity"," and "Child’s First Nam”, “Count”, and “Rank”.

A screenshot of a computer

Description automatically generated

1. I will use awk -F to extract the selected columns from the CSV file and create a new file containing the relevant data. This command extracts the columns 4 and 5 from the CSV file, assuming ',' as the delimiter, and redirects the output to a new file named "Popular\_Baby\_Names.csv".

.

A computer screen with text and numbers

Description automatically generated

My interesting in pursuing are Child’s First Name and Count. This is initial data of Hispanic name in 2011 follow the alphabet order.

A computer screen with white text

Description automatically generated

Initial Data: To gather initial data, I will calculate basic statistics such as the frequency of each name, the distribution of names over time, and any noticeable trends or patterns. This may involve counting the occurrences of specific names or analyzing the popularity of names in different regions.

Conclusion: Based on the initial data analysis, I will document any observed trends, correlations, or anomalies. These early conclusions will provide insights into the factors influencing baby name popularity and guide further analysis.

**HW2:**

The advanced analysis in HW2 delved deeper into understanding the variation in the number of unique ethnicities associated with baby names across different years. The MapReduce program was used to process the dataset and extract the number of unique ethnicities for each year, providing a more detailed perspective on ethnic diversity in baby naming trends over time.

A computer screen with white text

Description automatically generated

A screenshot of a directory

Description automatically generated

**Conclusion:**

* The number of unique ethnicities associated with baby names varies across different years. Some years might have a higher diversity of ethnicities compared to others, indicating changing trends in naming practices and cultural influences.
* Ethnic Diversity: The analysis highlights the presence of diverse ethnicities in baby naming patterns. This diversity reflects the multicultural nature of societies and the influence of various cultural backgrounds on naming choices.
* By analyzing the data using Hadoop, it becomes evident that there are fluctuations in the popularity of certain ethnicities over time, suggesting evolving societal preferences and cultural shifts.

**Compare:**

* The initial conclusions from HW1 provided a broad overview of baby naming trends, whereas the advanced analysis in HW2 offered a more understanding of ethnic diversity in naming practices over time. The combination of both analyses provides a comprehensive view of the factors influencing baby name popularity, including cultural, societal, and temporal influences.

**Summary:**

* The insights gained from both HW1 and HW2 offer valuable perspectives on the complexities of baby naming trends, emphasizing the multifaceted nature of naming practices influenced by various factors.