**Big Data Analysis on Census Data using**

**Hadoop**

Project Analysis Report

The dataset is a census data of a country on a yearly basis. The census data consists of the details of 6 lakh candidates of a country. The details are composed of:

1. Age

2. Education(highest qualification)

3. Marital status

4. Gender

5. Tax-filer status(tax plan they follow)

6. Income

7. Parents

8. Country of birth

9. Citizenship

10. Weeks-Worked.

On analyzing these columns we can categorize the data into 4 modules namely education, social welfare, financial and planning. In education module, all the data related to the education field will fall like the literacy rate etc. In social welfare the issues related to the public and their well-being will be considered. In financial and planning modules, economic related like income and tax related data will be included.

Module 1-

|  |  |
| --- | --- |
| Module | Education |
| Schema(columns) | Gender, education, age, citizenship, weeks-worked |
| Keys | Gender, citizenship, age, education |
| External tables | Age-group(age, category-name) |

Module 2-

|  |  |
| --- | --- |
| Module | Social-welfare |
| Schema(columns) | Marital-status, gender, parents, weeks-worked, age, citizenship |
| Keys | Gender, marital-status, parents |
| External tables | NA |

Module 3-

|  |  |
| --- | --- |
| Module | Financial |
| Schema(columns) | Citizenship, weeks-worked, taxfiler-status, gender, income |
| Keys | Citizenship, gender, taxfiler-status |
| External tables | Taxpercentage |

Module 4-

|  |  |
| --- | --- |
| Module | Planning |
| Schema(columns) | Citizenship, age, taxfiler-status |
| Keys | Age, taxfiler-status |
| External tables | NA |

After analyzing the columns of the dataset some of the problem statements can be stated according to the divided modules:

Education:

1. Find the candidates category-wise gender-wise

2. Age group-wise, gender-wise literacy rate

3. No of students studying from other countries and native students

4. No of students age-group some benchmark most preferred education

5. Find the no of candidates education category wise employed or not

6. Country literacy rate(US)

Social Welfare:

1. No of widow female candidates

2. No of children parents category-wise gender-wise

3. Age above 60 citizens and non-citizens

4. Sex-ratio

5. No of unemployable widows and divorced

Financial:

1. No of non-citizens working %

2. Gender wise income generated

3. Tax money generated

4. Money generated for non-citizens

5. Education category wise top and least salaries

Planning:

1. Next year senior citizens i.e., becoming age >60

2. List of tax payers and non-tax payers

3. No of voters will be upcoming in next 3 years

4. Citizens age above 23 having no employment and highest education

5. List of candidates age group wise