Aadhar Based Analysis using Hadoop

Group no - 06

B02 Rushit Anadkat

B16 Dhwani Karkar

B28 Hetul Parmar

B30 Nikhil Parmar

B36 Dhruv Patel

B45 Kwinsi Sadhu

Data Preparation

- **Data Selection**

  - We have collected the data from the github repository and we have put that data in the API in json format.

  - From that API we have created a CSV file.

- **Data Pre-processing**

  - After collecting the CSV we found missing values in some columns and some noisy data.

  - So we have done cleaning the data in python using the tool Google Colab.

- **Data Loading**

  - After preprocessing we have uploaded our CSV in HDFS.

- **Data Analysis**

  - After analyzing the dataset following queries have been generated.

- **Queries**
  - Identify the age of people whose Aadhaar card is not enrolled in "Allahabad bank".
  - Identify the Registrar name who has the maximum Aadhaar card generated from "Bagalkot District".
  - Identify in which city most male people have Aadhaar enrolled.
  - Identify Females who have generated aadhar and belong to the 18 to 30 age group.
  - Identify Which state has the most females registered from "Frontech system pvt ltd" or "financial information network".
  - Identify the people whose age is not more than 50, belong to Ecartes Technology Pvt. Ltd and their Registrar is CSC e-Governance Services India Limited.
  - Count the number of banks that have the most number of enrollment agencies, belong to state karnataka, district chikkaballapur and have people above age 20.
  - Identify the subdistrict where no aadhar card is generated from bank Uti Infrastructure Technology & Service Limited.
  - Display Registrar and enrollment agency whose age is greater than or equal to 18 and does not generate Aadhar card in Bihar state.
  - Display the sum of the total state of each unique district.