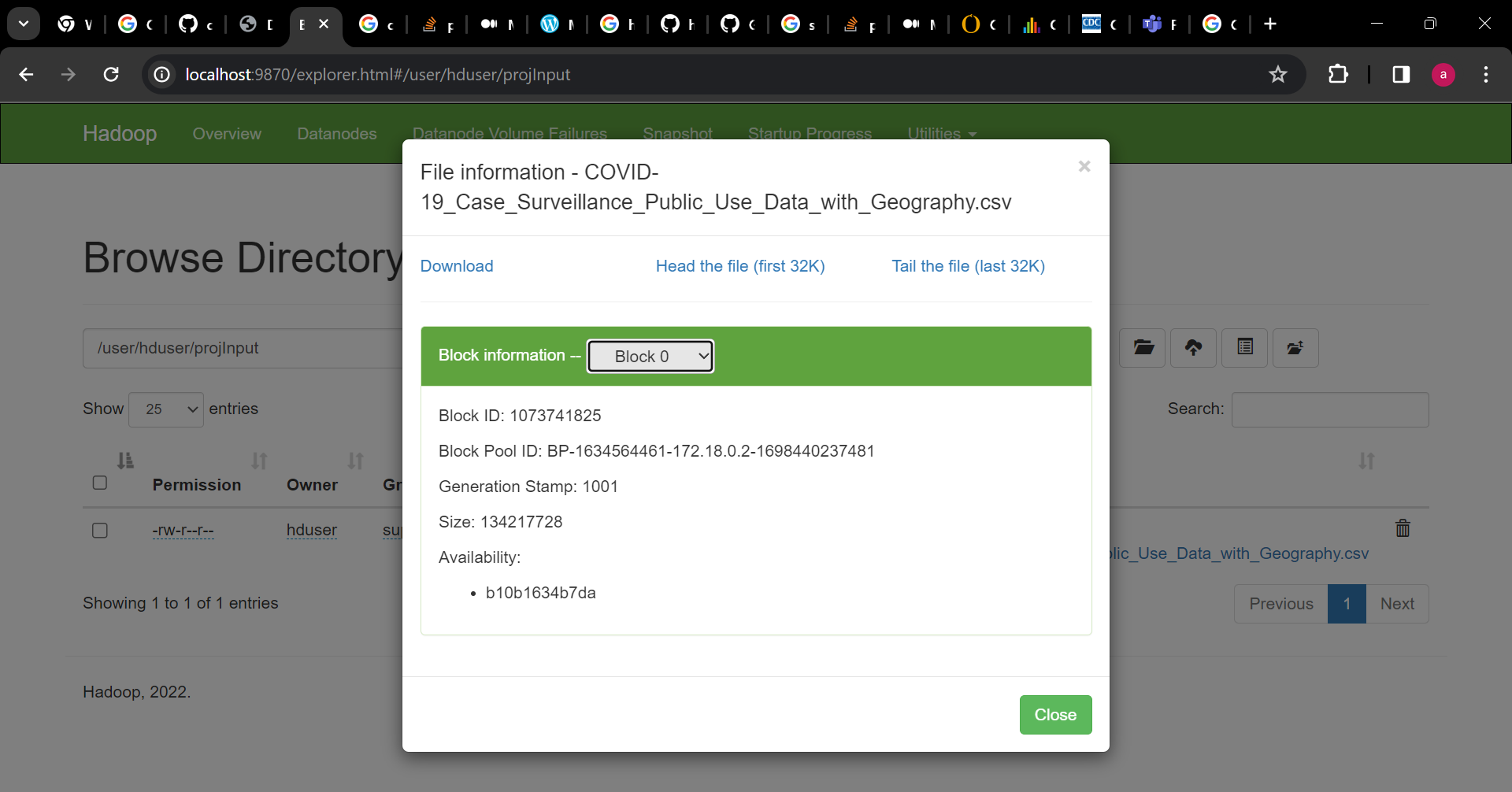
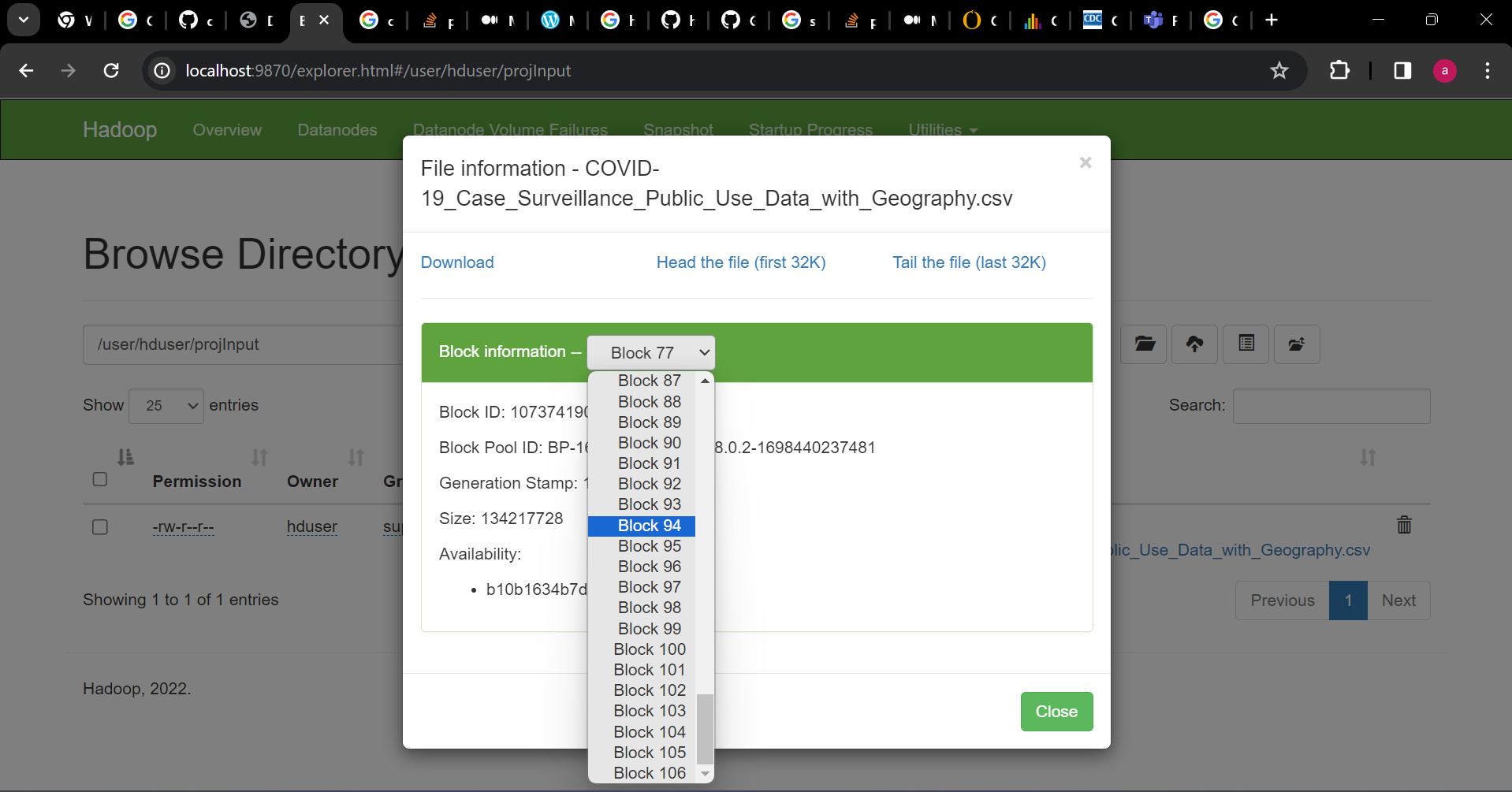
* Data is Collected from CDC Which was available for public use
  + <https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/ge62-5fe5>
* Other sources of data considered for the project
  + <https://covid19.who.int/region/searo/country/in>
  + <https://covid19.census.gov/>
  + <https://www.kaggle.com/datasets/imdevskp/corona-virus-report>
  + <https://datascience.nih.gov/covid-19-open-access-resources>
  + <https://health.google.com/covid-19/open-data/raw-data>
* As per scope of the project main analysis is on following points
  + Considering only one Country – USA
  + Demographic Area
  + Age factor in death rate
  + Monthly infection rate
  + Above factors help to analyses the risk factor for a particular person
* Reason for considering CDC data
  + Massive amount of data 14GB
  + Age factor available in four categories which will be useful for our analysis
    - 0 - 17 years
    - 18 to 49 years
    - 50 to 64 years
    - 65+ years
  + Demographic area is mentioned for major records with variable [res\_county]
  + Underlying condition has been mentioned
  + Confirmed death cases mentioned
  + Month covid infected
  + Ethnicity available
* Above mentioned variables are useful for determining the risk factor for a particular person.

Steps for project

* Data Collection from different sources
* Filter the data
* Determine the variables for assessing
  + res\_state [For State]
  + res\_county [For sub demographic area]
  + age\_group [Age Group]
  + sex [Gender]
  + race
  + ethnicity
  + underlying\_conditions\_yn
* Upload the into distributed filesystem, a total of 106 blocks of data have been formed





* Decide the filtering analysis.
* Map the data and form key value pairs.
* Decide sampling pattern.
* Algorithm for creating the pattern and calculating the risk factor according to the category.
  + If person is from an area with more deaths, lies in the high-risk age factor, having underlying medical condition the person is in high-risk condition. These results may help to take proper action.
  + Tracking monthly infection rate may also help to analyze the situation.