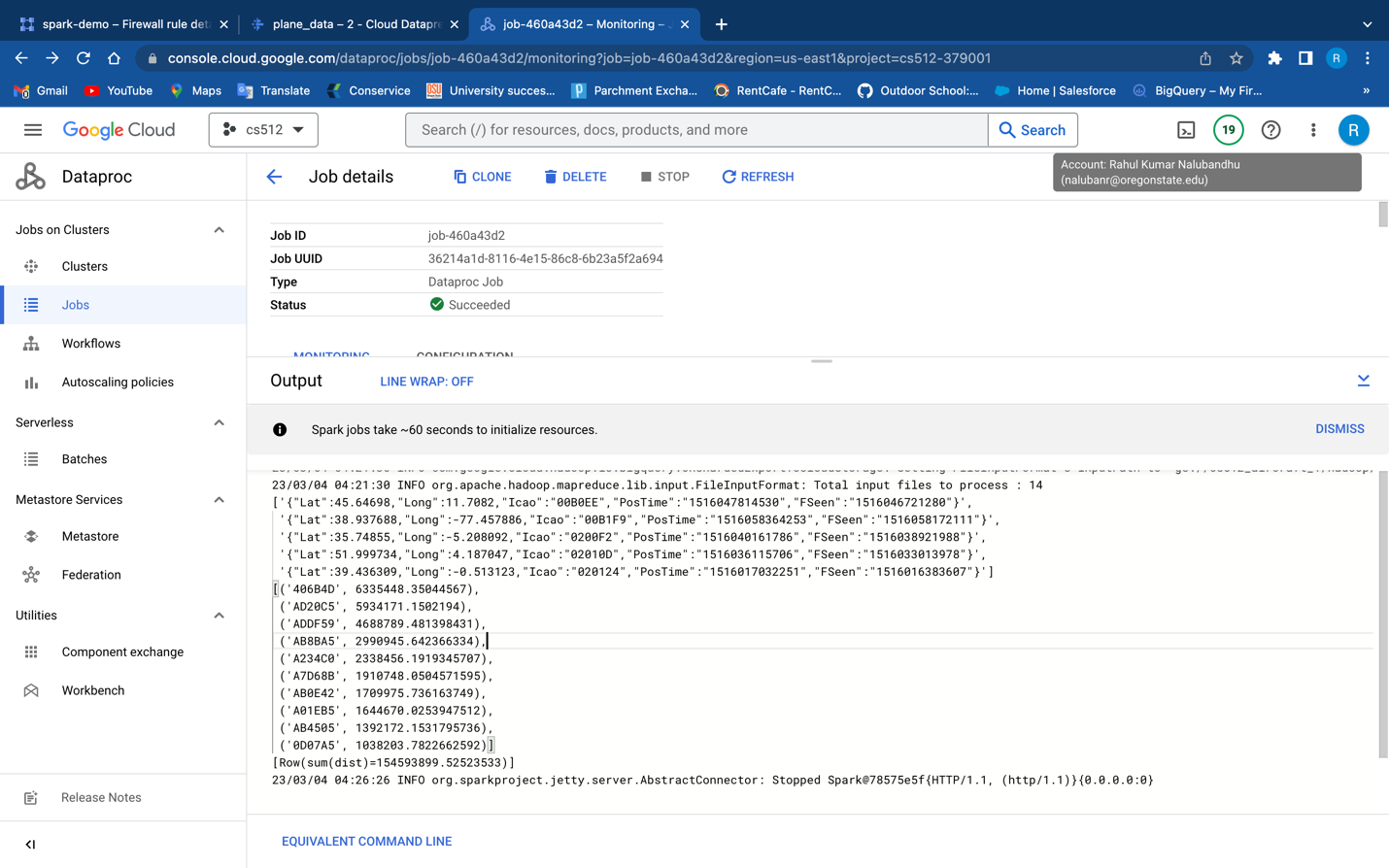
The below picture shows the total distance as (154593899.52523533). This is when the initial run made on the job by just using the given updated python file.



In the process of cleaning the data, I reviewed it manually and applied a filter to the "Icao" column. To accomplish this, I created a new column using "groupby" and "count" functions, which allowed me to generate a count of values for each unique "Icao" value. Upon examining the data, I observed that the count for "Icao" ranged from 1 to 4. To ensure that we only consider valid distance calculations, we need to have an even count for each "Icao" value. This is because we require two sets of latitude-longitude coordinates to calculate the distance between them. Consequently, I deleted any values from the new column that corresponded with odd numbers (1 and 3), leaving only even values. This approach enabled me to work with valid data to calculate the distance.

when I run this job in dataproc then I did see the drop in total distance which is a good sign to view that we filtered the data.

Graphical user interface, text, application, Word, email

Description automatically generated