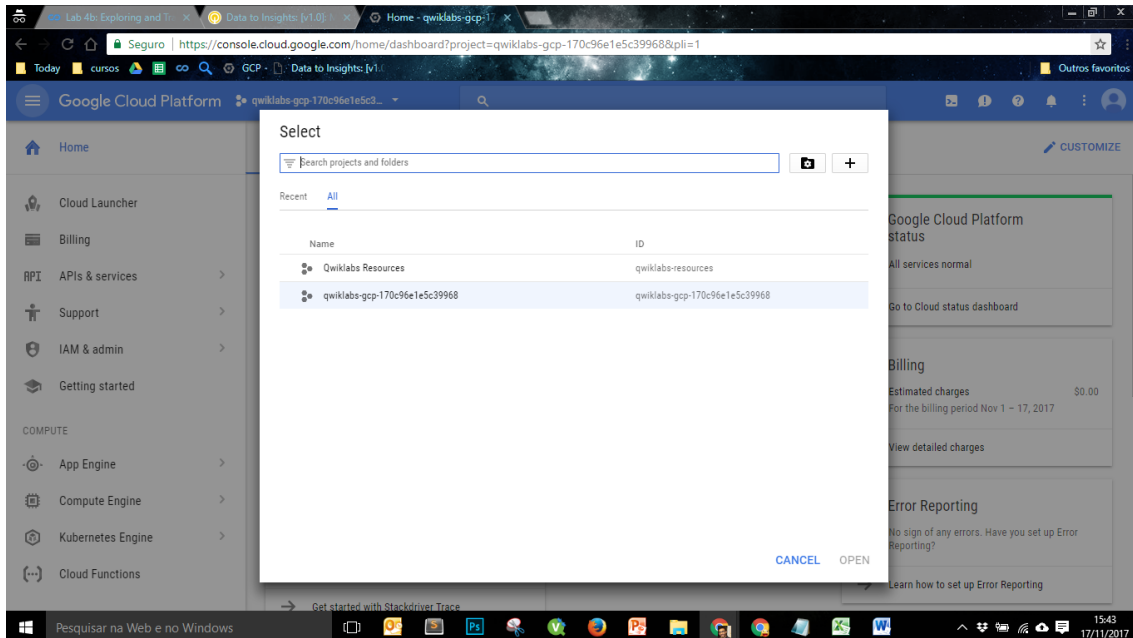


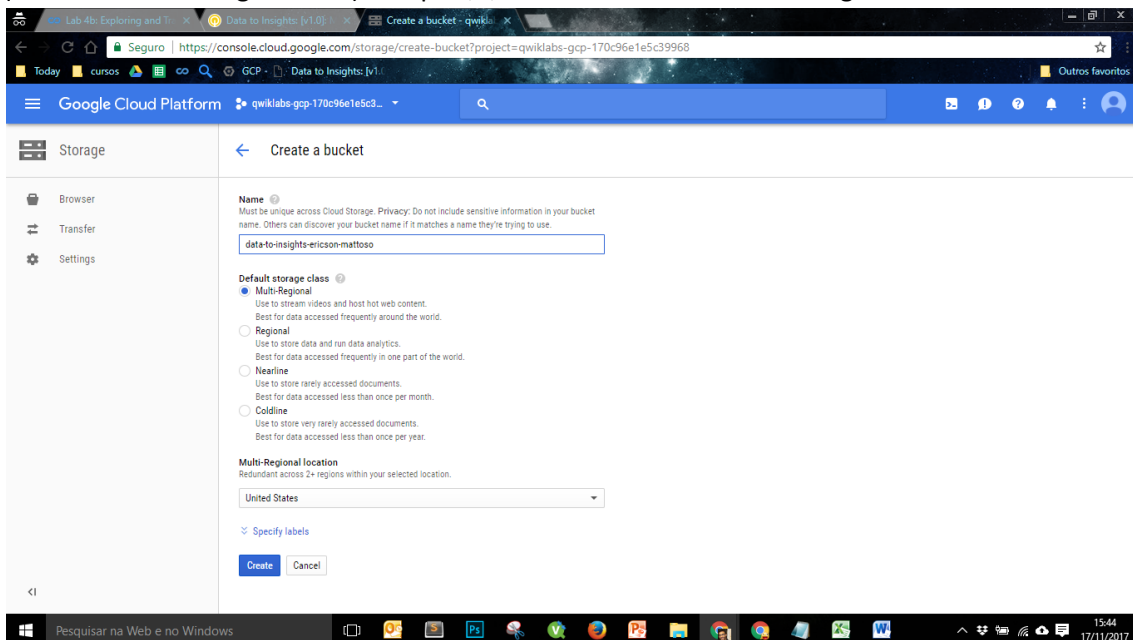
Let's go step by step.

First, I choose the test Data to Insights: [v1.0]: M05 L4ab: Transforming Data with Cloud Dataprep using the account google135671_student@qwiklabs.net. I am using here in Brazil. I followed the steps:

(Create a new Storage Bucket) - Steps 1,2,3



(Create a new Storage Bucket) - Steps 4,5,6 - Bucket name: data-to-insights-ericson-mattoso



(Create a new BigQuery Dataset to store Cloud Dataprep output) – Steps 1,2,3,4,5,6

When I tried to Access BigQuery I receive this message: *Access Denied: Project [qwiklabs-resources](mailto:google135671_student@qwiklabs.net): The user google135671_student@qwiklabs.net does not have*

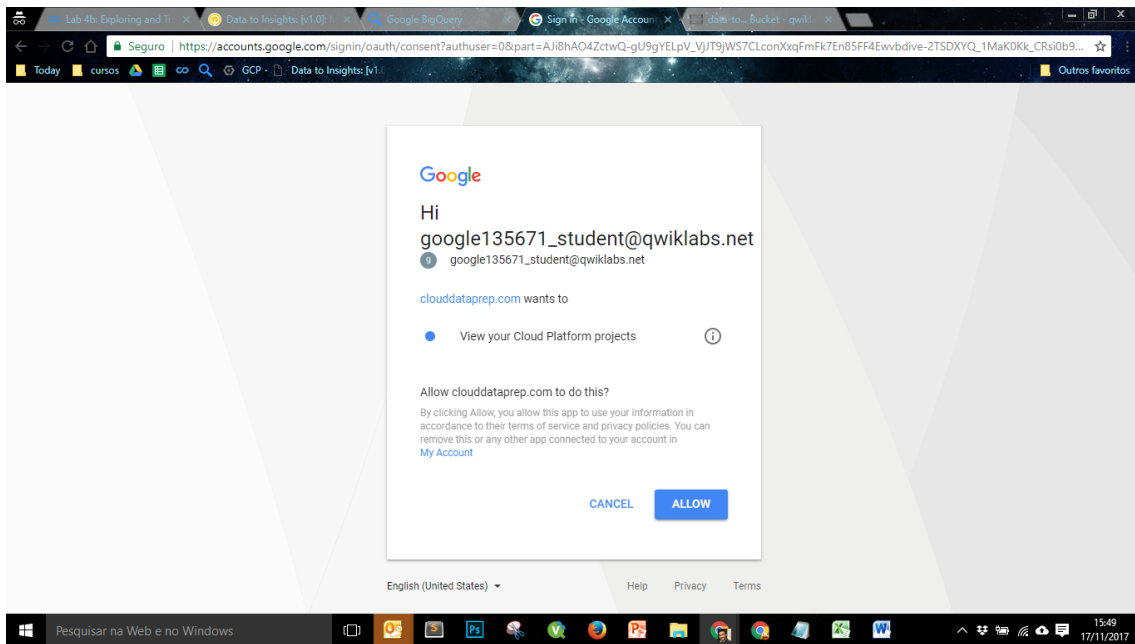
bigquery.job.list permission in project qwiklabs-resources. Dismiss. So, I click on *Dismiss*.

The image consists of two screenshots of the Google BigQuery web interface, showing the process of creating a new dataset.

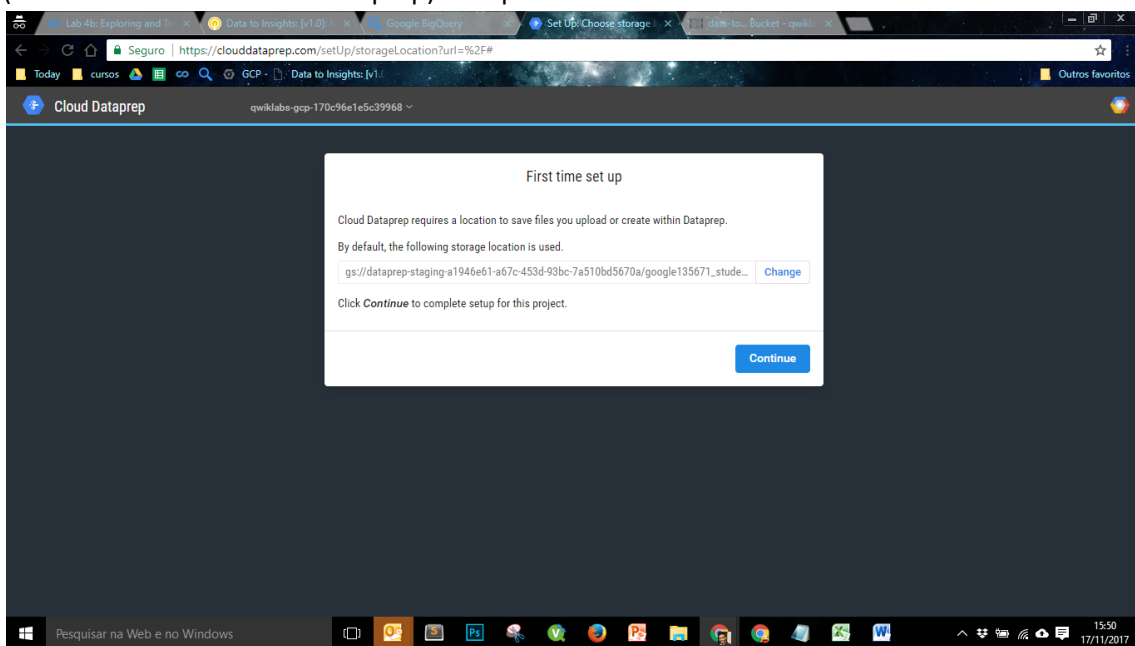
Top Screenshot: The browser address bar shows `https://bigquery.cloud.google.com/welcome/qwiklabs-resources`. A red error banner at the top states: "Access Denied: Project qwiklabs-resources: The user google135671_student@qwiklabs.net does not have bigquery.job.list permission in project qwiklabs-resources. Dismiss". The main content area says "Welcome to BigQuery!" and lists options to get started. On the left sidebar, under "Qwiklabs Resources", it says "No datasets found in this project. Please create a dataset or select a new project from the menu above." A dropdown menu is open for "Switch to project", showing "Qwiklabs Resources (qwiklabs-resources)" as the selected option, with a sub-option "qwiklabs-gcp-170c96e1e5c39968 (qwiklabs-gcp-170c96e1e5c39968)".

Bottom Screenshot: The browser address bar shows `https://bigquery.cloud.google.com/welcome/qwiklabs-gcp-170c96e1e5c39968`. The error banner is gone. The "Create Dataset" dialog box is open, showing the "Dataset ID" field with the value "irs_990", the "Data location" field with the value "unspecified", and the "Data expiration" field with the value "Never". The "OK" button is highlighted.

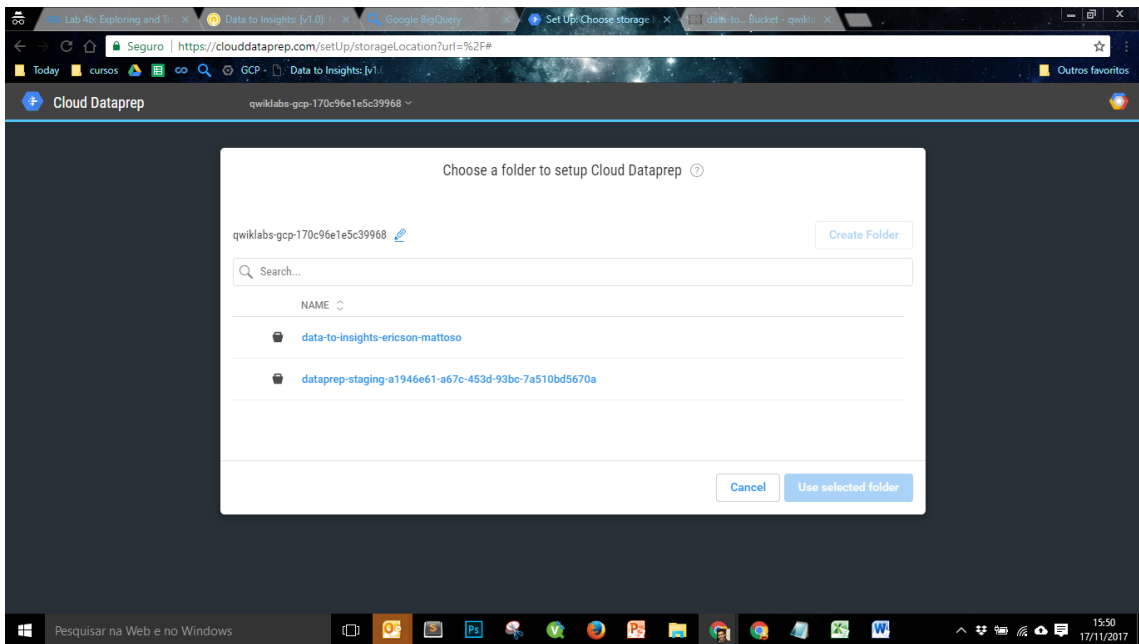
(Access and Launch Cloud Dataprep) – Steps 1,2,3,4



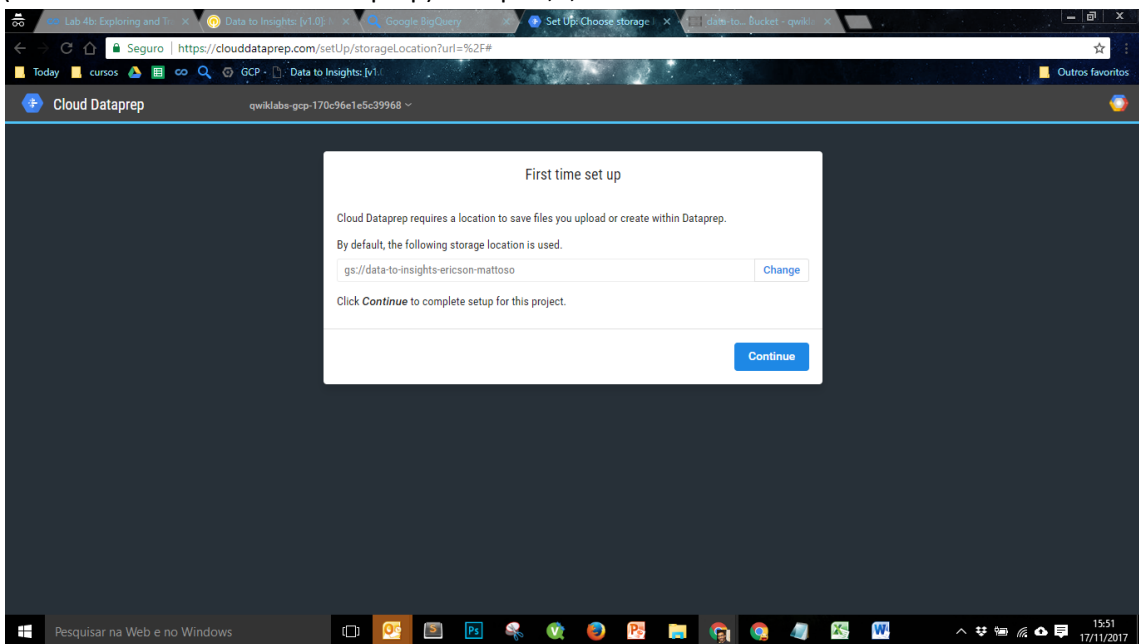
(Access and Launch Cloud Dataprep) – Steps between 4 and 5



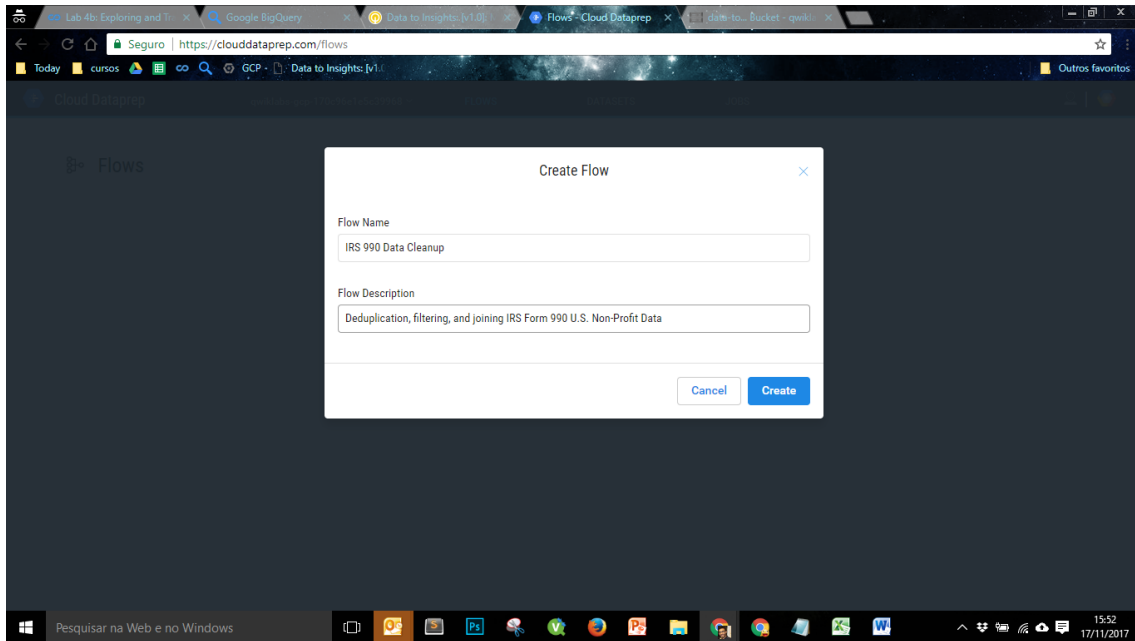
(Access and Launch Cloud Dataprep) – Steps between 4 and 5. I choose the bucket name “data-to-insights-ericson-mattoso”.



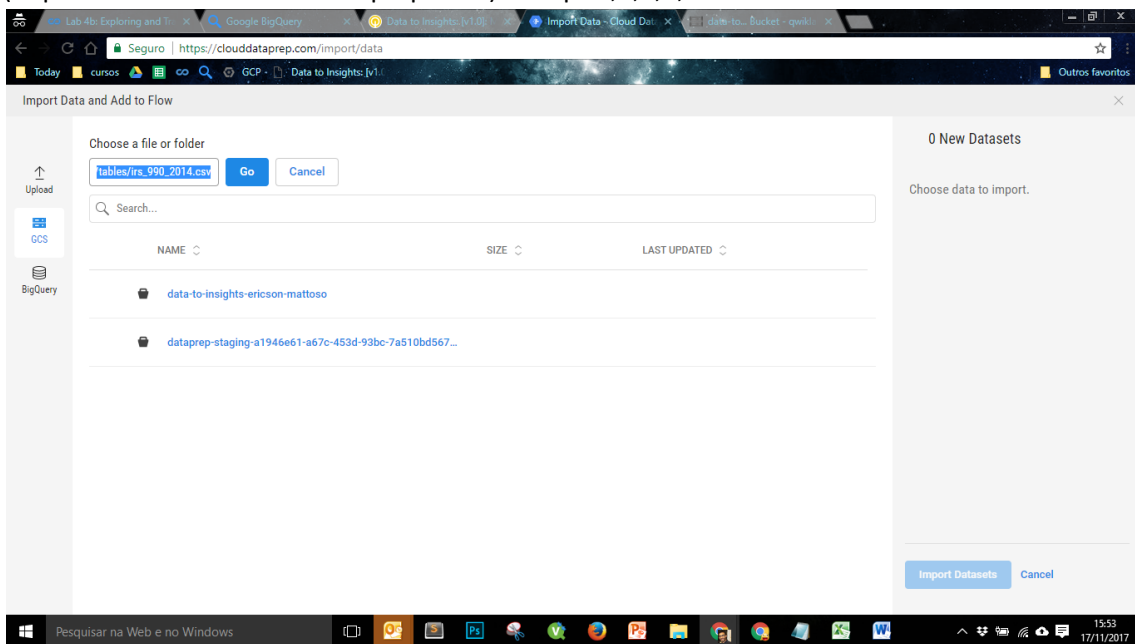
(Access and Launch Cloud Dataprep) – Steps 5,6,7



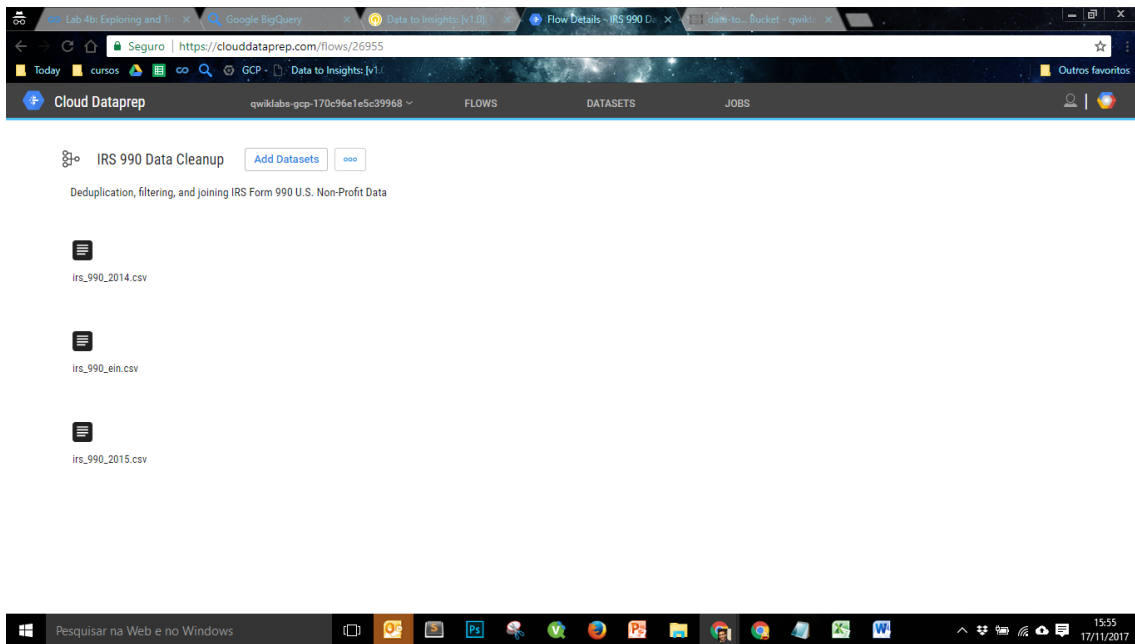
(Import IRS Dataset into a Dataprep Flow) – Steps 1,2



(Import IRS Dataset into a Dataprep Flow) – Steps 3,4,5,6,7.



(Import IRS Dataset into a Dataprep Flow) – Steps 3,4,5,6,7(Import IRS Dataset into a Dataprep Flow) – Steps 8,9,10,11,12,13.



Lab 4b: Exploring and Transforming data with Cloud Dataprep (Transforming Raw Data with a Dataprep Recipe) – Steps 1,2,3,4,5

Details

irs_990_ein.csv

[Add new Recipe](#)

Data Preview

ssn	ein	abc	name
460394864			OLD ELM SPRINGS HUTTERIAN BRETHREN
609090804			HERITAGE VILLAGE CIVIC ASSOCIATION
46326977			ELIZABETH D OVEY 1974 TRUST 031274
42621716			ORDER OF THE RECONCILIATION INC
236728945			AGNES D SITMAN PARK HOME TRUST
161516146			PROFESSIONAL MACINTOSH USERS GROUP
460229905			MAXWELL HUTTERIAN BRETHREN INC

Type: GCS
 File Size: 250.55MB
 Size: 28 columns - 5 types
 Location: /tables/irs_990_ein.csv
 Updated: Today at 3:54 PM
 Created: Today at 3:54 PM

(Build a Recipe to Transform your Raw Data) – Steps 1,2,3,4,5.

Grid Columns Find column Filters

1 Remove duplicate rows

28 Columns 67,474 Rows 5 Data Types

(Trim Unused Columns) – Steps 1,2,3,4,5,6.

Grid Columns Select Find column Filters

1 Remove duplicate rows
2 Drop tax_period
3 Drop status

26 Columns 67,474 Rows 5 Data Types

(Cleaning Historical Tables) - Steps 1,2,3,4,5,6.

(ADDING A NEW FIELD FOR CALENDAR YEAR) - Steps 1,2,3,4.

The screenshot shows the Google Cloud Data Studio interface. The main view displays a data table with the following columns: `calendar_year_filed`, `elf`, `ein`, `tax_pd`, and `subseccd`. The table contains 13,717 rows. On the right side, there is a 'Recipe' panel with two steps: '1 Remove duplicate rows' and '2 Create calendar_year_filed from 2015'. The interface also shows a 'Find column' search bar and a 'Filters' dropdown.

(UNIONING HISTORICAL TABLES TOGETHER FOR A TALLER DATASET) – Steps 1,2,3,4,5,6,7,8

The screenshot shows the 'Union' operation in Google Cloud Data Studio. It displays two datasets: 'irs_990_2015' (247 of 247 columns in Union) and 'irs_990_2014' (246 of 246 columns in Union). The 'Union Output' section shows the resulting 247 columns in the union, including 'calendar_year_filed', 'elf', 'ein', 'tax_pd', 'subseccd', and various identification numbers. The interface also shows a 'Match columns' dropdown and an 'Add datasets' button.

(DATE PARSING THROUGH DERIVED FIELDS) – Steps 1,2,3,4,5,6,7,8

The screenshot shows the Google Cloud Data Studio interface. At the top, there's a browser window with the URL <https://clouddataprep.com/data/26955/72430>. Below the browser, the Data Studio interface is visible. The main panel shows a data table with columns: calendar_year_filed, elf, ein, tax_pd, and tax_pd_cleaned. The 'tax_pd' column is highlighted in blue. On the right, the 'Recipe' panel is open, showing the 'Derive' transformation step. The formula entered is `merge([left(tax_pd, 4), right(tax_pd, 2), '01'], '-')`. The 'New column name' is set to 'tax_pd_cleaned'. The bottom status bar indicates 248 Columns, 21,335 Rows, and 5 Data Types.

(ADDING A JOIN TO LOOKUP DATA COLUMNS FROM ANOTHER DATASET) – Steps 1,2,3,4,5,6,7,8

The screenshot shows the Google Cloud Data Studio interface. At the top, there's a browser window with the URL <https://clouddataprep.com/data/26955/72430>. Below the browser, the Data Studio interface is visible. The main panel shows a data table with columns: ein, ein1, calendar_year_filed, elf, and tax_pd. The 'ein' column is highlighted in blue. On the left, the 'Join Keys' panel is open, showing the join configuration. The 'Join Keys' section shows a join on the 'ein' column. The 'Columns' section lists the columns from the datasets. The bottom status bar indicates 248 Columns, 21,335 Rows, and 5 Data Types.

(Running our Dataprep Job) – Steps 1,2,3,4,5,6,7,8

