Module 5

Cleaning and Transforming your Data

In this module we will:

- **Examine the 5 Principles of Dataset Integrity**
- Characterize Dataset Shape and Skew
- Clean and Transform Data using SQL
- Clean and Transform Data using a new UI: Introducing Cloud Dataprep

Google Cloud

Garbage in… garbage out

High quality datasets conform to strict integrity rules

1

Validity

Data conforms to your business rules

Challenges

Out of Range
Empty Fields
Data Mismatch

2

Accuracy

Data conforms to an objective true value.

Challenges

Lookup Datasets
Do Not Exist

3

Completeness

Create, save, and store datasets.

Challenges

Missing Data

4

Consistency

Derive insights from data.

Challenges

Duplicate Records
Concurrency Issues

5

Uniformity

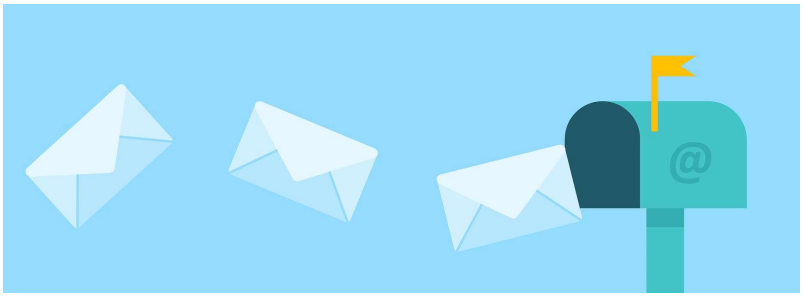
Explore and present data

Challenges

Same Units of Measurement

Google Cloud

Valid data follows constraints on uniqueness



what do these identifiers have in common? why were they set up that way?

Valid data corresponds to range constraints

Roll #	Value
1	2
2	2
3	6
4	5
5	1
6	7

Which value(s) are out of range?

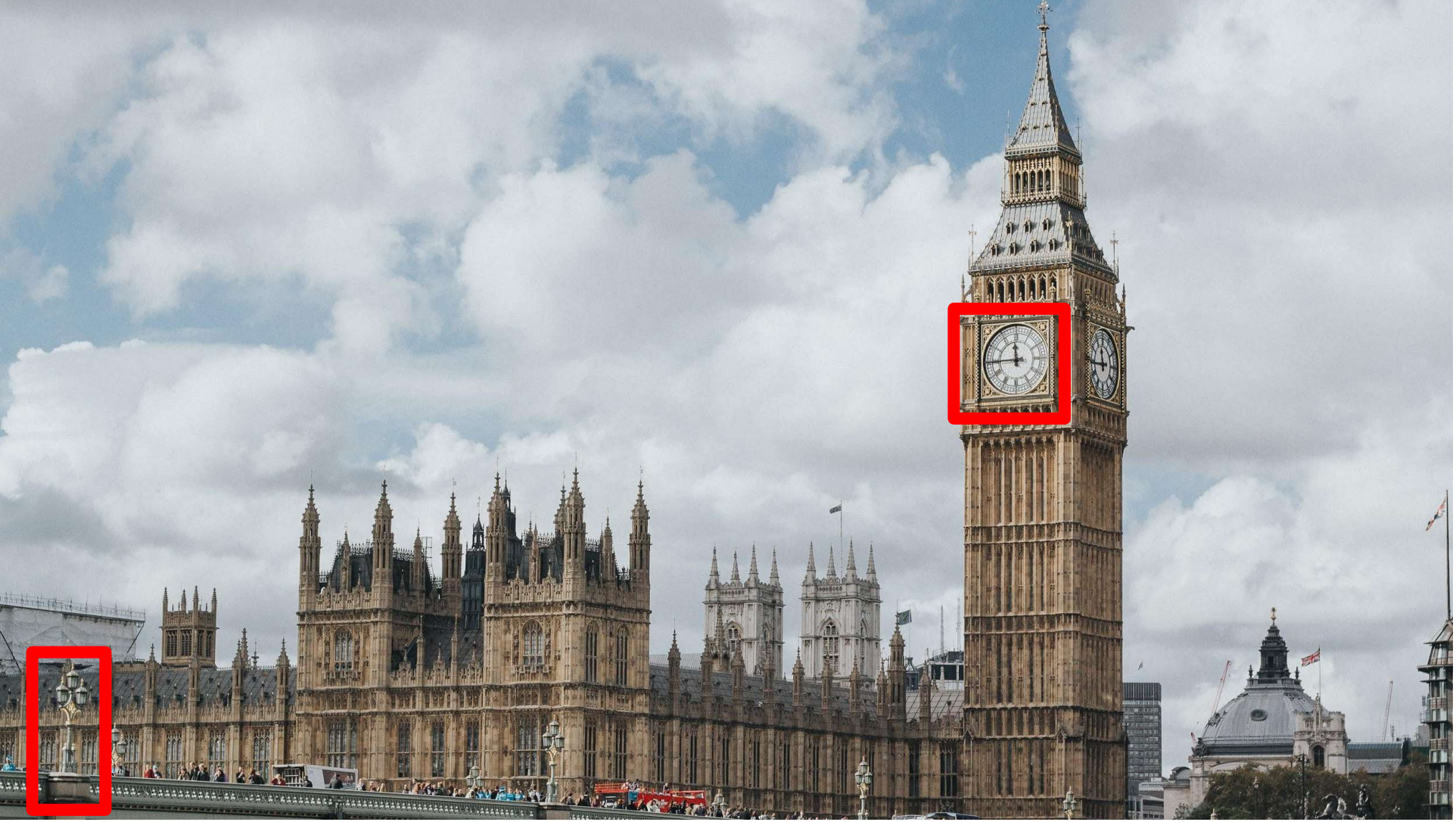
Accurate data matches to a known source of truth

U.S. States
Washington
Oregon
California
Hot Dog
Florida
Maine

Google Cloud

Lamps and Clocks?

Google Cloud



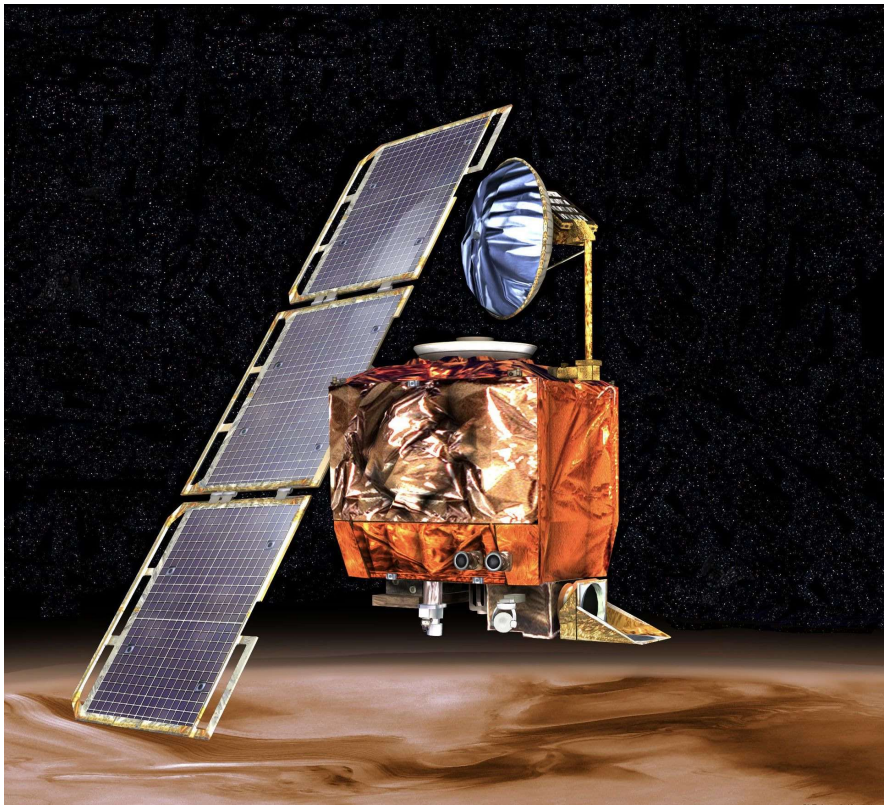
Consistent Data Ensures Harmony across Systems

House Address	Owner ID
123 ABC St	12

Owner ID	Owner Address
15	123 ABC St.
12	53rd Ave.

Who owns the house?

Uniformity in Data Means Measuring the Same Way



=$125 Million

In November 1999, NASA lost a Mars climate orbiter because of English vs Metric system measurements

Google Cloud

Module 5

Cleaning and Transforming your Data

In this module we will:

- Examine the 5 Principles of Dataset Integrity
- **Characterize Dataset Shape and Skew**
- Clean and Transform Data using SQL
- Clean and Transform Data using a new UI: Introducing Cloud Dataprep

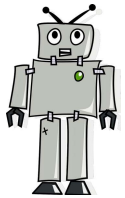
Lab: Explore and Shape data with Cloud Dataprep

Google Cloud

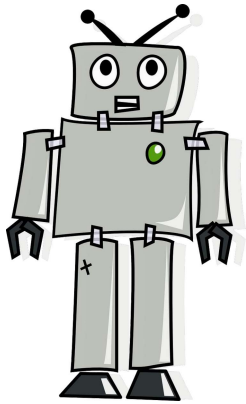
Understanding Dataset Shape

Number of Columns

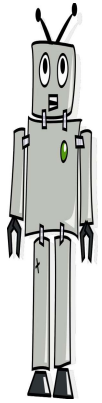
Number of Rows



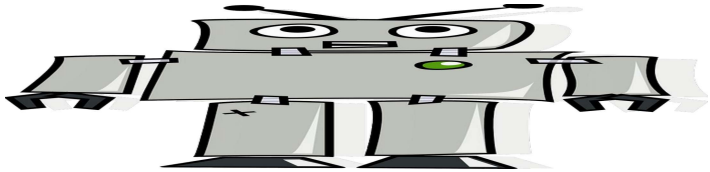
Small Dataset



Even Height and Width

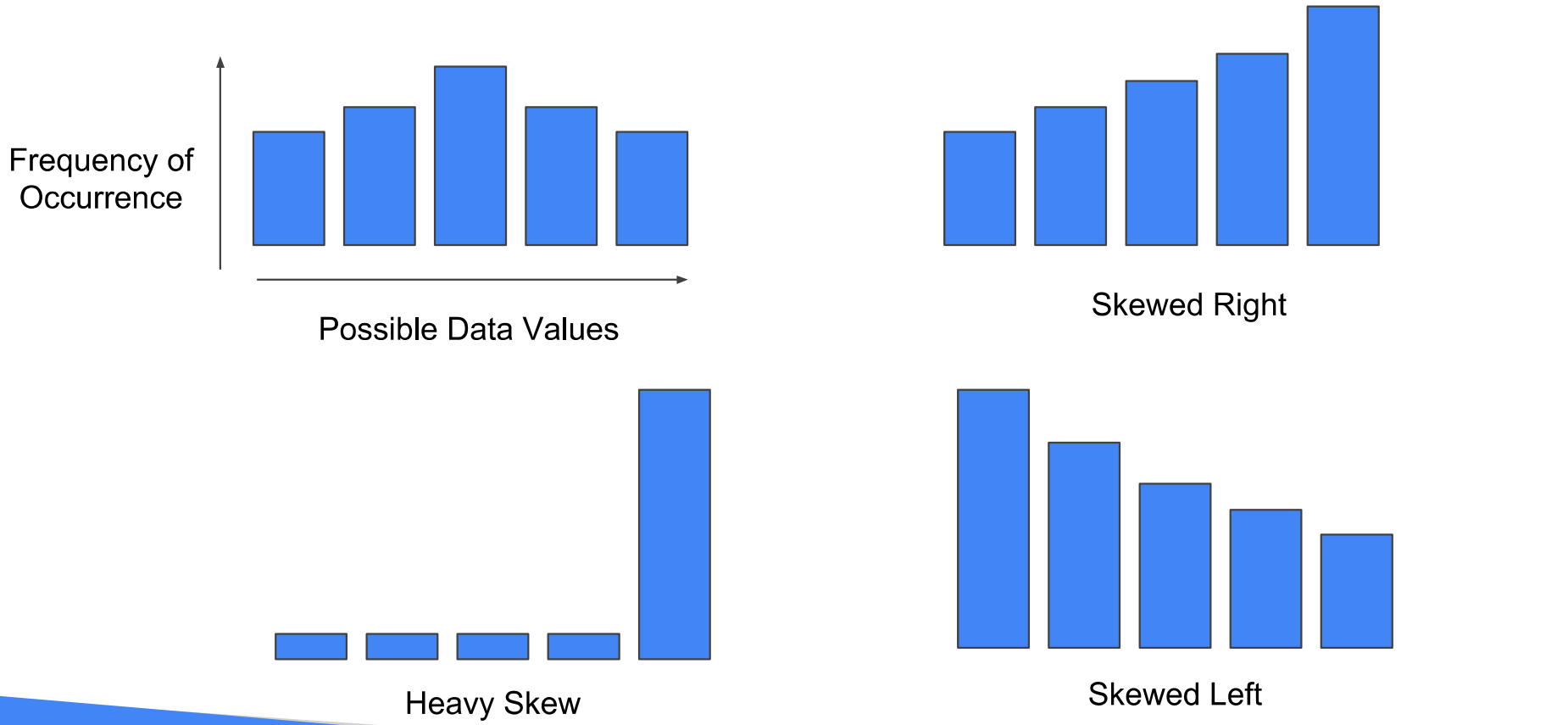


Taller than Wide



Wide but Short

Understanding Dataset Skew (Distribution of Values)



Module 5

Cleaning and Transforming your Data

In this module we will:

- Examine the 5 Principles of Dataset Integrity
- Characterize Dataset Shape and Skew
- **Clean and Transform Data using SQL**
- Clean and Transform Data using a new UI: Introducing Cloud Dataprep

Google Cloud

Clean and Transform Data with SQL

1

Validity

Data conforms to your business rules

Challenges

Out of Range

Empty Fields

Data Mismatch

- Setup Field Data Type Constraints
- Specify fields as NULLABLE or REQUIRED
- Proactively check for NULL values
- Check and Filter for Allowable Range values
  - SQL Conditionals: CASE WHEN, IF ( )
- Require Primary Keys / Relational Constraints in upstream source systems (remember, BigQuery is an analytics warehouse not your primary operational database)

Google Cloud

Clean and Transform Data with SQL

2

Accuracy

Data conforms to an objective true value.

Challenges

Lookup Datasets Do Not Exist

- Create test cases or calculated fields to check values
  - SQL: (quantity_ordered * item_price) AS sub_total
- Lookup values against an objective reference dataset
  - SQL: IN( ) with a subquery or JOIN

Google Cloud

Clean and Transform Data with SQL

3

Completeness

Create, save, and store datasets.

Challenges

Missing Data

- Thoroughly explore the existing dataset shape and skew and look for missing values
  - SQL: NULLIF( ) , IFNULL( ), COALESCE( )
- Enrich the existing dataset with others using UNIONs and JOINs
  - SQL: UNION, JOIN
  - Example: Multiple years of historical data are available for analysis

Google Cloud

Clean and Transform Data with SQL

4

Consistency

Derive insights from data.

Challenges

Duplicate Records

Concurrency Issues

- Store one fact in one place and use IDs to lookup
- Use String Functions to clean data
  - PARSE_DATE( )
  - SUBSTR( )
  - REPLACE( )

Google Cloud

Clean and Transform Data with SQL

5

Uniformity

Explore and present data

Challenges

Same Units of Measurement

- Document and comment your approach
- Use FORMAT ( ) to clearly indicate units
- CAST( ) data types to the same format and digits
- Label all visualizations appropriately

Google Cloud

Tricky NULLs when Filtering Out Missing Values

```
#standardSQL
SELECT * FROM
`bigquery-public-data.noaa_gsod.stations`
WHERE state IS NOT NULL
LIMIT 10
```

Why does the below query still show blank state values when we clearly filtered on IS NOT NULL?

Results | Explanation | Job Information

Row	usaf	wban	name	country	state	call	lat	lon	elev	begin	end
1	007011	99999	CWOS 07011				null	null		20120101	20121129
2	007005	99999	CWOS 07005				null	null		20120127	20120127
3	007025	99999	CWOS 07025				null	null		20120127	20120127
4	007044	99999	CWOS 07044				null	null		20120127	20120127
5	007047	99999	CWOS 07047				null	null		20120613	20120717
6	007083	99999	CWOS 07083				null	null		20120713	20120717
7	007034	99999	CWOS 07034				null	null		20121024	20121106
8	007084	99999	CWOS 07084				null	null		20121214	20121217
9	007094	99999	CWOS 07094				null	null		20121217	20121217

Table | JSON

First < Prev Rows 1 - 9 of 10

Google Cloud

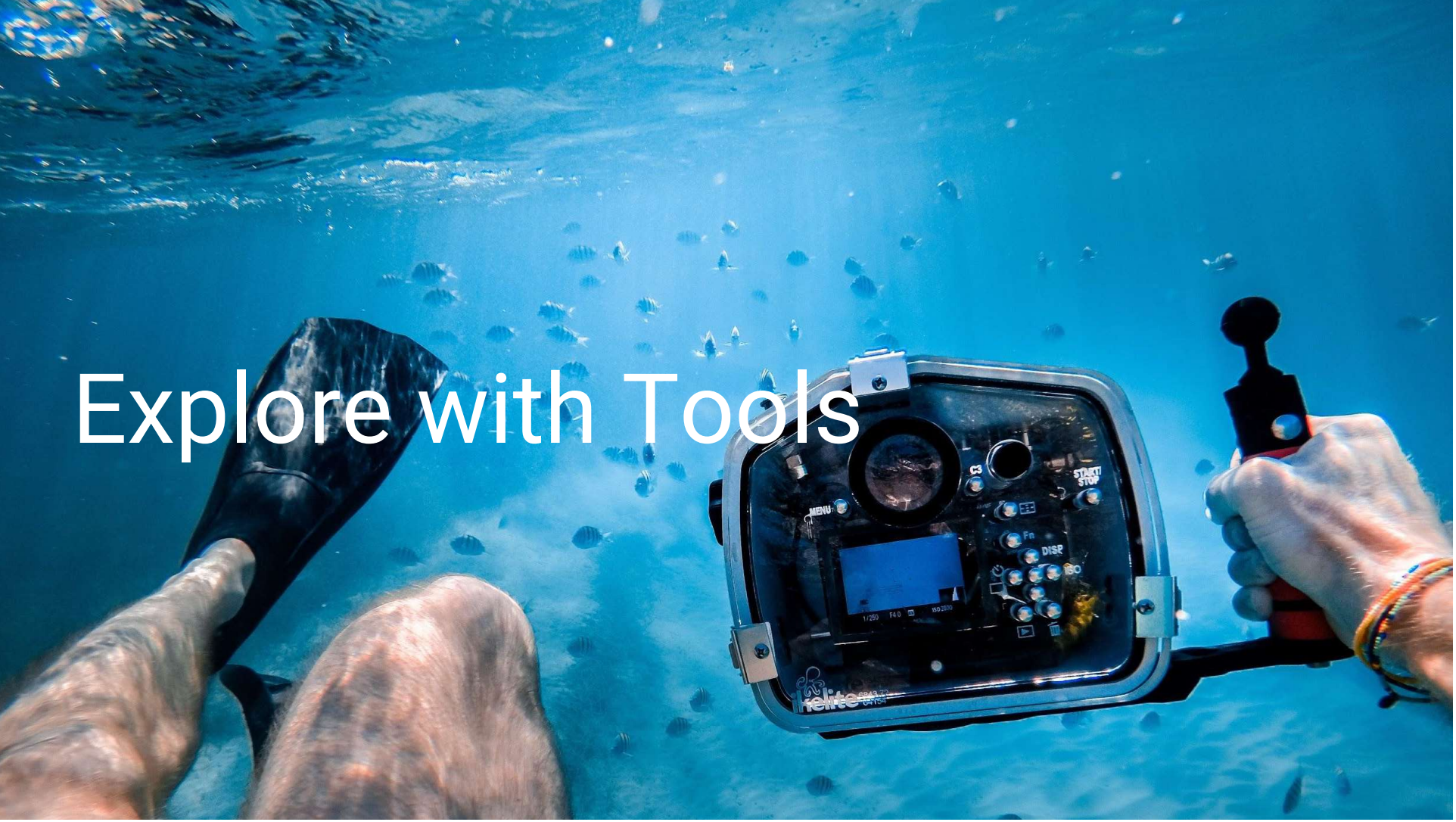
Module 5

Cleaning and Transforming your Data

In this module we will:

- Examine the 5 Principles of Dataset Integrity
- Characterize Dataset Shape and Skew
- Clean and Transform Data using SQL
- **Clean and Transform Data using a new UI: Introducing Cloud Dataprep**

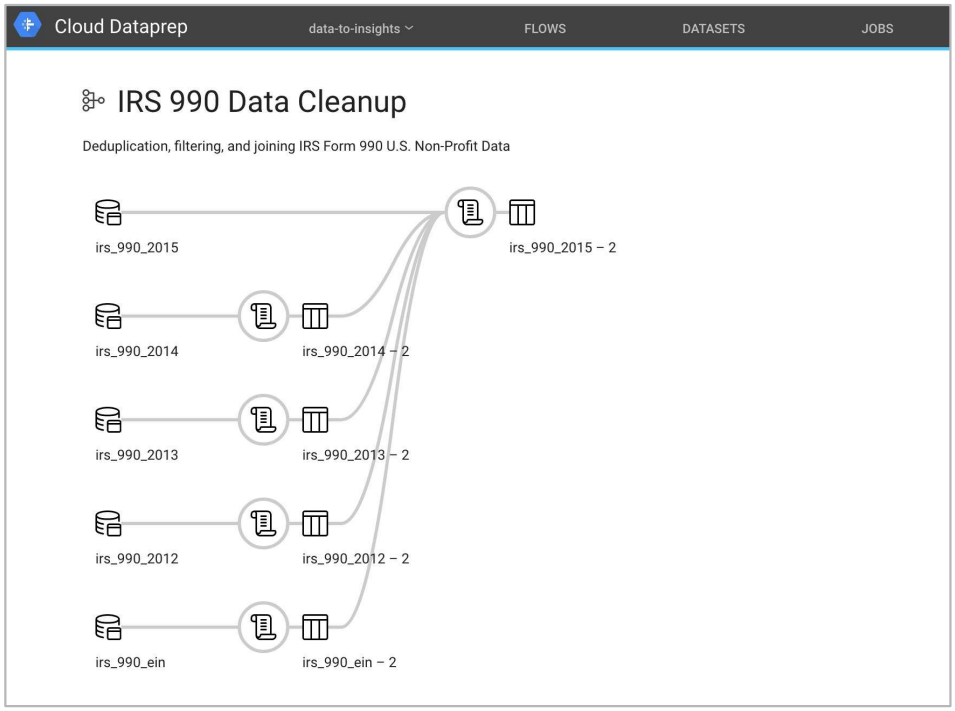
Google Cloud


Explore with Tools

Create Repeatable Data **Transformation Flows in a UI**

Use Flows to wrangle your data.

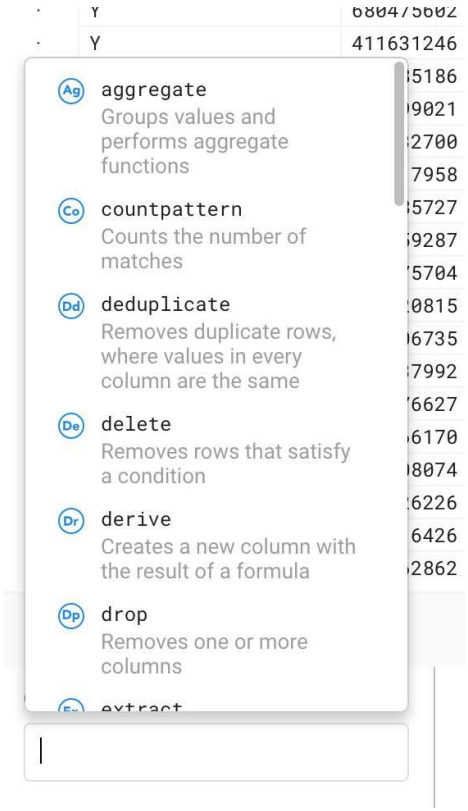
Create Flow



Google Cloud

Transform Data with a Variety of **Predefined Wranglers**

- Use the Cloud Dataprep GUI to create and preview data preparation steps
- Chain together multiple wranglers into a repeatable recipe
- Common tasks like record deduplication and derived fields



Google Cloud

Chain Transformation Rules Together into a **Recipe**

- Repeatable set of transformation steps build by chaining data wranglers together
- Jobs run against recipes
- Can include end-to-end steps from ingestion, transformation, aggregation, save to BigQuery

Sp Break into rows using `'\n'` as a delimiter

Sp Split `column1` into 246 columns on `/ /`

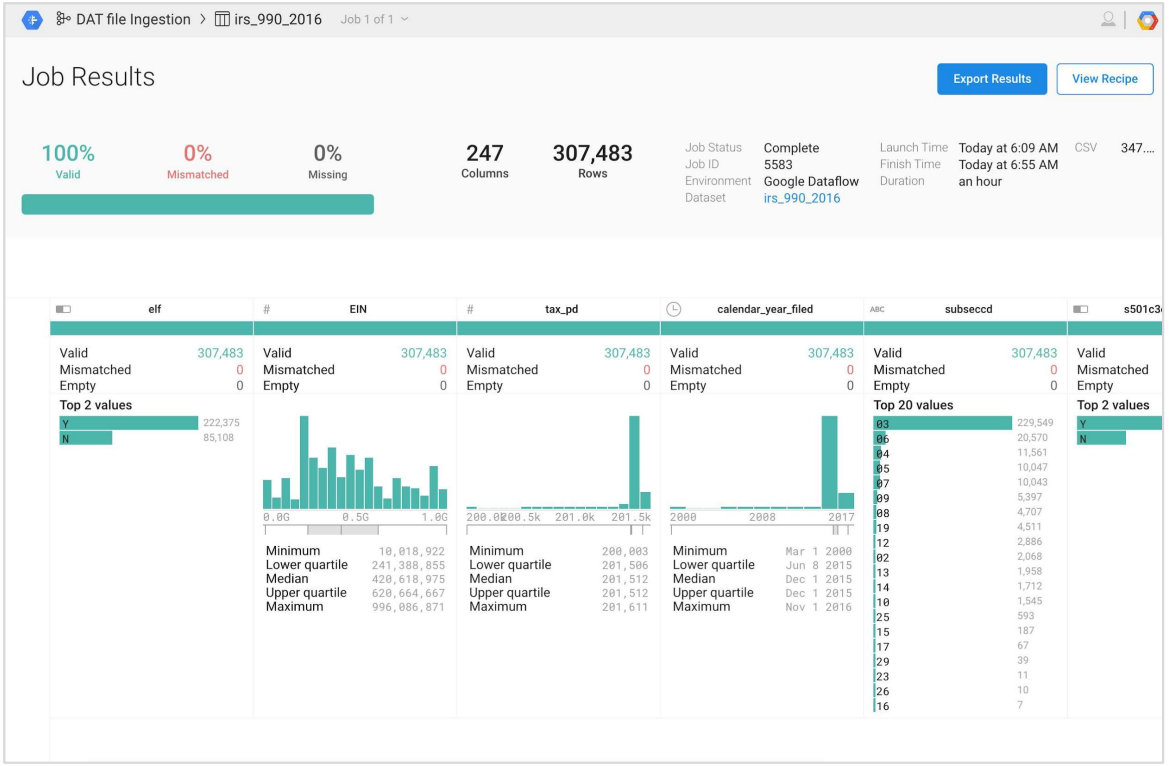
He Convert row 1 to header

Se Change `EIN` type to `Integer`

De Create `calendar_year_filed` from Concatenate 3 functions

Monitor Jobs and **Save Results** as a New Table in BigQuery

- Track completed and ongoing jobs
- See the data quality metrics for transformed datasets
- View histograms with summary statistics for each field



Google Cloud

Lab 4a

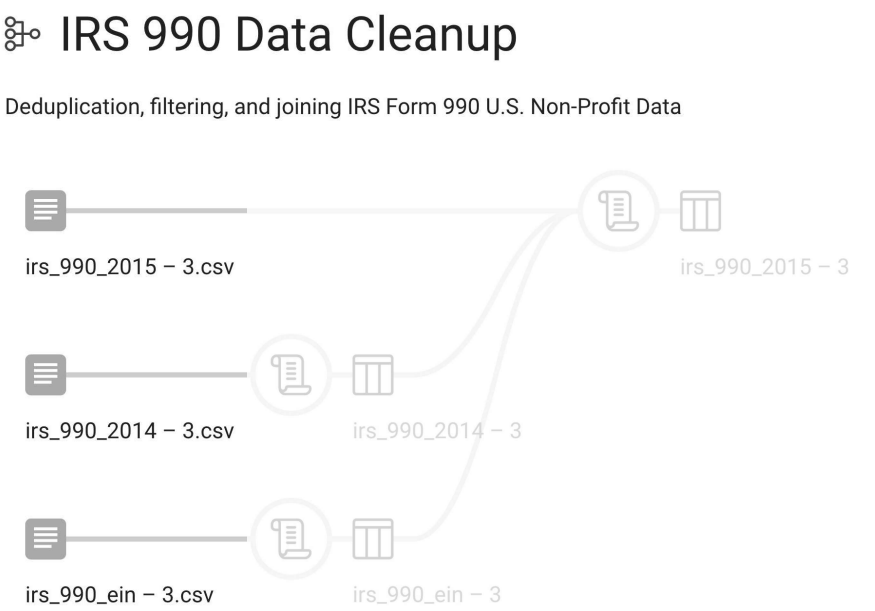
Explore and Load Data with Cloud Dataprep

Google Cloud

Transform your data with Cloud Dataprep

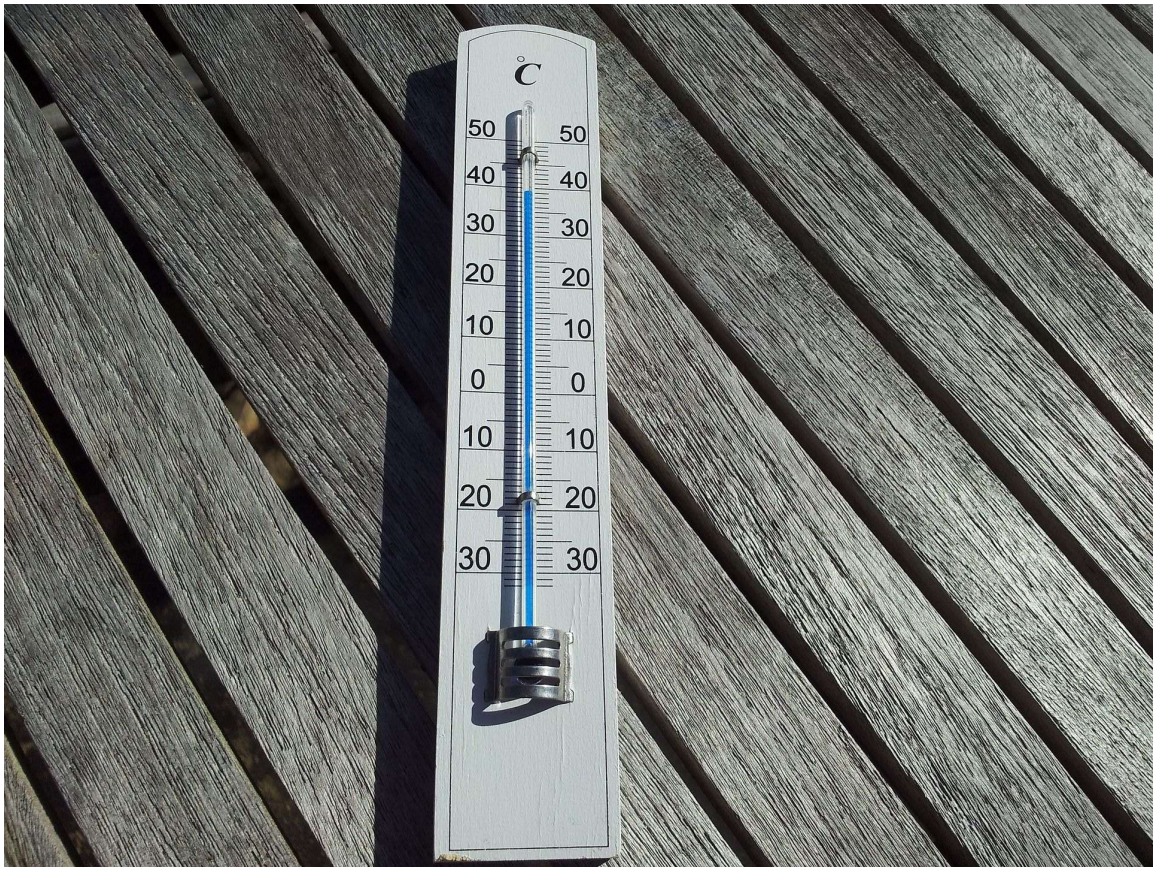
Cloud Dataprep is Google's self-service data preparation tool.

In the first part of this lab, we will load data sources as part of a new flow.



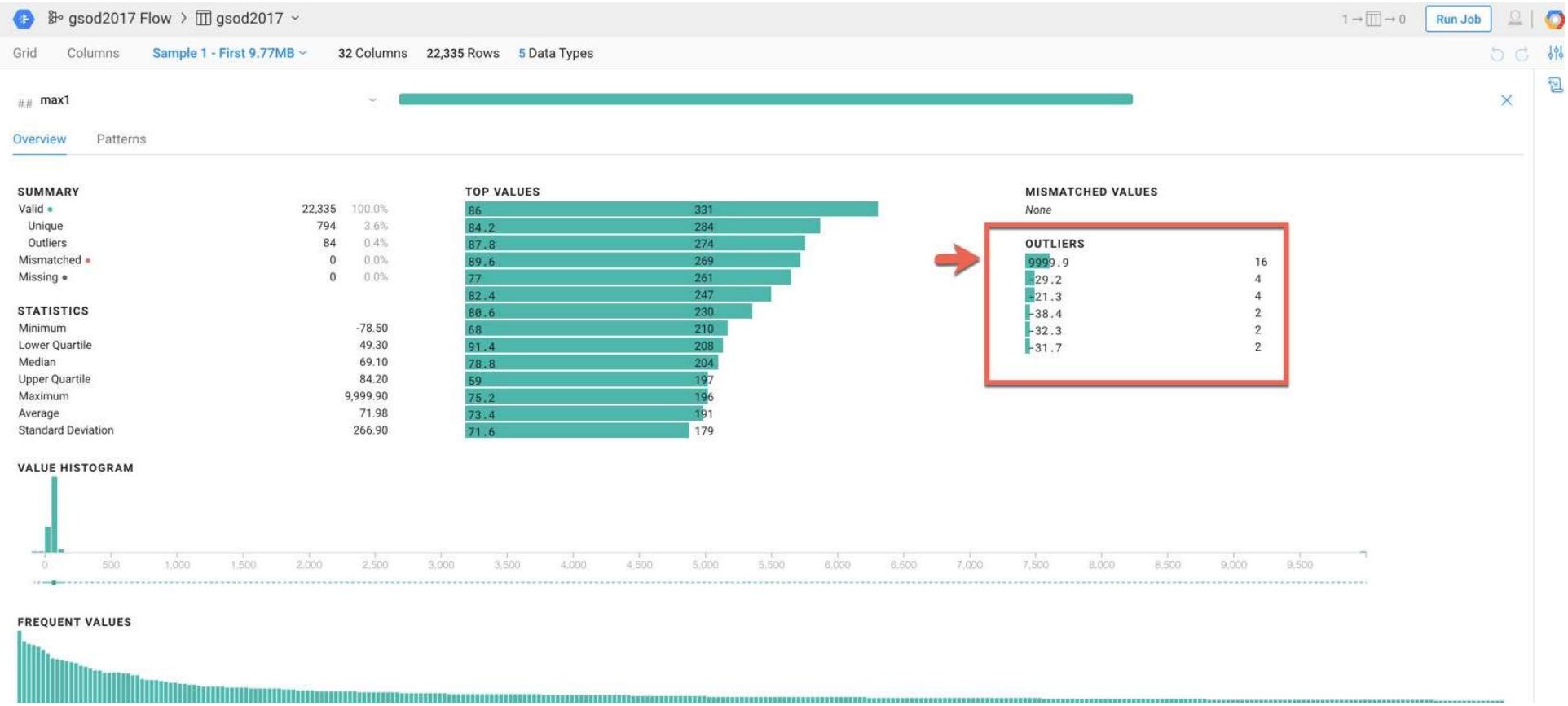
Google Cloud

Cleaning NOAA Temperature Data with Cloud Dataprep

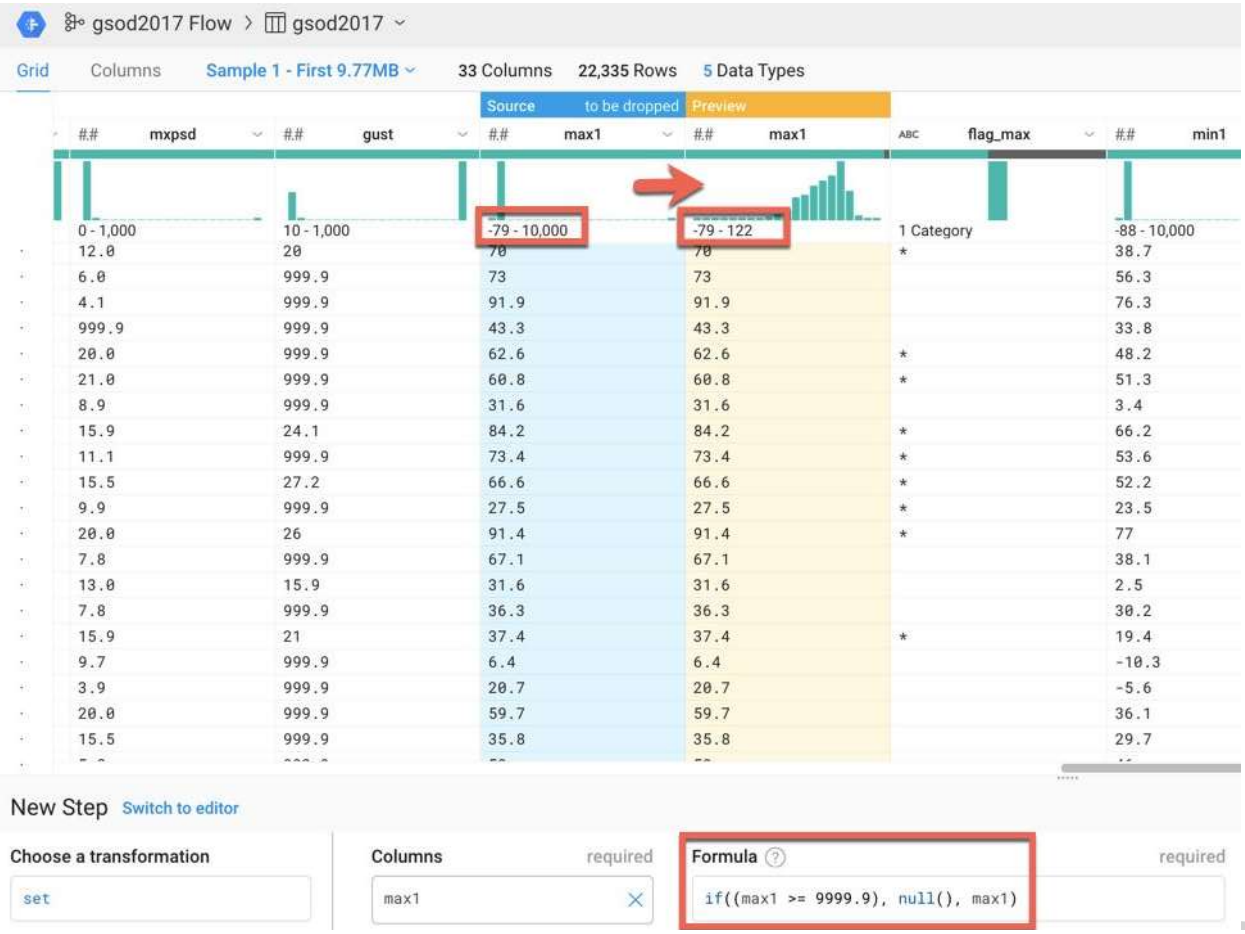


Google Cloud

Using **Column Details** Statistics Reveals Outlier Max Temperature

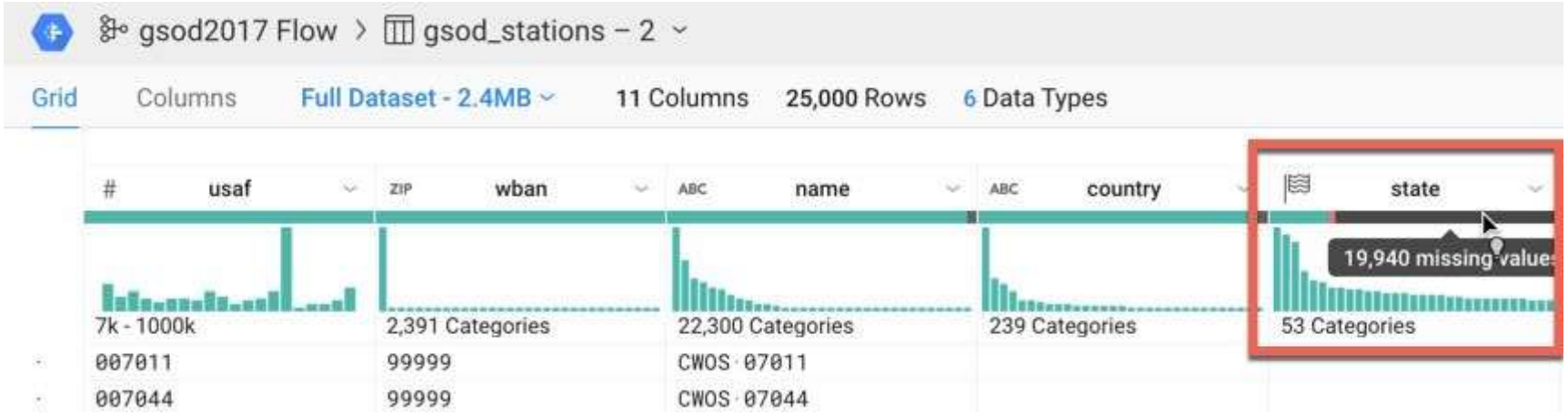


Set the Anomalous 9999.9 Temp Value to NULL with a Formula



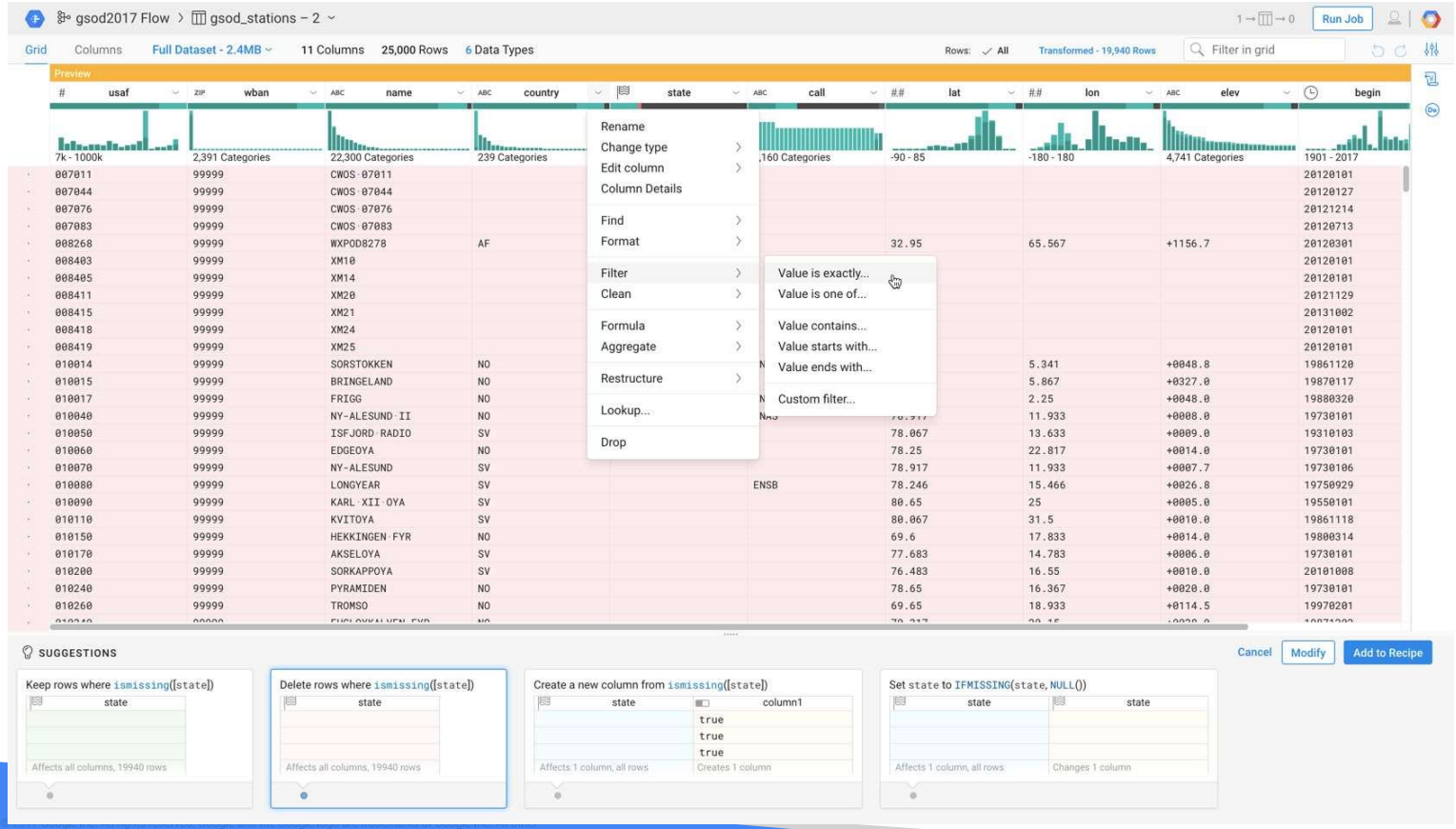
Google Cloud

Looking at the **Data Quality Bar** shows many **States missing**



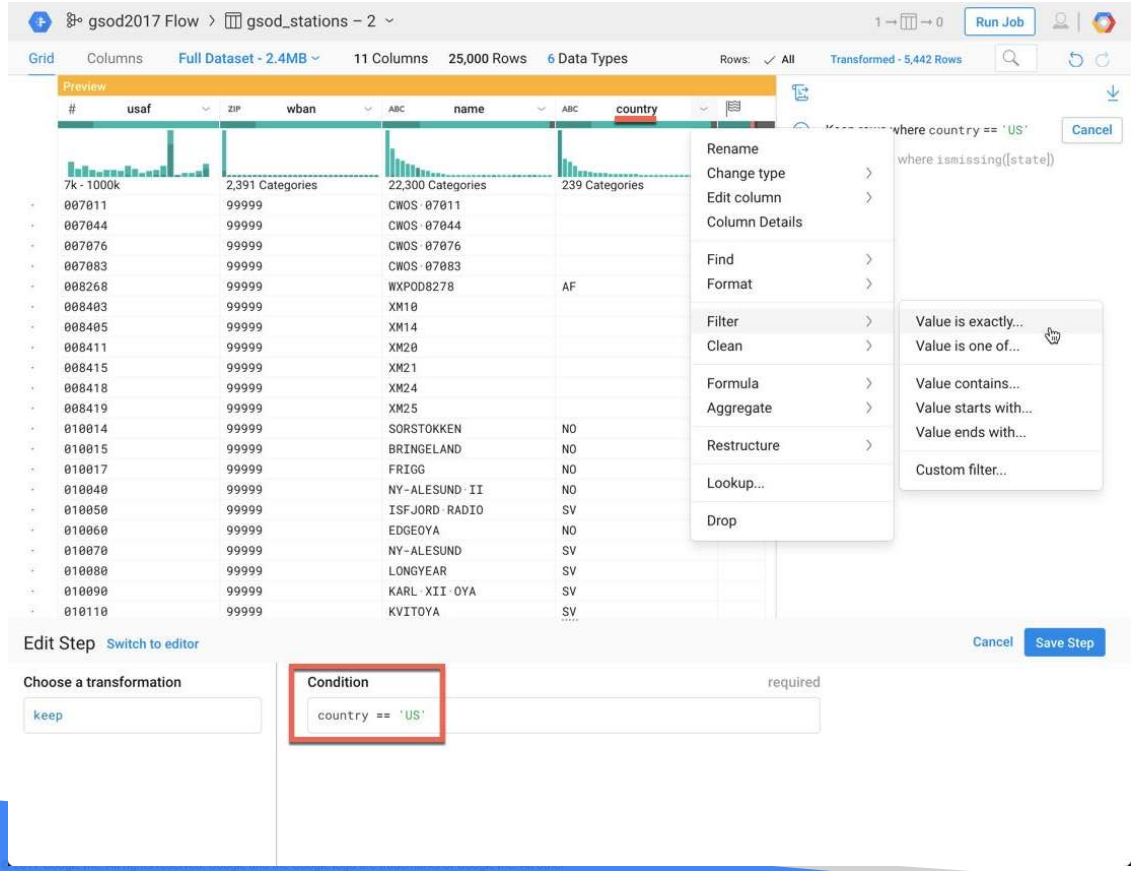
Google Cloud

Filter on U.S. Only Weather Recordings



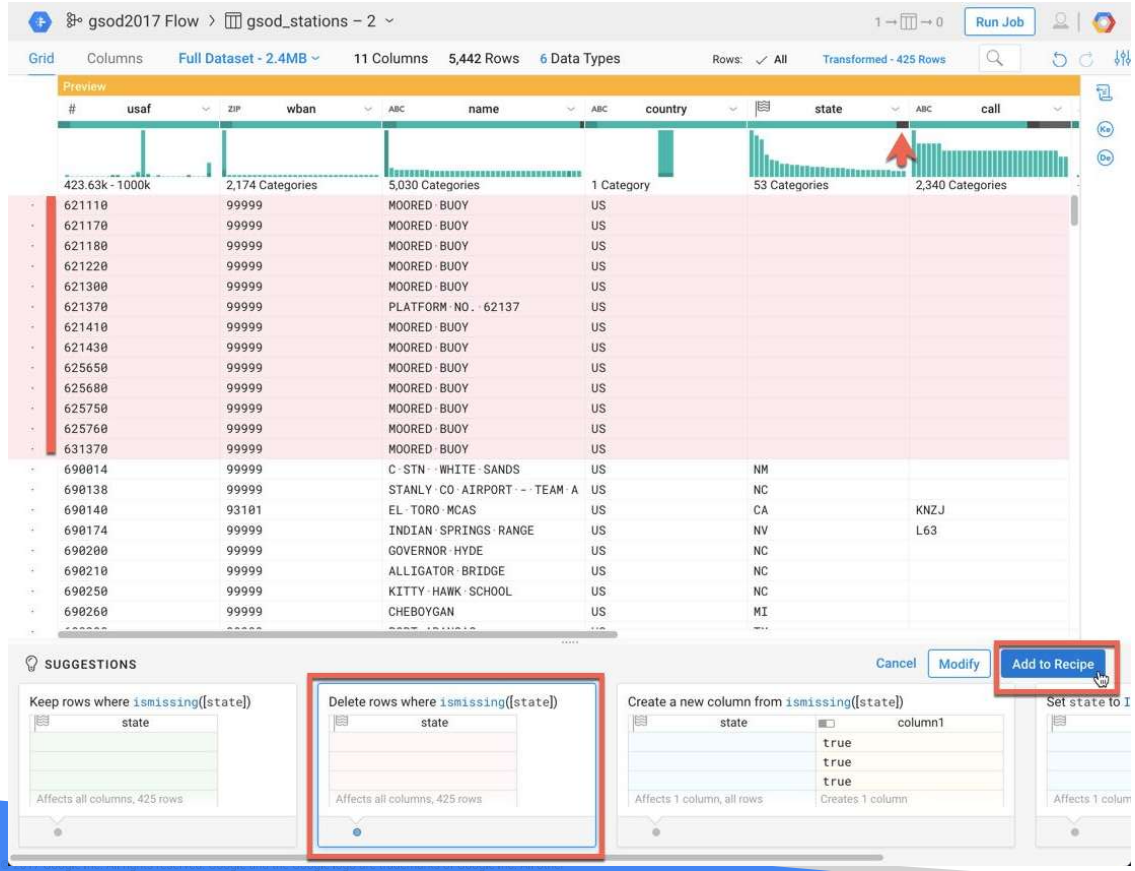
Google Cloud

Keep only U.S. Only Weather Recordings



Google Cloud

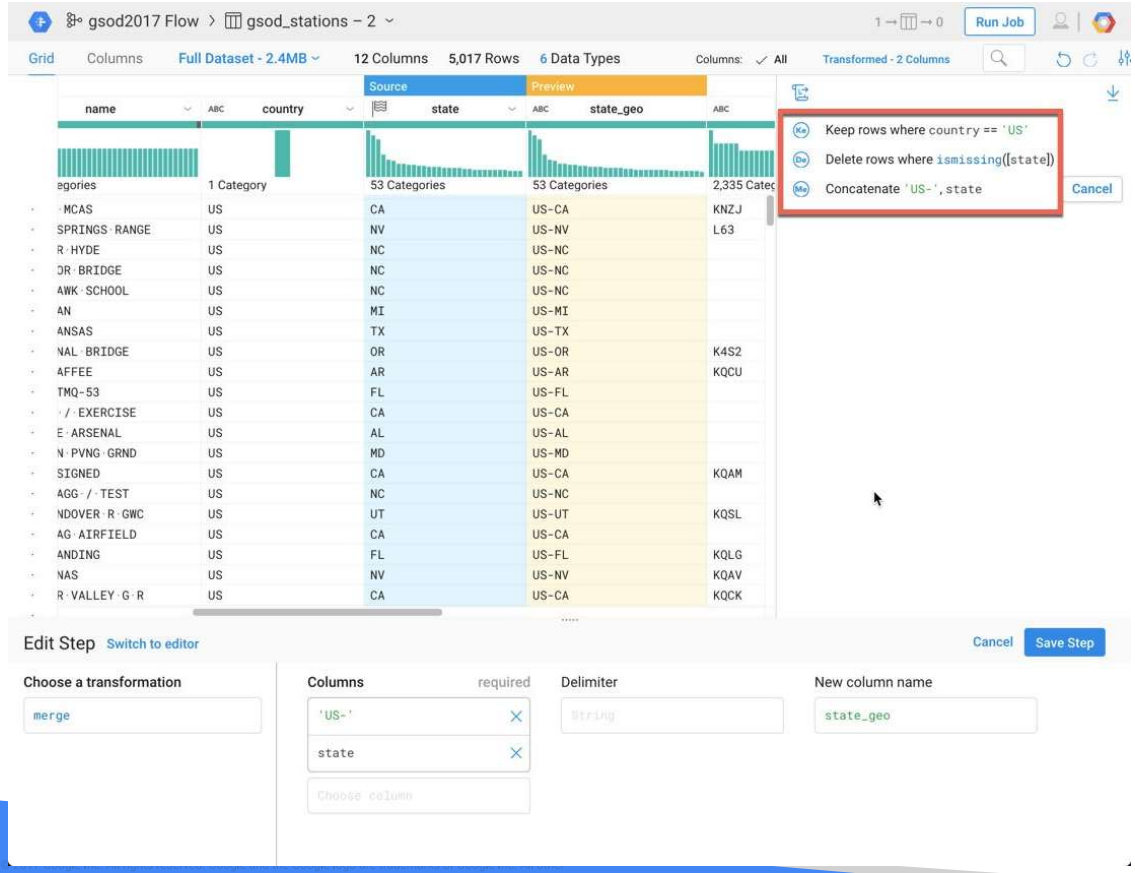
Delete Missing Data for the State Field



- Browse through automatic **suggestion** cards for transformation
- **Modify** to customize your own logic
- **Add to Recipe** when ready

Google Cloud

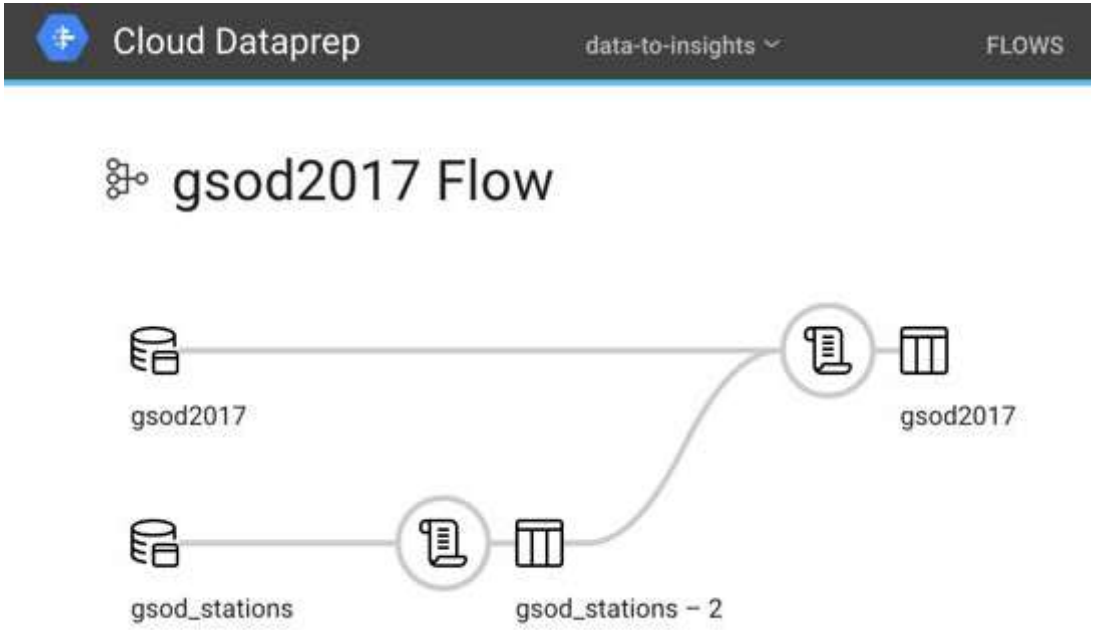
Review Final Recipe and Save



- Toggle open the right side bar to view the steps in your recipe
- Modify or remove steps as needed
- Click **Run Job** when you want to Execute

Google Cloud

Run the Flow which includes our Recipes and Outputs a Table



Google Cloud

Summary: Create clean datasets with SQL and/or Cloud Dataprep

Dataset integrity includes validity, accuracy, completeness, consistency, and uniformity

Explore your data to determine if there is heavy skew which could impact performance

Clean and transform your dataset by writing SQL statements

Clean and transform your dataset through the Cloud Dataprep UI

Google Cloud

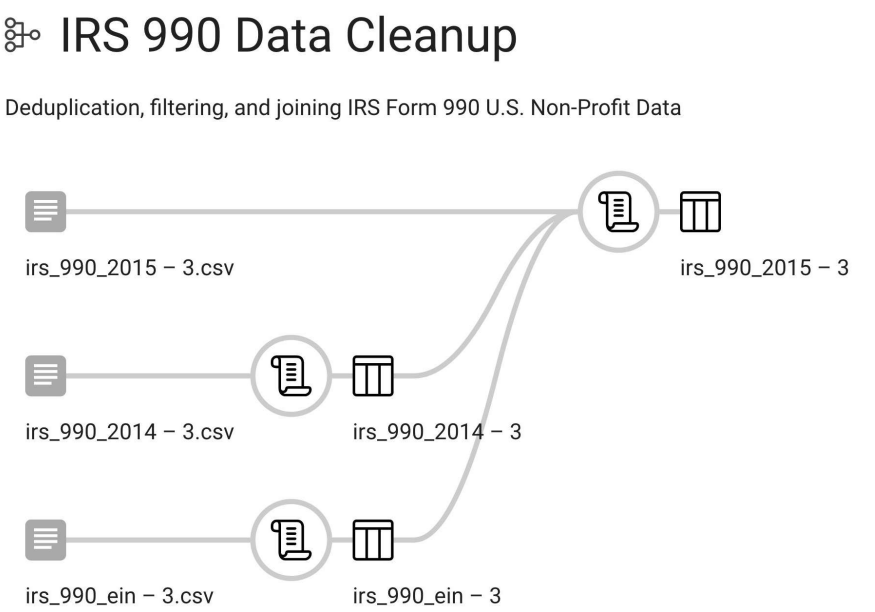
Lab 4b
Transform Data with Cloud Dataprep

Google Cloud

Transform your data with Cloud Dataprep

In the second part of this lab, we will clean, merge, and join our IRS datasets together.

Afterward we will execute our first Cloud Dataprep pipeline job.



Google Cloud