

Module 5

Cleaning and Transforming your Data

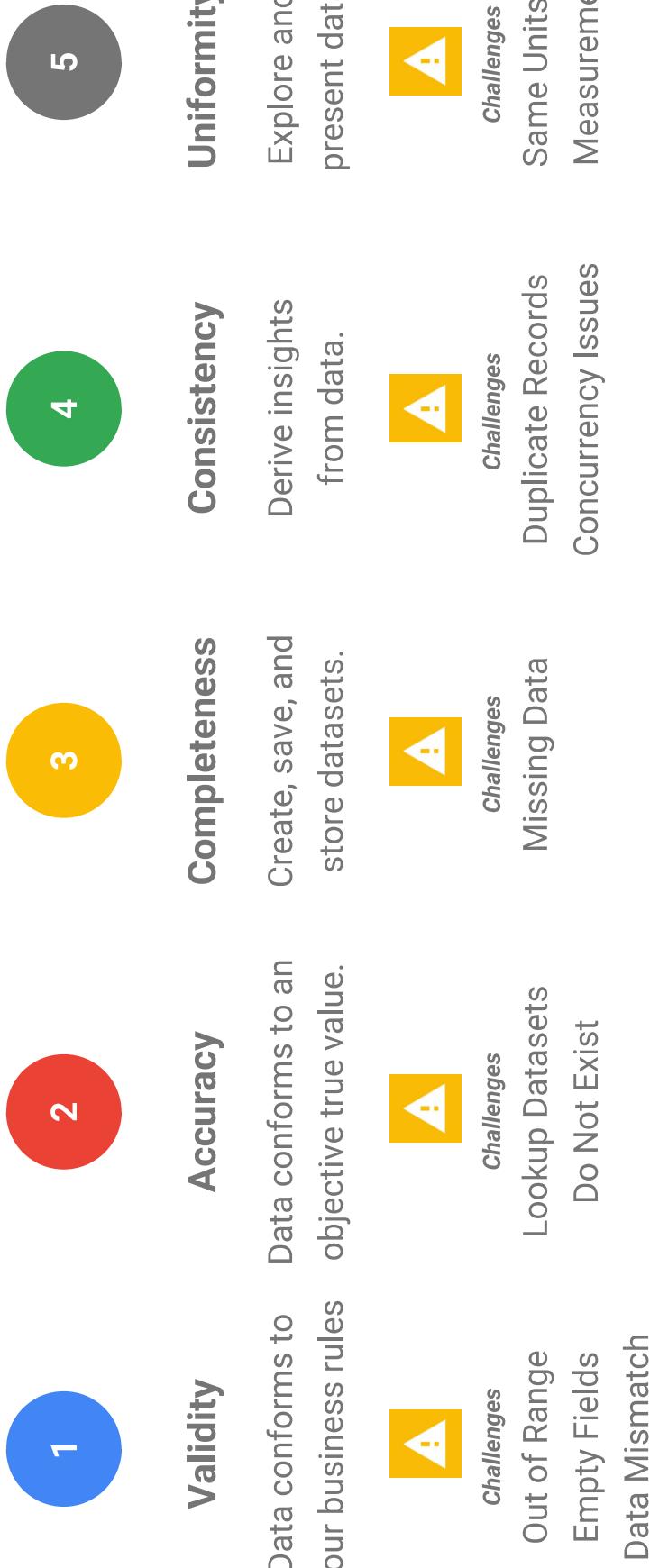
In this module we will:

- **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



Garbage in... garbage out

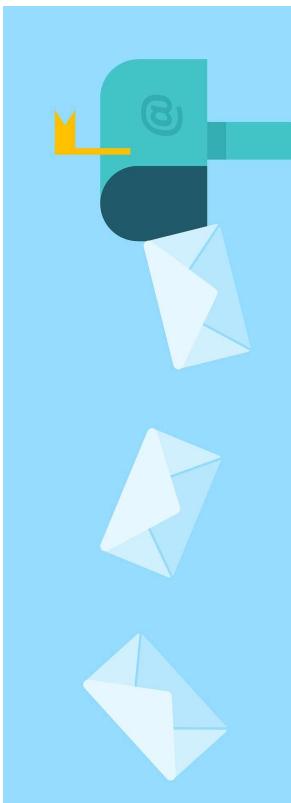
High quality datasets conform to strict integrity rules



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

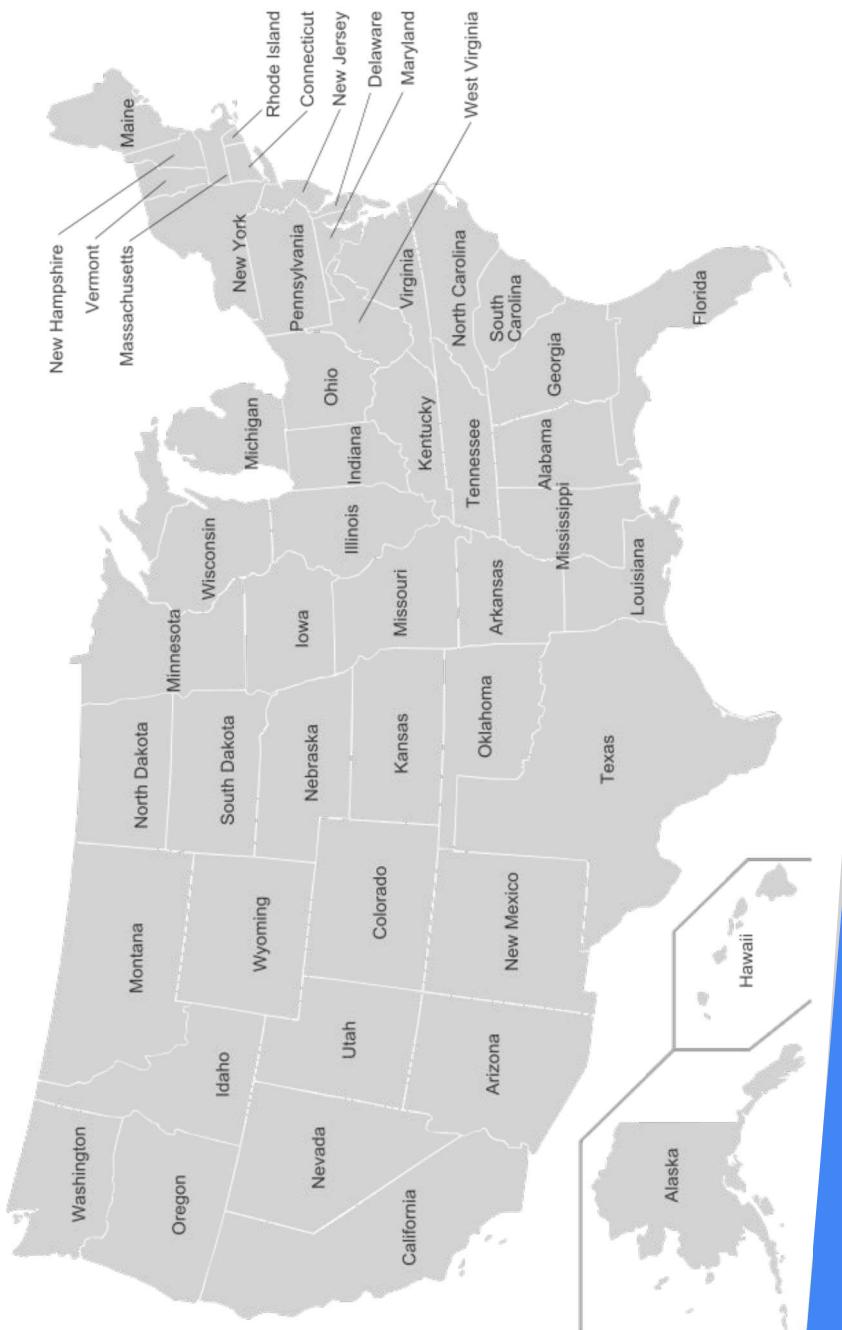


Roll #	Value
1	2
2	2
3	6
4	5
5	1
6	7

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



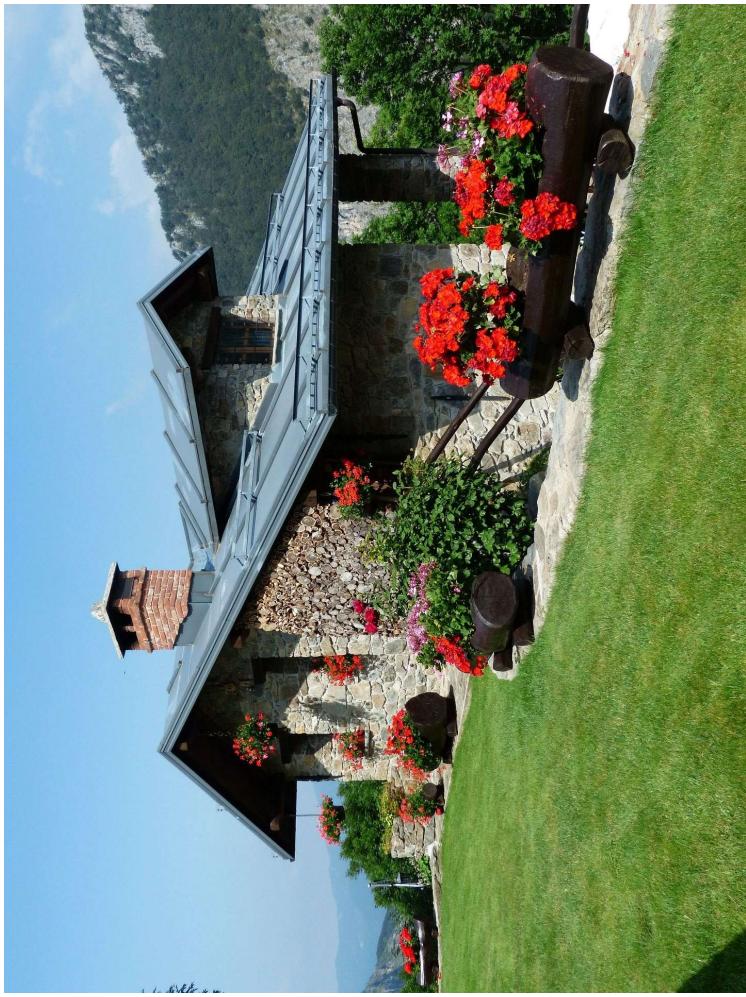
Lamps and Clocks?



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.

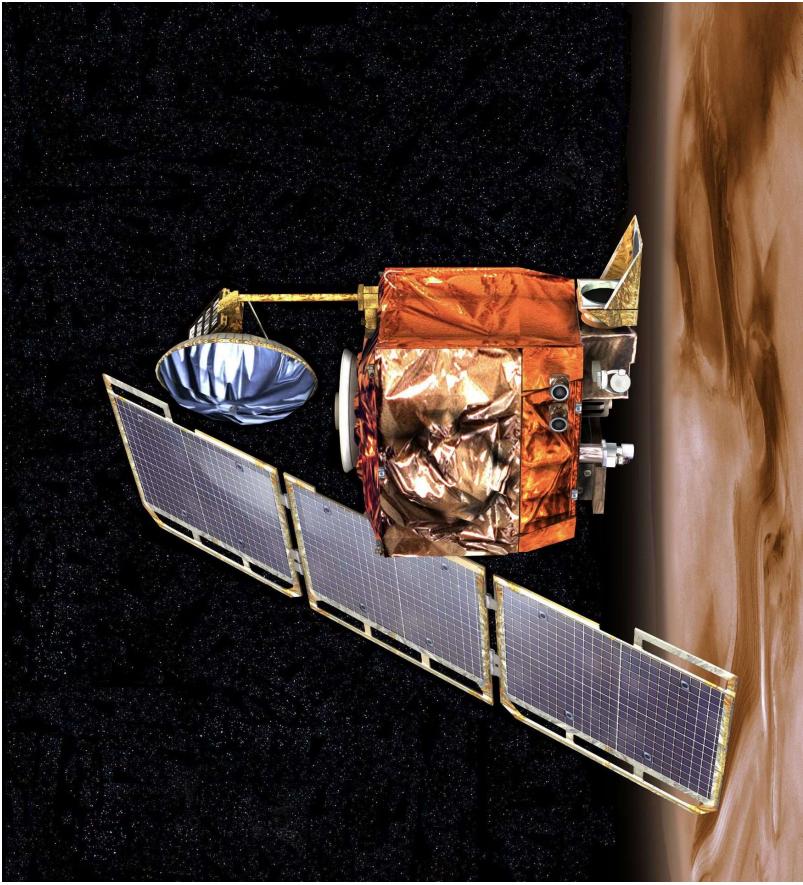
Who owns the house?



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



Module 5

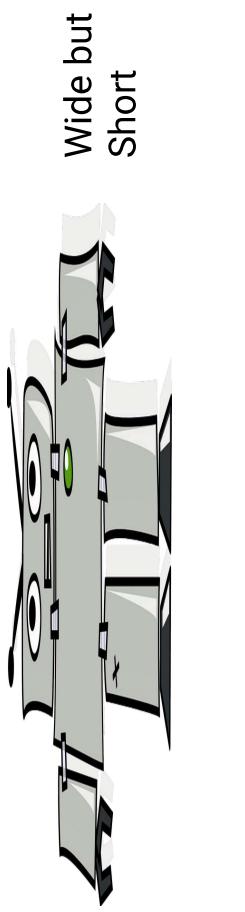
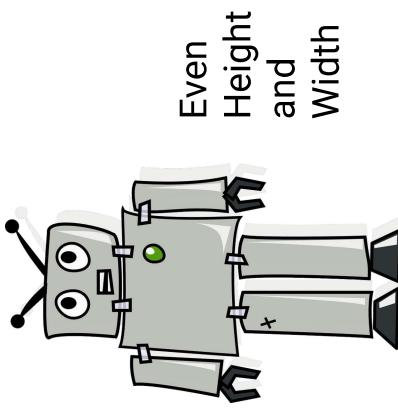
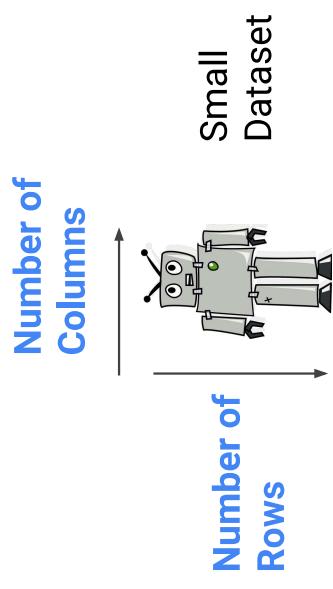
Cleaning and Transforming your Data

In this module we will:

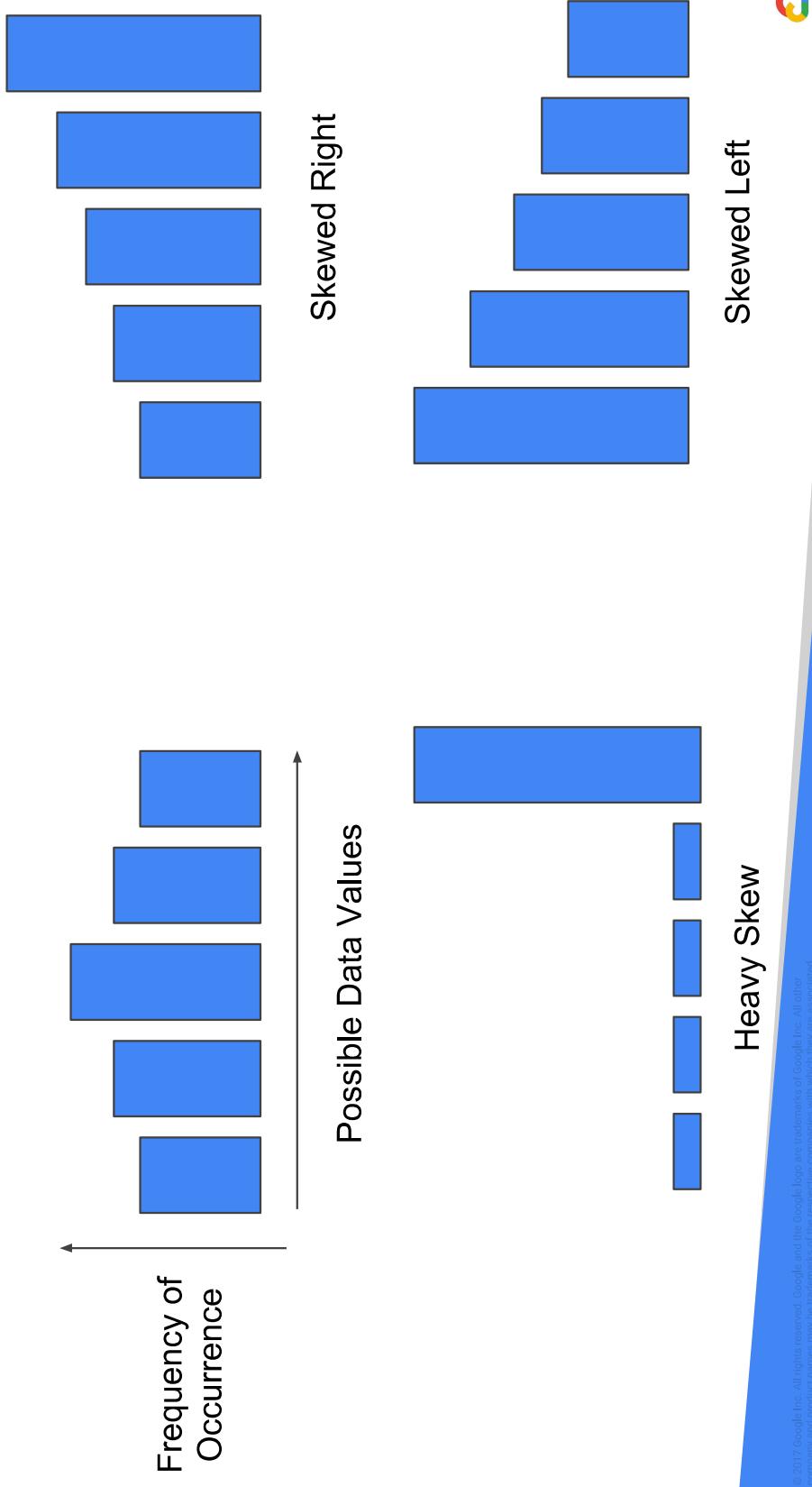
- Examine the 5 Principles of Dataset Integrity
- **Characterize Dataset Shape and Skew**
- Clean and Transform Data using SQL
- Clean and Transform Data using a new UI:
Introducing Cloud Dataprep

Lab: Explore and Shape data with Cloud Dataprep

Understanding Dataset Shape



Understanding Dataset Skew (Distribution of Values)



© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks or registered trademarks of the respective companies with which they are associated.

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

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)

Clean and Transform Data with SQL

2

- 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
- Accuracy**
- Data conforms to an objective true value.



challenges

Lookup Datasets

Do Not Exist

Clean and Transform Data with SQL

17

3

- 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



Challenges

Missing Data

Completeness

Create, save, and store datasets.

Clean and Transform Data with SQL

4

- Store one fact in one place and use IDs to lookup

Consistency

Derive insights
from data.

- Use String Functions to clean data
 - PARSE_DATE()
 - SUBSTR()
 - REPLACE()



Challenges

Duplicate Records

Concurrency Issues

Clean and Transform Data with SQL

5

Uniformity

Explore and present data

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



Challenges
Same Units of
Measurement

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

Module 5

Cleaning and Transforming your Data

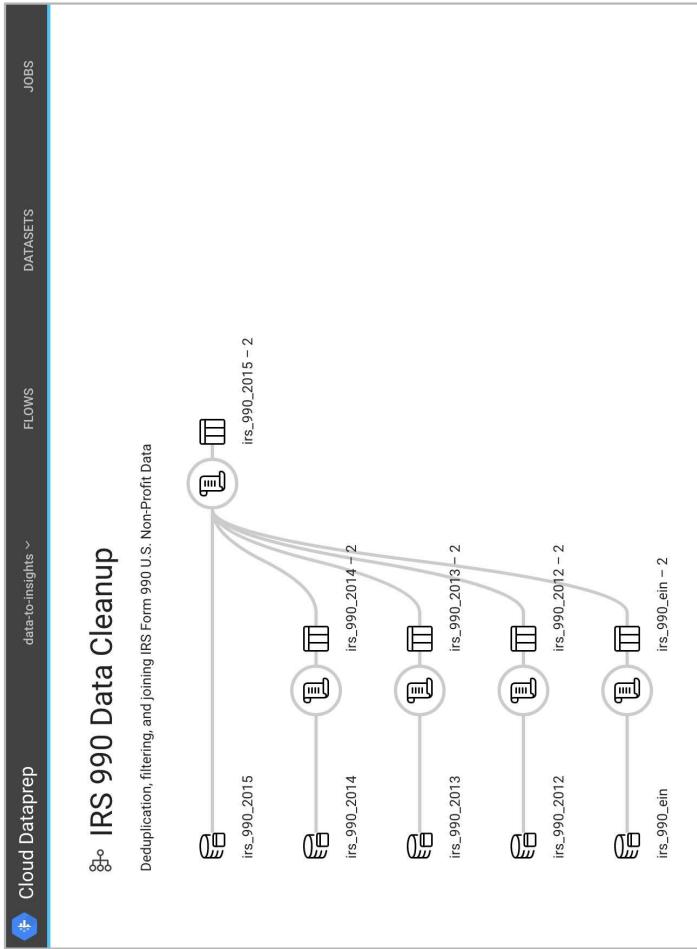
In this module we will:

- 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**

Explore with Tools



Create Repeatable Data Transformation Flows in a UI



Use Flows to wrangle your data.

Create Flow



© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks or registered trademarks of the respective companies with which they are associated.

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



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

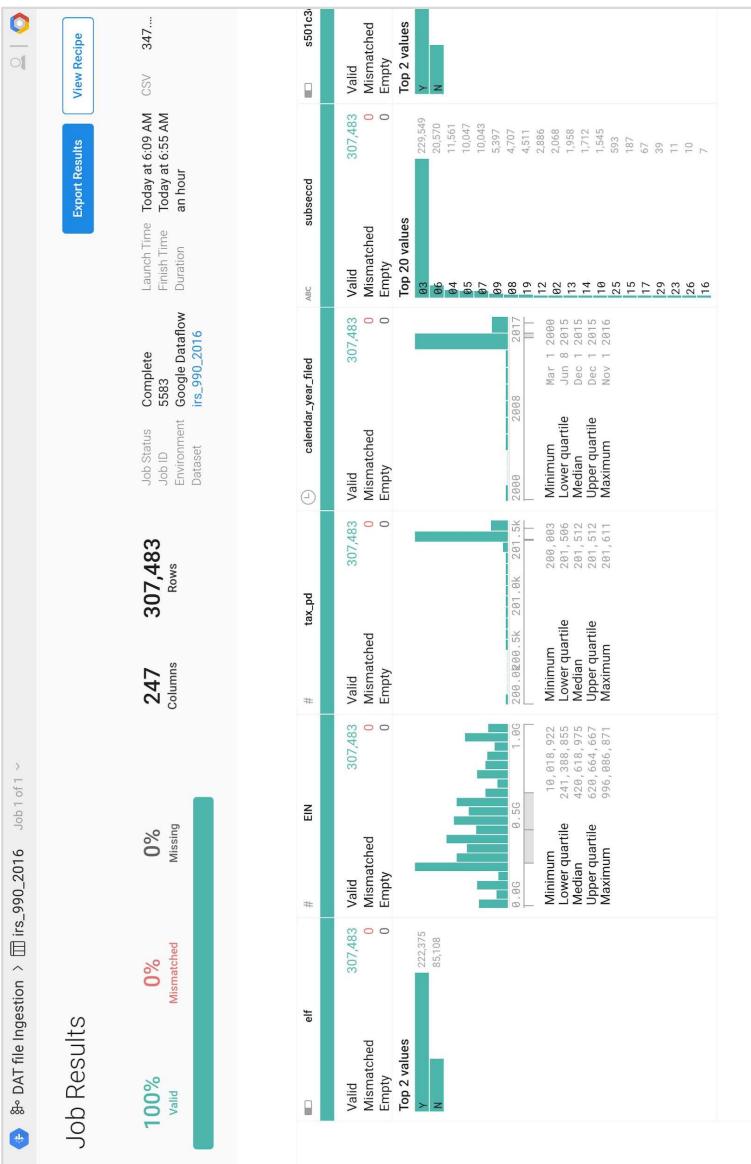


Monitor Jobs and Save Results as a New Table in BigQuery

26

- Track completed and ongoing jobs

- See the data quality metrics for transformed datasets
- View histograms with summary statistics for each field



Lab 4a

Explore and Load Data with Cloud Dataprep

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

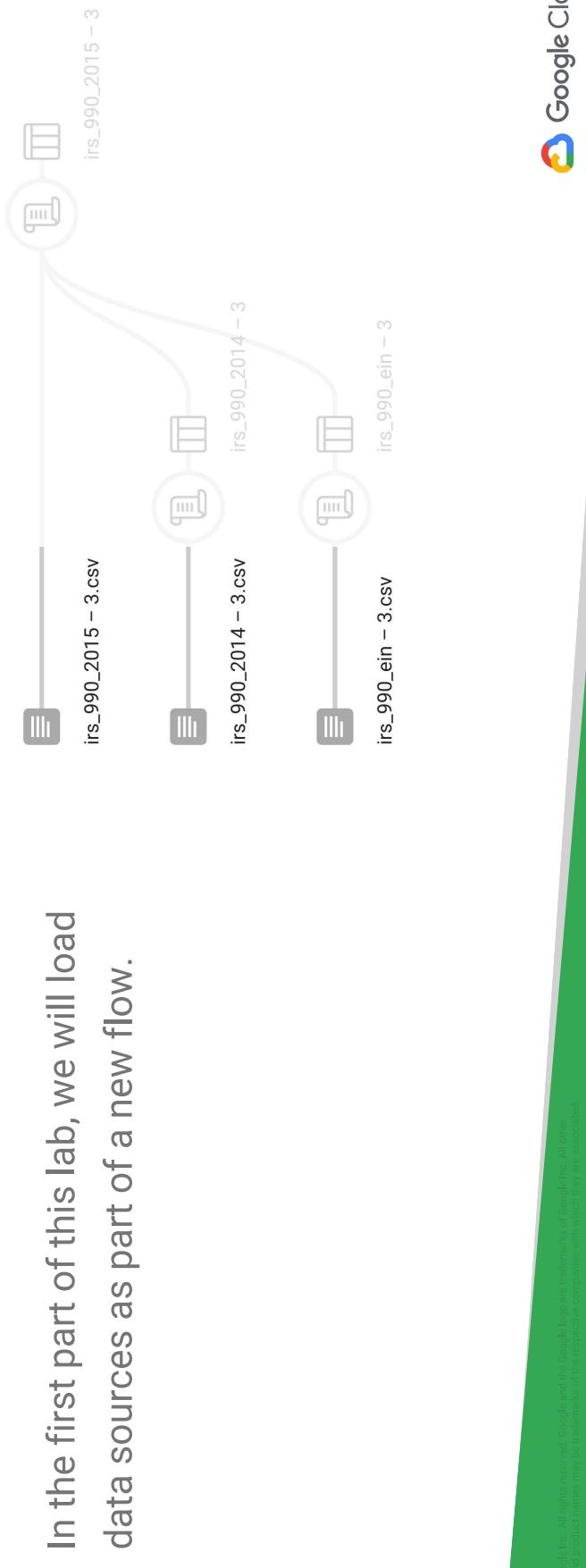
Google Cloud

Transform your data with Cloud Dataprep

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

IRS 990 Data Cleanup

Deduplication, filtering, and joining IRS Form 990 U.S. Non-Profit Data

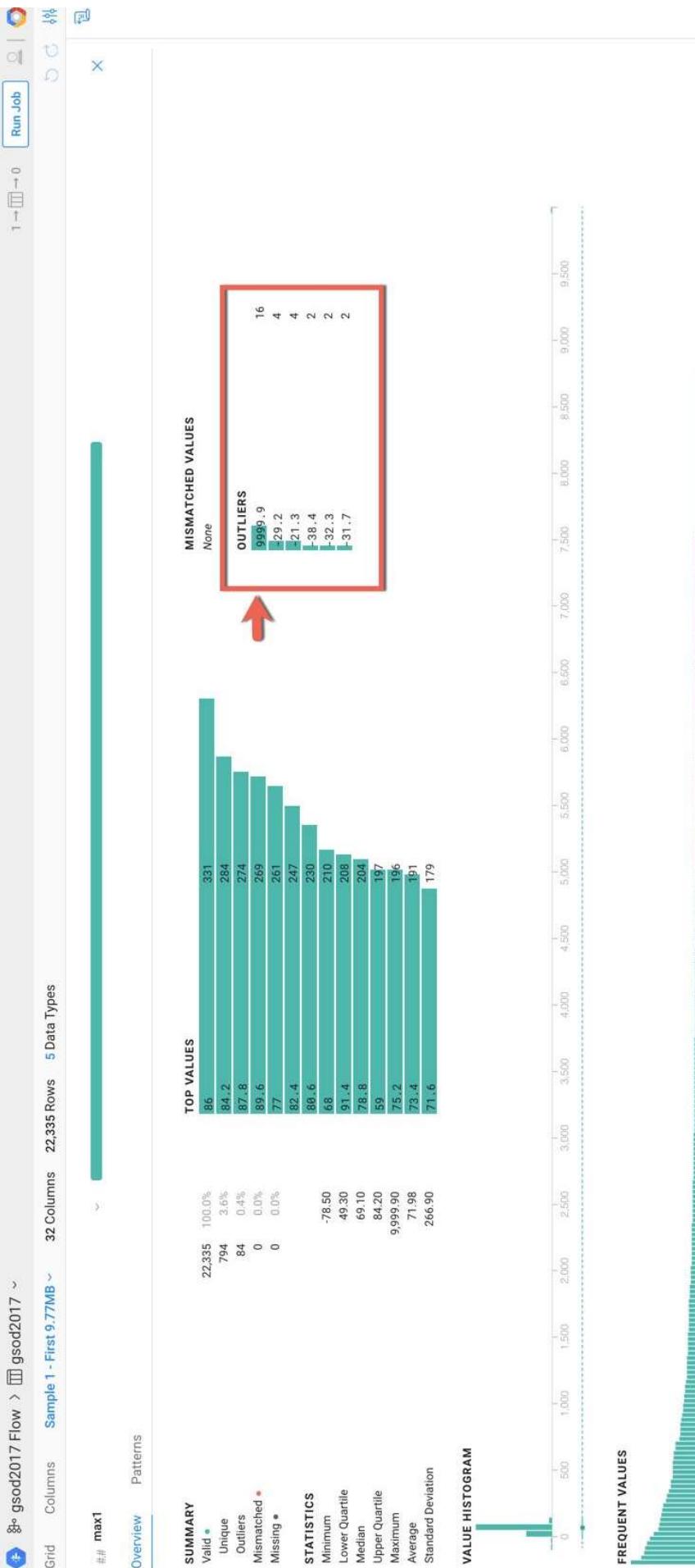




Cleaning NOAA Temperature Data with Cloud Dataprep



Using Column Details Statistics Reveals Outlier Max Temperature



Set the Anomalous 9999.9 Temp Value to NULL with a Formula

gsod2017 Flow > gsod2017

Grid Columns Sample 1 - First 9.77MB ~

Source	to be dropped	Preview
##	max1	max1
mpsd	##	##
gust	~	~
10.-1,000	-79.-10,000	-79.-122
12.0	78	78
6.0	73	73
4.1	91.9	91.9
999.9	43.3	43.3
999.9	62.6	62.6
20.0	*	*
21.0	60.8	60.8
8.9	31.6	31.6
15.9	84.2	84.2
11.1	73.4	73.4
15.5	66.6	66.6
27.2	27.5	27.5
9.9	91.4	91.4
20.0	67.1	67.1
7.8	31.6	31.6
13.0	36.3	36.3
7.8	37.4	37.4
15.9	6.4	6.4
9.7	20.7	20.7
3.9	59.7	59.7
20.0	35.8	35.8
15.5	~	~

5 Data Types

ABC flag_max ~ ### min1

1 Category *

Red boxes highlight the values 10.-1,000, -79.-10,000, and -79.-122.

New Step Switch to editor

Choose a transformation

Columns required

max set

Formula (2) required

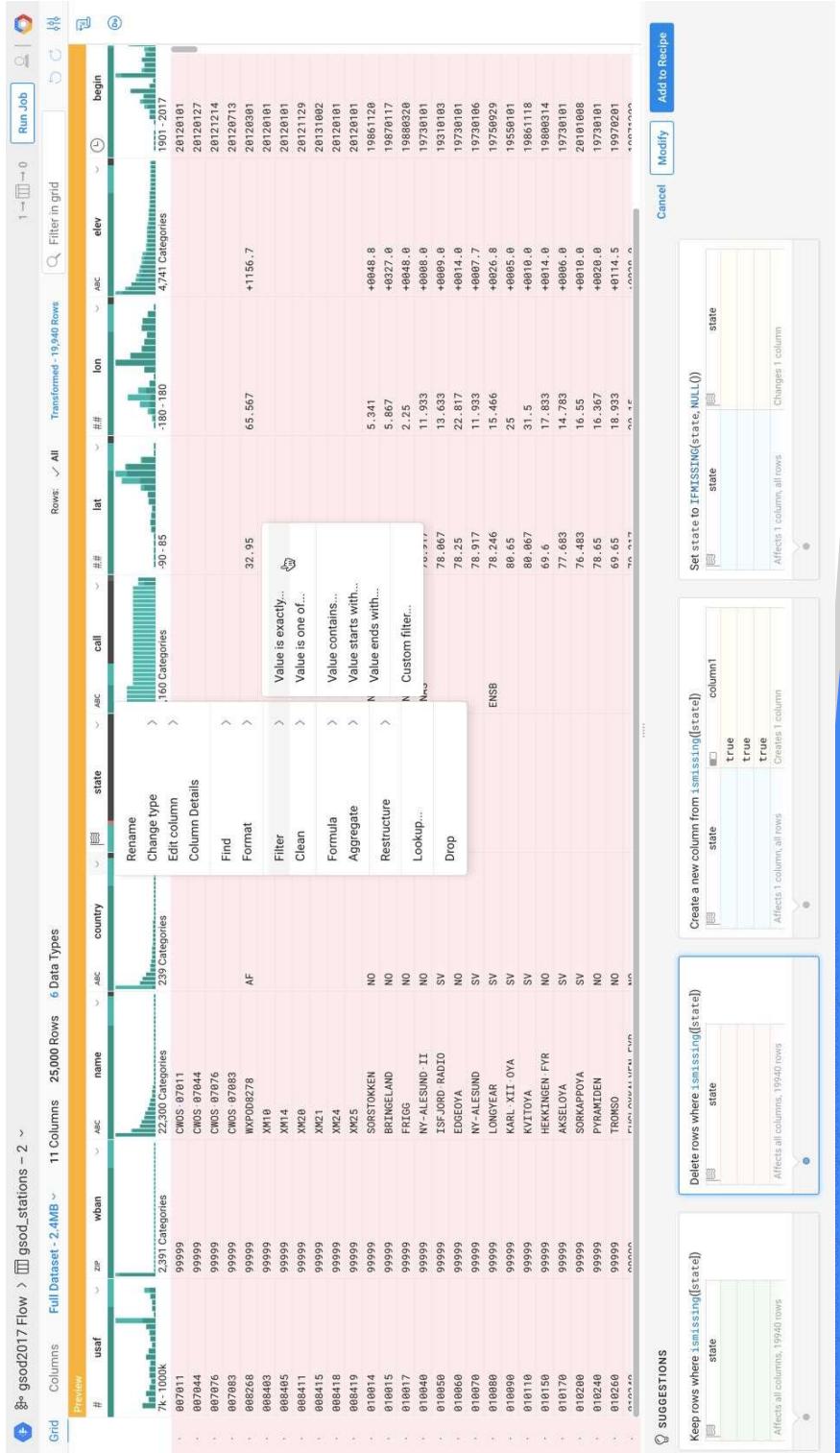
if((max1 >= 9999.9), null(), max1)

Google Cloud

Looking at the Data Quality Bar shows many States missing



Filter on U.S. Only Weather Recordings



Keep only U.S. Only Weather Recordings

The screenshot shows a data processing interface with the following details:

- Job Status:** 1 → 0 → 0 | Run Job
- Dataset:** Full Dataset - 2.41MB
- Columns:** 11 Columns
- Rows:** 25,000 Rows
- Data Types:** 6 Data Types
- Transformation Step:** Rows: ✓ All
- Condition:** country == 'US' (highlighted with a red box)
- Operations:** where 1 < missing(state) | Transform 5,442 Rows
- Options:** Preview, Grid, Columns, Run Job, Stop, Refresh
- Contextual Menu (over country column):**
 - Rename
 - Change type
 - Edit column
 - Column Details
 - Find
 - Format
 - Filter
 - Clean
 - Value is exactly...
 - Value is one of...
 - Formula
 - Aggregate
 - Restructure
 - LookUp...
 - Drop
- Transformation Editor:**
 - Condition: country == 'US'
 - Choose a transformation: Keep
 - Save Step
 - Cancel



company at a detailed level may be trademarks of the respective companies with which they are associated.

Delete Missing Data for the State Field

The screenshot shows a Google Cloud Data Studio interface with several cards:

- Grid Preview:** Shows a preview of a dataset with columns: #, usaf, zip, wban, name, ABC, country, state, ABC, and call. A red arrow points to the 'state' column header.
- Transformation Card:** Titled "Delete rows where ismissing([state])". It has a "Keep rows where ismissing([state])" section with a "state" checkbox, and a "Delete rows where ismissing([state])" section with a "state" checkbox. Both sections have "Affects all columns, 425 rows" notes.
- Transformation Card:** Titled "Create a new column from ismissing([state])". It has a "state" section with a "true" checkbox, and an "Affects 1 column, all rows" note.
- Transformation Card:** Titled "Set state to I". It has a "column1" section with a "true" checkbox, and an "Affects 1 column" note.
- Card:** Titled "Add to Recipe".

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



company at a specific time may be trademarks of the respective companies with which they are associated.

Review Final Recipe and Save

Keep rows where country == 'US'
Delete rows where isMissing(state)
Concatenate US-, state
2335 Categories

Save Step

Cancel

Grid Columns Full Dataset - 2.4MB ~ 12 Columns 5,017 Rows 6 Data Types Preview ABC state geo

name country ABC

agories 1 Category

MCAS US CA

SPRINGS RANGE US NV

R HYDE US NC

JR BRIDGE US NC

ANK SCHOOL US NC

AN US MI

ANSAS US TX

NAL BRIDGE US OR

AFFEE US AR

TNO-53 US FL

/ EXERCISE US CA

E ARSENAL US AL

N PING GRND US MD

SIGNED US CA

AGG - TEST US NC

DOVER R GMC US UT

AG AIRFIELD US CA

ANDING US FL

NAS US NV

R VALLEY G R US CA

US_

state

Choose column

Choose a transformation

merge

New column name

state_geo

Delimiter

US-UT

X

US-CA

X

US-FL

X

US-NV

X

US-CA

US-NC

KOAM

KOOL

KOAV

KICK

.....

Edit Step Switch to editor

Cancel Save Step



company at a detailed level may be problematic if the resulting documents with which they are associated

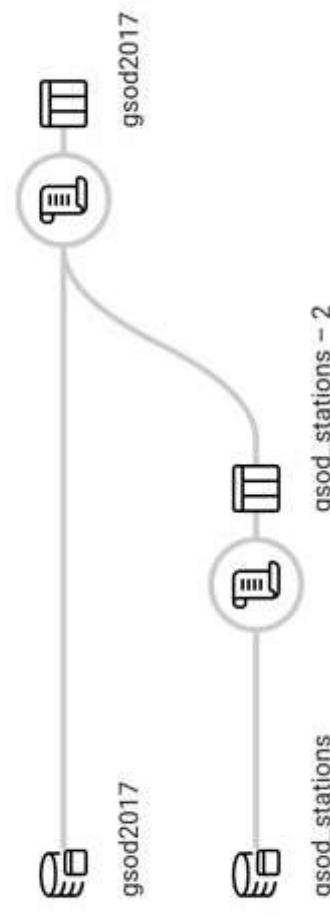
Run the Flow which includes our Recipes and Outputs a Table



Cloud Dataflow

data-to-insights ~

gsod2017 Flow



© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks or registered trademarks of the respective companies with whom they are associated.

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



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



Clean and transform your dataset by writing SQL statements



Clean and transform your dataset through the Cloud Dataprep UI



Lab 4b

Transform Data with Cloud Dataprep

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

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.

IRS 990 Data Cleanup

Deduplication, filtering, and joining IRS Form 990 U.S. Non-Profit Data together.

