Bay Area Bike Share Database

Final Project
Group 4
DATS6102 FALL 2016

The Data

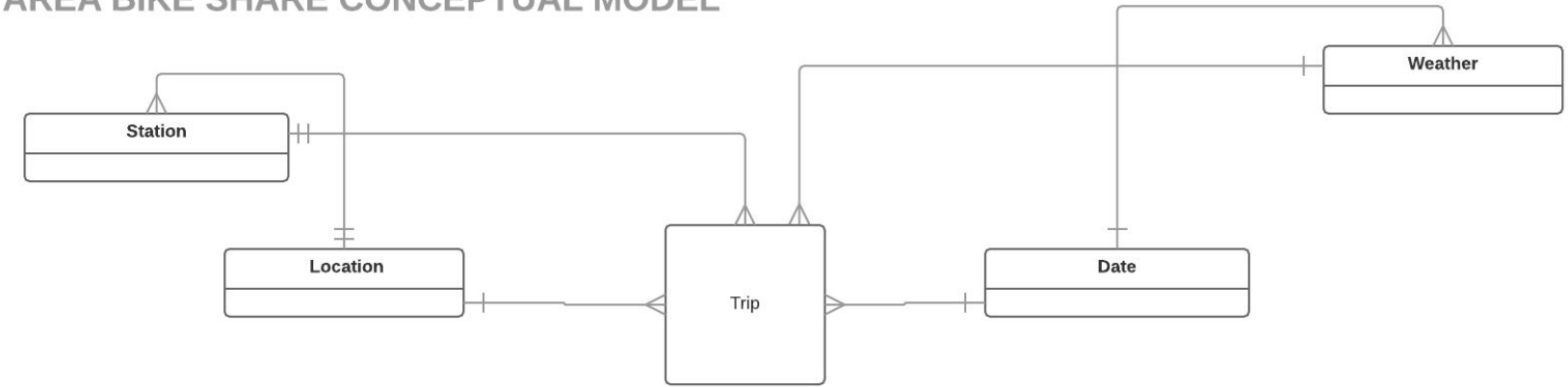
- Bay Area Bike Share, bike sharing system. 700 bikes and 67 stations throughout the San Francisco region
- Bay Area bikes are rented from and returned to any station in the system
- Operated through Motivate, an American firm focused on large-scale bike systems
- Year 3, September 2015 - August 2016

http://www.bayareabikeshare.com/open-data

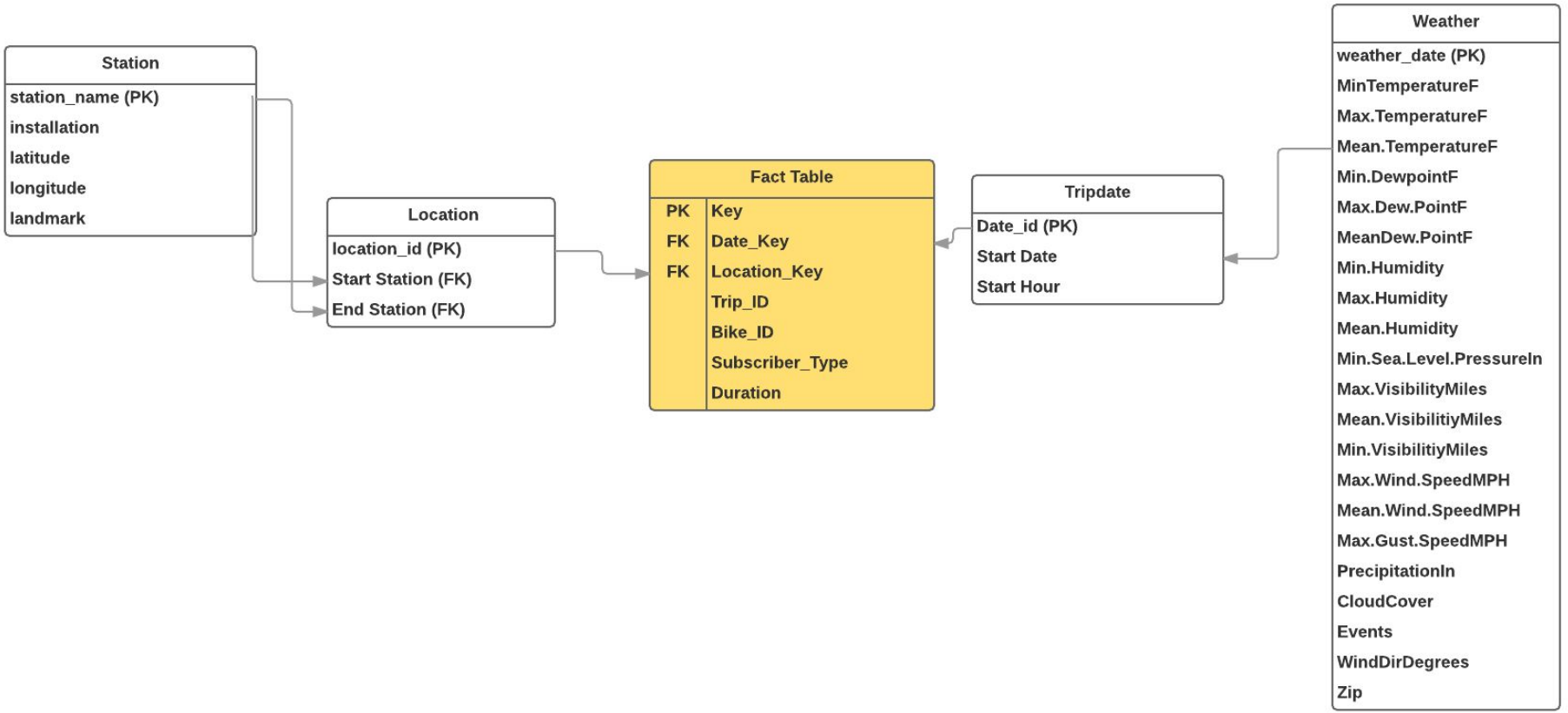
Raw Data

- Raw data consisted of four Excel files
- Station, status, trip, weather
- Each trip is anonymized and includes: bike number, trip start/end day and time, start and end stations
- Plenty to work with, also can be easily understood

BAY AREA BIKE SHARE CONCEPTUAL MODEL



BAY AREA BIKE SHARE DIMENSIONAL MODEL



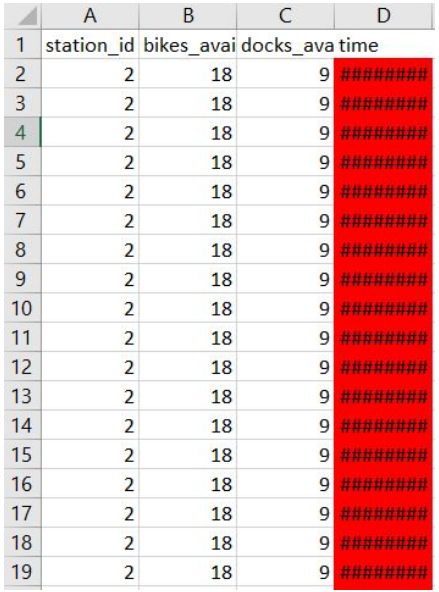
The Grain

- A trip taken by one user (subscriber, customer) using Bike Share
- Transactional Grain

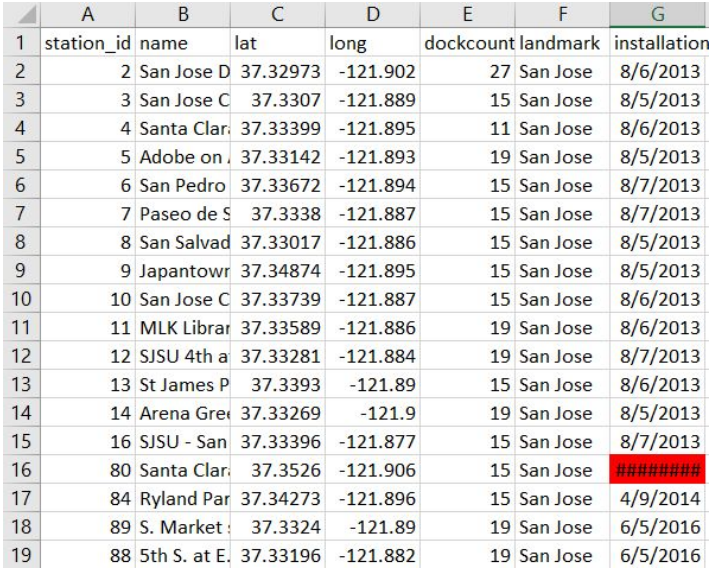
Fact Table	
PK	Key
FK	Date_Key
FK	Location_Key
	Trip_ID
	Bike_ID
	Subscriber_Type
	Duration

Profiling - Excel

- Subscriber type, two unique identifiers of subscriber and customer ID, lots of nulls
- Station ID, unique but lacking completeness w/ lots of null values.



	A	B	C	D
1	station_id	bikes_avai	docks_ava	time
2	2	18	9	########
3	2	18	9	########
4	2	18	9	########
5	2	18	9	########
6	2	18	9	########
7	2	18	9	########
8	2	18	9	########
9	2	18	9	########
10	2	18	9	########
11	2	18	9	########
12	2	18	9	########
13	2	18	9	########
14	2	18	9	########
15	2	18	9	########
16	2	18	9	########
17	2	18	9	########
18	2	18	9	########
19	2	18	9	########



	A	B	C	D	E	F	G
1	station_id	name	lat	long	dockcount	landmark	installation
2	2	San Jose D	37.32973	-121.902	27	San Jose	8/6/2013
3	3	San Jose C	37.3307	-121.889	15	San Jose	8/5/2013
4	4	Santa Clar	37.33399	-121.895	11	San Jose	8/6/2013
5	5	Adobe on	37.33142	-121.893	19	San Jose	8/5/2013
6	6	San Pedro	37.33672	-121.894	15	San Jose	8/7/2013
7	7	Paseo de S	37.3338	-121.887	15	San Jose	8/7/2013
8	8	San Salvad	37.33017	-121.886	15	San Jose	8/5/2013
9	9	Japantowr	37.34874	-121.895	15	San Jose	8/5/2013
10	10	San Jose C	37.33739	-121.887	15	San Jose	8/6/2013
11	11	MLK Librar	37.33589	-121.886	19	San Jose	8/6/2013
12	12	SJSU 4th a	37.33281	-121.884	19	San Jose	8/7/2013
13	13	St James P	37.3393	-121.89	15	San Jose	8/6/2013
14	14	Arena Gree	37.33269	-121.9	19	San Jose	8/5/2013
15	16	SJSU - San	37.33396	-121.877	15	San Jose	8/7/2013
16	80	Santa Clar	37.3526	-121.906	15	San Jose	########
17	84	Ryland Par	37.34273	-121.896	15	San Jose	4/9/2014
18	89	S. Market	37.3324	-121.89	19	San Jose	6/5/2016
19	88	5th S. at E.	37.33196	-121.882	19	San Jose	6/5/2016

Profiling - Excel

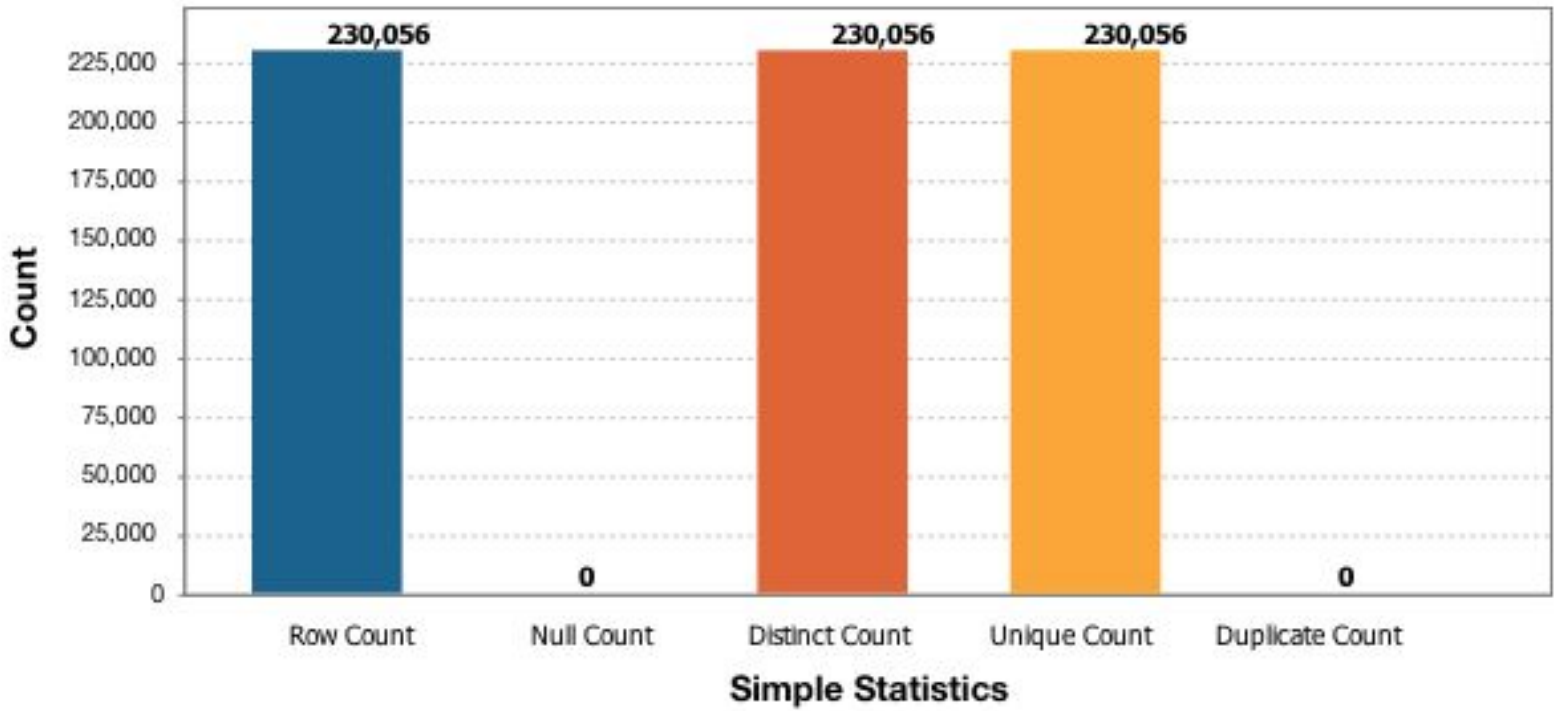
- Trip ID, all have uniqueness values, nulls galore

	A	B	C	D	E	F	G	H	I	J	K
1	Trip ID	Duration	Start Date	Start Station	Start Term	End Date	End Station	End Termi	Bike #	Subscriber	Zip Code
2	913465	746	########	San Francisco Caltrain 2 (330 Townsend)	69	########	San Francisco City Hall	58	238	Subscriber	94107
3	913466	969	########	Clay at Battery	41	########	Washington at Kearny	46	16	Subscriber	94133
4	913467	233	########	Davis at Jackson	42	########	Commercial at Montgomery	45	534	Subscriber	94111
5	913468	213	########	Clay at Battery	41	########	Steuart at Market	74	312	Subscriber	94107
6	913469	574	########	Steuart at Market	74	########	San Francisco Caltrain 2 (330 Townsend)	69	279	Subscriber	94107
7	913470	623	########	San Jose Diridon Caltrain Station	2	########	Japantown	9	261	Subscriber	95112
8	913471	746	########	Embarcadero at Bryant	54	########	Powell Street BART	39	436	Subscriber	94103
9	913472	1038	########	Townsend at 7th	65	########	Howard at 2nd	63	607	Subscriber	94107
10	913473	424	########	Market at 10th	67	########	Townsend at 7th	65	259	Subscriber	94102
11	913474	633	########	Embarcadero at Bryant	54	########	Embarcadero at Sansome	60	613	Subscriber	94105
12	913475	174	########	2nd at Folsom	62	########	Market at Sansome	77	449	Subscriber	94107
13	913476	777	########	Davis at Jackson	42	########	San Francisco Caltrain (Townsend at 4th)	70	86	Subscriber	94111
14	913477	454	########	Powell Street BART	39	########	San Francisco Caltrain 2 (330 Townsend)	69	440	Subscriber	94107
15	913478	228	########	Washington at Kearny	46	########	Market at Sansome	77	351	Subscriber	94133
16	913479	445	########	Civic Center BART (7th at Market)	72	########	San Francisco Caltrain (Townsend at 4th)	70	315	Subscriber	94103
17	913480	649	########	5th at Howard	57	########	Steuart at Market	74	559	Subscriber	94112
18	913481	459	########	San Francisco Caltrain (Townsend at 4th)	70	########	2nd at Folsom	62	579	Subscriber	94002
19	913482	337	########	Townsend at 7th	65	########	2nd at Townsend	61	287	Subscriber	94107

Profiling - Talend

Trip_ID

All contain unique values



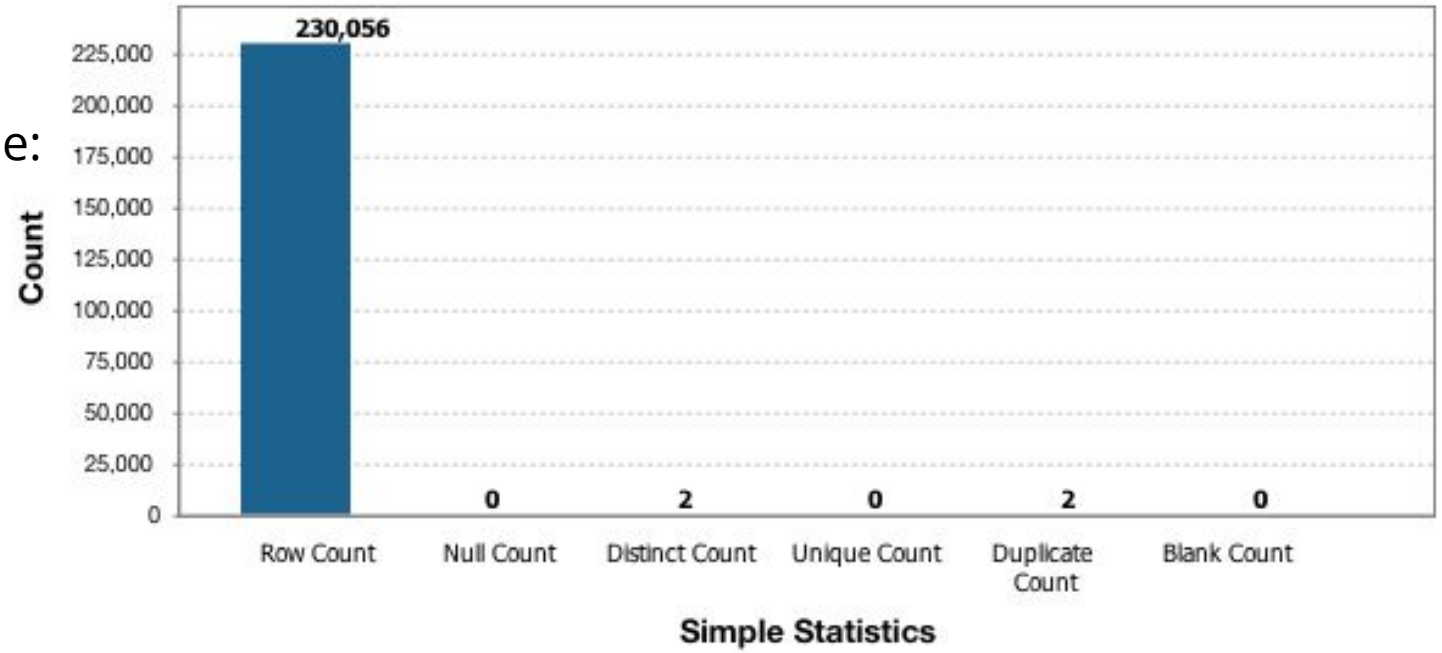
Profiling - Talend

Subscriber Type

2 Distinct, unique:

-Subscriber

-Customer



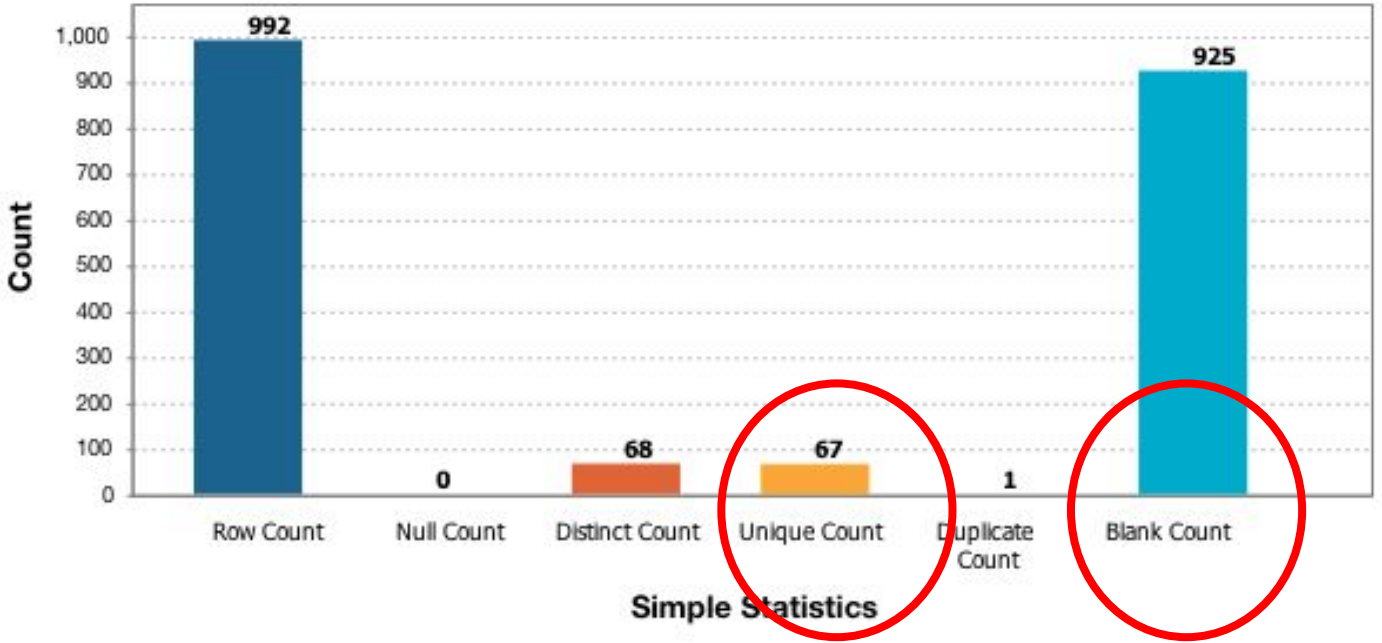
Profiling - Talend

Station ID

-Blank values confusing data

-67 unique

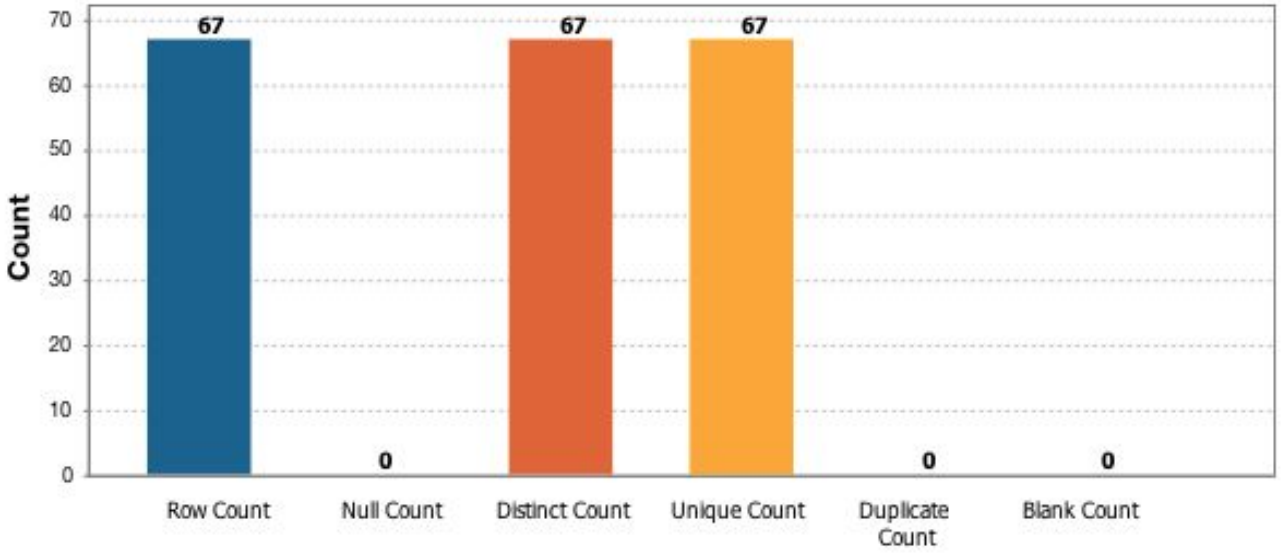
stations



Profiling - Talend

Station ID

-Removed blank values in FileDeliminated connection



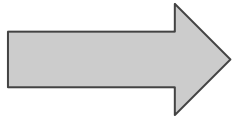
Data Quality Scorecard

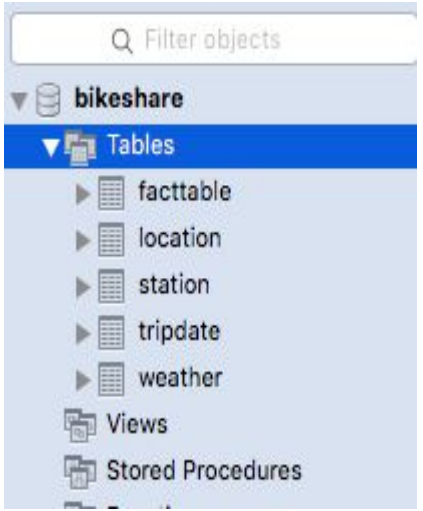
Extract, Transform, Load

1. Create Database & Tables in MySQL

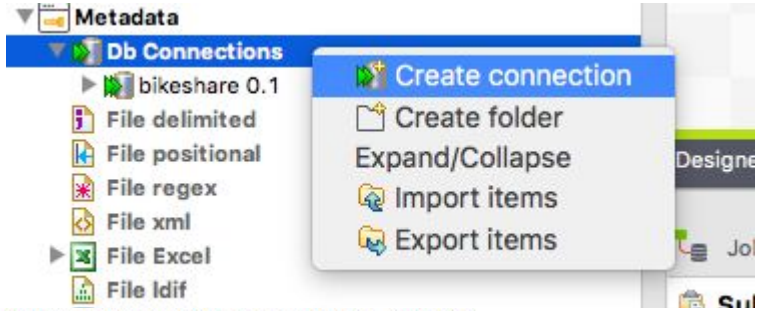
Query 1 | SQL File 5* | Limit to 200

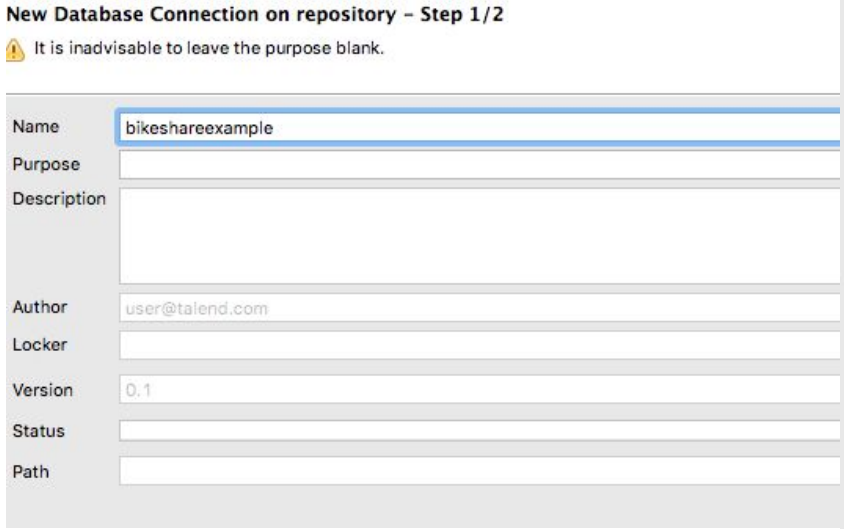
```
CREATE SCHEMA bikeshare;

use bikeshare;

CREATE TABLE weather(
weather_date VARCHAR(45),
Max_TemperatureF INT,
Mean_TemperatureF INT,
Min_TemperatureF INT,
Max_Dew_PointF INT,
MeanDew_PointF INT,
Min_DewpointF INT,
Max_Humidity INT,
Mean_Humidity INT,
Min_Humidity INT,
Max_Sea_Level_PressureIn FLOAT,
Mean_Sea_Level_PressureIn FLOAT,
Min_Sea_Level_PressureIn FLOAT,
Max_VisibilityMiles INT,
Mean_VisibilityMiles INT,
Min_VisibilityMiles INT,
Max_Wind_SpeedMPH INT,
Mean_Wind_SpeedMPH INT,
Max_Gust_SpeedMPH INT,
CloudCover INT,
WindDirDegrees INT,
PRIMARY KEY(weather_date)
);


CREATE TABLE station(
station_name VARCHAR(45),
latitude FLOAT
```

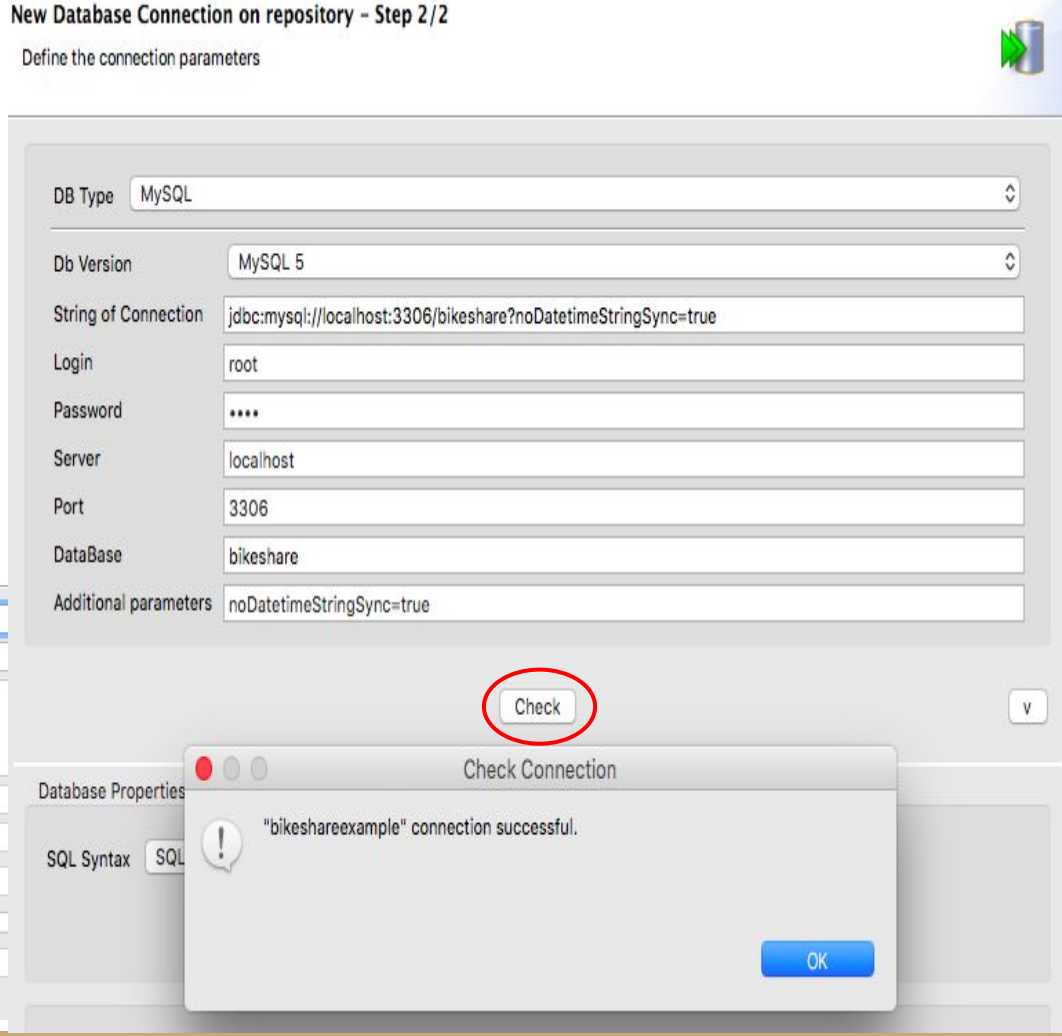




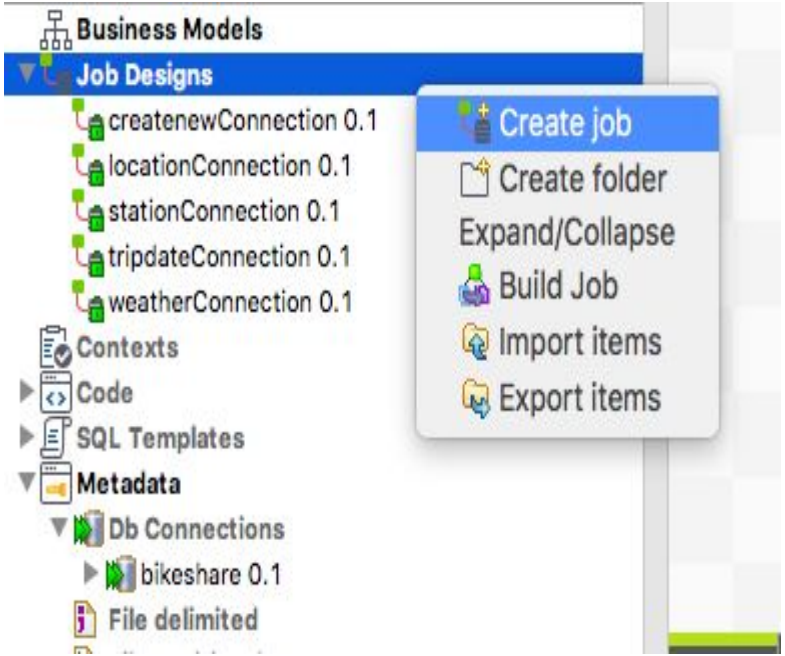
2. Create Connection in talend (MySQL)

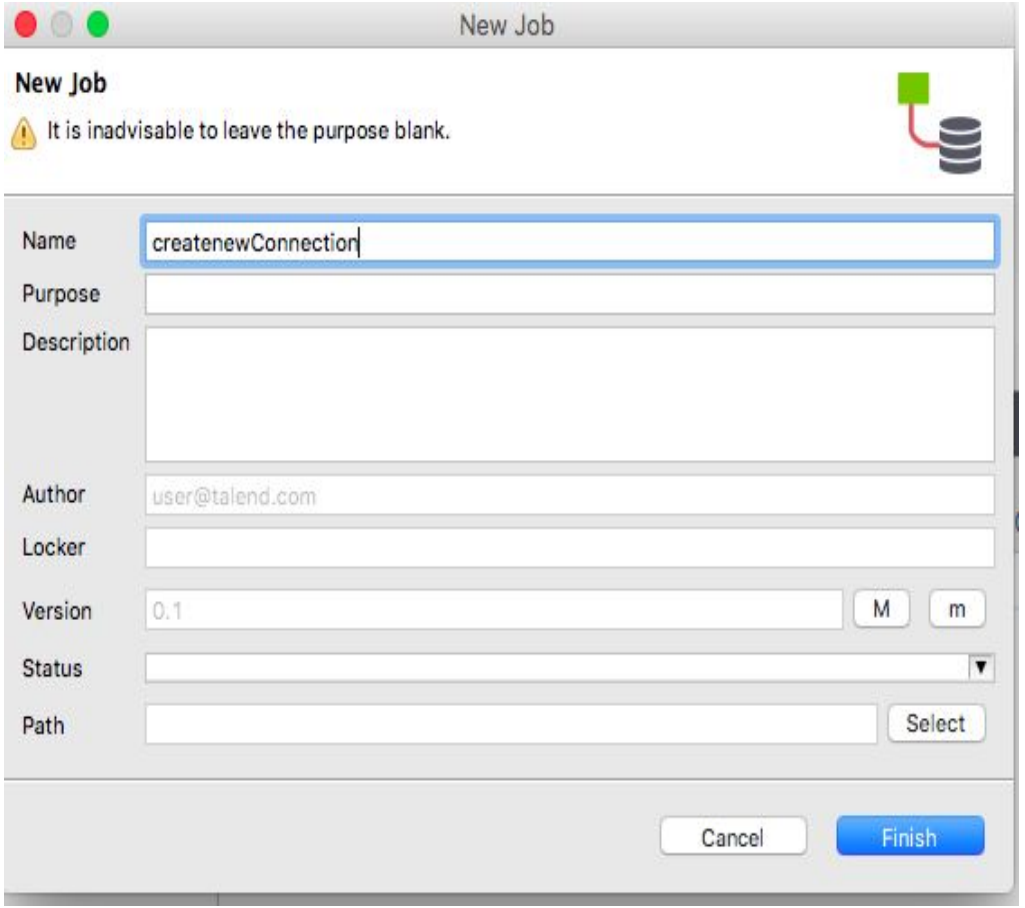






3. Create workspace in Talend





4. Create Excel file

File – Step 1 of 4

⚠ It is inadvisable to leave the purpose blank.

Name	tripdateexample
Purpose	
Description	
Author	user@talend.com
Locker	
Version	0.1
Status	
Path	

File – Step 2 of 4

Add a Metadata File on repository
Define the path of the file and the format settings

File Settings

Server: Localhost 127.0.0.1

File: /Users/ann/Documents/GW life/data warehousing/final project/trip_final.xlsx [Browse...]

☑ Read excel2007 file format(xlsx)

Generation mode: Memory-consuming(User mode)

File Viewer and Sheets setting

Set sheets parameters

☑ All sheets/DSelect sheet
☑ trip_final

Please select sheet (Sheet structure as schema guide) trip_fina

A	B	C	D	E	F	G	H	I
	Trip.ID	Duration	Start.St...	Start.Ter...	End.Stat...	End.Ter...	Bike..	Subscri...
5.0	913469.0	574.0	Steuart...	74.0	San Fran...	69.0	279.0	Subscriber
7.0	913471.0	746.0	Embarca...	54.0	Powell S...	39.0	436.0	Subscriber
9.0	913473.0	424.0	Market a...	67.0	Townse...	65.0	259.0	Subscriber
10.0	913474.0	633.0	Embarca...	54.0	Embarca...	60.0	613.0	Subscriber
16.0	913480.0	649.0	5th at H...	57.0	Steuart...	74.0	559.0	Subscriber
19.0	913483.0	518.0	Townse...	65.0	South V...	66.0	457.0	Subscriber
20.0	913484.0	352.0	Tempora...	55.0	2nd at T...	61.0	589.0	Subscriber
22.0	913486.0	314.0	Rengsto...	33.0	San Ant...	29.0	118.0	Subscriber
24.0	913488.0	274.0	Washing...	46.0	Market a...	76.0	329.0	Subscriber
27.0	913491.0	509.0	Harry Br...	50.0	2nd at T...	61.0	413.0	Subscriber
39.0	913507.0	613.0	San Fran...	70.0	Tempora...	55.0	548.0	Subscriber
40.0	913510.0	270.0	2nd at T...	61.0	San Fran...	70.0	502.0	Subscriber
46.0	913518.0	390.0	Broadwa...	82.0	Embarca...	51.0	496.0	Subscriber

File - Step 3 of 4

Add a Metadata File on repository
Define the setting of the parse job

File Settings

Encoding: UTF-8

Advanced separator(for number)

Thousands separator:

Decimal separator:

Rows To Skip

If any rows must be ignor

Header 2

Footer

Metadata column setting

First column: 1

Last column:

Limit Of Rows

If the number of lines mus

Limit

Preview | Output

Set heading row as column names | Refresh Preview

Column 0	Trip.ID	Duration	Start.Station	Start.Terminal	End.Statio
5	913469	574	Steuart at Market	74	San Franci
7	913471	746	Embarcadero at Bryant	54	Powell Str
9	913473	424	Market at 10th	67	Townsend
10	913474	633	Embarcadero at Bryant	54	Embarcad
16	913480	649	5th at Howard	57	Steuart at
19	913483	518	Townsend at 7th	65	South Van
20	913484	352	Temporary Transbay Terminal (Howard at Beale)	55	2nd at Tov
22	913486	314	Rengstorff Avenue / California Street	33	San Anton
24	913488	274	Washington at Kearny	46	Market at

File - Step 4 of 4

Add a Schema on repository
Define the Schema

Name: metadata

Comment:

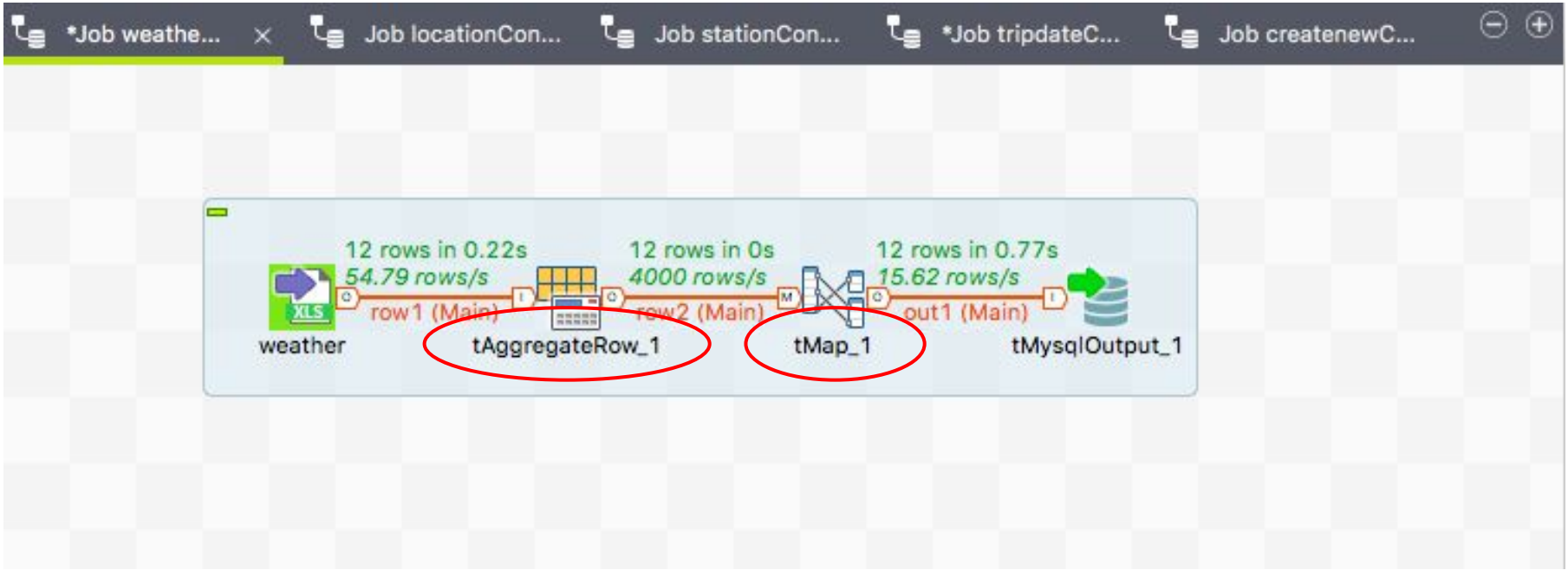
Schema

Click to update schema preview | Guess

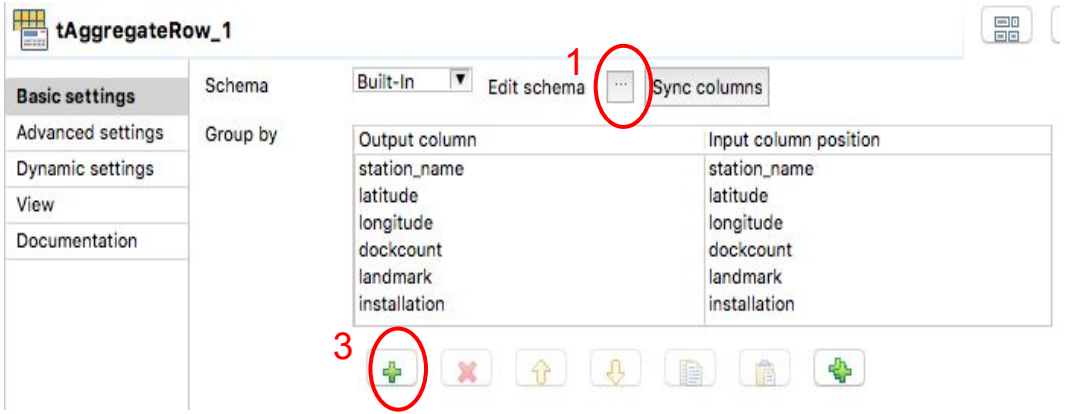
Description of the Schema

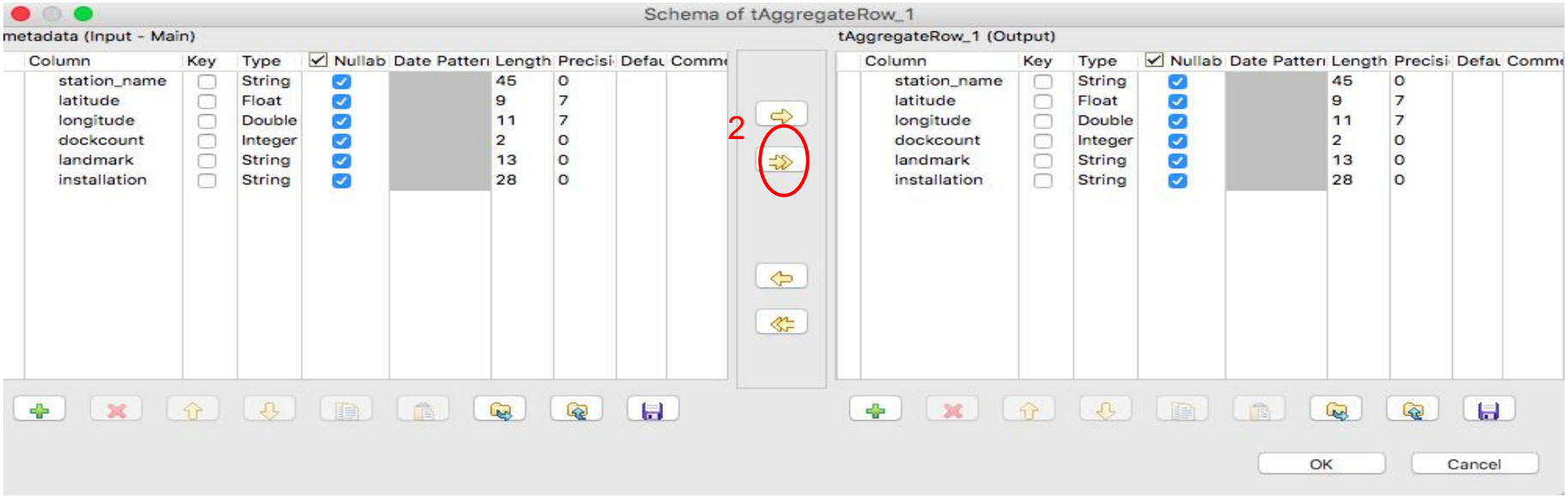
Column	Key	Type	Nullab	Date Pattern (Ctrl+Space	Length	Precision	Default	Comment
Column0	☑	Integer	☑		3	0		
_13469	☐	Integer	☑		6	0		
_74	☐	Integer	☑		4	0		
Steuart_at_Market	☐	String	☑		45	0		
_4	☐	Integer	☑		2	0		
San_Francisco_Caltrain_2__330_T...	☐	String	☑		45	0		
_9	☐	Integer	☑		2	0		
_79	☐	Integer	☑		3	0		
Subscriber	☐	String	☑		10	0		
_4107	☐	Integer	☑		5	0		
Wed Sep 09 00 00 00 EDT 2015	☐	String	☑		28	0		

5. Drag to workspace

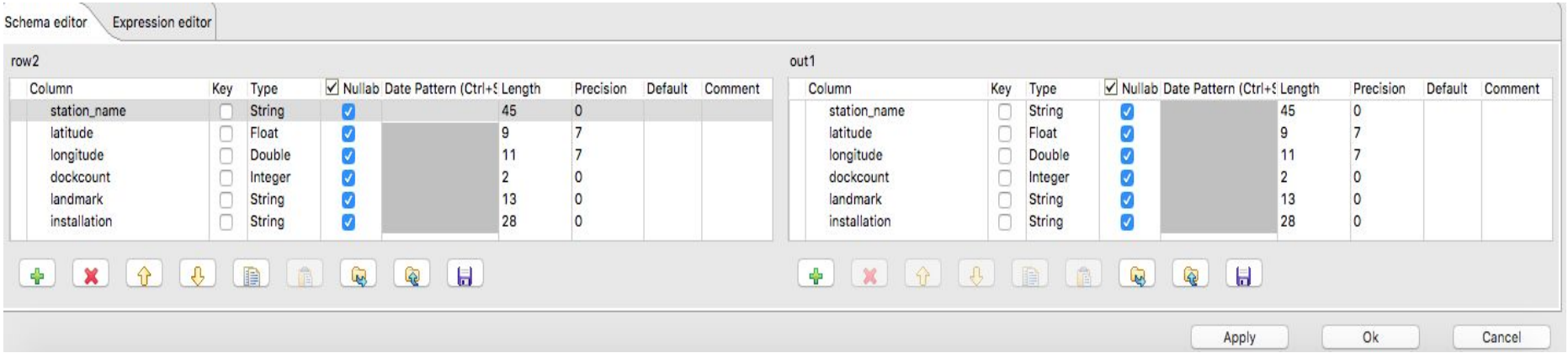


ETL process-tAggregate

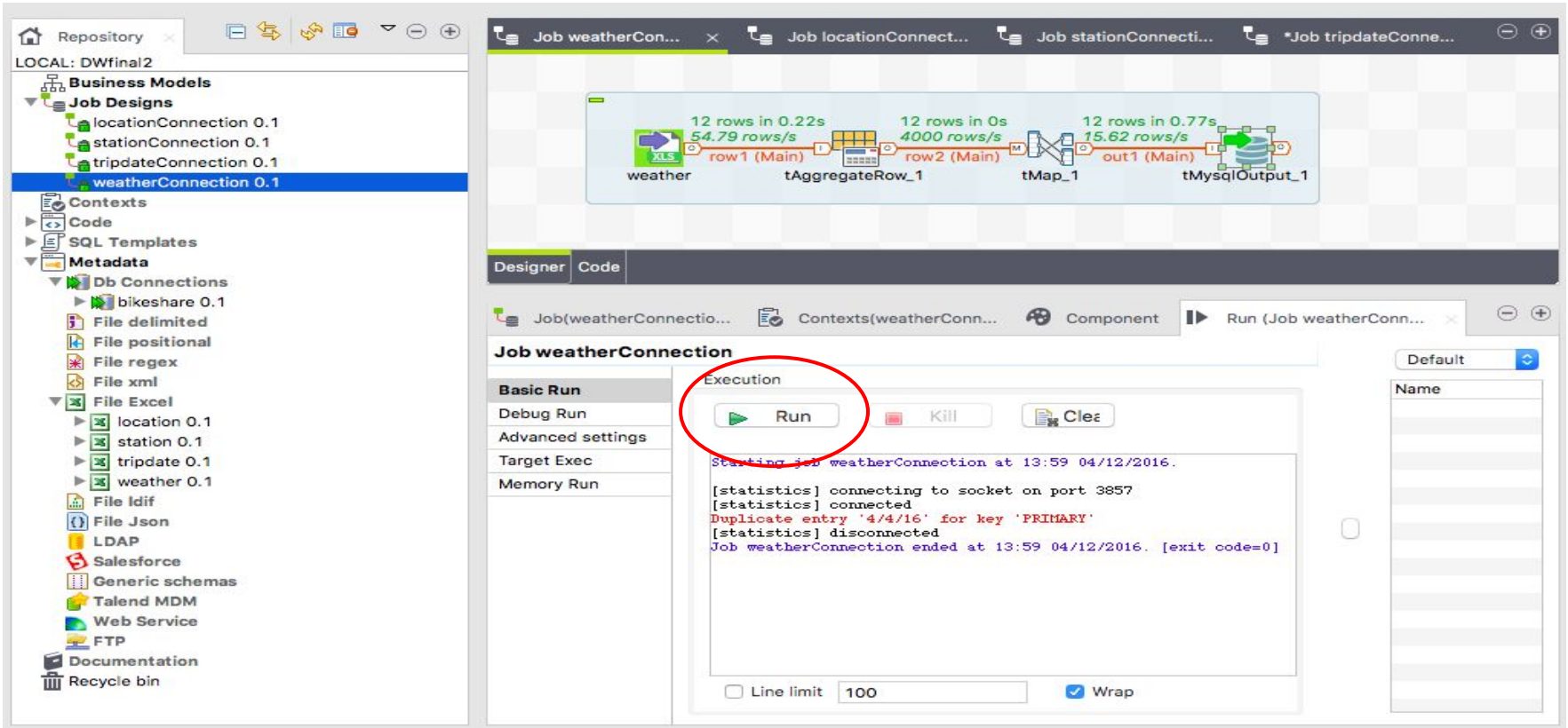




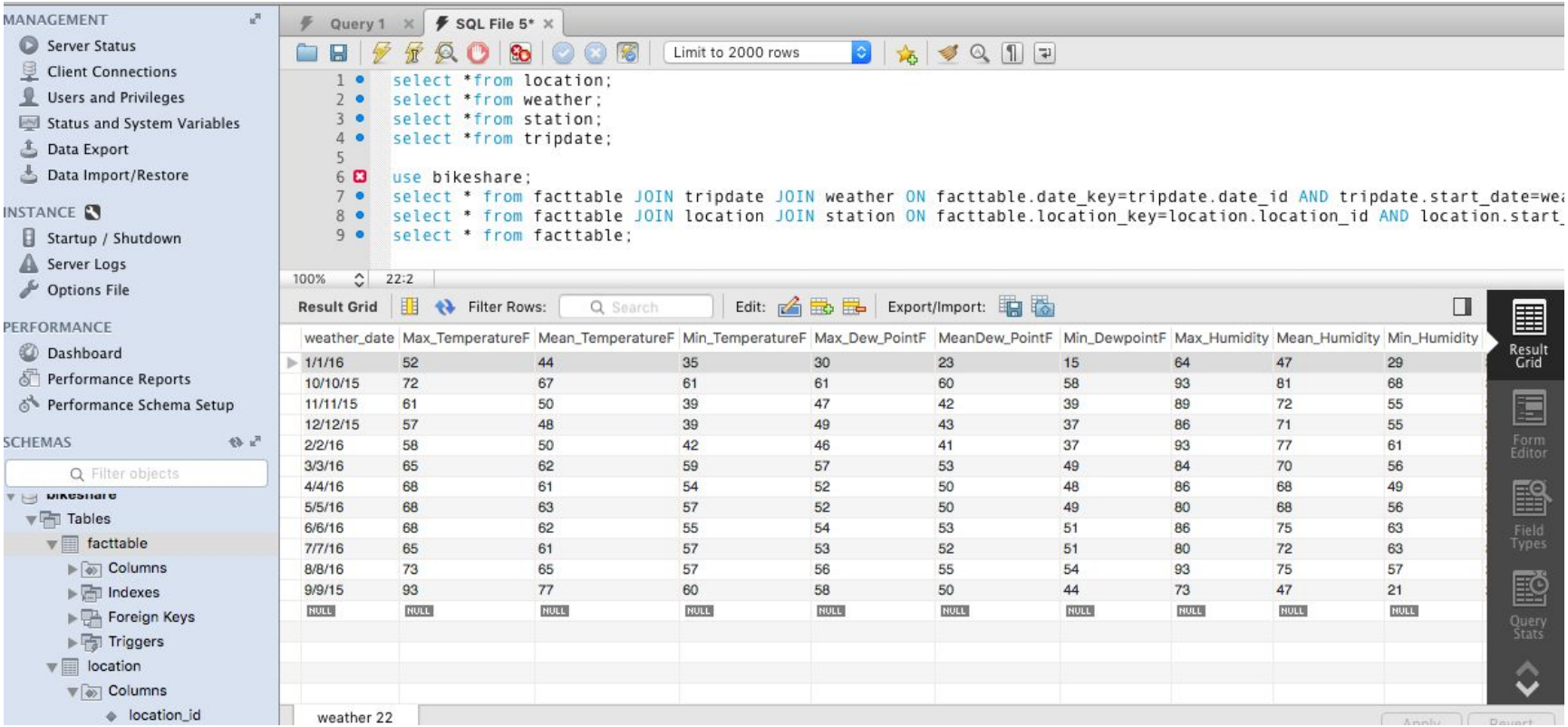
ETL process-tMap



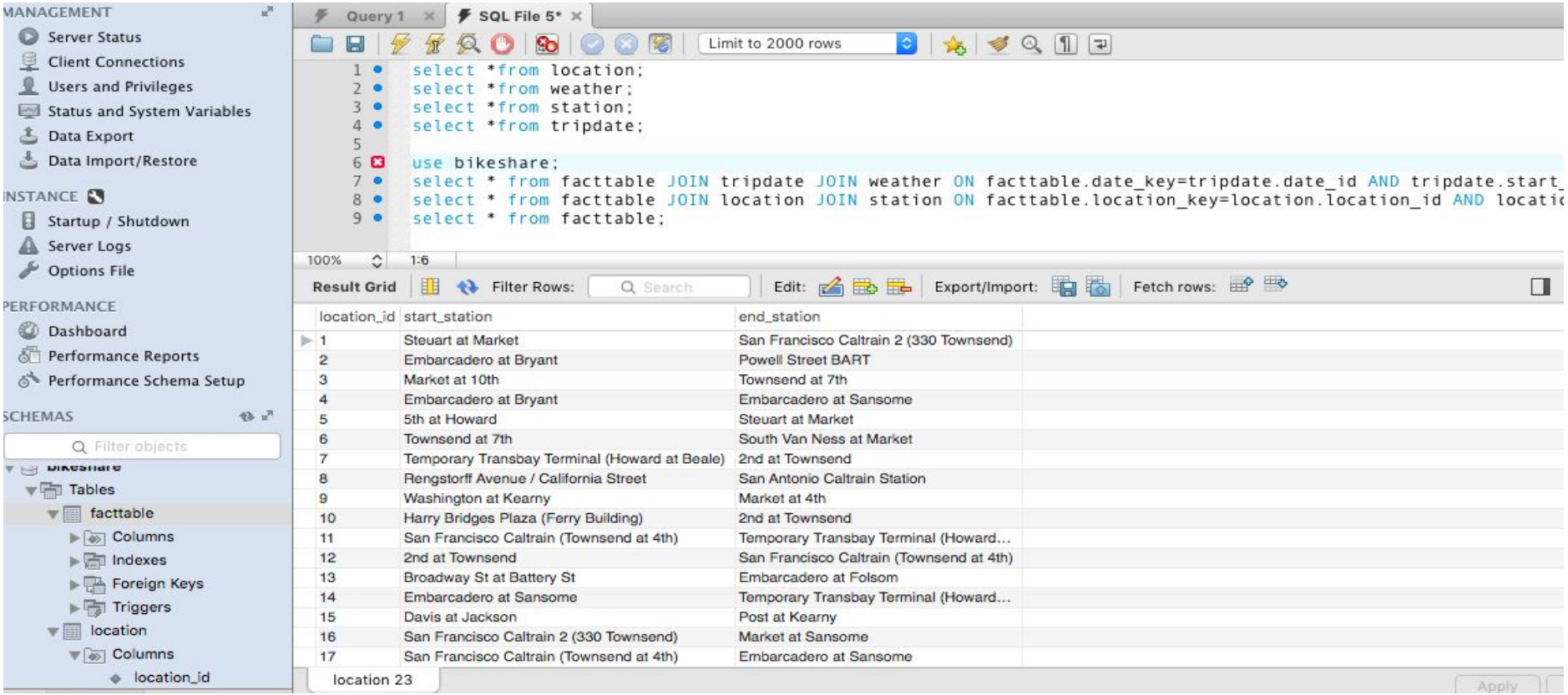
Talend weather



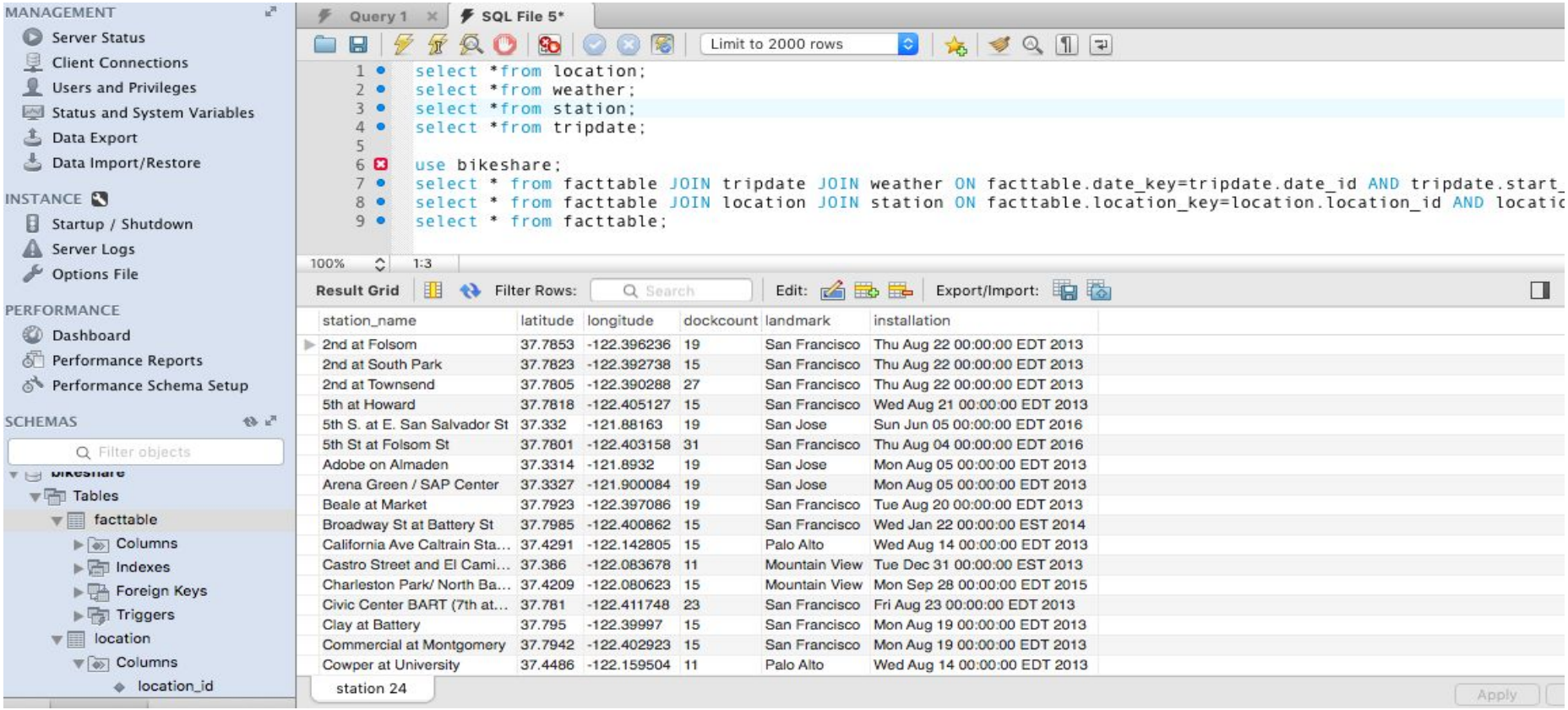
6. Result in MySQL (weather)



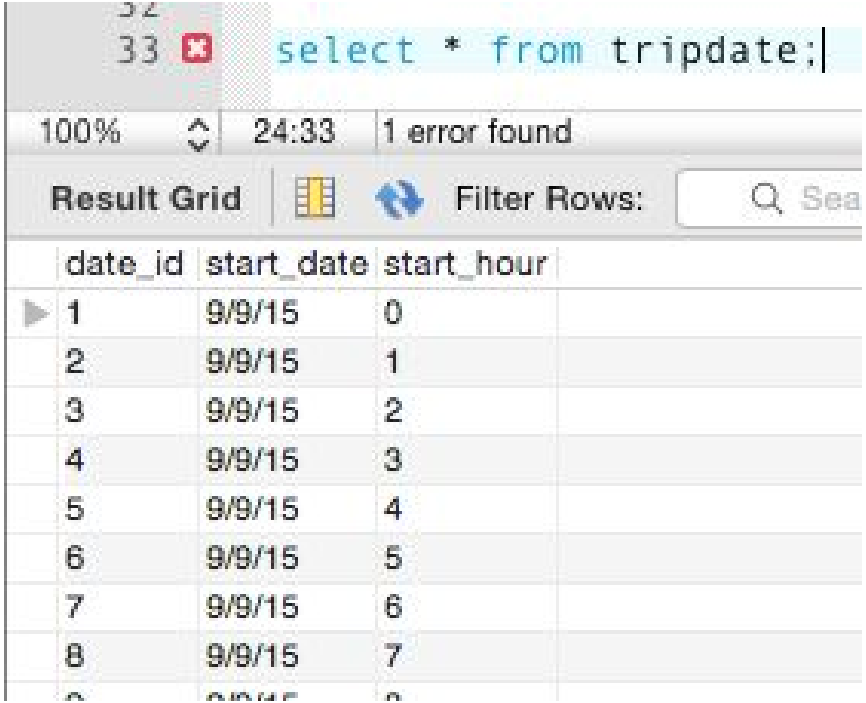
Result in MySQL (location)



Result in MySQL (station)



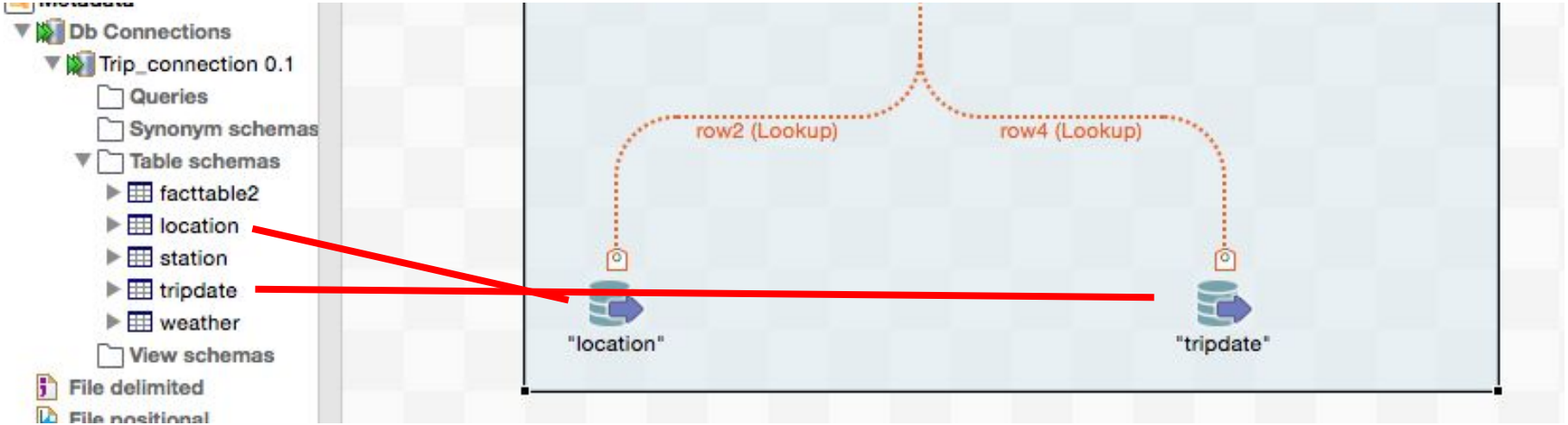
Result in MySQL (tripdate)



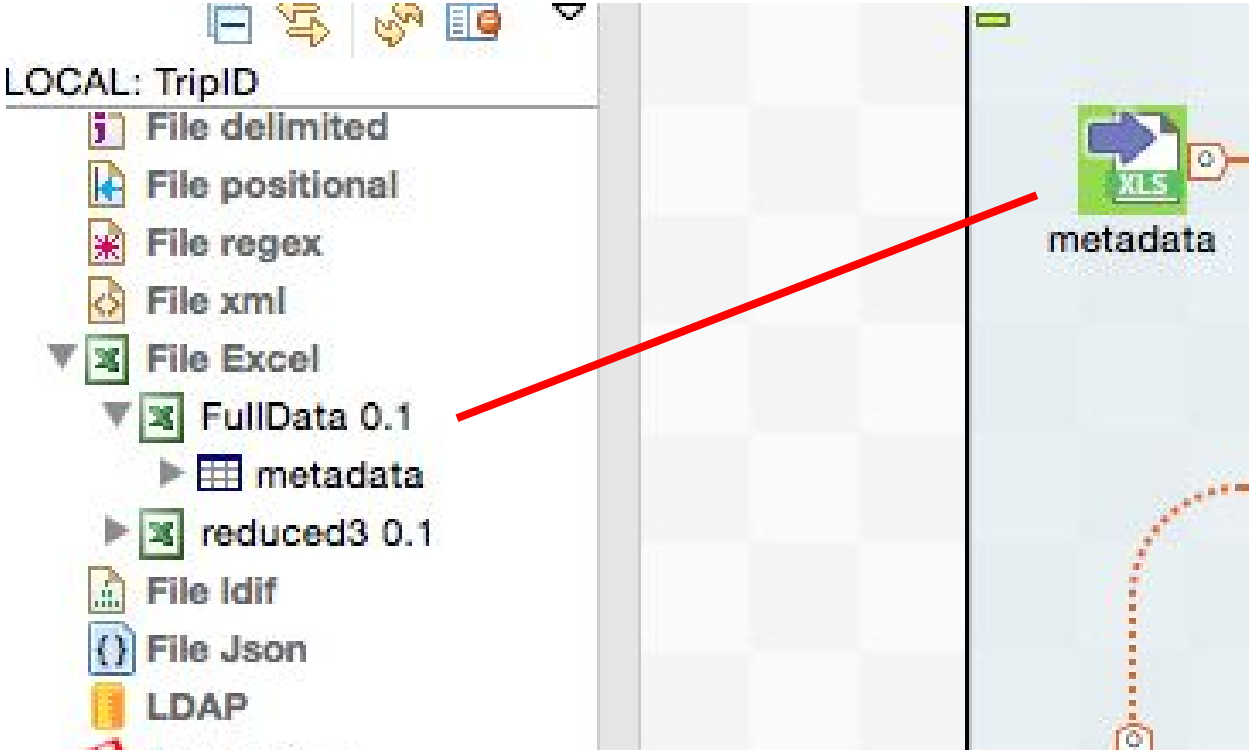
date_id	start_date	start_hour
1	9/9/15	0
2	9/9/15	1
3	9/9/15	2
4	9/9/15	3
5	9/9/15	4
6	9/9/15	5
7	9/9/15	6
8	9/9/15	7

ETL for the Fact Table

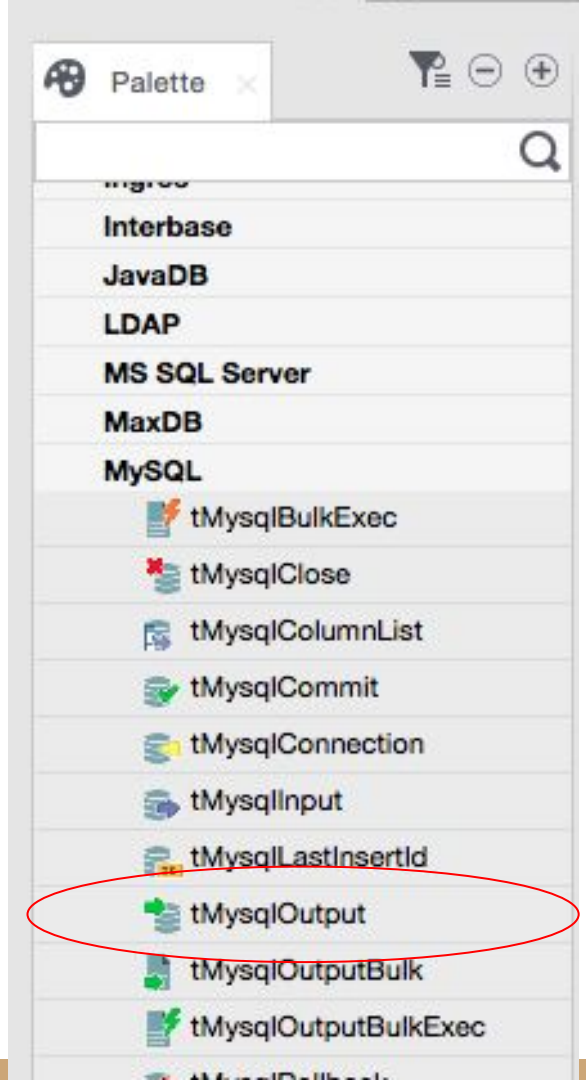
Establish connection with MySQL db



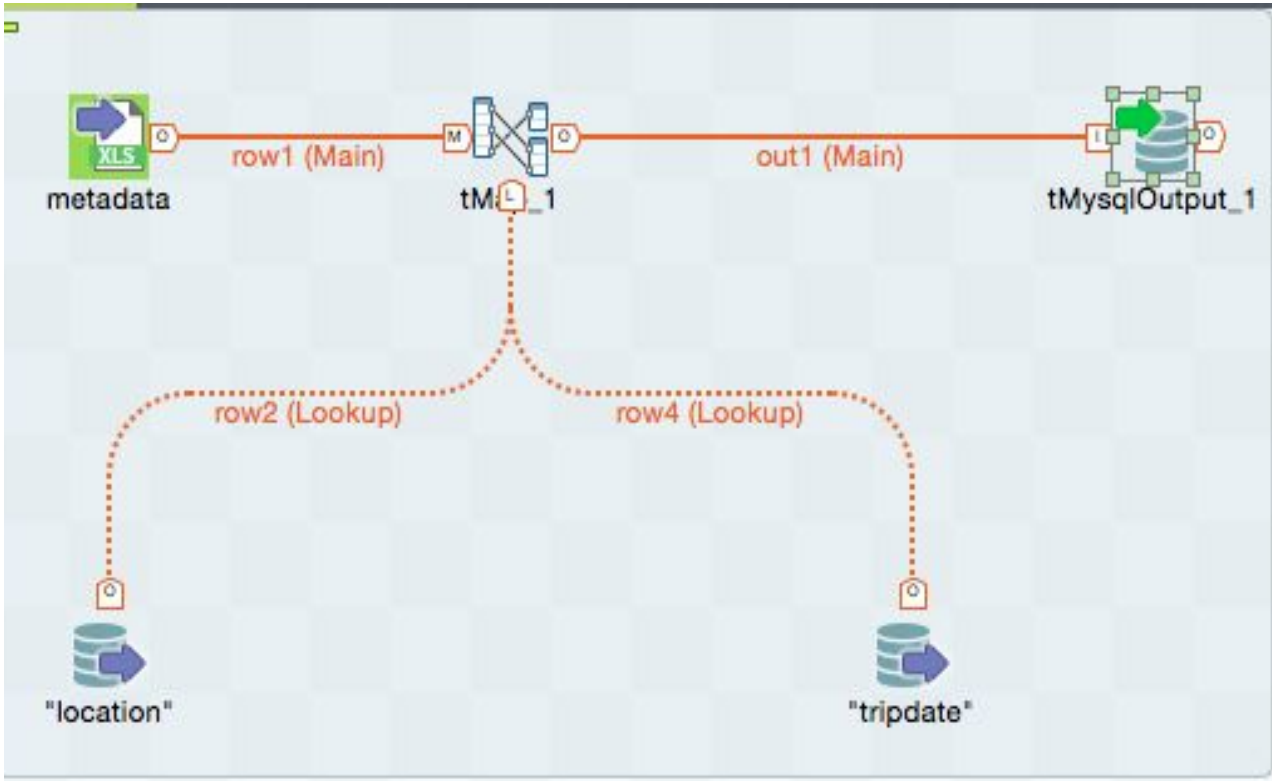
Establish metadata connection



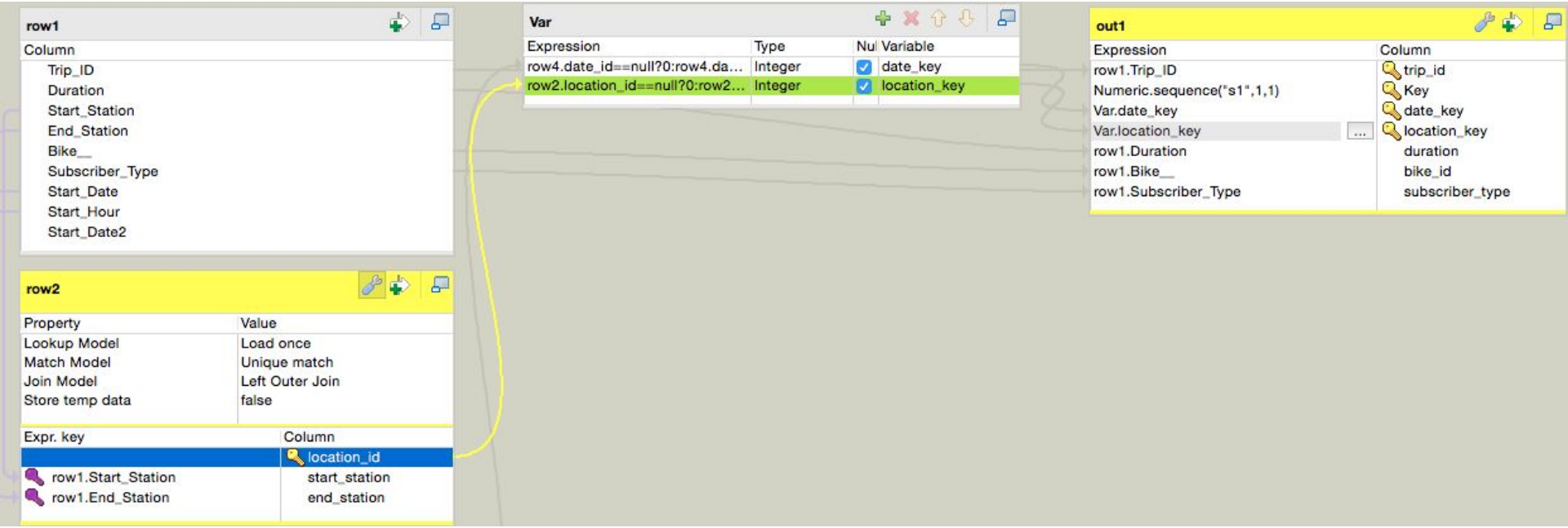
MySQL output



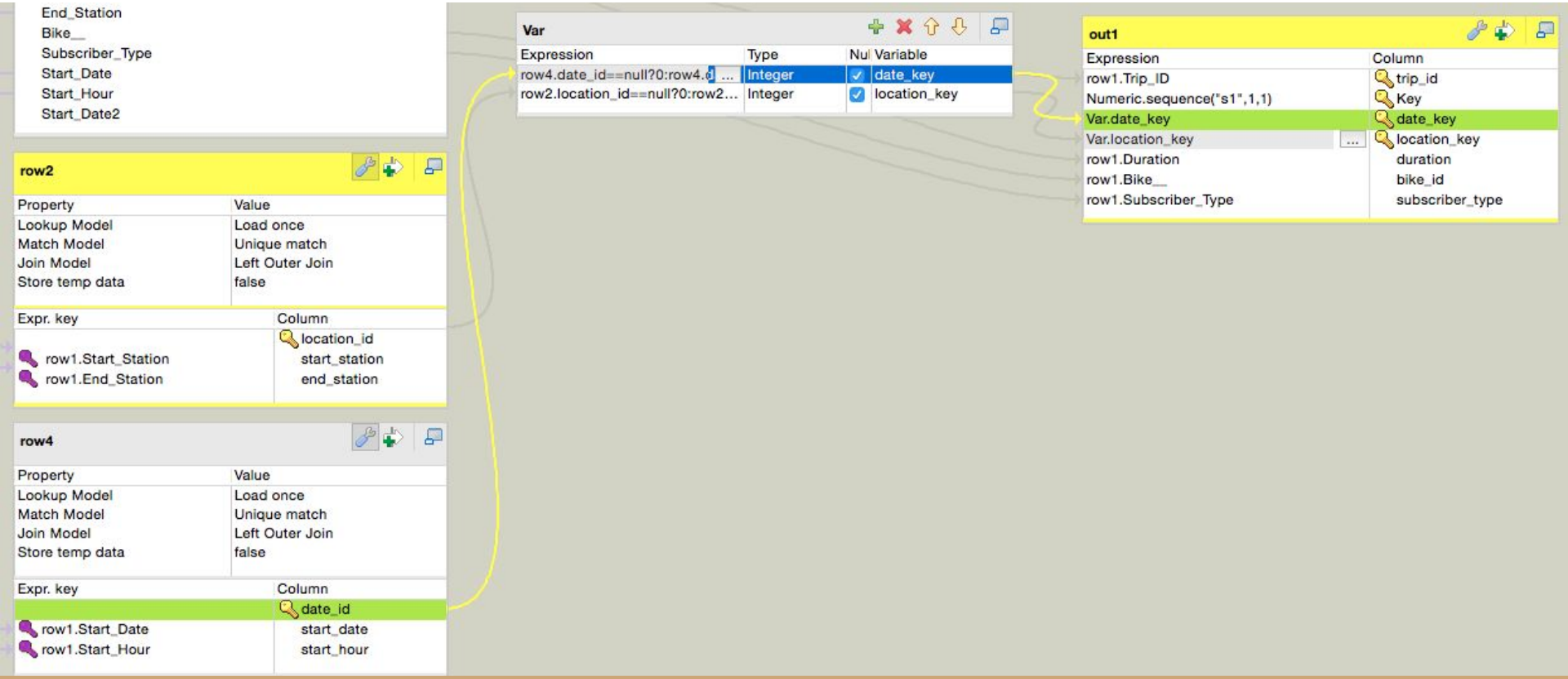
Add tMap Processing Function



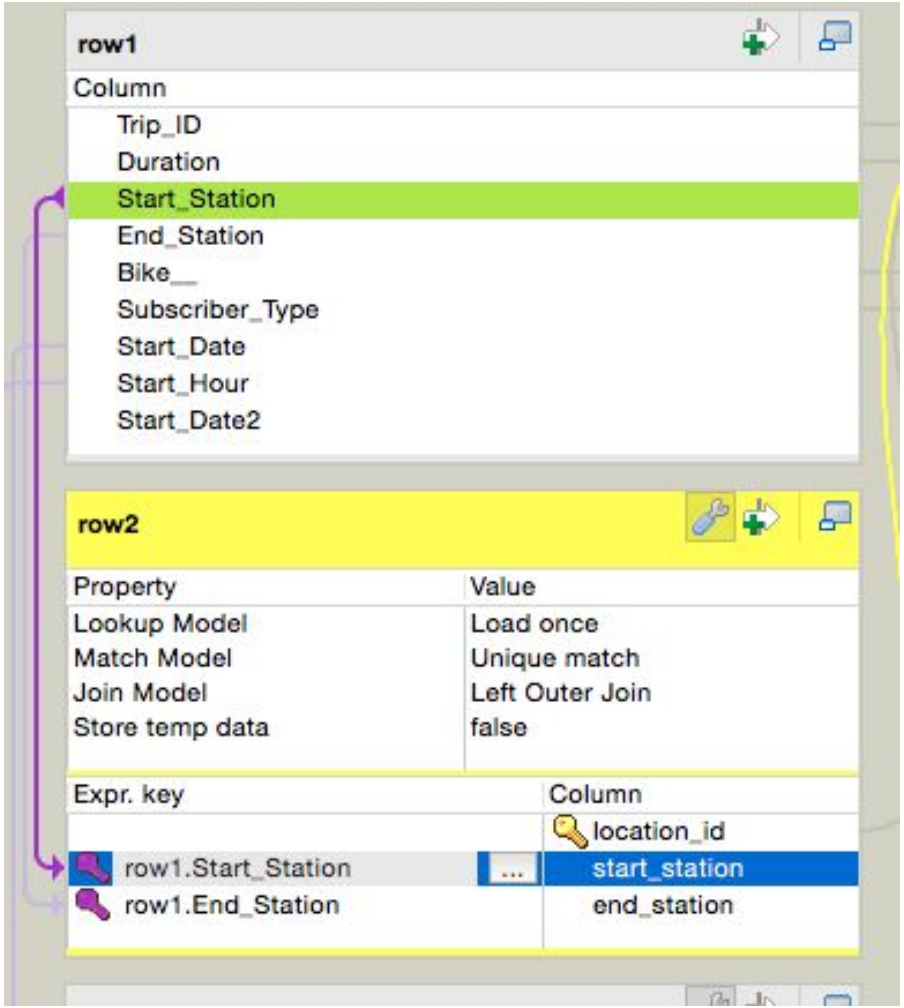
Location key



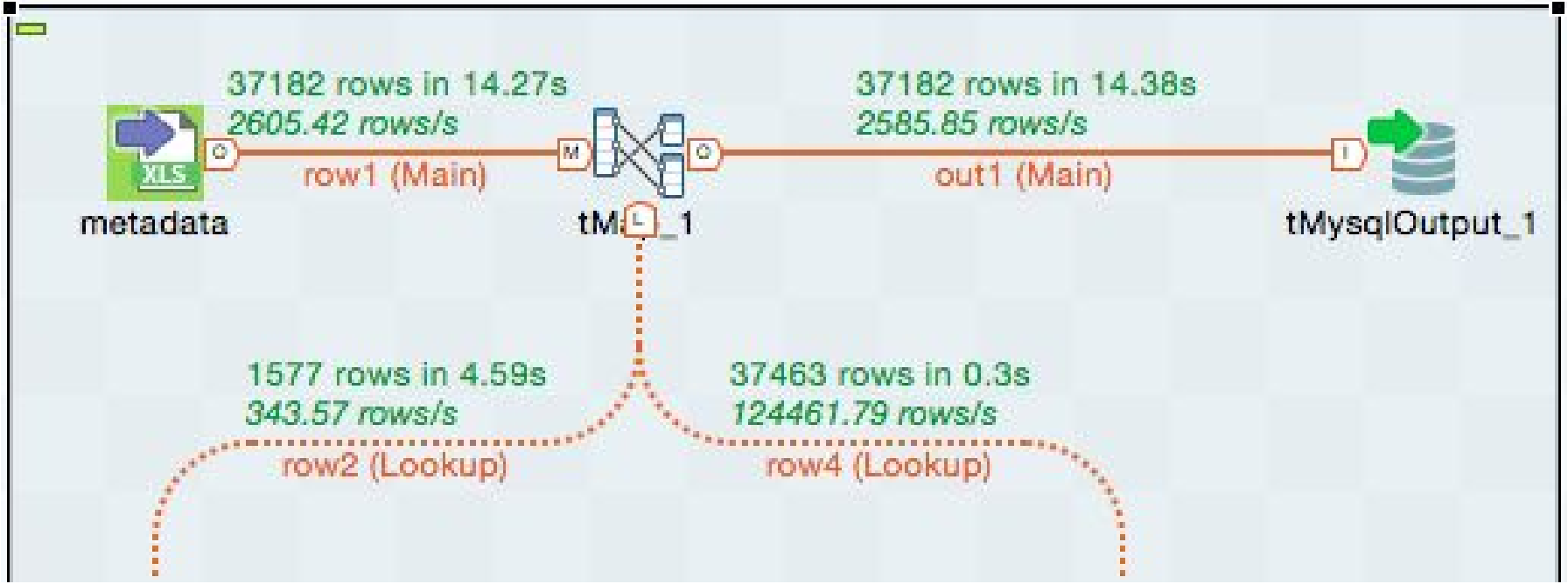
Date key


End_Station
Bike__
Subscriber_Type
Start_Date
Start_Hour
Start_Date2
row2
Property
Value
Lookup Model
Load once
Match Model
Unique match
Join Model
Left Outer Join
Store temp data
false
Expr. key
Column
location_id
row1.Start_Station
start_station
row1.End_Station
end_station
row4
Property
Value
Lookup Model
Load once
Match Model
Unique match
Join Model
Left Outer Join
Store temp data
false
Expr. key
Column
date_id
row1.Start_Date
start_date
row1.Start_Hour
start_hour
Var
Expression
Type
Nul
Variable
row4.date_id==null?0:row4.d
Integer
date_key
row2.location_id==null?0:row2...
Integer
location_key
out1
Expression
Column
row1.Trip_ID
trip_id
Numeric.sequence("s1",1,1)
Key
Var.date_key
date_key
Var.location_key
location_key
row1.Duration
duration
row1.Bike__
bike_id
row1.Subscriber_Type
subscriber_type

Joining metadata and MySql database



After running the job



MySQL output

```
31  select * from facttable2;
```

100% 1:31 1 error found

Result Grid | Filter Rows: Search | Edit:

trip_id	date_key	location_key	duration	bike_id	subscriber_type	Key
913469	2	1	574	279	Subscriber	1
913471	3	2	746	436	Subscriber	2
913473	5	3	424	259	Subscriber	3
913474	5	4	633	613	Subscriber	4
913480	6	5	649	559	Subscriber	5
913483	7	6	518	457	Subscriber	6
913484	7	7	352	589	Subscriber	7
913486	7	8	314	118	Subscriber	8
913488	7	9	274	329	Subscriber	9

Talend benefits and problems

Benefits

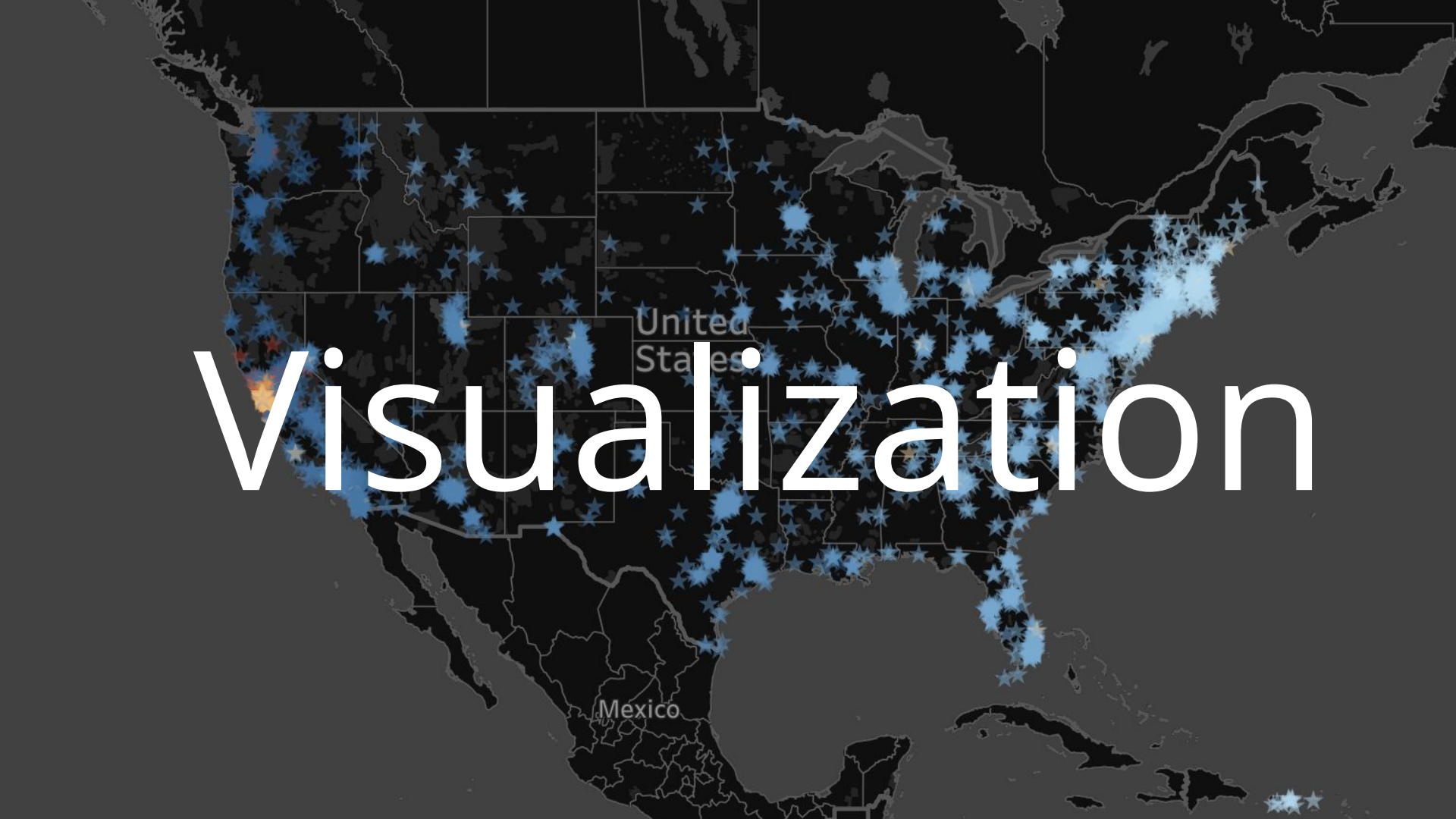
- When we change the settings of MySQL tables, we can easily reload all the data into Talend.
- Talend handles large source files.
- Eventually, after debugging, we got the results we wanted.

Problems

- Not intuitive when attempting to configure
- Many errors during when running the job, and not easy to fix
  - Errors during ELT process are understandable only for someone experienced in Java
  - Troubleshooting process alone took 3 days

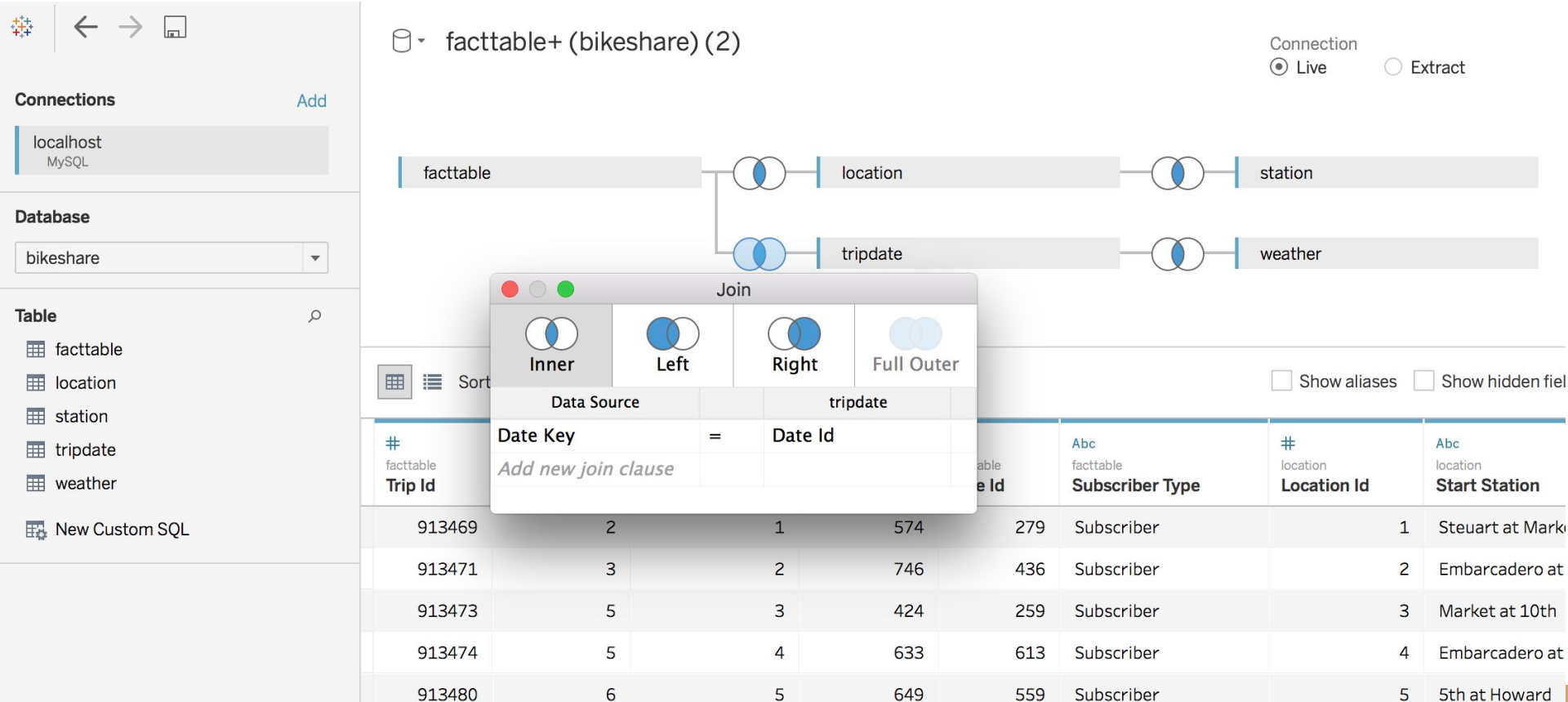
Benefits and problems cont.

- Explains why relationship between Business and IT is so important (we were acting as both in this project)
- Those skilled in ETL can save hours for those in analytics, BI, etc.
- Worth investing in a configurable ETL tool
- Familiarity with SQL made the process move quicker

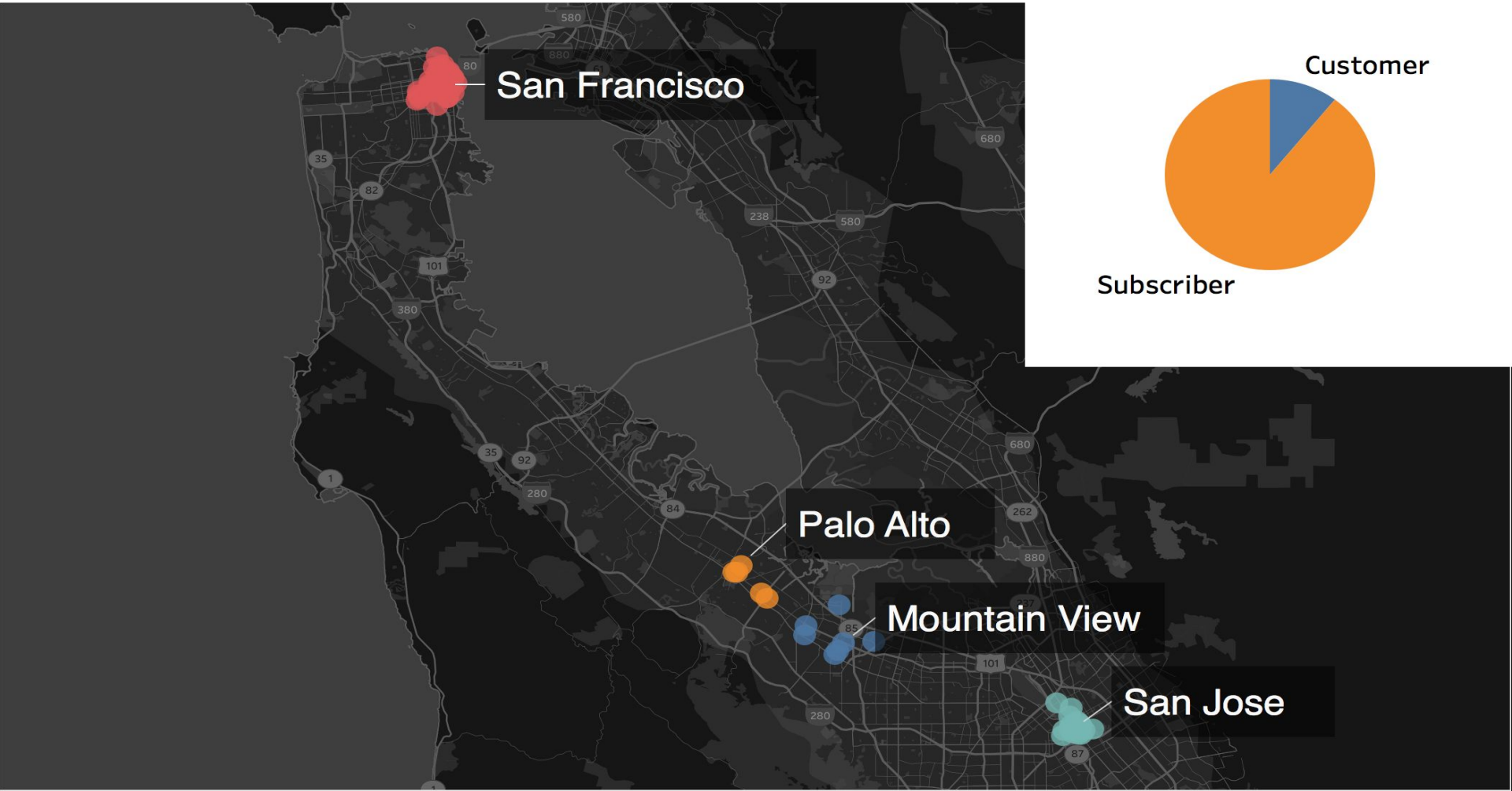


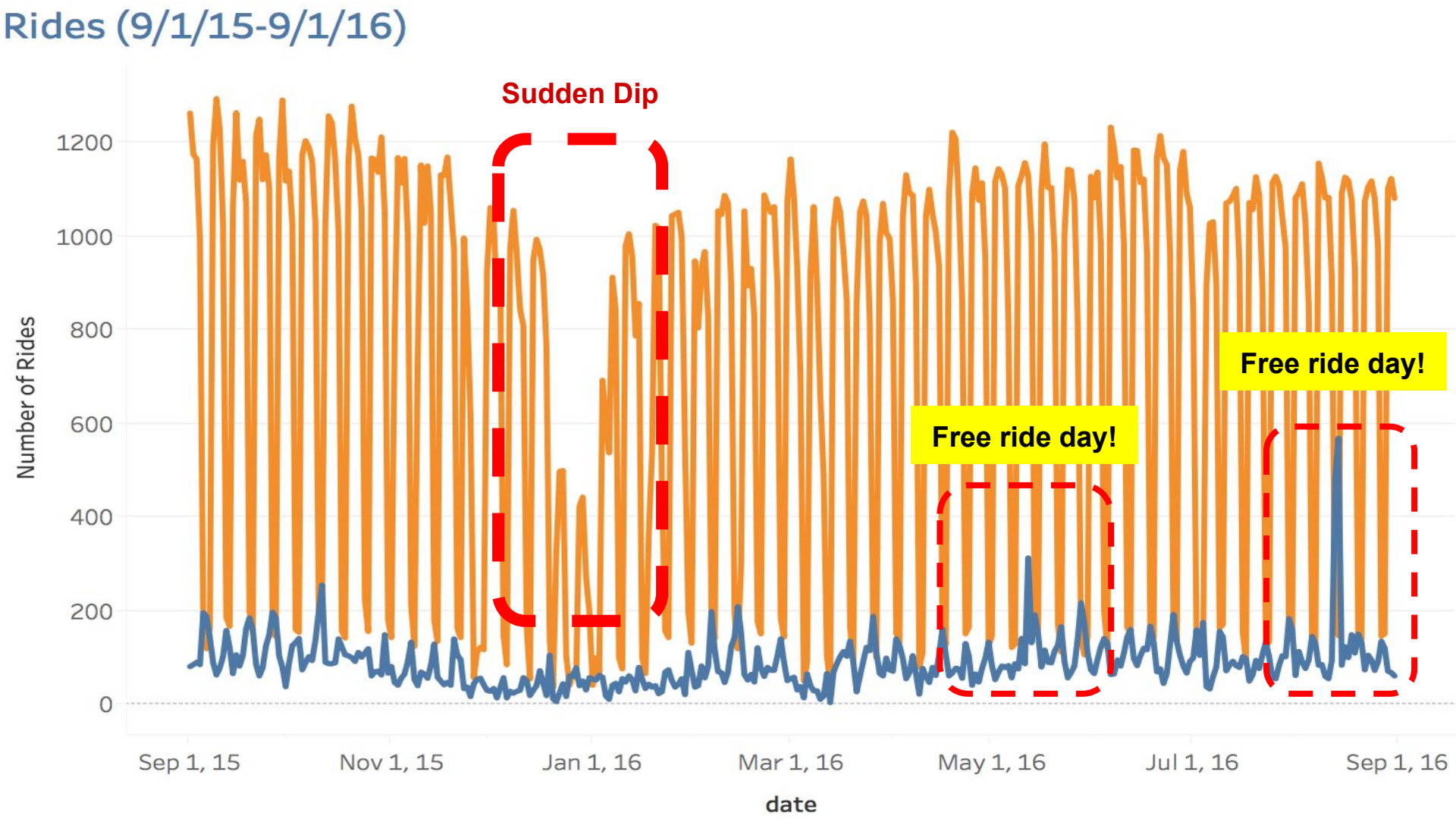
Visualization

Connect MySQL to Tableau

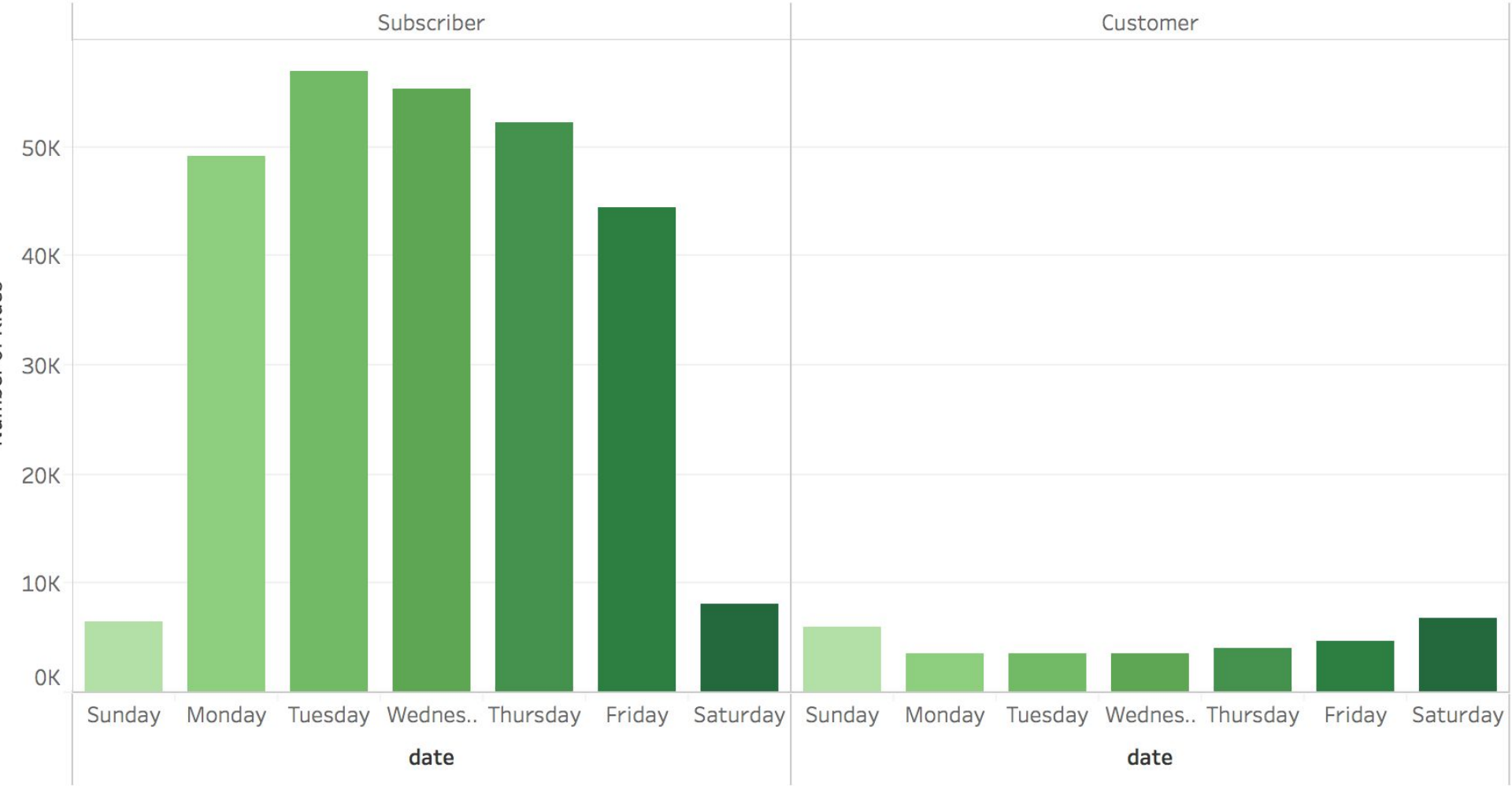


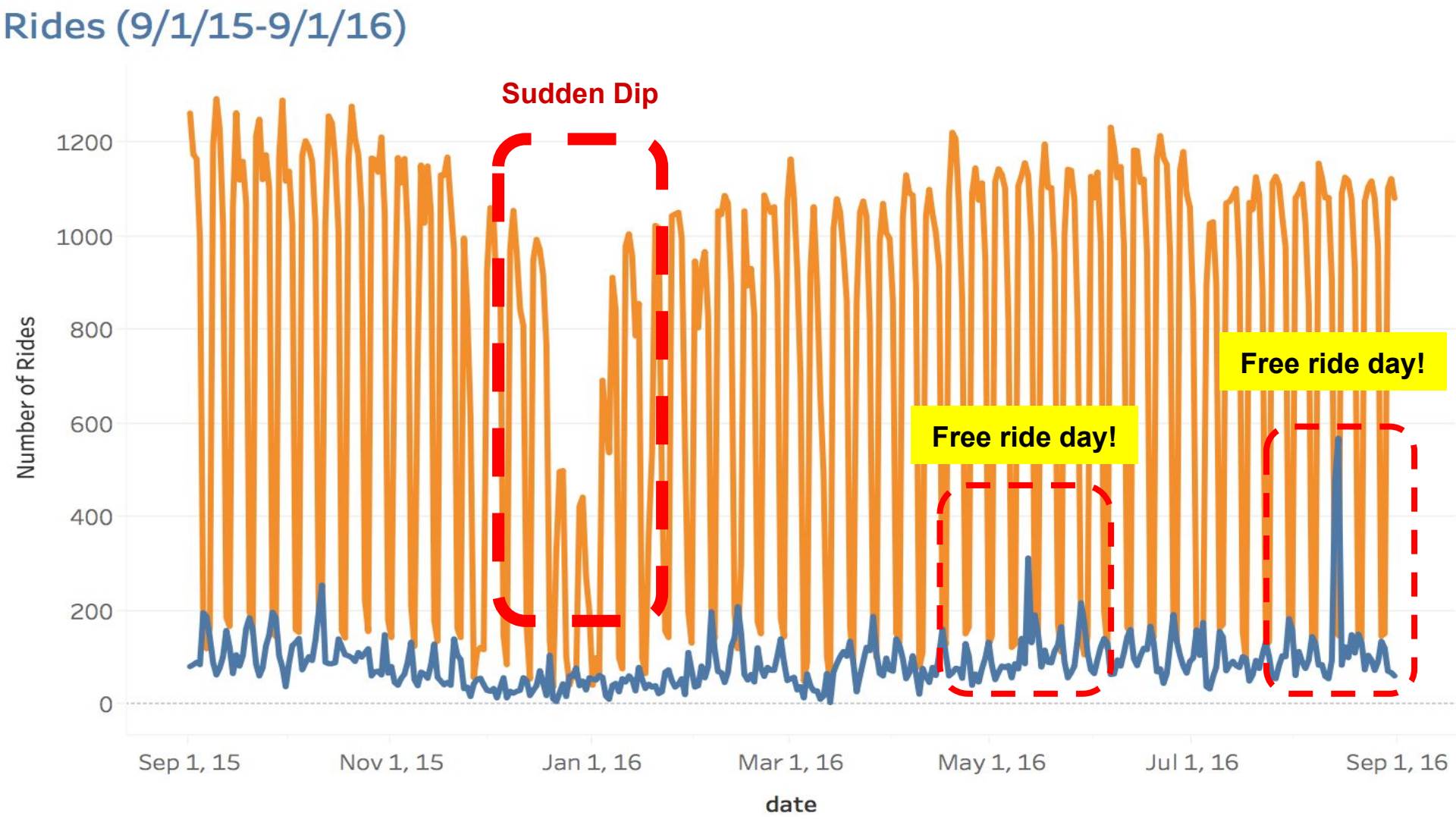
Station location



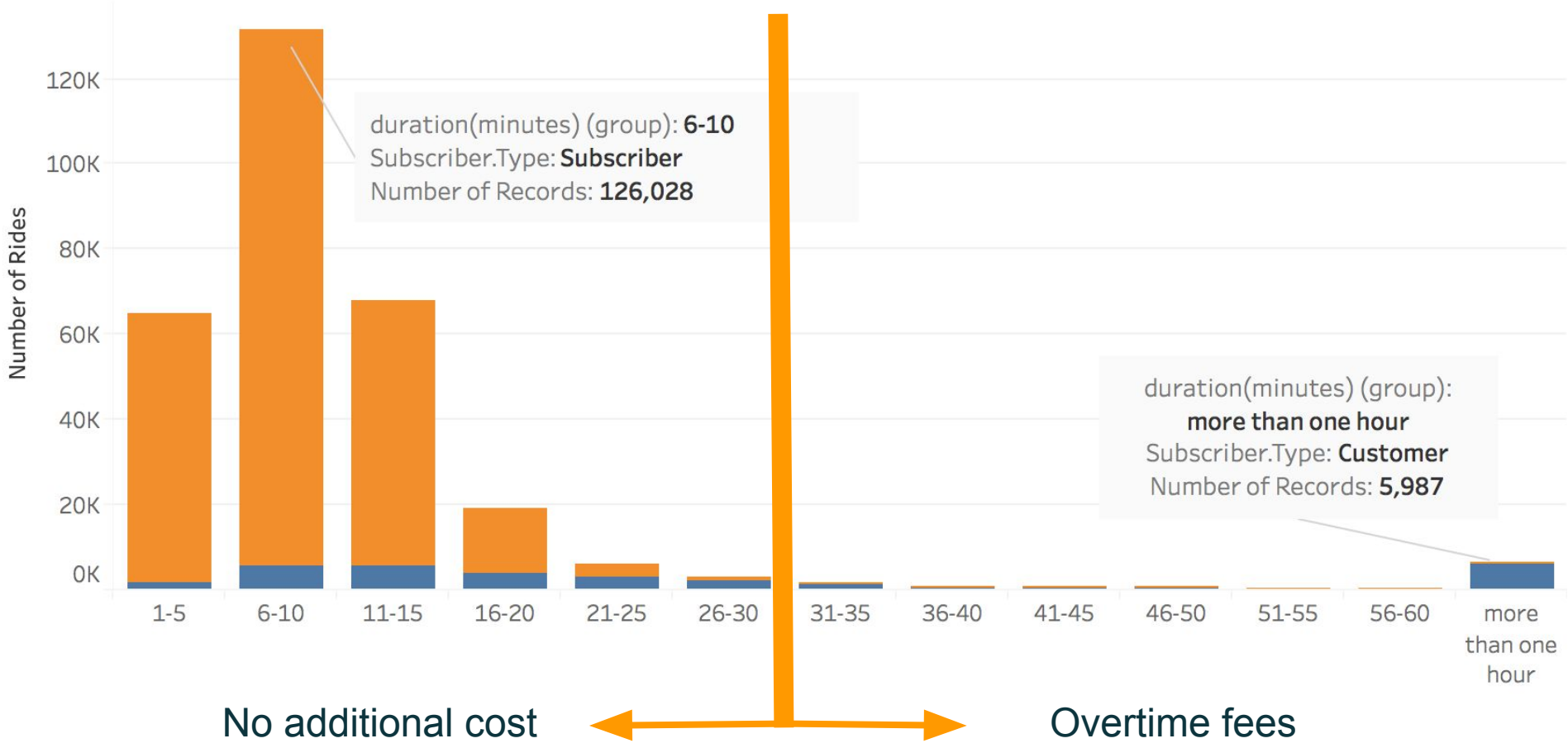

Rides (9/1/15-9/1/16)
Sudden Dip
Free ride day!
Free ride day!
Number of Rides
1200
1000
800
600
400
200
0
Sep 1, 15
Nov 1, 15
Jan 1, 16
Mar 1, 16
May 1, 16
Jul 1, 16
Sep 1, 16
date

Weekday analysis of subscribers & customers

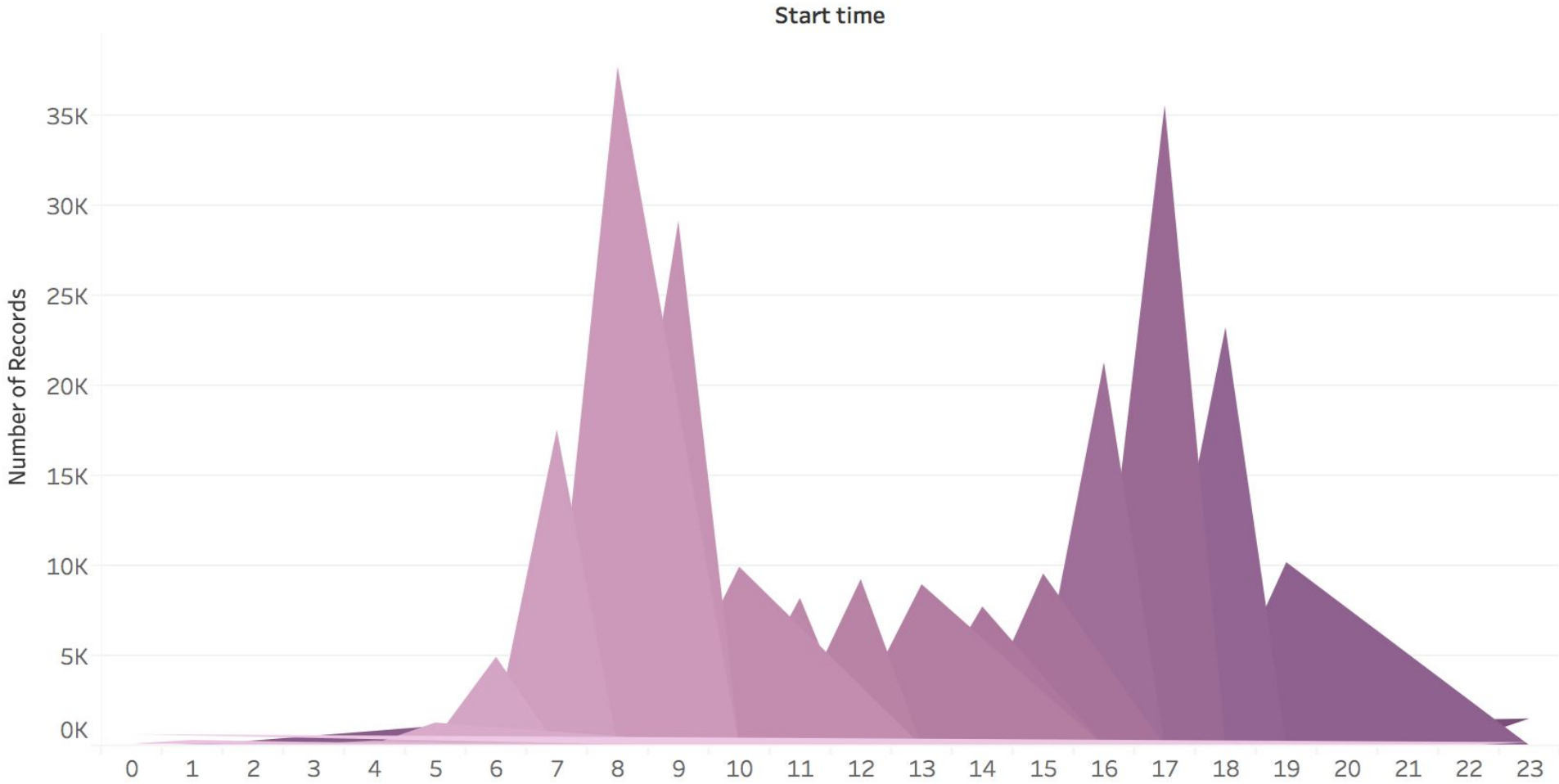



Rides (9/1/15-9/1/16)
Sudden Dip
Free ride day!
Free ride day!
Number of Rides
1200
1000
800
600
400
200
0
Sep 1, 15
Nov 1, 15
Jan 1, 16
Mar 1, 16
May 1, 16
Jul 1, 16
Sep 1, 16
date

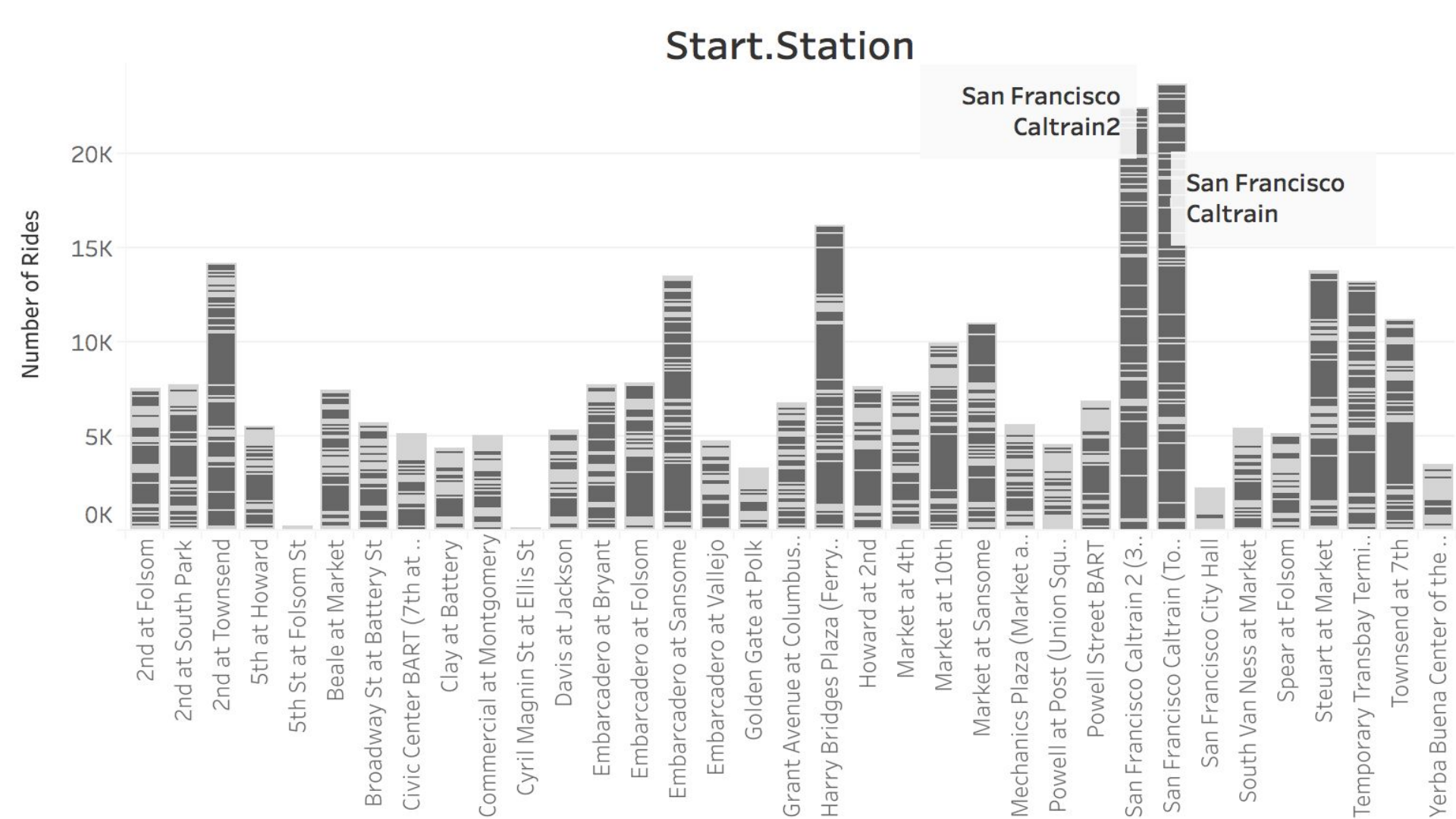
Durations(miniutes)

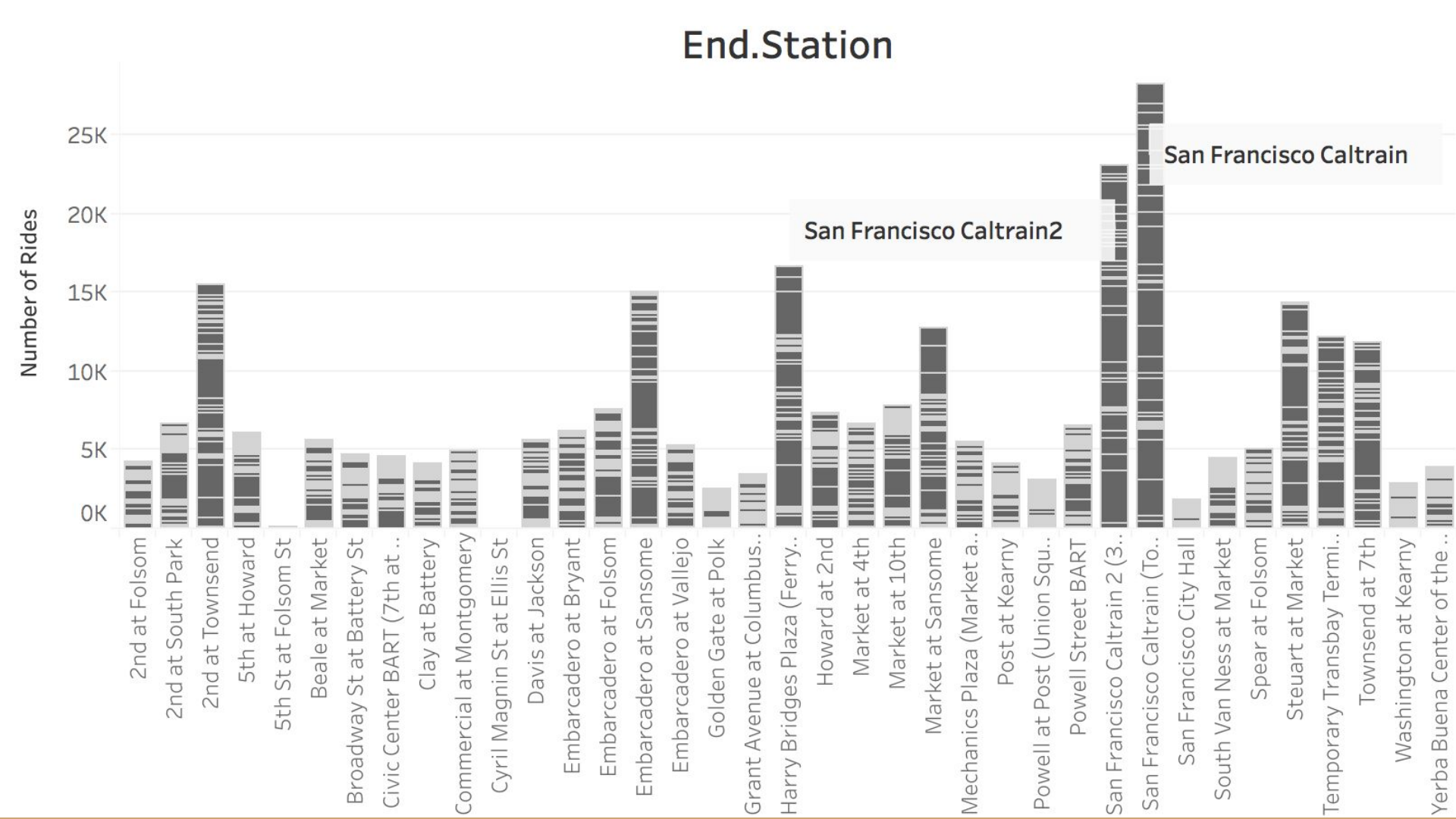


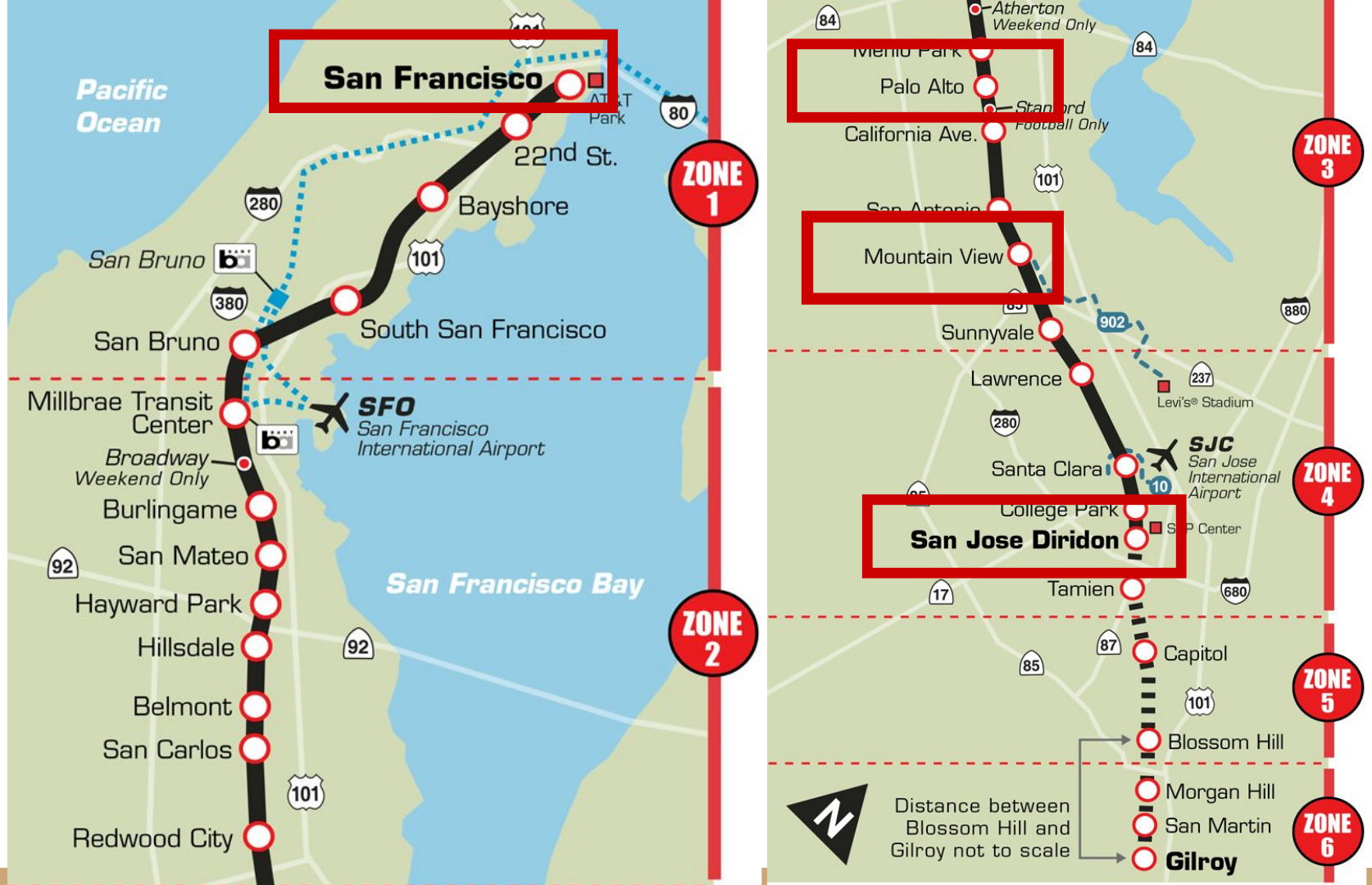
Rides appear in rush hours



San Francisco


Start.Station
Number of Rides
20K
15K
10K
5K
0K
San Francisco Caltrain2
San Francisco Caltrain
2nd at Folsom
2nd at South Park
2nd at Townsend
5th at Howard
5th St at Folsom St
Beale at Market
Broadway St at Battery St
Civic Center BART (7th at ..
Clay at Battery
Commercial at Montgomery
Cyril Magnin St at Ellis St
Davis at Jackson
Embarcadero at Bryant
Embarcadero at Folsom
Embarcadero at Sansome
Embarcadero at Vallejo
Golden Gate at Polk
Grant Avenue at Columbus..
Harry Bridges Plaza (Ferry..
Howard at 2nd
Market at 4th
Market at 10th
Market at Sansome
Mechanics Plaza (Market a..
Powell at Post (Union Squ..
Powell Street BART
San Francisco Caltrain 2 (3..
San Francisco Caltrain (To..
San Francisco City Hall
South Van Ness at Market
Spear at Folsom
Steuart at Market
Temporary Transbay Termi..
Townsend at 7th
Yerba Buena Center of the ..


End.Station
Number of Rides
25K
20K
15K
10K
5K
0K
San Francisco Caltrain
San Francisco Caltrain2
2nd at Folsom
2nd at South Park
2nd at Townsend
5th at Howard
5th St at Folsom St
Beale at Market
Broadway St at Battery St
Civic Center BART (7th at ..
Clay at Battery
Commercial at Montgomery
Cyril Magnin St at Ellis St
Davis at Jackson
Embarcadero at Bryant
Embarcadero at Folsom
Embarcadero at Sansome
Embarcadero at Vallejo
Golden Gate at Polk
Grant Avenue at Columbus..
Harry Bridges Plaza (Ferry..
Howard at 2nd
Market at 4th
Market at 10th
Market at Sansome
Mechanics Plaza (Market a..
Post at Kearny
Powell at Post (Union Squ..
Powell Street BART
San Francisco Caltrain 2 (3..
San Francisco Caltrain (To..
San Francisco City Hall
South Van Ness at Market
Spear at Folsom
Steuart at Market
Temporary Transbay Termi..
Townsend at 7th
Washington at Kearny
Yerba Buena Center of the ..


Pacific Ocean
San Francisco
AT&T Park
22nd St.
Bayshore
South San Francisco
San Bruno
Millbrae Transit Center
Broadway
Weekend Only
Burlingame
San Mateo
Hayward Park
Hillsdale
Belmont
San Carlos
Redwood City
SFO
San Francisco International Airport
San Francisco Bay
ZONE 1
ZONE 2
Atherton
Weekend Only
Menlo Park
Palo Alto
Stanford
Football Only
California Ave.
San Antonio
Mountain View
Sunnyvale
Lawrence
Levi's® Stadium
Santa Clara
SJC
San Jose International Airport
College Park
SAP Center
San Jose Diridon
Tamien
Capitol
Blossom Hill
Morgan Hill
San Martin
Gilroy
Distance between Blossom Hill and Gilroy not to scale
ZONE 3
ZONE 4
ZONE 5
ZONE 6

Key insights

Subscribers are mainly working population, who rides on the rush hours on weekdays for less than 20 min.

Customers are mainly tourists, who rides on the weekends for more than one hour.

Bike stations near caltrain station are the most popular stations. Many people may use bike-caltrain combinations in their routes between their houses and offices.

Thanks & Questions