Bigquery

```
library(bigrquery)
library(tidyverse)
```

```
## -- Attaching packages --------

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts -----------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

When using bigrquery interactively, you'll be prompted to authorize bigrquery in the browser.

- login the Google Cloud Platform
- create a new project
- enable BigQuery API
- add public data

```
bigrquery::bq_auth().
```

```
# replace it with your project id
project <- "adept-vigil-269305"
result <- bq_project_query(
  project,
  "SELECT * FROM `bigquery-public-data.samples.gsod` LIMIT 100;")
```

```
## Using an auto-discovered, cached token.
## To suppress this message, modify your code or options to clearly consent to the use of a cached toke
## See gargle's "Non-interactive auth" vignette for more details:
## https://gargle.r-lib.org/articles/non-interactive-auth.html
## The bigrquery package is using a cached token for randy.cs.lai@gmail.com.
```

```
bq_table_download(result)
```

```
## # A tibble: 100 x 31
##    station_number wban_number  year month   day mean_temp num_mean_temp_s~
##             <int>       <int> <int> <int> <int>     <dbl>            <int>
##  1          38110       99999  1929    12    11      52.8                4
##  2          30750       99999  1930     1    13      37.2                4
```

```
##  3          36010       99999  1930    10     7      53                 4
##  4          39800       99999  1931     9     2      52.6               5
##  5         726810       24131  1931     9    18      67.4              24
##  6         726810       24131  1931     5    30      72.8              24
##  7         726815       24106  1932     7    15      70.9              24
##  8         726815       24106  1932     5     6      53.6              24
##  9         726810       24131  1932    12    10       4.5              24
## 10         726815       24106  1932     9    29      62.1              24
## # ... with 90 more rows, and 24 more variables: mean_dew_point <dbl>,
## #   num_mean_dew_point_samples <int>, mean_sealevel_pressure <dbl>,
## #   num_mean_sealevel_pressure_samples <int>, mean_station_pressure <dbl>,
## #   num_mean_station_pressure_samples <int>, mean_visibility <dbl>,
## #   num_mean_visibility_samples <int>, mean_wind_speed <dbl>,
## #   num_mean_wind_speed_samples <int>, max_sustained_wind_speed <dbl>,
## #   max_gust_wind_speed <dbl>, max_temperature <dbl>,
## #   max_temperature_explicit <lgl>, min_temperature <dbl>,
## #   min_temperature_explicit <lgl>, total_precipitation <dbl>,
## #   snow_depth <dbl>, fog <lgl>, rain <lgl>, snow <lgl>, hail <lgl>,
## #   thunder <lgl>, tornado <lgl>
```

Upload dataset

You could upload via the web interface or using `bq_` functions.

```
mydataset <- bq_dataset(project, "mydataset")
bq_dataset_create(mydataset)
bq_dataset_exists(mydataset)
```

Let's try to upload the `mtcars` dataset and pretend that it is huge.

```
ta <- bq_table(mydataset, "mtcars")
bq_table_create(
  ta,
  friendly_name = "Motor Trend Car Road Tests",
  description = "The data was extracted from the 1974 Motor Trend US magazine",
  labels = list(category = "example")
)
bq_table_exists(ta)

cars <- mtcars %>%
  mutate(cyl = as_factor(cyl), vs = as_factor(vs), am = as_factor(am))
bq_table_upload(ta, cars, fields = as_bq_fields(cars))
```

Now, let's have some fun.

There are three interfaces provided by `bigrquery`. - Low level API overy REST - DBI - dplyr

*bq_*

```
result <- bq_project_query(
  project,
  "SELECT * FROM `adept-vigil-269305.mydataset.mtcars` where `mpg` < 30")
bq_table_download(result)
```

```
## # A tibble: 28 x 11
##      mpg cyl    disp    hp  drat    wt  qsec vs    am     gear  carb
##    <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <chr> <dbl> <dbl>
##  1  24.4 4      147.    62  3.69  3.19  20   1     0         4     2
##  2  22.8 4      141.    95  3.92  3.15  22.9 1     0         4     2
##  3  21.5 4      120.    97  3.7   2.46  20.0 1     0         3     1
##  4  21.4 6      258    110  3.08  3.22  19.4 1     0         3     1
##  5  18.1 6      225    105  2.76  3.46  20.2 1     0         3     1
##  6  19.2 6      168.   123  3.92  3.44  18.3 1     0         4     4
##  7  17.8 6      168.   123  3.92  3.44  18.9 1     0         4     4
##  8  18.7 8      360    175  3.15  3.44  17.0 0     0         3     2
##  9  14.3 8      360    245  3.21  3.57  15.8 0     0         3     4
## 10  16.4 8      276.   180  3.07  4.07  17.4 0     0         3     3
## # ... with 18 more rows
```

```
library(DBI)
con <- dbConnect(
  bigquery(),
  project = project,
  dataset = "mydataset"
)
```

DBI

```
con %>% dbGetQuery("SELECT * FROM `adept-vigil-269305.mydataset.mtcars` WHERE `mpg` < 30")
```

```
## # A tibble: 28 x 11
##      mpg cyl    disp    hp  drat    wt  qsec vs    am     gear  carb
##    <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <chr> <dbl> <dbl>
##  1  24.4 4      147.    62  3.69  3.19  20   1     0         4     2
##  2  22.8 4      141.    95  3.92  3.15  22.9 1     0         4     2
##  3  21.5 4      120.    97  3.7   2.46  20.0 1     0         3     1
##  4  21.4 6      258    110  3.08  3.22  19.4 1     0         3     1
##  5  18.1 6      225    105  2.76  3.46  20.2 1     0         3     1
##  6  19.2 6      168.   123  3.92  3.44  18.3 1     0         4     4
##  7  17.8 6      168.   123  3.92  3.44  18.9 1     0         4     4
##  8  18.7 8      360    175  3.15  3.44  17.0 0     0         3     2
##  9  14.3 8      360    245  3.21  3.57  15.8 0     0         3     4
## 10  16.4 8      276.   180  3.07  4.07  17.4 0     0         3     3
## # ... with 18 more rows
```

dplyr

```
con %>% tbl("mtcars") %>%
  filter(mpg < 30) %>%
  collect()
```

```
## # A tibble: 28 x 11
##      mpg cyl    disp    hp  drat    wt  qsec vs    am     gear  carb
##    <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <chr> <dbl> <dbl>
##  1  24.4 4      147.    62  3.69  3.19  20   1     0         4     2
##  2  22.8 4      141.    95  3.92  3.15  22.9 1     0         4     2
```

```
##  3  21.5 4      120.    97  3.7   2.46  20.0 1     0         3     1
##  4  21.4 6      258    110  3.08  3.22  19.4 1     0         3     1
##  5  18.1 6      225    105  2.76  3.46  20.2 1     0         3     1
##  6  19.2 6      168.   123  3.92  3.44  18.3 1     0         4     4
##  7  17.8 6      168.   123  3.92  3.44  18.9 1     0         4     4
##  8  18.7 8      360    175  3.15  3.44  17.0 0     0         3     2
##  9  14.3 8      360    245  3.21  3.57  15.8 0     0         3     4
## 10  16.4 8      276.   180  3.07  4.07  17.4 0     0         3     3
## # ... with 18 more rows
```

```
SELECT * FROM `adept-vigil-269305.mydataset.mtcars` WHERE `mpg` < 30;
```

Table 1: Displaying records 1 - 10

mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3

Running linear regression in Bigquery

```
CREATE VIEW `mydataset.mtcars2` AS
  SELECT *,
  RAND() as `train`
  FROM `adept-vigil-269305.mydataset.mtcars`
```

```
CREATE MODEL `mydataset.mtcars_model`
OPTIONS
  (model_type='linear_reg',
    input_label_cols=['mpg']) AS
SELECT
  `mpg`,
  `cyl`,
  `disp`,
  `hp`,
  CAST(`gear` AS string) AS `gear`
FROM
  `adept-vigil-269305.mydataset.mtcars2`
WHERE
  `train` < 0.9  -- select rows randomly
```

If you want to delete the model

```
DROP MODEL `mydataset.mtcars_model`;
```

To do prediction

```
SELECT * FROM ML.PREDICT(MODEL `adept-vigil-269305.mydataset.mtcars_model`, (
  SELECT
    `cyl`,
    `disp`,
    `hp`,
    CAST(`gear` AS string) AS `gear`
  FROM `adept-vigil-269305.mydataset.mtcars2` WHERE `train` >= 0.9
))
```

Reference

BigQuery: https://cloud.google.com/bigquery-ml/docs/reference/standard-sql