# ETL - Extract, Transfer and Load

# Moving the data to a convenient structure

# into valid\_commits

# Legacy sql

# Allow Large Results

Select

c.repo\_name as repo\_name

, commit

, message

, subject

, author.date

, author.name

, author.email

, difference.old\_path

From

flatten(flatten([bigquery-public-data:github\_repos.commits], repo\_name), difference) as c

Join

[citric-trees-187317:repo\_attributes.valid\_repos\_a] as v

On c.repo\_name = v.repo\_name

# into valid\_commits\_2018

# Legacy sql

# Allow Large Results

Select \*

From

[citric-trees-187317:repo\_attributes.valid\_commits]

Where year(author\_date) = 2018

# into valid\_files

# Standard sql

# Allow Large Results

Select

F.repo\_name as repo\_name

, ref

, path

, lower(reverse(substr(reverse(path), 0, strpos(reverse(path),'.')))) as extension

, mode

, id

, symlink\_target

From

`bigquery-public-data.github\_repos.files` as f

Join

`citric-trees-187317.repo\_attributes.valid\_repos\_a` as v

On f.repo\_name = v.repo\_name

# into valid\_commits\_by\_key

# Legacy sql

# Allow Large Results

Select

repo\_name

, Commit

, max(message) as message # Message is the use. Max is used for aggregation

, max(author\_date) as author\_date

, max(author\_name) as author\_name

, max(author\_email) as author\_email

From

[citric-trees-187317.repo\_attributes.valid\_commits]

Group by

repo\_name

, Commit