1 BigQuery SQL

- work with date
  - EXTRACT(DAY FROM date_expression)
- arithmatic operation with numeric data
- string functions
  - check if contains ampersand sign: `CONTAINS_SUBSTR(name, '&')`
  - string append: `||`

2 command line

- upload data to bucket (`my_bucket`)
  - upload a file to bucket: `gsutil cp data1.csv gs://my_bucket`
  - upload a folder to bucket: `gsutil cp -r myfolder/ gs://my_bucket`
  - list files in my bucket: `gsutil ls gs://my_bucket`
- work with BigQuery (`bq`)
  - list datasets: `bq ls`
  - view a table in a database: `bq show projectId:datasetId.tableId`
  - run a query:

```
bq query --use_legacy_sql=false \
'SELECT *
FROM talbe'
```

- BigQuery architecture
  - decoupling/denormolization of storage and compute. BigQuery is based on a number of foundational technologies deployed in Google. Google has created a replicated distributed storage system that is based on a couple of systems. One is Colossus (globally distributed storage system) and Spanner (globally scalable relational database, for managing metadata about the data we store in BigQuery). The compute side is made up of two Google technologies. One is called Dremel which is a system that takes our SQL queries and breaks them down into distinct executable chunks and then produces the results based on those chunks and merges the results and delivers the results to us. Dremel does this by creating or deploying this work in jobs. A job is an abstraction for running a query or other operations within BigQuery. The system that manages jobs is known as Borg. Borg is a predecessor of Kubernetes. A lot of ideas that we're developed in Borg kind of influence the design of Kubernetes as well.

    Because the storage and the compute are separated. It's really important that we're able to quickly move data between storage and compute. And that's done using the Jupiter network. Jupiter network is network hardware that provides for petabit scale throughput and it is a key to successfully decoupling storage and compute.