11/26/18 Notes

Applications of SQL

- **Partitioning** - segmenting the tables by a specific field to improve performance on associated queries
    - **Syntax** - `CREATE TABLE ... PARTITION BY [field]`
    - More information on partitioned table management, creation, querying: https://cloud.google.com/bigquery/docs/partitioned-tables
- **CASE Keyword** - used as a multi-conditional statement that always returns a value;
    - **Syntax** - `CASE [expression] WHEN [value] THEN [result when expression evaluates to value] [WHEN ...] ELSE [default result] END`

Usage of DataFlow

- **Only one** worker in DataFlow is required to read from BigQuery. *(Why?)*
- Worker count increases/decreases according to the performance of the job
- for testing DoFn's, debugging with print statements is fine.
    - with **DirectRunner**, `print(...)` will work fine and will return in console
    - with **DataflowRunner**, `logging.info(...)` is required since these jobs are not only remote but distributed (running on several workers)

Advice

- **unit testing** - running small parts of a job to make sure all parts do what you think it does
- **cleaning CloudShell** - remove old output logs and outdated scripts to prevent confusion