Personal Project Report

1. Motivation

In my home country, each province has their own famous industries, because each province varies in history, culture, geography and even policies. For instance, my home province is specifically outstanding in food, education and culture industries. So I am wondering if in the US, each state has characteristic industries as well. And Yelp is a very helpful and straightforward platform that collect and display the data

2. About the Data

I downloaded the data from Kaggle.

In the downloaded folder, five json files are included, and two of them are utilized. They are Business and Reviews. The size of Business is 138.3M, and that of Reviews is 5.3G.

Here are the samples of Business and Review in sequence.


```
root: {} 14 items
  business_id: f9NumwFMBDn751xgFiRbNA
  name: The Range At Lake Norman
  address: 10913 Bailey Rd
  city: Cornelius
  state: NC
  postal_code: 28031
  latitude: 35.4627242
  longitude: -80.8526119
  stars: 3.5
  review_count: 36
  is_open: 1
  attributes: {} 6 items
    BusinessAcceptsCreditCards: True
    BikeParking: True
    GoodForKids: False
    BusinessParking: {'garage': False, 'street': False, 'validated': False, 'lot': True, 'valet': False}
    ByAppointmentOnly: False
    RestaurantsPriceRange2: 3
  categories: Active Life, Gun/Rifle Ranges, Guns & Ammo, Shopping
  hours: {} 7 items
    Monday: 10:0-18:0
    Tuesday: 11:0-20:0
    Wednesday: 10:0-18:0
    Thursday: 11:0-20:0
    Friday: 11:0-20:0
    Saturday: 11:0-20:0
    Sunday: 13:0-18:0
```

```
root: {} 9 items
  review_id: xQY8N_XvtGbearJ5X4QryQ
  user_id: OwjRMXRC0KyPrIlcjaXeFQ
  business_id: -MhfebM0QIsKt87iDN-FNw
  stars: 2
  useful: 5
  funny: 0
  cool: 0
  text: As someone who has worked with many museums, I was eager to visit this gallery on my most recent trip to Las Vegas. When

  Tucked away near the gelateria and the garden, the Gallery is pretty much hidden from view. It's what real estate agents would

  That being said, you can still see wonderful art at a gallery of any size, so why the two *s you ask? Let me tell you:

  * pricing for this, while relatively inexpensive for a Las Vegas attraction, is completely over the top. For the space and the
  * it's not kid friendly at all. Seriously, don't bring them.
  * the security is not trained properly for the show. When the curating and design teams collaborate for exhibitions, there is a

  At such a *fine* institution, I find the lack of knowledge and respect for the art appalling.
  date: 2015-04-15 05:21:16
```


3. **BigQuery**
   a. How did I load the data to BigQuery?
   The raw data is loaded into GCP storage, after processing, the final data is loaded into BigQuery by Java Code

```java
// 5. Write results to GCS as well as to BigQuery.
PCollection<String> predictionJson =
        cas.apply(MapElements.into(TypeDescriptors.strings()).via((CategoryOuterClass.Category data) -> {
            try {
                return ProtoUtils.getJsonFromMessage(data, omitSpace: true);
            } catch (InvalidProtocolBufferException e) {
                e.printStackTrace();
            }
            return null;
        }));
predictionJson.apply(TextIO.write().to(config.getWritePathToPredictionData()).withNumShards(1));

// 6. Write to BigQuery.
if (options.getExportToBigQuery()) {
    cas.apply(BigQueryIO.<CategoryOuterClass.Category>write().to(Main.DEST_TABLE)//
            .withWriteDisposition(BigQueryIO.Write.WriteDisposition.WRITE_TRUNCATE)//
            .withCreateDisposition(BigQueryIO.Write.CreateDisposition.CREATE_IF_NEEDED) //
            .withSchema(new TableSchema().setFields(BQ_FIELDS))//
            .withFormatFunction((SerializableFunction<CategoryOuterClass.Category, TableRow>) input //
                    -> new TableRow().set("state",input.getState()).set("title", input.getTitle())//
                    .set("num_of_good_reviews", input.getNumOfGoodReview()).set("stars", input.getAverageStar())));
}
p.run().waitUntilFinish();
```

b. The schema of data in BigQuery

The table in BigQuery should look like this.

State	Categories	Good_reviews_count	Stars
NC	Active Life	Unknown	Unknown
NC	Gun/Rifle Ranges	Unknown	Unknown
NC	Guns & Ammo	Unknown	Unknown
NC	Shopping	Unknown	Unknown

c. Analytics

This table is produced for users to do some further analytics. For example, I want to know what are the top five industries in Nevada and Alabama, with respect to number of good reviews, then I would use this Query

```
select*from (select *,RANK() over (partition by state order by num_of_good_reviews DESC )
as rank from Project_03.PP_table where (state='NV' or state='AB') )where rank <=5;
```

And this is the result.

Row	state	title	num_of_good_reviews	stars	rank
1	AB	restaurants	5166	4.132617	1
2	AB	nightlife	1989	4.1233697	2
3	AB	bars	1989	4.1233697	2
4	AB	breakfast & brunch	1661	4.196049	4
5	AB	food	1616	4.07613	5
6	NV	restaurants	680405	4.068974	1
7	NV	food	245833	4.1761446	2
8	NV	nightlife	200654	4.0524697	3
9	NV	bars	177514	4.026728	4
10	NV	american (new)	143299	4.021118	5

Alternatively, I can get the top five industries with average ratings.

```
select*from (select *,RANK() over (partition by state order by stars DESC )
as rank from Project_03.PP_table where (state='NV' or state='AB') )where rank <=5;
```

And have a slightly different result.

Row	state	title	num_of_good_reviews	stars	rank
1	AB	himalayan/nepalese	212	4.617778	1
2	AB	ice cream & frozen yogurt	344	4.487363	2
3	AB	gastropubs	115	4.452381	3
4	AB	breweries	115	4.452381	3
5	AB	macarons	111	4.3656716	5
6	NV	home health care	198	4.9949493	1
7	NV	rafting/kayaking	176	4.9829545	2
8	NV	mountain biking	176	4.9829545	2
9	NV	property management	300	4.981132	4
10	NV	watch repair	422	4.9663577	5

d. Reason for BigQuery
I plan to provide a table so that anyone can use SQL to manipulate and do some analytics, and this is the reason I need BigQuery.

4. **Beam / Dataflow**
   a. Pre-process Data

I created proto messages for Business, Review and Category to extract and hold useful fields.

```
message Business{

    string business_id = 1;

    string state = 2;

    float stars = 3;

    repeated string category = 4;

    int64 good_review_count = 5;
}
```

```
message Review{

    string business_id = 1;

    float stars = 2;

    string review_id = 3;
}
```

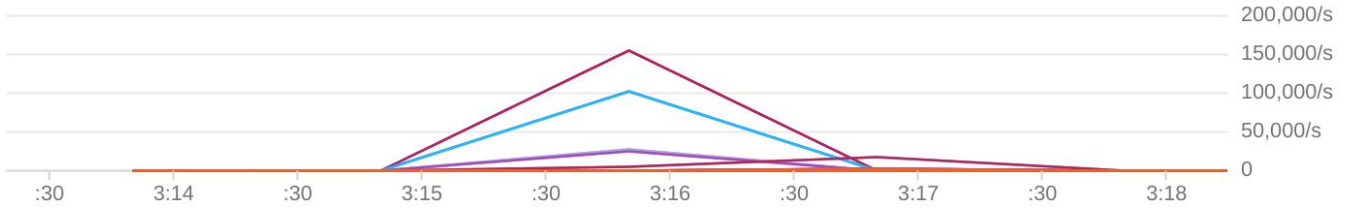
```
message Category{
    string state = 1;

    string title = 2;

    int64 num_of_good_review = 3;

    float average_Star = 4;
}
```

I used PTransform to process the raw data. For this part, please refer to code: https://github.com/Yuyao-Xie/cs686-Personal-Project/tree/master/java/dataflow/src/main/java/edu/usfca/dataflow/transforms.

e. Job Metrics

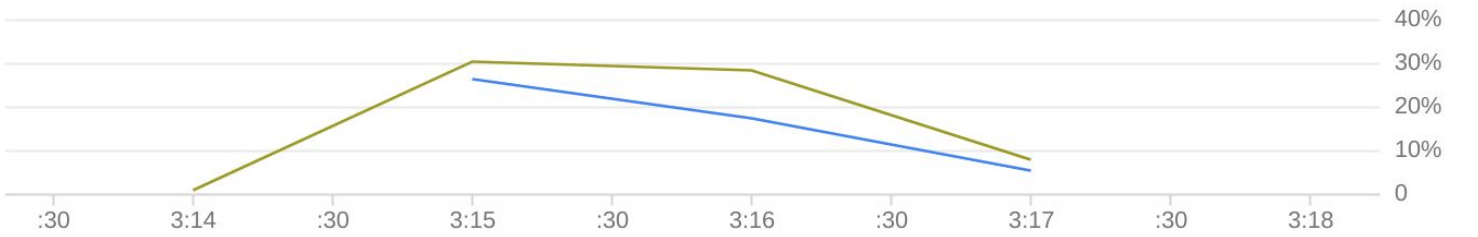
i. Job metrics.

Throughput (elements/sec) Create alerting policy



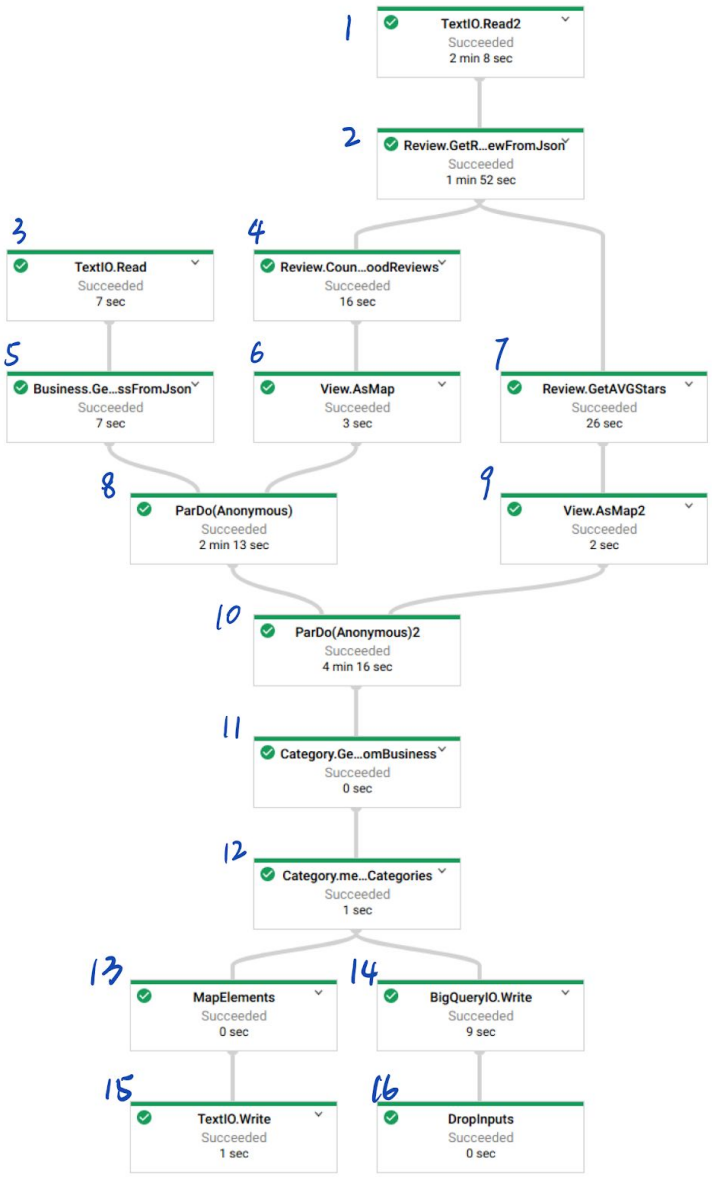
Name	Value
BigQueryIO.Write/BatchLoads/Create.Values/Read(CreateSource)	0
BigQueryIO.Write/BatchLoads/CreateJobId	0
BigQueryIO.Write/BatchLoads/GetTempFilePrefix	0
BigQueryIO.Write/BatchLoads/JobIdCreationRoot/Read(CreateSource)	0

CPU utilization (All Workers) Create alerting policy



Name	Value
yelpjob-84009-05041513-ld6i-harness-kpnm	7.53%
yelpjob-84009-05041513-ld6i-harness-r82c	5.07%

ii. Job Steps



For step 3 and 1 read Business and Review from json strings. Step 2 and 5 produce PCollections of Review and Business. Step 4 counts good reviews for each business, and step 7 computes average rates for each business, and step 6 and 9 turn the result of 4 and 7 into PCollectionView for reference. Then the PCollection of business produced from step 5, takes the two PCollectionViews from step 6 and 9 as side inputs and produces PCollection of categories in step 11. Up to now, each business message is turned into a category message. Obviously, this mapping does not make any sense, then in sep 12, categories with the same title and state are merged together. Finally, in step 14, all categories are exported into BigQuery, and the table example is in 3-b.

5. Link

Link to GCP: yelpJob
Link to BigQuery: PP_table