Please finish this in a databricks notebook, and upload your notebook in .ipynb format.

1. About CDR Data

CDR is the name of a **Call Detail Record** .

The data is located on WeCloudData's S3 bucket, the folder is CDR. The detailed S3 link is *s3://weclouddata/datasets/telecom/CDR*.

In the CDR folder, each month has a sub-folder, and each day has a single file. Like this:

```
aws s3 ls s3://weclouddata/datasets/telecom/CDR/cdr_by_grid_december/
2018-08-15 21:27:35  298715145 sms-call-internet-mi-2013-12-01.txt
2018-08-15 21:27:35  341919663 sms-call-internet-mi-2013-12-02.txt
2018-08-15 21:27:35  353947238 sms-call-internet-mi-2013-12-03.txt
2018-08-15 21:27:35  352032545 sms-call-internet-mi-2013-12-04.txt
2018-08-15 21:27:35  353519447 sms-call-internet-mi-2013-12-05.txt
2018-08-15 21:32:13  354028475 sms-call-internet-mi-2013-12-06.txt
2018-08-15 21:32:40  307172220 sms-call-internet-mi-2013-12-07.txt
2018-08-15 21:32:45  293663723 sms-call-internet-mi-2013-12-08.txt
2018-08-15 21:32:45  338963272 sms-call-internet-mi-2013-12-09.txt
2018-08-15 21:32:45  349269739 sms-call-internet-mi-2013-12-10.txt
2018-08-15 21:37:27  351378560 sms-call-internet-mi-2013-12-11.txt
2018-08-15 21:37:28  353930588 sms-call-internet-mi-2013-12-12.txt
2018-08-15 21:37:31  351454734 sms-call-internet-mi-2013-12-13.txt
2018-08-15 21:38:09  310080286 sms-call-internet-mi-2013-12-14.txt
2018-08-15 21:38:12  296653479 sms-call-internet-mi-2013-12-15.txt
```

In the above example, you can see: under cdr folder, there a sub-folder cdr_by_grid_december; and under the December sub-folder, each day has a file.

Each file has the same columns, they are:

- **Square id**: the id of the square that is part of the Milano GRID; TYPE: numeric
- **Time interval**: the beginning of the time interval expressed as the number of millisecond elapsed from the Unix Epoch on January 1st, 1970 at UTC. The end of the time interval can be obtained by adding 600000 milliseconds (10 minutes) to this value. TYPE: numeric
- **Country code**: the phone country code of a nation. Depending on the measured activity this value assumes different meanings that are explained later. TYPE: numeric
- **SMS-in activity**: the activity in terms of received SMS inside the Square id, during the Time interval and sent from the nation identified by the Country code. TYPE: numeric
- **SMS-out activity**: the activity in terms of sent SMS inside the Square id, during the Time interval and received by the nation identified by the Country code. TYPE: numeric
- **Call-in activity**: the activity in terms of received calls inside the Square id, during the Time interval and issued from the nation identified by the Country code. TYPE: numeric
- **Call-out activity**: the activity in terms of issued calls inside the Square id, during the Time interval and received by the nation identified by the Country code. TYPE: numeric

Internet traffic activity: the activity in terms of performed internet traffic inside the Square id, during the Time interval and by the nation of the users performing the connection identified by the Country code . TYPE: numeric

2. Project Requirement

Finish the entire project in the Databricks.

1. Mount CDR data from S3 bucket to Databricks 'hdfs' files system's /mnt/ folder. So your mounted folder will be like /mnt/cdr; if need help to mount the data, please refer to this code.
2. Please read the AWS ACCESS_KEY, SECRET_ACCESS_KEY from 'https://wcd-de-labs-files.s3.amazonaws.com/key.json' using library requests.
3. Only use the sub-folder --- cdr_by_grid_december in the CDR folder;
4. We will use top 5 days data to do analysis, which means you will read 5 files from the sub-folder.

3. Tasks

Please finish the following tasks in databricks:

1. **Change the column names**. The initial column names are using '-', please replace '-' with '_'.
2. **Add a new column.** create a new column 'sms_ratio' show the ratio of 'sms-in-activity/sms-out-activity'.
3. **Create a date column:** we need change the 'time_interval' column to timestamp first ('time_interval'/1000 --> change to timestamp type), and then change the date foramt to 'yyyy/MM/dd'.
4. **Calculate summaray statistics at the square_id level.** Create a dataframe calculate the aggregation of:
    - sms_in_activity ==> mean
    - sms_out_activity ==> mean
    - call_out_activity ==> min
    - internet_activity ==> max
    - all records ==> count
5. **Find the min and max:** Group by 'square_id', find out the min and max value in columns ['sms_in_activity', 'sms_out_activity','internet_activity','call_in_activity','call_out_activity'].
6. **Create an summary table**:
    - Generate an aggregate table to summary how many sms, call and internet activities in each country each day. Be careful of the value null in the columns. If the cell is null, it is not counted as an activity.
    - Write the dataframe to the 'tmp' folder in parquet format.

    - (Optional) Try to mount the your own aws s3 bucket to databricks, and write the file to your own s3 bucket.
7. **Create a dataframe rank internet activity with Window function.** Based on the dataframe from Task 6, use window function, partition by coutry_code, rank the total internect activities of each day.