Skills tests

1. ETL task contains three phases: Extract, Transform and Load.

Extract Phase:

The extraction phase aims to convert the data into a single format appropriate for transformation processing. Most data-warehousing projects combine data from different source systems. In our case, the data is already available in unique format and we don't have to combine data from different source systems. We are reading data from csv file.

Transform Phase:

In the data transformation stage, rules are applied to the extracted data in order to prepare it for loading into the end target. We don't know how clean or dirty the data is and we have to take care about this problem.

This is one line of csv file:
blogId, views, clicks
1, 10, 3

Let view be one search for some blog which Id is blogId. Let click be number of blog's visit which Id is blogId.

Special cases include:
-missing blog's id: in this case we can't make predictions about daily traffic so we drop rows from csv file

-views negative values: in this case, we need to check if this value is valid. Correction is done with same value without negative sign.

-clicks negative values: in this case, we need to check if this value is valid. Correction is done with same value without negative sign.
-missing view value: in this case, I was looking for two values for current blog's id, using previous and next value. Missing value is filled with an average of these two values. If before or after missing view value doesn't exist two inputs with the same blog Id that have valid view values, row with missing value is deleted.
-missing click value: in this case, I was looking for two values for current blog's id, using previous and next value. Missing click is filled with an average of these two values. If before or after missing click value doesn't exist two inputs with the same blog Id that have valid view values, row with missing value is deleted.

In output, table contains new column date, with different date for every row so BI/Analytics tool can make predictions about blog traffic by day.

Load Phase:

This phase loads the data into the end target, which may be a csv file or a data warehouse. In this phase, it is important that all validation is made before loading data into warehouse. I have already applied these rules in transform phase. Also, we can disable integrity checking in the target database tables during the load, dropping the indices before the load and recreate them after the load. Removing duplicates using distinct may be slow in the database; so, it makes sense to do it outside. On the other side, if using distinct significantly decreases the number of rows to be extracted, then it makes sense to remove duplications as early as possible in the database before unloading data. This situation can occur in our problem, because it’s possible to have identically records, as thay are informing about number of visits and clicks for some blog.

For implementation, I have used IDE Jupyter, Python 3.0 language and Anaconda, an open source distribution for Python.

2. I used report for table name from csv_file.
I used isnumeric function because in csv file exists one row with non numeric AP_amount value.
2a:

```
select department_family, sum(isnumeric(AP_amount) select AP_amount else select 0)

from report

group by department_family
```

2b:

```
select department_family, expense_type, sum(isnumeric(AP_amount) select AP_amount else select 0)
from report
group by department_family, expense_type
```

3.

```
with auxiliary_table (expense_type, expense_area, amount) as
(

        select expense_type, expense_area, sum(AP_amount)
        from report
        group by expense_type, expense_area

)


create table expense_per_area_and_type (expense_type varchar(255) not null, expense_area varchar(255) not null, amount int not null, primary_key(expense_type, expense_area)

insert into expense_per_area_and_type
select *
from auxiliary_table
```