ClickHouse Lab

During this lab you have to implement Data Warehouse (DWH) using ClickHouse (CH) and its techniques such as Materialized View (MV) and Distributed tables. Max score for this lab - 20.

Dataset

Dataset is presented by a parquet file with users' transactions. Path to this file:
`/nfs/shared/clickhouse_data/transactions_12M.parquet`

Dataset sample:

user_id_out	user_id_in	important	amount	datetime
2781	3343	0	199.2	2018-09-02 17:25:12
2789	3343	0	566.33	2018-11-26 11:29:26
2838	3343	0	85.42	2018-09-05 19:59:22
2850	3343	0	850.74	2018-02-19 14:47:41
2860	3343	0	238.35	2018-10-16 00:58:21
2872	3343	0	940.16	2018-09-17 08:24:18
2874	3343	0	308	2018-12-13 11:59:50
2878	3343	0	709.32	2018-11-20 10:35:57
2891	3343	0	121.71	2018-11-27 03:59:52
2939	3343	0	240.06	2018-08-27 20:03:52

Dataset properties:

- ~20% transactions are important (important == 1)
- Total records amount - 12 millions

Task

1. You have to choose **2 or more** MVs. The MVs list is located below.
2. Upload the data into the CH cluster. Table for uploaded data has to be the MergeTree family. To distribute data over the cluster you have to use the Distributed engine and sharding expression.
3. Implement the chosen MVs. Also you are able to create extra tables with different engines if you need them. The number of extra tables should be reasonable.

Requirements

1. All the tables (including MVs) must be sharded. I.e. your database does not have any table that keeps all the data from the cluster. In case of View usage, your view shouldn't request all the data to aggregate it on the initiator server.
2. Table schemas should be reasonable (for example, defining a String column to store digits is a bad idea).
3. MVs return correct results.
4. To get the results, users will be able to use any server of the cluster.

Queries examples that users may request:
SELECT user_id, balance FROM users_balance WHERE balance < 0 LIMIT 20;

*SELECT * FROM avg_out_month WHERE user_id = 12345;*

SELECT user_id, max(trans_sum) FROM transactions_sum WHERE user_id IN (123, 456, 789, ...) GROUP BY user_id;

MVs options

1. Average amount for incoming and outcoming transactions by months and days for each user.
2. The number of important transactions for incoming and outcoming transactions by months and days for each user.
3. The sums for incoming and outcoming transactions by months for each user.
4. Users' saldo for the current moment.

Passing procedure

You have to prepare a short pdf report that contains:

a. Table schemas
b. Chosen sharding expression justification
c. List the chosen MVs and provide their creation queries.

This report should be located in your nfs home directory (on gateway.st: `/nfs/home/<login>/lab-clickhouse.pdf`).

Deadline - 10.06 23:59

PLAGIARISM IS PUNISHABLE UP TO NON-ADMISSION TO THE EXAM! SO WE DO THE TASKS ON OUR OWN!