FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION

OF HIGHER EDUCATION

ITMO UNIVERSITY

Report

on the practical task

Performed by

Daniil Nelubin

J41322c

Accepted by

Teryoshkin Sergey

St. Petersburg

2021

1. What have you done?

It was created distributed tables with corresponding columns and its types getting from the initial JSON’s files. Created tables were filled with data from initial JSON files. A few materialized views were created. It was implemented Kafka buffer.

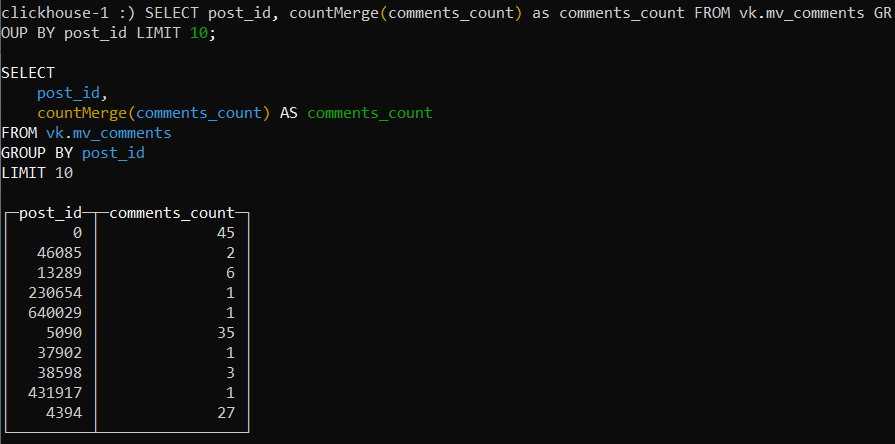
1. How do you sharded the data over the cluster and why did you chose this way?

Using documentation and hash xxHashxx64 for sharding key. Without hash it does not work as well as hashing provides high performance

1. Which MVs have you created?

There were created several materialized views:

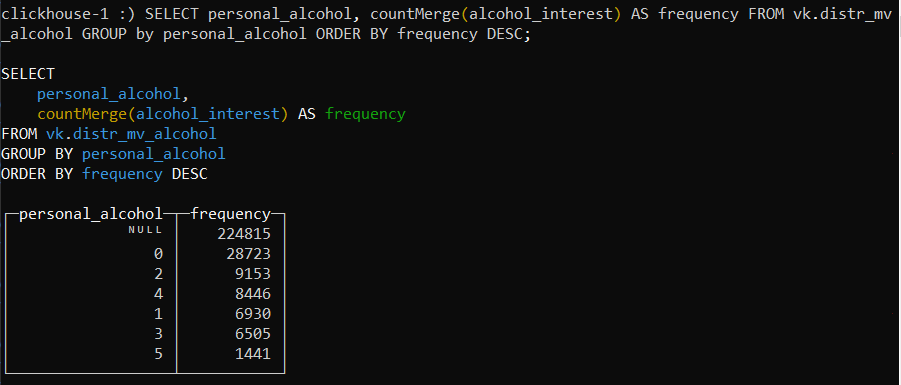
* 1. Top-10 comments count



* 1. The most popular first name in vk



* 1. Distribution of alcohol interest among vk’s users



1. How did you set up the buffer and why this approach?

To read data from a Kafka topic to a ClickHouse table, we need three things:

* A target MergeTree table to provide a home for ingested data
* A Kafka engine table to make the topic look like a ClickHouse table
* A materialized view to move data automatically from Kafka to the target table

SELECT is not particularly useful for reading messages (except for debugging), because each message can be read only once. It is more practical to create real-time threads using materialized views.

