Clickhouse lab

Выполнил: Доржу Начын Шолбанович, J4140, ndorzhu_414205.

1. Средняя сумма входящих и исходящих транзакций по месяцам и дням для каждого пользователя.

По месяцам

1) Создал материализованное представление average_amount_by_month и распределенную таблицу distributed_average_amount_by_month, которая ссылается на представление.

```
clickhouse-6.clickhouse.clickhouse.svc.cluster.local :) DESCRIBE ndorzhu_414205.distributed_average_amount_by_month

DESCRIBE TABLE ndorzhu_414205.distributed_average_amount_by_month

Query id: 2cec4e63-d880-498a-9d7b-97c5b973c7ea

┌─name─────┬─type────┬─default_type─┬─default_expression─┬─comment─┬─codec_expression─┬─ttl_expression─┐
│ user_id  │ Int64   │              │                    │         │                  │                │
│ date     │ String  │              │                    │         │                  │                │
│ avg_in   │ Float64 │              │                    │         │                  │                │
│ avg_out  │ Float64 │              │                    │         │                  │                │
└──────────┴─────────┴──────────────┴────────────────────┴─────────┴──────────────────┴────────────────┘

4 rows in set. Elapsed: 0.003 sec.

clickhouse-6.clickhouse.clickhouse.svc.cluster.local :)
```

```
clickhouse-6.clickhouse.clickhouse.svc.cluster.local :) SELECT * FROM ndorzhu_414205.distributed_average_amount_by_month LI
MIT 10

SELECT *
FROM ndorzhu_414205.distributed_average_amount_by_month
LIMIT 10

Query id: 8c82ce5a-4ef3-4b9c-85bd-89c787a4c273

┌─user_id─┬─date────┬─avg_in─┬─avg_out─┐
│       1 │ 01-2018 │ 512.75 │  470.73 │
│       1 │ 02-2018 │ 459.23 │  489.53 │
│       1 │ 03-2018 │ 503.93 │   470.5 │
│       1 │ 04-2018 │ 490.56 │  532.66 │
│       1 │ 05-2018 │ 494.69 │  498.66 │
│       1 │ 06-2018 │ 476.93 │  535.05 │
│       1 │ 07-2018 │ 500.28 │  476.06 │
│       1 │ 08-2018 │ 522.28 │  463.56 │
│       1 │ 09-2018 │ 476.53 │  519.57 │
│       1 │ 10-2018 │  506.8 │  478.15 │
└─────────┴─────────┴────────┴─────────┘

10 rows in set. Elapsed: 0.016 sec.

clickhouse-6.clickhouse.clickhouse.svc.cluster.local :)
```

- user_id - Идентификатор пользователя.
- Date - Месяц и год в формате MM-YYYY, к которым относятся средние значения транзакций.
- avg_in - Средняя сумма входящих транзакций для данного пользователя в указанный месяц.
- avg_out - Средняя сумма исходящих транзакций для данного пользователя в указанный месяц.

2) Обоснование выбора шардирующего выражения

Поле user_id уникально для каждой записи и часто используется в запросах как основной ключ для агрегации данных. В случае, если система обслуживает миллионы пользователей, распределение данных по user_id позволит равномерно распределять запросы. Например, данные будут параллельно обрабатываться на всех узлах кластера, что значительно ускорит выполнение запросов.

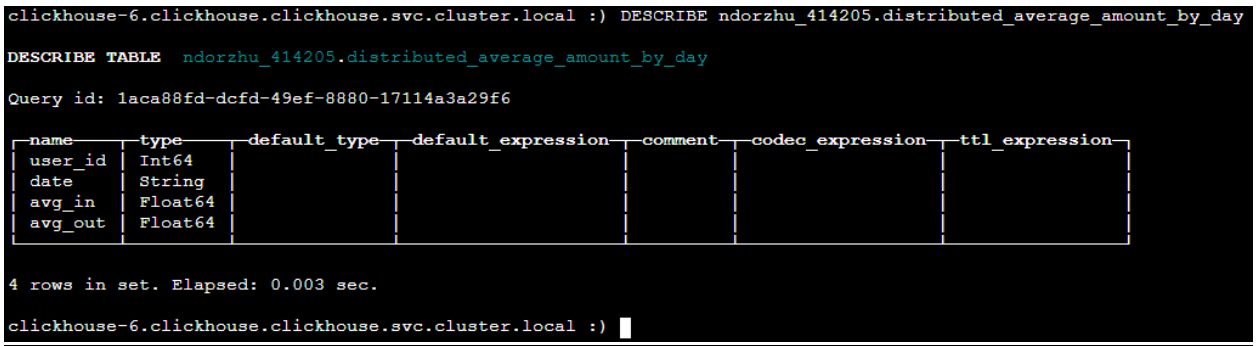
3) Перечислите выбранные MVS и укажите запросы на их создание.

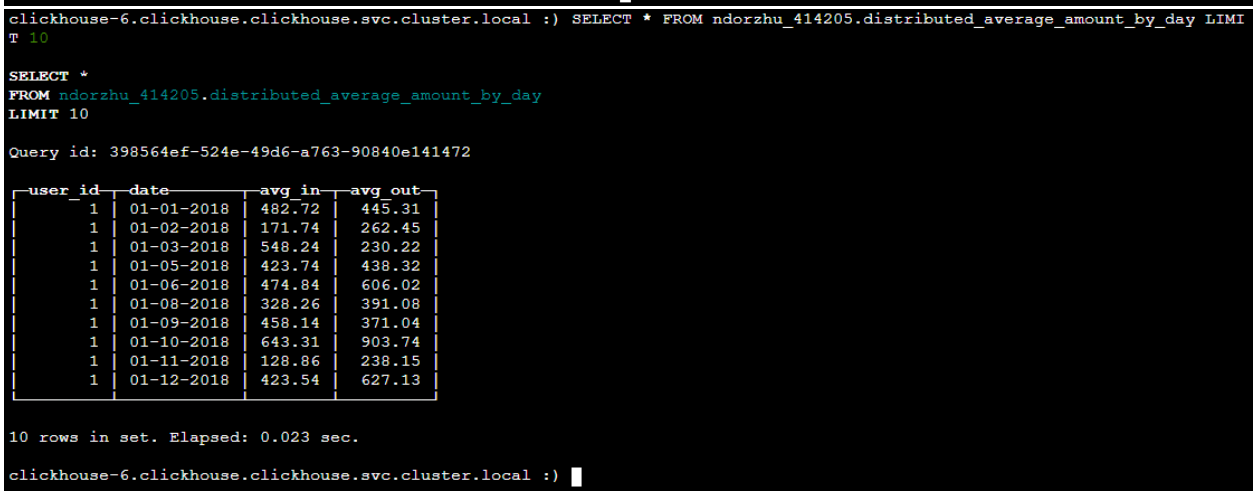
```
CREATE MATERIALIZED VIEW ndorzhu_414205.average_amount_by_month
ON CLUSTER kube_clickhouse_cluster
ENGINE = AggregatingMergeTree
ORDER BY (user_id, date) AS
WITH count_in AS (
   SELECT
      user_id_in AS user_id,
      formatDateTime(datetime, '%m-%G') AS date,
      ROUND(AVG(amount), 2) AS avg_in
   FROM ndorzhu_414205.distributed_users_transactions
   GROUP BY
      user_id,
      date
),
count_out AS (
   SELECT
      user_id_out AS user_id,
      formatDateTime(datetime, '%m-%G') AS date,
      ROUND(AVG(amount), 2) AS avg_out
   FROM ndorzhu_414205.distributed_users_transactions
   GROUP BY
      user_id,
      date
)
SELECT
   tci.user_id AS user_id,
   tci.date,
   avg_in,
   avg_out
FROM count_in AS tci
INNER JOIN count_out AS tco ON (tco.user_id = tci.user_id) AND (tco.date = tci.date)
ORDER BY
   tci.user_id,
   date;

CREATE TABLE ndorzhu_414205.distributed_average_amount_by_month
ON CLUSTER kube_clickhouse_cluster AS ndorzhu_414205.average_amount_by_month
ENGINE = Distributed(kube_clickhouse_cluster, ndorzhu_414205, average_amount_by_month);
```

По дням

1) Создал материализованное представление average_amount_by_day и распределенную таблицу distributed_average_amount_by_day, которая ссылается на представление.





- user_id - Идентификатор пользователя.
- Date - Дата, за которую рассчитывается средняя сумма транзакций.
- avg_in - Средняя сумма входящих транзакций пользователя за указанный день.
- avg_out - Средняя сумма исходящих транзакций пользователя за указанный день.

2) Обоснование выбора шардирующего выражения

Использование user_id дает равномерное распределение нагрузки между узлами кластера, что помогает предотвратить перегрузку отдельных узлов. Такой подход облегчает масштабирование системы путем добавления новых узлов кластера без значительной переработки данных.

3) Перечислите выбранные MVS и укажите запросы на их создание.

```
CREATE MATERIALIZED VIEW ndorzhu_414205.average_amount_by_day
ON CLUSTER kube_clickhouse_cluster
ENGINE = AggregatingMergeTree
ORDER BY (user_id, date) AS
WITH count_in AS (
  SELECT
    user_id_in AS user_id,
    formatDateTime(datetime, '%d-%m-%G') AS date,
    ROUND(AVG(amount), 2) AS avg_in
  FROM ndorzhu_414205.distributed_users_transactions
  GROUP BY
    user_id,
    date
),
count_out AS (
  SELECT
    user_id_out AS user_id,
```

```
    formatDateTime(datetime, '%d-%m-%G') AS date,
    ROUND(AVG(amount), 2) AS avg_out
  FROM ndorzhu_414205.distributed_users_transactions
  GROUP BY
    user_id,
    date
)
SELECT
  tci.user_id AS user_id,
  tci.date,
  avg_in,
  avg_out
FROM count_in AS tci
INNER JOIN count_out AS tco ON (tco.user_id = tci.user_id) AND (tco.date = tci.date)
ORDER BY
  tci.user_id,
  date;

CREATE TABLE ndorzhu_414205.distributed_average_amount_by_day
ON CLUSTER kube_clickhouse_cluster AS ndorzhu_414205.average_amount_by_day
ENGINE = Distributed(kube_clickhouse_cluster, ndorzhu_414205, average_amount_by_day);
```

2. Количество важных транзакций для входящих и исходящих транзакций по месяцам и дням для каждого пользователя.

По месяцам

1) Создал материализованное представление important_number_by_month и распределенную таблицу distributed_important_number_by_month, которая ссылается на представление.

```
clickhouse-6.clickhouse.clickhouse.svc.cluster.local :) DESCRIBE ndorzhu_414205.distributed_important_number_by_month

DESCRIBE TABLE  ndorzhu_414205.distributed_important_number_by_month

Query id: eb4098ca-0b9e-4d03-b4cd-f4a398e74d2a

┌─name─────┬─type───┬─default_type─┬─default_expression─┬─comment─┬─codec_expression─┬─ttl_expression─┐
│ user_id  │ Int64  │              │                    │         │                  │                │
│ date     │ String │              │                    │         │                  │                │
│ cnt_in   │ UInt64 │              │                    │         │                  │                │
│ cnt_out  │ UInt64 │              │                    │         │                  │                │
└──────────┴────────┴──────────────┴────────────────────┴─────────┴──────────────────┴────────────────┘

4 rows in set. Elapsed: 0.003 sec.

clickhouse-6.clickhouse.clickhouse.svc.cluster.local :)
```

```
clickhouse-6.clickhouse.clickhouse.svc.cluster.local :) SELECT * FROM ndorzhu_414205.distributed_important_number_by_month
LIMIT 10

SELECT *
FROM ndorzhu_414205.distributed_important_number_by_month
LIMIT 10

Query id: 0f844945-8d53-42bf-bf0d-f8d944a1a160

┌─user_id─┬─date────┬─cnt_in─┬─cnt_out─┐
│       1 │ 01-2018 │     40 │      40 │
│       1 │ 02-2018 │     66 │      34 │
│       1 │ 03-2018 │     18 │      28 │
│       1 │ 04-2018 │     62 │      38 │
│       1 │ 05-2018 │     52 │      36 │
│       1 │ 06-2018 │     42 │      38 │
│       1 │ 07-2018 │     34 │      52 │
│       1 │ 08-2018 │     24 │      26 │
│       1 │ 09-2018 │     44 │      38 │
│       1 │ 10-2018 │     40 │      52 │
└─────────┴─────────┴────────┴─────────┘

10 rows in set. Elapsed: 0.016 sec.

clickhouse-6.clickhouse.clickhouse.svc.cluster.local :)
```

- user_id - Идентификатор пользователя.

- date - Месяц и год, за который рассчитывается количество транзакций важного типа.
- cnt_in - Количество важных входящих транзакций для пользователя за указанный месяц.
- cnt_out - Количество важных исходящих транзакций для пользователя за указанный месяц.

2) Обоснование выбора шардирующего выражения

Поле user_id идентифицирует каждого пользователя, а также часто используется в запросах как основной ключ для агрегации данных. Использование user_id обеспечивает равномерное распределение данных по кластерам. В случае обработки данных о транзакциях с учетом их важности, распределение данных по user_id обеспечит эффективное выполнение запросов и анализа важных транзакций пользователей в различные периоды времени.

3) Перечислите выбранные MVS и укажите запросы на их создание.

```
CREATE MATERIALIZED VIEW ndorzhu_414205.important_number_by_month
ON CLUSTER kube_clickhouse_cluster
ENGINE = AggregatingMergeTree
ORDER BY (user_id, date) AS
WITH count_in AS (
  SELECT
    user_id_in AS user_id,
    formatDateTime(datetime, '%m-%G') AS date,
    COUNT(amount) AS cnt_in
  FROM ndorzhu_414205.distributed_users_transactions
  WHERE important = 1
  GROUP BY
    user_id,
    date
),
count_out AS (
  SELECT
    user_id_out AS user_id,
    formatDateTime(datetime, '%m-%G') AS date,
    COUNT(amount) AS cnt_out
  FROM ndorzhu_414205.distributed_users_transactions
  WHERE important = 1
  GROUP BY
    user_id,
    date
)
SELECT
  tci.user_id AS user_id,
  tci.date AS date,
  cnt_in,
  cnt_out
FROM count_in AS tci
INNER JOIN count_out AS tco ON (tco.user_id = tci.user_id) AND (tco.date = tci.date)
ORDER BY
  date;

CREATE TABLE ndorzhu_414205.distributed_important_number_by_month
ON CLUSTER kube_clickhouse_cluster AS ndorzhu_414205.important_number_by_month
ENGINE = Distributed(kube_clickhouse_cluster, ndorzhu_414205, important_number_by_month);
```

По дням

1) Создал материализованное представление important_number_by_day и распределенную таблицу distributed_important_number_by_day, которая ссылается на представление.

```
clickhouse-6.clickhouse.clickhouse.svc.cluster.local :) DESCRIBE ndorzhu_414205.distributed_important_number_by_day

DESCRIBE TABLE ndorzhu_414205.distributed_important_number_by_day

Query id: 5f86c06b-c614-42e2-a8df-aa81729647eb

┌─name─────┬─type───┬─default_type─┬─default_expression─┬─comment─┬─codec_expression─┬─ttl_expression─┐
│ user_id  │ Int64  │              │                    │         │                  │                │
│ date     │ String │              │                    │         │                  │                │
│ cnt_in   │ UInt64 │              │                    │         │                  │                │
│ cnt_out  │ UInt64 │              │                    │         │                  │                │
└──────────┴────────┴──────────────┴────────────────────┴─────────┴──────────────────┴────────────────┘

4 rows in set. Elapsed: 0.003 sec.

clickhouse-6.clickhouse.clickhouse.svc.cluster.local :)
```

```
clickhouse-6.clickhouse.clickhouse.svc.cluster.local :) SELECT * FROM ndorzhu_414205.distributed_important_number_by_day LI
MIT 10

SELECT *
FROM ndorzhu_414205.distributed_important_number_by_day
LIMIT 10

Query id: 8a2424e5-4440-4a94-aba9-9da669cf0bbf

┌─user_id─┬─date───────┬─cnt_in─┬─cnt_out─┐
│       1 │ 01-01-2018 │      2 │       4 │
│       1 │ 01-02-2018 │      2 │       2 │
│       1 │ 02-02-2018 │     10 │       2 │
│       1 │ 02-04-2018 │      2 │       2 │
│       1 │ 02-06-2018 │      2 │       2 │
│       1 │ 03-08-2018 │      2 │       2 │
│       1 │ 03-09-2018 │      2 │       2 │
│       1 │ 03-11-2018 │      4 │       2 │
│       1 │ 03-12-2018 │      2 │       6 │
│       1 │ 04-01-2018 │      2 │       2 │
└─────────┴────────────┴────────┴─────────┘

10 rows in set. Elapsed: 0.016 sec.

clickhouse-6.clickhouse.clickhouse.svc.cluster.local :)
```

- user_id - Идентификатор пользователя.
- date - Дата, за которую рассчитывается количество важных транзакций.
- cnt_in - Количество важных входящих транзакций для пользователя за указанный день.
- cnt_out - Количество важных исходящих транзакций для пользователя за указанный день.

2) Обоснование выбора шардирующего выражения

Выбор user_id в качестве шардирующего выражения для создания материализованного представления и распределенной таблицы был логичным, с учетом производительности, масштабируемости и балансировки нагрузки. Этот подход позволяет эффективно обрабатывать данные и выполнять запросы в системах с высокой производительностью и масштабируемостью.

3) Перечислите выбранные MVS и укажите запросы на их создание.

```
CREATE MATERIALIZED VIEW ndorzhu_414205.important_number_by_day
ON CLUSTER kube_clickhouse_cluster
ENGINE = AggregatingMergeTree
ORDER BY (user_id, date) AS
WITH count_in AS (
  SELECT
    user_id_in AS user_id,
    formatDateTime(datetime, '%d-%m-%G') AS date,
```

```
    COUNT(amount) AS cnt_in
  FROM ndorzhu_414205.distributed_users_transactions
  WHERE important = 1
  GROUP BY
    user_id,
    date
),
count_out AS (
  SELECT
    user_id_out AS user_id,
    formatDateTime(datetime, '%d-%m-%G') AS date,
    COUNT(amount) AS cnt_out
  FROM ndorzhu_414205.distributed_users_transactions
  WHERE important = 1
  GROUP BY
    user_id,
    date
)
SELECT
  tci.user_id AS user_id,
  tci.date AS date,
  cnt_in,
  cnt_out
FROM count_in AS tci
INNER JOIN count_out AS tco ON (tco.user_id = tci.user_id) AND (tco.date = tci.date)
ORDER BY
  date;

CREATE TABLE ndorzhu_414205.distributed_important_number_by_day
ON CLUSTER kube_clickhouse_cluster AS ndorzhu_414205.important_number_by_day
ENGINE = Distributed(kube_clickhouse_cluster, ndorzhu_414205, important_number_by_day);
```