ClickHouse Mark Cache

5 minute primer

Link to slides ^^ (will also show at the end)

How ClickHouse stores data (MergeTree engine)

- Data in MergeTree table is stored in "parts" (each INSERT produces a part)
- Parts are continuously merged
- Each part stores data in primary key order
- Each part has its own directory
- Within that directory each column has its own "bin" and "mrk" files
  - `default/my_table/20200202_5042_6949_1107/userId.bin`
  - `default/my_table/20200202_5042_6949_1107/userId.mrk`
  - …
  - `default/my_table/20200202_5042_6949_1107/timestamp.bin`
  - `default/my_table/20200202_5042_6949_1107/timestamp.mrk`

How ClickHouse finds data

- Each part has its own "primary.idx" file (one per directory)
- Index stores value of the primary key for each N-th row (sparse)
- Mark files store location (byte offset) of N-th row in each column file
- Entire range with `index_granularity` rows must be read for each key
- The unit at which data is retrieved from part files is a block
    - Blocks usually aligned with marks
    - Compression applied at block level

SELECT Performance Considerations

- Primary index always in RAM
- Marks in LRU cache configured with `mark_cache_size` server setting
  - Default 5GiB
- Must be populated before accessing data
- Total size of marks:
  - Grows with each new column
  - Inverse of `index_granularity`
- Keeping marks cache hot increases SELECT performance

Napkin Math

- For our data sets marks are ~0.1% of data bytes
- For 70TB of data that is 70GB of marks
- Look at sizes or "mrk" files or check the `system.parts.mark_bytes`
- Try dropping the cache to see its impact:
    - `clickhouse :) SYSTEM DROP MARK CACHE ;`

Thanks!

ClickHouse documentation is fantastic:
- https://clickhouse.tech/docs/en/development/architecture/#merge-tree
- https://clickhouse.tech/docs/en/operations/table_engines/mergetree/#mergetree-data-storage
- https://clickhouse.tech/docs/en/operations/system_tables/#system_tables-parts
- https://clickhouse.tech/docs/en/operations/server_settings/settings/#server-mark-cache-size
- https://clickhouse.tech/docs/en/query_language/system/#query_language-system-drop-mark-cache

Google slides ^^