DigiClips Database Analysis

Pavan Manjunath

Implementation Plan: Data Retention & Optimization

Goal Description

The primary goal is to optimize the dc database by enforcing strict data retention policies (deleting records > 90 days), improving search relevance and performance, streamlining database backups, and enhancing data integrity through validation.

Info

TvVTT and RadioClips store timestamps as VARCHAR. To efficiently delete records older than 90 days, these MUST be converted to DATETIME. This is a potentially blocking operation for large tables. Use of STR_TO_DATE in delete queries is possible but performs poorly.

Proposed Changes

Phase 1: Data Retention & Cleanup (Priority)

We will focus on automating the cleanup of old data to maintain database performance and compliance.

1. Schema Updates for Date Handling

- Target Initial Tables(later on extend to others after verification) : TvVTT, RadioClips, RadioClipsB.
- Action:
    1. Add a new TStamp_dt DATETIME column.
    2. Populate it using STR_TO_DATE(TStamp, format).
    3. Rename/Swap to replace the old column. Note: If table size is massive, we may need to do this in batches.(so this is optional)

2. Logging Infrastructure

- New Table: cleanup_logs (or enhance existing cleanup_log).
    - Columns: id, table_name, deleted_count, execution_time, status, timestamp.

3. Stored Procedures for Cleanup

- Optimize/Create: Cleanup_SRT_OldRecords and Cleanup_Radio_OldRecords.

- Logic:
    - Delete simple records > 90 days.
    - Log the number of deleted rows to cleanup_logs.
    - Handle SRT, SRT21, TvVTT, RadioClips.

4. Automation(to be reviewed)

- Create MySQL Event Daily_Data_Purge to call the cleanup procedures every 24 hours.

Phase 2: Database Dump Management(can be modified acc post Phase 1 results)
Improve the backup process to be more robust and organized.

- Structure: Create specific directories
- Split Dumps:
    - Schema Dump: Full structure of all tables (including massive ones like SRT).
    - Data Dump (Config/Ref): Data for smaller tables (Stations, Users, Config).
    - Exclusion: Do NOT dump data for SRT or RadioClips tables (too large).

Phase 3: Search Optimization (based on results from indexing)
Improve the quality and speed of search results.

- Relevance Ranking: Modify Bool_Search (or creating Relevant_Search) to order results by:
    1. Match frequency (count of hits).
    2. Recency (newer items first).
    3. Category weighting (if applicable).
- Indexes: Review execution plans for search queries and add specific secondary indexes where full-text search is insufficient.

Phase 4: Data Validation

- Validation Procedures:
    - Update Insert procedures to validate Email format (regex).

    - Validate State codes against the State reference table.
- Error Logging:
    - Create Validation_Errors table to capture rejected inserts/updates for review.

Verification Plan

Automated Tests

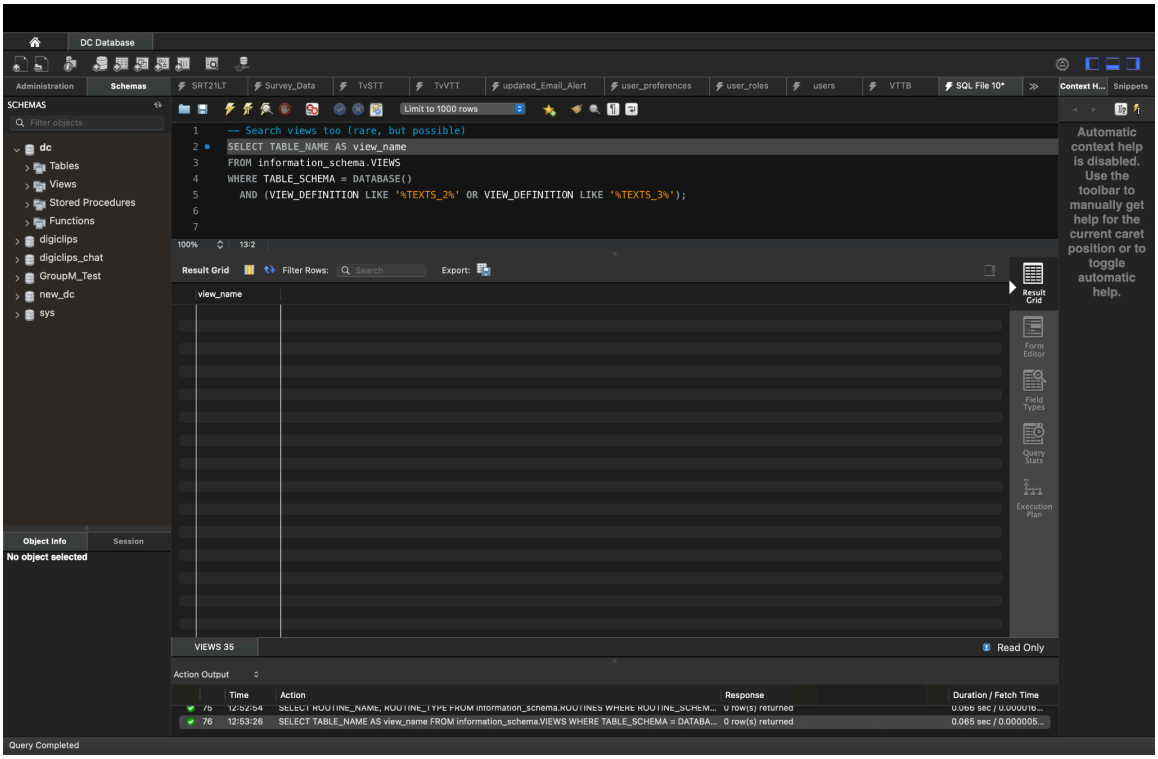
- Retention Test: Insert dummy records dated 91 days ago, run the procedure, verify they are gone and logged.
- Validation Test: Attempt to insert invalid emails/states and verify rejection/logging.

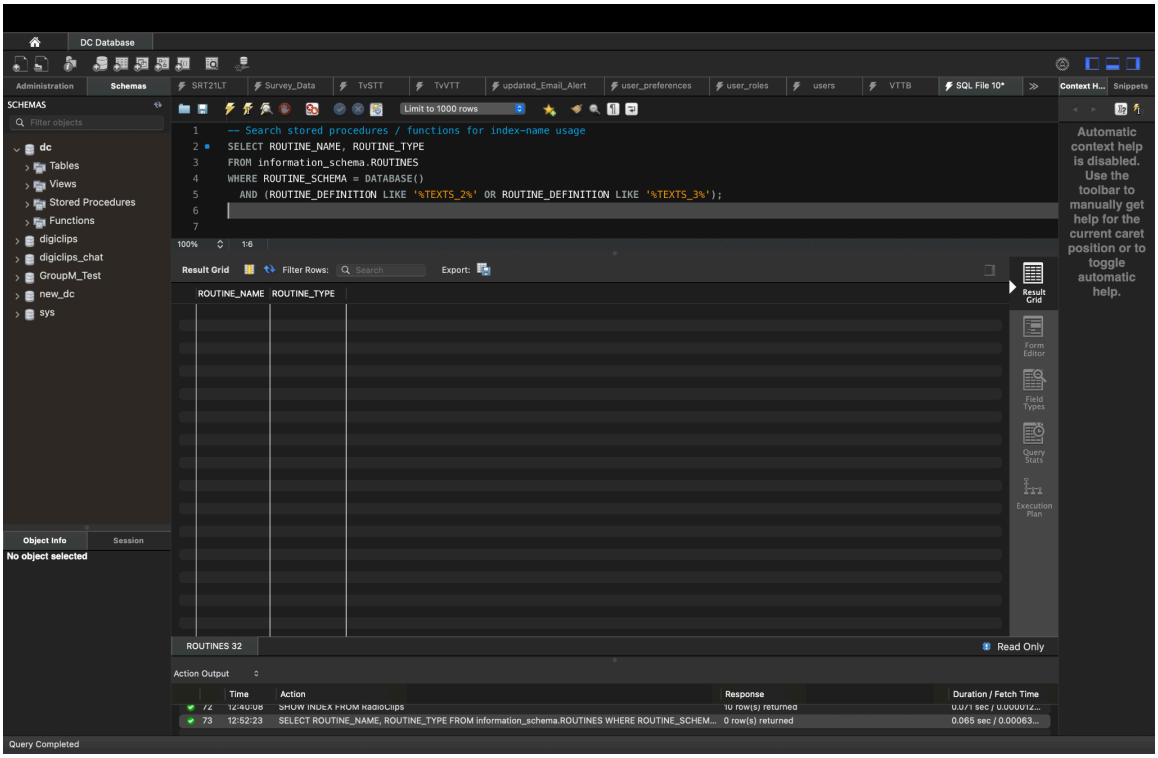
Manual Verification

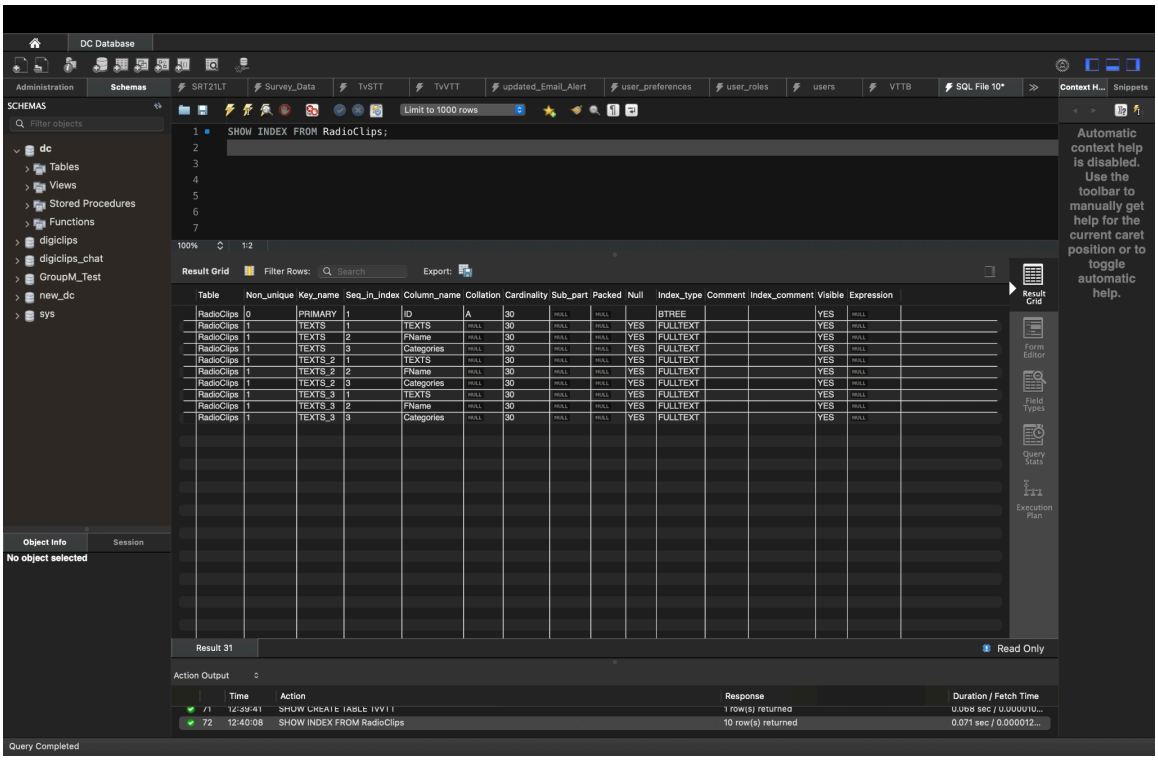
- Search Quality: Run sample queries before and after optimization to compare result relevance.

All the additional tables to be added can have different table names, just mentioned them for an example.

Queries:




DC Database
Administration
Schemas
SCHEMAS
Filter objects
dc
Tables
Views
Stored Procedures
Functions
digiclips
digiclips_chat
GroupM_Test
new_dc
sys
Object Info
Session
No object selected
SRT21LT
Survey_Data
TvSTT
TvVTT
updated_Email_Alert
user_preferences
user_roles
users
VTTB
SQL File 10*
Context H...
Snippets
Limit to 1000 rows
1 -- Search stored procedures / functions for index-name usage
2 SELECT ROUTINE_NAME, ROUTINE_TYPE
3 FROM information_schema.ROUTINES
4 WHERE ROUTINE_SCHEMA = DATABASE()
5 AND (ROUTINE_DEFINITION LIKE '%TEXTS_2%' OR ROUTINE_DEFINITION LIKE '%TEXTS_3%');
6
7
100%
1:6
Result Grid
Filter Rows:
Search
Export:
ROUTINE_NAME
ROUTINE_TYPE
Result Grid
Form Editor
Field Types
Query Stats
Execution Plan
ROUTINES 32
Read Only
Action Output
Time
Action
Response
Duration / Fetch Time
72 12:40:08 SHOW INDEX FROM RadioClips 10 row(s) returned 0.071 sec / 0.000012...
73 12:52:23 SELECT ROUTINE_NAME, ROUTINE_TYPE FROM information_schema.ROUTINES WHERE ROUTINE_SCHEM... 0 row(s) returned 0.065 sec / 0.00063...
Query Completed
Automatic context help is disabled. Use the toolbar to manually get help for the current caret position or to toggle automatic help.


DC Database
Administration
Schemas
SCHEMAS
Filter objects
dc
Tables
Views
Stored Procedures
Functions
digiclips
digiclips_chat
GroupM_Test
new_dc
sys
Object Info
Session
No object selected
SRT21LT
Survey_Data
TvSTT
TvVTT
updated_Email_Alert
user_preferences
user_roles
users
VTTB
SQL File 10*
Context H...
Snippets
Limit to 1000 rows
1 SHOW INDEX FROM RadioClips;
2
3
4
5
6
7
100%
1:2
Result Grid
Filter Rows:
Search
Export:
Table Non_unique Key_name Seq_in_index Column_name Collation Cardinality Sub_part Packed Null Index_type Comment Index_comment Visible Expression
RadioClips 0 PRIMARY 1 ID A 30 BTREE YES
RadioClips 1 TEXTS 1 TEXTS 30 YES FULLTEXT YES
RadioClips 1 TEXTS 2 FName 30 YES FULLTEXT YES
RadioClips 1 TEXTS 3 Categories 30 YES FULLTEXT YES
RadioClips 1 TEXTS_2 1 TEXTS 30 YES FULLTEXT YES
RadioClips 1 TEXTS_2 2 FName 30 YES FULLTEXT YES
RadioClips 1 TEXTS_2 3 Categories 30 YES FULLTEXT YES
RadioClips 1 TEXTS_3 1 TEXTS 30 YES FULLTEXT YES
RadioClips 1 TEXTS_3 2 FName 30 YES FULLTEXT YES
RadioClips 1 TEXTS_3 3 Categories 30 YES FULLTEXT YES
Result Grid
Form Editor
Field Types
Query Stats
Execution Plan
Result 31
Read Only
Action Output
Time
Action
Response
Duration / Fetch Time
72 12:40:08 SHOW INDEX FROM RadioClips 10 row(s) returned 0.071 sec / 0.000012...
Query Completed
Automatic context help is disabled. Use the toolbar to manually get help for the current caret position or to toggle automatic help.