Reporting Tables Optimization

Sprint 1 Report

1. Overview:

The NewspaperResults and MagazineResults tables form the core of DigiClips' reporting layer. They store publication-level results and media analytics that feed dashboards and summary reports. The purpose of this sprint was to validate schema design, review query patterns, and ensure proper indexing to optimize reporting performance.

2. Schema Review Summary

a. NewspaperResults: Column name → Data Type
   - ID → INT
   - Title → VARCHAR(1000)
   - Author → VARCHAR(100)
   - Summary → VARCHAR(5000)
   - PublishDate → DATETIME
   - NewspaperLink → VARCHAR(1000)
   - ImageURL → VARCHAR(1000)
   - UpdateDate → DATETIME
   - GUID → VARCHAR(1000)
   - AddedDate → DATETIME

b. MagazineResults: Column name → Data Type
   - ID → INT
   - Title → VARCHAR(1000)
   - Author → VARCHAR(100)
   - Summary → VARCHAR(5000)
   - PublishDate → DATETIME
   - MagazinerLink → VARCHAR(1000)
   - ImageURL → VARCHAR(1000)
   - UpdateDate → DATETIME
   - GUID → VARCHAR(1000)
   - AddedDate → DATETIME

Notes: Non-nullables include ID, Title, Author, NewspaperLink/MagazineLink, GUID

3. Index Review and Verification

Name → Type: Purpose

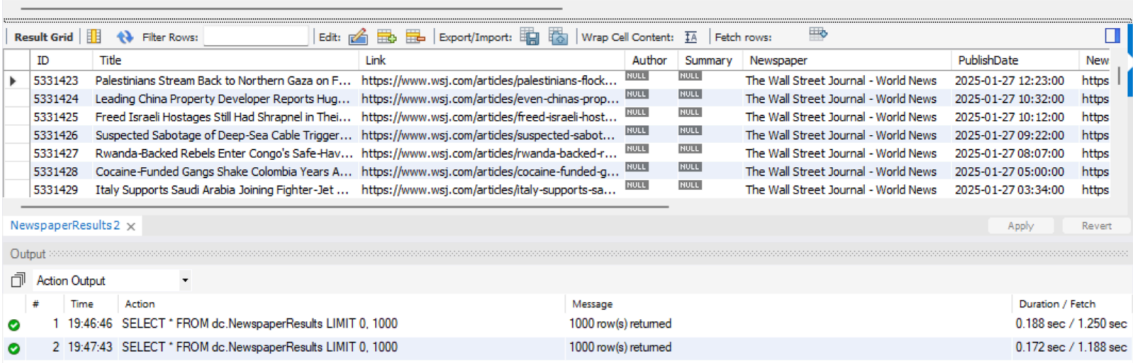
PRIMARY → **PRIMARY:** On the ID column, ensures unique record identification
GUID_UNIQUE → **UNIQUE:** On the GUID column. Avoids duplicate ingestion
idx_Newspaper_Results_Summary → **FULLTEXT:** Enables full-text search within article summaries
idx_MagazineResults → **FULLTEXT:** Enables full-text search within article summaries

4. Connected Tables:

Both NewspaperResults and MagazineResults are not connected to any tables. There are no foreign keys.

5. Test Results:
    - Scenario 1: Data Query Speed: Returned 1000 rows in 0.172 seconds

```
SELECT * FROM dc.NewspaperResults;
SELECT * FROM NewspaperResults WHERE date BETWEEN '2024-01-01' AND '2024-12-31';
```



Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content: | Fetch rows:

ID	Title	Link	Author	Summary	Newspaper	PublishDate	New
5331423	Palestinians Stream Back to Northern Gaza on F…	https://www.wsj.com/articles/palestinians-flock…	NULL	NULL	The Wall Street Journal - World News	2025-01-27 12:23:00	https
5331424	Leading China Property Developer Reports Hug…	https://www.wsj.com/articles/even-chinas-prop…	NULL	NULL	The Wall Street Journal - World News	2025-01-27 10:32:00	https
5331425	Freed Israeli Hostages Still Had Shrapnel in Thei…	https://www.wsj.com/articles/freed-israeli-host…	NULL	NULL	The Wall Street Journal - World News	2025-01-27 10:12:00	https
5331426	Suspected Sabotage of Deep-Sea Cable Trigger…	https://www.wsj.com/articles/suspected-sabot…	NULL	NULL	The Wall Street Journal - World News	2025-01-27 09:22:00	https
5331427	Rwanda-Backed Rebels Enter Congo's Safe-Hav…	https://www.wsj.com/articles/rwanda-backed-r…	NULL	NULL	The Wall Street Journal - World News	2025-01-27 08:07:00	https
5331428	Cocaine-Funded Gangs Shake Colombia Years A…	https://www.wsj.com/articles/cocaine-funded-g…	NULL	NULL	The Wall Street Journal - World News	2025-01-27 05:00:00	https
5331429	Italy Supports Saudi Arabia Joining Fighter-Jet …	https://www.wsj.com/articles/italy-supports-sa…	NULL	NULL	The Wall Street Journal - World News	2025-01-27 03:34:00	https

NewspaperResults2 | Apply | Revert

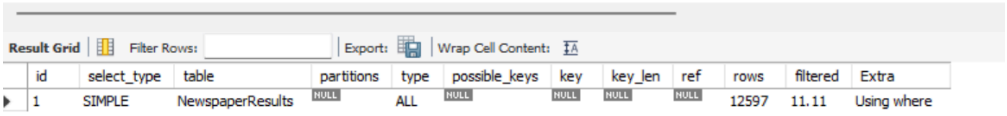
Output

Action Output

#	Time	Action	Message	Duration / Fetch
1	19:46:46	SELECT * FROM dc.NewspaperResults LIMIT 0, 1000	1000 row(s) returned	0.188 sec / 1.250 sec
2	19:47:43	SELECT * FROM dc.NewspaperResults LIMIT 0, 1000	1000 row(s) returned	0.172 sec / 1.188 sec

- Scenario 2 Sentiment Lookup: There is no sentiment filer in both of these tables
- Scenario 3 Index Verification: For both MagazineResults and NewspaperResult, there is no index at PublishDate.

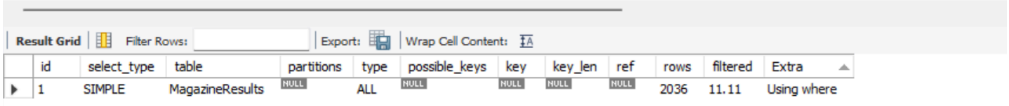
```
SELECT * FROM dc.NewspaperResults;
EXPLAIN SELECT *
FROM dc.NewspaperResults
WHERE PublishDate BETWEEN '2024-01-01' AND '2024-12-31';
```



Result Grid | Filter Rows: | Export: | Wrap Cell Content:

id	select_type	table	partitions	type	possible_keys	key	key_len	ref	rows	filtered	Extra
1	SIMPLE	NewspaperResults	NULL	ALL	NULL	NULL	NULL	NULL	12597	11.11	Using where

```
SELECT * FROM dc.MagazineResults;
EXPLAIN SELECT *
FROM dc.MagazineResults
WHERE PublishDate BETWEEN '2024-01-01' AND '2024-12-31';
```



Result Grid | Filter Rows: | Export: | Wrap Cell Content:

id	select_type	table	partitions	type	possible_keys	key	key_len	ref	rows	filtered	Extra
1	SIMPLE	MagazineResults	NULL	ALL	NULL	NULL	NULL	NULL	2036	11.11	Using where

6. Key Findings

a. Schema Consistency:

The two tables share nearly identical structures, which supports unified reporting but also introduces redundancy that could be normalized in future releases.

b. Data Type Validation:

All columns use appropriate types for their data. However, VARCHAR(1000) for GUIDs and links may be over allocated, optimization to VARCHAR(255) would suffice.

c. Index Coverage:
   Full-text indexes exist on Summary, useful for keyword search.

7. Recommendations

- Add index on PublishDate. Currently, The EXPLAIN output shows a full table scan (type = ALL) with no index being used. This indicates that MySQL is scanning all ~2,000+ rows even though the query filters on PublishDate.
- The current FULLTEXT indexes (idx_Newspaper_Results_Summary, idx_MagazineResults) enable text search over Summary, which is beneficial for keyword-based reporting. However, FULLTEXT indexes can become large and should be monitored for performance overhead during inserts/updates.
- Keep GUID_UNIQUE index as it ensures data integrity.