Big Data Processing and Analysis MLDS

Διδάσκων: Μίνως Γαροφαλάκης

Τελική Αναφορά Project

Νεαμονιτάκη Ζωγραφούλα Ιωάννα 2020030088

25 Αυγούστου 2025

1 Introduction

Σε αυτό το project μελετήθηκε η απόδοση ερωτημάτων SQL σε σχέση με την τεχνική σύνοψης reservoir sampling. Το dataset είναι από το database "Discogs" το οποίο περιέχει πληροφορίες για κυκλοφορίες κομματιών σε βινύλια, κασέτες κλπ. καθώς και πληροφορίες για τους καλλιτέχνες. Μελετήθηκαν queries στο dataset "releases" αλλά και join queries με το dataset "artists". Στόχος αυτού του project είναι ο έλεγχος απόδοσης του reservoir sampling με διαφορετικό δείγμα (1% και 10%).

2 Περιβάλλον υλοποίησης

Η ανάλυση πραγματοποιήθηκε με χρήση Apache Flink 2.0.0 και του PyFlink Table API. Η εκτέλεση έγινε σε batch mode, με Python runtime ορισμένο στο /home/fneon/flink-2.0.0/bin/venv/bin/python. Για την επεξεργασία των δεδομένων φορτώθηκαν αρχεία CSV σε προσωρινά tables, τα οποία ορίστηκαν μέσω SQL DDL με τον connector filesystem.

3 Dataset And Queries

Το dataset βρέθηκε στην πλατφόρμα kaggle.com. Το Discogs είναι ένα online database και marketplace για ηχογραφήσεις, συμπεριλαμβανομένων εμπορικών κυκλοφοριών, διαφημιστικών εκδόσεων και εκδόσεων bootleg ή off-label. Το project έγινε πάνω σε δύο αρχεία CSV, το releases.csv που περιλαμβάνει όλες τις απαραίτητες πληροφορίες για κάθε κυκλοφορία (50 πεδία) και το αρχείο artists.csv που περιέχει πληροφορίες για τους καλλιτέχνες (11 πεδία). Τα πιο σημαντικά πεδία που χρησιμοποιήθηκαν φαίνονται παρακάτω:
Releases:

- `artists_artist_id`= key id του καλλιτέχνη της κυκλοφορίας.
- `artists_artist_name` = το όνομα του καλλιτέχνη της κυκλοφορίας.
- `release_formats_format_name` = τι format είναι κυκλοφορία, δηλαδή αν είναι cd, vinyl κλπ.
- `release_genres_genre` = τι είδος μουσικής είναι η κυκλοφορία.
- `release_styles_style` = τι μουσικό στυλ έχει η κυκλοφορία.
- `releases_release_country`= η χώρα κυκλοφορίας του release.
- `releases_release_id` = key id της κυκλοφορίας.
- `releases_release_released` = η ημερονία κυκλοφορίας του release.
- `releases_release_title` = ο τίτλος της κυκλοφορίας.
- `sub_tracks_track_duration` = η διάρκεια του κάθε subtrack της κυκλοφορίας.

Artists:

- `artist_groups_name` = μπάντες ή γκρουπς όπου ο καλλιτέχνης μπορεί να έχει συμμετάσχει.
- `artist_members_name` = αν ο artist είναι μπάντα, αναφέρεται στα μέλη της μπάντας.
- `artists_artist_id` = key id του καλλιτέχνη (σε αυτό το κλειδί πάνω γίνεται το join) .

Για να γίνει αυτή η μελέτη επιλέχθηκε ένα σετ από queries έτσι ώστε να συγκριθούν διαφορετικά είδη ερωτημάτων, όπως average, group by και group by με where clause. Παρακάτω αναλύονται τα queries για το releases dataset:
Q1. Top 5 Formats
Q2. Top 10 Styles
Q3. Top 20 Artists with the most releases
Q4. Metal Artists with most releases
Q5. Artists with most genres
Q6. Top 10 Artists in Greece with most releases
Q7. Average number of tracks in all releases
Q8. Average number of releases per year and average number of releases in year 1999

Joins Queries:
Q9. Καλλιτέχνες σε σχέση με το genre τους
Q10. Καλλιτέχνες σε σχέση με την χώρα τους
Q11. Top 10 artists with the most releases and the number of bands they have participated in.

3.1 Queries In Dataset "releases.csv"

Παρακάτω εξηγούνται λίγο πιο αναλυτικά τα queries:

Q1. Στόχος: Να βρεθούν τα 5 πιο συχνά format κυκλοφοριών. Αυτό το query υλοποιήθηκε για τον έλεγχο ορθότητας των ερωτημάτων καθώς είναι γνωστό ότι οι περισσότερες κυκλοφορίες στο discogs είναι βινύλια, cd, και μετά κασέτες.

- Φιλτραρίστηκαν εκτός τα πεδία όπου το `releases_release_country` ήταν null ή κενό.
- Το dataset ομαδοποιήθηκε με βάση το `release_formats_format_name`.
- Υπολογίστηκε το πλήθος κυκλοφοριών για κάθε format.
- Τα αποτελέσματα ταξινομήθηκαν κατά φθίνουσα σειρά.
- Επιλέχθηκαν οι top 5 formats.

Q2. Στόχος: Να βρεθούν τα 10 πιο συχνά μουσικά στυλ στο dataset.

- Φιλτραρίστηκαν εκτός τα πεδία όπου το `release_styles_style` ήταν null ή κενά.
- Τα στυλ εντοπίστηκαν χρησιμοποιώντας μια UDF συνάρτηση (`has_style`). Η συνάρτηση αυτή ελέγχει αν το στυλ υπάρχει μέσα σε ένα JSON array ή σε string. Αν βρεθεί το στυλ, επιστρέφει True.
- Το dataset ομαδοποιήθηκε με βάση το πεδίο `release_styles_style`.
- Υπολογίστηκε το πλήΘος των κυκλοφοριών.
- Τα αποτελέσματα ταξινομήθηκαν κατά φθίνουσα σειρά.
- Επιλέχθηκαν τα 10 πιο συχνά στυλ.

Q3. Στόχος: Να βρεθούν οι 20 καλλιτέχνες με τις περισσότερες κυκλοφορίες.

- Φιλτραρίστηκαν εκτός τα πεδία όπου το `artists_artist_name` ήταν null ή κενό.
- Οι καλλιτέχνες εντοπίστηκαν χρησιμοποιώντας μια UDF συνάρτηση (`has_artist`). Η συνάρτηση αυτή ελέγχει αν ο καλλιτέχνης υπάρχει μέσα σε ένα JSON array ή σε string. Αν βρεθεί ο καλλιτέχνης, επιστρέφει True.
- Το dataset ομαδοποιήθηκε με βάση το `artists_artist_name`.

- Ταξινομήθηκαν οι διακριτές κυκλοφορίες (`DISTINCT releases_release_id`).
- Τα αποτελέσματα ταξινομήθηκαν κατά φθίνουσα σειρά.
- Επιλέχθηκαν οι top 20 καλλιτέχνες.

Q4. Στόχος: Να βρεθούν οι top 10 καλλιτέχνες με τις περισσότερες κυκλοφορίες με style Metal.

- Φιλτραρίστηκε εκτός το πεδίο όπου το `artists_artist_name` είναι null ή κενό.
- Φιλτράρισμα των κυκλοφοριών που έχουν το style "Metal" χρησιμοποιώντας μια UDF συνάρτηση `has_style`.
- Ομαδοποίηση με βάση το `artists_artist_name`.
- Υπολογίστηκε ο αριθμός των κυκλοφοριών (`COUNT(*)`) για κάθε καλλιτέχνη.
- Τα αποτελέσματα ταξινομήθηκαν κατά φθίνουσα σειρά.
- Επιλέχθηκαν οι top 10 Metal καλλιτέχνες.

Q5. Στόχος: Να βρεθούν οι καλλιτέχνες που έχουν συμμετάσχει σε περισσότερα διαφορετικά είδη μουσικής.

- Φιλτραρίστηκαν εκτός τα πεδία `release_genres_genre` ή `artists_artist_name` ήταν null ή κενά.
- Ομαδοποίηση με βάση το artists_artist_name.
- Υπολογίστηκε ο αριθμός διαφορετικών genres (`COUNT(DISTINCT release_genres_genre)`) για κάθε καλλιτέχνη.
- Τα αποτελέσματα ταξινομήθηκαν κατά φθίνουσα σειρά.
- Επιλέχθηκαν οι top 10 καλλιτέχνες με τα περισσότερα genres.

Q6. Στόχος: Να βρεθούν οι 10 καλλιτέχνες με τις περισσότερες ελληνικές κυκλοφορίες.

- Φιλτραρίστηκαν εκτός τα πεδία όπου το `releases_release_country` είναι null ή κενά.
- Οι ελληνικές κυκλοφορίες εντοπίστηκαν χρησιμοποιώντας μια UDF συνάρτηση (`has_country`). Η συνάρτηση αυτή ελέγχει αν η χώρα υπάρχει μέσα σε ένα JSON array ή σε string. Αν βρεθεί η χώρα, επιστρέφει True.
- Ομαδοποίηση με βάση το πεδίο `artists_artist_name`.
- Υπολογίστηκε το πλήθος των κυκλοφοριών.
- Τα αποτελέσματα ταξινομήθηκαν κατά φθίνουσα σειρά.
- Επιλέχθηκαν οι top 10 καλλιτέχνες στην Ελλαδα.

Q7. Στόχος: Να βρεθεί ο μέσος όρος αριθμού κομματιών ανά κυκλοφορία.

- Ο αριθμός των tracks σε κάθε κυκλοφορία υπολογίζεται από την UDF συνάρτηση `track_count` η οποία παίρνει το πεδίο `tracklist_track_duration`, το οποίο περιέχει το χρονική διάρκεια κάθε κομματιού στην κυκλοφορία, και επιστρέφει το μέγεθος του array με τις χρονικές διάρκειες των κομματιών.
- Υπολογίζευαι ο μέσος όρος των αριθμών των tracks των κυκλοφοριών.

Q8. Στόχος: Να βρεθεί ο μέσος όρος των κυκλοφοριών ανά έτος και να εκτυπωθεί το πλήθος κυκλοφοριών ενός τυχαίου έτους (επιλέχθηκε το έτος 1999 τυχαία).

1. CTE: Δημιουργήθηκε ένας προσωρινός πίνακας `year_counts` και έγιναν τα παρακάτω:
   - Με την χρήση του `SUBSTRING(releases_release_released, 1, 4)` πάρθηκε το κομμάτι του έτους από το πεδίο `releases_release_released`.
   - Φιλτραρίστηκαν εκτός τα πεδία όπου το `releases_release_released` είναι null ή κενό.
   - Ομαδοποίηση με βάση το έτος κυκλοφορίας.
   - Υπολογίστηκε ο αριθμός μοναδικών κυκλοφοριών ανά έτος με το `COUNT(DISTINCT releases_release_id)`.
2. Βασικό SELECT:
   - Υπολογίστηκε ο μέσος όρος κυκλοφοριών ανά έτος με το `AVG(release_count)`.
   - Εντοπίστηκε το πλήθος των κυκλοφοριών του έτους 1999 με το `MAX(CASE WHEN year_part = '1999' THEN release_count END)`.

3.2 Join Queries

Σε αυτό το σημείο της εργασίας εξετάστηκαν join queries. Το dataset "releases.csv" περιείχε το πεδίο "artists_artist_id" στο οποίο υπήρχαν ids υπό την μορφή JSON array, το οποίο ήταν δύσκολο να επεξεργαστεί κατευθείαν το SQL query. Ήταν αναγκαία η υλοποιήση της UDF συνάρτησης `extract_all_ids`, η οποία παίρνει ως όρισμα ένα string και επιστρέφει ένα άλλο string με τα artists_artist_id διαχωρισμένα από κόμματα. Το JSON array είχε την μορφή: `["123", "456"]` κι η συνάρτηση το μετατρέπει σε ένα string της μορφής `'123,456'`, το οποίο βοηθάει τον διαχωρισμό των ids μετά από το SQL query. Αν το string που δεχθεί η συνάρτηση δεν είναι JSON αλλά είναι απλά string τότε επιστρέφει το αρχικό string. Αφού επεξηγήθηκε η λειτουργηκότητα της συνάρτησης, μπορούν να αναλυθούν τα join queries.

Join Q1. Στόχος: Να βρεθούν τα top genres ανά καλλιτέχνη.

- Η extract_all_ids παίρνει τα artist ids του κάθε release. Αν είναι JSON array, τα μετατρέπει σε string.
- Το `SPLIT(..., ',') + UNNEST` κάνει split τα IDs βάση το κόμμα και το `CROSS JOIN UNNEST(...)` δημιουργεί μια νέα γραμμή για κάθε artist που περιέχεται σε ένα release.
- Οι καινούργιες γραμμές αποθηκεύονται σε έναν προσωρινό πίνακα με όνομα `T` και στήλη `artist_id`.
- Κάνει το JOIN ανάμεσα στο `T.artist_id` και το `artists.artists_artist_id` με σκοπό να εμφανιστούν τα ονόματα των καλλιτεχνών.
- Φιλτράρει εκτός τα πεδία όπου το `release_genres_genre` και `artists_artist_name` είναι null ή κενό.
- Ομαδοποιεί τα αποτελέσματα με βάση `release_genres_genre` και `artists_artist_name`.
- Μετράει πόσα διαφορετικά releases έχει κάθε artist σε κάθε genre.
- Επιστρέφει τα top 20 (artist, genre) συνδυασμούς με τα περισσότερα releases.

Join Q2. Στόχος: Να βρεθούν οι καλλιτέχνες με τις περισσότερες κυκλοφορίες ανά χώρα.

- Η extract_all_ids παίρνει τα artist ids του κάθε release. Αν είναι JSON array, τα μετατρέπει σε string.
- Το `SPLIT(..., ',') + UNNEST` κάνει split τα IDs βάση το κόμμα και το `CROSS JOIN UNNEST(...)` δημιουργεί μια νέα γραμμή για κάθε artist που περιέχεται σε ένα release.
- Οι καινούργιες γραμμές αποθηκεύονται σε έναν προσωρινό πίνακα με όνομα `T` και στήλη `artist_id`.
- Κάνει το JOIN ανάμεσα στο `T.artist_id` και το `artists.artists_artist_id` με σκοπό να εμφανιστούν τα ονόματα των καλλιτεχνών.
- Φιλτράρει εκτός τα πεδία όπου το `releases_release_country` και `artists_artist_name` είναι null ή κενό.
- Ομαδοποιεί τα αποτελέσματα με βάση `releases_release_country` και `artists_artist_name`.
- Μετράει πόσα releases έχει κάθε artist σε κάθε χώρα. Προσοχή εδώ δεν μετράει διαφορετικά releases.
- Επιστρέφει τα top 10 (artist, country) συνδυασμούς με τα περισσότερα releases.

Join Q3. Στόχος: Να βρεθούν οι artist ή οποίοι δεν είναι μπάντες με τα περισσότερα releases και να εμφανιστούν δίπλα το πλήθος των συγκροτημάτων που έχουν συμμετάσχει. Το query αποτελείται από 3 επιμέρους queries:

- Το πρώτο query χρησιμοποιεί την συνάρτηση `extract_all_ids`, λειτουργεί ακριβώς με τον ίδιο τρόπο όπως προ αναφέρθηκε, και επισρέφει τον αριθμό των γκρουπς που έχει συμμετάσχει ένας καλλιτέχνης, μαζί με το id του, το όνομα του, και το `group_name_id` του. Το αποτέλεσμα που επιστρέφετε ονομάζεται `artist_groups`.
- Το δεύτερο query επιστρέφει τους arists με το CROSS JOIN όπως και τα προηγούμενα queries. Οπότε έχουμε ένα πίνακα για τον οποίο έχουμε πληροφορία για το `release_id` σε σχέση με το artist_id χωρίς να έχουν JSON arrays. Το αποτελέσμα ονομάζεται `group_releases`.
- Το τελευταίο query κάνει join μεταξύ των 2 queries και κάνει ομαδοποιήση βάση το `artist_id` και το αποτέλεσμα των πόσων groups έχει συμμετάσχει ο κάθε καλλιτέχνης και εμφανίζει τα αποτελέσματα σε φθίνουσα σειρά έτσι ώστε να δούμε τους καλλιτέχνες με τα περισσότερα releases αφού έχουμε κάνει πριν το `COUNT(DISTINCT gr.releases_release_id)` για να μετρήσουμε τον αριθμό των κυκλοφοριών.

4 The Sampling Technique

Σε αυτήν την εργασία εξετάστηκε η τεχνική του Reservoir Sampling καθώς και του Join Synopses μέσω του sampled CSV αρχείου με την τεχνική του reservoir sampling.

4.1 Reservoir Sampling In Q1-Q8

Η εργασία αξιοποιεί τη μέθοδο του Reservoir Sampling για την τυχαία επιλογή δείγματος μεγέθους k από ένα πολύ μεγάλο αρχείο CSV (το οποίο δεν μπορεί πρακτικά να φορτωθεί ολόκληρο στη μνήμη). Στην περίπτωση της εργασίας το αρχείο ήταν περίπου 24 GB. Για να γίνει το Reservoir Sampling δημιουργήθηκε η συνάρτηση `reservoir_sampling_csv(file_path, k)`, η οποία διαβάζει σειριακά το αρχείο CSV, χρησιμοποιώντας την τεχνική όπως περιγράφεται παρακάτω:

- Τα πρώτα k rows μπαίνουν απευθείας στο array με όνομα `reservoir`.
- Για κάθε επόμενο row στη θέση i, γίνεται generate ένας τυχαίος αριθμός j από το διάστημα [0, i].
- Αν το j είναι μικρότερο από k, τότε το στοιχείο στη θέση j του array `reservoir` αντικαθίσταται με το νέο row.
- Όταν τελειώσουν τα rows η συνάρτηση επιστρέφει το array.

Με αυτόν τον τρόπο εξασφαλίζεται ότι κάθε row έχει την ίδια πιθανότητα επιλογής στο τελικό δείγμα, χωρίς να χρειάζεται να φορτωθεί όλο το dataset στη μνήμη.

Για την εργασία μεταμορφώθηκαν δύο reservoir sampled αρχεία. Ένα με δείγμα k=1,793,822, που είναι περίπου το 10% του full dataset και ένα με δείγμα k=179,382 που είναι περίπου το 1% του full dataset. Το full dataset έχει 17,938,222 rows.

4.2 Join Synopses Using Reservoir Sampling

Η βασική ιδέα των join synopses είναι πως ξεκινάει από ένα δειγματοληπτημένο υποσύνολο ενός πίνακα (στην περίπτωση της εργασίας τα releases) και στη συνέχεια κατασκευάζονται αντίστοιχα δειγματοληπτημένα αποτελέσματα από τις αλυσίδες foreign key joins με άλλους πίνακες του schema. Έτσι υπάρχει η δυνατότητα να απαντηθούν ερωτήματα join πάνω σε πολύ μεγάλα δεδομένα, χωρίς να χρειάζεται να πρόσβαση σε ολόκληρο το dataset.

Στον κώδικα της εργασίας φορτώθηκε το sampled SCV αρχείο σε ένα προσωρινό view `sampled_releases` και στην συνέχεια με την χρήση της συνάρτησης `extract_all_ids` που αναλύθηκε πιο πάνω έγινε `CROSS JOIN UNNEST`. Αυτό δημιουργεί ξεχωριστή γραμμή για κάθε artist που σχετίζεται με το release και γίνεται για να επιτραπεί το join με τον πλήρη πίνακα artists. Στην συνέχεια, επιλέχθηκαν μόνο εκείνοι οι artists από τον πλήρη πίνακα artists που συμμετέχουν στο δείγμα των releases. Παρακάτω φαίνεται ο κώδικας:

```
CREATE TEMPORARY VIEW sampled_artists AS
SELECT *
FROM artists a
WHERE a.artists_artist_id IN (
    SELECT DISTINCT T.artist_id
    FROM sampled_releases r
    CROSS JOIN UNNEST(SPLIT(extract_all_ids(r.artists_artist_id), ',')) AS T(artist_id)
)
```

5 Αποτελέσματα

Παρακάτω φαίνονται τα αποτελέσματα των queries για το full dataset, 10% sampled dataset και 1% sampled dataset.

5.1 Q1: Top 5 Formats

Ο πίνακας παρακάτω δείχνει τα αποτελέσματα για το query με τα πιο συχνά formats για ολόκληρο dataset, 10% reservoir sampling, και 1% reservoir sampling.

Release Style	Full Dataset	10% Sample	1% Sample
Vinyl	7,201,212	720,329	71,839
CD	4,601,487	460,600	46,036
File	2,269,908	226,738	22,535
Cassette	1,435,550	143,287	14,498
CDr	702,552	70,144	7,067

Table 1: Πιο συχνά formats για το full dataset, για 10% sampled dataset και 1%.

Όπως φαίνεται στο πινακάκι τα αποτελέσματα είναι τα ίδια και στο full dataset και στα 2 sampled datasets. Το συγκεκριμένο query επιλέχθηκε για να παρουσιαστεί η τυχαίοτητα επιλογής του reservoir sampling που δίνει ίση πιθανότητα σε όλα τα entries για να μπουν στο reservoir. Αλλή μία σημαντική παρατήρηση στο συγκεκριμένο query είναι πως τα βινύλια καλύπτουν σχεδόν το 40% του dataset και τα CD καλύπτουν σχεδόν το 25% οπότε έτσι κι αλλιώς είναι δύσκολο να αλλάξει η σειρά παρουσιάσης των κυκλοφοριών. Τέλος όσο μεγαλύτερο είναι το μέγεθος ενός γεγονότος, τόσο πιο πιστά αποτυπώνεται σε τυχαία δείγματα. Αυτό φαίνεται και σε μερικά από τα επόμενα queries.

5.2 Q2: Top 10 Release Styles

Το παρακάτω πινακάκι δείχνει τα 10 μουσικά styles με τις περισσότερες πληροφορίες για ολόκληρο το dataset, 10% reservoir sampling, και 1% reservoir sampling.

Release Style	Full Dataset	10% Sample	1% Sample
Pop Rock	278,704	27,781	2,643
Vocal	252,125	25,344	2,493
Country	232,826	23,267	2,293
House	209,430	20,962	2,077
Folk	168,786	16,855	1,631
Punk	166,061	16,624	1,702
Chanson	156,069	15,297	1,586
Romantic	151,705	15,132	1,482
Techno	148,295	14,659	1,511
Alternative Rock	143,076	14,313	—

Table 2: Top 10 release styles για full dataset και reservoir samples.

Όπως και στο πρώτο query βλέπουμε πως και τα 2 reservoir samples έχουν αρκετές εγγραφές με αποτέλεσμα να είναι στην ίδια σειρά με το πλήρες dataset. Όταν μιλάμε για χιλιάδες εγγραφές τα reservoir samples, ακόμα και το 1% παραμένει αξιόπιστο. Το μόνο αρνητικό είναι πως στο 1% δεν εμφανίστηκε το Alternaite Rock σαν μουσικό στυλ, οπότε όταν οι εγγραφές πλησιάζουν πιο κοντά προς τις εκατοντάδες, αρχίζει να μην υπάρχει τόσο καλό representasion των δεδομένων.

5.3 Q3: Top 20 Artists with Most Releases

Το παρακάτω πινακάκι δείχνει τους 20 καλλιτέχνες με τις περισσότερες κυκλοφορίες. Ο πίνακας έχει χωριστεί σε δύο επιμέρους πίνακες για ευκολότερη σύγκριση.

Artist (Full)	Count	Artist (10%)	Count
Various	189,255	Various	19,015
Unknown Artist	73,790	Unknown Artist	7,380
The Beatles	31,823	The Beatles	3,257
Elvis Presley	23,724	Elvis Presley	2,381
The Rolling Stones	23,525	The Rolling Stones	2,346
No Artist	16,840	No Artist	1,718
Pink Floyd	15,359	Pink Floyd	1,531
Led Zeppelin	12,280	Led Zeppelin	1,269
Queen	12,252	Queen	1,198
David Bowie	12,185	David Bowie	1,167
Depeche Mode	12,145	Depeche Mode	1,138
Frank Sinatra	11,825	Frank Sinatra	1,137
Bob Dylan	11,759	Bob Dylan	1,200
Iron Maiden	11,307	Iron Maiden	1,105
Elton John	10,422	Elton John	1,052
Madonna	10,416	Madonna	1,040
Metallica	10,301	Metallica	962
Deep Purple	9,922	Deep Purple	994
U2	9,702	U2	989
Kiss	9,429	Kiss	946

Table 3: Top 20 artists: Full dataset vs 10% sample.

Artist (Full)	Count	Artist (1%)	Count
Various	189,255	Various	1,919
Unknown Artist	73,790	Unknown Artist	750
The Beatles	31,823	The Beatles	309
Elvis Presley	23,724	The Rolling Stones	242
The Rolling Stones	23,525	Elvis Presley	221
No Artist	16,840	No Artist	177
Pink Floyd	15,359	Pink Floyd	157
Led Zeppelin	12,280	Queen	151
Queen	12,252	David Bowie	136
David Bowie	12,185	Led Zeppelin	122
Depeche Mode	12,145	Depeche Mode	116
Frank Sinatra	11,825	Bob Dylan	113
Bob Dylan	11,759	Deep Purple	108
Iron Maiden	11,307	Madonna	103
Elton John	10,422	Frank Sinatra	102
Madonna	10,416	Iron Maiden	101
Metallica	10,301	AC/DC	101
Deep Purple	9,922	Metallica	101
U2	9,702	U2	94
Kiss	9,429	Kiss	94

Table 4: Top 20 artists: Full dataset vs 1% sample. Οι λανθασμένες κατατάξεις σημειώνονται με κόκκινο.

Στο συγκεκριμένο query επιλέχθηκαν να εμφανιστούν 20 καλλιτέχνες καθώς στο 1% reservoir sample φτάνει να έχει εγγραφές κάτω από 500, στο οποίο φαίνεται με κόκκινο χρώμα πως αρχίζουν οι διατάξεις να αλλάζουν, κι άλλοι καλλιτέχνες που θα έπρεπε να βρίσκονται πιο κάτω στην κατάταξη, βρίσκονται υψηλότερα. Αυτό συμβαίνει καθώς τα απότελέσματα των κυκλοφοριών είναι όλο και λιγότερα στο πλήρες dataset. Δεν εξασφαλίζεται η λειτουργικότητα του 1%, το οποίο είναι πολύ μικρότερο δείγμα. Όμως φαίνεται πως το δείγμα 10% είναι αρκέτα αντιπροσωπευτικό. Μία ενδιάφερουσα παρατήρηση είναι πως παρόλο που οι καλλιτέχνες Elvis Presley και Rolling Stones έχουν πολλές εγγραφές (23,724 και 23,525) το 1% reservoir sampling αλλάζει την κατάταξη τους καθώς έχουν πολύ κοντινά counts στο full dataset.

5.4 Q4: Top 10 Metal Artists

Το παρακάτω πινακάκι δείχνει τους 10 Metal καλλιτέχνες με τις περισσότερες κυκλοφορίες. Ο πίνακας έχει χωριστεί σε δύο επιμέρους πίνακες για ευκολότερη σύγκριση. Οι νέες εμφανίσεις σημειώνονται με μπλε και οι λανθασμένες κατατάξεις με κόκκινο.

Artist (Full)	Count	Artist (10%)	Count
Iron Maiden	9,842	Iron Maiden	981
Metallica	3,352	Metallica	319
Judas Priest	2,729	Judas Priest	265
Black Sabbath	2,012	Black Sabbath	197
Accept	1,724	Accept	174
Ozzy Osbourne	1,423	Ozzy Osbourne	140
King Diamond	1,364	King Diamond	142
Mercyful Fate	1,363	Mercyful Fate	138
Cannibal Corpse	1,265	Cannibal Corpse	139
Helloween	1,221	Dio (2)	121

Table 5: Top 10 Metal artists: Full dataset vs 10% sample.

Artist (Full)	Count	Artist (1%)	Count
Iron Maiden	9,842	Iron Maiden	87
Metallica	3,352	Metallica	25
Judas Priest	2,729	Judas Priest	30
Black Sabbath	2,012	Black Sabbath	19
Accept	1,724	Accept	18
Ozzy Osbourne	1,423	Ozzy Osbourne	17
King Diamond	1,364	Cannibal Corpse	17
Mercyful Fate	1,363	Manowar	14
Cannibal Corpse	1,265	Helloween	14
Helloween	1,221	Korn	12

Table 6: Top 10 Metal artists: Full dataset vs 1% sample.

Κάτι καινούργιο που φαίνεται στο συγκεκριμένο query είναι η εμφάνιση νέων καλλιτεχνών κι όχι απλά η αλλάγη κατατάξης λόγο διαφορετικού represantation των δεδομένων. Το συγκεκριμένο query έχει και where clause, όπου φιλτράρονται οι καλλιτέχνες οι οποίοι χαρακτηρίζονται με το genre "Metal". Παρόλο που θα περιμέναμε με λιγότερες εμφανίσεις στο πλήρες dataset να μην υπάρχει και τόση καλή αντιπροσώπευση των δεδομένων, στο 10% είναι αξιόπιστα τα αποτελέσματα εκτός από την νέα εμφάνιση του καλλιτέχνη Dio. Επίσης το 1% είναι αρκετά αξιόπιστο δεδομένου ότι τα δεομένα έχουν count της τάξης του 10.

5.5 Q5: Artists with most genres

Το παρακάτω πινακάκι δείχνει τους καλλιτέχνες που έχουν κυκλοφορίες με τα περισσότερα διαφορετικά genres. Ο πίνακας έχει χωριστεί σε δύο επιμέρους πίνακες για ευκολότερη σύγκριση. Οι νέες εμφανίσεις σημειώνονται με μπλε και οι λανθασμένες κατατάξεις με κόκκινο.

Artist (Full)	Count	Artist (10%)	Count
Various	2,038	Various	713
Unknown Artist	930	Unknown Artist	337
No Artist	291	No Artist	79
James Last	152	James Last	69
Elvis Presley	139	Elvis Presley	71
Ennio Morricone	124	Prince	53
Prince	113	Harry Belafonte	46
Harry Belafonte	111	Ennio Morricone	45
Santana	110	Madonna	43
101 Strings	101	Michael Jackson	43

Table 7: Artists with most genres: Full dataset vs 10% sample.

Artist (Full)	Count	Artist (1%)	Count
Various	2,038	Various	204
Unknown Artist	930	Unknown Artist	108
No Artist	291	No Artist	25
James Last	152	James Last	25
Elvis Presley	139	Elvis Presley	24
Ennio Morricone	124	José Feliciano	20
Prince	113	Mantovani And His Orchestra	19
Harry Belafonte	111	Michael Jackson	18
Santana	110	Ennio Morricone	17
101 Strings	101	Prince	16

Table 8: Artists with most genres: Full dataset vs 1% sample.

Καθώς το παραπάνω query είναι πιο περίπλοκο από τα προηγούμενα, τα αποτέλσμετα δεν είναι τόσο αξιοπιστά και στα δύο reservoir samples μετά τον πέμπτο καλλιτέχνη με τις περισσότερες κυκλοφορίες με διαφορετικά genres. Και τα δύο reservoir samples έχουν εμφανίσεις της τάξης του 10 το οποίο επηρεάζει αρκετά τα αποτελέσματα.

5.6 Q6: Top 10 Artists in Greece with most releases

Το παρακάτω πινακάκι δείχνει τους 10 καλλιτέχνες με τις περισσότερες κυκλοφορίες στην Ελλάδα. Ο πίνακας έχει χωριστεί σε δύο επιμέρους πίνακες για ευκολότερη σύγκριση. Οι νέες εμφανίσεις σημειώνονται με μπλε και οι λανθασμένες κατατάξεις με κόκκινο.

Artist (Full)	Count	Artist (10%)	Count
Various	2,171	Various	228
Mikis Theodorakis	522	Mikis Theodorakis	56
Στέλιος Καζαντζίδης	506	Unknown Artist	51
Unknown Artist	488	Manos Hadjidakis	37
Manos Hadjidakis	398	Στέλιος Καζαντζίδης	34
The Beatles	296	Anna Vissi	29
Γιάννης Πουλόπουλος	260	The Beatles	27
Anna Vissi	257	The Sisters Of Mercy	24
Στράτος Διονυσίου	225	Πόλυ Πάνου	22
Iron Maiden	223	Γιάννης Πουλόπουλος	22

Table 9: Top 10 artists in Greece: Full dataset vs 10% sample.

Artist (Full)	Count	Artist (1%)	Count
Various	2,171	Various	20
Mikis Theodorakis	522	Unknown Artist	5
Στέλιος Καζαντζίδης	506	The Beatles	5
Unknown Artist	488	Manos Hadjidakis	5
Manos Hadjidakis	398	Στέλιος Καζαντζίδης	4
The Beatles	296	Rotting Christ	4
Γιάννης Πουλόπουλος	260	Γιάννης Πουλόπουλος	4
Anna Vissi	257	Μανώλης Αγγελόπουλος	4
Στράτος Διονυσίου	225	Τόλης Βοσκόπουλος	4
Iron Maiden	223	Ελεύθεροι	3

Table 10: Top 10 artists in Greece: Full dataset vs 1% sample.

Το "Various" και ο Μίκης Θεοδωράκης εμφανίζονται σωστά μόνο στο 10% datasets, άρα μόνο η πιο συχνή κατηγορία διατηρείται στο παραπάνω query. Στο 10% sample το αποτέλεσμα είναι πιο κοντά στο πραγματικό, αλλά ήδη φαίνονται αλλαγές στην κατάταξη από τη στοχαστικότητα της δειγματοληψίας. Στο 1% φαίνεται πολύς "θόρυβος" και λάθος αποτελέσματα μόνο επειδή λόγο της τυχαιότητας του reservoir sampling, έτυχε να έχουν περισσότερες εμφανίσεις κάποια entries σε σχέση με τα πραγματικά.

5.7 Q7 & Q8: Average number of tracks per release & releases per year

Το παρακάτω πινακάκι δείχνει τον μέσο όρο κομματιών ανά κυκλοφορία, τον μέσο όρο κυκλοφοριών ανά έτος και τις κυκλοφορίες του 1999.

Metric	Full	10% Sample	1% Sample
Avg tracks per release	3	3	3
Avg releases per year	100,571	10,958	1,118
Releases in 1999	258,078	25,940	2,551

Table 11: Average tracks per release and releases per year (full dataset vs reservoir samples).

Παρατηρείται ότι ο μέσος όρος κομματιών ανά κυκλοφορία δεν αλλάζει στα διαφορετικά samples, γεγονός το οποίο δείχνει ότι τέτοιες αναλογικές μετρικές είναι σταθερές απέναντι στο sampling. Οι μετρικές που αφορούν πλήθος κυκλοφοριών (ανά έτος ή σε συγκεκριμένο έτος όπως το 1999) μειώνονται αναλογικά με το μέγεθος του δείγματος, επιβεβαιώνοντας ότι το reservoir sampling διατηρεί την κατανομή στο χρόνο. Έτσι, ακόμα και μικρά δείγματα δείχνουν αξιόπιστα τις γενικές τάσεις του dataset.

6 Join Queries

6.1 Join Query 1: Top Genres per Artist (20 entries)

Ο παρακάτω πίνακας δείχνει τους Top-20 καλλιτέχνες ανά είδος με βάση τις κυκλοφορίες, για το Full Dataset και τα 10% και 1% Samples. Οι νέες εμφανίσεις σημειώνονται με μπλε και οι λανθασμένες κατατάξεις με κόκκινο.

Artist (Genre)	Full	Artist (Genre)	10% Sample
Ludwig van Beethoven (Classical)	53,265	Ludwig van Beethoven (Classical)	5,388
Wolfgang Amadeus Mozart (Classical)	49,189	Wolfgang Amadeus Mozart (Classical)	5,000
Johann Sebastian Bach (Classical)	41,217	Johann Sebastian Bach (Classical)	4,128
Pyotr Ilyich Tchaikovsky (Classical)	27,687	Pyotr Ilyich Tchaikovsky (Classical)	2,761
Johannes Brahms (Classical)	23,232	Johannes Brahms (Classical)	2,351
Franz Schubert (Classical)	21,398	Franz Schubert (Classical)	2,198
Berliner Philharmoniker (Classical)	18,758	Berliner Philharmoniker (Classical)	1,875
The Rolling Stones (Rock)	18,327	The Rolling Stones (Rock)	1,850
The Beatles (Rock)	17,959	The Beatles (Rock)	1,816
Frédéric Chopin (Classical)	17,751	Frédéric Chopin (Classical)	1,768
London Symphony Orchestra (Classical)	16,161	London Symphony Orchestra (Classical)	1,591
Joseph Haydn (Classical)	15,407	Joseph Haydn (Classical)	1,558
Antonio Vivaldi (Classical)	14,484	Antonio Vivaldi (Classical)	1,455
Herbert von Karajan (Classical)	14,480	Herbert von Karajan (Classical)	1,485
Georg Friedrich Händel (Classical)	14,227	Georg Friedrich Händel (Classical)	1,399
Pink Floyd (Rock)	13,824	Pink Floyd (Rock)	1,395
Wiener Philharmoniker (Classical)	13,622	Wiener Philharmoniker (Classical)	1,363
Antonín Dvořák (Classical)	13,470	Antonín Dvořák (Classical)	1,320
Robert Schumann (Classical)	13,127	Robert Schumann (Classical)	1,296
Philharmonia Orchestra (Classical)	12,294	Giuseppe Verdi (Classical)	1,252

Table 12: Top-20 Artists by Genre: Full dataset vs. 10% sample

Artist (Genre)	Full	Artist (Genre)	1% Sample
Ludwig van Beethoven (Classical)	53,265	Ludwig van Beethoven (Classical)	498
Wolfgang Amadeus Mozart (Classical)	49,189	Wolfgang Amadeus Mozart (Classical)	485
Johann Sebastian Bach (Classical)	41,217	Johann Sebastian Bach (Classical)	402
Pyotr Ilyich Tchaikovsky (Classical)	27,687	Pyotr Ilyich Tchaikovsky (Classical)	275
Johannes Brahms (Classical)	23,232	Johannes Brahms (Classical)	239
Franz Schubert (Classical)	21,398	Franz Schubert (Classical)	209
Berliner Philharmoniker (Classical)	18,758	Frédéric Chopin (Classical)	196
The Rolling Stones (Rock)	18,327	Berliner Philharmoniker (Classical)	183
The Beatles (Rock)	17,959	The Rolling Stones (Rock)	178
Frédéric Chopin (Classical)	17,751	London Symphony Orchestra (Classical)	171
London Symphony Orchestra (Classical)	16,161	The Beatles (Rock)	168
Joseph Haydn (Classical)	15,407	Joseph Haydn(Classical)	164
Antonio Vivaldi (Classical)	14,484	Georg Friedrich Händel (Classical)	149
Herbert von Karajan (Classical)	14,480	Wiener Philharmoniker (Classical)	145
Georg Friedrich Händel (Classical)	14,227	Antonín Dvořák (Classical)	140
Pink Floyd (Rock)	13,824	Pink Floyd (Rock)	139
Wiener Philharmoniker (Classical)	13,622	Herbert von Karajan (Classical)	139
Antonín Dvořák (Classical)	13,470	Robert Schumann (Classical)	139
Robert Schumann (Classical)	13,127	Felix Mendelssohn-Bartholdy (Classical)	138
Philharmonia Orchestra (Classical)	12,294	Deep Purple (Rock)	132

Στο 10% sample τα αποτελέσματα παραμένουν σε μεγάλο βαθμό ίδια με το πλήρες dataset, με μικρές διαφορές στις χαμηλότερες θέσεις της κατάταξης. Στο 1% sample, οι top καλλιτέχνες εξακολουθούν να εμφανίζονται σωστά, ωστόσο παρατηρούνται έντονες ανακατατάξεις στις μεσαίες και χαμηλές θέσεις, καθώς και νέες λανθασμένες εμφανίσεις. Αυτό δείχνει ότι ενώ το reservoir sampling αποτυπώνει αξιόπιστα τις πιο σημαντικές συμπεριφορές ακόμη και σε μικρά δείγματα, για queries που θέλουν πιο λεπτομέρη αποτύπωση του αποτελέσματος είναι αναγκαία η ύπαρξη μεγαλύτερου δείγματος ώστε να ελαχιστοποιηθούν τα αποτελέσματα της τυχαιότητας του sampling.

Παρακάτω φαίνονται οι γραφικές παραστάσεις των μετρήσεων του πλήρες dataset του Join Query 1 σε σχέση με το 10% και 1% reservoir sampling:

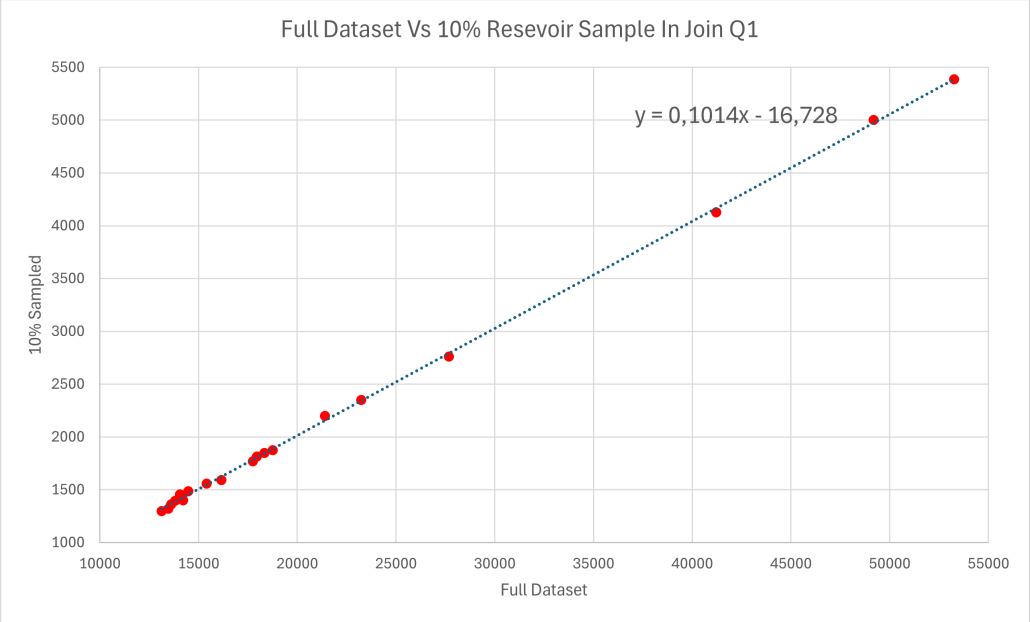

Figure 1: Γραφική Παράσταση Αποτελεσμάτων Πλήρους Dataset σε Σχέση με το 10% Reservoir Sample

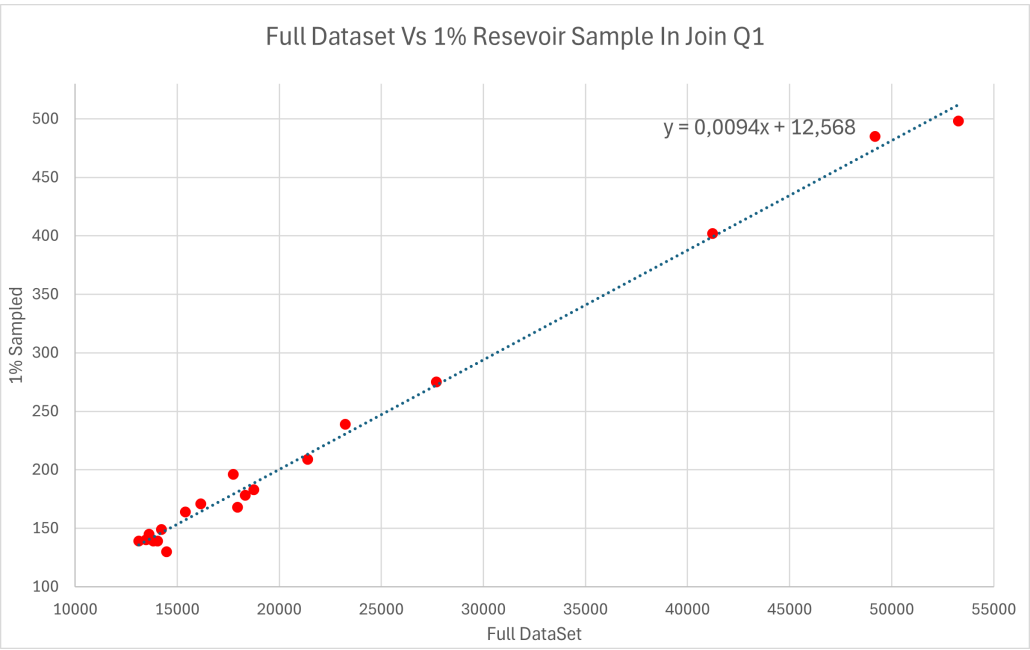

Figure 2: Γραφική Παράσταση Αποτελεσμάτων Πλήρους Dataset σε Σχέση με το 1% Reservoir Sample

Η γραφική παράσταση του πλήρους dataset σε σύγκριση με το 10% reservoir sample δείχνει μια γραμμική σχέση ($y = ax + b$), με τον συντελεστή a να προσεγγίζει το αναμενόμενο 0.1. Αυτό σημαίνει πως το reservoir sampling στο 10% αποτυπώνει με μεγάλη ακρίβεια την κατανομή των δεδομένων, διατηρώντας τη δομή του πλήρους dataset.

Όσον αφορά την γραφική παράσταση του 1% δείγματος φαίνεται πως ακόμα υπάρχει μια γραμμική συσχέτιση με το πλήρες dataset. Η κλίση της ευθείας (y=0.0094x+12.568) αποκλίνει πιο έντονα από την ιδανική τιμή 0.01. Αυτό σημαίνει ότι η δομή των δεδομένων δεν αποτυπώνεται με την ίδια συνέπεια όπως στο 10% δείγμα.

Παρακάτω φαίνεται ο πίνακας των σχετικών σφαλμάτων για τα reservoir samples 10% και 1%:

Artist(Genre)	**Full**	**10%**	**1%**	**Relative Error 10%**	**Relative Error 1%**
Ludwig van Beethoven (Classical)	53,265	5,388	498	1.15	6.51
Wolfgang Amadeus Mozart	49,189	5,000	485	1,65	1.40
Johann Sebastian Bach (Classical)	41,217	41,28	402	0.15	2.47
Pyotr Ilyich Tchaikovsky (Classical)	27,687	2,761	275	0.28	0.68
Johannes Brahms (Classical)	23,232	2,351	239	1.20	2.88
Franz Schubert (Classical)	21,398	2,198	209	2.72	2.33
Berliner Philharmoniker (Classical)	18,758	1,875	183	0.04	2.44
The Rolling Stones (Rock)	18,327	1,850	178	0.94	2.88
The Beatles (Rock)	17,959	1,816	168	1.12	6.45
Federic Chopin (Classical)	17,751	1,768	196	0.40	10.42
London Symphony Orchestra (Classical)	16,161	1,591	171	1.55	5.81
Joseph Haydn (Classical)	15,407	1,558	164	1.12	6.45
Antonio Vivaldi (Classical)	14,484	1,485	130	2.53	10.25
Herbert von Karajan (Classical)	14,048	1,455	139	3.57	1.05
Georg Friedrich Handel (Classical)	14,227	1,399	149	1.67	4.73
Pink Floyd (Rock)	13,824	1,395	139	0.91	0.55
Wiener Philharmoniker (Classical)	13,622	1,363	145	0.06	6.45
Antonín Dvořák (Classical)	13,470	1,320	140	2.00	3.93
Robert Schumann (Classical)	13,127	1,296	139	1.27	5.89

Table 13: Full Set Vs 10% and 1% With Relative Error In The Samplings

Για το 10% δείγμα, τα σχετικά σφάλματα είναι πολύ χαμηλά, συνήθως κάτω από 2%, με μέγιστο σφάλμα την τιμή 3,57%. Αυτό ενισχύει την παρατήρηση ότι το reservoir sampling σε αυτό το ποσοστό δίνει αξιόπιστα αποτελέσματα για queries που σχετίζονται με πιο σημαντικούς καλλιτέχνες.

Για το 1% δείγμα, βλέπουμε μεγάλες διακυμάνσεις. Κάποιες τιμές είναι ακόμα σωστές όπως ο Tchaikovsky με σχετικό σφάλμα 0,68%. Άλλες όμως έχουν τεράστια απόκλιση όπως ο Chopin με σχετικό σφάλμα 10,42% και ο Vivaldi με 10,25%. Αυτό συμβαίνει καθώς η τυχαιότητα του αλγορίθμου φαίνεται πολύ πιο έντονα σε μικρότερα δείγματα με αποτέλεσμα τα λιγότερο συχνά ονόματα να αλλοιώνονται σημαντικά.

6.2 Join Query 2: Artist Names per Country

Ο παρακάτω πίνακας δείχνει τους Top 10 καλλιτέχνες ανά χώρα με βάση τις κυκλοφορίες, για το Full Dataset και το 10% Sample. Οι νέες εμφανίσεις σημειώνονται με μπλε και οι λανθασμένες κατατάξεις με κόκκινο.

Artist	Country	Full	Artist	Country	10% Sample
Lata Mangeshkar	India	16,914	Lata Mangeshkar	India	1,723
Asha Bhosle	India	13,414	Elvis Presley	Europe	1,438
Elvis Presley	Europe	13,040	Asha Bhosle	India	1,269
W.A. Mozart	Germany	11,248	W.A. Mozart	Germany	1,098
Mohammed Rafi	India	10,626	Elvis Presley	US	1,029
J.S. Bach	Germany	10,318	Mohammed Rafi	India	1,020
Elvis Presley	US	10,220	J.S. Bach	Germany	1,013
L.v. Beethoven	Germany	10,081	L.v. Beethoven	Germany	1,009
Kishore Kumar	India	9,053	Bing Crosby	US	918
The Beatles	US	8,623	Frank Sinatra	US	896

Table 14: Top 10 artists per country: Full dataset vs. 10% sample

Artist	Country	Full	Artist	Country	1% Sample
Lata Mangeshkar	India	16,914	Lata Mangeshkar	India	199
Asha Bhosle	India	13,414	Asha Bhosle	India	159
Elvis Presley	Europe	13,040	W.A. Mozart	Germany	121
W.A. Mozart	Germany	11,248	J.S. Bach	Germany	116
Mohammed Rafi	India	10,626	Mohammed Rafi	India	100
J.S. Bach	Germany	10,318	Kishore Kumar	India	100
Elvis Presley	US	10,220	L.v. Beethoven	Germany	96
L.v. Beethoven	Germany	10,081	Elvis Presley	Europe	89
Kishore Kumar	India	9,053	Elvis Presley	US	88
The Beatles	US	8,623	Stevie Wonder	US	87

Table 15: Top 10 artists per country: Full dataset vs. 1% sample

Ως προς την ορθότητα του συγκεκριμένου query, οι πιο διάσημοι καλλιτέχνες με τις περισσότερες κυκλοφορίες, όπως ο Elvis Presley και οι Beatles, επειδή ήταν παγκοσμίως γνωστοί, έβγαλαν διαφορετικές κυκλοφορίες σε διαφορετικές χώρες και για αυτό ο Elvis Presley φαίνεται δύο φορές στο συγκεκριμένο πινανάκι, αφού είχε κυκλοφορίες και στην Ευρώπη και στις Ηνωμένες Πολιτείες. Η χώρα με τις περισσότερες κυκλοφορίες είναι η Ινδία και αυτό είναι λογικό καθώς ο πληθυσμός της Ινδίας είναι αρκετά μεγάλος και οι καλλιτέχνες της Ινδίας βγάζουν δίσκους μόνο για την Ινδία και όχι παγκοσμίως, επομένως έχουν πάρα πολλές κυκλοφορίες σε μία χώρα.

Όσον αφορά τα αποτελέσματα του query φαίνεται πως υπάρχουν αναδιατάξεις και νέες προσθήκες σε σχέση με το πλήρες dataset επειδή δεν υπάρχει αρκετά μεγάλος αριθμός αποτελεσμάτων για να υπάρχουν αρκετά entries στα reservoir samples.

6.3 Join Query 3: Top 10 artists with the most releases and the number of bands they have participated in

Ο παρακάτω πίνακας δείχνει τους 10 καλλιτέχνες με τις περισσότερες κυκλοφορίες και ταυτόχρονα σε πόσες μπάντες έχουν συμμετάσχει, για το Full Dataset και το 10% Sample.

Name	cnt_groups	Releases	Name	Releases 10%
Elvis Presley	3	34,852	Elvis Presley	3,502
Elton John	3	21,657	Elton John	2,117
Frank Sinatra	12	21,069	Frank Sinatra	2,103
Bob Dylan	9	16,106	Bob Dylan	1,637
Michael Jackson	5	15,027	Ella Fitzgerald	1,525
Ella Fitzgerald	6	14,855	Michael Jackson	1,494
Johnny Cash	6	14,672	Bing Crosby	1,467
Bing Crosby	5	14,387	Eric Clapton	1,397
Eric Clapton	15	13,888	Nat King Cole	1,385
Nat King Cole	10	13,874	Marvin Gaye (2)	1,380

Table 16: Top 10 artists with most releases: Full dataset vs. 10% sample

Name	cnt_groups	Releases	Name	Releases 10%
Elvis Presley	3	34,852	Elvis Presley	323
Elton John	3	21,657	Elton John	220
Frank Sinatra	12	21,069	Frank Sinatra	211
Bob Dylan	9	16,106	Bing Crosby	173
Michael Jackson	5	15,027	Bob Dylan	164
Ella Fitzgerald	6	14,855	Michael Jackson	162
Johnny Cash	6	14,672	Ella Fitzgerald	160
Bing Crosby	5	14,387	Marvin Gaye (2)	160
Eric Clapton	15	13,888	Nat King Cole	155
Nat King Cole	10	13,874	Eric Clapton	140

Table 17: Top 10 artists with most releases: Full dataset vs. 1% sample

Φαίνεται στο παραπάνω query πως καθώς έχουμε περισσότερα entries στα samples, τα αποτελέσματα είναι καλύτερα σε σχέση με το full dataset.

7 Χρόνοι Εκτέλεσης

Παρακάτω φαίνονται οι χρόνοι εκτέλεσης των queries. Οι χρόνοι είναι ενδεικτικοί, δηλαδή τα queries δεν εκτελέστηκαν πολλές φορές με σκοπό να παρουσιαστεί ο μέσος όρος τους. Παρόλα αυτά δείχνουν μια καλή εκτιμήση του χρόνου εκτέλεσης.

Queries	**Time Full(s)**	**Time 10% (s)**	**Time 1%(s)**
Q1	2m 59s 964ms	19s 767ms	1s 650ms
Q2	3m 9s 66ms	26s 383ms	13s 416ms
Q3	4m 31s 431ms	25s 394ms	2s 373ms
Q4	3m 17s 136ms	27s 550ms	4s 869ms
Q5	3m 36s 668ms	27s 655ms	2s 369ms
Q6	3m 47s 171ms	26s 527ms	10s 898ms
Q7	2m 59s 18ms	20s 769ms	4s 603ms
Q8	2m 56s 288ms	29s 934ms	2s 116ms

Table 18: Χρόνοι Εκτέλεσης Στα Πρώτα Queries

Παρακάτω παρουσιάζονται οι χρόνοι εκτέλεσης των Join Queries:

Join Queries	**Time Full(s)**	**Time 10% (s)**	**Time 1%(s)**
Join Q1	21m 2s 747ms	4ms 25s 486ms	1m 4s 0ms
Join Q2	14m 25s 980ms	3m 14s 369ms	50s 352ms
Join Q3	23m 19s 323ms	3m 46s 650ms	1m 19s 246ms

Table 19: Χρόνοι Εκτέλεσης Join Queries

Queries 1-8
Αρχικά η διαφοροποιήση στους χρόνους δείχνει ότι ο βαθμός επιτάχυνσης των χρόνων εκτέλεσης εξαρτάται από την πολυπλοκότητα και το είδος του query, δηλαδή το Q2 απαιτεί περισσότερο χρόνο για την ολοκλήρωση του ακόμα και στο 1% δείγμα. Ο χρόνος εκτέλεσης στο πλήρες dataset κυμαίνεται από περίπου 3 έως 4,5 λεπτά ανά query. Στο 10% δείγμα, οι χρόνοι μειώνονται αρκετά, φτάνοντας περίπου τα 20 με 30 δευτερόλεπτα και στο 1% δείγμα οι χρόνοι πέφτουν στα 1 με 13 δευτερόλεπτα.

Join Queries 1-3
Τα Join queries είναι προφανώς πιο χρονοβόρα: στο πλήρες dataset κυμαίνονται από 14 έως 23 λεπτά, στο δείγμα 10% οι χρόνοι ολοκλήρωσης των queries πέφτουν στα 3 με 4 λεπτά και στο 1% είναι περίπου στο 1 λεπτό.

Συμπερασματικά όσο πιο μικρό το δείγμα, τόσο μικρότερος ο χρόνος ολοκλήρωσης των queries αλλά και τόσο μεγαλύτερος ο κίνδυνος απώλειας δεδομένων.

8 Conclusion

Καταλήγουμε μετά από όλα αυτά τα πειράματα και τις μετρήσεις πως το reservoir sampling με δείγμα 10% είναι ιδιαίτερα ικανοποιητικό για ερωτήματα συχνότητας εμφάνισης και εξαιρετικά αποτελεσματικό για aggregate queries. Μία σημαντική παρατήρηση είναι ότι όσο μεγαλύτερος είναι ο αριθμός των εγγραφών στο πλήρες dataset, τόσο πιο αξιόπιστα είναι και τα αποτελέσματα που προκύπτουν από το δείγμα 10%.

Αντίθετα, στο 1% δείγμα παρατηρήθηκαν σημαντικές αποκλίσεις, ειδικά σε περιπτώσεις όπου το πλήρες dataset περιείχε λιγότερα entries. Παρ' όλα αυτά, ακόμη και στο 1%, τα aggregate queries παρήγαγαν ικανοποιητικά αποτελέσματα.

Για τα join queries, το δείγμα 10% παρουσίασε πολύ καλύτερη σχέση κόστους/αποτελέσματος σε σχέση με το πλήρες dataset, καθώς ο χρόνος ολοκλήρωσης των join queries έφτασε να είναι ως και 5 με 6 φορές μικρότερος. Για το 1% δείγμα τα αποτελέσματα δεν ήταν τόσο καλά, όπως φάνηκε και από το σχετικό σφάλμα το οποίο έφτασε μέχρι και την τιμή 10%. Παρ' όλ' αυτά είναι μια καλή λύση για να δωθεί μία γενική εικόνα για το αποτέλεσμα του query σε μικρό χρονό ολοκλήρωσης.

Συνοψίζοντας, το reservoir sampling αποτελεί μια αποτελεσματική τεχνική για aggregate queries και queries συχνότητας εμφάνισης προσφέροντας τον καλύτερο συμβιβασμό μεταξύ ακρίβειας και ταχύτητας.

9 References

- Reservoir Sampling Wikipedia https://en.wikipedia.org/wiki/Reservoir_sampling
- Geeks For Geeks Code: https://www.geeksforgeeks.org/dsa/reservoir-sampling/
- Professor's lecture on Join Synopses
- Professor's lecture on Reservoir Sampling
- Python Table API and SQL Queries https://nightlies.apache.org/flink/flink-docs-master/docs/dev/python/table/intro_to_table_api/
- Releases CSV https://www.kaggle.com/datasets/ofurkancoban/discogs-datasets-january-2025?select=discogs_20250201_releases.csv
- Artists CSV https://www.kaggle.com/datasets/ofurkancoban/discogs-datasets-january-2025?select=discogs_20250201_artists.csv