



PADDYPOWER.

 betfair

# Calitatea datelor



Defectele datelor pot fi corectate încă dinainte sau chiar în timpul raportării.

Acest lucru este posibil folosind pași intermediari în procesarea datelor.

- Tool-uri externe - teste, verificări, etc
- Verificări interne - cod, loguri sau metrici

Calitatea datelor este foarte importantă

- Detectarea încercărilor de fraudă
- Detectarea încercărilor de spălare de bani
- Protejarea clienților

# Calitatea datelor



Calitatea datelor are mai multe caracteristici:

- Acuratețe și precizie
- Validitate și legitimitate
- Consistență și fiabilitate
- Relevantă și actualitate
- Completitudine și exhaustivitate
- Disponibilitate și accesibilitate
- Granularitate și unicitate

# Calitatea datelor



Data trecută am populat tabelele cu date de test.

Este important să verificăm calitatea datelor cu care vom lucra mai departe.

De exemplu, pentru fiecare cont ar trebui să existe o singură intrare în baza de date.

---

```
SELECT
    COUNT(*),
    full_name,
    email_address
FROM
    `music-streaming-332016.star_schema.dim_account`
GROUP BY 2, 3
HAVING COUNT(*) > 1;
```

# Calitatea datelor



```
SELECT
  COUNT(*),
  song_title
FROM
  `music-streaming-332016.star_schema.dim_song`
GROUP BY 2
HAVING COUNT(*) > 1;
```

Processing location: us-west1

Query results

Query complete (0.3 sec elapsed, 415.9 KB processed)

Job information

f0\_ song\_title

2 delectus quia eos

2 voluptatem excepturi itaque

2 quia et accusantium

2 est et voluptas

2 excepturi a reprehenderit

Același lucru e valabil și pentru celelalte tabele.

Este corectă aceasta verificare?

# Calitatea datelor



```
SELECT
    s.stream_id,
    a.full_name,
    a.country,
    song.song_title,
    d.date
FROM
    `music-streaming-332016.star_schema.fact_streams` s
JOIN
    `music-streaming-332016.star_schema.dim_account` a
ON
    a.account_id = s.account_id
JOIN
    `music-streaming-332016.star_schema.dim_song` song
ON
    CAST (s.song_id AS integer) = song.song_id
JOIN
    `music-streaming-332016.star_schema.dim_date` d
ON
    s.stream_date_id = d.date_id
WHERE
    date = '2019-06-05'
    AND full_name = 'Karli Johnston';
```

Cum putem interpreta acest rezultat?

Row	stream_id	full_name	country	song_title	date
1	51431	Karli Johnston	Romania	distinctio dolore et	2019-06-05
2	51431	Karli Johnston	Romania	distinctio dolore et	2019-06-05

# Calitatea datelor



- Defectele datelor pot fi corectate înainte sau chiar la pasul de raportare
- Acest lucru este posibil folosind Tableau Prep (preprocesare / pregatire)
- Sau folosind direct funcții în BigQuery