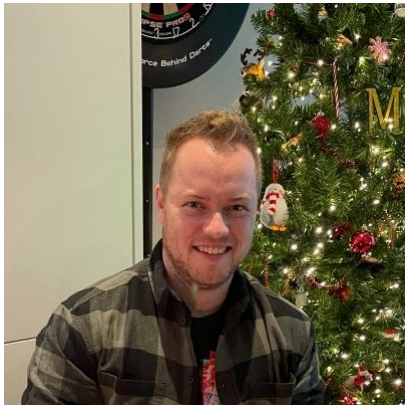
BigQuery Nested and Repeated Fields

Mark McCracken
The GCU Live

Agenda

- Who am I?
- What are nested and repeated records?
- Hands on!
- How has this been useful in a real-world scenario?
- Hands on!
- Pros and Cons
- Where can I learn more?



Mark McCracken

- Senior Engineering Manager at The Guardian, managing engineers in 3 teams across Identity, Transparency and Consent, and Personalisation. Largely using AWS
- Previously Head of Solution Engineering in Vodafone's Analytics Data and Services global team, designing solutions on GCP and managing Architects
- 6 Years in engineering roles across BI, Data Engineering, Software Engineering, Architecture, Networking & Infrastructure

The Guardian

vodafone

London Stock Exchange

LiveScore

gamesys

What are nested and repeated records?

Special formats for data in BigQuery tables, that make it easier to navigate.

We can have structs, or records, that are not repeated, which simply group related fields in a table.

We can store multiple values in an array field, or a repeated field, without requiring a join to another table, or resort to using strings to represent multiple values.

We can model data in a table more accurately, without requiring extra tables for one-to-many relationships.

https://cloud.google.com/bigquery/docs/nested-repeated

Hands on!

How has this been useful in a real-world scenario?

Google analytics is a popular way to record what happens on your website or app, and what actions a user takes, and things they view

Exporting this to bigquery makes heavy use of nested and repeated fields, because it makes sense

Querying this can be quite complex though, and the data won't come in a format that makes sense for your business domain

On livescore.com, we're interested in the custom dimensions of what a user is viewing, we've got Google Tag Manager to create these custom dimensions to analyse what a user is looking at, so we want to reformat this data, in a way that's easy for analysts to query effectively

Hands on!

Pros and Cons

This is a new skill to learn! You add complexity for data engineers, to save effort for data analysts, so engineering needs to get comfortable with creating and using this. Analysts have a learning curve to make the most of it, but after this learning curve, things are a lot easier and more standardised for a single table

This is Bigquery best practice for denormalising data, rather than using table joins

Consider bigquery limitations: only 100MB per row. This might sound massive, and in most cases it’s way more than adequate. However we had some unexpected behaviour that caused some difficulties, with users behaving outside the norm.

Where can I learn more?

- BigQuery docs - using arrays
- BQ docs - Best Practice - Use nested and repeated fields
- Exploring a powerful SQL pattern: ARRAY_AGG, STRUCT and UNNEST
- Code for this demo