How to Connect 100M+ Records to Tableau

Problem

We have a Gold table with 100+ million records that needs to be accessible in Tableau. Direct connection will be too slow.

My Recommended Solution

Main Approach: Don't connect directly to the big table

Instead, create smaller tables that Tableau can handle:

1. **Create summary tables** - aggregate the data by day/week/month
2. **Use Tableau extracts** - for when users need detailed data
3. **Keep live connections** only for small aggregated tables

Step-by-Step Plan

Step 1: Create Aggregated Tables

Step 2: Set Up Tableau Connections

- Connect live to summary tables (these are small and fast)
- Create extracts from main table for detailed analysis
- Set up incremental extract refresh (only pull new/changed data)

Step 3: Optimize Performance

- Add indexes on date and other filter columns
- Use context filters in Tableau to limit data early
- Set default date ranges (like "last 30 days")

Data Refresh Strategy

- **Summary tables**: Refresh every hour or daily
- **Tableau extracts**: Refresh daily or weekly
- **Live connections**: Real-time (but only to small summary tables)

Tools I'd Use

- **Database**: Whatever we currently have (Snowflake, BigQuery, etc.)
- **Scheduling**: Airflow or native database scheduling
- **Tableau**: Published data sources with proper refresh schedules

Trade-offs I Considered

Performance vs Fresh Data

- Chose performance because most business users are okay with daily updates
- Keep hourly refresh for executive dashboards that need fresher data

Simple vs Complex

- This approach is more complex than direct connection
- But it's necessary to handle 100M+ records without terrible performance

What Could Go Wrong

- Extract refreshes might fail - need monitoring and alerts
- Summary tables might not have the detail users want - may need to create more granular aggregations
- Costs could increase with more storage - but better than unusable dashboards

Success Measurements

- Dashboards load in under 10 seconds
- Extract refreshes complete successfully 95%+ of the time
- Users actually use the dashboards (engagement metrics)

Timeline

- **Week 1**: Create summary tables and test performance
- **Week 2**: Build Tableau data sources and first dashboards
- **Week 3**: Test with users and fix issues
- **Week 4**: Go live and monitor

Questions I Still Have

- What are the most common ways users will filter this data?
- How fresh does the data really need to be?
- Are there specific time periods or regions that get queried more often?

These answers would help me optimize the aggregation strategy better.