Databases for data analytics

https://github.com/evidencebp/databases-course/

Introduction to Performance

Semantic time units

- I’ll google something
- Coffee
- Lunch
- Over night
- Over weekend
- Rather long
- Too long

Best Performance Improvement

- The best developer is the one that sees an unreasonable and replies with WTF
- **Please** do not say it literally
- Justification really helps
- Alternative solution help even more

“Cheating” optimization options

- Donald Knuth: “premature optimization is the root of all evil”
- Stronger hardware
- Sampling and filtering
- Pre-compute

Important factors

- Query
- Table size
- Exponential growth
- Disk usage
- Hardware
- Indices
- Network
- Physical layer
- Expression optimization
- Normalization

Identifying performance issues

- Performance schema and other monitoring tools
  - It is hard to guess where your performance issue really are
- Execution plan and EXPLAIN ANALYZE

Indices

- Indices reduce search from O(n) to O(log(n))
- That is very effective when one search few values by comparison
- Not that effective if many values are required or some computation is done on the values
- Indices require storage and make update slower, getting worse over time
- Therefore it is common to REMOVE indices before massive inserts, creating them afterward

Normalization - avoid it…

- Normalization require joining tables, hurting performance
- Instead, denormalization can trade off storage and running time
- Some pre-computations (e.g. BI cubes, directors_genres), can also boost performance

Exercises

- Copy movies to a table without indices
- Compare execution plan and time on both tables of
  - Select of a specific id
  - Select of all movie of odd id
  - Select of all movies where 5000 < id < 10000
- Compare execution plan and time of select distinct year and select count(*) from group by year