Database Management for Data Scientists

5 SQL Performance Video - Additional Info & Corrigenda

1. Introduction

This document contains additional information and corrects some errors in the video "5 SQL Performance".

1.1. Slide 5 "Hardware components"

Additionally, netwerk components are even slower than RAM and Disks. If you retrieve a large result, that query takes even longer.

1.2. Slide 15 "Table Joins"

The access pattern from Nonclustered/Secondary index to the Heap/Clustered Index is performed internally by the RDBMS.

Imagine these two statements for the video sequence:

a) Search for surname "Beck"
Retrieve firstname
	`SELECT p.Firstname`
	`FROM person AS p`
	`WHERE p.Surname = 'Beck';`

b) Search for surname "Beck"
Retrieve primary key (PK),
e.g. for a join
	`SELECT pa.City`
	`FROM person AS p`
	`  INNER JOIN person_address AS pa`
	`  ON p.PersonId = pa.PersonId`
	`WHERE p.Surname = 'Beck';`

1.3. Slide 25 "Query Optimizer"

The query optimizer is part of the RDBMS. It is responsible for creating an execution plan. The actual execution of the query is done by other RDBMS components.

1.4. Slide 26 "Algebraic Transformation"

Additional information to the customer example why the query optimizer used a cross join:

- The query optimizer was not able to find a transformation for the complex SQL statement.
- As a fallback (or last resort), the optimizer chose a cross join to start with and then use all the selection and projection to remove unnecessary data.

1.5. Slide 36 "Date Range search"

Please note that the operators "BETWEEN … AND …" are also a bad choice when searching for dates.

BETWEEN … AND … is transformed into >= AND <=. It includes lower and upper boundaries!

1.6. Workbench – Time 45:40

The estimated number is roughly 900’000 or 900K, not 9’000.

1.7. Workbench – Time 48:40

The example for one month with time fraction is wrong. The example shows a time fraction with day “01-01-2021”. I should have used “2021-01-31 23:59:59” to show a problem statement with one month.