Fact-Based Indexing

Lothar Flatz – Senior Principal Consultant

May I introduce myself

» 25 Year Oracle Database experience (starting with Version 5)

» 15 Years Oracle Employee

» Oak Table Member

» Ex-Real World Performance Group

» Oracle ACE

» Signature Project: PVSS (CERN)

» US 8103658 B2 patent with Björn Engsig

» Senior Principal Consultant at Diso AG

About Diso AG

» Founded in 1996, 40 employees

» Located in Gümligen (Bern), Switzerland

» Portfolio:

- – Cloud Computing
- – Software Development
- – IT Consulting
- – Managed Services
- – Project Implementation
- – System Optimization

Fact-based Indexing

One of Codd's famous rules:

» The system must support an online, inline, relational catalog that is accessible to authorized users by means of their regular query language. That is, users must be able to access the database's structure (catalog) using the same query language that they use to access the database's data «

Codd, Edgar Frank (14 October 1985), "Is Your DBMS Really Relational?" ComputerWorld

Fact-based Indexing

The database has valuable information

Fact-based Indexing

Let's go and find it!

Raiders of the data dictionary –

Indexing for the workload

The idea

A consequential coffee break

"I bet, that at minimum 30% of our indexes are superfluous. I would prefer to have them all double checked."

OG_h_

Indexing for the workload

Some basic rules, please…

… no religious discussions!

What we have

The usual malaise

» Developers define indexes

» They use a rule of thumb or their gut feeling

» Quite often they never see the application under real load

Indexing for the workload

Our case

» Very complex data model

» Never ending performance trouble

We want to stop the issues once and for all!

Why work?

Can we automate this?

» First, while the automated tools reduce the complexity of the physical design process, it is still nontrivial to identify a representative workload that can be used to drive the physical design in its entirety. Second, automated tools do not consider all factors that impact physical design (e.g., the impact of replication architectures) «

Bruno, N. and Chaudhuri, S. 2007. Physical design refinement: The 'merge-reduce' approach. ACM Trans. Database Syst. 32, 4 (Nov. 2007)

Team up

Who can contribute?

» Developer (software vendor)

» DBAs

» Database consultant

Quality

Target

» Complete

» Non-redundant

» Minimal

» Efficient

Patch or build ?

Not incremental, but complete

Drop every index, do it new from scratch

First Step

Give it Structure

» Naming convention

» Index building rules based on design principles

» Foreign keys (to index or not to index)

» Physical attributes (table spaces)

We need to cover processing

Collect data

» Minimum a month

» Collect on daily, weekly and monthly processing

» Various sources

Find the metainformation

Sources

» Sys.col_usage$

» V$sql_plan

» DBA_HIST_SQL_PLAN

» V$SQL

» Column statistics

» Sys_col_group_usage$?

https://blogs.oracle.com/optimizer/entry/how_do_i_know_what_extended_statistics_are_needed_for_a_given_workload

What is important?

Column Statistics

```
COLUMN_NAME                      EQUALITY_PREDS RANGE_PREDS LIKE_PREDS EQUIJOIN_PREDS
---------------- -------------- ----------- ---------- -------------- ------------
KUND_ID                                     129          0          0            165
KUND_STATUS                                 165          0          0              0
KUND_CLASSID                                  0          0          0            154
KUND_MANDANT                                107          1          0             10
KUND_TYPE                                   104          0          0              0
KUND_FAMILIENNR                              90          0          0              0
KUND_GEBURTSTAG                              84          0          0              0
KUND_POTFAMILIENNR                           84          0          0              0
KUND_VERSICHERTENNR                          67          0          0              1
```

What is selective?

Column Statistics

Column Name	Distinct Values	Density	NUM_BUCKETS
KUND_VERSICHERTENNR	3,191,680	0	1
KUND_GPANR	2,683,264	0	1
KUND_FAMILIENNR	1,265,369	0	1
KUND_BEZEICHNER	1,228,595	0	200
KUND_STRASSE	449,653	0	198
KUND_POTFAMILIENNR	197,732	0	200
KUND_NAME	157,774	0	199
KUND_ROOTID	68,922	0	1
KUND_VORNAME	66,494	0	195
KUND_GEBURTSTAG	36,520	0	200
KUND_KVNR	10,227	0	116
KUND_ORT	8,964	0	174
KUND_PLZ	4,242	0	200
KUND_MANDANT	28	0	27
KUND_TYPE	6	0	5

Search combinations

Collected from v$sql_plan

```
COL_STR                                                            Number Queries  EXECUTIONS
----------------------------------------------------------------- --------------- ----------
KUND_ID                                                                        84    3408270
KUND_TYPE,NLS_UPPER(KUND_BEZEICHNER),NLS_UPPER(KUND_KVNR),NLS_UPPER(KO           1     372568
NT_VERSICHERTENNR)

KUND_FAMILIENNR,KUND_ID,KUND_MANDANT                                            17     175198
KUND_FAMILIENNR,KUND_ID,KUND_MANDANT,KUND_POTFAMILIENNR                         17      27890
KUND_ID,KUND_POTFAMILIENNR                                                       2      24946
KUND_TYPE,NLS_UPPER(KUND_NAME),NLS_UPPER(KUND_VORNAME)                           1      11920
KUND_FAMILIENNR                                                                  2      11879
KUND_ID,KUND_TYPE                                                                1      10029
KUND_FAMILIENNR,KUND_POTFAMILIENNR                                               2       6654
KUND_GEBURTSTAG,KUND_TYPE,NLS_UPPER(KUND_PLZ)                                    1       3862
KUND_TYPE,KUND_VERSICHERTENNR                                                    2       2936
KUND_POTFAMILIENNR                                                               2        368
```
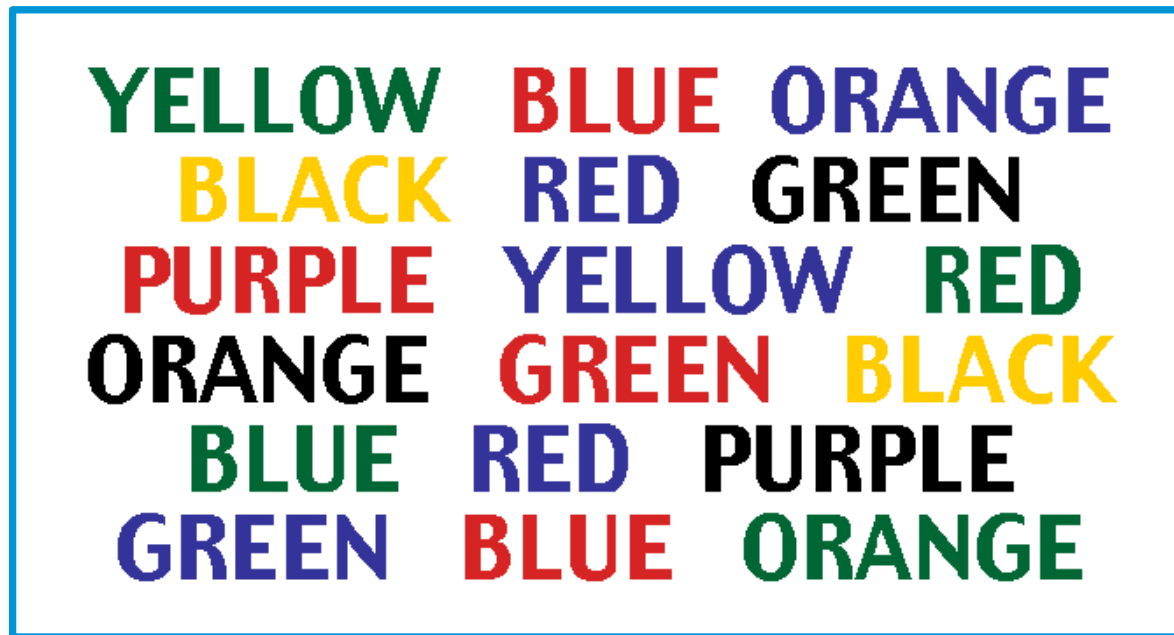
What was a feeling

All information is on the table...

... we just have to organize it

Brain block, got stuck?

Look at the chart and say the **COLOR** and not the word.

Your right brain tries to say the color while your left brain insist on reading the word.

# Words blocking the flow

How do I get out of it?

Let go your conscious self and …. act on instinct.

Order and conquer

There it is …

COLUMN_NAME	EQUALITY_PREDS	RANGE_PREDS	LIKE_PREDS	EQUIJOIN_PREDS	Rank
KUND_ID	129	0	0	165	1
KUND_STATUS	165	0	0	0	2
KUND_CLASSID	0	0	0	154	3
KUND_MANDANT	107	1	0	10	4
KUND_TYPE	104	0	0	0	5
KUND_FAMILIENNR	90	0	0	0	6
KUND_GEBURTSTAG	84	0	0	0	7
KUND_POTFAMILIENNR	84	0	0	0	
KUND_VERSICHERTENNR	67	0	0	1	

KUND_FAMILIENNR,KUND_ID,KUND_MANDANT

KUND_ID,KUND_MANDANT, KUND_FAMILIENNR
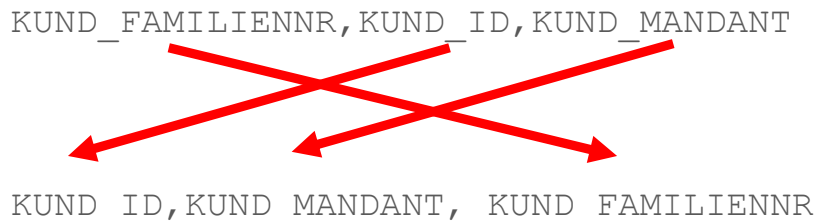
All you need is some repeating structure.

Arrange the columns such that reuse is most likely

Reduce

Eliminate superfluous indexes

```
{last name, first name, Age}
{last name, first name}
{last name}
```

Merge

Combine indexes when possible

```
{last name, first name, City}
{last name, first name, ZIP}
```

{last name, first name, City, ZIP}

Compare: Jonathan Lewis's Index usage

Indexing for the Workload

Implement

» Generate create index scripts

» Drop all existing indexes
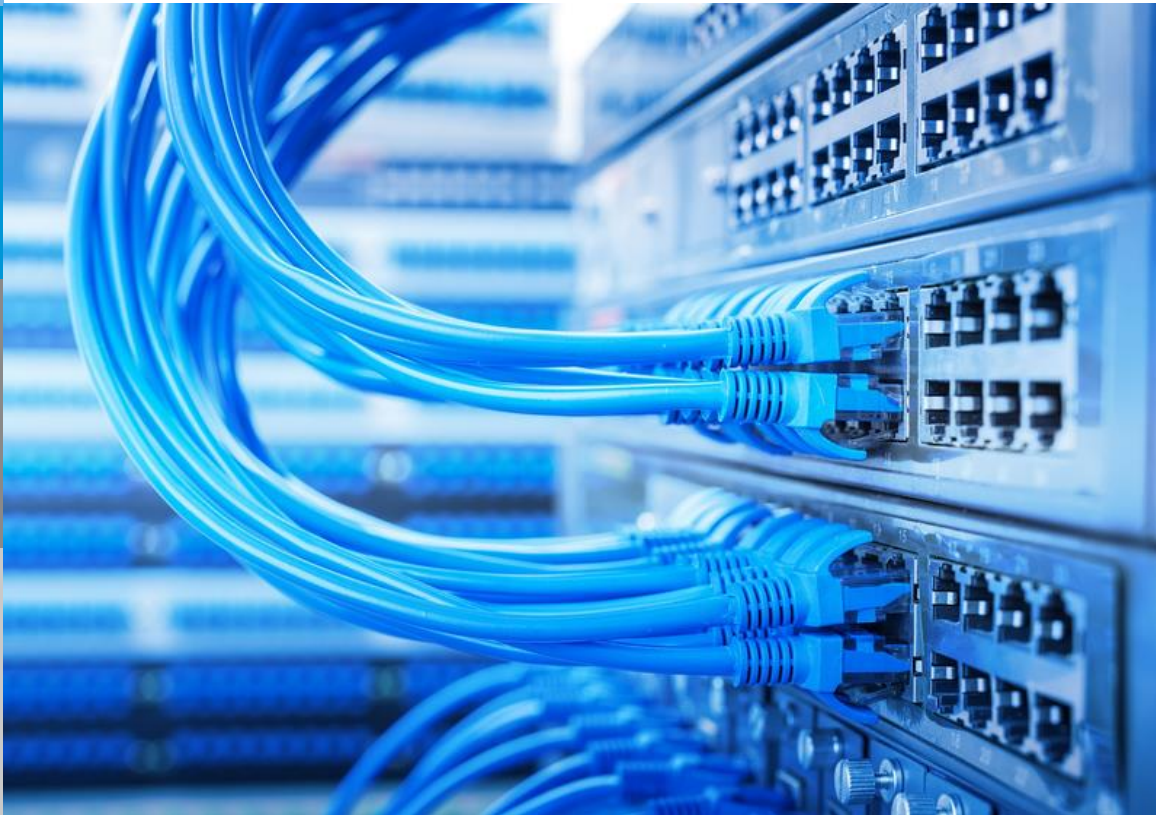
» Build indexing from scratch

Safty first

Test

» Loadtests (Batch)

» RAT

Wrap up

Close remaining gaps

» Missing indexes

» Improvable indexes

» Manual patching

Indexing for the Workload

Results

» Less index space (30%)

» Faster batch programs

» Higher plan stability

» 90% less memory according to software vendor recommendation

Finally

Done !

» The issues did not come back

» They were resolved once and for all

Indexing for the workload

7 years later

» The DBAs at (company where indexing was changed) are looking so relaxed. I want that too! «

A DBA group leader running that same software visiting the site

Conclusions

No Fairy Tales:

There are no alternative to hard facts

Thank you very much for your attention!

Do not hesitate to contact me for more information.

Lothar Flatz
Diso AG   Phone +41 31 958 90 90
Morgenstrasse 1   info@diso.ch
3073 Gümligen / CH   www.diso.ch