PASS SUMMIT 2014

Niko Neugebauer, OH22

ETL Patterns

with Clustered Columnstore

PASS NOVEMBER 4-7 | SEATTLE, WA
The Conference for SQL Server Professionals.

Please silence cell phones

Explore Everything PASS Has to Offer

24 HOURS OF PASS

Free SQL Server and BI Web Events

PASS SQL saturday

Free 1-day Training Events

PASS SQLRally

Regional Event

PASS SUMMIT

This is Community

PASS BUSINESS ANALYTICS CONFERENCE

Business Analytics Training

PASS CHAPTERS

Local User Groups Around the World

Session Recordings

PASS CommunityCONNECTOR

PASS Newsletter

PASS VIRTUAL CHAPTERS

Free Online Technical Training

Session Evaluations

Your feedback is important and valuable.

3 ways to access

Submit by 11:59 PM EST Friday Nov. 7 to **WIN** prizes

Evaluation Deadline:
11:59 PM EST, Sunday Nov. 16

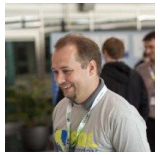
Go to
passsummit.com/evals

Download the GuideBook App and search: **PASS Summit 2014**

Follow the QR code link displayed on session signage throughout the conference venue and in the program guide

PASS

Niko Neugebauer



Microsoft Data Platform Professional

OH22 (http://www.oh22.net)

SQL Server MVP

Founder & Co-Founder of 3 Portuguese PASS Chapters

In-Memory VC co-lead

Blog: http://www.nikoport.com

Twitter: @NikoNeugebauer

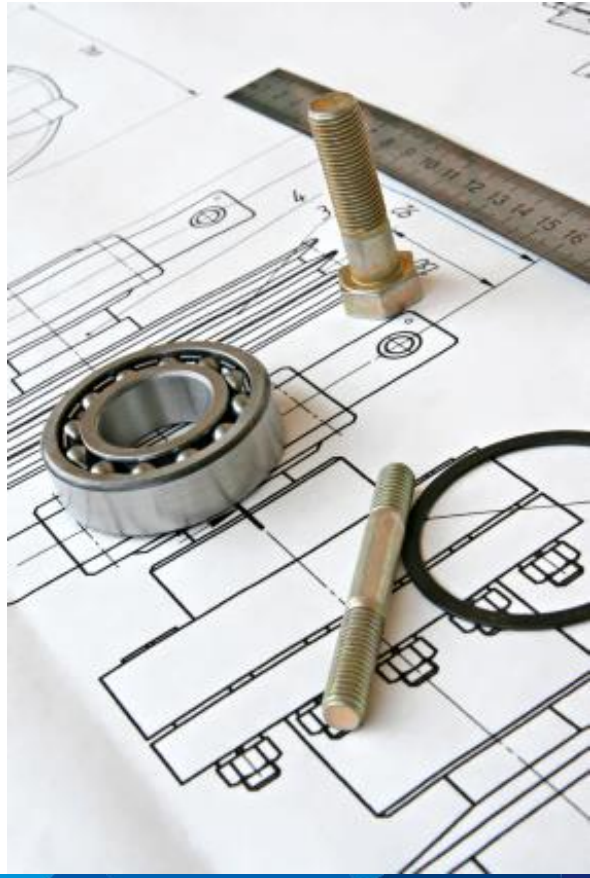
LinkedIn: http://pt.linkedin.com/in/webcaravela

Email: info@webcaravela.com

Oha! Since this is 400-level session:

And so I assume:

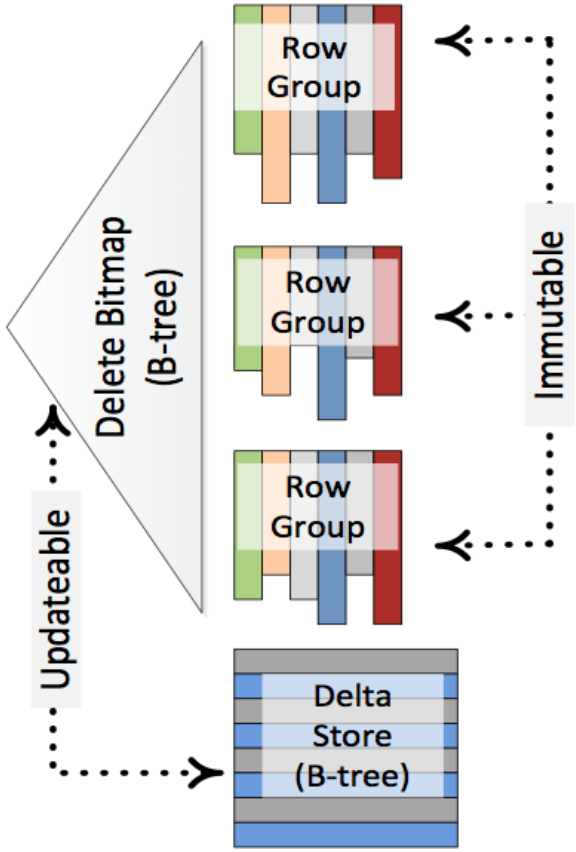
- You know the Columnstore Structure Elements, such as **Row Groups, Delta-Stores, & Deleted Bitmaps**.
- You do understand Segment Elimination.
- You have heard about Tuple Mover. ☺
- You know what Columnstore Dictionary is, what types of dictionaries do exist and why.
- You know what a Batch Mode is.



Clustered Columnstore:

- Delta-Stores (open & close)
- Deleted Bitmap
- Update work as a

DELETE + INSERT



General Thoughts:

Working with CCI requires a lot of resources, and so you might want/need to:

- Use **Resource Governor** – control memory grants and impact on CPU & IO. (You should consider capping the number of cores & memory given to Columnstore.)
- Use **Partitioning** – it helps you to control the impact.

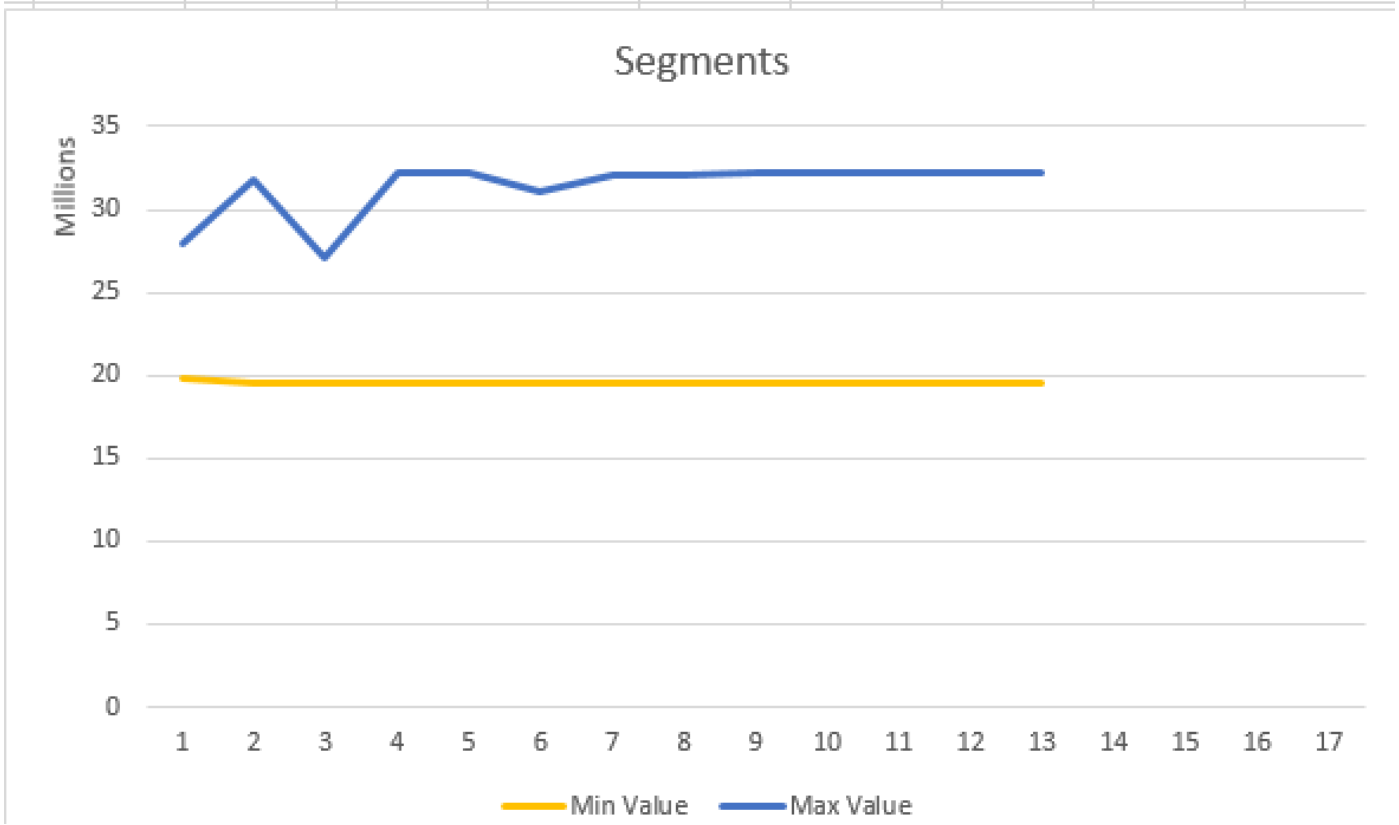
ETL, ELT, LTE – My Today's Agenda

- **E** – **Extract** Data from Columnstore
- **T** – **Transform** (Maintain)
- **L** – **Load data** into Columnstore

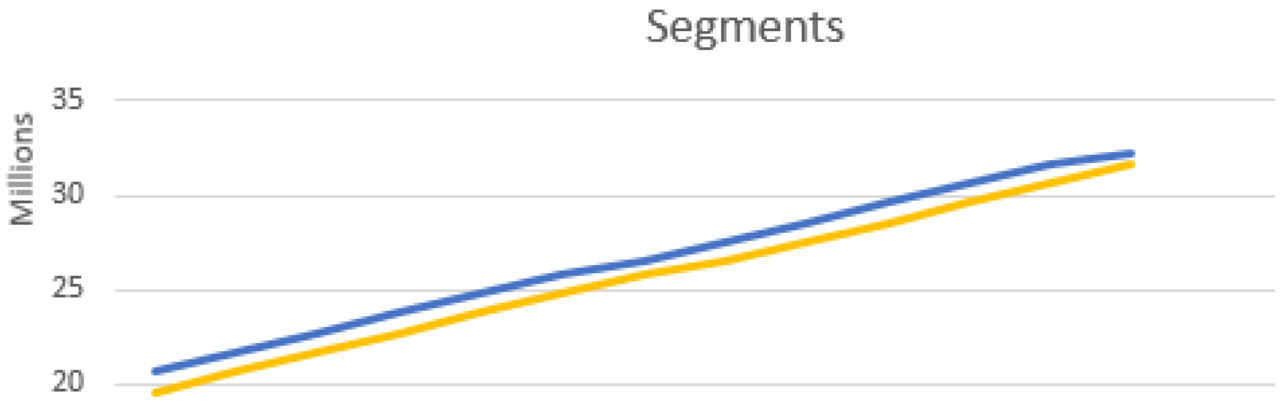
Extract

- Use Batch Processing, unless you prove that it does not help (You will need DOP>=2) Ω
- Avoid Delta Stores (Force Tuple Mover to close & compress them with Reorganize with (CLOSE_ALL_DELTA_STORES)) Ω

Consider Example of this Column:



Extraction with Segment Clustering



- Use Column Cluserting (You will need 1 to keep it perfect (ref: 2014 RTM + CU2)) **Ω**
- Use Partitioning:

  it helps Column Clustering Rebuilds with DOP = 1
  (Resource & impact optimization)

Extract continued

CCI Rebuild is a *half-online*, this means that you can read data but writing new info is not available.

- Use SELECT INTO, which became parallel execution plan in SQL Server 2014

Transform (Maintain)

- Do **not** rebuild CCI unless you really need it.
- **Columnstore is not a Rowstore.**
- What is the percentage of your
  data that generally became *deleted*?
- What is the improvement
  that you are looking for ? **Ω**

Transform (Maintain)

Rebuild if you have a good window and need to realign column clustering.

If Rebuilding frequently, use partitioning – it helps you to minimize impact and control better

Create your own Columnstore Resource Group in Resource Governor Ω

Load

- Use Resource Governor
- Use Partitioning (especially for Switching In)
- Always use Batch Processing, unless proven that you don't need it
  (You will need DOP>=2)
- Use BULK Load API (with more than 102.400 rows)
- Avoid using open Delta-Stores
- Use Parallel T-SQL (SELECT INTO)

Load Strategies

Columnstore Demo

Load Scenarios

Things to bear in mind:

Regarding Dictionaries:

- Global Dictionaries are created only on the first load/build and than they are never updated
- The Sampling is Max (1 Million Rows, 1%)
- Single Threaded

Load Scenarios

Global Dictionaries are not your Global Saviours:

- You need to have similar data in your column in order to get their usage
- You might actually get much more compact & efficient Local Dictionaries

Questions?

Thank you!

Links:

My blog series on Columnstore Indexes (40+ Blogposts):

- http://www.nikoport.com/columnstore/

Remus Rusanu Introduction for Clustered Columnstore:

- http://rusanu.com/2013/06/11/sql-server-clustered-columnstore-indexes-at-teched-2013/

White Paper on the Clustered Columnstore:

- http://research.microsoft.com/pubs/193599/Apollo3%20-%20Sigmod%202013%20-%20final.pdf

Connect Items

Implement Batch Mode Support for Row Store:

- https://connect.microsoft.com/SQLServer/Feedback/Details/938021

Multi-threaded rebuilds of Clustered Columnstore Indexes break the sequence of pre-sorted segment ordering (Order Clustering):

- https://connect.microsoft.com/SQLServer/Feedback/Details/912452

Columnstore Segments Maintenance – Remove & Merge:

- https://connect.microsoft.com/SQLServer/Feedback/Details/930664