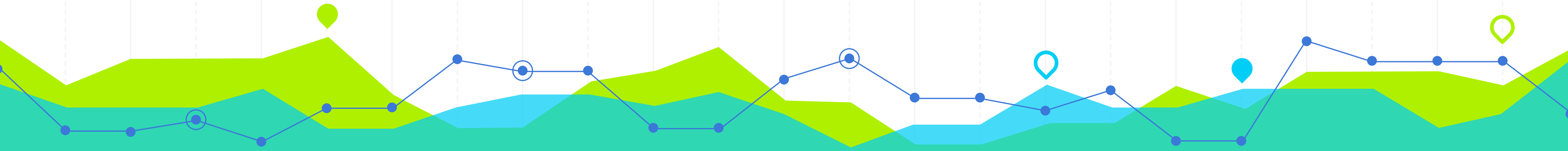

Mining Statistics for Data Insights

Deborah Melkin (she\her)

EightKB, August 8, 2024

About Me

- 20+ years as a DBA
- Mainly work with SQL Server & OLTP
- Data Platform WIT co-leader
- WITspiration co-founder
- Redgate Community Ambassador
- Microsoft MVP – Data Platform

About Me – The Important Stuff

- I'm a long time member of the alto section.
- I go to bluegrass jams regularly.
- I've been learning guitar and mandolin.
- I am a bit of a musical theater geek.
- My husband and I do geeky things with our dog.

Backstory

I had a work proj
was looking for d
profiling & qualit
information on se
tables …

rts work on his
s session and
Erin Stellato’s
as a refresher…

> *Could I have used statistics to get the data I needed from my project?*

How to Test this Theory?

- Understand the Basics
  - What do statistics contain
  - How they’re created
  - How they’re updated
- What do statistics look like in running workloads
- Combine statistics data with other metadata

Terminology

- Data Profiling
  - Data profiling is the process of examining, analyzing, and creating useful summaries of data. The process yields a high-level overview which aids in the discovery of data quality issues, risks, and overall trends... More specifically, data profiling sifts through data to determine its legitimacy and quality. (*Talend*)

Terminology

- Data Quality
  - Data quality measures how well a dataset meets criteria for accuracy, completeness, validity, consistency, uniqueness, timeliness and fitness for purpose, and it is critical to all data governance initiatives within an organization. (*IBM*)

Why look at Statistics

- Stats summarize the data in the table
- Why not use Indexes instead of Statistics?
  - Indexes say where to find the data (performance)
  - Stats say what is in the data (distribution)
- The alternative is to directly query table…

Let’s go to Demos…

Is this a solution to my work project?

- Yes!

Can answer questions directly about uniqueness.

Joining to other meta data and query plans starts answering data profiling questions

- But it’s not perfect…

Completeness has to be implied in some cases.

Doesn’t fully answer some of the data quality issues.

Statistics may not cover everything and extra work may be needed.

Additional Thoughts & Conclusions

- May not want to do this from production but copy data to a different database to query.
    - Turn off auto update of stats first!!!!
    - Best time to get data may be after index rebuilds and statistic updates
- This is only snapshot information.
    - Changes over time may need to be tracked differently but could be done.

Microsoft Learn Resources

- Statistics
- DBCC SHOWSTATISTICS
- Limitations when re-indexing

Resources

- IBM Definition of Data Quality
- Talend Definition of Data Profiling
- Erin Stellato: Demystifying Statistics (EightKB 2020)
- Andy Yun: A Query Tuner’s Practical Guide to Statistics
- Kendra Little: Can I Use Statistics to Design Indexes
- Brent Ozar: How to Download the Stack Overflow Database

Resources – Filtered Index

- Kimberly Tripp: Filtered Indexes and Filtered Stats Might Become Seriously Out-of-Date
- Aaron Bertrand: Filtered Indexes
- Paul White: Optimizer Limitations with Filtered Indexes

Any Other Questions?

- Social Media handles:
  @dgmelkin (@dataplatform.social) (.bsky.social)
- Email: dgmelkin@gmail.com
- Blog: DebtheDBA.wordpress.com
- Github: https://github.com/DebtheDBA/

Thanks for coming!