PASS SUMMIT 2012

Demystifying Database Statistics

Erin Stellato
Senior Consultant
SQLskills.com

PASS

PASS

http://www.erinstellato.com/statistics

twitter

@erinstellato

SQL skills.com

immerse yourself in sql server

http://www.sqlskills.com/blogs/erin

erin@sqlskills.com

MVP
Microsoft®
Most Valuable
Professional

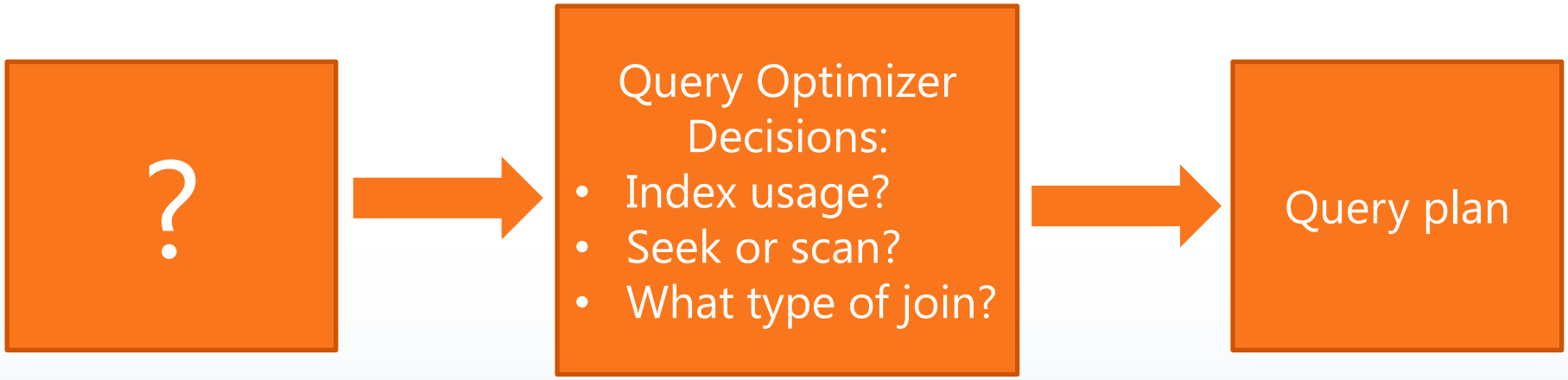
Abstract

Database statistics are not made up numbers, like the statistics you may hear on TV or read in a news article. They are critical metadata used by SQL Server's query optimizer. In this session, we will look at statistics from the ground up and remove the mystery that surrounds them. Demos and real-world examples will be used to explain what statistics are, why you should care about them, and what you can control.

By the end of the session, you will understand:

- How to view statistics and interpret the histogram
- How and when to update statistics
- How statistics affect the query optimizer
- When you can let SQL Server manage statistics, and when you need to step in

Why Should We Care About Statistics?



How Are Statistics Created?

"…over the supplied column or set of columns of a table or indexed view…"

For every index (and indexed view)

- Created automatically

For individual columns

- Created automatically or manually

For multiple columns

- Created manually

For filtered values in a column

- Created manually

Use the sys.stats catalog view to list all statistics for a table

How Are Statistics Updated?

"…Up-to-date statistics allow the optimizer to accurately assess the cost of different query plans, and choose a high-quality plan."

As part of index rebuilds

- Only updates index statistics
- Equivalent of a full scan (100% sample)

Manually (UPDATE STATISTICS or sp_updatestats)

How Are Statistics Updated?(2)

Automatically

- A minimum of 500 + 20% of the rows
  - Exception: Trace Flag 2371 (2008R2 SP1+)
- Not a full scan
- Can be immediate or asynchronous
- Index, column or filtered stats

Remember that updating statistics causes queries to recompile*

How often should I update statistics?

Let SQL Server manage statistics?

Know your data

- Volume
- Rate of change
- Skew of data

Let SQL Server tell you about your data

- Profiler, Xevents for auto stats update (volatile)
- sys.dm_db_index_physical_stats for volume
- sys.dm_db_index_usage stats for update
  - Remember that updates will occur more often for a new database/new tables

There is no perfect option, the goal is to provide a better understanding of what your data is like

How do I know what sample to use?

Use the default sample

Use 100% for all

Use 100% for columns with very skewed distribution

Lower the threshold for the default update (Trace Flag)

Sample Size Does Matter

Index Rebuild = 100% sample

UPDATE STATISTICS

- WITH FULLSCAN = 100% sample
- WITH SAMPLE number = % or number of rows
- WITH RESAMPLE = inherited sample
- (default) = less than 10%

What if the value doesn't exist in the histogram? (move this to later)

But what ARE statistics *really*?

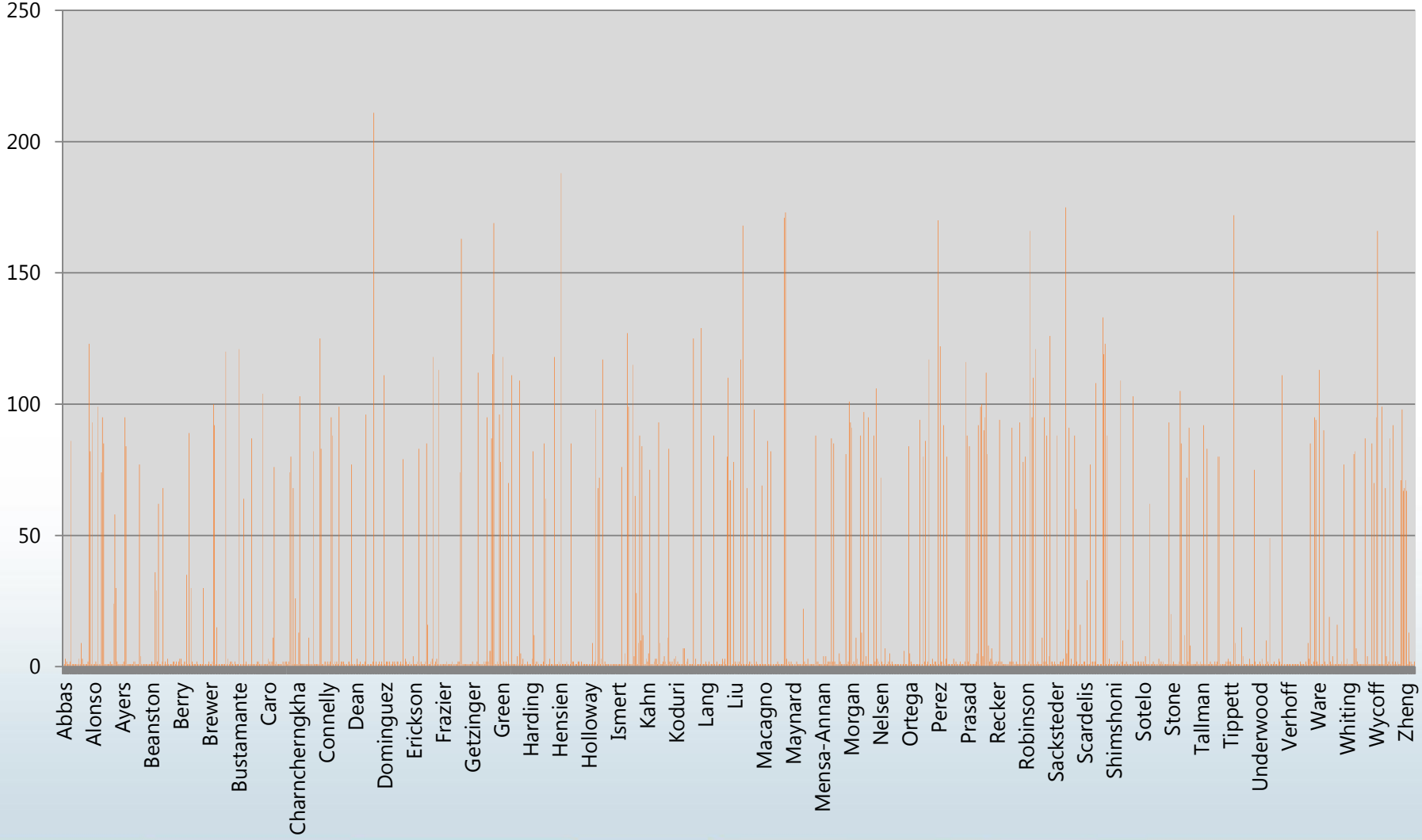
"…a histogram and associated density groups (collections)…"

Density: uniqueness of values with a set of data, calculated as (1/number distinct rows)

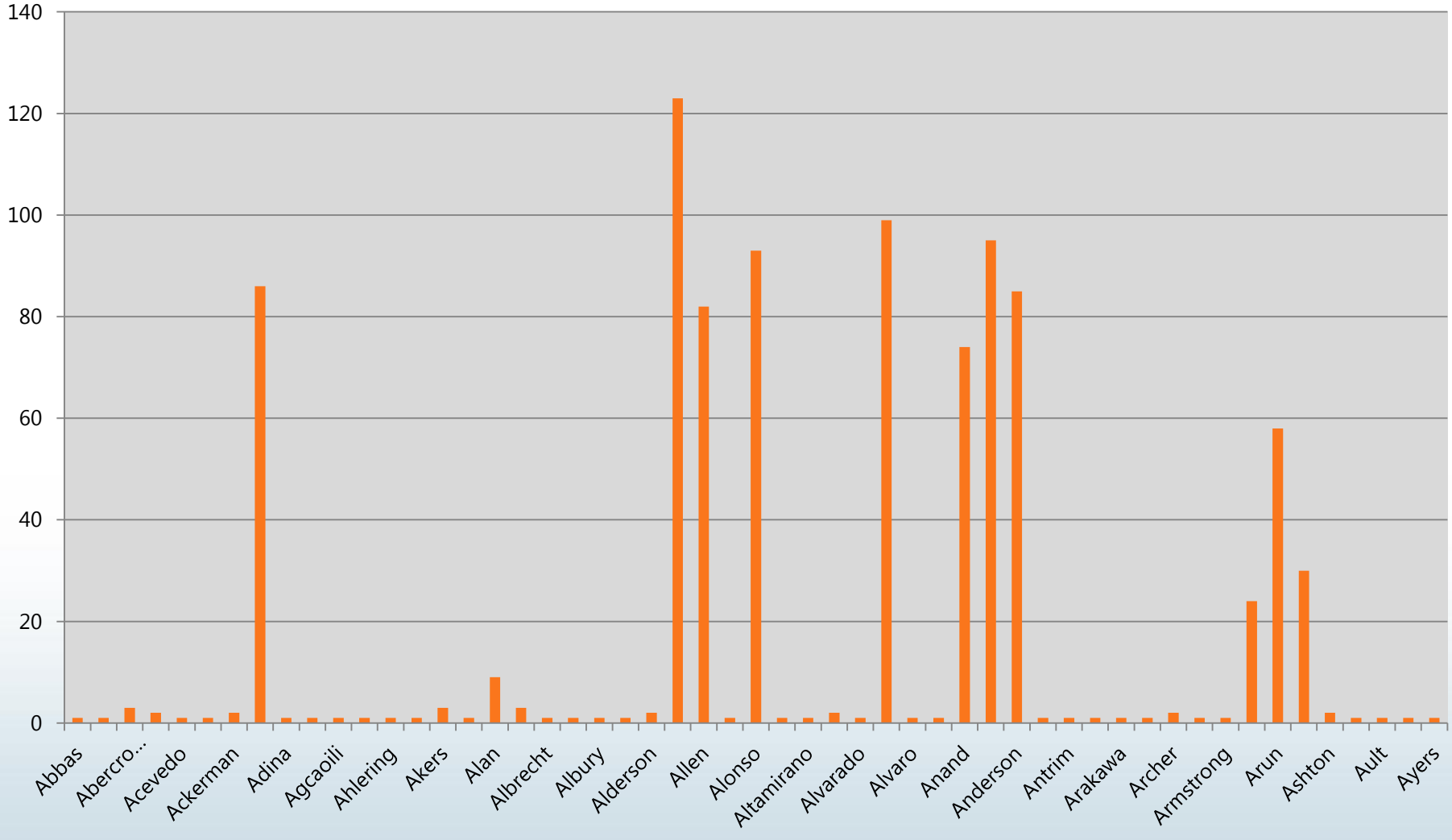
- 1/5000 = .0002 density

Histogram: represents the distribution of values for a set of data

Another way to think of a histogram



Another way to think of a histogram(2)



Statistics and the Optimizer

"The query optimizer uses this statistical information to choose the most efficient plan for retrieving or updating data."

Statistics affect what plan the optimizer chooses

There may be only a few ways to get it right, but thousands of ways to get it wrong

It's your job to set up the optimizer for success

Statistics (as defined by BOL)

"…a histogram and associated density groups (collections) over the supplied column or set of columns of a table or indexed view … The query optimizer uses this statistical information to choose the most efficient plan for retrieving or updating data. Up-to-date statistics allow the optimizer to accurately assess the cost of different query plans, and choose a high-quality plan."

Thank you

for attending this session and
the 2012 PASS Summit in Seattle