The E, the A and the V

How I Learned to Stop Worrying and Love the Hybrid

Evaluation Forms

- Don’t forget to fill out the evaluation forms
  - Speakers
  - Sponsors
  - Event

Transact-SQL Guru

- Core competencies
  - Algorithms
  - Query tuning
  - Database design
- Known for
  - Unorthodox ideas
  - Thinking outside the box
  - *"The guy other MVPs turn to for help"*
  - Passionate about the Community

Microsoft CERTIFIED Trainer

Microsoft CERTIFIED Solutions Expert Data Platform

MVP Microsoft Most Valuable Professional

PASS

redgate ingeniously simple tools

Microsoft

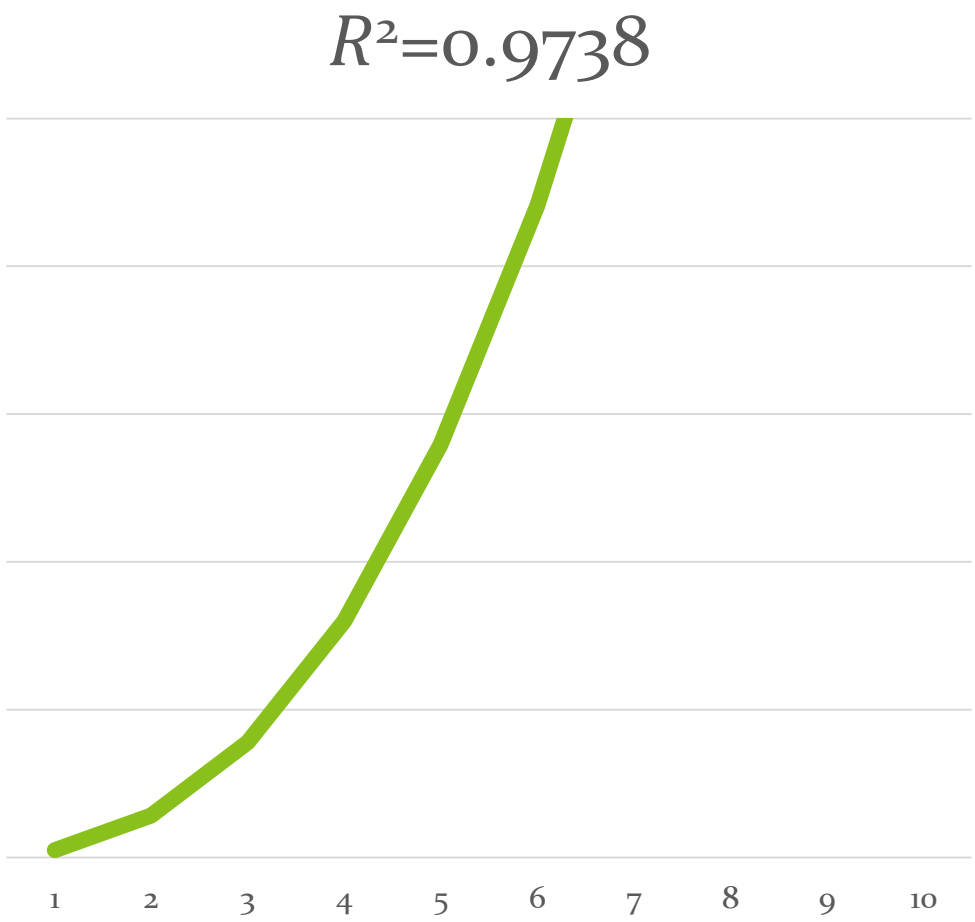
SolidQ

Content Disclaimer

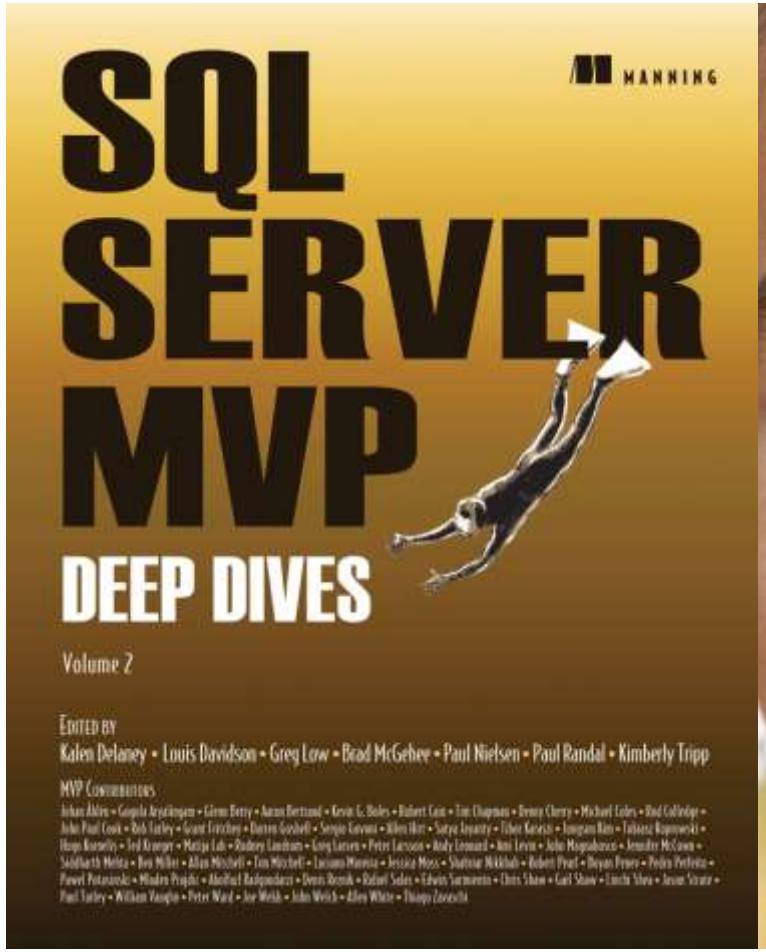
- Protecting privacy
  - Classified information
  - Data may be obfuscated
  - Diagrams may be conceptual
- Transactional data is 500 TB
  - Total database size is 1.5 PB
- Continuous improvements

Background

- Bad database design
- Slow algorithm
- Inefficient storage planning

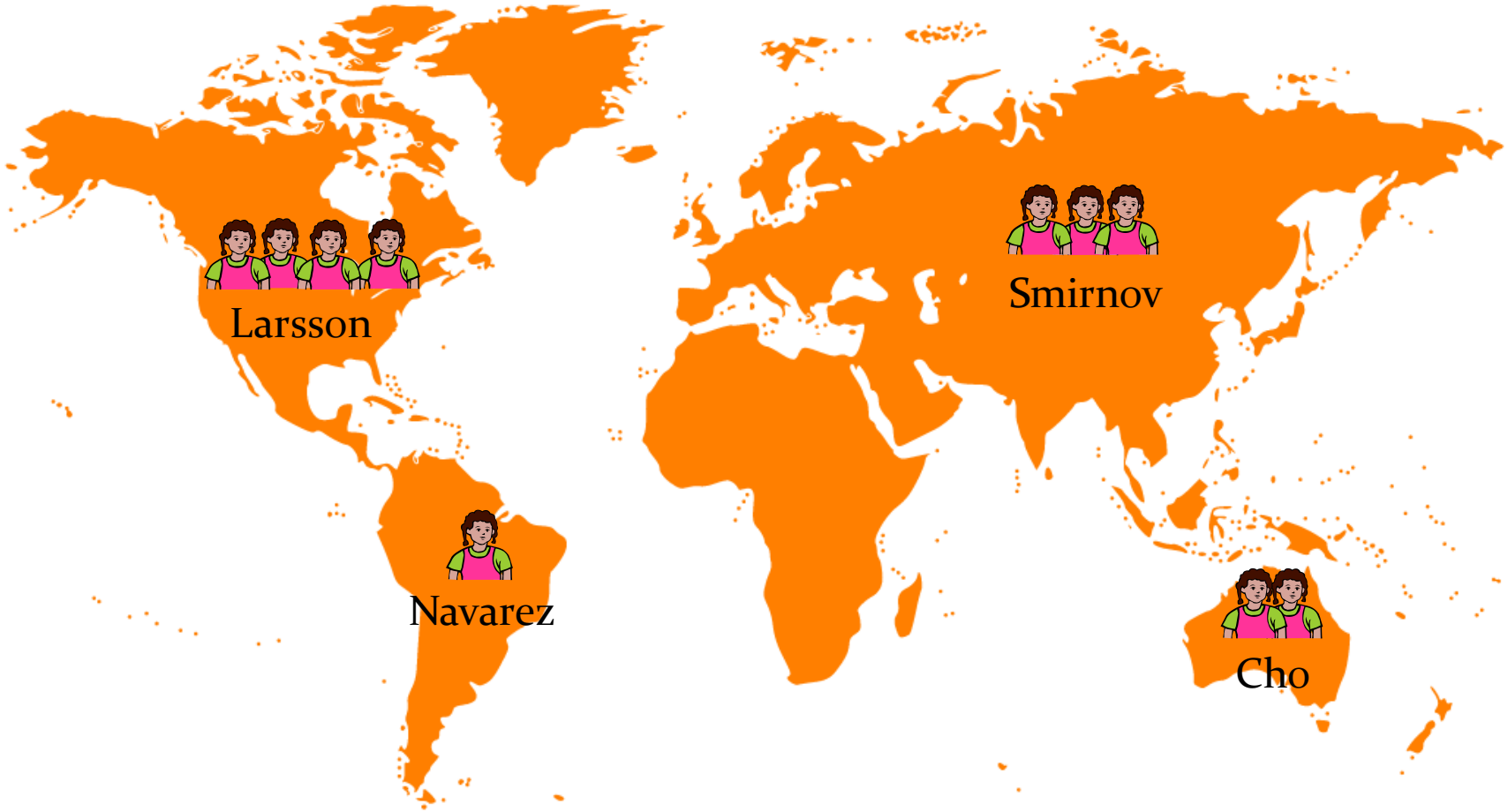


SQL Server MVP Deep Dives 2



http://www.manning.com/delaney

Where in the World is Isabelle?



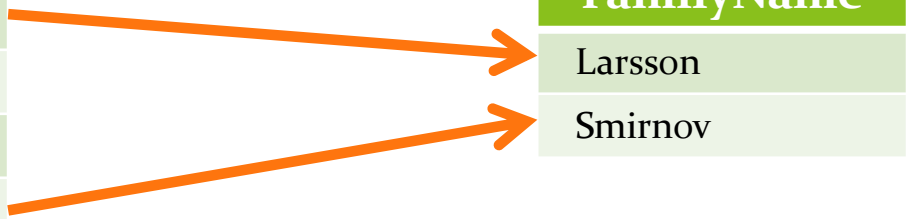
The Simplest Relational Division There Is

FamilyName	FirstName	Age
Navarez	Emilia	12
Larsson	Filippa	10
Larsson	Isabelle	6
Larsson	Samuel	4
Larsson	Cecilia	3
Smirnov	Isabelle	9
Smirnov	Alexander	7
Smirnov	Maxim	2
Cho	Aoi	6
Cho	Hiroto	4

```
SELECT FamilyName
FROM   dbo.Families
WHERE  FirstName = 'Isabelle'
```

Query Result for 1 Row and 1 Column filter

FamilyName	FirstName	Age
Navarez	Emilia	12
Larsson	Filippa	9.5
Larsson	Isabelle	6
Larsson	Samuel	4
Larsson	Cecilia	2.5
Smirnov	Isabelle	9
Smirnov	Alexander	7
Smirnov	Maxim	2
Cho	Aoi	6
Cho	Hiroto	4



FamilyName
Larsson
Smirnov

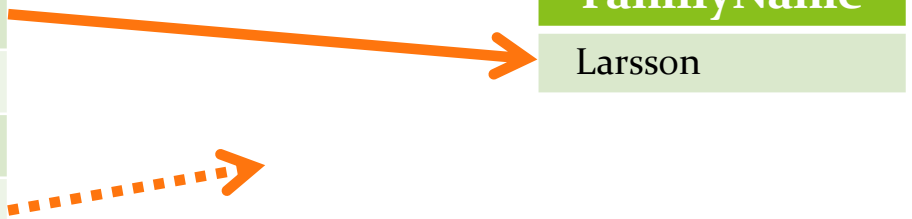
Two Column Filter

FamilyName	FirstName	Age
Navarez	Emilia	12
Larsson	Filippa	10
Larsson	Isabelle	6
Larsson	Samuel	4
Larsson	Cecilia	3
Smirnov	Isabelle	9
Smirnov	Alexander	7
Smirnov	Maxim	2
Cho	Aoi	6
Cho	Hiroto	4

```
SELECT FamilyName
FROM   dbo.Families
WHERE  FirstName = 'Isabelle'
       AND Age = 6
```

Two Column Filter

FamilyName	FirstName	Age
Navarez	Emilia	12
Larsson	Filippa	9.5
Larsson	Isabelle	6
Larsson	Samuel	4
Larsson	Cecilia	2.5
Smirnov	Isabelle	9
Smirnov	Alexander	7
Smirnov	Maxim	2
Cho	Aoi	6
Cho	Hiroto	4



FamilyName
Larsson

Unqualified vs Qualified Query Statement

```
SELECT FamilyName
FROM   dbo.Families
WHERE  FirstName = 'Isabelle'
       AND FirstName = 'Samuel'
```

```
SELECT FamilyName
FROM   dbo.Families
WHERE  FirstName = 'Isabelle'
       OR FirstName = 'Samuel'
```

Standard Query for Two Row Filter

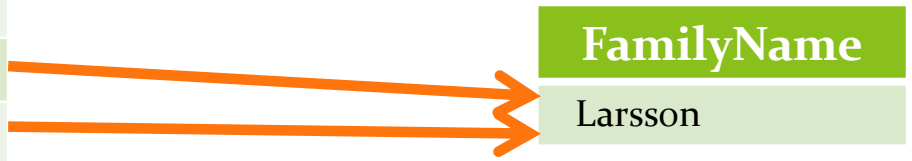
```
SELECT FamilyName
FROM   dbo.Families
WHERE  FirstName = 'Isabelle'

INTERSECT

SELECT FamilyName
FROM   dbo.Families
WHERE  FirstName = 'Samuel'
```

Two Row Filter Visualized

FamilyName	FirstName	Age
Navarez	Emilia	12
Larsson	Filippa	9.5
Larsson	Isabelle	6
Larsson	Samuel	4
Larsson	Cecilia	2.5
Smirnov	Isabelle	9
Smirnov	Alexander	7
Smirnov	Maxim	2
Cho	Aoi	6
Cho	Hiroto	4



FamilyName
Larsson
Larsson

Opinions About Standard EAV Data Model

- OTLT and EAV are “generic” approaches that are seductive to programmers, but are actually a bad idea for most databases. Resist the temptation! – *Tony Andrews 2004*
- EAV structures are like drugs; in small quantities and used in the proper circumstances, they can be beneficial. However, too much will kill you. – *Thomas (Seven24) 2009*
- The main characteristic of EAV schemas is that they have a “store everything, query nothing” property. You can store whatever you like, because it's so generic. Because it's so generic, you can't query it, because nothing really ever means something. – *Roland Bouman 2009*
- The EAV data model is an anti-pattern – *Kennie Pontoppidan 2014*

EAV Data Model In Theory

- Entity
  - The item described by the *Attribute* and the *Value*.
- Attribute
  - The information type describing the *Entity*.
- Value
  - The value of the *Attribute*.

EAV Data Model In Practice

Entity		Attribute	Value
	Antonija Mišura	Body	Athletic
	Antonija Mišura	Career	Pro
	Antonija Mišura	Gender	Female
	Antonija Mišura	Country	Croatia
	Jason Momoa	Career	Actor
	Jason Momoa	Gender	Male
	Jason Momoa	Hair	Dreadlock
	Kate Beckinsale	Career	Actress
	Kate Beckinsale	Gender	Female
	Kate Beckinsale	Married	Yes

Communicating With the EAV Data Model

- Natural English queries
  - Give me all documents where
    - Manager is John Doe
    - Manager is Jane Doe and project is World Domination
- Machine equivalent
  - SELECT * FROM EAV WHERE
    - Manager=John Doe
    - Manager=John Doe,Jane Doe|Project=World Domination
    - Manager=John Doe|Manager=Jane Doe|Project=World Domination

Demonstration

Let Us Do a Basic Search On the EAV Data Model

Implementations and PoC's

- Japanese-Swedish telecom company (version 1)
  - About 100 million rows
    - Original 1,300 seconds
    - Rewrite 2 seconds
- County medical facility (version 2)
  - Projected one billion patients notes
    - Original 65 seconds (200 million rows, 300 000 reads)
    - Rewrite 0.005 seconds (1 billion rows, 4 reads)

Hybrid Data Model Design Goals

- Flexible
  - Extend Natural English Query
- Versatile
  - Introduce quality mindset
- Smaller
  - Use less storage space
- Faster
  - Make algorithm predictable

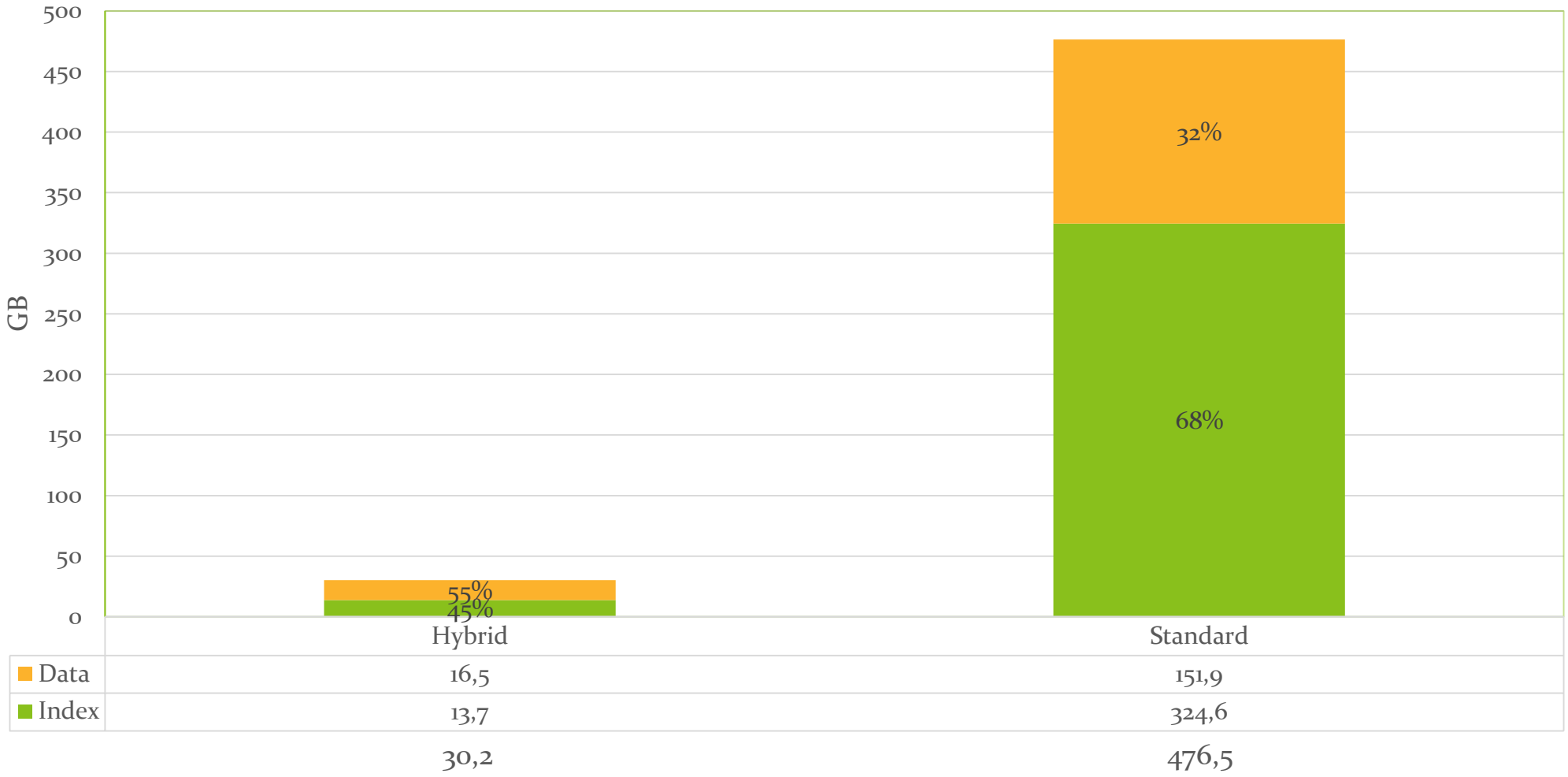
Communication With the EAV Data Model

- Extended natural English queries
  - Give me all documents where
    - Manager is either John Doe *or* Jane Doe
    - Manager is either John Doe *or* Jane Doe, *and* project is World Domination
    - Manager is either John Doe *or* Jane Doe, *and* project status is either Completed, Assigned *or* Approved
    - Project city begins with “Lon”.
  - Give me all persons where
    - Gender is Female *and* Pet is either Cat, Parrot *or* Rabbit, *and* Driver’s License is True.
    - Ignore persons where Education is University, *or* City is either Stockholm *or* London.

Introduce Quality Mindset

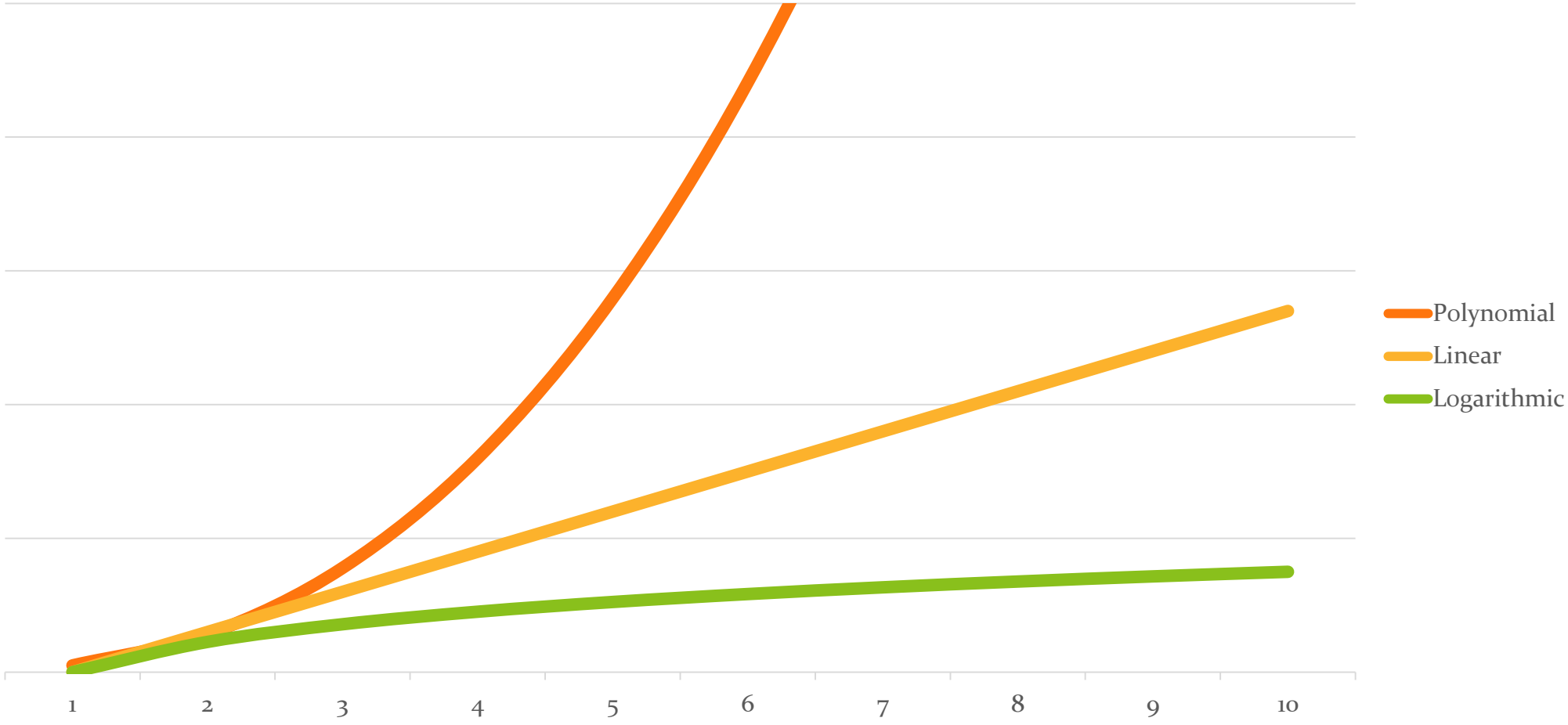
- Status flag
  - Approved
- Naïve heuristics (v2)
  - Better approximations
- Statistics snapshot (v3)
  - Up to date statistics

Use Less Storage Space



	Hybrid	Standard
Data	16,5	151,9
Index	13,7	324,6
	30,2	476,5

Algorithm Complexity



What are the Necessary Steps?

Required Changes

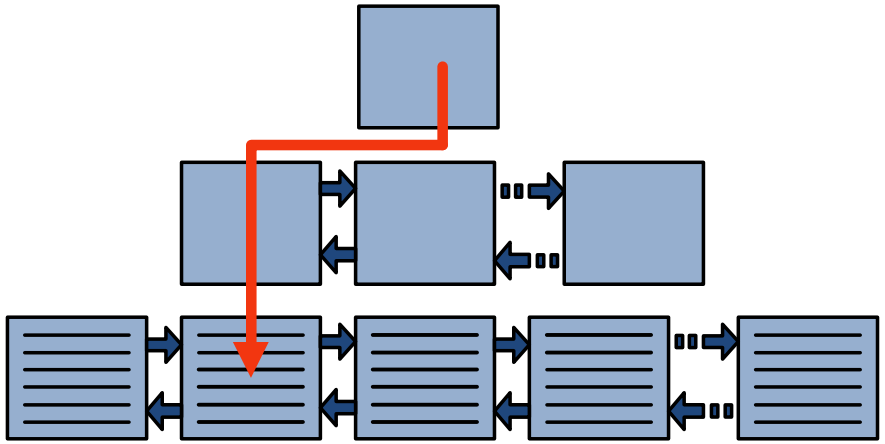
- Indexing strategy
- Normalization
- Centralized search engine
- Short-circuit logic
- Heuristics

Edition Perks

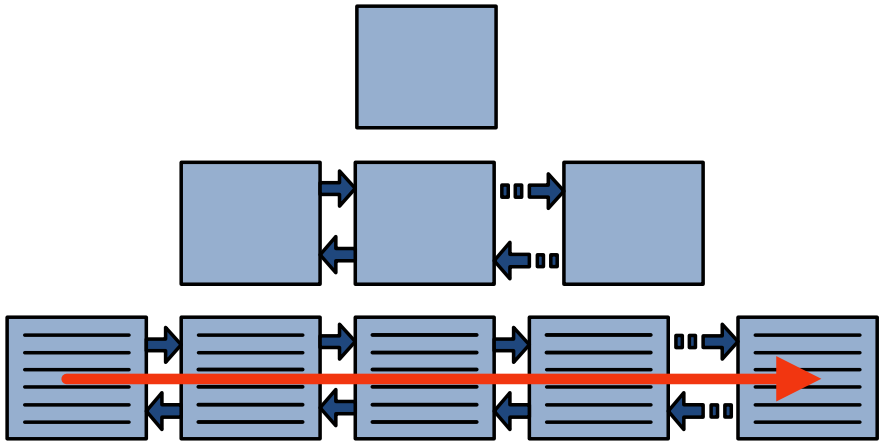
- *Data compression*
- *Partitioning*

Index Access Methods

Seek

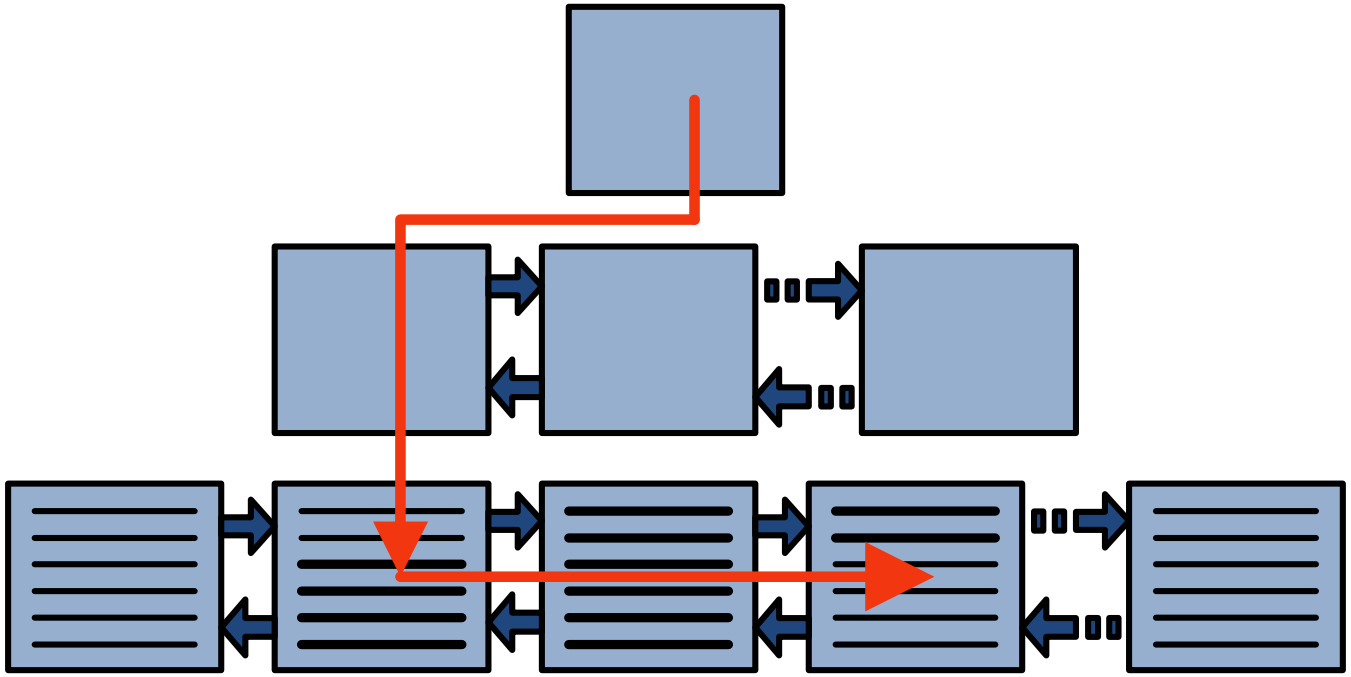


Scan



Clustered Index Seek and Ordered Partial Scan

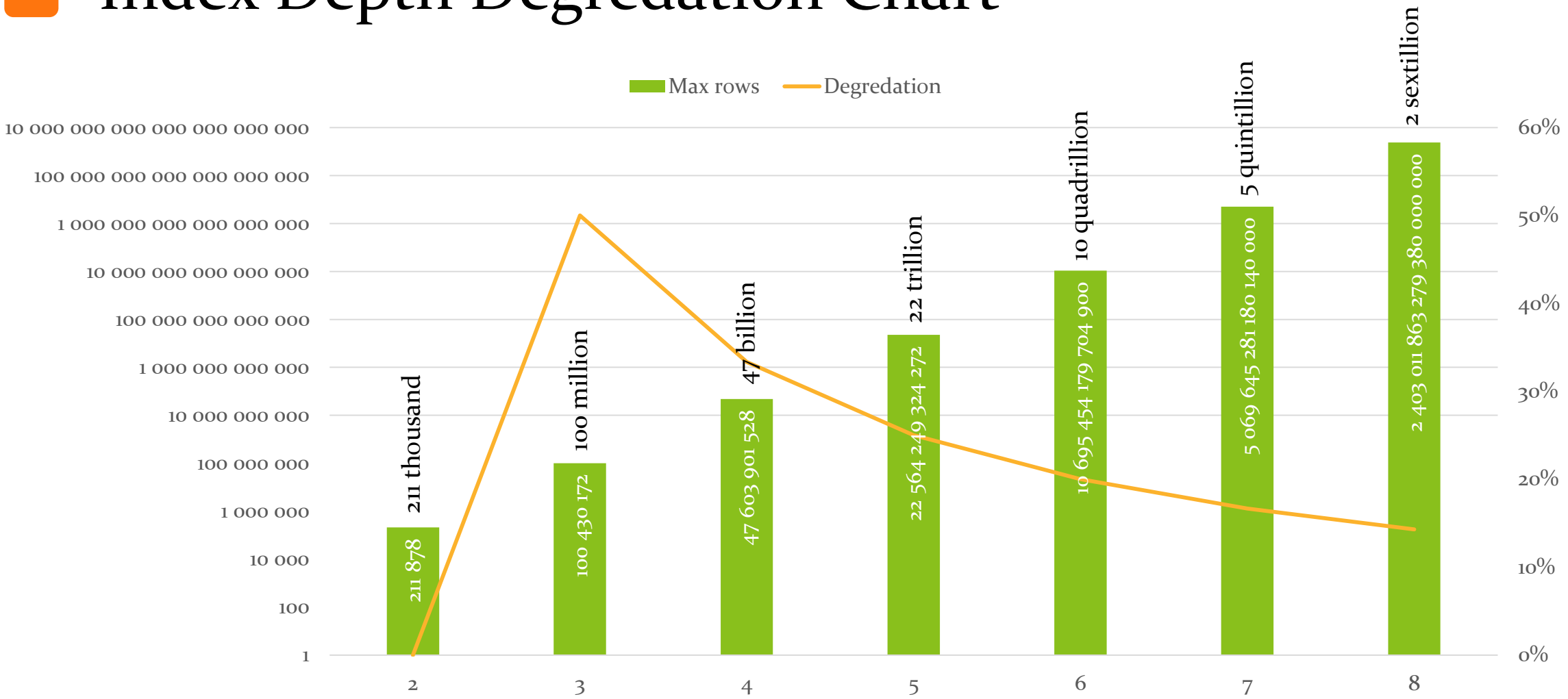
- An efficient clustered index
  - NUSE



Uncompressed Clustered Index Math

- Row size 10 bytes
- Leaf nodes
  - 8,060 / (10 + 8) = 447 rpp
- Key size 9 bytes
- Non-leaf nodes
  - 8,060 / (9 + 8) = 474 rpp
- Additional storage overhead
  - 0.2%
- Leaf level
  - 1,000,399,680 rows
  - 2,238,031 pages
- Intermediate levels
  - 4,722 pages
  - 10 pages
- Root level
  - 1 page

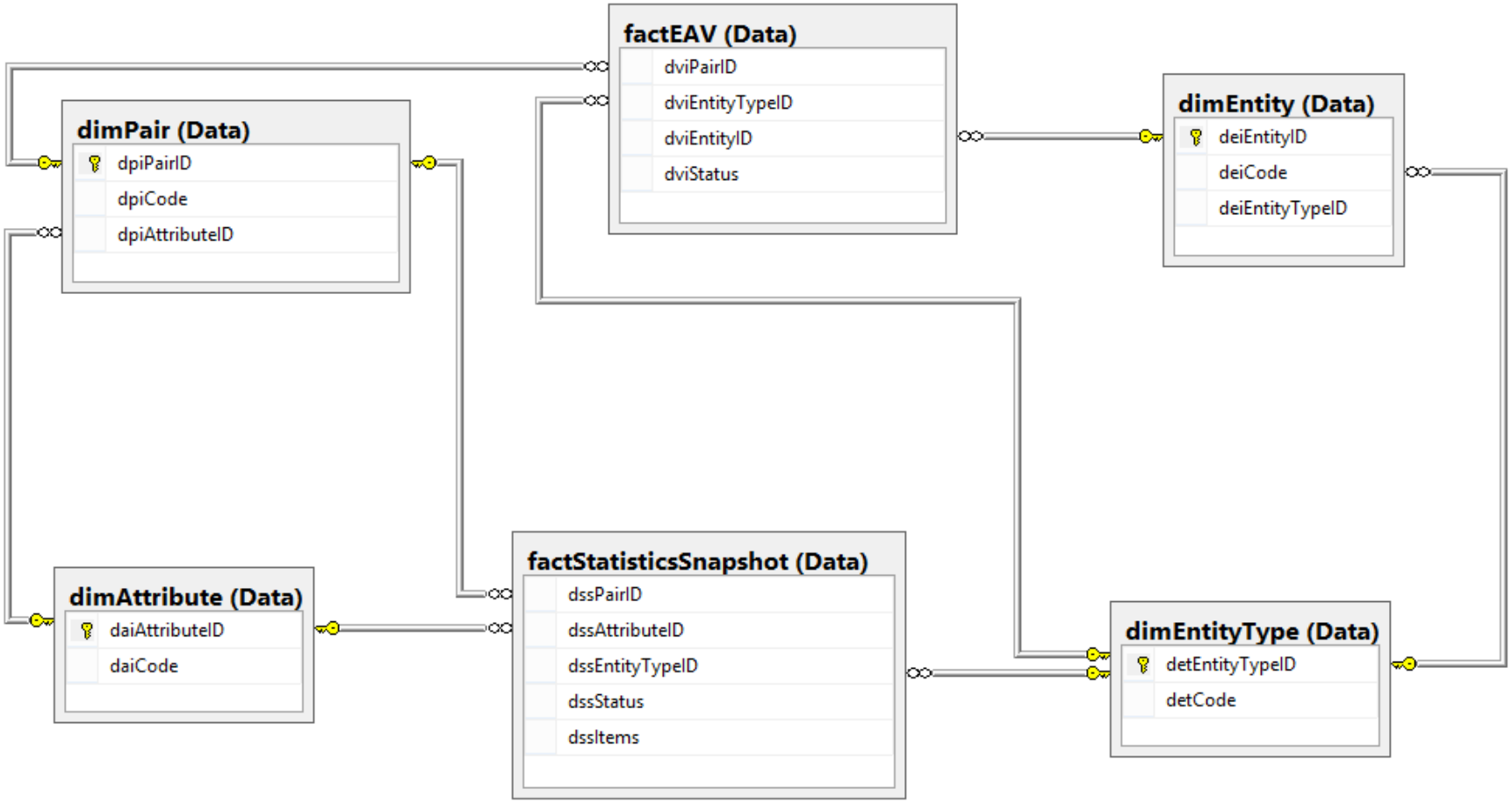
Index Depth Degredation Chart



There Is No Such Thing As Free Lunch

- Clustered Index Insert

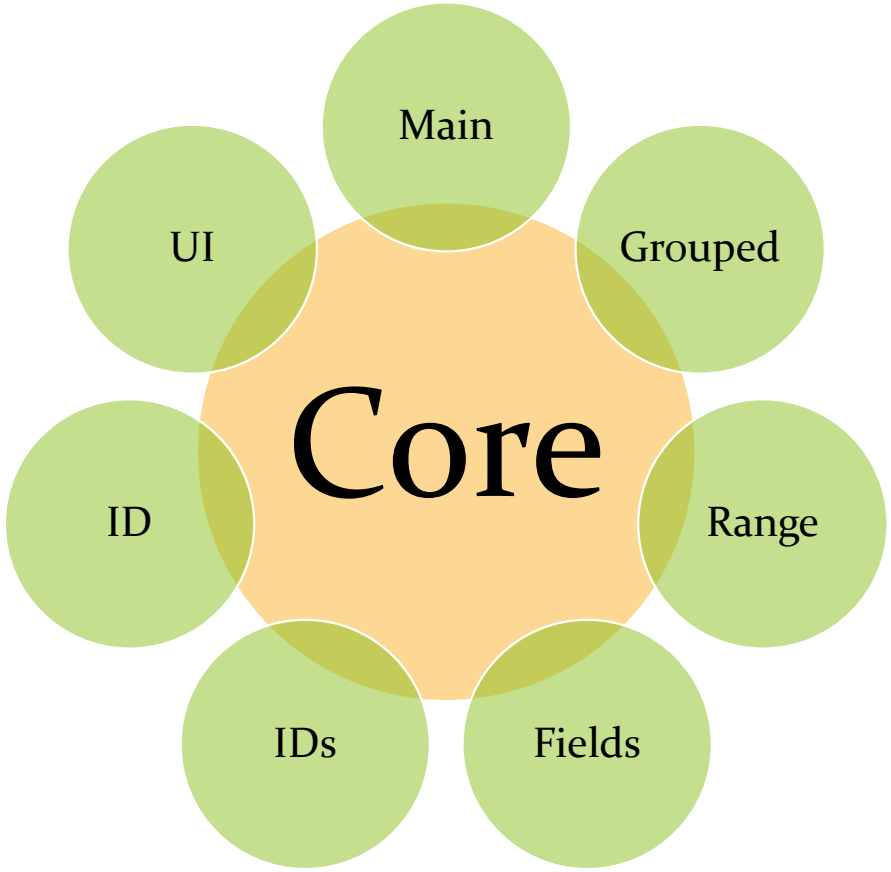
Hybrid Data Model Schema



Enterprise Edition Features

- Data Compression
- Partitioning

Centralized Search Engine



- Differentiate between
  - Selection (σ)
  - Projection (π)

Attributes Domain

- Single attribute
  - SSN=2014-03-03-1234
- Multiple attributes (v1)
  - Education=Candidate|Jurisdiction|City=Los Angeles
  - Education=PhD|Jurisdiction=*|City=Los Angeles
- Multi-valued attributes (v2)
  - Membership=Gold,Silver|Debit=Yes
- Duplicate attributes (v3)
  - Membership=Gold|Membership=Silver|Debit=Yes

How to Tokenize the Fact Pairs

```
WITH cteList(ListNumber, Attribute, LowerLimit, UpperLimit)
AS (
        SELECT          DENSE_RANK() OVER (ORDER BY l.n) AS ListNumber,
                        RTRIM(LTRIM(COALESCE(a.n.value('(a[1])', 'VARCHAR(48)'), ''))) AS Attribute,
                        RTRIM(LTRIM(COALESCE(b.n.value('(b[1])', 'VARCHAR(512)'), ''))) AS LowerLimit,
                        RTRIM(LTRIM(COALESCE(b.n.value('(b[2])', 'VARCHAR(512)'), ''))) AS UpperLimit
        FROM            (
                                VALUES  (CAST('<l>' + REPLACE(@AttributeList, '|', '</l><l>') + '</l>' AS XML))
                        ) AS x(n)
        OUTER APPLY     x.n.nodes('(l)') AS l(n)
        OUTER APPLY     (
                                VALUES  (CAST('<a>' + REPLACE(l.n.value('(.)', 'VARCHAR(MAX)'), '=', '</a><a>') + '</a>' AS XML))
                        ) AS a(n)
        OUTER APPLY     a.n.nodes('(a[2])') AS p(n)
        OUTER APPLY     (
                                VALUES  (CAST('<v>' + REPLACE(p.n.value('(.)', 'VARCHAR(MAX)'), ',', '</v><v>') + '</v>' AS XML))
                        ) AS w(n)
        OUTER APPLY     w.n.nodes('(v)') AS v(n)
        OUTER APPLY     (
                                VALUES  (CAST('<b>' + REPLACE(v.n.value('(.)', 'VARCHAR(MAX)'), ':', '</b><b>') + '</b>' AS XML))
                        ) AS b(n)
)
```

http://blogs.solidq.com/SQL-Server-pa-svenska/Post.aspx?ID=16

Direct Short-Circuit Logic

- Positive search attributes
  - City=New York,New Delhi|Approved=Yes
- Negative search attributes
  - Self-employed=Yes|City=New%

Indirect Short-Circuit Logic

- Positive search attributes
  - City=New York,New Delhi,London|Approved=Yes
- Negative search attributes
  - Self-employed=Yes|City=New%

Naïve Heuristics (v2)

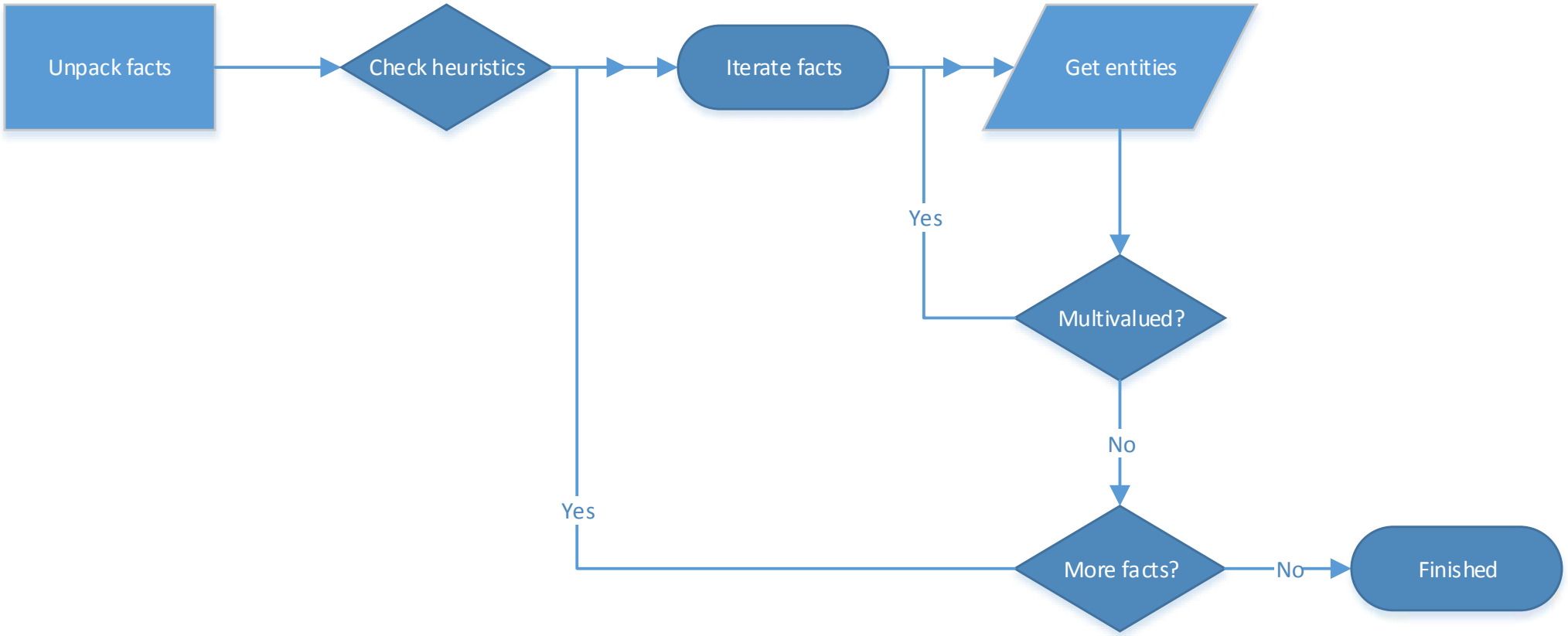
- Experience based
- Fallible
- Limited
- Facilitate quick estimates

ValueID	Code	AttributeID
1	Male	1
2	Yes	2
3	Female	1
4	Copenhagen	3
5	Yes	4
6	Sweden	5
7	Main Street	6
8	2014-03-01-1234	19

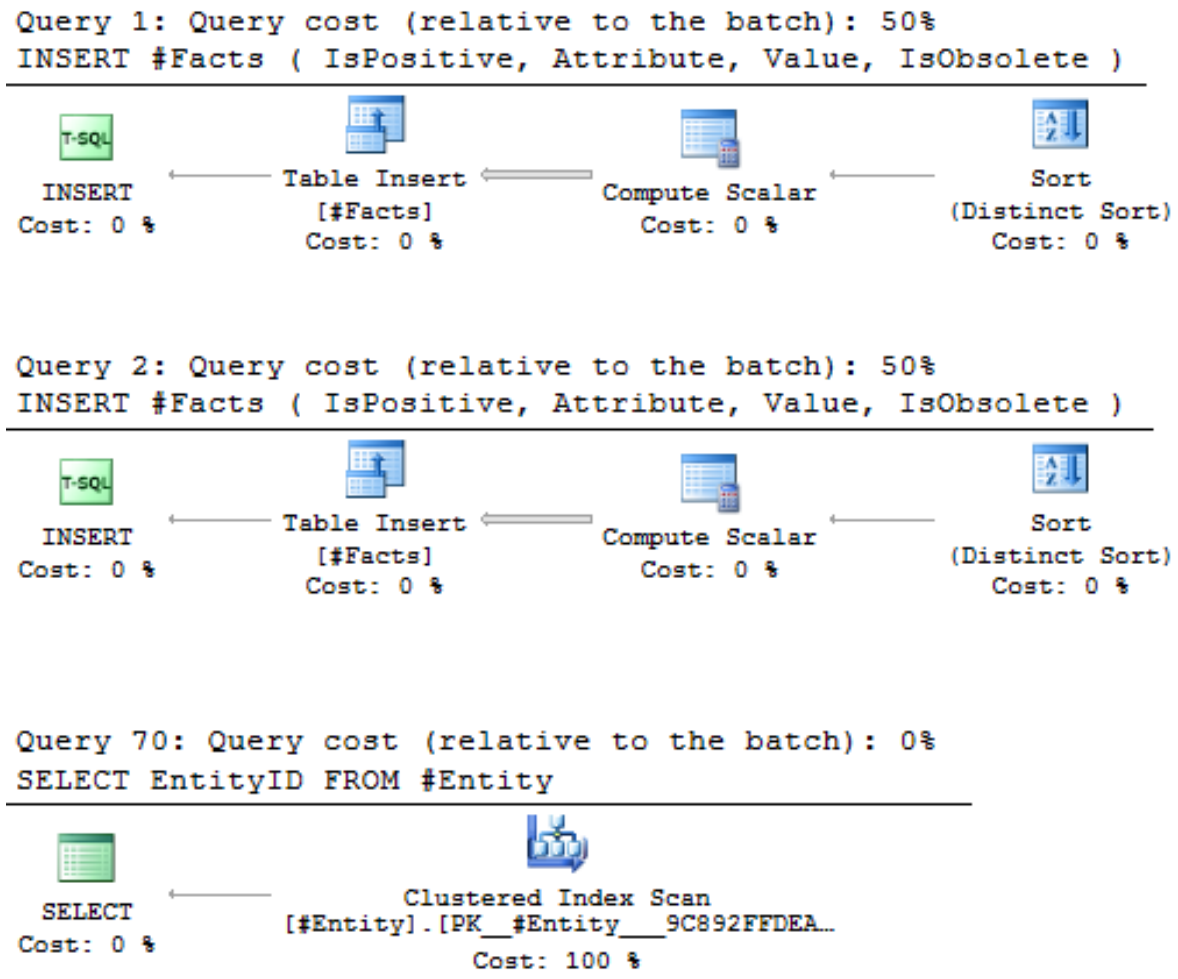
Statistics Snapshot (v3)

ValueID	AttributeID	EntityTypeID	Status	Items
28 111	1	1	1	12 300
29 069	2	1	1	74 437
1	2	1	1	1 222 911
81	2	1	0	7 845
245	2	1	0	11 004
436	2	1	1	2
680	2	1	1	412 665
791	2	1	1	37 128

Stored Procedure Algorithm



Execution Plans



Demonstration

I Feel The Need… The Need For Speed!

Help Me Make a Difference

- Vote on Microsoft Connect and make your voice heard!
  - http://connect.microsoft.com/SQLServer/feedback/details/670531/move-t-sql-language-closer-to-completion-with-a-divide-by-operator
  - http://bit.ly/u3318c

Questions?

Tak fordi du lytter

Bedankt voor het luisteren

Thank you for listening

Tack för att ni lyssnade

Köszönöm, hogy meghallgattak

Hvala za poslušanje

Social Media

SwePeso

http://www.sqltopia.com

http://weblogs.sqlteam.com/peterl

http://blogs.solidq.com/SQL-Server-pa-svenska

swepeso@sqltopia.com

plarsson@solidq.com

Peso and *Pesomannen* are obsolete