Why Does the Db2 Optimizer Need a Data Scientist?

Sheryl Larsen

Worldwide Db2 Z Solutions

BMC

AI will be Pervasive by 2020

By 2020, **30%** of data centers that fail to apply AI and machine learning effectively in support of enterprise business will cease to be operationally and economically viable.

Gartner Report: The IT Implications of the 2018 CIO Survey for I&O Leaders
By Dave Russell, Hank Marquis Published : 8 March 2018


D
b
2

The Changing Mainframe SQL Optimization Landscape

Mainframe organizations <u>must</u> take a new approach aimed at self-tuning SQL.

You need:

- ✔ Target SQL that fits the mold
- ✔ Proactive data gathering
- ✔ Model behavior
- ✔ Confirm change = improvement
- ✔ Deploy

The Db2 Optimizer
How Does it Decide so Fast?

Good Input

- 35 years of catalog statistics refinement
- Ability to use some real time information
- Ability to refine scope of data collection - STATSFEEDBACK

Cost-based Smarts

- 35 years of algorithm refinement
- Creates a cost model for every query
- **Defaults** are used when query values are **unknown**

How close does the optimizer get with '?' or ':hv'?

The Trillions of Optimizer Cost-based Results

Good for Everybody

Great for a Few

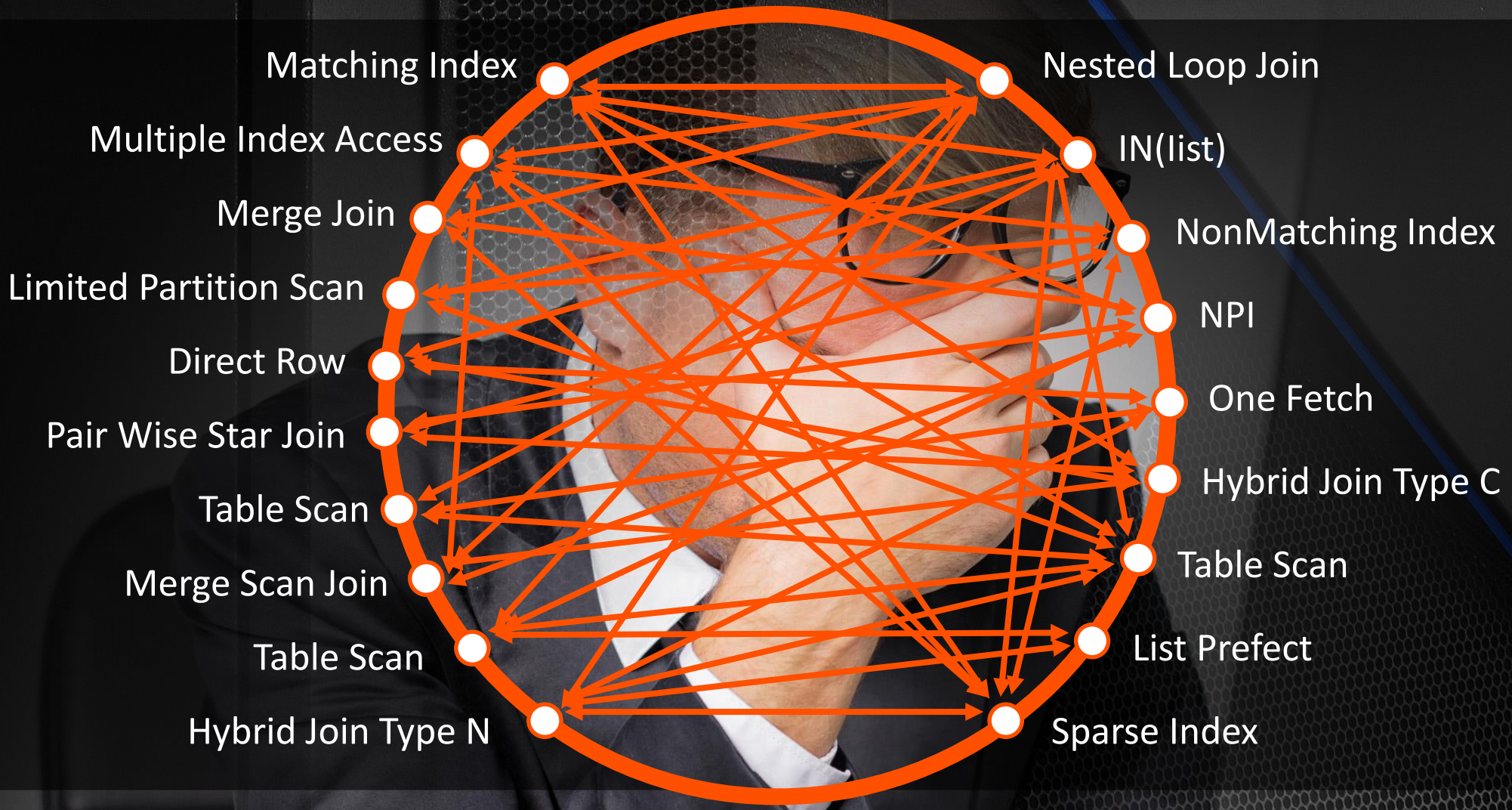
Default Statistics

WHERE >? WHERE BETWEEN ? AND ?

COLCARDF	Factor for <, <=, >, >=	Factor for LIKE or BETWEEN
>=100,000,000	1/10,000	3/100,000
>=10,000,000	1/3,000	1/10,000
>=1,000,000	1/1,000	3/10,000
>=100,000	1/300	1/1,000
>=10,000	1/100	3/1,000
>=1,000	1/30	1/100
>=100	1/10	3/100
>=2	1/3	1/10
=1	1/1	1/1
<=0	1/3	1/10

How Close to Reality?

There Are Many Ways to Get to Your Data



The Answer: Personalize Your Optimizer's Decisions

Technology needed:

- Learns patterns from workload data collected in your unique operating environment

- Uses derived insight in determining optimal access paths for SQL statements

Built on top of the IBM Machine Learning for z/OS (MLz) stack

Leverages MLz services ***without requiring data scientist support*** –
Db2 generates model training data, deploys and monitors
and retrains models via MLz services

- Db2ZAI product ID: 5698-CGN

IDEA: Augment the Db2 Z Optimizer with AI/Machine Learning!

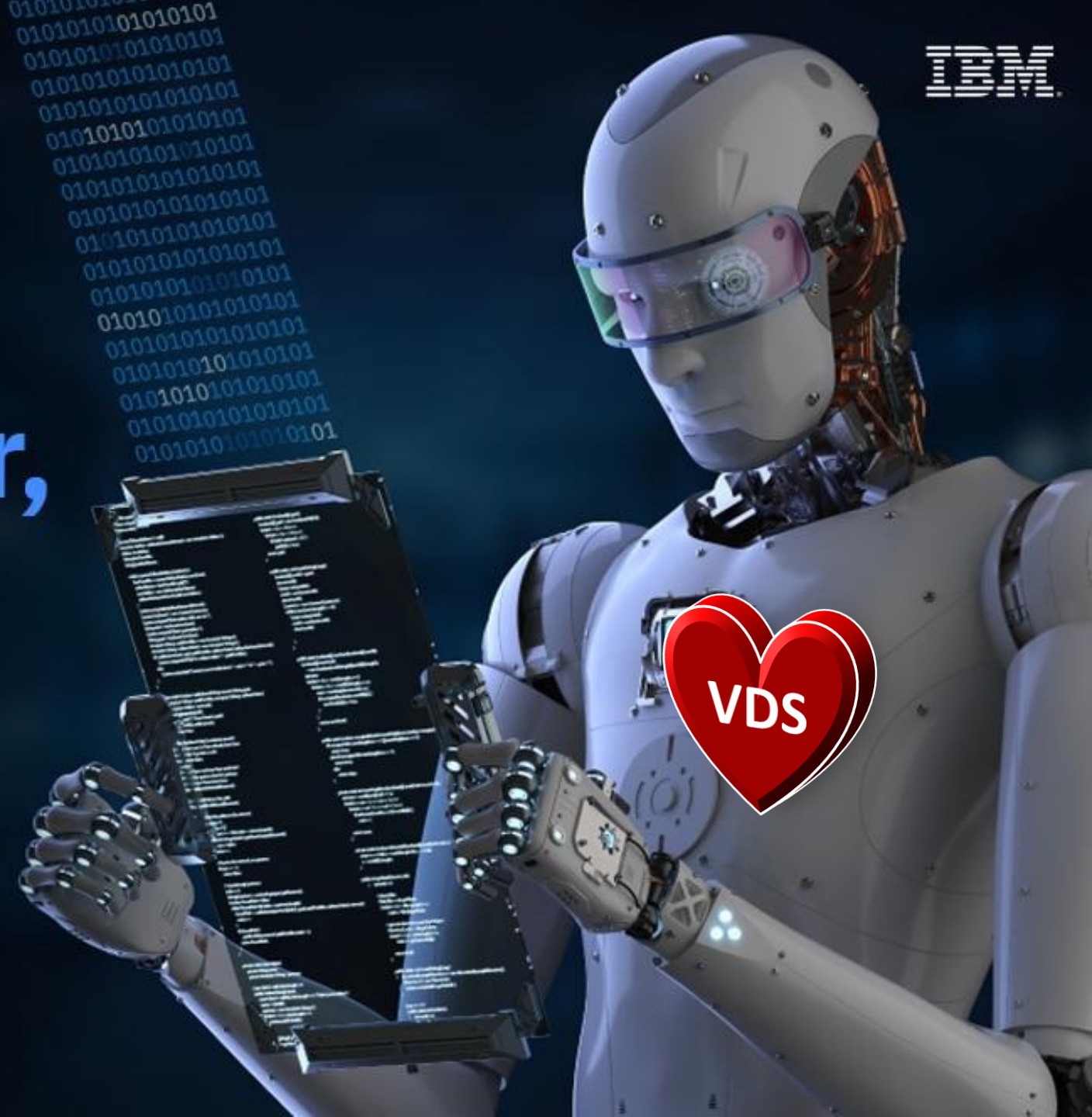
1. Correct estimates used for :hv and ?
2. Add OPIMIZE FOR n Rows when # of rows fetched is learned
3. Examine Sort behavior to optimize memory usage
4. Optimize parallelism in packages using history

The Db2 Z Optimization Team Took Action: Announcing IBM Db2 AI for z/OS 1.1 + Announcing IBM Db2 AI for z/OS 1.1.0.1 + Announcing IBM Db2 AI for z/OS V1.2!

IBM Db2® AI for z/OS® aka Db2ZAI

IBM **Db2**

AI Makes Db2 **Better, Smarter, Faster**



IBM Db2 AI for z/OS

#Db2ZAI

VDS – Virtual Data Scientist

Has the data

- Catalog statistics
- Deep execution statistics
- History

Knows which algorithm to use

- Classification for known patterns
- Linear Regression for Date/Time sequencing
- Models for random behavior

Learns from modeling and scoring

- Watches 100 executions

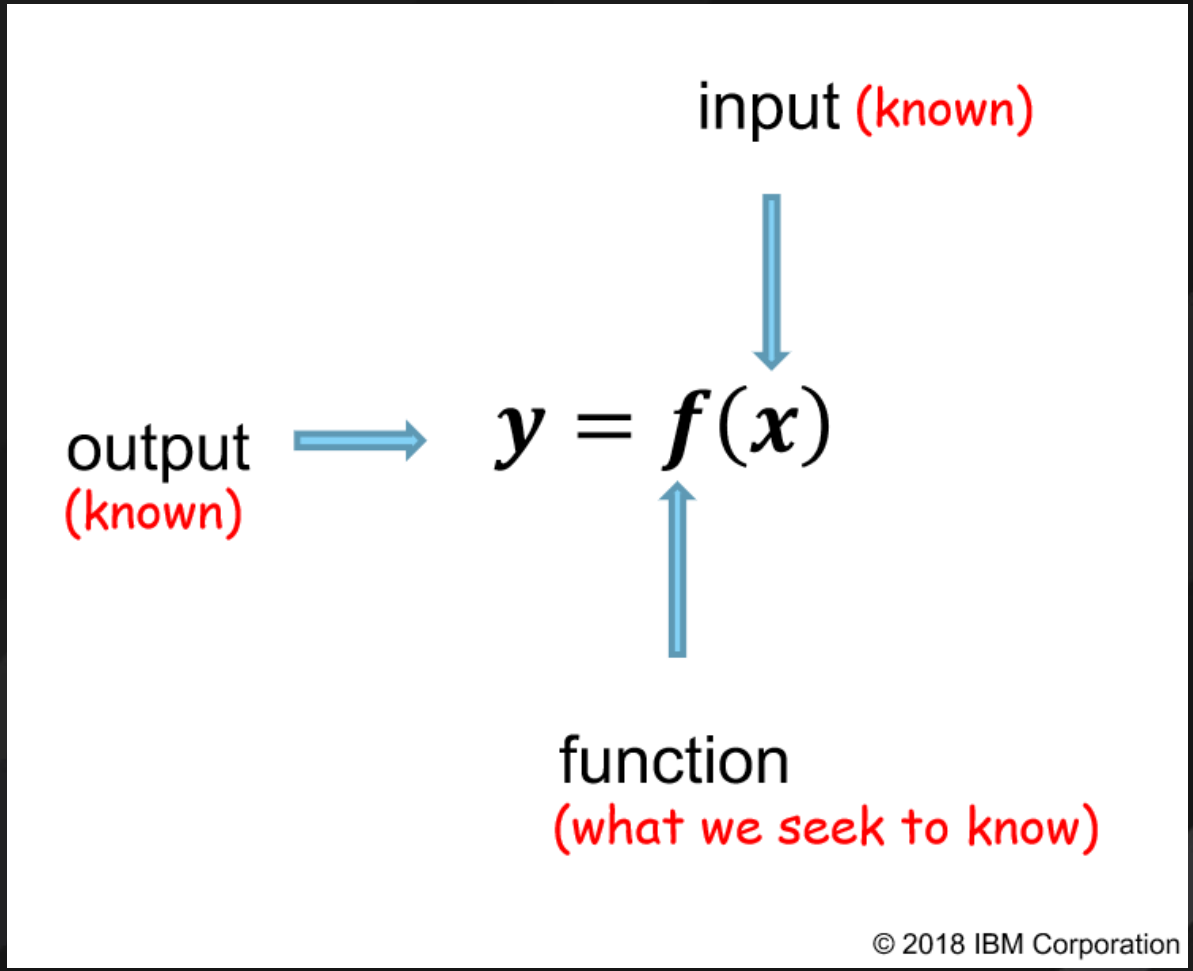
Provides solutions

- A list of ready packages
- Db2ZAI SQL Performance dashboard

Cleans up after itself

- Keep models current and removes old behavior

Supervised Learning

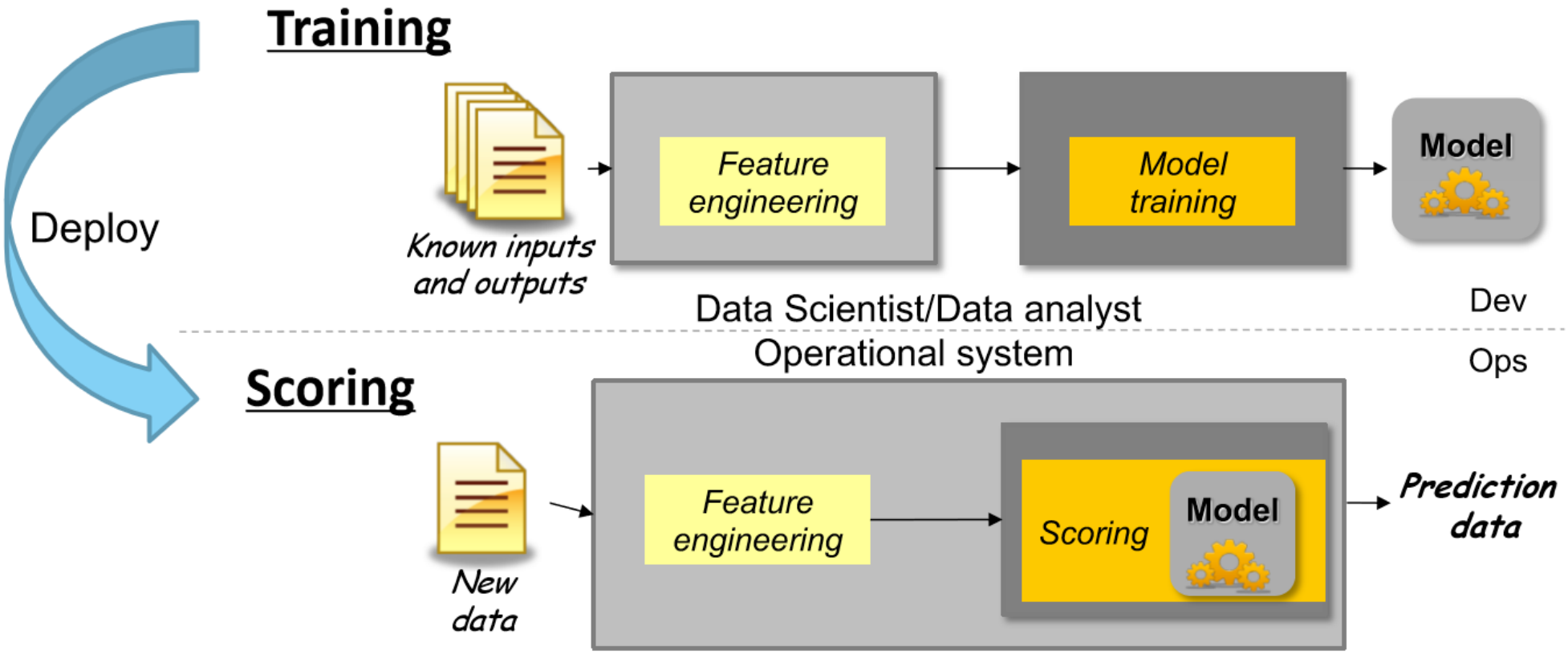


Catalog Stats +
Other Internal Stats

Last execution results
+ Tons of history

Db2 optimizer: predict next host variable value? Or cost estimate for an access path.

Use predictive models: training, deployment, scoring - IBM



Feature = more useful info for prediction

Feature = inputs to Model for scoring SQL

Fill in Unknown Values - :hv or ?

Customized Filter Factors

For STATIC use REBIND

For DYNAMIC uses PREPARE

PACKAGE Selection Screen

INCLUDE/EXCLUDE

Recommended List

Learn from the workload .....

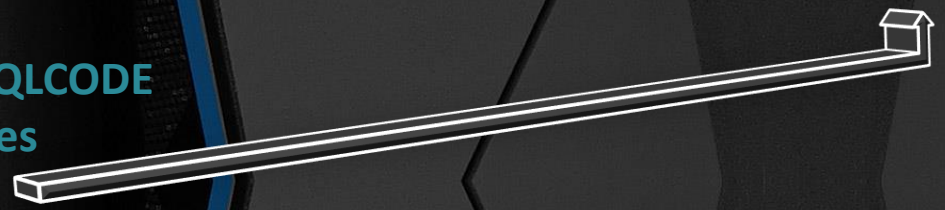
```
DECLARE   C_BRWZUM3 CURSOR FOR
SELECT   COL1, COL2, COL3, COL4,
FROM   BRWZUM
WHERE   ((COL1 = :COL1-LAST
   AND   COL2 = :COL2-LAST
   AND   COL3 = :COL3-LAST
   AND   COL4 > :COL4-LAST )
OR
      (COL1 = :COL1-LAST
   AND   COL2 = :COL2-LAST
   AND   COL3 > :COL3-LAST )
   OR
      (COL1 = :COL1-LAST
   AND   COL2 > :COL2-LAST )
   OR
      (COL1 > :COL1-LAST))
ORDER BY COL1,COL2,COL3,COL4
```

Applies To

Any query with :hv or ?

Predict # of Rows Qualifying

OPTIMIZE FOR n ROWS



Input

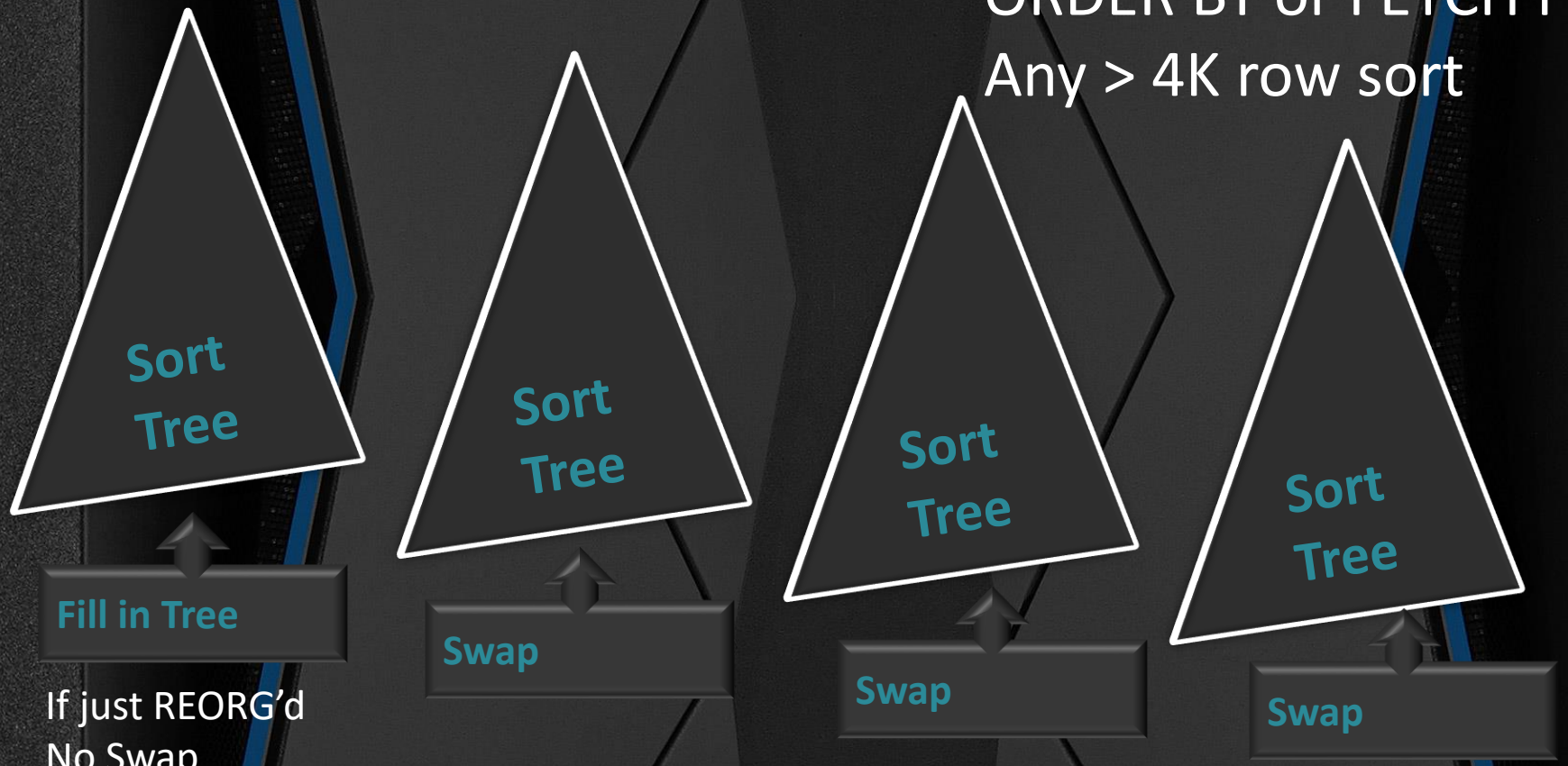
Track last fetched + SQLCODE
Repeat 100 times
Take AVG #

Applies To

Queries qualifying many rows,
But retrieving only a few

Optimize Sort Tree Usage and Memory

SQL with DISTINCT
ORDER BY or FETCH FIRST large rows
Any > 4K row sort



If just REORG'd
No Swap

Optimize Parallelism in non-OLTP Queries

DEGREE = ‘ANY’

DSNZPARM CDDSSRDEF = ‘ANY’

Input

Transactions > 120ms

Never < 10ms

Output

Reduced ELAPSED

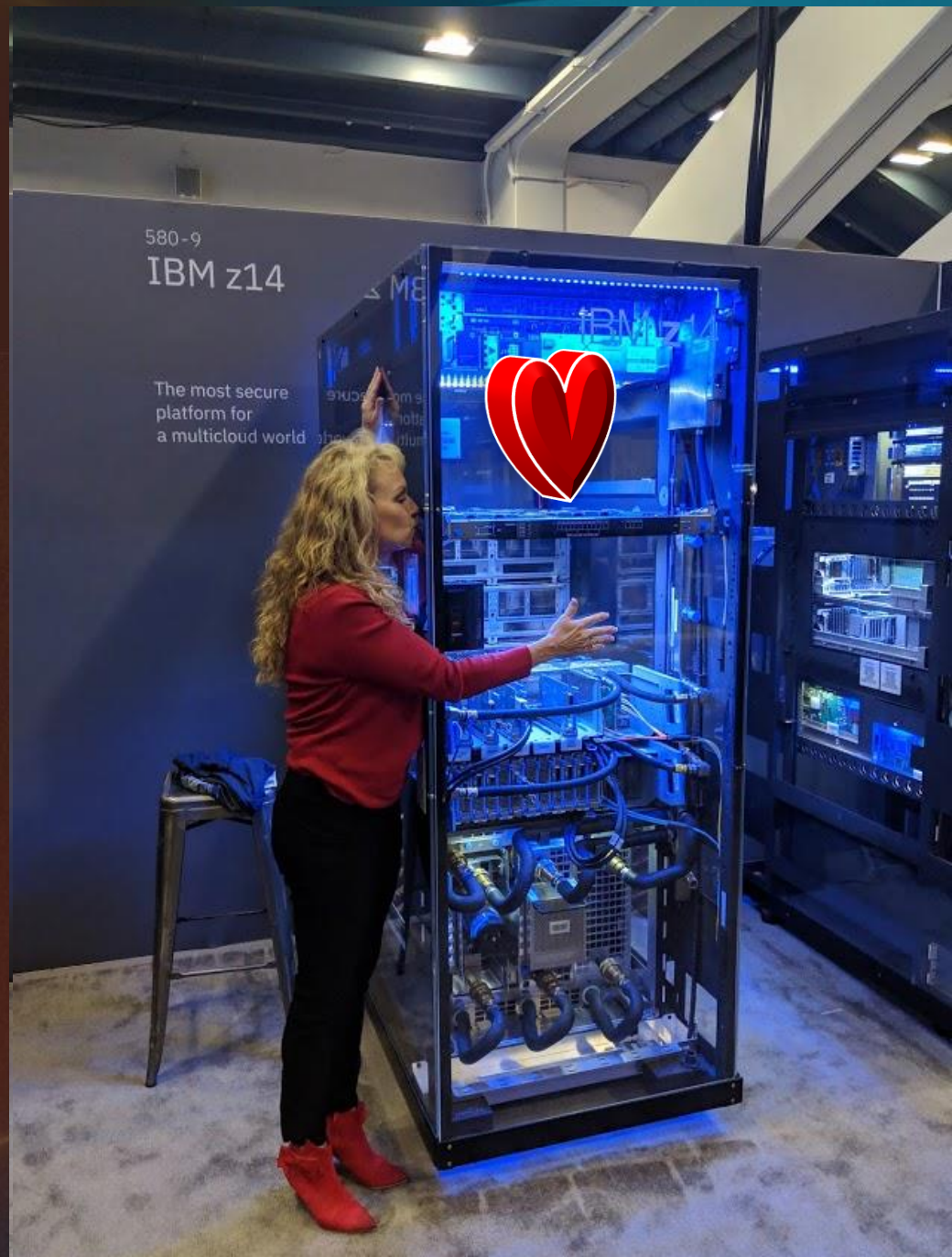
Slight CPU Increase

Db2ZAI: Augments the Db2 Z Optimizer with AI/Machine Learning!

1. Fill in “unknown” values in queries – Use Classification, Linear Regression and Model random behavior to correct estimates
2. Predict number of rows processed and add OPIMIZE FOR n = Optimal Rows
3. Examine Sort behavior to optimize memory usage
4. Optimize Parallelism in non-OLTP packages

Imagine Running with A VDS!

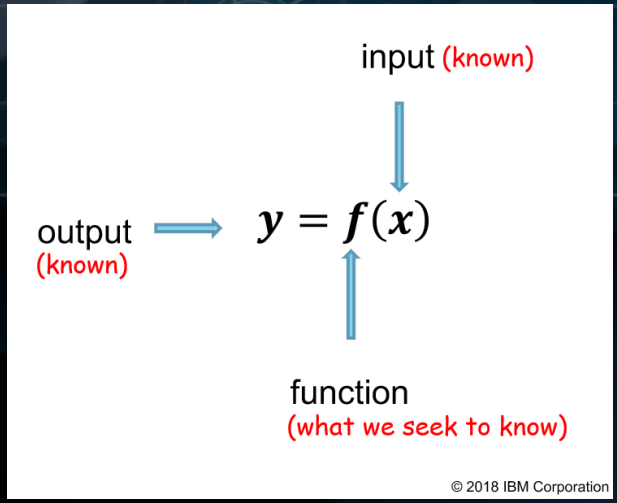
What % of your Db2 workload qualifies?



Other Use Cases

Automating package fallback – learn what went wrong and process with regression

Any problem you want to solve:



Thank You