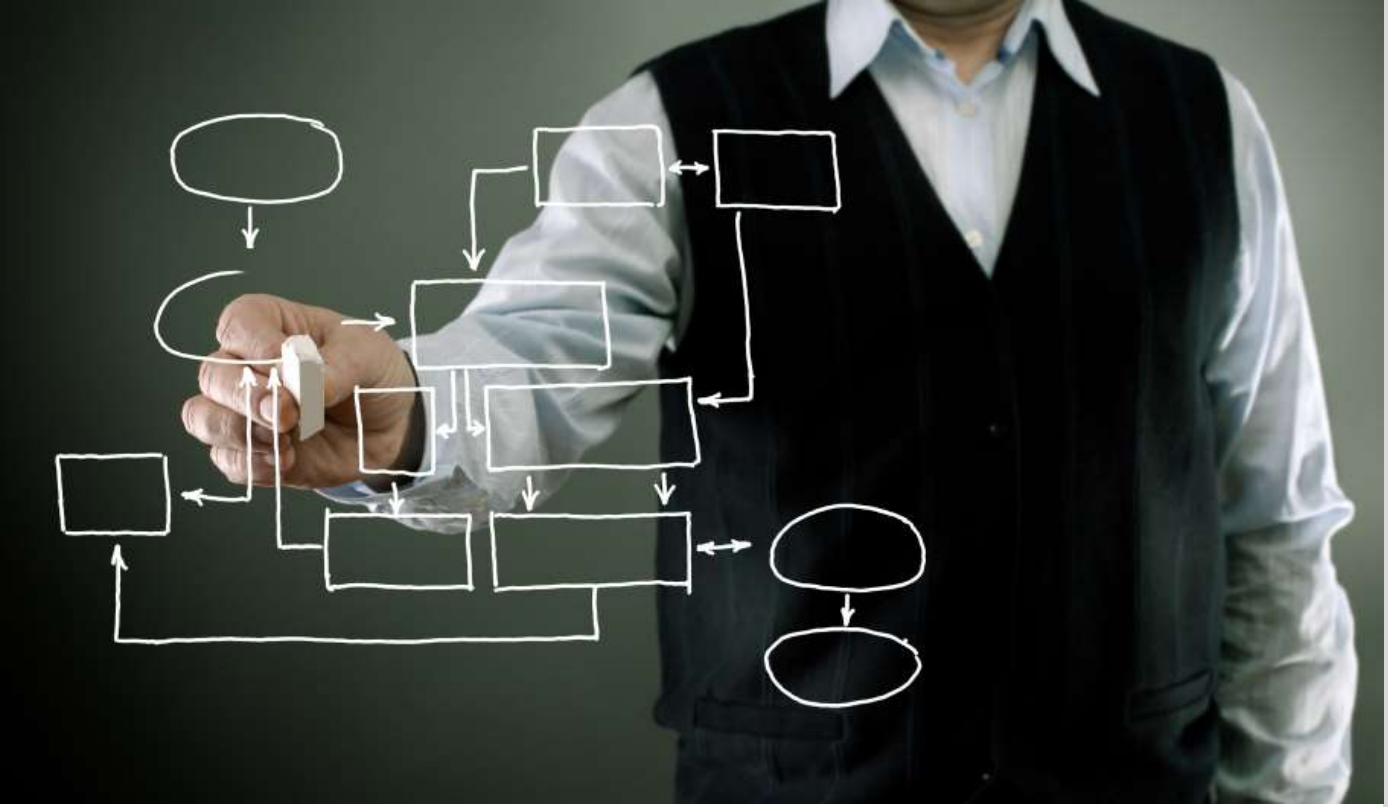


PERFORMANCE OF INDUSTRIAL SENSOR DATA PERSISTENCE IN DATA VAULT

Florian Bachinger, Jan Zenisek, Lukas Kammerer, Martin Stimpfl, Gabriel Kronberger

HEAL
HEURISTIC AND EVOLUTIONARY ALGORITHMS LABORATORY

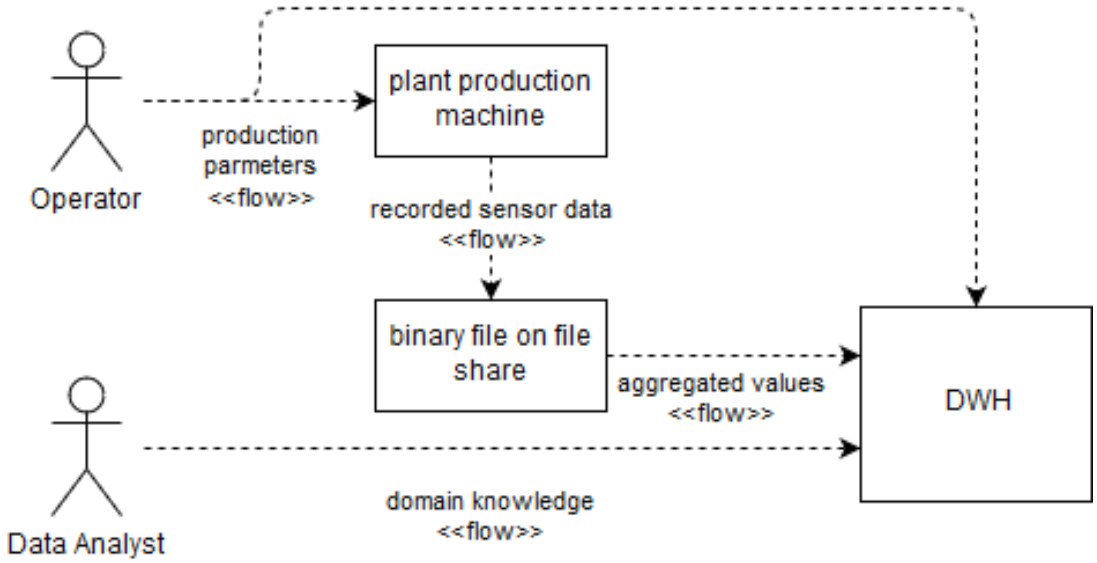
Contact:
Florian Bachinger MSc
FH OOE - School of Informatics, Communications and Media
Heuristic and Evolutionary Algorithms Lab (HEAL)
Softwarepark 11, A-4232 Hagenberg

e-mail:
florian.bachinger@fh-hagenberg.at
Web:
https://heal.heuristiclab.com
https://www.symreg.at

Motivation

- **Storing industrial sensor data**
  - Generic storage model
  - Performant read and write access
  - Compression of data
  - Versioning and audit information
- **As basis for reporting and machine learning applications**
  - Easy and fast transformation of stored data
- **Analysis of scalability for long term deployment and production use**

Scenario description



- Production machine records high frequency sensor data
- Manual preprocessing of data through domain expert (data analyst)
- Versioned, audited storage of sensor data in DWH

Data warehousing

- **DWH modeling**
  - "A data warehouse is a **subject-oriented**, **integrated**, nonvolatile, and **time-variant collection of data** … ." [1]
- **Data Vault modeling approach**

Indicators according to Hultgren [2]:

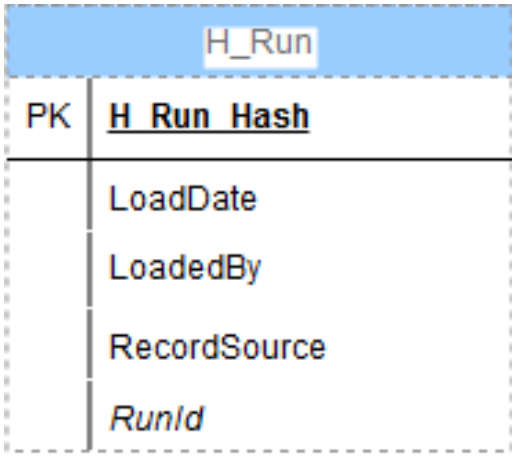
- Integration of multiple heterogeneous data sources
- Accurate and complete time-slice history
- Frequent addition of new sources or change of existing ones

[1] Inmon, W. H. (2002). Building the data warehouse. John Wiley & Sons, Inc., page 31
[2] Hultgren, H. (2012). Modeling the agile data warehouse with data vault. New Hamilton.

Data Vault - Overview

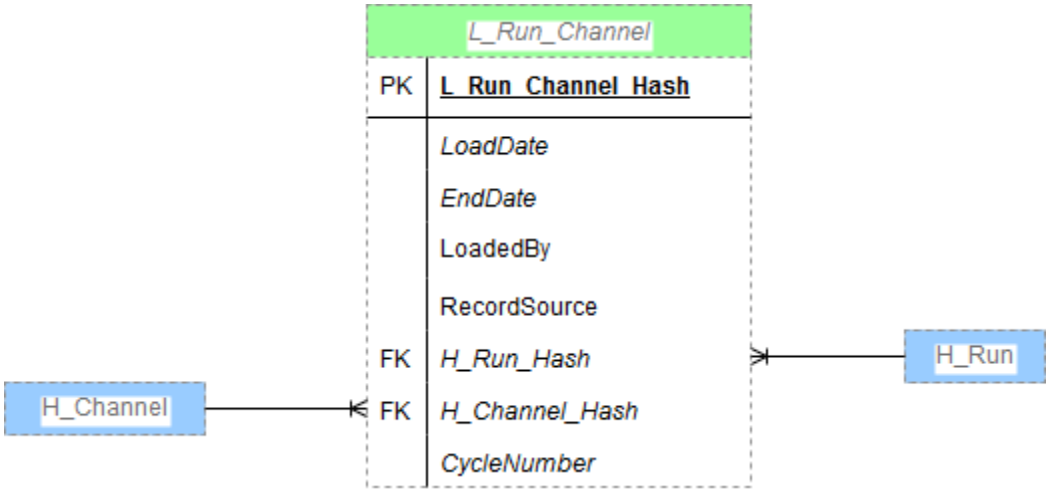
- **DWH modeling approach**
- **Compliant with RDBMS restrictions**
- Three different RDBMS table types: **Hubs, Links, Satellites**
- **Audit information**
  - Record source
  - Loaded by
  - Load date
  - Edited By
- **Hashed business keys**
  - Improved performance (compared to variable string length)
  - Easier parallel import (compared to identity insert)

Data Vault – Hub – modeling business concepts



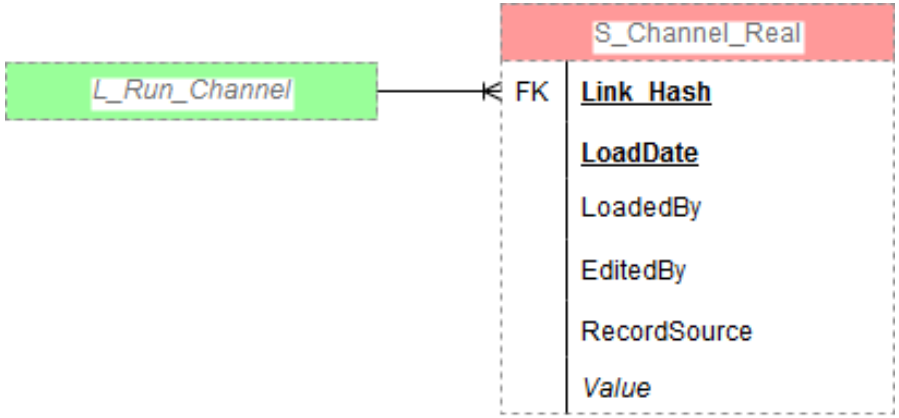
- **One logical business concept**
- **Stores only the business key for unique identification**
- **Only first occurrence of the key creates an entry**

Data Vault – Link – modeling business relations



- **Logical connection between to business concept instances**
- Every connection is of possible **n:m cardinality**
- No longer valid connections are represented by a defined **EndDate**
- **Primary key is a hash of all Foreign Key values**

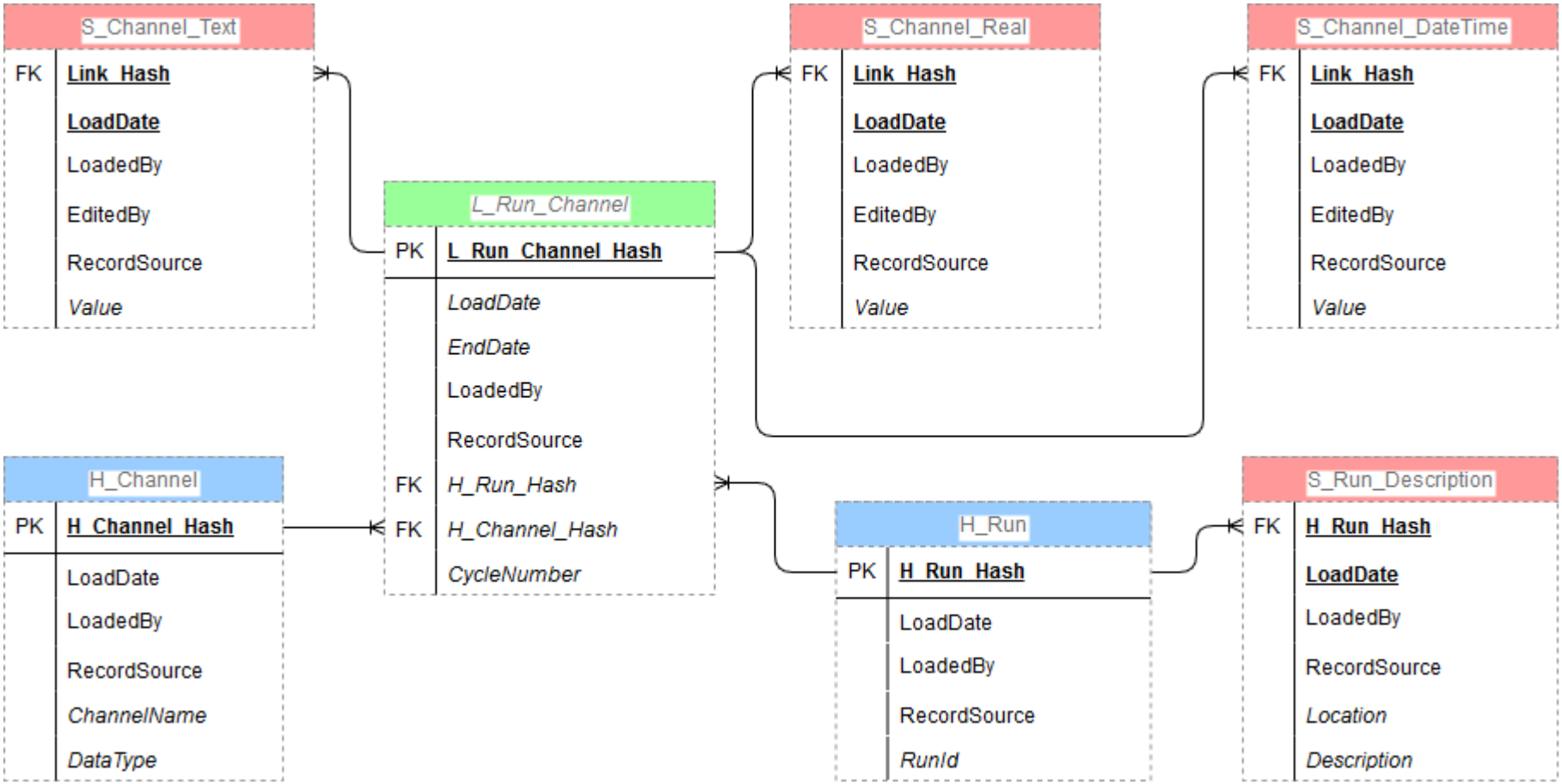
Data Vault – Satellite – modeling business data



- **Satellite can be connected to a Link or Hub**
- **LoadDate is part of the unique constraint to allow time slices**
- Adding new data attributes to a business entity **by adding a new satellite**
- Separation of Satellites by following indicator encouraged[1]:
  - **Subject area / data context**
  - **Rate of change**
  - **Source system**
  - **Type of data**

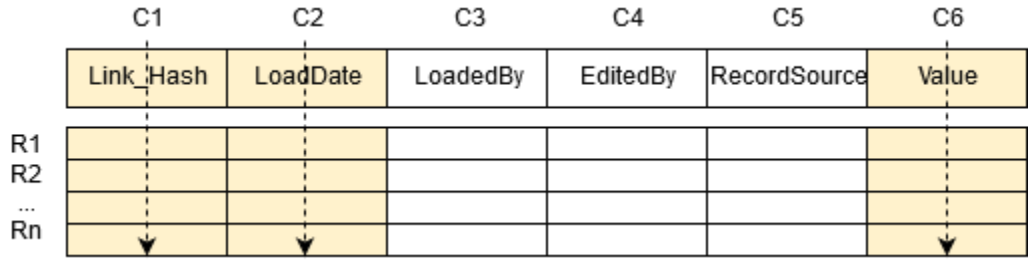
[1] Hultgren, H. (2012). Modeling the agile data warehouse with data vault. New Hamilton.

Devised sensor storage model

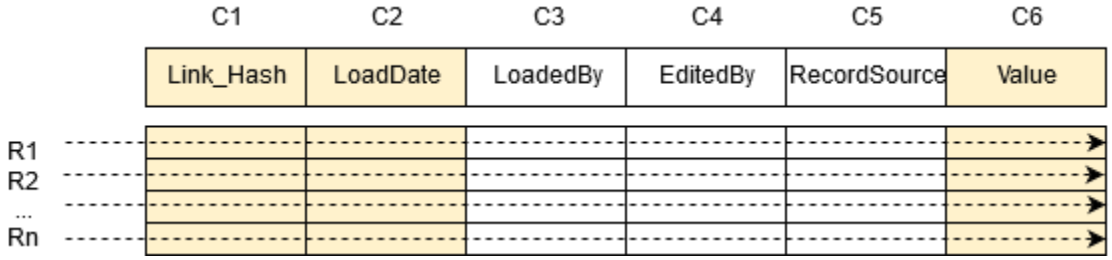


Performance benefits of clustered columnar store index

Columnar storage



Row storage



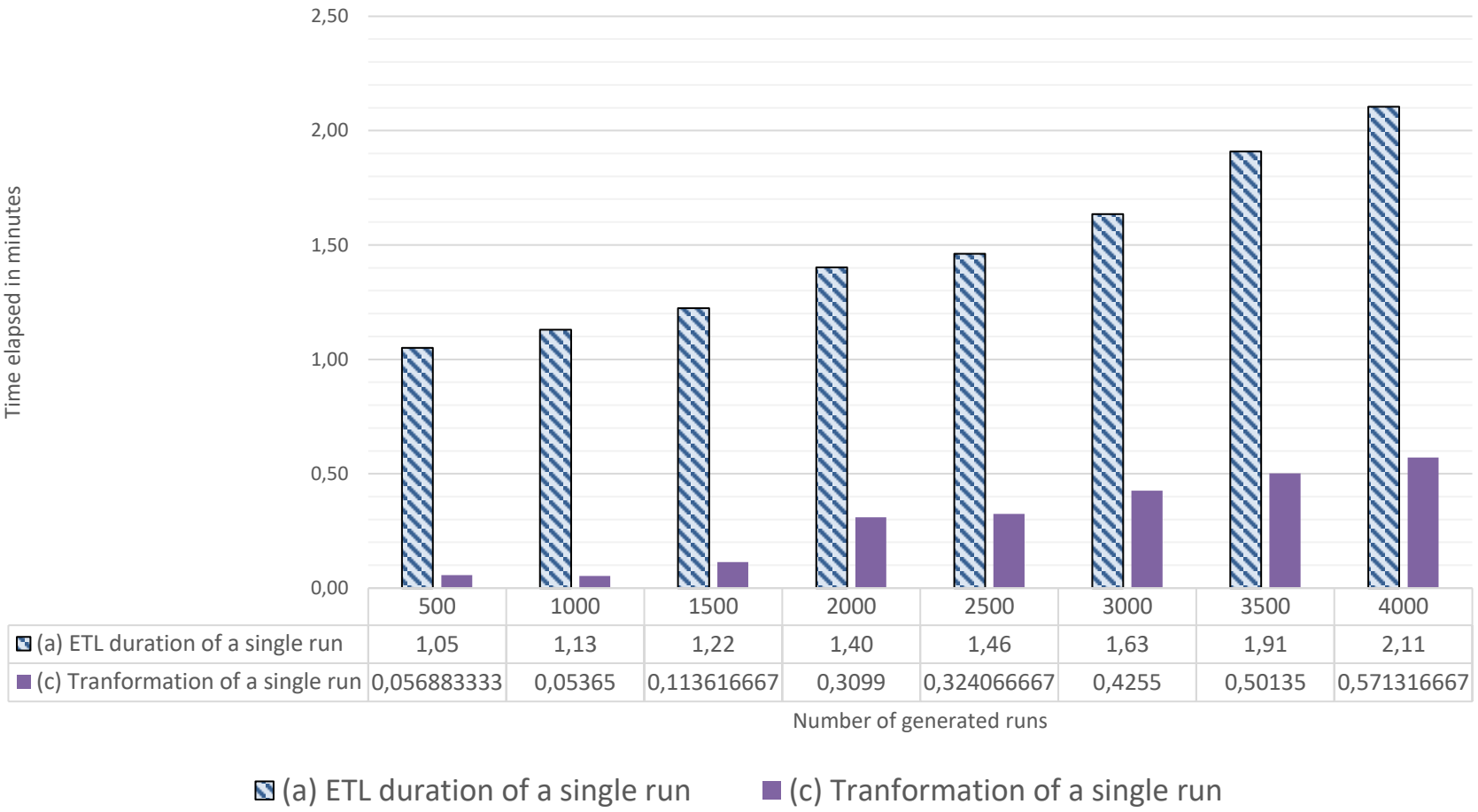
How to transform to usual format

```
SELECT * INTO dm.RunData
FROM (SELECT c.[Name] AS ChannelName
   , p.RunNr
   , l.CycleNumber
   ,  d.[value] AS DataValue
 FROM dv.L_Run_Channel l
 JOIN dv.H_Channel c
  ON l.H_Channel_Hash = c.H_Channel_Hash
 JOIN dv.H_Run p
  ON l.H_Run_Hash = p.H_Run_Hash
 LEFT JOIN dv.S_Channel_Double d ON d.L_Run_Channel_Hash = l.L_Run_Channel_Hash
  AND c.DataType = 'Double'
  AND d.LoadDate = (SELECT MAX(LoadDate)
                    FROM dv.S_Channel_Double sub WHERE sub.L_Run_Channel_Hash = l.L_Run_Channel_Hash)) p
 PIVOT( MIN(DataValue) FOR ChannelName IN ( Sensor1, Sensor2, Sensor3, Sensor4) AS PivotTable;
```

Transpose generic channel oriented storage into usual format

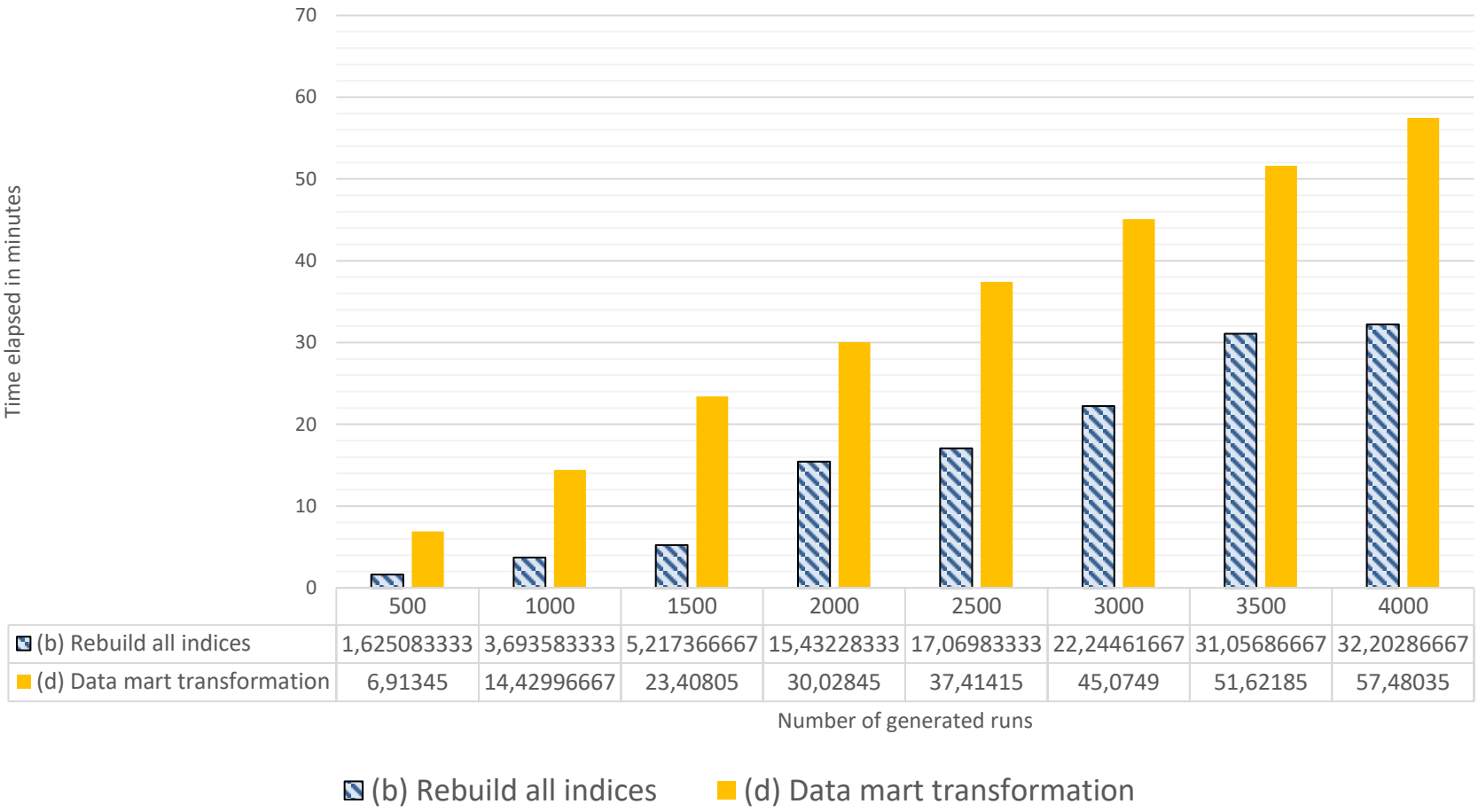
RunNr	CycleNumber	Sensor1	Sensor2	Sensor3	Sensor4
RunNr_123	1	54,6268845	0,12314744	0,48683135	0,99353153
RunNr_123	2	57,3680382	0,11766822	0,48689716	0,99393608
RunNr_123	3	59,2105484	0,11411818	0,48687127	0,99374263

ETL Timings

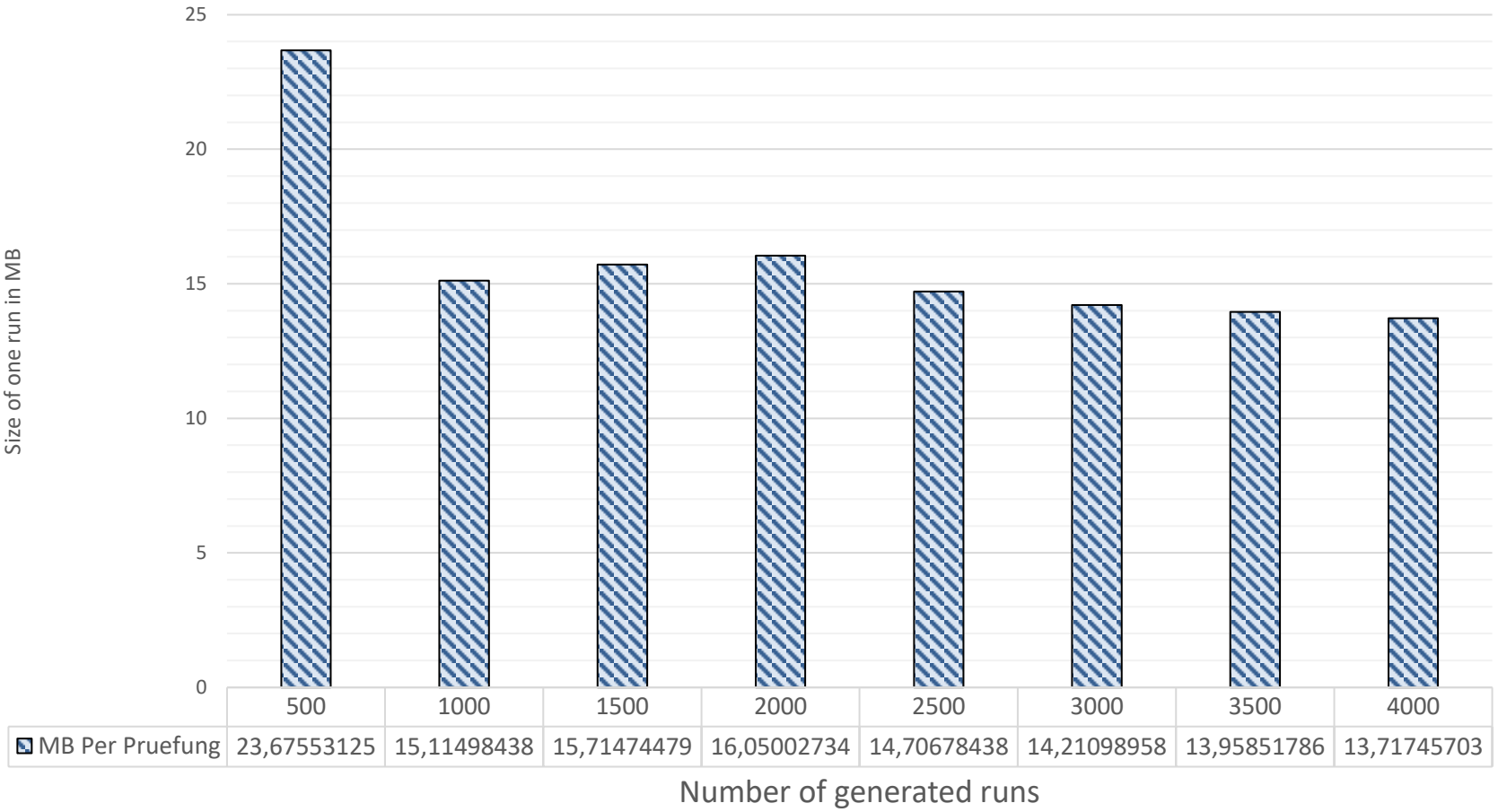


	500	1000	1500	2000	2500	3000	3500	4000
(a) ETL duration of a single run	1,05	1,13	1,22	1,40	1,46	1,63	1,91	2,11
(c) Tranformation of a single run	0,056883333	0,05365	0,113616667	0,3099	0,324066667	0,4255	0,50135	0,571316667

SQL Operation Timings



Required storage space



	500	1000	1500	2000	2500	3000	3500	4000
MB Per Pruefung	23,67553125	15,11498438	15,71474479	16,05002734	14,70678438	14,21098958	13,95851786	13,71745703

Conclusion

- **Propose a DWH model to store industrial sensor data**
  - Modeled in Data Vault
  - Still highly performant on off the shelf hardware for 355.200.000 sensor values
- **… where versioning and audit information is required**
- **… where the transformed data is read frequently**

Questions?