Tackling Semantic Shift in Industrial Streaming Data Over Time

Lisa Ehrlinger, Christian Lettner, Johannes Himmelbauer

DI Lisa Ehrlinger
Senior Researcher Data Analysis Systems / Software Competence Center Hagenberg
Senior Researcher / Johannes Kepler University Linz
+43 50 343 836
lisa.ehrlinger@scch.at
www.scch.at

SCCH is an initiative of

JOHANNES KEPLER UNIVERSITY LINZ

SCCH is located in

Data Quality Research at JKU and SCCH

- Johannes Kepler University (JKU) Linz
  - Senior researcher in research group of a.Univ.-Prof. Wolfram Wöß
  - DQ tool survey: https://arxiv.org/abs/1907.08138 (Ehrlinger et al. 2019)
  - DQ tool DQ-MeeRKat: https://github.com/lisehr/dq-meerkat
  - Talks at MIT Chief Data Officer and Information Quality Symposium 2019 and 2020

- Software Competence Center Hagenberg GmbH (SCCH)
  - Lead of research focus "Data Management and Data Quality"
  - Research on DQ issues with industrial companies (e.g., KTM)
  - DQ tool: A DaQL to Monitor Data Quality in Machine Learning Applications
    International Conference on Database and Expert Systems Applications. Springer, Cham (Ehrlinger et al. 2019)

© Software Competence Center Hagenberg GmbH

s c c h
software competence center
hagenberg

The Software Competence Center Hagenberg (SCCH) is

part of the software park Hagenberg

At JKU / Open Innovation Center

Non-profit organization for **data science** and **software science**

Founded 1999

~ 80 employees

> 7 Mio. € turnover

COMET Center

Union of SCCH partner companies

Semantic Shift – A Data Quality Problem

- Linguistics: “**semantic shift**”
  - Also: “semantic change”, “semantic drift”
  - Evolution of word meaning over time (Bloomfield 1933)
- Machine learning (ML) research: “**concept drift**”
  - Drift in the target variable predicted by a ML model (Widmer & Kubat 1996)
- Data quality (DQ) research
  - A lot of research into DQ dimensions (cf. Wang & Strong 1996)
  - Related DQ terms:
    - “identity” and “rigity”, referring to the stability of a variable (Guarino and Welty 2002)
    - “timeliness”, which describes how current data is for a task at hand (Heinrich et al. 2018)

© Software Competence Center Hagenberg GmbH

Industrial Application Scenario (1)

- Austrian manufacturing company
  - Mass production of plastic and multi-material parts with **injection molding machines**
  - Injection molding = complex physical-chemical process

- ML Project with SCCH: monitor the stability of the production process
  - Avoid machine damage & perform countermeasures as early as possible
  - Data-driven solution to ensure production quality using …
    - Stream data processing, classical ML algorithms, outlier detection, causal discovery, etc.
  - Requirement for ML: algorithms expect data to be in a standardized format

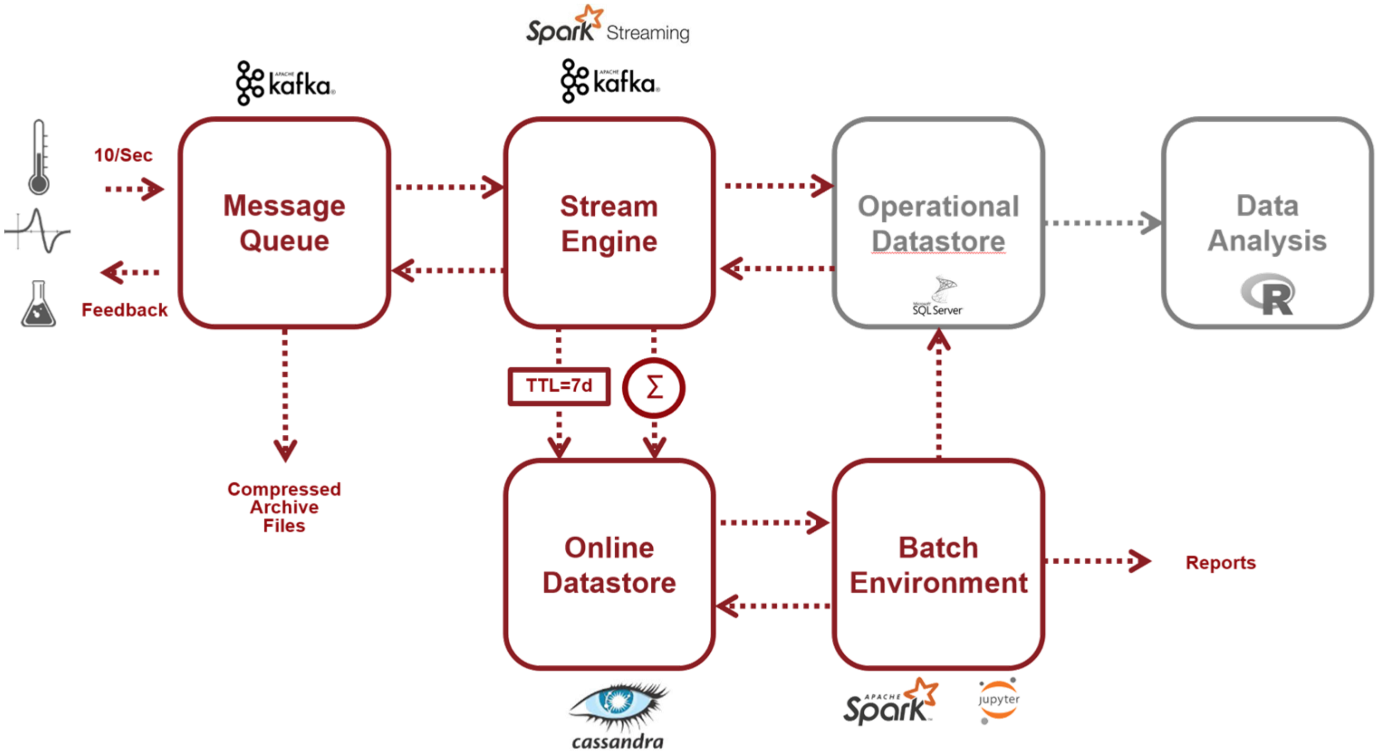
Industrial Application Scenario (2)

- Injection molding machines almost exclusively from same vendor
  - Shipped with standardized API → high level of data consistency
  - Process data logged into the “MES system”

- Issues with semantic shift
  - There exist different machine types and versions
  - Identical machines (same type + version) might still have different firmware
  - Variables in process log schema undergo semantic shift (over time)
    - Example: with a firmware update, measurements of pressure sensor are changed from storage in bar to millibar (updated for higher granularity)
  - Ignoring semantic shift yields to **wrong ML results**!

© Software Competence Center Hagenberg GmbH

L* System Architecture



© Software Competence Center Hagenberg GmbH

scch software competence center hagenberg

L* Online Datastore

- Apache Cassandra
  - Column-based
  - Optimized for large amounts of data

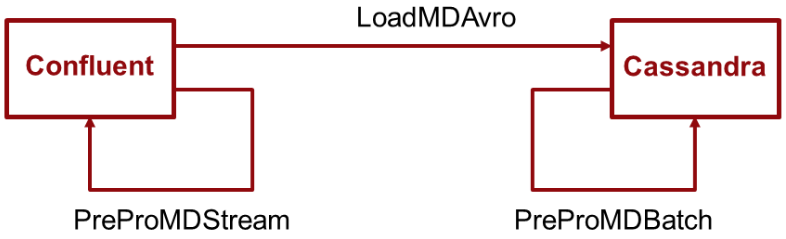
cassandra

```
create table MDavro (
          jahr int,
          seriennummer int,
          interval int,
          zeitpunkt timestamp,
          value blob,
          primary key((jahr, seriennummer,
             interval), zeitpunkt));

create table MD (
          jahr int,
          seriennummer int,
          metric text,
          zeitpunkt timestamp,
          value text,
          primary key((jahr, seriennummer,
             metric), zeitpunkt));
```

© Software Competence Center Hagenberg GmbH

L* Data Processing to handle Semantic Shift



- 3 Spark jobs
  - Data preprocessed according to defined rules to handle semantic shift (cf. next slide)
- Stream engine
  - Encoded with Apache Avro data serialization
  - `LoadMD-Avro` receives machine data → decode → store to Cassandra
  - `PreProMDStream` receives machine data → decode → preprocessing with rules → returns data to Confluent
- Batch environment
  - `PreProMDBatch` reads data from Cassandra (requires start and end point as batch interval) → preprocessing with rules → returns data to Cassandra

© Software Competence Center Hagenberg GmbH

Excerpt of Semantic Shift Processing Rules

MD_paramname	Process_paramname	Machine_type	Scale	Offset	Lag	Datatype
Process_value_1	Mode stopped	T1, T2, T3	1	0	0	Bool
Process_value_2	Mote Starting	T1, T2, T3	1	0	0	Bool
Process_value_3	Mode Production	T1, T2	1	0	0	Bool
Process_value_4	Product Counter	T1, T2, T3	1	0	0	Long
Process_value_5	Process Temperature 1	T1, T2	1	0	0	Float
Process_value_6	Process Pressure	T1, T2	1	0	0	Float
Process_value_3	Mode Production Phase 1	T3	1	0	0	Bool
Process_value_7	Mode Production Phase 2	T3	1	0	0	Bool
Process_value_5	Process Temperature 1	T3	1.8	32	0	Float
Process_value_6	Process Temperature 2	T3	1.8	32	0	Float
Process_value_7	Process Temperature 1 Previous	T3	1.8	32	1	Float
Process_value_8	Process Pressure	T3	1	0	0	Float

© Software Competence Center Hagenberg GmbH

scch software competence center hagenberg

L* Performance

- Deployment in productive environment → handle Big Data
- Performance evaluation: 28.8 million records
  - Avg.: 42.2 measurement values / record

Spark Data Stream	Unit	Throughput (unit/sec)	Storage (byte/unit)	Storage (disk space in GB)
LoadMDAvro	Records	358	182	5.01 GB
PreProMDBatch	Values	174,343	4.6	6.49 GB
PreProMDStream	Values	4,816	-	-

software competence center hagenberg

Outlook

- Data preprocessing system L* to handle semantic shift in data streams
- Rule-based solution most common in DQ tools (cf. Ehrlinger et al. 2019)

- Ongoing and future work
  - Extend rule-based system with **semantic solution** to achieve a higher degree of automation
  - Investigate **DQ assessment for streaming data** from a more general viewpoint → develop DQ metrics specific for data streams

© Software Competence Center Hagenberg GmbH

s c c h
software competence center hagenberg

Contact

DI Lisa Ehrlinger
Senior Researcher Data Analysis Systems
+43 50 343 836
lisa.ehrlinger@scch.at
www.scch.at

References

1. L. Bloomfield, *Language*. Allen & Unwin, 1933.
2. A. Tsymbal, "The Problem of Concept Drift: Definitions and Related Work," *Computer Science Department, Trinity College Dublin*, vol. 106, no. 2, 2004, p. 58.
3. G. Widmer and M. Kubat, "Learning in the Presence of Concept Drift and Hidden Contexts," *Machine Learning*, vol. 23, no. 1, 1996, pp. 69–101.
4. M. Klenner and U. Hahn, "Concept Versioning: A Methodology for Tracking Evolutionary Concept Drift in Dynamic Concept Systems," in ECAI, vol. 94. PITMAN, 1994, pp. 473–477.
5. L. Ehrlinger, E. Rusz, and W. W¨oß, "A Survey of Data Quality Measurement and Monitoring Tools," 2019, https://arxiv.org/abs/1907.08138

© Software Competence Center Hagenberg GmbH