Daimler TSS

Ein Unternehmen der Daimler AG

# LECTURE @DHBW: DATA WAREHOUSE

# PART L: DATA CATALOG


## ANDREAS BUCKENHOFER, DAIMLER TSS


DHBW
Duale Hochschule
Baden-Württemberg

# ABOUT ME

**Andreas Buckenhofer**
Senior DB Professional
andreas.buckenhofer@daimler.com

Since 2009 at Daimler TSS
Department: Big Data
Business Unit: Analytics

https://de.linkedin.com/in/buckenhofer

https://twitter.com/ABuckenhofer

DOAG https://www.doag.org/de/themen/datenbank/in-memory/

DHBW http://wwwlehre.dhbw-stuttgart.de/~buckenhofer/

XING https://www.xing.com/profile/Andreas_Buckenhofer2

**1995** — **Datenbanken**
- Admin
- Tuning
- Entwicklung
- Modellierung

**1998** — **Oracle DB**
- ab V7.3.4

**1998** — **DWH / CRM**
- Architektur
- Integration
- Reporting

**2004** — **RZ Migration**
- 2 Jahre UK

**2006** — **Siebel CRM**
- ab V7.8

**2006** — **Replication / CDC**
- Oracle CDC

**2007** — **SQL Server**
- ab V2005
- SSIS

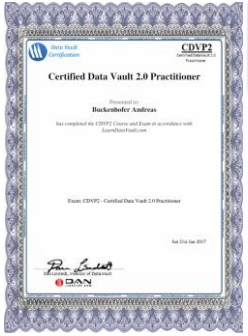
**2009** — **Data Vault**

**2010** — **Cognos**
- ab V8.4

**2012** — **DB2**
- ab V9.7

**2013** — **Hadoop**
- Architektur
- Datenintegration
- SQL-on-Hadoop

**2015** — **IIDR**
- 10.2.1




ORACLE Certified Professional

ORACLE Certified Expert

# ANDREAS BUCKENHOFER, DAIMLER TSS GMBH

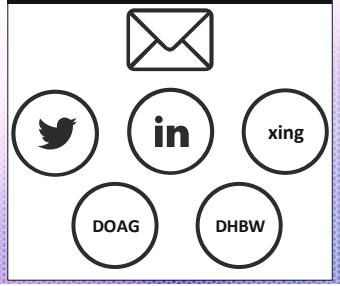
*“Forming good abstractions and avoiding complexity is an essential part of a successful data architecture”*

Data has always been my main focus during my long-time occupation in the area of data integration. I work for **Daimler TSS** as Database Professional and Data Architect with over 20 years of experience in Data Warehouse projects. I am working with Hadoop and NoSQL since 2013. I keep my knowledge up-to-date - and I learn new things, experiment, and program every day.

I share my knowledge in internal presentations or as a speaker at international conferences. I'm regularly giving a full lecture on Data Warehousing and a seminar on modern data architectures at Baden-Wuerttemberg Cooperative State University DHBW. I also gained international experience through a two-year project in Greater London and several business trips to Asia.

I’m responsible for In-Memory DB Computing at the independent German Oracle User Group (DOAG) and was honored by Oracle as ACE Associate. I hold current certifications such as "Certified Data Vault 2.0 Practitioner (CDVP2)", "Big Data Architect“, „Oracle Database 12c Administrator Certified Professional“, “IBM InfoSphere Change Data Capture Technical Professional”, etc.

Contact/Connect



# NOT JUST AVERAGE: OUTSTANDING.

As a 100% Daimler subsidiary, we give 100 percent, always and never less. We love IT and pull out all the stops to aid Daimler's development with our expertise on its journey into the future.

Our objective: We make Daimler the most innovative and digital mobility company.

# INTERNAL IT PARTNER FOR DAIMLER

- Holistic solutions according to the Daimler guidelines
  - IT strategy
  - Security
  - Architecture
- Developing and securing know-how
- TSS is a partner who can be trusted with sensitive data

As subsidiary: **maximum added value** for Daimler

- Market closeness
- Independence
- Flexibility (short decision making process, ability to react quickly)

# LOCATIONS

**Daimler TSS Germany**

7 locations

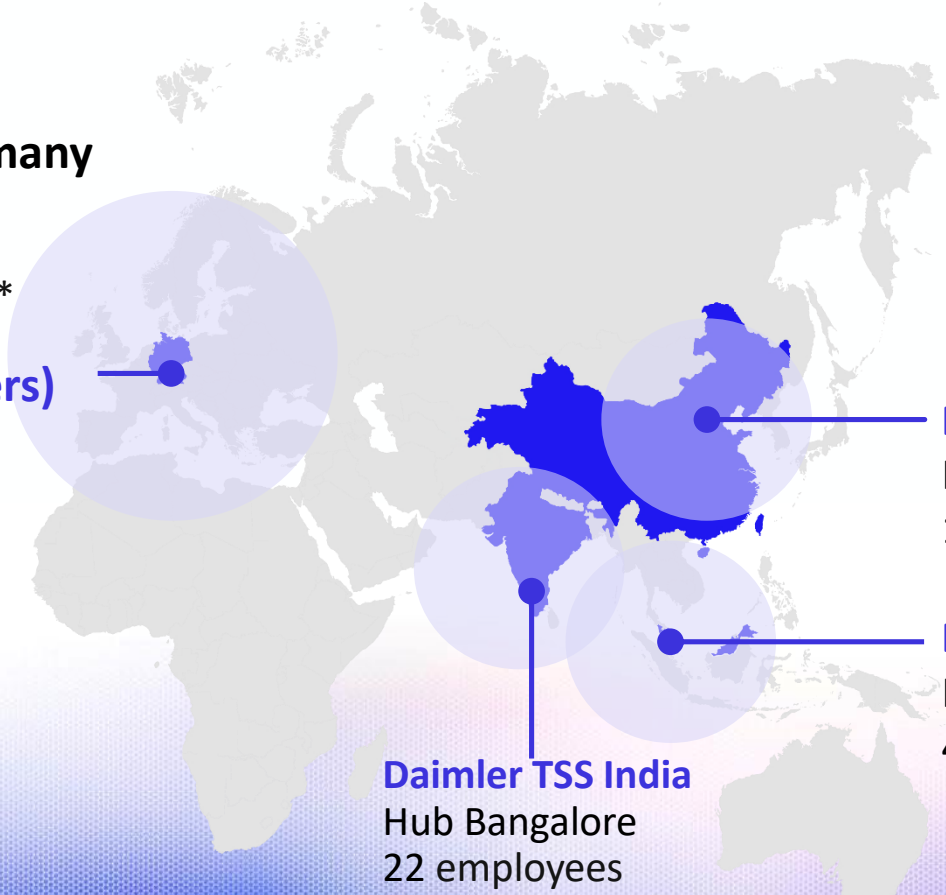
1000 employees*

**Ulm (Headquarters)**

**Stuttgart**

**Berlin**

**Karlsruhe**



**Daimler TSS China**
Hub Beijing
10 employees

**Daimler TSS Malaysia**
Hub Kuala Lumpur
42 employees

**Daimler TSS India**
Hub Bangalore
22 employees

* as of August 2017

# WHAT YOU WILL LEARN TODAY

- After the end of this lecture you will be able to
  - Explain metadata
    - Technical
    - Business
    - operational
  - Explain data catalogs
  - Understand use cases for metadata

# MAKING IT EASIER TO DISCOVER DATASETS AVAILABLE SINCE 05-SEP-2018

Google Dataset Search Beta

Nach Datensätzen suchen

Ausprobieren boston education data oder weather site:noaa.gov

Source: Google announcement https://www.blog.google/products/search/making-it-easier-discover-datasets/

# MAKING IT EASIER TO DISCOVER DATASETS AVAILABLE SINCE 05-SEP-2018

Google Dataset Search

taxi trips

Info

Fee

D
Taxi Trips
catalog.data.gov
data.wu.ac.at
Am Feb 3, 2018 aktualisiert

kaggle 2014 New York City Taxi Trips
www.kaggle.com
Am Jul 20, 2017 aktualisiert

C
Taxi Trips
data.cityofchicago.org
Am Aug 12, 2017 aktualisiert

W
Taxi Trips - Dashboard
data.wu.ac.at
Am Nov 17, 2016 aktualisiert

kaggle New York City Taxi Trips -

## Taxi Trips

catalog.data.gov data.wu.ac.at

**Datensatz aktualisiert** Feb 3, 2018

**Datensatz bereitgestellt von**
data.cityofchicago.org

**Beschreibung**

Taxi trips reported to the City of Chicago in its role as a regulatory agency. To protect privacy but allow for aggregate analyses, the Taxi ID is consistent for any given taxi medallion number but does not show the number, Census Tracts are suppressed in some cases, and times are rounded to the nearest 15 minutes. Due to the data reporting process, not all trips are reported but the City believes that most are. See http://digital.cityofchicago.org/index.php/chicago-taxi-data-released for more information about this dataset and how it was created.

# FIND THE RIGHT DATA

- With data science and analytics on the rise and under way to being democratized, the **importance of being able to find the right data** to investigate hypotheses and derive insights is paramount
  Source: https://www.zdnet.com/article/google-can-now-search-for-datasets-first-research-then-the-world
- Google Dataset search helps to find external data
  - Schema.org defines open metadata format; dataset itself may not be open/free
  - Search engines can interpret the format
  - Ranking of data
  - Help users discover where the data is and user can access it directly from the source

**What about internal data?**

# WHAT IS METADATA?


**Data about other data**

# TYPES OF METADATA (1)

## Business Metadata

- Definition of business vocabulary and relationships
- Definition of the value range
- Linkage to physical representation

GL Account Number

The ten digit account number for general ledger. Sometimes referred to as the account ID. This value is of the form L-FIIIIVVVV.

Business

Technical

Database = DB2

Schema = NAACCT

Table = DLYTRANS

Column = ACCT_NO

data type = char(11)

# TYPES OF METADATA (2)

## Report and ETL metadata

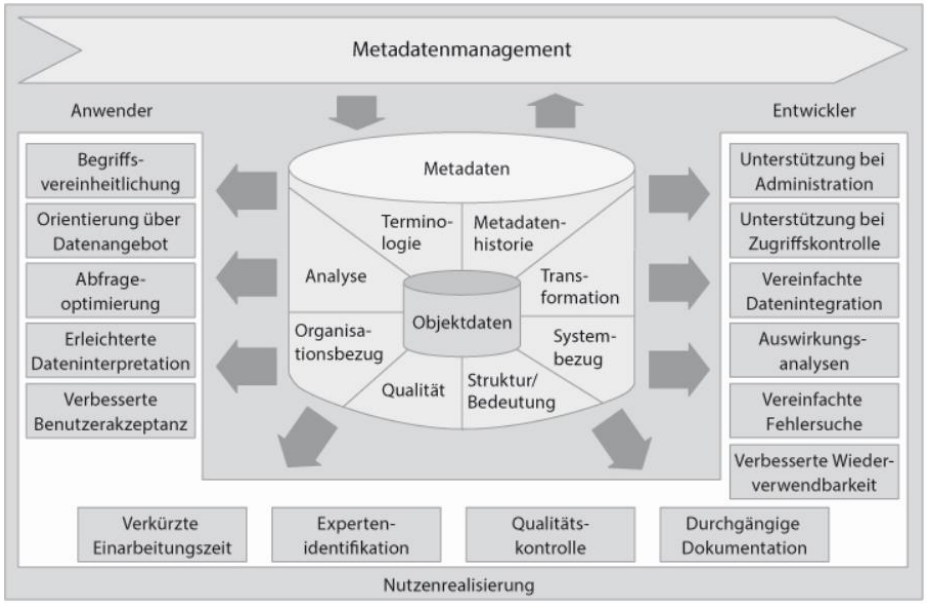
- Report definitions
- Data sources
- Column definitions
- Computations

## Logical and physical metadata of data model

- Table structure
- Definition of columns
- Relationships between tables and columns
- Dimension hierarchy

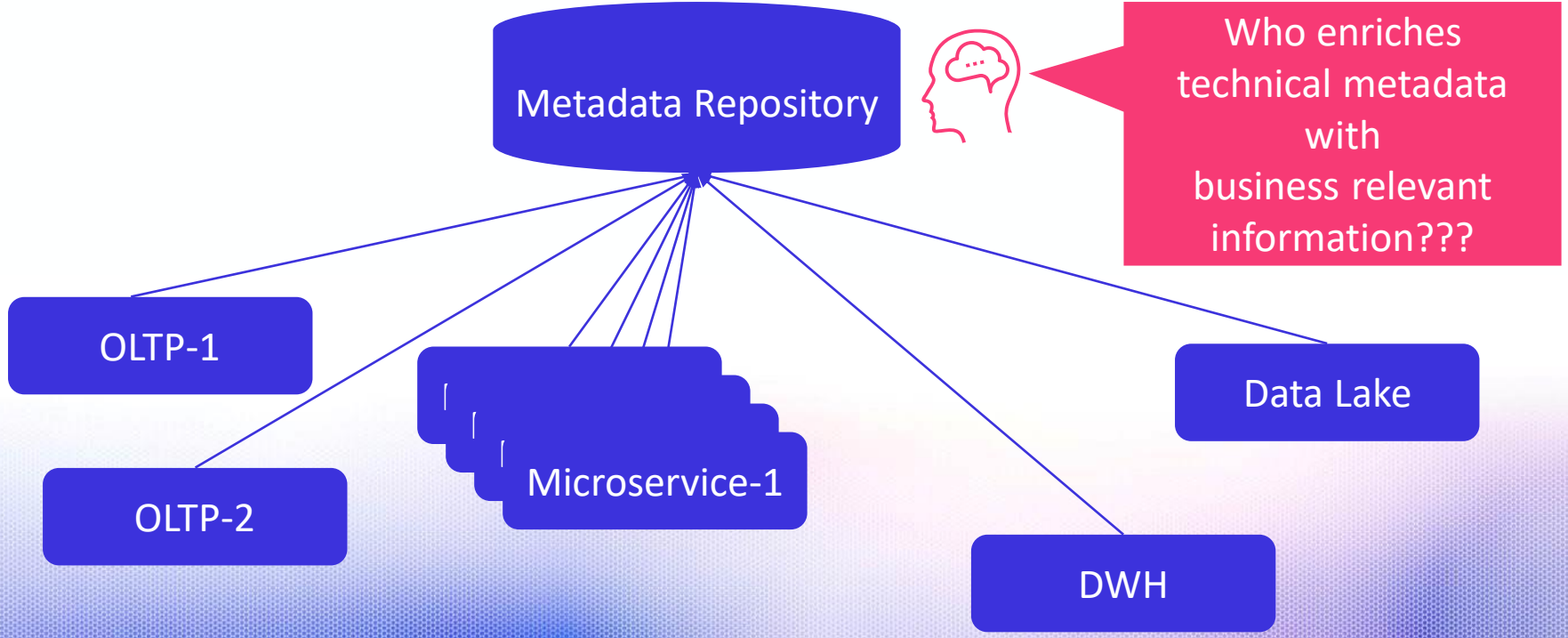
# BENEFITS OF METADATA MANAGEMENT

- Data Lineage and dependencies
- Generating and controlling DWH processes
- Improve SW development quality
- Increase comprehensibility of KPIs




Source: Detlef Apel: Datenqualität erfolgreich steuern, dpunkt 2015, chapter 14

# TECHNICAL METADATA MANAGEMENT VERY OFTEN NOT SUCCESSFUL



# OVER 75%

- Of time is spent for
- Say they least enjoy

# DATA PREPARATION

## Data Consumers

# DOES METADATA MANAGEMENT PROVIDE ANSWERS TO SUCH QUESTIONS ACROSS THE WHOLE WORKFLOW?

What table contains production dates?

How is this column calculated?

Is the data reliable?

How to get access to the data?

Is FIN unique?

| Find | Understand | Trust | Access | Write |
|---|---|---|---|---|
| Search for data | | | Work with data | |

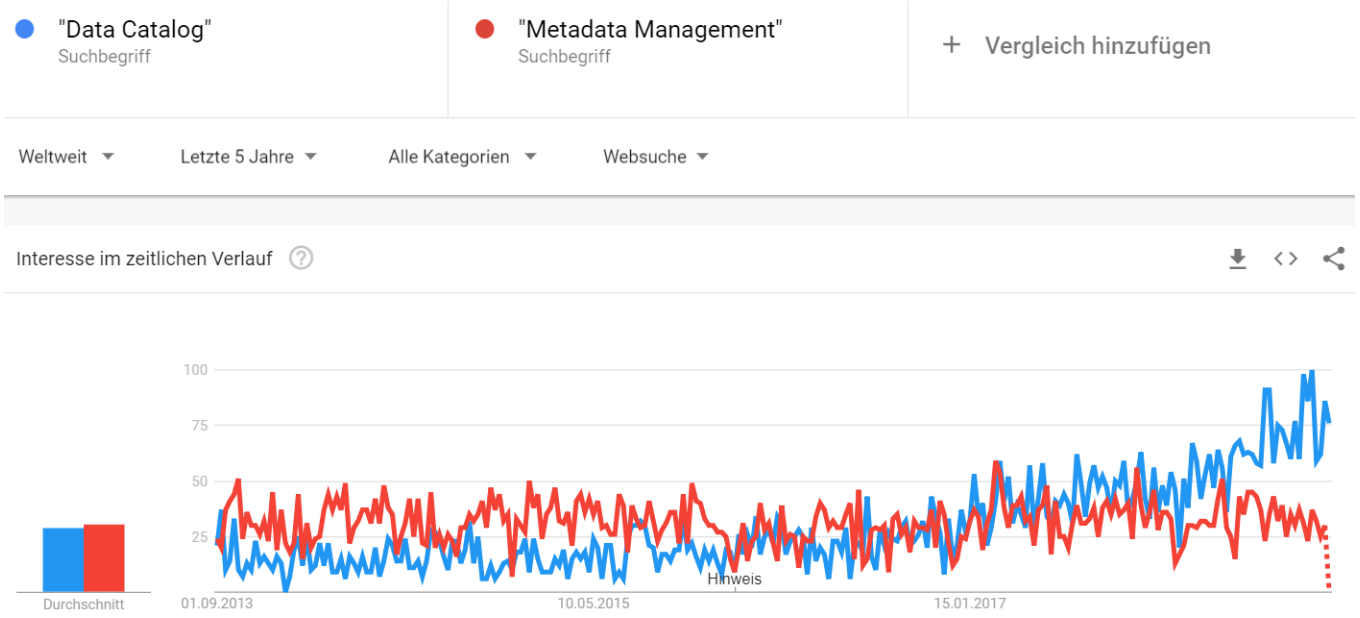
What tables are important?

Who knows about the data?

What is the difference between production_date and prod_dt?

How to join the tables?

# DATA CATALOG A HOT TOPIC

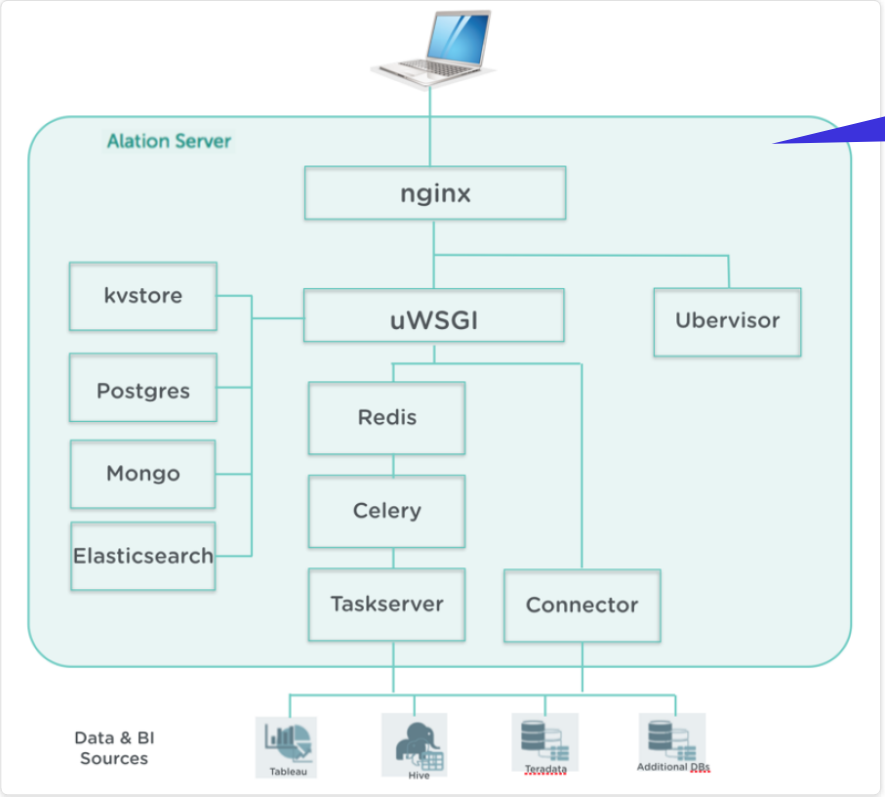



- New Data Catalog vendors are entering the market
- Established vendors rebrand and enrich their existing tools

# EVALUATION CRITERIA

- Technical Metadata
- Business Metadata incl. Glossary
- Tagging (Linkage)
- Collective Intelligence(Collaboration)
- Search
- Security
- Source connectors
- Data profiling
- Data access
- Lineage
- API
- Versioning
- Architecture
- Components
- Prerequisites
- Licencing

# ALATION ARCHITECTURE




Not just an RDBMS for structured metadata, but also storage engines for text data

# CATALOGING SOURCE SYSTEMS
# MANY FORMATS = MANY CONNECTORS

- RDBMS (Oracle, Db2, SQL Server, Teradata, …)
- Hadoop (HDFS, Hive, …; on-premises, Cloud)
- NoSQL DBs
- Files (Excel, csv, …)
- Powerdesigner, Erwin, and other data modeling tools

# METADATA IMPORT USED TO BE SIMPLE WITH RDBMS

```
Sep 25 00:03:12 Maple.local mdworker[19184]: code validation failed in the p
Domain=NSOSStatusErrorDomain Code=-67062 "The operation couldn't be complete
{SecCSArchitecture=ppc}
Sep 25 00:03:12 Maple.local mdworker[19184]: code validation failed in the p
Domain=NSOSStatusErrorDomain Code=-67062 "The operation couldn't be complete
{SecCSArchitecture=x86_64}
Sep 25 00:04:03 Maple.local CalendarAgent[664]: [com.apple.calendar.store.lo
because of content-type: [text/html; charset=UTF-8].]
Sep 25 00:04:05 --- last message repeated 1 time ---
Sep 25 00:04:05 Maple.local garcon[19162]: Garcon destroyed (0 alive).
Sep 25 00:04:08 Maple com.apple.xpc.launchd[1] (com.apple.imfoundation.IMRem
available on this platform.
Sep 25 00:04:08 Maple.local locationd[623]: NETWORK: requery, 0, 0, 0, 0, 11
Sep 25 00:04:09 Maple.local Safari[9673]: CFPropertyListCreateFromXMLData():
line 3. Parsing will be abandoned. Break on _CFPropertyListMissingSemicolon
Sep 25 00:04:09 --- last message repeated 4 times ---
Sep 25 00:04:09 Maple com.apple.xpc.launchd[1] (com.apple.imfoundation.IMRem
available on this platform.
Sep 25 00:04:10 Maple.local Safari[9673]: CFPropertyListCreateFromXMLData():
line 3. Parsing will be abandoned. Break on _CFPropertyListMissingSemicolon
Sep 25 00:04:40 --- last message repeated 4 times ---
Sep 25 00:04:43 Maple com.apple.xpc.launchd[1] (com.apple.imfoundation.IMRem
available on this platform.
Sep 25 00:04:44 Maple com.apple.xpc.launchd[1] (com.apple.imfoundation.IMRem
available on this platform.
Sep 25 00:04:47 --- last message repeated 2 times ---
```

Where is the data and where is the metadata in this logfile?

Data Lake: decentralized control of the data

# DATA LAKE / HADOOP

- Easy approach: Access Hive Metastore and import metadata
- Prerequisite: all data/files in HDFS require Hive access
- But unrealistic prerequisite
  - Many logs are just dumped into the file system
  - Interpreting ALL files by catalog SW unrealistic, too.
    - Huge computing power
    - Huge number of variations (Cloud, on-premises, SW versions) lacks support of vendors for catalog SW
  - Sources should deliver metadata

# CATALOGING @GOOGLE

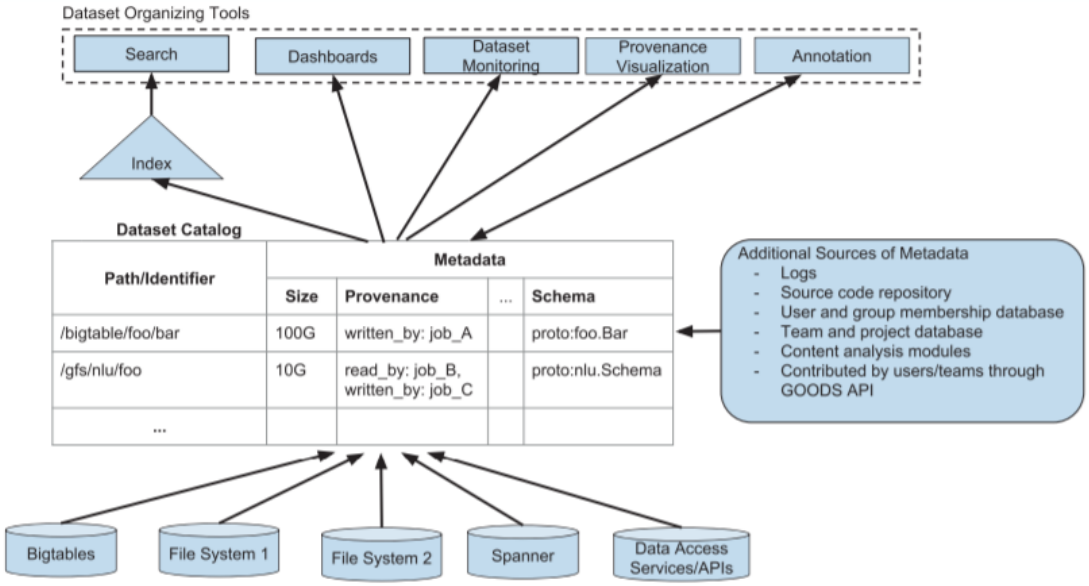



**Figure 1: Overview of Google Dataset Search (Goods). The figure shows the Goods dataset catalog that collects metadata about datasets from various storage systems as well as other sources. We also infer metadata by processing additional sources such as logs and information about dataset owners and their projects, by analyzing content of the datasets, and by collecting input from the Goods users. We use the information in the catalog to build tools for search, monitoring, and visualizing flow of data.**

Source: https://ai.google/research/pubs/pub45390

Heavy usage of Automation and Machine Learning

# CATALOGING AT NETFLIX, TWITTER, LINKEDIN, ETC.

| Company | Link |
| --- | --- |
| Netflix (Metacat) | https://medium.com/netflix-techblog/metacat-making-big-data-discoverable-and-meaningful-at-netflix-56fb36a53520<br>https://github.com/Netflix/metacat |
| Twitter | https://blog.twitter.com/engineering/en_us/topics/insights/2016/discovery-and-consumption-of-analytics-data-at-twitter.html |
| LinkedIn (WhereHows) | https://github.com/linkedin/WhereHows<br>https://github.com/linkedin/WhereHows/wiki |
| Google (Goods) | https://ai.google/research/pubs/pub45390<br>https://www.buckenhofer.com/2016/10/goods-how-to-post-hoc-organize-the-data-lake/ |
| Uber | https://eng.uber.com/databook/ |
| ebay | https://www.ebayinc.com/stories/blogs/tech/bigdata-governance-hive-metastore-listener-for-apache-atlas-use-cases/ |

# CATALOGING @UBER

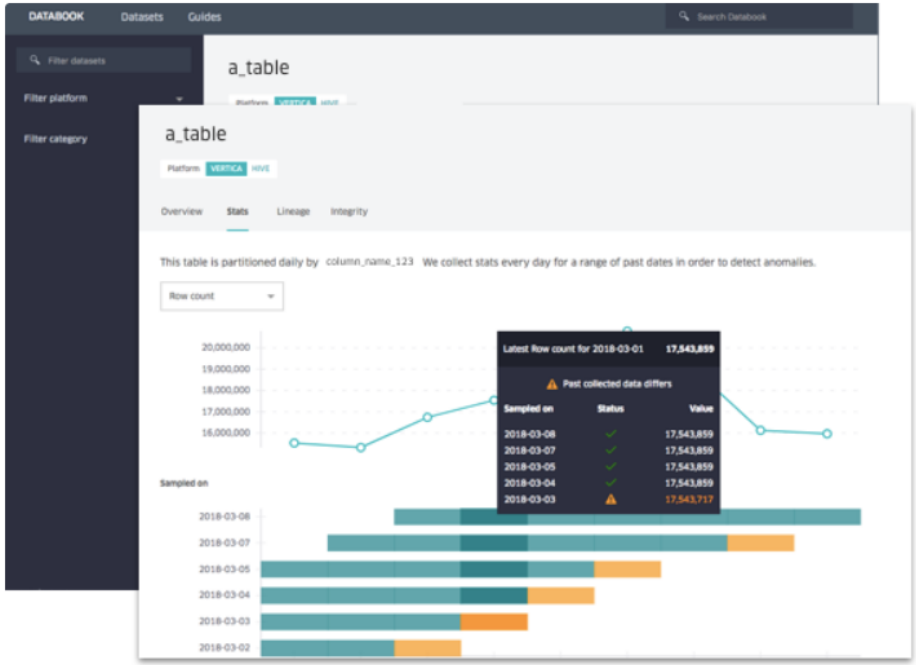



Figure 1. Databook is Uber's in-house platform that surfaces and manages metadata about internal data locations and owners.

Source: https://eng.uber.com/databook/

# CATALOGING @TWITTER

owner_name/

**Dataset Name** HDFS data source - active

Detail

Graph

Status

Comments

*Write a one line summary for this App*

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore v aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsu dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

LAST EDITED A MONTH AGO BY SRIKANTH THIAGARAJAN - edit - history

**Info**

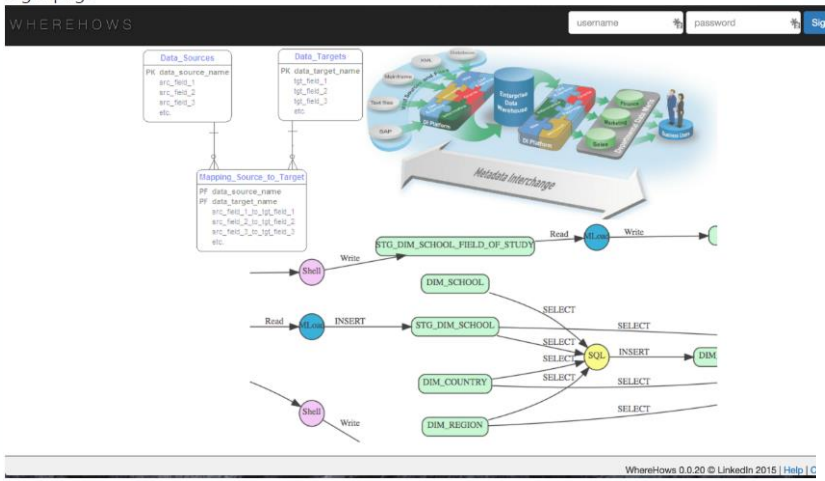
| | |
|---|---|
| **Owner** | Owner Name |
| **Frequency** | daily |
| **Data Range** | Unknown |
| **Thrift Class** | com.twitter.owner.data.DataClass |
| **HDFS Path** | /user/owner_name/dataset_name/yyyy/mm/dd- |

Example Usage

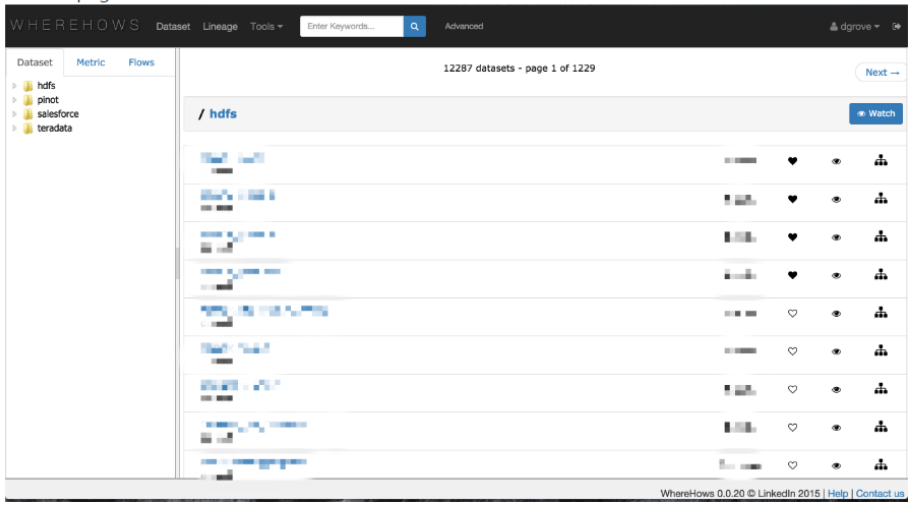
Source: https://blog.twitter.com/engineering/en_us/topics/insights/2016/discovery-and-consumption-of-analytics-data-at-twitter.html

# CATALOGING @LINKEDIN (OPEN SOURCE)

login page




dataset page




Source: https://github.com/LinkedIn/Wherehows/wiki

# CATALOGS ARE EVERYWHERE … GOOGLE, AMAZON

INVENTORY


USER EXPERIENCE

# INVENTORY VS USER EXPERIENCE

## Suppliers provide inventory

- A catalog should list everything that is actually available

## Consumers require user experience

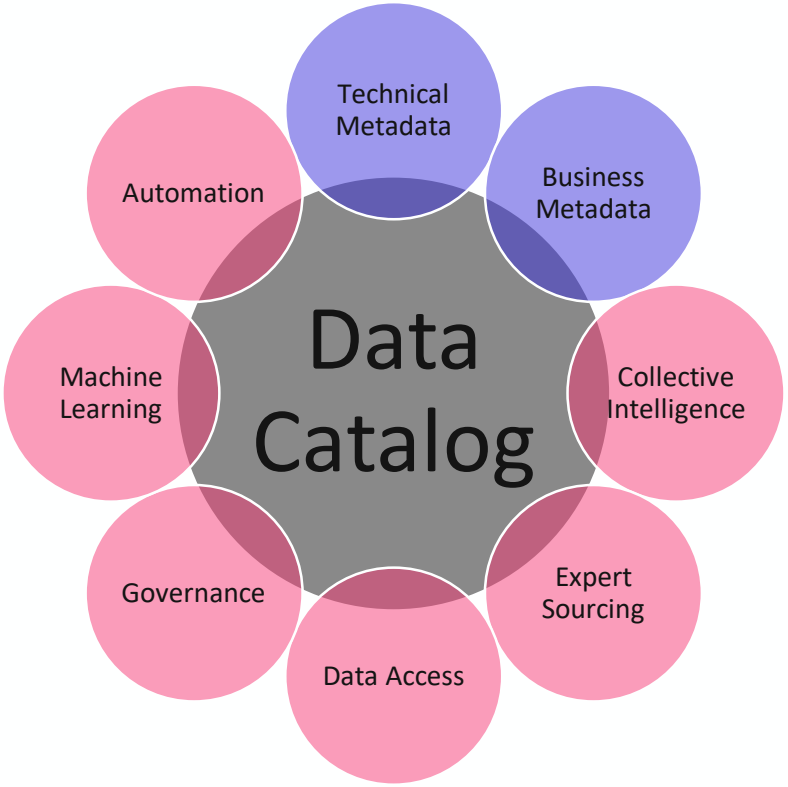
- A catalog should provide data usage statistics, ratings, data samples, statistical profiles, lineage, lists of users and stewards, and tips on how the data should be interpreted

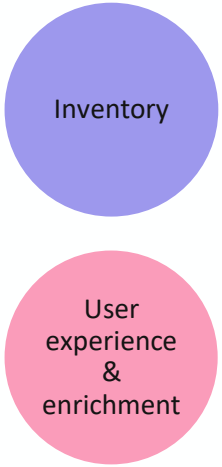
# AUTOMATION, CROWD KNOWLEDGE, AND EXPERTS

- ☹ limitation of permissions to a trusted group
  - A trusted group documents few datasets very well
  - But most of the metadata is not documented
  - Failure of many past approaches
- ☺ Automation, crowd knowledge and experts required
  - Automation to get a broad coverage and use existing information like query logs
  - Crowd to increase broad coverage
  - Experts to confirm or reject „guesses“

-> Combination of coverage and accuracy

# DATA CATALOG – AMAZON FOR INFORMATION



# CATALOG SEARCH

DAIMLER | Data Catalog | Glossaries | Track | Search | Admin | Help

plant

Showing Everything

- VTCOVO / VTCOVO_DBA — SL_PLANT_DIM
- TestAqua / aqua — t_part_plant
- TestAqua / aqua . t_plant_ml — plant_nm, Plant Name
- TestAqua / aqua — t_md_plant_asn
- TestAqua / aqua . t_plant_ml — plant_cd, Plant Code
- VTCOVO / vtcovo_dba . sl_plant_dim — plant_id, Plant Identify
- TestAqua / aqua — t_test_result_plant
- TestAqua / aqua — t_plant_ml
- TestAqua / aqua — t_pue_plant_ml

Data
Queries
Articles
Conversations

VT... / V... BA . SL_PLANT_DIM

## SL_PLANT_DIM

*Please provide a title*

Star | Watch | Compose | Open With | More... | Conversatio

Overview | Columns 3 | Samples 0 | Filters 0 | Joins 0 | Lineage | Queries 0

### Source Comments

Plant Dimension

### Description

*No description*

### Columns (3 of 3)

| | Column | Title | Type | Popularity |
|---|---|---|---|---|
| 1 | PLANT_ID | Plant Identify | varchar2(4) No Nulls | |
| 2 | PLANT_TXT | Plant Text | varchar2(255) | |
| 3 | CITY | City | varchar2(25) | |

### Content (0 of 0)

*No sample data*

### Published Queries (0 of 0)

### Top Users

*No top users*

### Data Stewards

*No data stewards*

### Tags

*No tags*

### Properties

OWNER
VTCOVO_DBA

### Relevant Articles

*No mentioning articles*

### Use Cases

*No use cases*

# SCHEMA AND ITS TABLES

VeDoc ML / ZUOC_ML

ZUO(

*Please provide a title*

Sta

Overview | Queries 0

## Description

*No description*

## Tables

| | | Table | Title | Popularity | Columns | Row |
|---|---|---|---|---|---|---|
| ★ | +1 | F_VEHICLE | | | 99 | 1000 |
| ☆ | | PIT_VEH_S | Pit Vehicle Sales Area Information | | 12 | 0 |
| ☆ | | F_MA_AXLE | | | 18 | 1000 |
| ☆ | | F_VEHICLE | | | 28 | 0 |
| ☆ | | P_PART_CU | | | 8 | 6779 |
| ☆ | | PIT_MA_AX | Pit Moeglichen Auspraegungen Axle Product Date | | 16 | 0 |
| ☆ | | PIT_MA_EN | Pit Moeglichen Auspraegungen Engine | | 45 | 0 |

# TABLE AND ITS COLUMNS WITH SAMPLE DATA

VeDoc ML / ZUOC_ML . F_VEHICLE_MANAGEMENT_BASE

1 Endorsement

## F_VEHICLE

*Please provide a title*

Starred | Watch | Compose | Open With | More...

Overview | Columns 99 | Samples 100 | Filters 0 | Joins 0 | Lineage | Queries 0

### Description

*No description*

### Sample Columns (3 of 99)

| | Column | Title | Type | Popularity |
|---|---|---|---|---|
| 0 | BUSINESS_DATE | Business Date | DATE | |
| 1 | LOAD_DATE | Load Date | TIMESTAMP | |
| 2 | JOB_ID | Job Identifyer | BIGINT | |

See all 99 columns...

### Sample Content (3 of 10,000+)

| N | PRODUCTSERIES_ID | VEHICLE_FIN | STATUS_LIFECYCLE_ID | BUSINESS_DATE | DATE_OFTECHNICALSTATE | LIGH |
|---|---|---|---|---|---|---|
| | Product Series Identi... | Vehicle Fin | Status Lifecycle Identi... | Business Date | | Lights |
| | 3062 | 3MB6344001M050739 | 859 | 2018-05-10 | 1899-12-31 | NULL |
| | 2973 | 3MB3681541M050119 | 859 | 2018-05-10 | 1899-12-31 | NULL |
| | 2973 | 3MB3681501M037735 | 859 | 2018-05-10 | 1899-12-31 | NULL |

See 100 rows...

### Top Users

*No top users*

### Data Stewards

*No data stewards*

### Tags

*No tags*

### Properties

CREATED
Nov 11 at 8:51pm

LAST ALTERED
Apr 19 at 10:26pm

OWNER
IVEDOCBI

### Relevant Articles

*No mentioning articles*

# COLUMNS AND RELATIONSHIPS

**F_VEHICLE**
*Please provide a title*

Starred | Watch | Compose | Open With | More...

Overview | Columns 99 | Samples 100 | Filters 0 | Joins 0 | Lineage | Queries 0

Columns — Show 25 rows — Filter

| # | Column | Title | Type | Popularity |
|---|---|---|---|---|
| 0 | BU | Busi | DATE | |
| 1 | LO | Load | TIMESTAMP | |
| 2 | JO | Job I | BIGINT | |
| 3 | DA | Date | DATE | |
| 4 | DA | Date | DATE | |
| 5 | DA | | DATE | |
| 6 | DA | | DATE | |
| 7 | DA | Date | DATE | |
| 8 | DA | | DATE | |
| 9 | DA | | DATE | |
| 10 | DA | Date | DATE | |
| 11 | DA | Date | DATE | |
| 12 | DA | | DATE | |
| 13 | DA | Date | DATE | |
| 14 | DA | Date | DATE | |

# LEGAL TAGS
# GDPR AND OTHER REGULATIONS

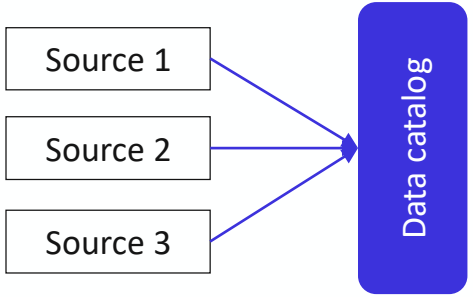
Associate legal tags

- Articles 16-21
- Identify data
  - Right to erasure
  - Right to be forgotten

# CENTRAL VS LOCAL DATA CATALOGS

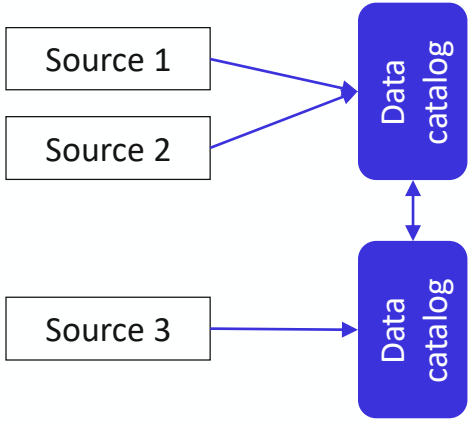
**Central data catalog**

- Integrated views
- Mammoth task
- No redundancy




**Local data catalogs (reality)**

- Legal requirements
- Feasibility
- Tool support very weak ☹



# BIMA-STUDIE 2018 (BARC + SOPRA STERIA CONSULTING) DIGITIZATION HOT SPOTS

- Data quality and meta data management
- Domain knowledge
- Data culture

# IS THE DATA CATALOG A “METADATA MANAGEMENT RELOADED”?

Name it as you like, but there are some critical developments

- Automation, Collective intelligence and expert knowledge
  - Enable crowd sourcing and get help from other users
  - Help to understand quality of data and usage of datasets
  - Rating of information
  - Web application for search / collaboration and API to access metadata
- Governance and legal framework for e.g. GDPR scenarios
  - Capture metadata for security and end-user data consumption
  - Identify the owner of the dataset and get access to source data

Daimler TSS

Daimler TSS GmbH
Wilhelm-Runge-Straße 11, 89081 Ulm / Telefon +49 731 505-06 / Fax +49 731 505-65 99
tss@daimler.com / Internet: www.daimler-tss.com/ Intranet-Portal-Code: @TSS
Sitz und Registergericht: Ulm / HRB-Nr.: 3844 / Geschäftsführung: Martin Haselbach (Chairperson), Steffen Bäuerle