IJK
Institut für Journalistik und Kommunikationsforschung

DATA DRIVEN BUSINESS MODELS

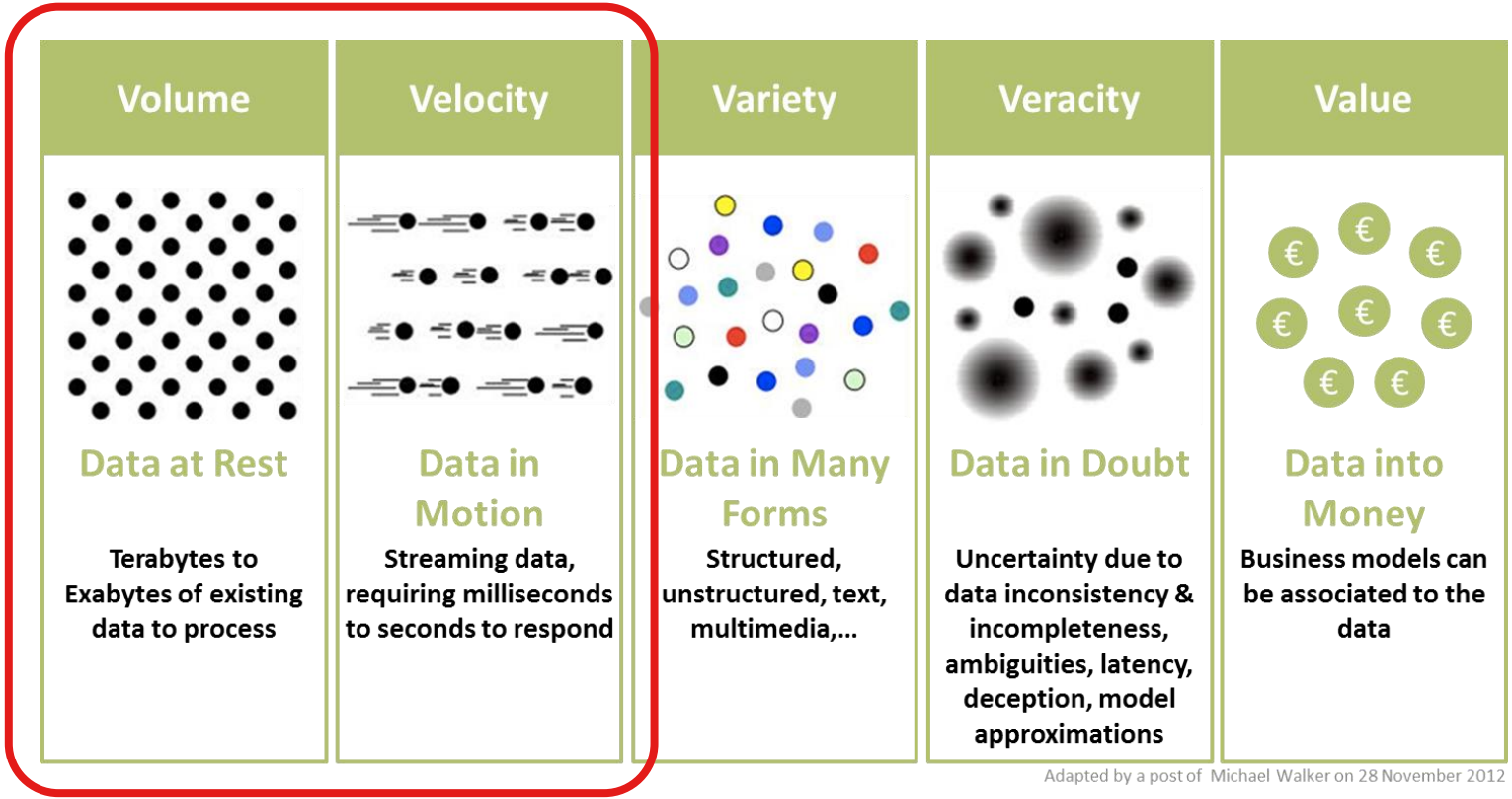
Big Data Technology, Data Science

Nikolaus Pohle, HMTH Hannover, IJK 2017

Store and Process

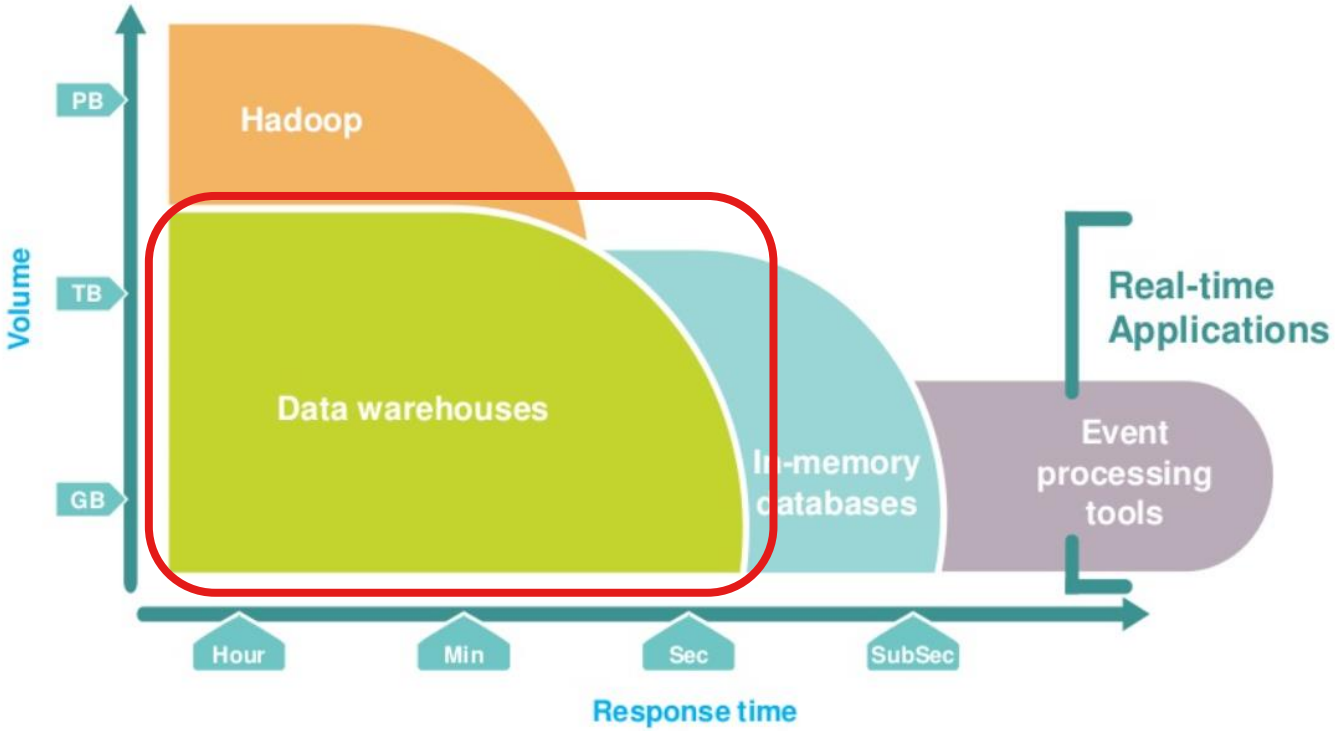
The 5 V‘s of Big Data

IJK
Institut für Journalistik und Kommunikationsforschung



Adapted by a post of Michael Walker on 28 November 2012

https://www.kdnuggets.com/2017/05/must-know-common-data-quality-issues-big-data.html

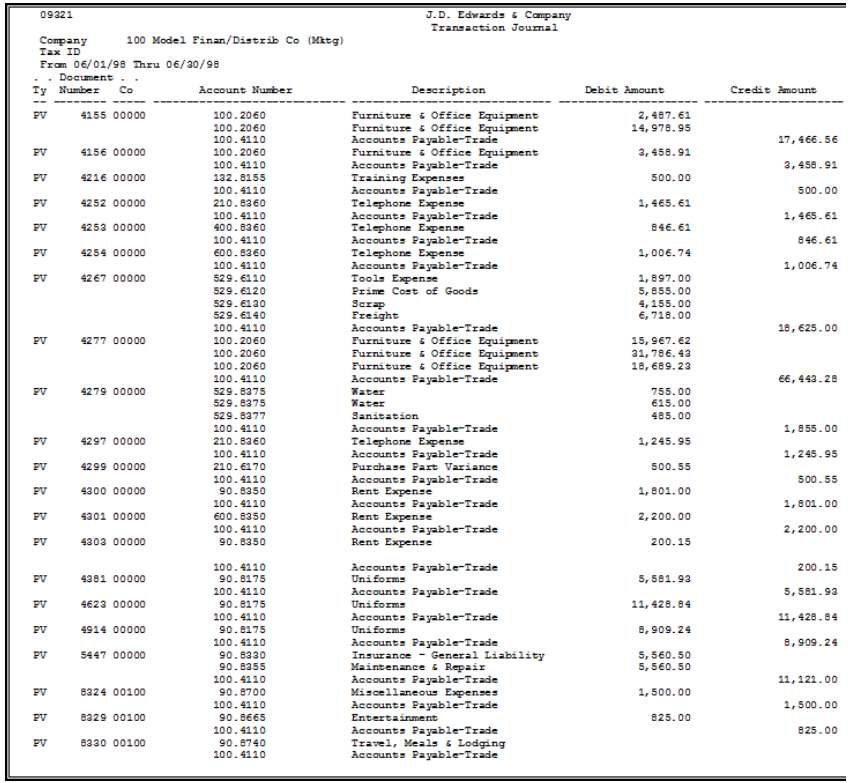


https://www.slideshare.net/capgemini/impact-of-big-data-on-analytics?ref=

OLTP – Transactional Workload

IJK
Institut für Journalistik und Kommunikationsforschung

- Most business maintain a variety of databases to keep track of what happens in the organization.
- These databases are optimized for consistency and speed on a „transactional level“. It‘s fast and safe to update individual records about stock items, accounts receivable etc.
- They are not very good to create summaries and reports because this requires to scan all records.



09321 J.D. Edwards & Company
Transaction Journal

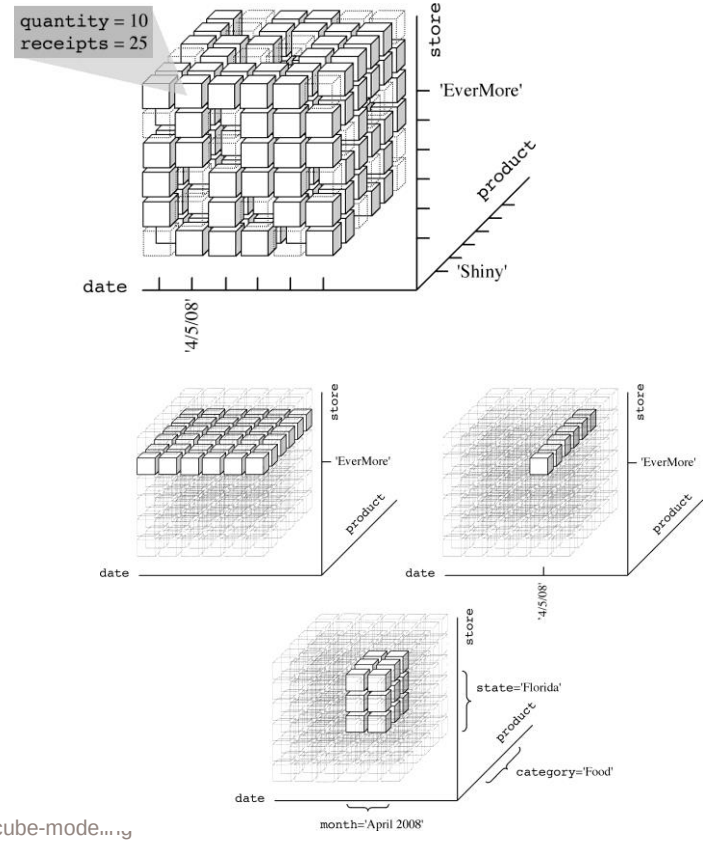
Company 100 Model Finan/Distrib Co (Mktg)
Tax ID
From 06/01/98 Thru 06/30/98

Ty	Document Number	Co	Account Number	Description	Debit Amount	Credit Amount
PV	4155	00000	100.2060	Furniture & Office Equipment	2,487.61	
			100.2060	Furniture & Office Equipment	14,978.95	
			100.4110	Accounts Payable-Trade		17,466.56
PV	4156	00000	100.2060	Furniture & Office Equipment	3,458.91	
			100.4110	Accounts Payable-Trade		3,458.91
PV	4216	00000	132.8155	Training Expenses	500.00	
			100.4110	Accounts Payable-Trade		500.00
PV	4252	00000	210.8360	Telephone Expense	1,465.61	
			100.4110	Accounts Payable-Trade		1,465.61
PV	4253	00000	400.8360	Telephone Expense	846.61	
			100.4110	Accounts Payable-Trade		846.61
PV	4254	00000	600.8360	Telephone Expense	1,006.74	
			100.4110	Accounts Payable-Trade		1,006.74
PV	4267	00000	529.6110	Tools Expense	1,897.00	
			529.6120	Prime Cost of Goods	5,855.00	
			529.6130	Scrap	4,155.00	
			529.6140	Freight	6,718.00	
			100.4110	Accounts Payable-Trade		18,625.00
PV	4277	00000	100.2060	Furniture & Office Equipment	15,967.62	
			100.2060	Furniture & Office Equipment	31,786.43	
			100.2060	Furniture & Office Equipment	18,689.23	
			100.4110	Accounts Payable-Trade		66,443.28
PV	4279	00000	529.8375	Water	755.00	
			529.8375	Water	615.00	
			529.8377	Sanitation	485.00	
			100.4110	Accounts Payable-Trade		1,855.00
PV	4297	00000	210.8360	Telephone Expense	1,245.95	
			100.4110	Accounts Payable-Trade		1,245.95
PV	4299	00000	210.6170	Purchase Part Variance	500.55	
			100.4110	Accounts Payable-Trade		500.55
PV	4300	00000	90.8350	Rent Expense	1,801.00	
			100.4110	Accounts Payable-Trade		1,801.00
PV	4301	00000	600.8350	Rent Expense	2,200.00	
			100.4110	Accounts Payable-Trade		2,200.00
PV	4303	00000	90.8350	Rent Expense	200.15	
			100.4110	Accounts Payable-Trade		200.15
PV	4381	00000	90.8175	Uniforms	5,581.93	
			100.4110	Accounts Payable-Trade		5,581.93
PV	4623	00000	90.8175	Uniforms	11,428.84	
			100.4110	Accounts Payable-Trade		11,428.84
PV	4914	00000	90.8175	Uniforms	8,909.24	
			100.4110	Accounts Payable-Trade		8,909.24
PV	5447	00000	90.8330	Insurance - General Liability	5,560.50	
			90.8355	Maintenance & Repair	5,560.50	
			100.4110	Accounts Payable-Trade		11,121.00
PV	8324	00100	90.8700	Miscellaneous Expenses	1,500.00	
			100.4110	Accounts Payable-Trade		1,500.00
PV	8329	00100	90.8665	Entertainment	825.00	
			100.4110	Accounts Payable-Trade		825.00
PV	8330	00100	90.8740	Travel, Meals & Lodging		
			100.4110	Accounts Payable-Trade		

OLAP – Analytical Workload

IJK
Institut für Journalistik und Kommunikationsforschung

- Instead of scanning through all records for each analytical query specialized databases create „cubes“ that hold pre-aggregated sums, averages etc. for all possible combinations of dimensions.
- In additon some dimensions might be hierarchical which allows to aggregate on several hierarchy levels (e.g. day, month, year)

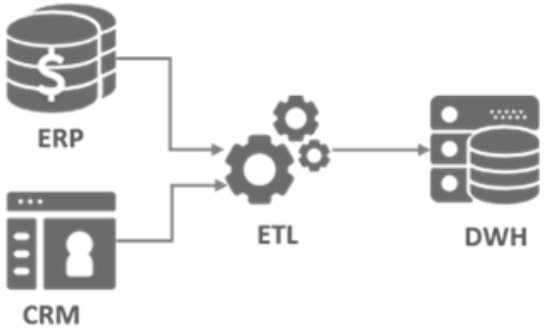


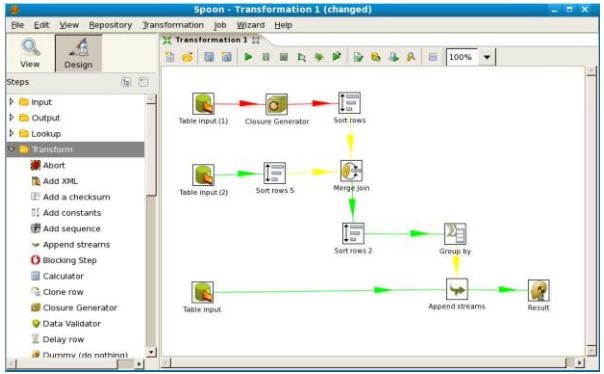
http://searchdatamanagement.techtarget.com/feature/Advantages-of-the-multidimensional-database-model-and-cube-mode...g

ETL – Extract, Transform, Load

IJK
Institut für Journalistik und Kommunikationsforschung

- For defining and peridocally updating OLAP cubes there are specialized tools available that help to
- connect and find relevant data in operational OLTP systems (Extract)
- clean, augment and aggregate data across dimensions and levels (Transform)
- Store aggregated results into the data warehouse (Load)
- The ETL pipelines tend to become complex very quickly and are notoriously hard to maintain

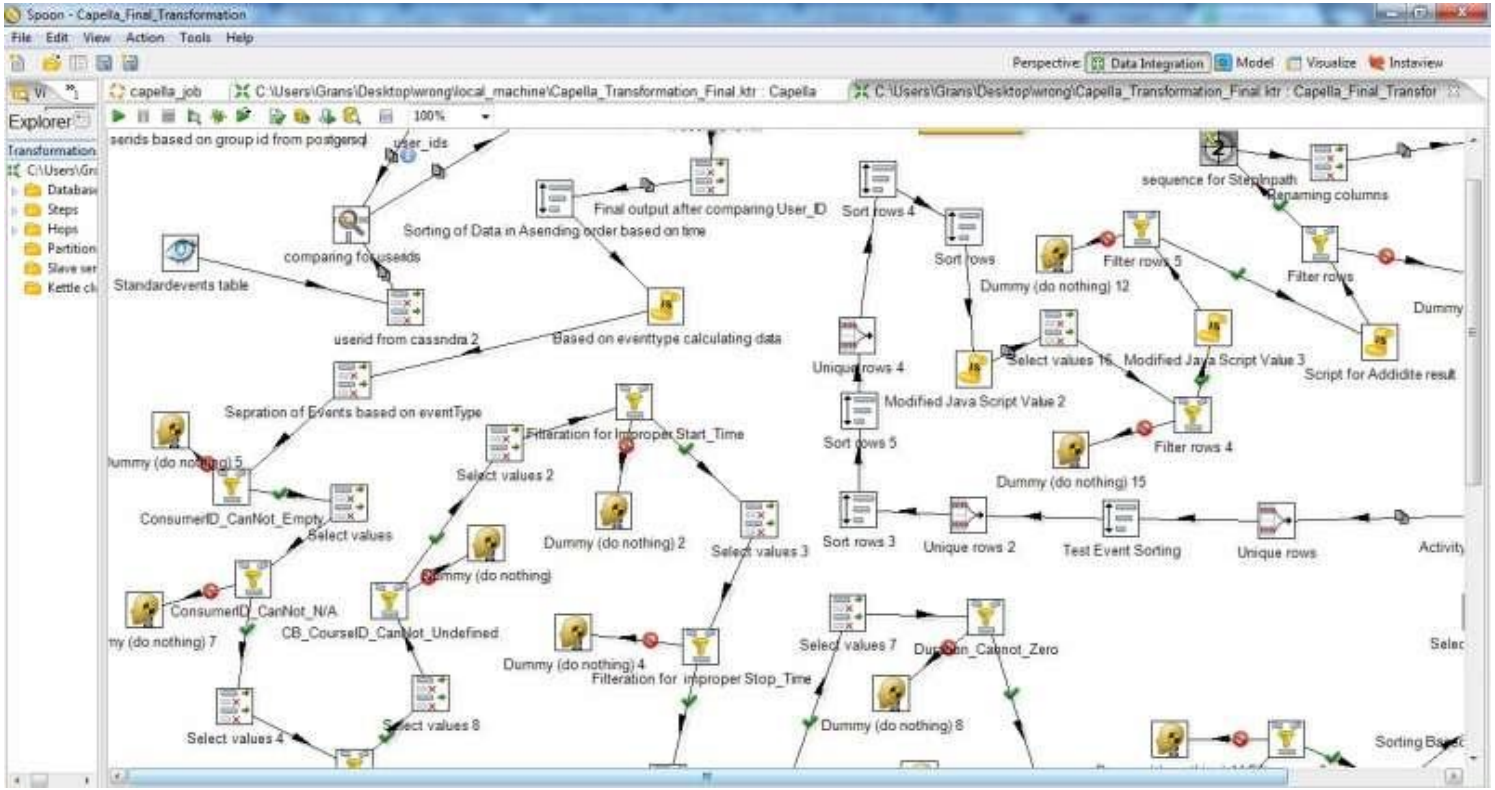




http://www.bizcubed.com.au/pentaho-data-integration-for-database-developers-this-year-training-in-june-sydney/

ETL – Extract, Transform, Load

IJK
Institut für Journalistik und Kommunikationsforschung

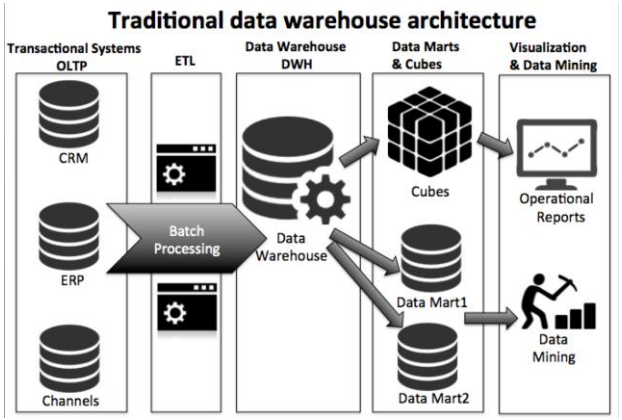


https://www.truelancer.com/portfolio/pentaho-kettle-complex-transformation-example-293

What‘s wrong with that?

IJK
Institut für Journalistik und Kommunikationsforschung

- Data Warehouses do contain aggregate data only. Lack of transactional details.
- Aggregates and Data Prep is „fixed“. Long turn around time for changes.
- Data often is stale. Update frequency often limited to 1-2 days or worse.
- Integration of new data sources can be a tedious process. Some (unstructured) data types cannot be integrated at all
- ETL and DWH are often owned by a team which becomes a bottleneck quickly
- Specialized hardware and commercial tools are very costly (typically starting at >$500k)



- **Data Warehouses cannot keep up with the changing requirements**

Increased number and variety of data sources that generate large quantities of data.

Realization that data is “too valuable” to delete.

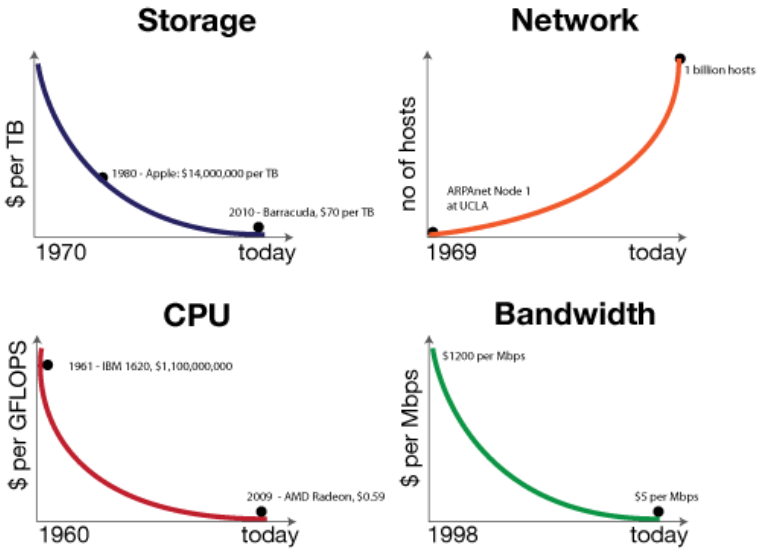
Dramatic decline in the cost of hardware, especially storage.

http://slideplayer.com/slide/10757535/

Technology allows to Scale Out

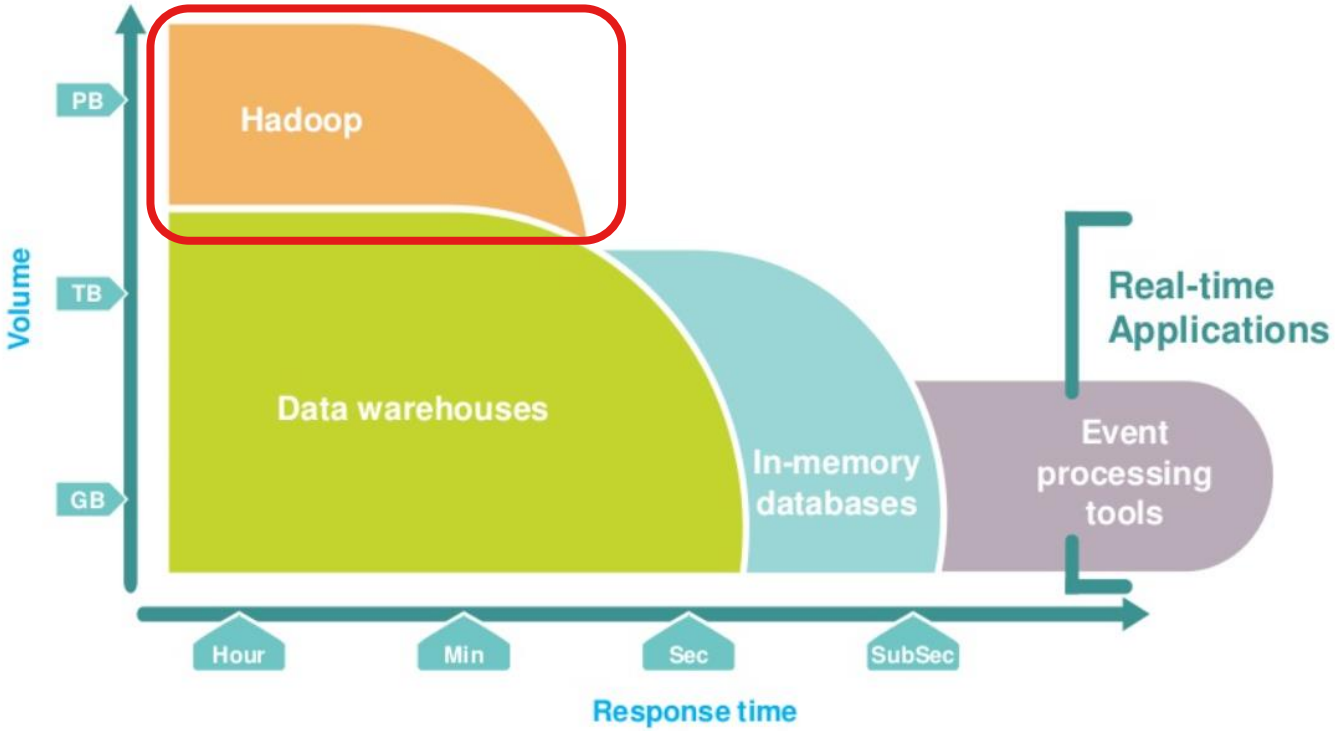
IJK
Institut für Journalistik und Kommunikationsforschung

- Historically storage was very expensive. You ought to be very clever about how to minimize storage requirements.
- Traditional OLAP systems therefore use ETL to aggregate early andd transfer data to where the logic is.
- Using commodity hardware, cost of storage and network bandwidth is dramatically decreasing



- Keep all the data and run analytics directly on the raw data.

http://radar.oreilly.com/2011/08/building-data-startups.html

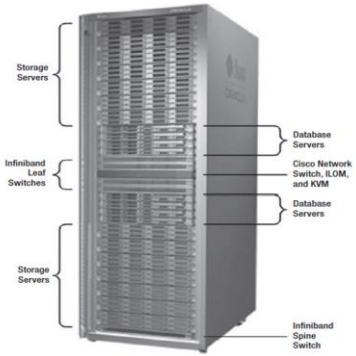


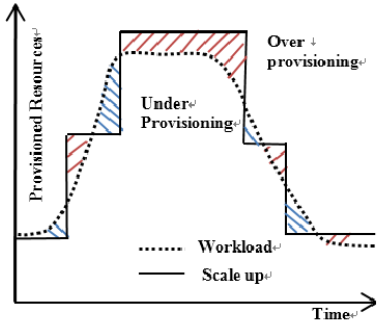
https://www.slideshare.net/capgemini/impact-of-big-data-on-analytics?ref=

Scale-Up vs Scale-Out

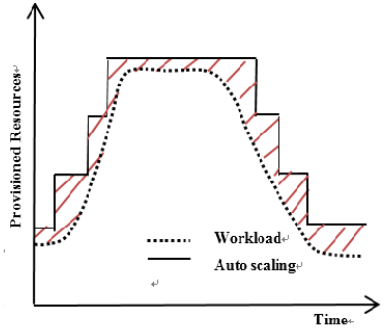
IJK
Institut für Journalistik und Kommunikationsforschung

- Traditional systems have been able to scale up. You have to buy a bigger machine to prepare for heavier workloads.
- Economically very challenging to plan for. What is the right timing? What do you do with the old machine?
- The decline of network cost allows for a much more „elastic“ approach by combining many small machines on-demand.





(a) Scale-up instance type (capacity)



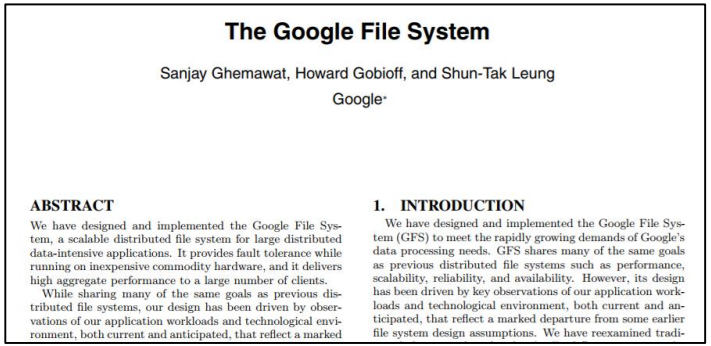
(b) Scale-out in instance quantity

https://www.semanticscholar.org/paper/Cloud-Performance-Modeling-with-Benchmark-Evaluati-Hwang-Bai/457b3796442d5944259d49f2f39c92767ffc3041

Add Software to the mix.

- If you can access all your data in its raw form you have maximum flexibility.
- Because it‘s a lot of data your need to distribute the load across many machines
- This happens to be very similar to what search engines do.

- Google popularized the idea of „distributed storage“ and „distributed processing“ by publishing papers about how they approach the search problem



The Google File System

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung

Google*

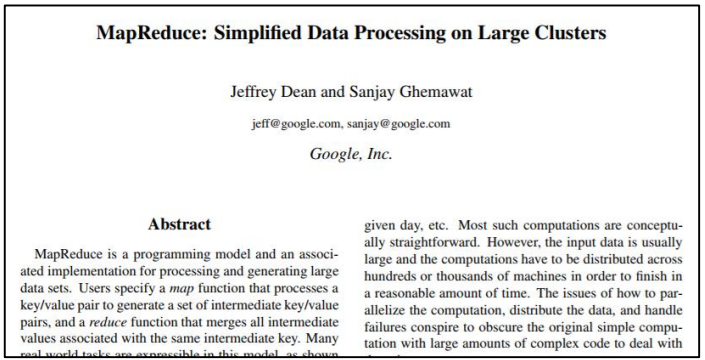
ABSTRACT

We have designed and implemented the Google File System, a scalable distributed file system for large distributed data-intensive applications. It provides fault tolerance while running on inexpensive commodity hardware, and it delivers high aggregate performance to a large number of clients.

While sharing many of the same goals as previous distributed file systems, our design has been driven by observations of our application workloads and technological environment, both current and anticipated, that reflect a marked

1. INTRODUCTION

We have designed and implemented the Google File System (GFS) to meet the rapidly growing demands of Google's data processing needs. GFS shares many of the same goals as previous distributed file systems such as performance, scalability, reliability, and availability. However, its design has been driven by key observations of our application workloads and technological environment, both current and anticipated, that reflect a marked departure from some earlier file system design assumptions. We have reexamined tradi-



MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.

Abstract

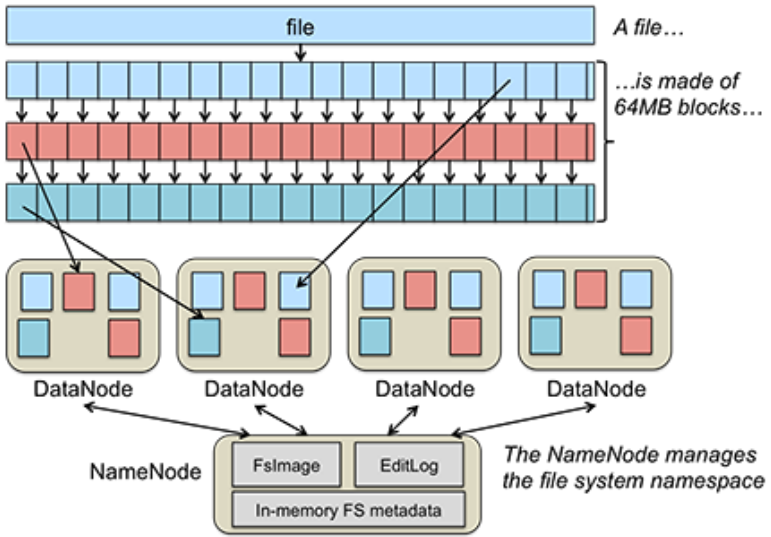
MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown

given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures conspire to obscure the original simple computation with large amounts of complex code to deal with

Hadoop implements an Open Source Version of GFS

- Yahoo Engineers read the Google Papers and start building their own version called Hadoop

- Data is stored in flat files which are split up and distributed across many (small) machines. Blocks are stored redundantly because failure of single machines is inevitable.
- Adding / Removing machines allows to scale up and down.



https://apachesparktraining.wordpress.com/2014/07/31/introduction-to-the-hadoop-ecosystem/

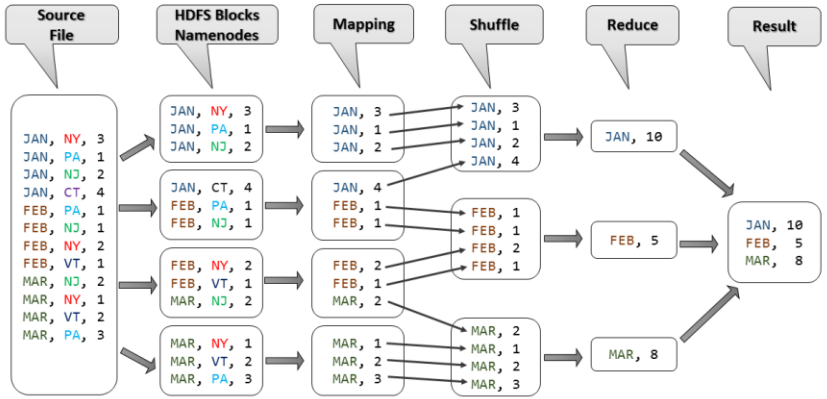
Hadoop implements an Open Source Version of M/R

IJK
Institut für Journalistik und Kommunikationsforschung

- The same Yahoo engineers also build a data processing engine into Hadoop utilizing the „Map Reduce“ approach.

- Very simple and pretty flexible
- Very easy to parallelize
- Leverages data locality of HDFS

- Hard to apply to complex questions



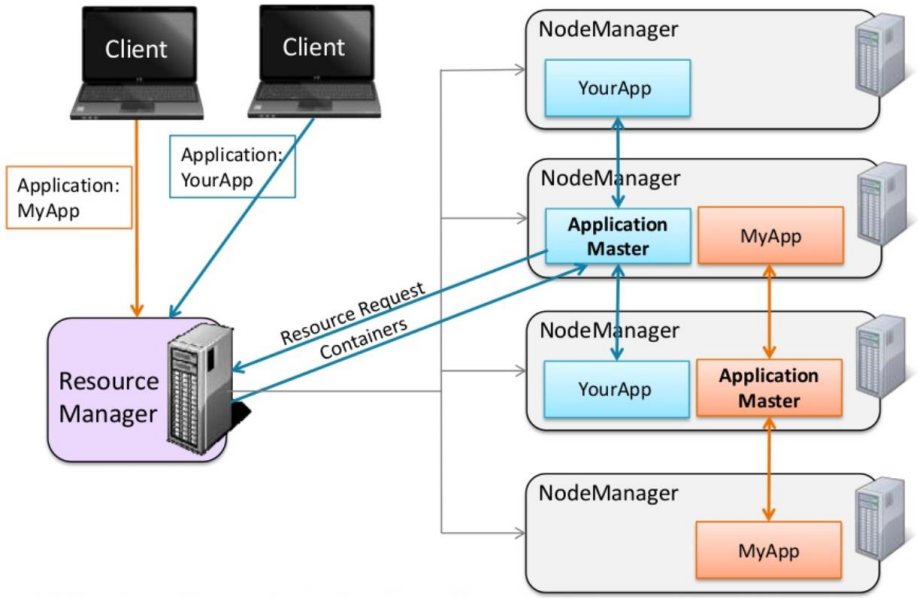
https://apachesparktraining.wordpress.com/2014/07/31/introduction-to-the-hadoop-ecosystem/

Storage is close but separate from Processing

IJK
Institut für Journalistik und Kommunikationsforschung

- Data has inertia – it‘s hard to move
- Big Data even more so
- If data sits in a distributed system it‘s easier to run your analytical application where the data is then moving data to the application.

- Hadoop supports this by leveraging info about „data locality“



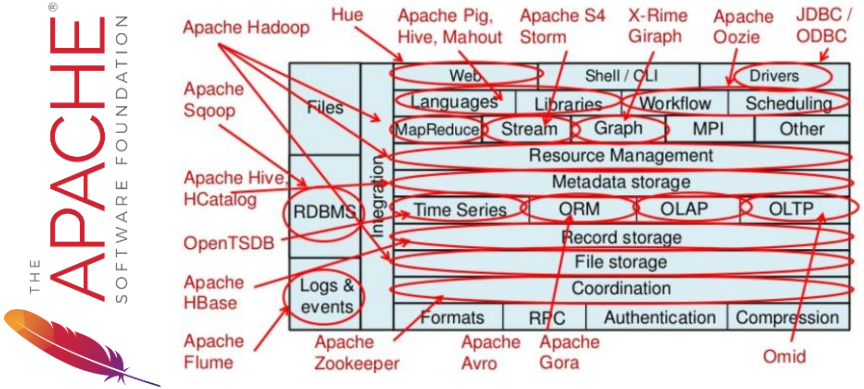
https://www.slideshare.net/cloudera/introduction-to-yarn-and-mapreduce-2

Apache Foundation vs Commercial Distributions

IJK
Institut für Journalistik und Kommunikationsforschung

- Many relevant technology components for Big Data are Open Source. Core elements of the stack are maintained under the umbrella of the Apache Software Fundation.

- Enterprise users typically need a stable stack that is proven to works well together. Also they need a phone numer in case things get



HORTONWORKS

cloudera

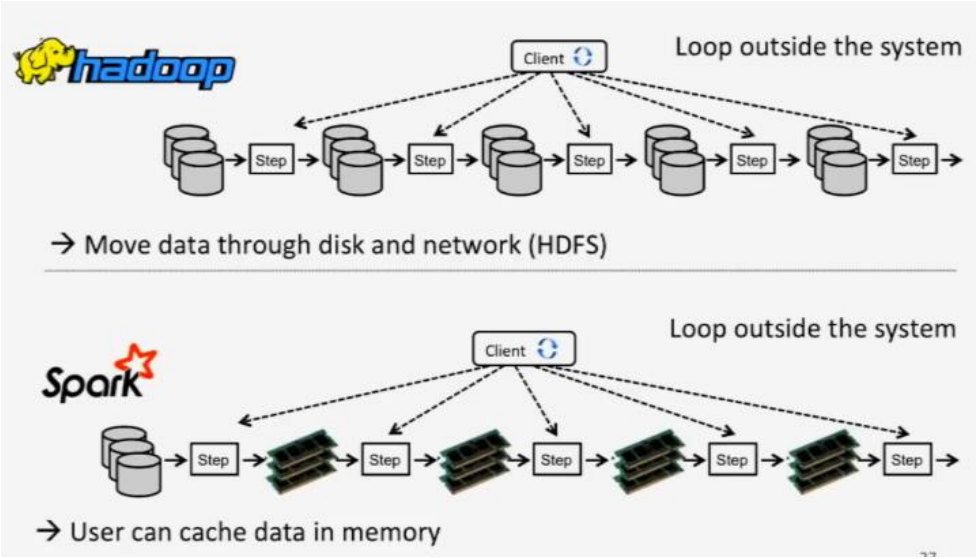
MAPR

https://www.slideshare.net/cloudera/the-hadoop-stack-then-now-and-in-the-future-eli-collins-charles-zedlewski-cloudera (from 2011!)

Processing becomes more powerful when in-memory

IJK
Institut für Journalistik und Kommunikationsforschung

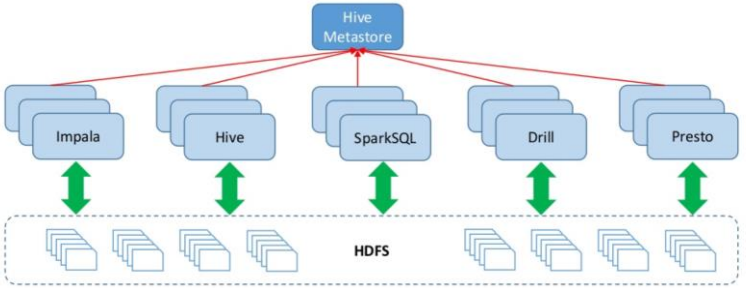
- Map Reduce heavily relies on writing temporary results to disc.
- Spark leverages fast memory to cache temporary data between steps.
- Spark offers a richer API that is allows for more complex steps than Map Reduce would enable.
- Multiple language bindings: JAVA, Scala, Python, R
- Spark can even translate SQL queries to its API calls.



https://www.nextplatform.com/2015/02/22/flink-sparks-next-wave-of-distributed-data-processing/

Storage vs. Query Engines

- Applications are either custom made using Python, JAVA, etc. or written in a high level scripting language like PIG or SQL
- SQL allows easy integration with BI tools which are often SQL-based
- There are competing engines optimized for specific use cases.



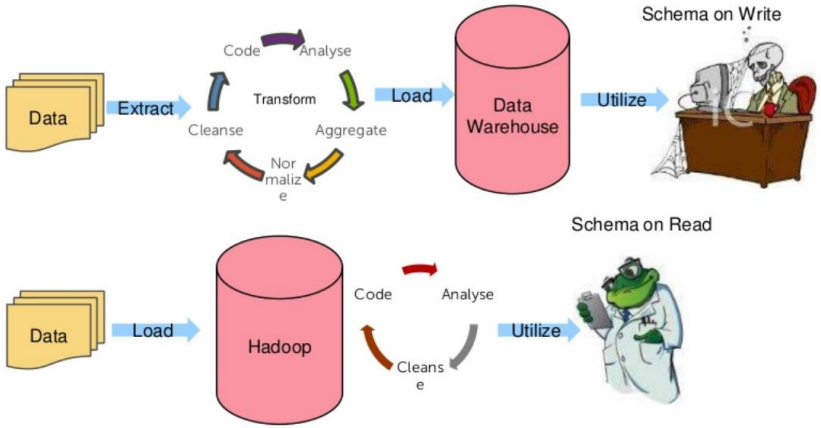
- The way you store your data greatly impacts the efficiency of your query.
- column-based vs. row-based

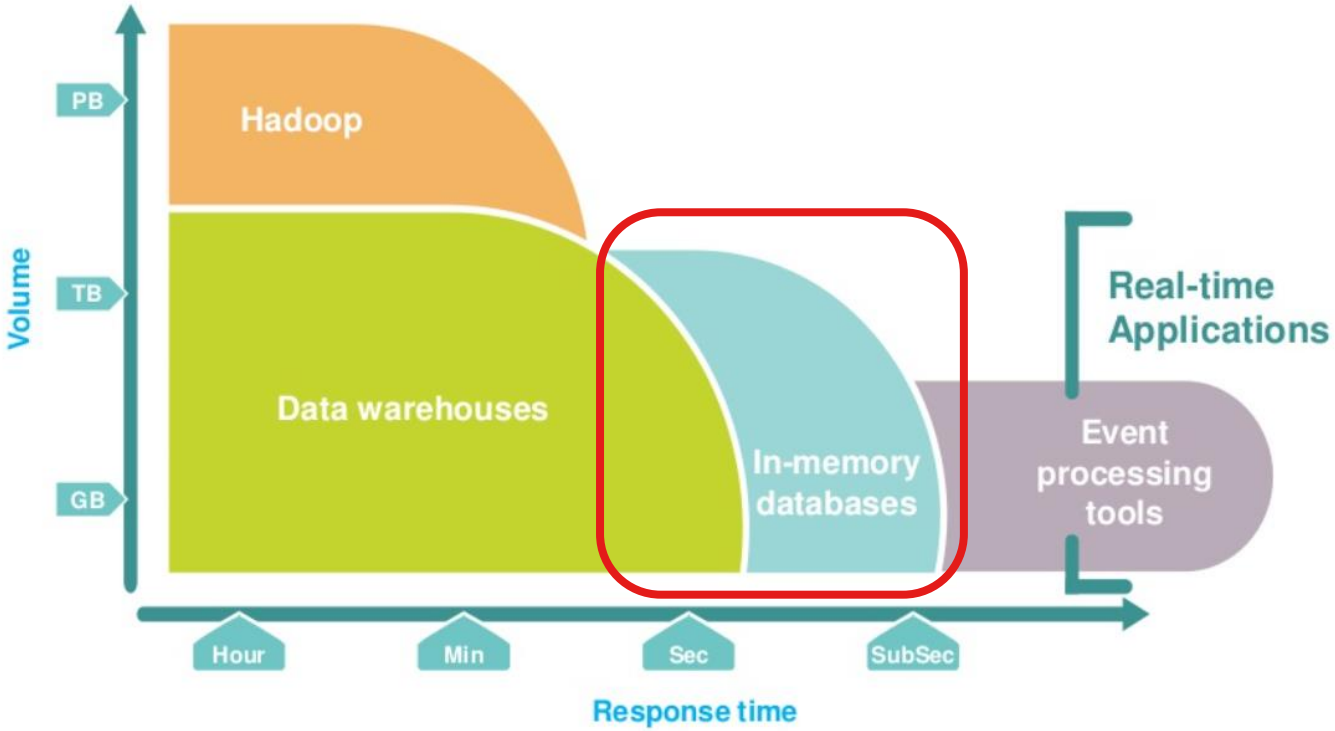
https://www.jowanza.com/blog/which-hadoop-file-format-should-i-use

„Schema on Read“

- Storing all the raw data prevents you from knowing all analytical use cases upfront.

- You cannot and want not to optimize for certain queries early in the process (=while writing the data).
- Instead you decide how to interprete the data once you know what the actual question is (=while reading the data)



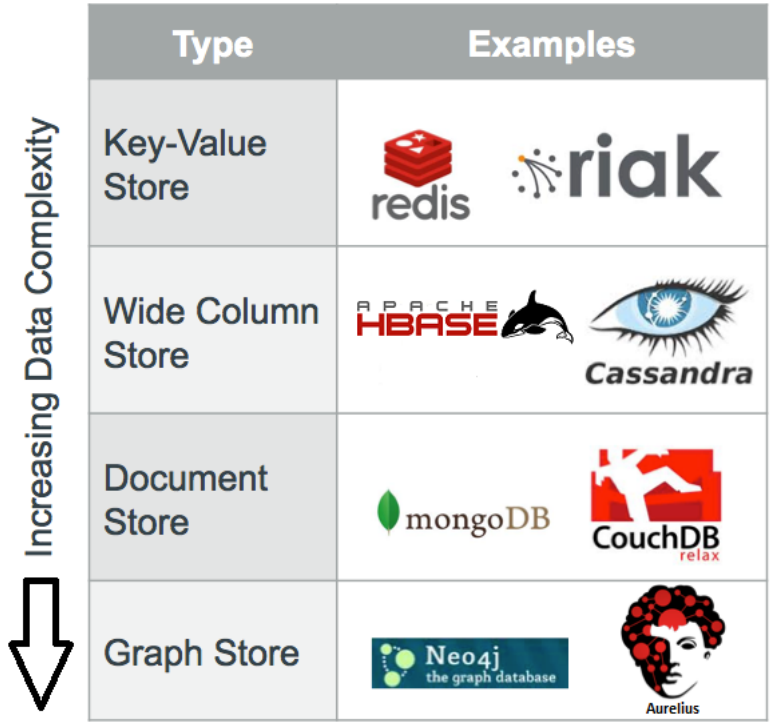


https://www.slideshare.net/capgemini/impact-of-big-data-on-analytics?ref=

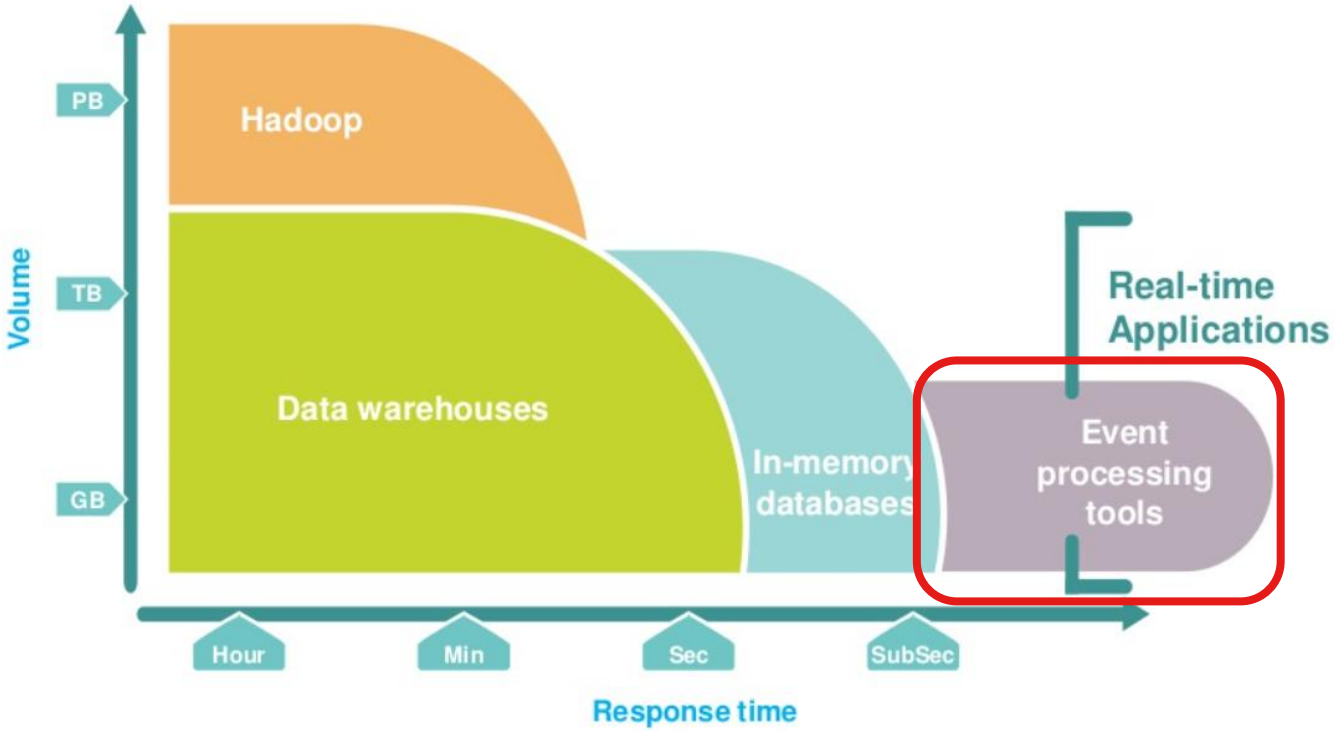
NoSQL

- Sometimes storing data in a flat file is not enough. Especially if you need to optimize for certain access patterns.
- A lot of innovation happend in this area during the last 10 years.
- Some of these (e.g. HBase) do integate with Hadoop for data storage and metadata management.

- Most of these engines are of limited use for analytical queries.



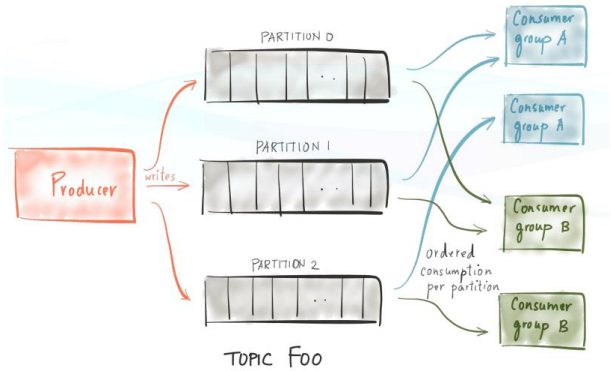
https://www.linkedin.com/pulse/comprehensive-analysis-nosql-vs-rdbms-rassul-fazelat/



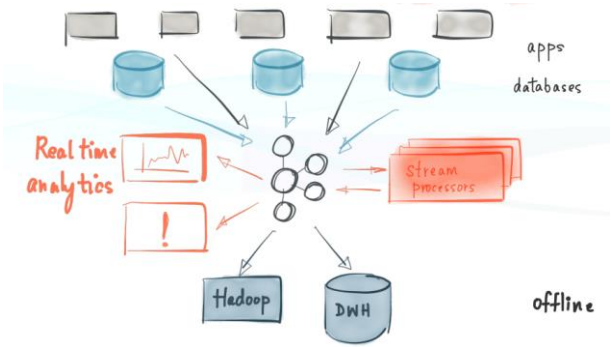
https://www.slideshare.net/capgemini/impact-of-big-data-on-analytics?ref=

Stream and Realtime

- The Hadoop ecosystem has been optimized for high throughput and for that sacrificied latency.
- For workloads that do require very low latency there are data systems that apply analytical functions while data is streaming in. Often these system use the noition of a message based Publish / Subscribe Queue.



- Hard problems:
- Exactly once delivery.
- Order of Delivery.



STORM

APACHE kafka

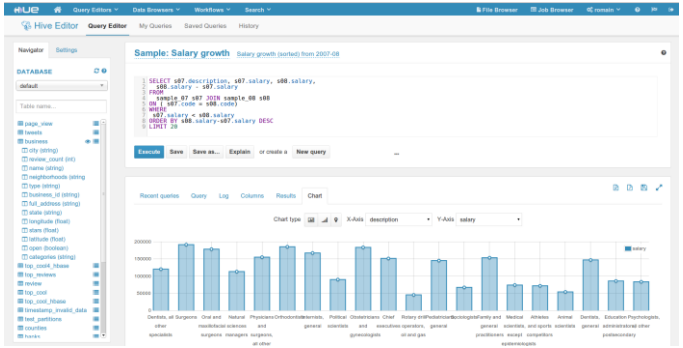
https://blog.parse.ly/post/3886/pykafka-now/

Analyse and Visualize

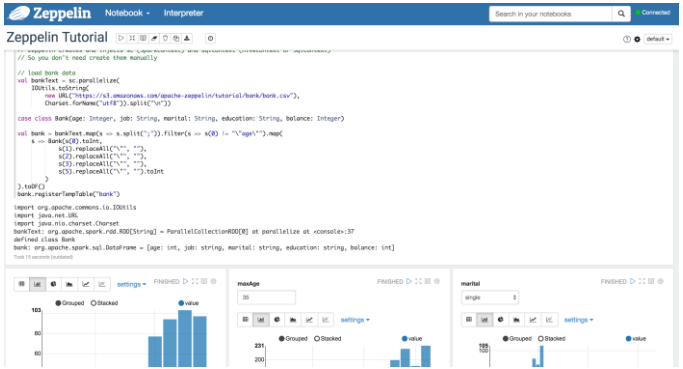
Simplistic Hadoop User Interfaces

IJK
Institut für Journalistik und Kommunikationsforschung

- Apache Hue
- HDFS File Browser
- SQL Query Workbench
- Design and Schedule workflows
- Manage Processing Job



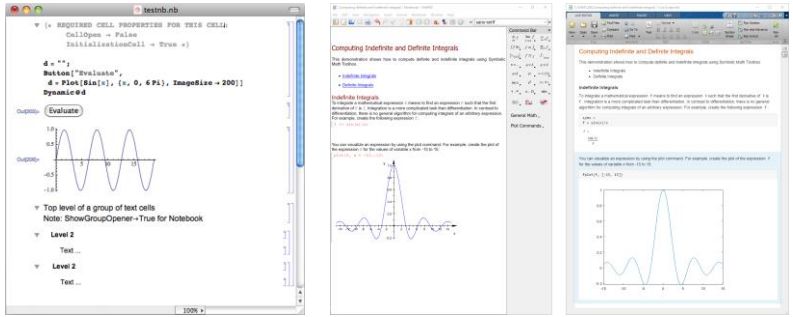
- Apache Zeppelin
- Interactive Data Science Notebooks
- Data Visualizations and Forms
- Supports multiple interpreters: R, Python, SQL



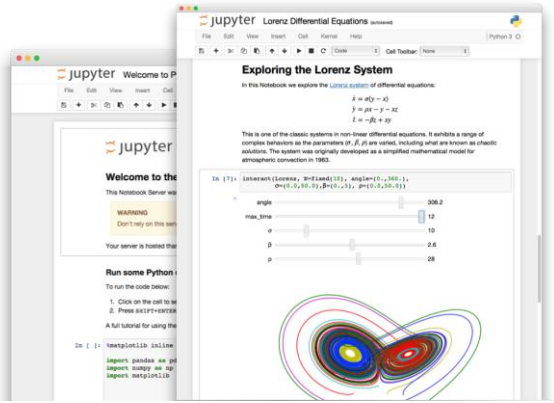
Notebooks are pretty popular too

IJK
Institut für Journalistik und Kommunikationsforschung

- Notebooks mix text, graphics and code
- Originates from MATLAB and Mathematica
- Browser-based IPython started ca. 2001
- Renamed to Apache Jupyther.



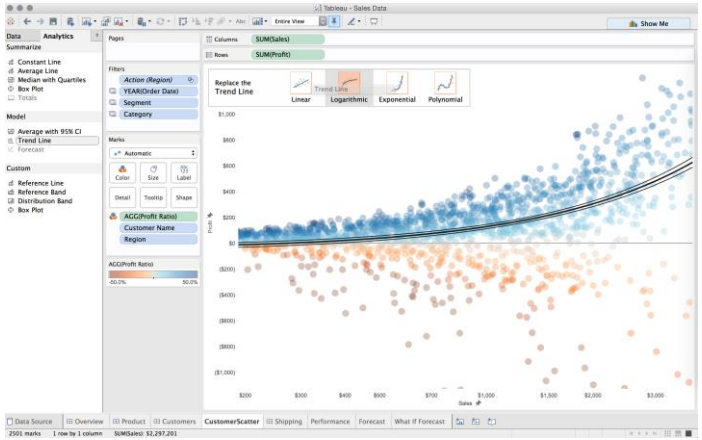
- Interactive execution of >40 languages
- Output of code execution inline
- Notebooks are stored as JSON files which makes it easy to version them
- Great tool for collaboration



BI Tooling vs Data Science Tooling

IJK
Institut für Journalistik und Kommunikationsforschung

- Self-Service Visual Data Exploration and Dashboard Generation
- Department Level users and subject matter experts without engineering or database background.



Power BI

REVOLUTION ANALYTICS

SPSS

pandas

$y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$

Qlik

tableau

platfora

sas

rapidminer

python

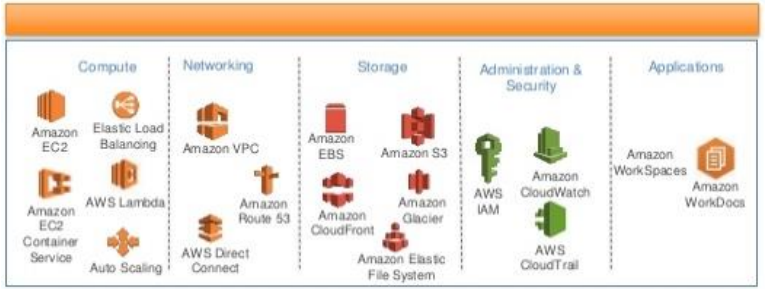
Data in the Clouds

Cloud

IJK
Institut für Journalistik und Kommunikationsforschung

- Public Cloud providers offer unlimited scale for global storage and compute
- Typically storage is independent from compute so you can scale both independently

- In top of the infrastructure there is a plethora of managed services, among them databases and big data clusters



Cloud Managed Data Services

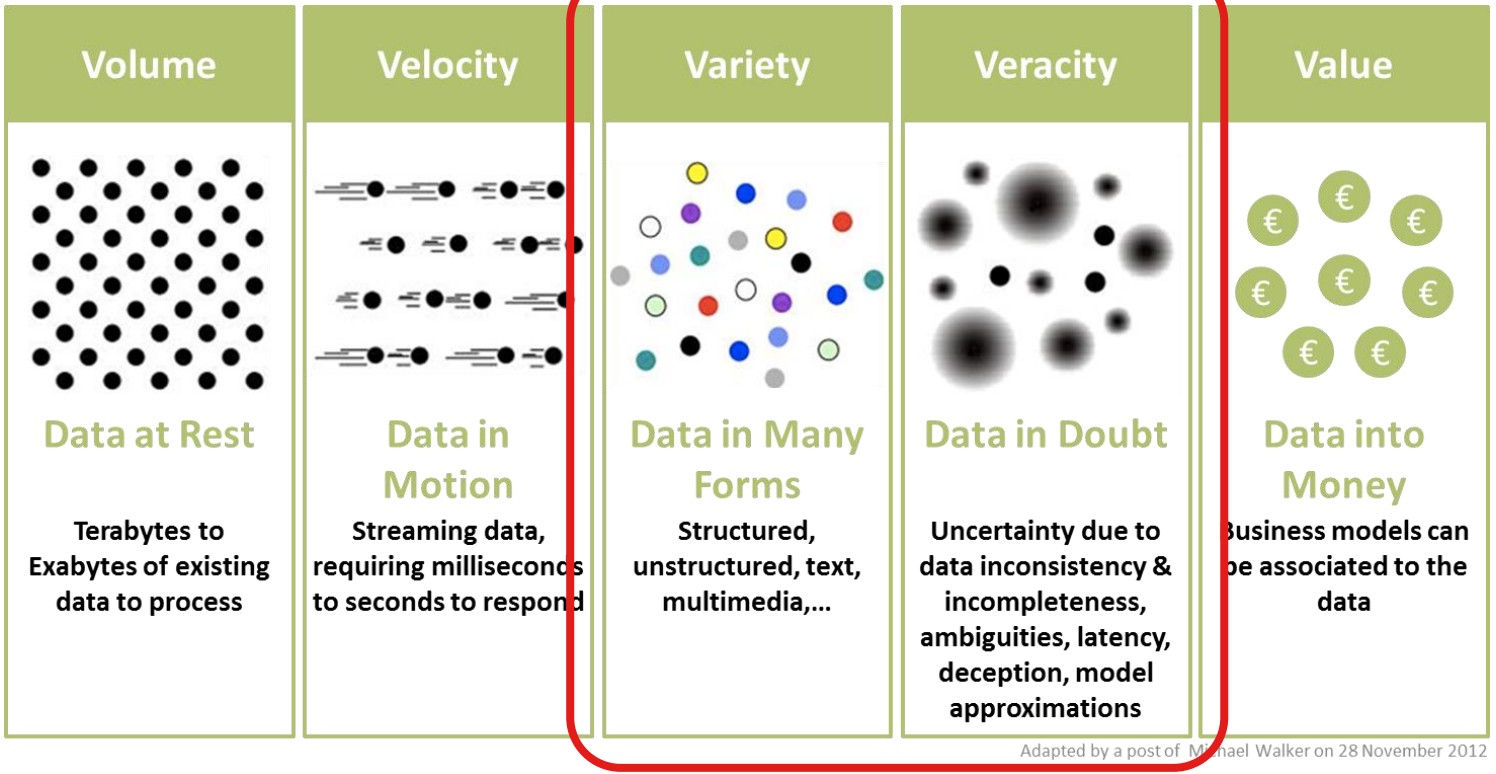
IJK
Institut für Journalistik und Kommunikationsforschung

				
Compute	*n/a*	EC2	Compute	GCE
Storage	*HDFS*	S3	Blob	GCS
Database	*HBase*	DynamoDB	CosmosDB	BigQuery
Processing	*Hadoop, Spark*	EMR, Glue	HDInsight	DataProc
Stream	*Kafka*	Kinesis	Stream Analytics	DataFlow
Notebook	*Zeppelin*	Athena	Notebooks	DataLab
AI	*MLlib*	AWS AI	Azure ML	TensorFlow

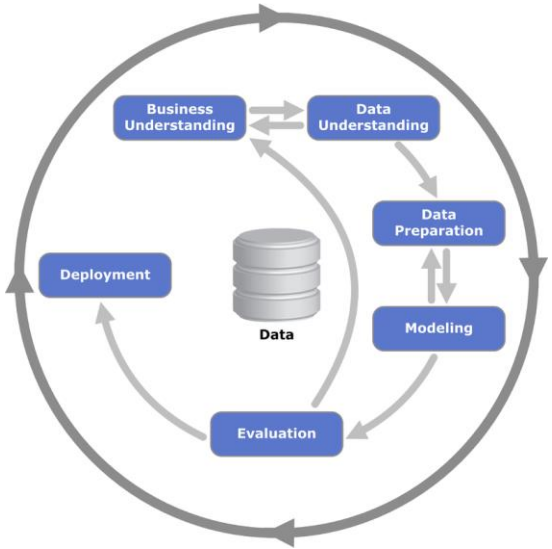
Describe and Govern

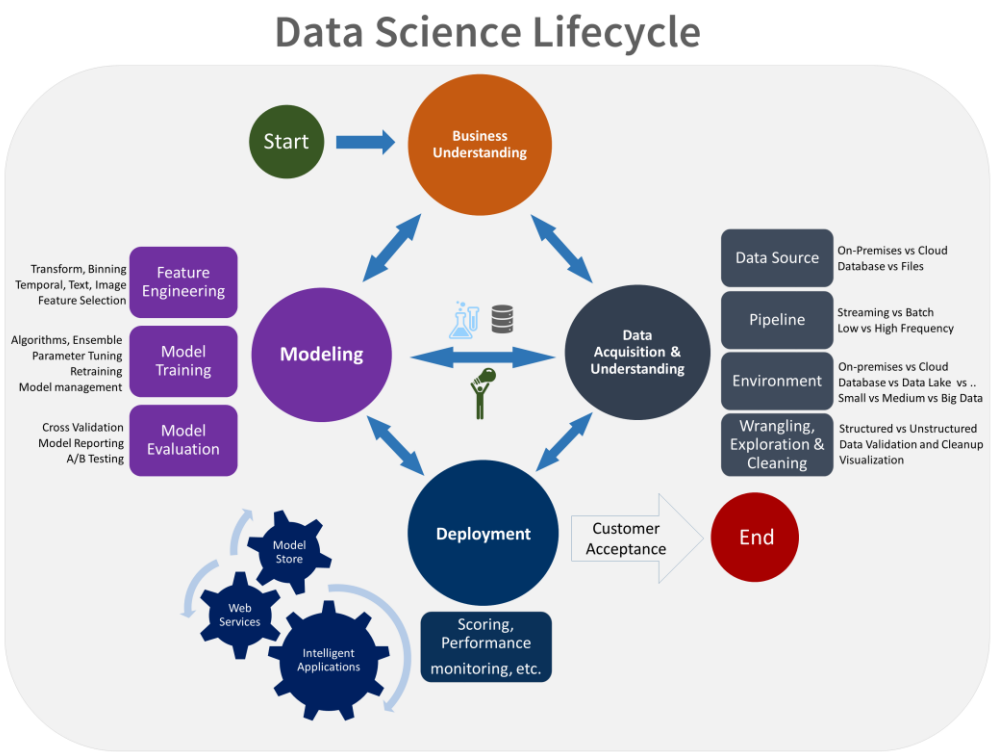
The 5 V‘s of Big Data

IJK
Institut für Journalistik und Kommunikationsforschung



https://www.kdnuggets.com/2017/05/must-know-common-data-quality-issues-big-data.html





https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining
https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview

Example 1: GfK Data Activation

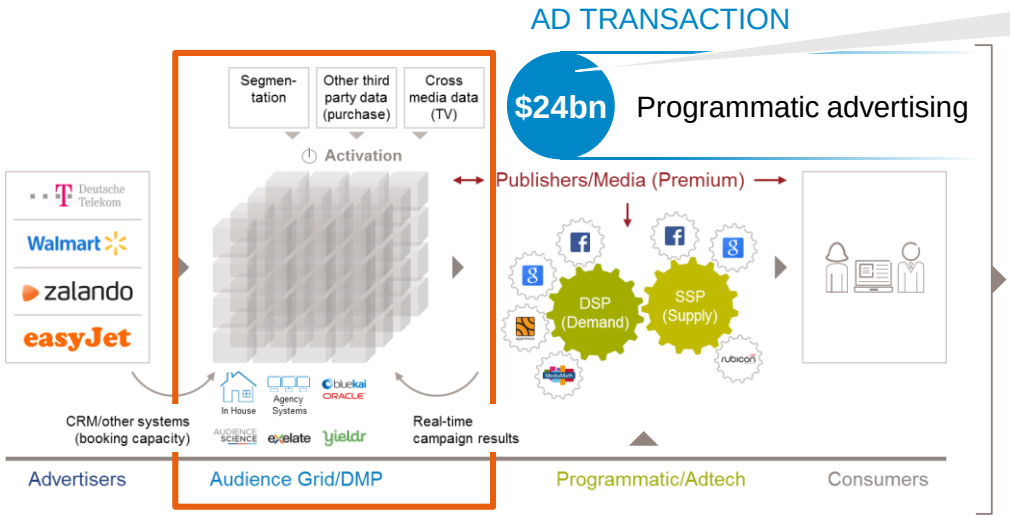
Working with DMP partners as platforms allows GfK to address new business

IJK
Institut für Journalistik und Kommunikationsforschung

Marketing Services Initiative

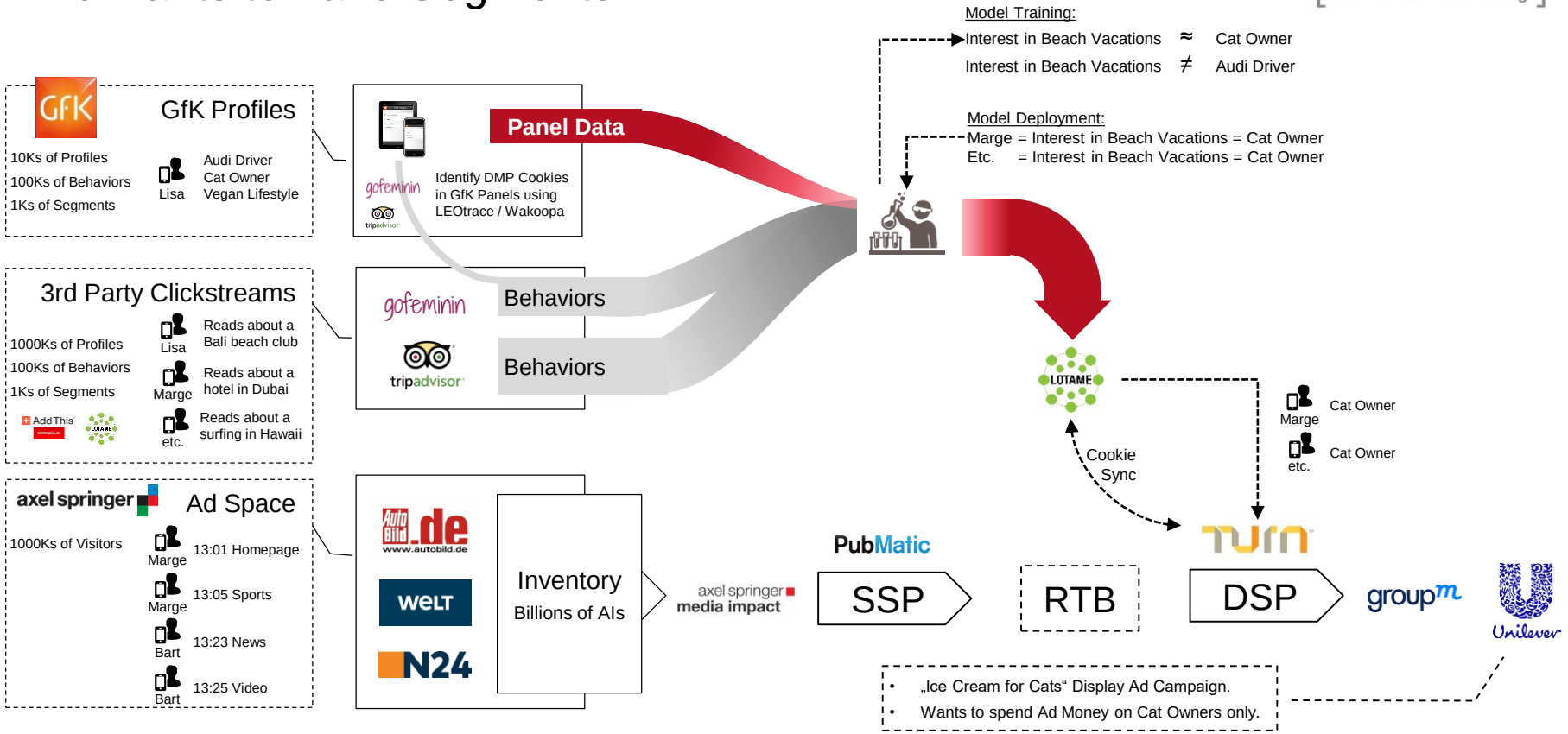
Working with DMPs opens up four different growth opportunities

1. Standard Audience Segments
2. Custom Audience Segments
3. Personalization, Creative Ad Optimization
4. Measurement and Attribution

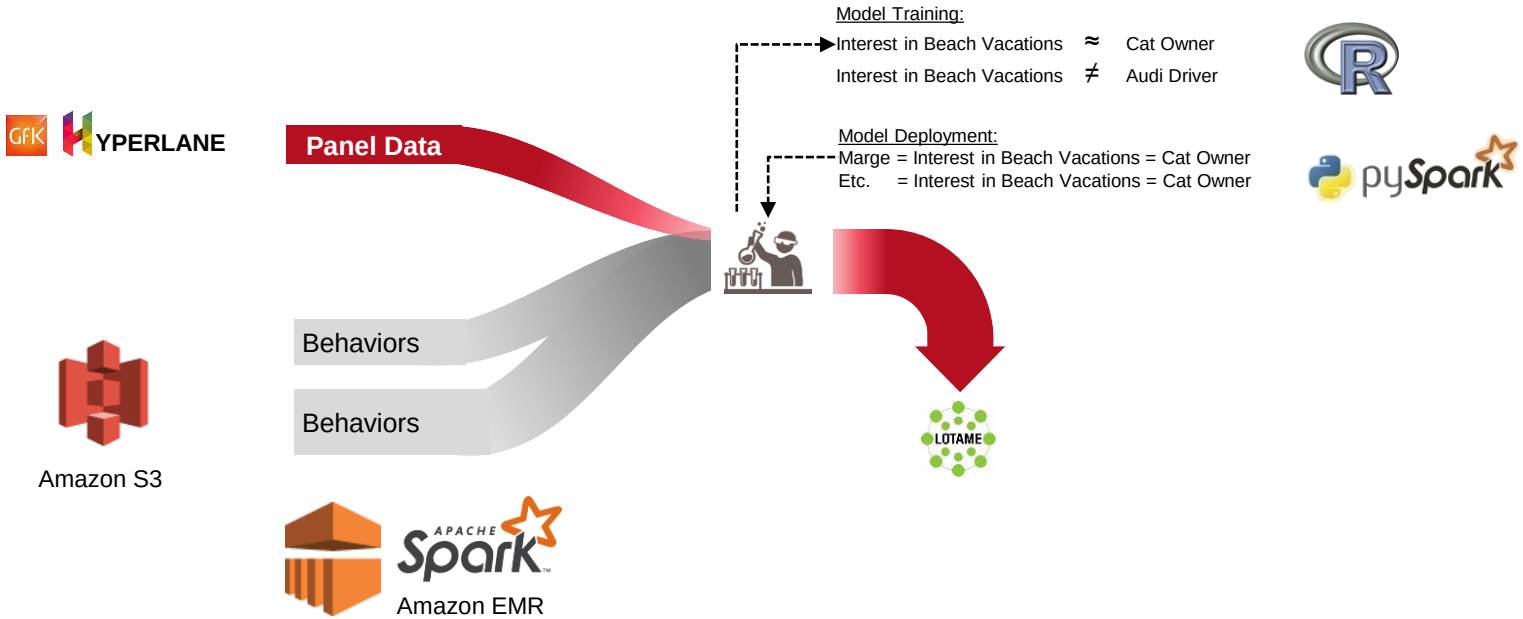


- Cross-media reach (xGRP)
- Online and offline purchases (ROI)
- Brand lift

Who wants to have Segments?



Who wants to have Segments?



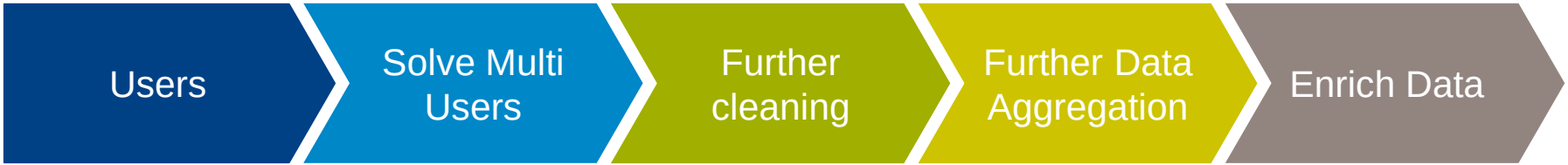
Modelling Lotame Data
Preprocessing:

IJK
Institut für Journalistik und Kommunikationsforschung

- Aggregating hierarchy nodes (optional)
- Remove duplicated columns

- Delete sparse Cookies / Behaviours

- Add Geomarketing Data

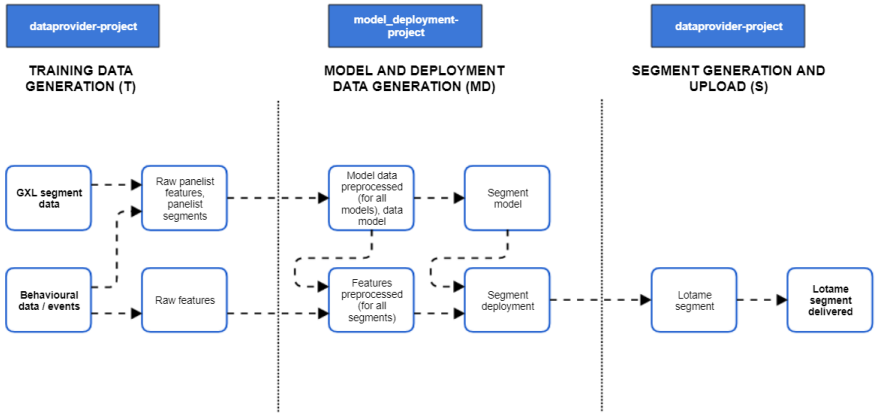


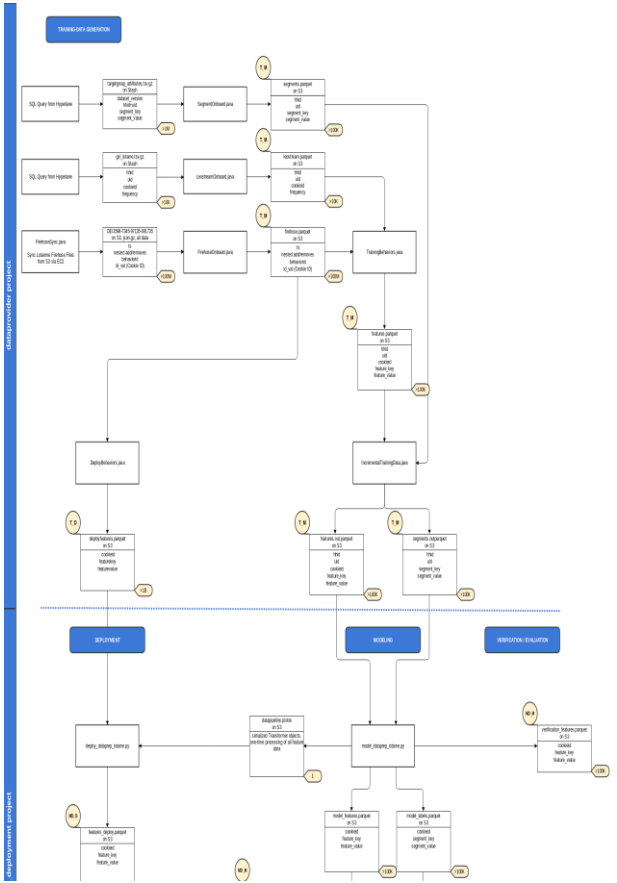
- If a cookie has multiple uid's we solve the mulit-user problem the following way (If the behavior differs)
- Assign the behavior randomly (with observed frequencies)

- Principal Component Analysis
- Aggregation by Correlation
- Sober (developed within GfK)

Pipelines quickly become complex

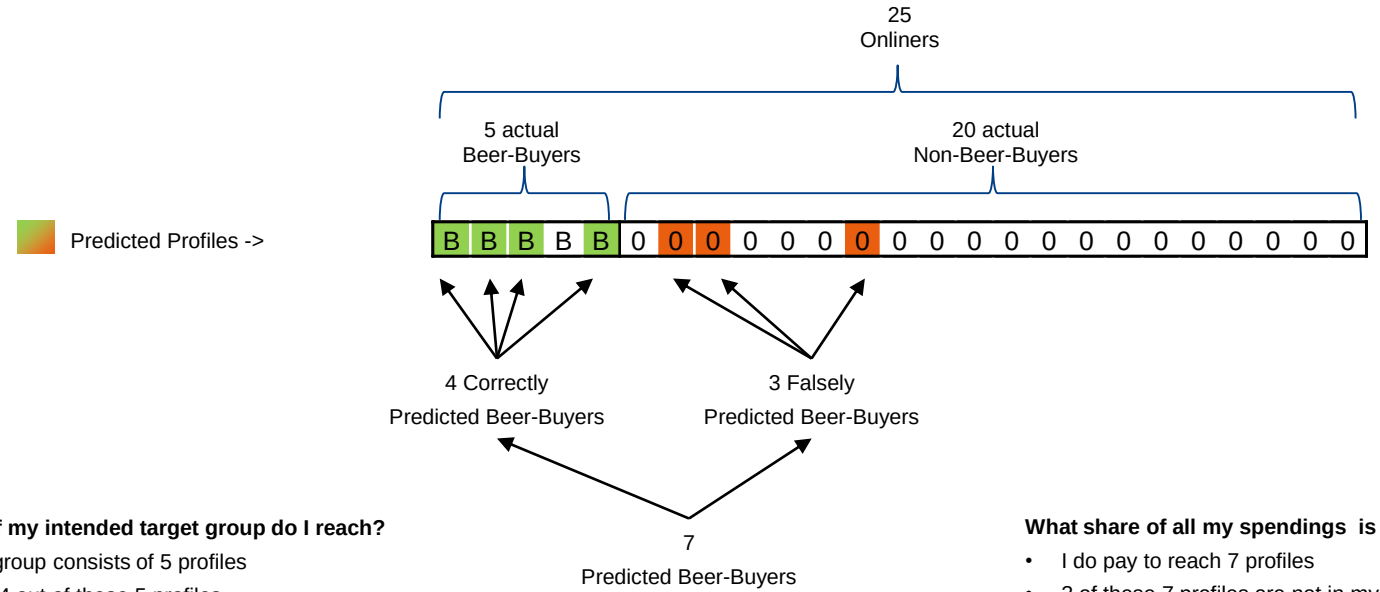
IJK
Institut für Journalistik und Kommunikationsforschung





How do we define our Quality?

IJK
Institut für Journalistik und Kommunikationsforschung



What share of my intended target group do I reach?

- My target group consists of 5 profiles
- I do reach 4 out of these 5 profiles
- this means I cover 80% of my target group

What share of all my spendings is wasted?

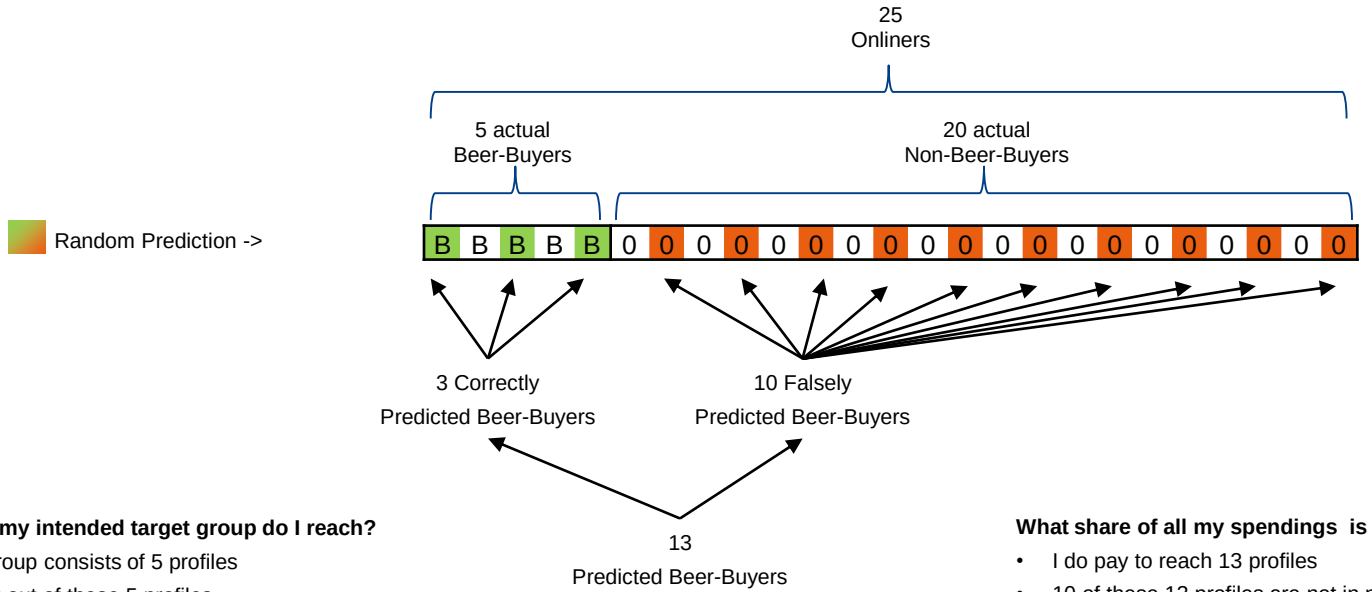
- I do pay to reach 7 profiles
- 3 of these 7 profiles are not in my target group
- this means 43% of my spendings are wasted

$$Hit\ Rate = \frac{TP}{P} = \frac{True\ Positive}{Actual\ Target\ Group} = \frac{4}{5} = 80\%$$

$$Waste = \frac{FP}{TP+FP} = \frac{False\ Positive}{All\ Predicted} = \frac{3}{7} = 43\%$$

What is good quality? Better than the random baseline:

IJK
Institut für Journalistik und Kommunikationsforschung



What share of my intended target group do I reach?

- My target group consists of 5 profiles
- I do reach 3 out of these 5 profiles
- this means I cover 60% of my target group

What share of all my spendings is wasted?

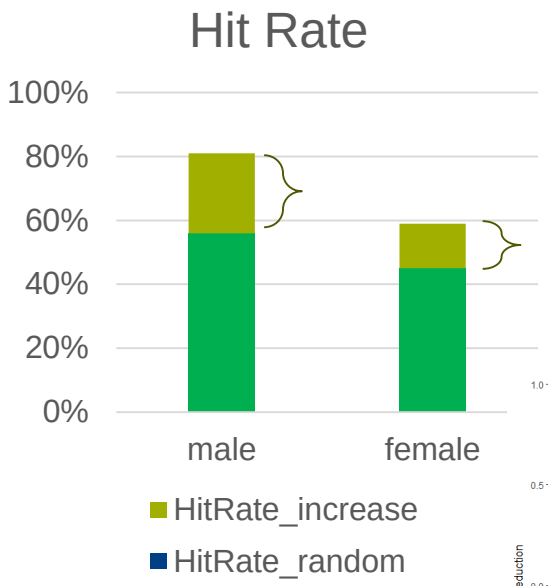
- I do pay to reach 13 profiles
- 10 of these 13 profiles are not in my target group
- this means 77% of my spendings are wasted

$$Hit\ Rate = \frac{TP}{P} = \frac{True\ Positive}{Actual\ Target\ Group} = \frac{3}{5} = 60\%$$

$$Waste = \frac{FP}{TP+FP} = \frac{False\ Positive}{All\ Predicted} = \frac{10}{13} = 77\%$$

How do we visualize these relations in a consice way?

IJK
Institut für Journalistik und Kommunikationsforschung

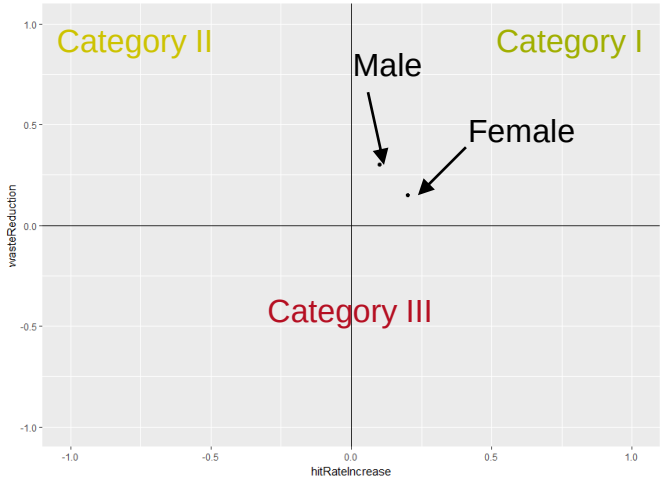


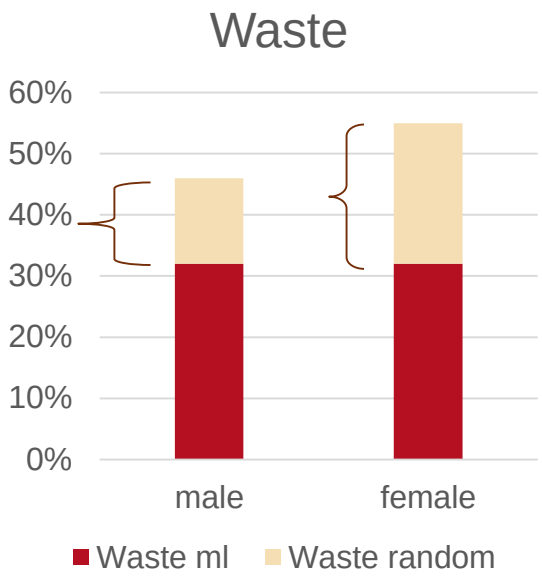
Hit-Rate increase

- Male: 25%
- Female: 14%

Waste Reduction

- Male: 14%
- Female: 23%



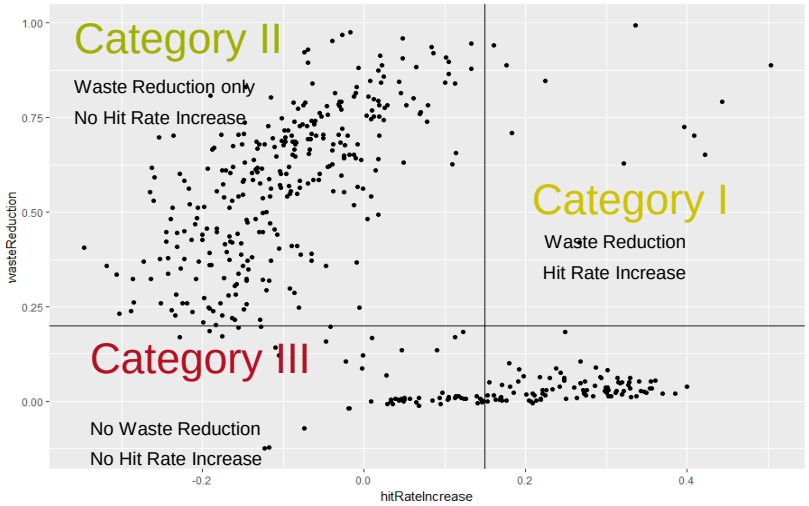


All numbers are illustrative only

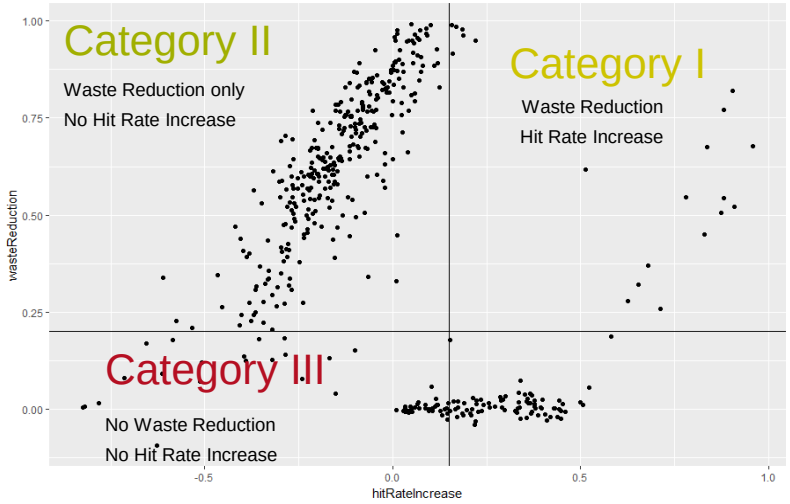
Overview

IJK
Institut für Journalistik und Kommunikationsforschung

Raw Data Input A



Raw Data Input B



Data and modelling quality is better with Input A.

Example 2: GfK Crossmedia Link / GfK Hyperlane

GXLL: A platform of single source cross-device digital measurement panels

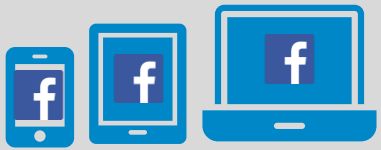
IJK
Institut für Journalistik und Kommunikationsforschung

360° View on Consumers

Claimed, sometimes enriched through fusion
(e.g. to industry-specific currencies such as FRS for the Financial Services Industry)

Device behavior and digital ad exposure

Passively measured



Additional digital ad exposure

Data link with partners, such as Facebook

TV viewing behavior and ad exposure

Passively measured or claimed; sometimes through fusion

Other media behavior and ad exposure

Claimed (print, radio)

Purchases

Claimed or measured by Consumer Panels sometimes enriched through fusion

GfK's core global data assets

IJK
Institut für Journalistik und Kommunikationsforschung

Survey based TV exposure

MS: MarketingScan

Single Source

Data type	UK	USA	DE	NL	IT	FR	PL	RU	TU	INDO	BR	SP	ME	ARG	CO
TV exposure and behaviour			GXL	GXL	Sinottica	MS	GXL	GXL	GXL	GXL	GXL	Netquest	Netquest	Netquest	Netquest
Desktop usage behaviour	GXL	GXL	GXL	GXL	Sinottica		GXL	GXL	GXL		GXL				
Mobile/tablet usage data	GXL	GXL	GXL	GXL	Sinottica		GXL	GXL	GXL	GXL	GXL	Netquest	Netquest	Netquest	Netquest
HH and individual demographic and attitudinal segments	GXL	MRI	GXL	GXL	Sinottica	MS	GXL	GXL	GXL	GXL	GXL	Netquest	Netquest	Netquest	Netquest
HH and individual purchase information			GXL	GXL	Sinottica	MS	GXL	GXL	GXL						
Purchasing power															
Point of sale/ retail data															

One size doesn't fit all. Journeys are different, allowing a large or narrow range of possibilities to influence on consumers in the purchase process

IJK
Institut für Journalistik und Kommunikationsforschung

Depending on the category, the Journey may be active for a longer or shorter time period

Highly Involved

Paul is looking for a new Coffee machine, and search internet to learn more about the different types that exist, look at the machines available in his usual hypermarket during his weekly shopping trip and asks friends & colleagues for tips….

Autopilot

Martha discovers she's out of beef flavoured bouillon cubes when preparing for dinner, and puts it on the shopping list for next shopping trip. She goes through the store leaflets to see if there's anything on promotion – one never knows – ….

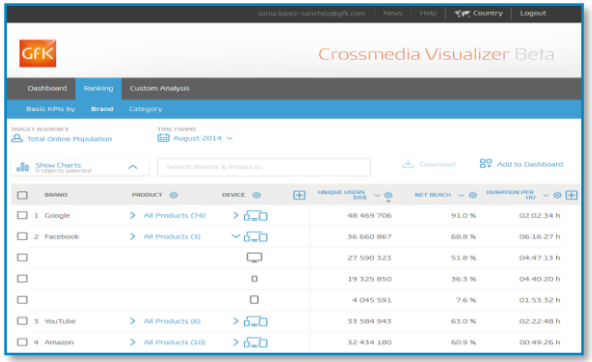
Impulse

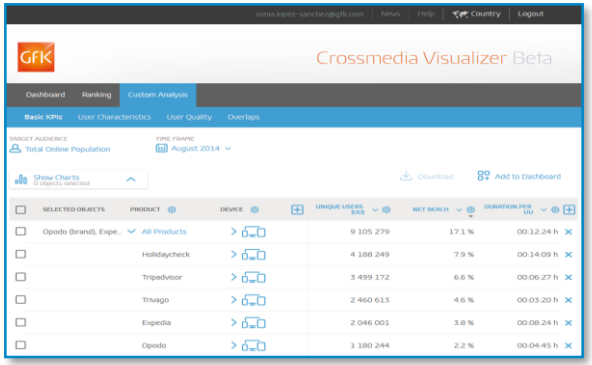
It's September, but really hot still. Jack decided to walk to work. "40' walk would do me well" he was thinking. However, he didn't realise it was that hot until he saw the ad for Pellegrino in front of the Relay kiosk. "I definitely need a bottle of water" he said to himself and went into the kiosk…

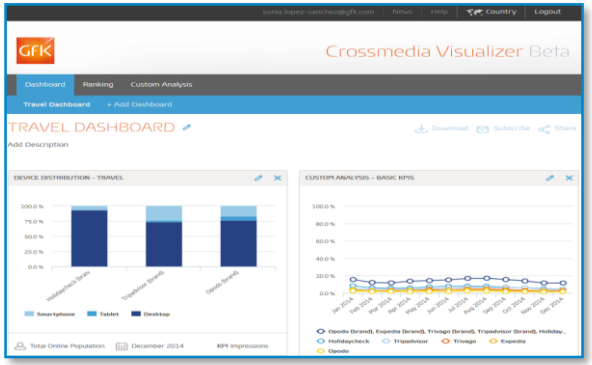
Central Dashboards available for crossdevice website and mobile app usage accross target groups and categories…

IJK
Institut für Journalistik und Kommunikationsforschung

Available Features* (I)







Top ranking: websites & apps
desktop + smartphone + tablet

- ranking of websites & apps in combination with target group filter
- various KPIs can be evaluated per devices or in total

websites & apps:
Category view & custom analysis

- helps to build up your own relevant universe
- enables flexible combinations & filter for brands and products as well as categories

Dashboard
starting point for analysis & navigation

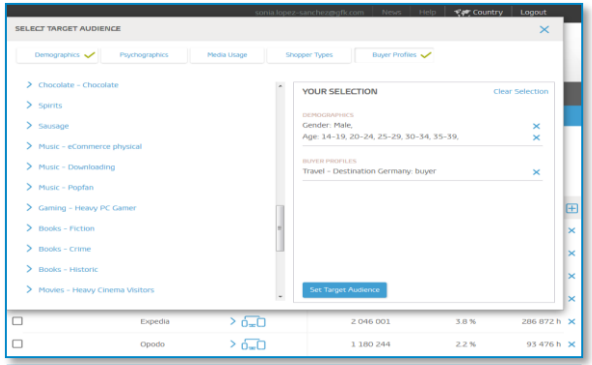
- central dashboard for snapshot views and navigation
- outlines most important KPIs at a glance
- data will be automatically updated

Additional features will be implemented gradually (until end of 2015), back data available from January 2014, several features on demand.

…which is highly appropriate to continuously track and understand target segments‘ online and mobile behavior

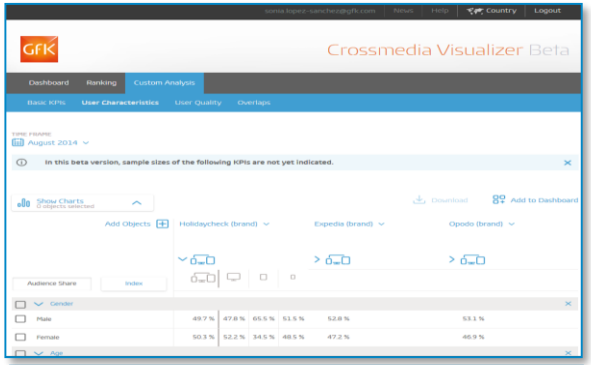
IJK
Institut für Journalistik und Kommunikationsforschung

Available Features* (II)



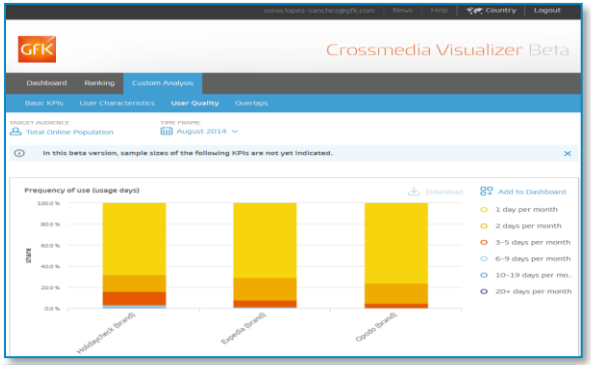
Target Group Filter

- pre defined buyer target groups & offline media target groups can be selected for all further analysis
- individual target groups on demand



Demographics: audience profiles

- who are the users of websites & apps
- analyze the data to focus your actions on your target audience



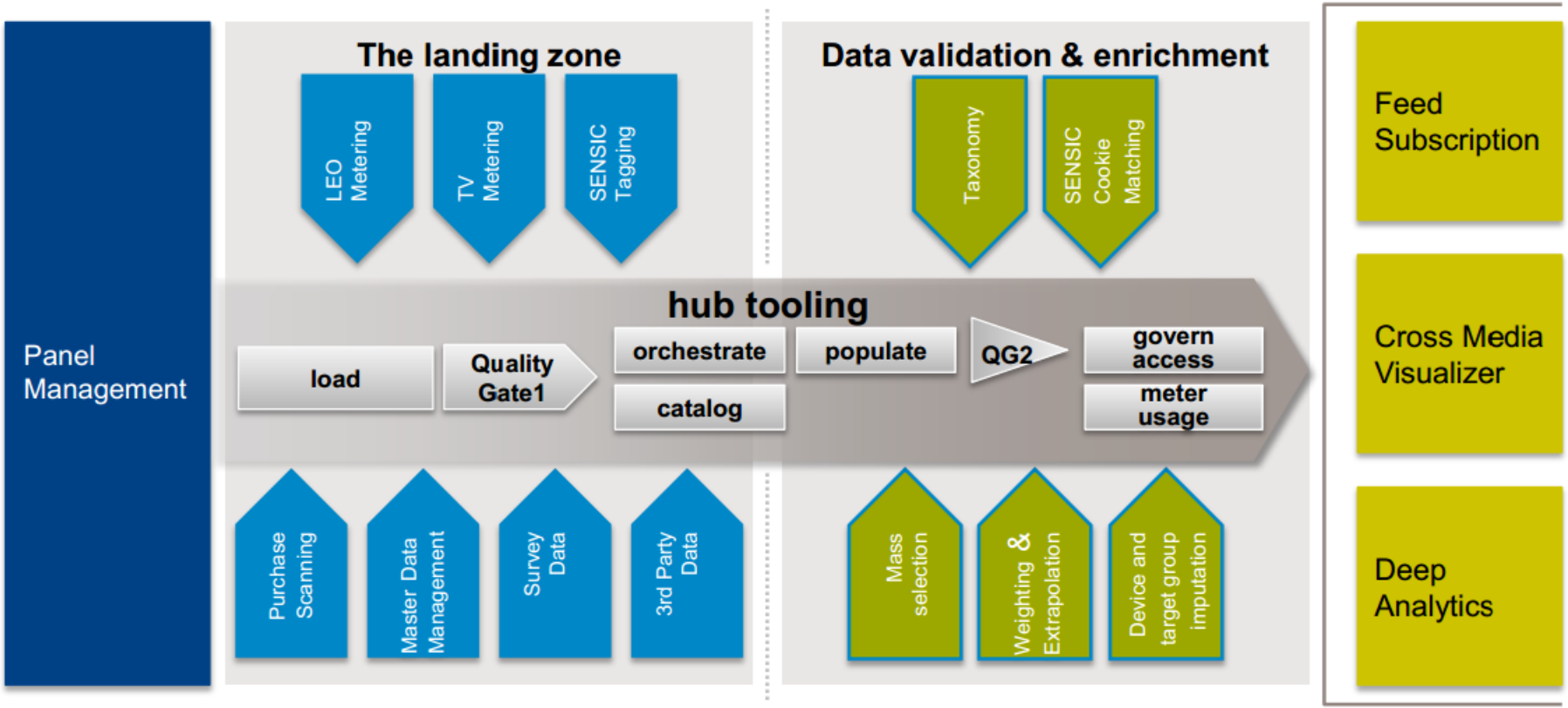
User Quality

- explore the quality of visitors and their intensity of use
- visit frequency
- reach w/wo bouncers
- lost / recurring / new users
- number of usage days in a month

* Additional features will be implemented gradually (until end of 2015), back data available from January 2014, several features on demand

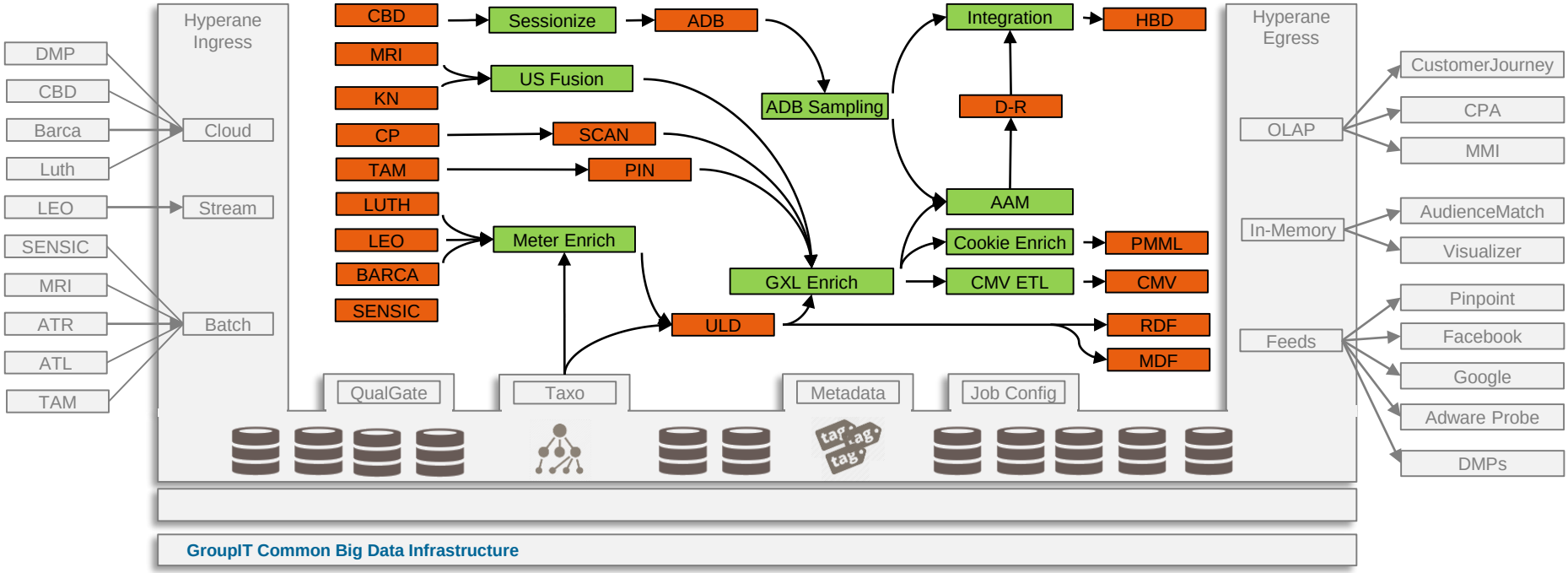
Hyperlane: GXL Business Application

IJK
Institut für Journalistik und Kommunikationsforschung



Hyperlane Apps: Data Assets and Job Flows

IJK
Institut für Journalistik und Kommunikationsforschung



Hyperlane: Broader Context

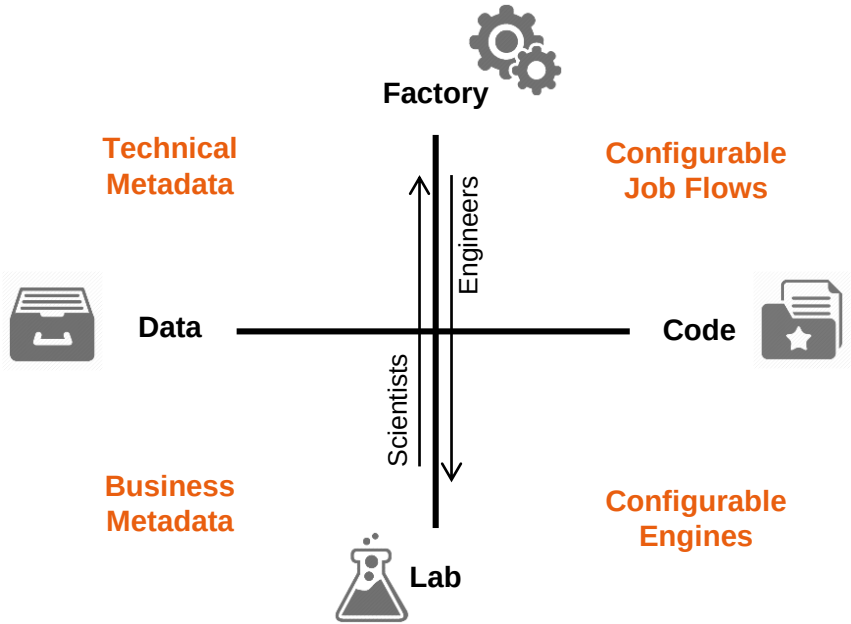
IJK
Institut für Journalistik und Kommunikationsforschung

Sculley et al (2014):

Undeclared Consumers:
Consuming the output of a given prediction model as an input to another component of the system. Changes will very likely impact these other parts.

Unstable Data Dependencies:
Some input signals are unstable, meaning that they qualitatively change behavior over time.

Static Analysis of Data:
On teams with many engineers, or if there are multiple interacting teams, not everyone knows the status of every single feature.



Sculley et al (2014):

Glue Code:
Using self-contained solutions often results in a glue code system design pattern, in which a massive amount of supporting code is written to get data into and out of general-purpose packages.

Pipeline Jungles:
The system for preparing data in an ML-friendly format may become a jungle of scrapes, joins, and sampling steps, often with intermediate files output.

Configuration Debt:
In a mature system which is being actively developed, the number of lines of configuration can far exceed the number of lines of the code.

http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/43146.pdf

Hyperlane: Functional Scope

IJK
Institut für Journalistik und Kommunikationsforschung

Offline Business Process
Find suitable data science methods to deal with peculiarities of the data

Data Science

Online Business Process
Continously load and prepare data so it integrates well with science method

Data Engineering

Job Flow Automation
Collaboration between Sciencists and Engineers to build scalable data apps that cover Online Business Processes. Configure, schedule and monitor for country setups based on templates.

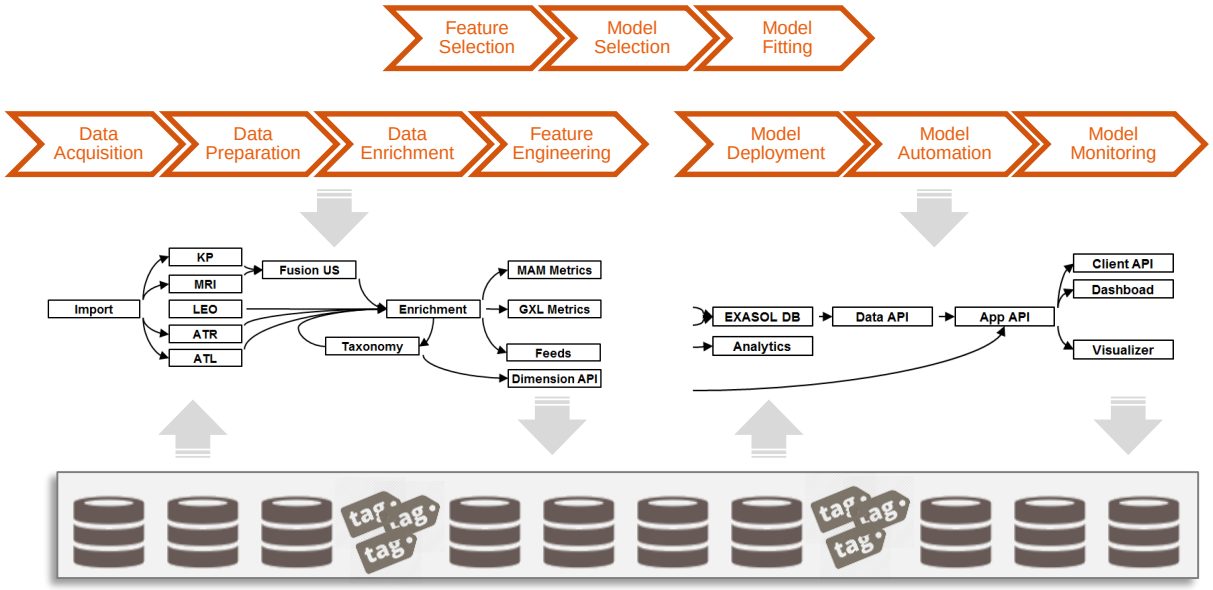
Job Flow Ops

Data Governance
Ingest cumulative and evolving Data Sets for which the quality needs to be monitored and documented. Business and technical meta data is tagged.

Data Ops

Data Enrichment
Check quality of ingested data sets and approve for automation. Curate Taxonomy and maintain links to data.

Coding Ops

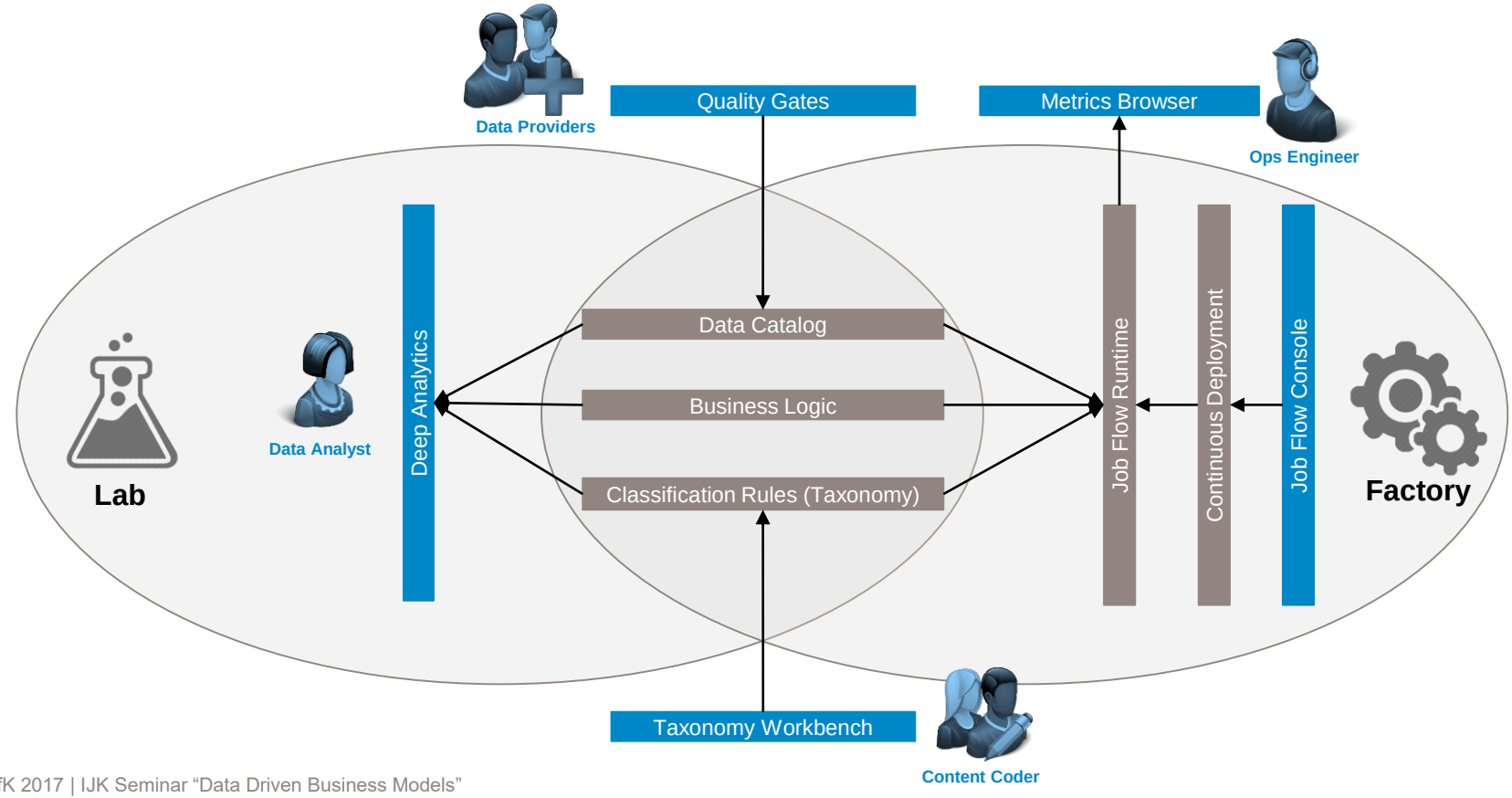


Quality Gates

Taxonomy

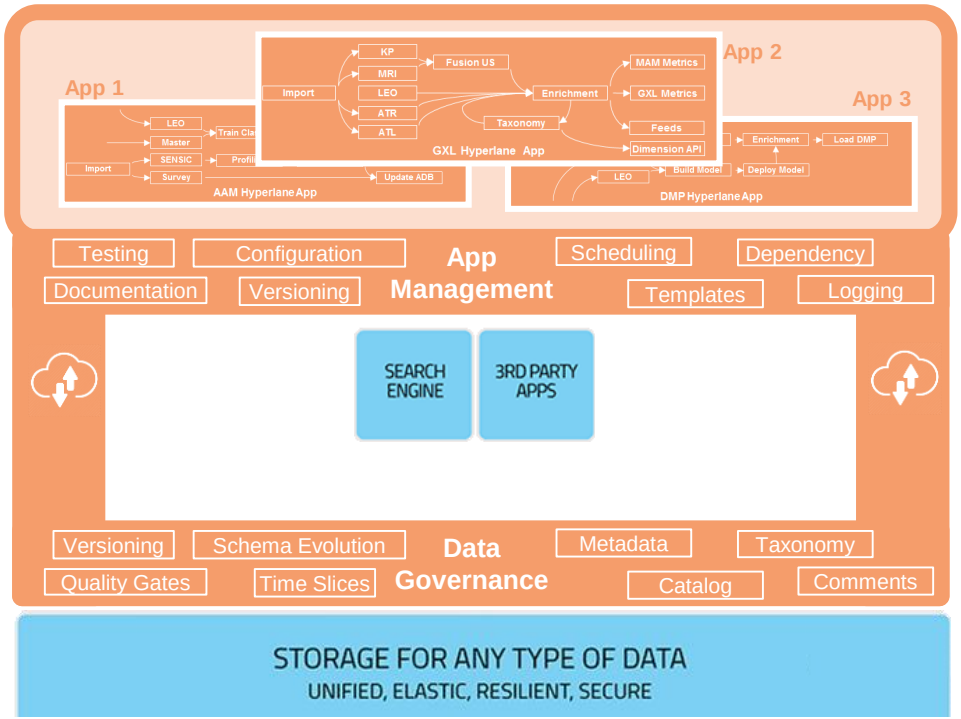
Hyperlane: Functional Scope

IJK
Institut für Journalistik und Kommunikationsforschung



Hyperlane: Platform vs. Apps

IJK
Institut für Journalistik und Kommunikationsforschung



Hyperlane hosts Data Apps

that combine Data Engineering and Data Science Models into automated Pipelines. Hyperlane Apps are configurable so they can be reused in multiple setups. Data Sets can be selectively fed into Apps based on their meta data. Tooling is available to manage configs, schedule execution and monitor progress.

Hyperlane manages Data Sets

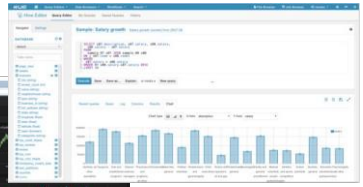
that can be annotated with technical and Business Meta Data to keep track of lineage and quality. The meta data powers tooling to search, navigate and query arbitrary data sets.

Hyperlane productionizes Data-Science

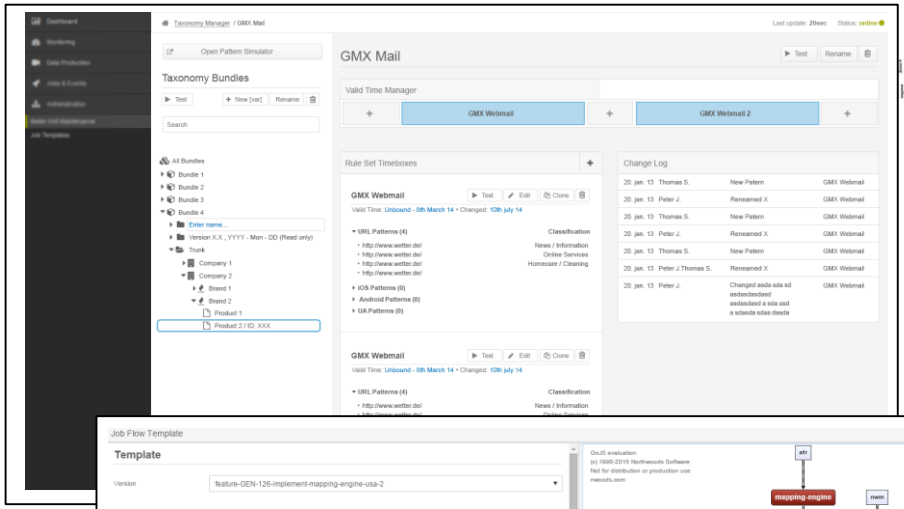
capabilities by extending GfK‘s Common Big Data Technology Stack (Cloudera CDH5) into a platform for GXL and other projects with smiliar goals. It features an open architecture based on standard technologies.

Hyperlane: Tooling

Monitoring

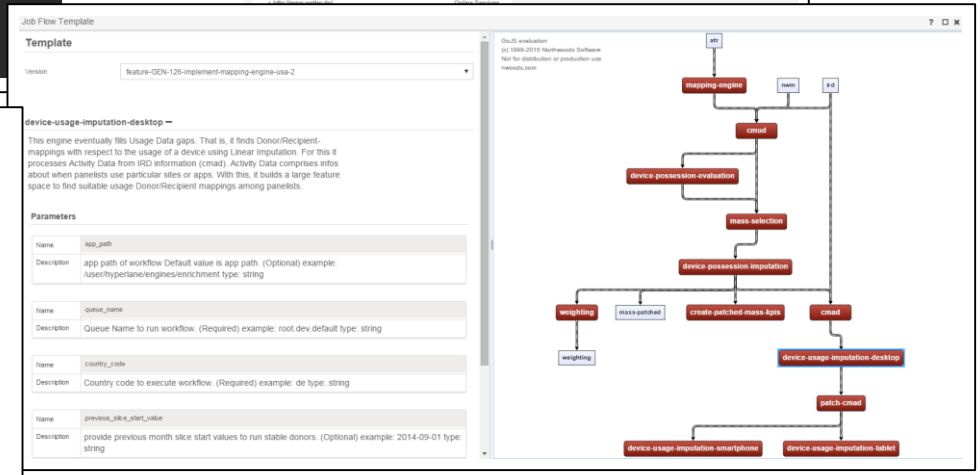


Analytics



Taxonomy Workbench

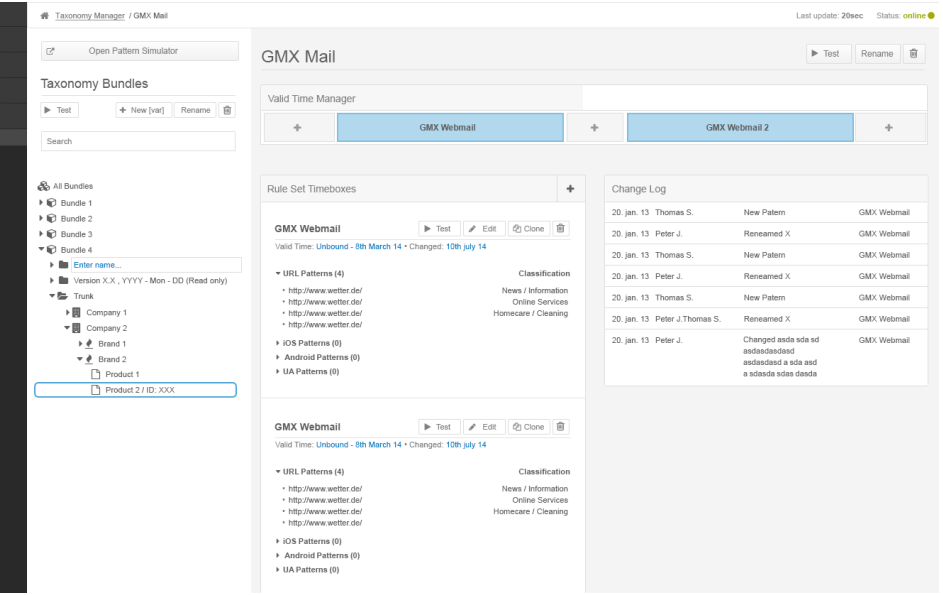
Data Set Browser



Job Flow Configuration Management

Hyperlane Tooling: Taxonomy Workbench

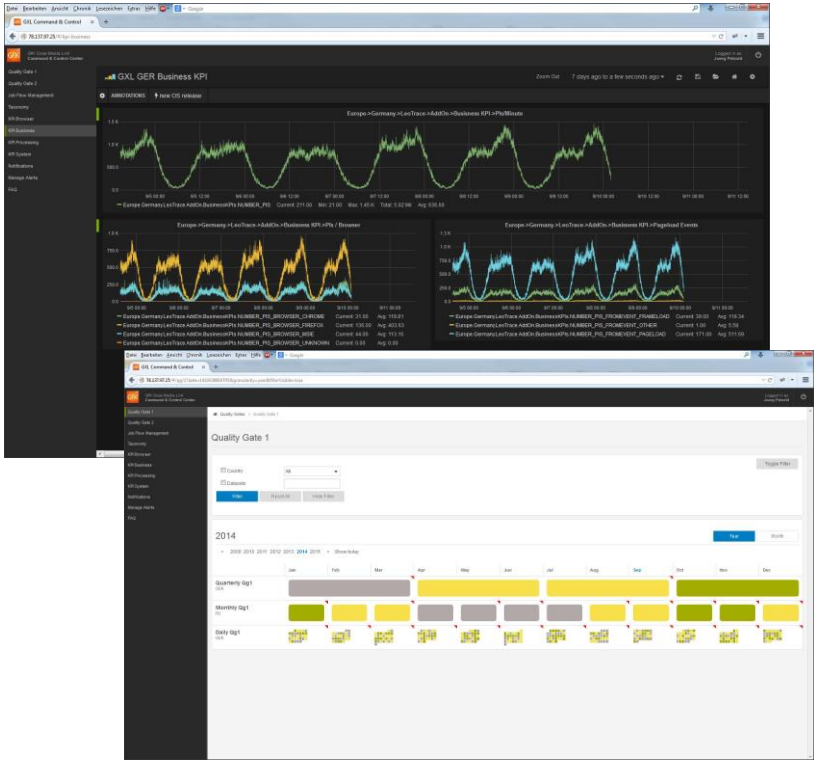
IJK
Institut für Journalistik und Kommunikationsforschung



A tool that supports our Hub-based Coding Team to efficiently manage the A2C taxonomy.

- Specify URL Patterns and Mobile App IDs that comprise Products& Brands
- Map A2C Categories to Products
- Manage existing Pattern Sets as well as emerging URLs and App IDs
- Support of Custom Patterns for specific clients (Patching the Default Taxonomy)

Hyperlane Tooling: Monitoring

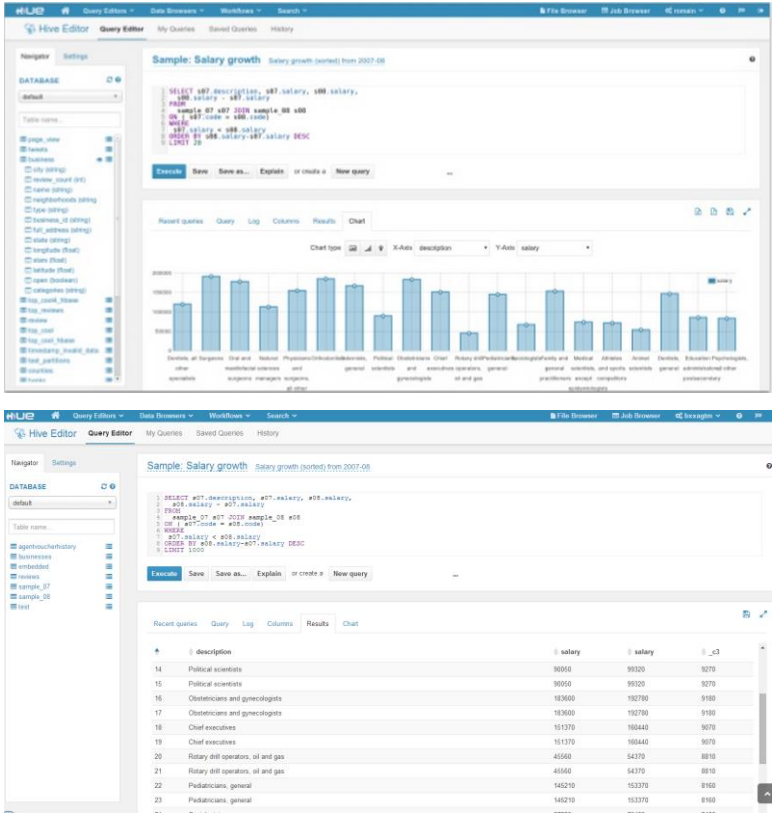


A centralized Monitoring System that provides transparency about data collection and data production processes.

- Visualization of events as time series
- Browser to manage hundreds of metrics
- Compare and Overlay different metrics
- Calculate Delta to time periods in the past
- Define Thresholds to receive alerts
- Setup customized Dashboards

Hyperlane Tooling: Deep Data Analytics

IJK
Institut für Journalistik und Kommunikationsforschung



A web-based analytics console that allows analysts to answer advanced questions by browse, query, aggregate and visualize data.

- Interactively browse existing data sets
- Run queries using well-known SQL
- Maintain a repository of Query Templates
- Merge data sets and apply custom weights
- Replace parts of the existing analytics tool chain and provide an environment for collaboration.

Wrap Up