Lucerne University of
Applied Sciences and Arts

HOCHSCHULE
LUZERN

Wirtschaft

Modern Data Engineering in the Cloud

Dr. Christian Dollfus
Dozent

T direkt +41 41 228 22 54
christian.dollfus@hslu.ch

Luzern

Dr. Pavlin Mavrodiev
Dozent

T direkt +41 76 733 61 66
pavlin.mavrodiev@alumni.ethz.ch

PART 1 : FOUNDATIONS IN DATA ENGINEERING

Introduction and Motivation

- Introductory words
- What is Data Engineering?
- Motivation and Value Proposition - a historic overview
- How does Data Engineering look like in many companies?
- What are features and advantages of workflow-based ETL?
- Motivation for the use of ETL tools.
- Data Engineering and Business Process Automation - similarities and differences

Learning Goals

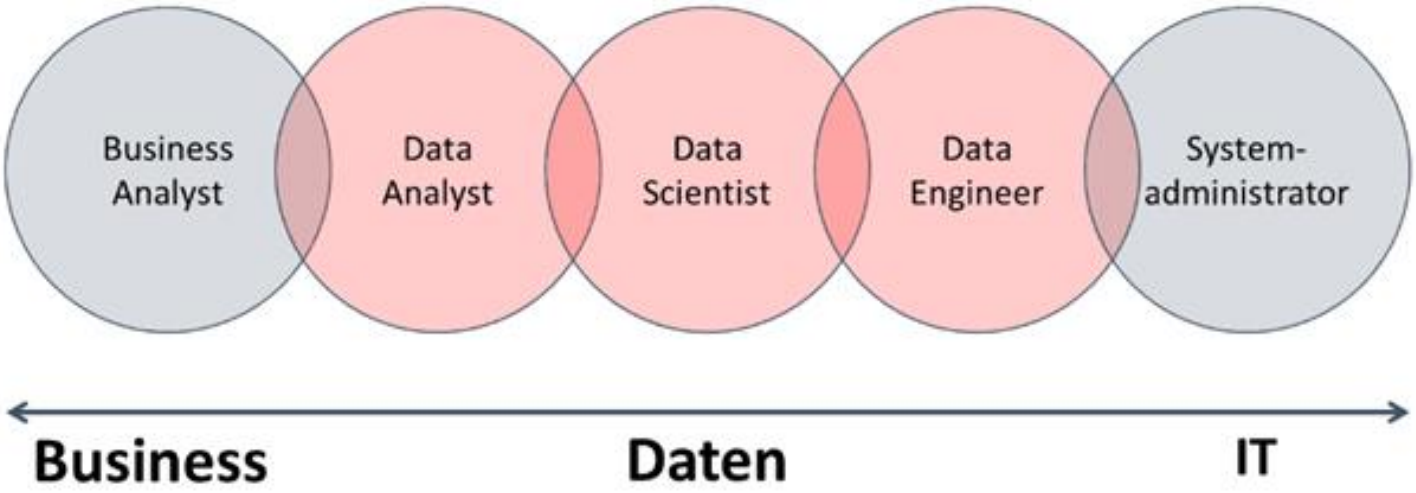
-After the 5 days you are able to **engineer automatic batch based and realtime Data-Pipelines** with PDI and KAFKA

-You have hands-on experience with 2 most used techniques in DE

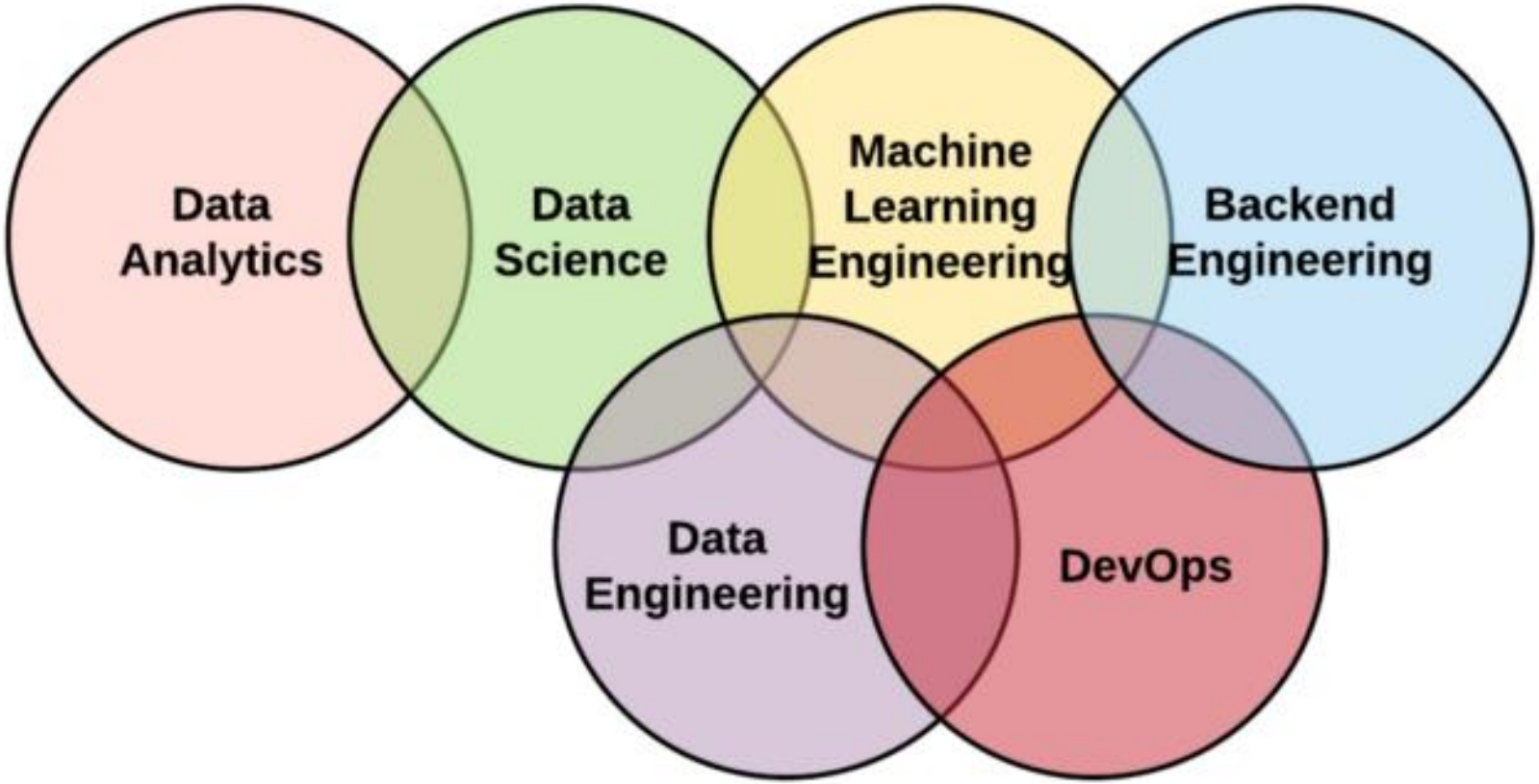
-You have an **overview and foundation** of the huge field **of Data Engineering** namly on the processual part (storage is also a topic of Data Lab 1 and 2)

-You have the **possibility to dive into the topic** if you use it. You got information there.

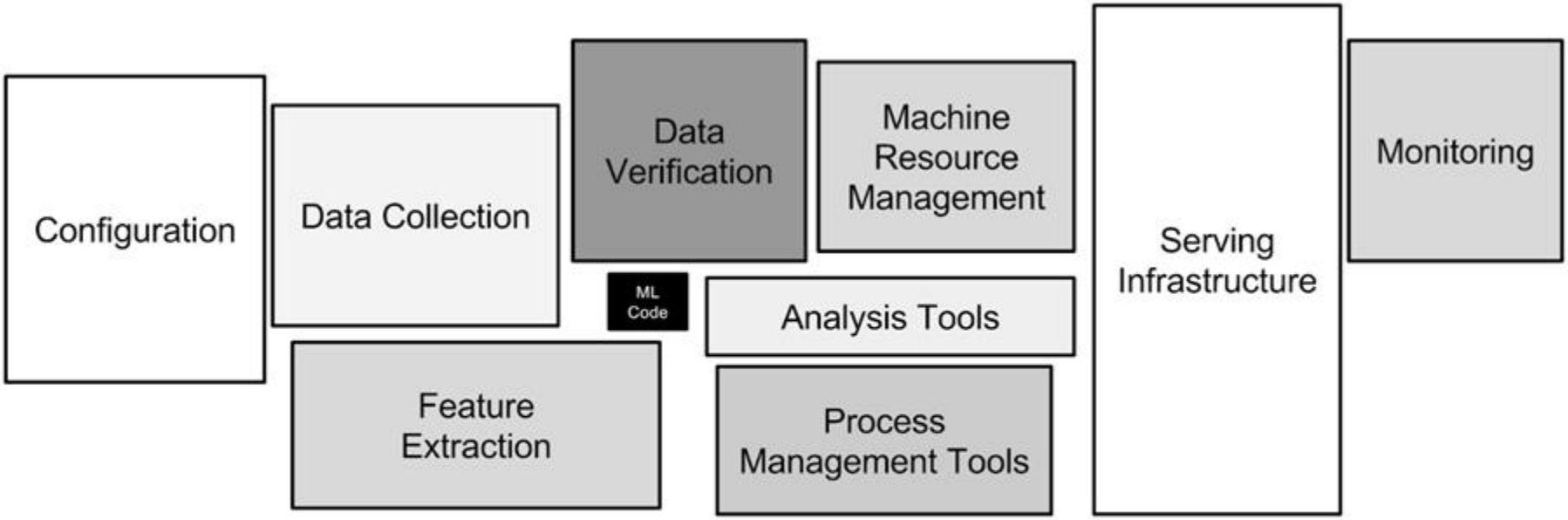
Introduction and Motivation: Roles



Introduction and Motivation: Areas

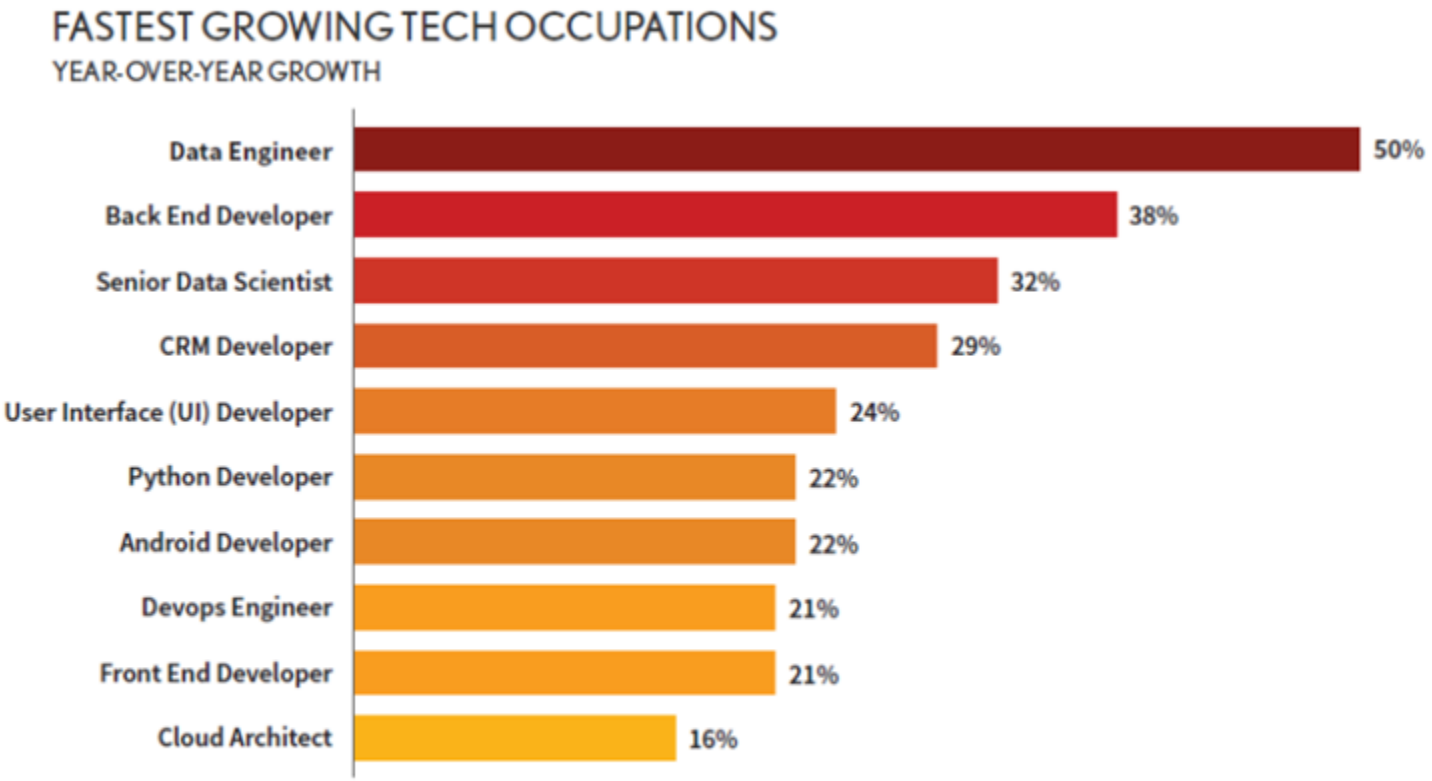


Source: https://towardsdatascience.com/data-science-is-boring-1d43473e353e



Source: Sculley, D., Holt, G., Golovin, D. et al. Hidden Technical Debt in Machine Learning Systems

Introduction and Motivation

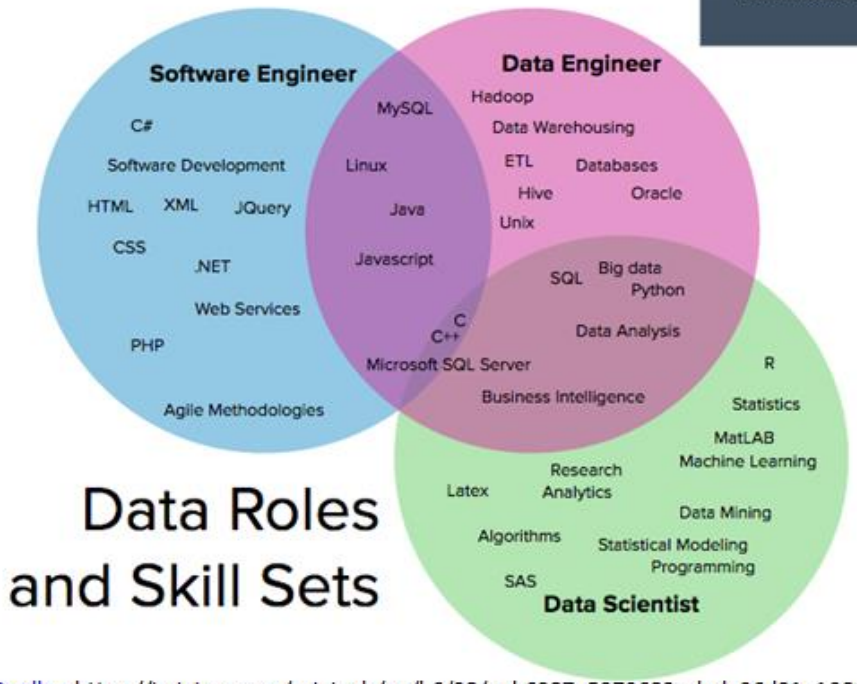


Source: The Dice 2020 Tech Job Report

Introduction and Motivation

Data Analyst	Data Scientist	Data Engineer
Schwerpunkt: BWL	**Schwerpunkt:** Mathematik	**Schwerpunkt:** Informatik
Wichtigste Skills: • Excel, ggf. SQL • BI-Anwendungen • Kommunikation	**Wichtigste Skills:** • Statistik • R oder Python • Datenbanken • Datenvisualisierung	**Wichtigste Skills:** • Data Warehouse • ETL (Extraction, Transformation & Load) • Datenbank-Design
Aufgaben: • Reporting • Geschäftsentwicklung	**Aufgaben:** • Insights • Predictive Analytics	**Aufgaben:** • Data Pipeline • Datenarchitektur

'jobs-karriere/data-scientist-oder-data-engineer-was-ist-der-unterschied



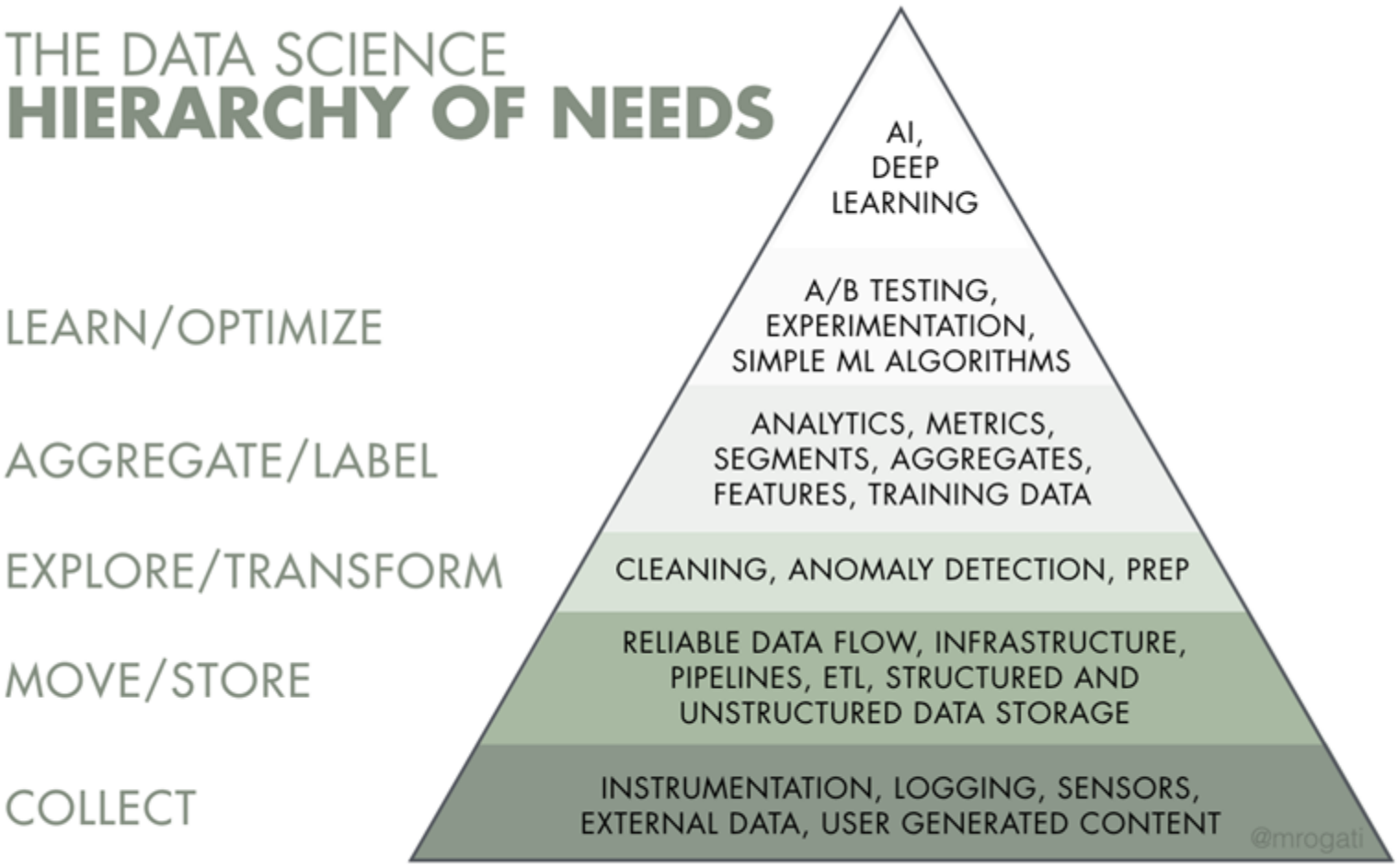
Quelle : https://i.pinimg.com/originals/aa/b6/98/aab6987a5979683edada36d01e13862d.png

If **data science** is the discipline of **making data *useful***, then you can think of **data engineering** as the discipline of **making data *usable***.

Data engineers are the heroes who provide behind-the-scenes infrastructure support for Data Scientists

Data scientists are the data-wranglers, while data engineers are the data-pipeline-wranglers

Quelle : https://www.linkedin.com/posts/hamed-zitoun-machine-learning-freelance_datascience-dataengineering-machinelearning-activity-6730899989536505857-MCDo/ and https://towardsdatascience.com/data-science-without-any-data-6c1ae9509d92



Quelle : https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

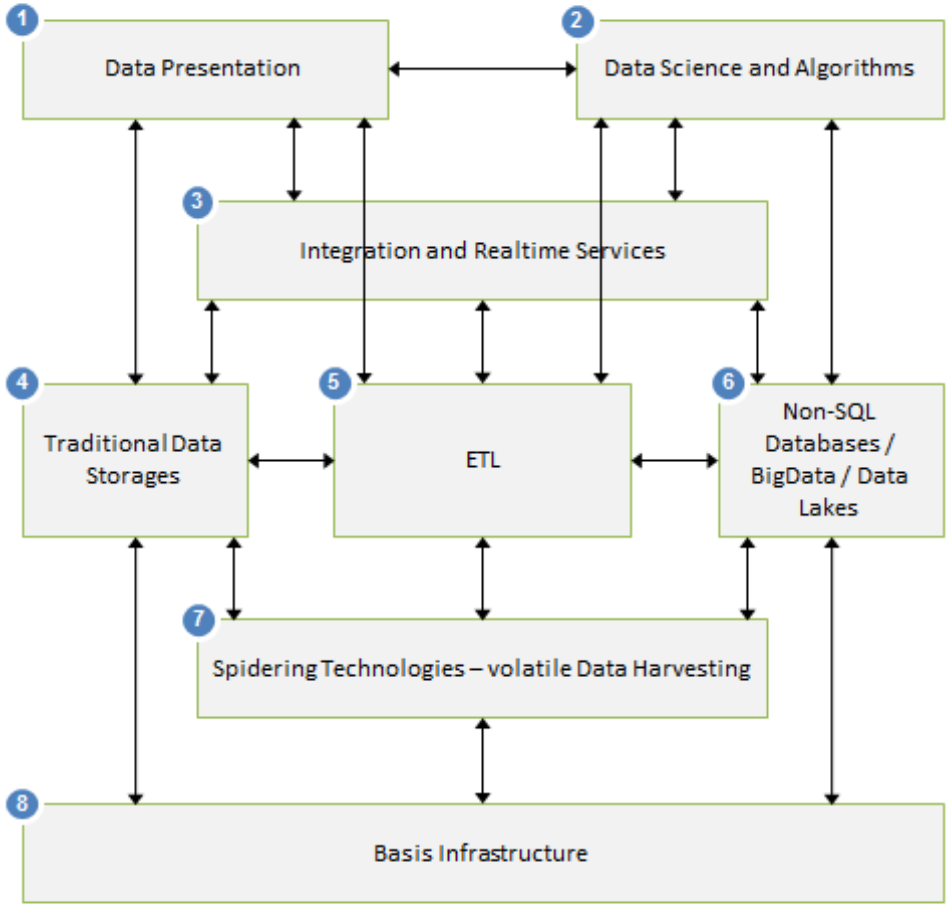
Introduction and Motivation

Volume + Velocity + Variety + Veracity = Value

Quelle : https://medium.com/dev-genius/what-you-need-to-know-before-you-become-a-data-engineer-career-advise-503b95e7a3cf

Introduction and Motivation

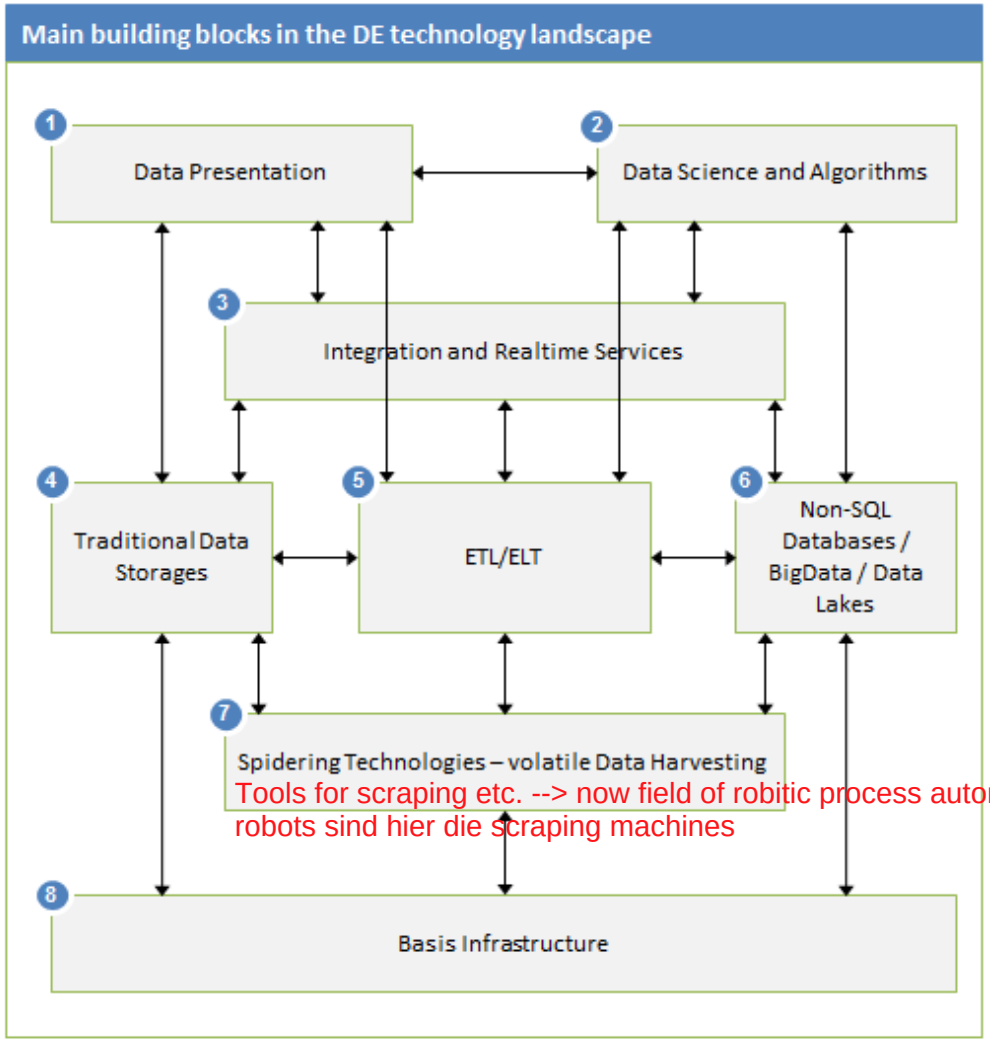
Main building blocks in the DI technology landscape

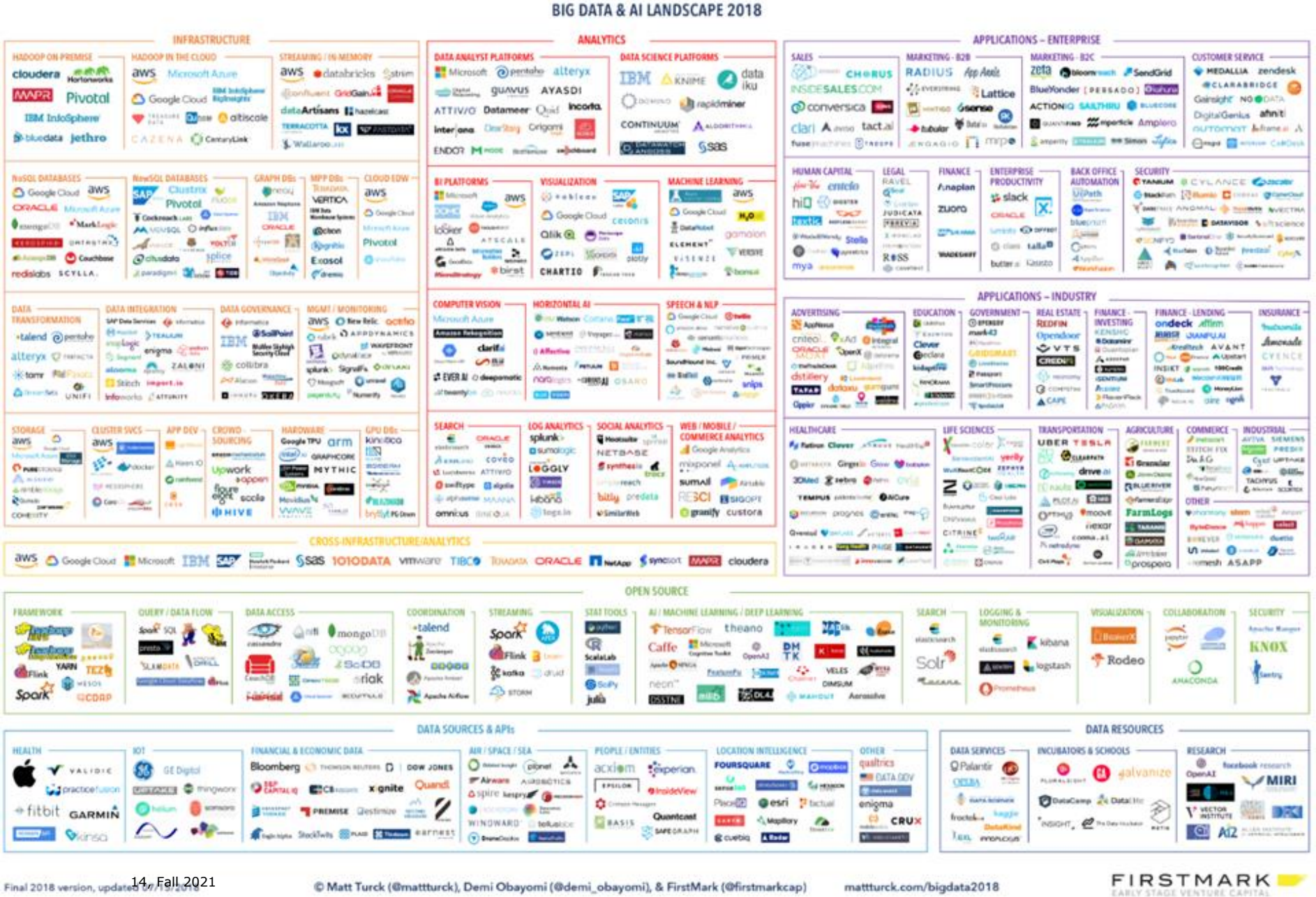


Building blocks

1. Structured and unstructured presentation of data such as Reporting, Cockpits, Dashboards and other data visualization.
2. Deep learning, artificial intelligence, neuronal networks, machine learning, algorithms (e.g. regression, clustering, decision trees) and bot's.
3. Entire underlying technology in order to allow for real-time enterpise data integration.
4. Traditional relational and multidimensional data storages such as SQL, DWH and BI technologies including in-memory approaches.
5. Data extraction, transformation and loading including robots for data provisioning automation approaches and bot's
6. Non-SQL data storages such as big data technologies, MapReduce, Hadoop including in-memory approaches.
7. Web Crowling techniques to get unstructured data from different sources such as ontology techniques, interfaces to web-services as well as related bot technologies.
8. Entire underlying infrastructure such as hardware, network, OS-layer, orchestration as well as new technologies such as blockchain.

Introduction and Motivation





Introduction and Motivation: How does Data Engineering look like in many companies ?

Source: https://blog.userfeed.io/informed-userfeed-product-feedback

Introduction and Motivation: How does Data Engineering look like in many companies ?

Use Excel as Data Scientist only as a presentation-tool

		MONATE												letzte 4 Wochen			
Produkt	Jahr	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Okt	Nov	Dec	9	10	11	12
Besuche aller Webseiten	2014	702'041	558'186	576'874	592	0	0	0	0	0	0	0	0	119'092	128'916	150'799	12
	2013	705'818	537'505	574'537	527'663	705'955	835'130	833'372	799'296	681'926	529'972	371'207	393'298	130'256	130'387	133'314	12
	Verhaltnis	99.46%	103.85%	100.41%	0.11%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	91.43%	98.87%	113.12%	96
Umsatz Pauschalreisen	2014	3'706'499	3'549'532	3'891'273	0	0	0	0	0	0	0	0	0	797'622	814'920	882'998	92
	2013	2'838'527	2'523'908	3'239'348	3'142'750	3'178'890	4'662'563	5'571'988	4'839'891	4'259'929	2'595'905	1'280'367	1'592'120	800'863	752'441	711'269	68
	Verhaltnis	130.58%	140.64%	120.13%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	99.60%	108.30%	124.14%	134
Buchungen Pauschalreisen	2014	1'317	1'286	1'469	0	0	0	0	0	0	0	0	0	312	303	329	
	2013	1'014	980	1'269	1'313	1'364	1'803	2'164	2'068	1'722	1'023	483	507	299	300	261	
	Verhaltnis	129.88%	131.22%	115.76%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	104.35%	101.00%	126.05%	122
Umsatz Linienfluge	2014	231'819	151'634	200'121	0	0	0	0	0	0	0	0	0	43'551	40'898	50'914	5
	2013	254'603	131'606	162'818	210'350	243'237	183'768	214'069	199'159	215'188	188'514	137'561	113'704	47'333	39'961	31'733	2
	Verhaltnis	91.05%	115.22%	122.91%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	92.01%	102.34%	160.44%	171
Buchungen Linienfluge	2014	160	124	166	0	0	0	0	0	0	0	0	0	38	41	33	
	2013	175	104	131	163	201	155	200	151	172	156	96	85	37	28	30	
	Verhaltnis	91.43%	119.23%	126.72%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	102.70%	146.43%	110.00%	148
Umsatz	2014																

Result Report Switzerland

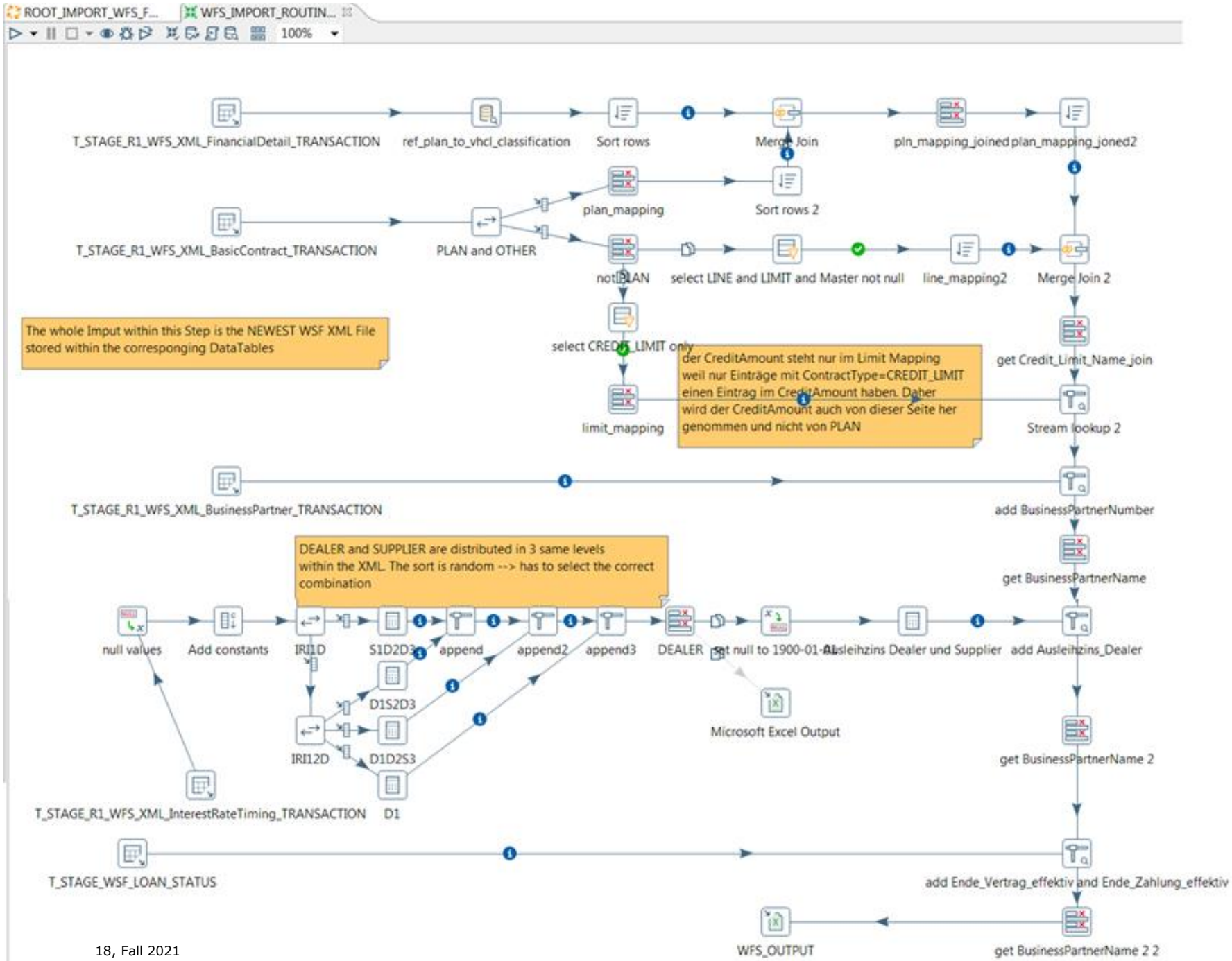
	Rolling Forecast			Test	Customers							
					Active			New			Pot. Prosp.	
	Curr. Si.	Next Si.	Quotes	Re	Bu	Re	%r.	Bu	Re	%r.	Pfl	
	69'002	249'538	207'500	8	17	30	176.47	9	1	11.11		
	100'558	179'475	535'400	38	29	27	93.1	12	7	58.33		
	153'775	170'492	315'000	41	49	45	91.84	12	11	91.67		
	104'890	187'684	213'750	33	29	28	96.55	9	12	133.33		
	94'287	36'775	393'500	20	35	18	51.43	8	6	75		
	236'220	311'782	1'071'745	79	68	69	101.47	34	28	82.35		
TOTAL MUCH		5'424'000.00	3'780'585.98	69.70	758'731.64	1'135'746.75	2'736'894.97	219.00	227.00	217.00	95.59	84.00	65.00	77.38	0.0

Total reported consultants
Total consultants' factors for av.
% budget on period 92.3076923

Introduction and Motivation: Motivation for the use of ETL tools. Data Engineering and Business Process Automation - similarities/differences

I choose a lazy person to do a hard job. Because a lazy person will find an easy way to do it.
-Bill Gates

- Velocity and agility
- Cost of development → predefined functions
- «ease of use» without programming knowledge
- Independen from programming langugage → more people can participate in the process
- Overview of process → Directed acyclic graph
- Maintenance cost
- Documentation
- 1000 rows JAVA Code in 1 workflow (Ex. WFS_IMPORT_ROUTINE_FOR_DWH_STORE_PREPARATION


ROOT_IMPORT_WFS_F...
WFS_IMPORT_ROUTIN...
100%
T_STAGE_R1_WFS_XML_FinancialDetail_TRANSACTION
ref_plan_to_vhcl_classification
Sort rows
Merge Join
pln_mapping_joined
plan_mapping_joned2
plan_mapping
Sort rows 2
T_STAGE_R1_WFS_XML_BasicContract_TRANSACTION
PLAN and OTHER
notPLAN
select LINE and LIMIT and Master not null
line_mapping2
Merge Join 2
The whole Imput within this Step is the NEWEST WSF XML File stored within the corresponging DataTables
select CREDIT_LIMIT only
der CreditAmount steht nur im Limit Mapping weil nur Einträge mit ContractType=CREDIT_LIMIT einen Eintrag im CreditAmount haben. Daher wird der CreditAmount auch von dieser Seite her genommen und nicht von PLAN
get Credit_Limit_Name_join
limit_mapping
Stream lookup 2
T_STAGE_R1_WFS_XML_BusinessPartner_TRANSACTION
add BusinessPartnerNumber
get BusinessPartnerName
DEALER and SUPPLIER are distributed in 3 same levels within the XML. The sort is random --> has to select the correct combination
null values
Add constants
IRI1D
S1D2D3
append
append2
append3
DEALER
set null to 1900-01-01
Ausleihzins Dealer und Supplier
add Ausleihzins_Dealer
D1S2D3
Microsoft Excel Output
IRI12D
D1D2S3
get BusinessPartnerName 2
T_STAGE_R1_WFS_XML_InterestRateTiming_TRANSACTION
D1
T_STAGE_WSF_LOAN_STATUS
add Ende_Vertrag_effektiv and Ende_Zahlung_effektiv
WFS_OUTPUT
get BusinessPartnerName 2 2

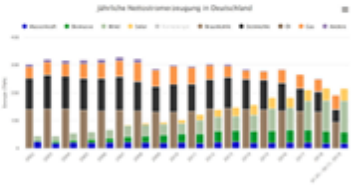
Introduction and Motivation

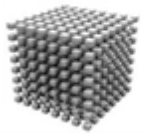
Elements of a Data Pipeline:

An ETL/ELT-Tool deals with all Tasks

Very heterogenous

Can be the same technology provider/vendor





Data Sourcing
Data Wrangling
Data Preparation

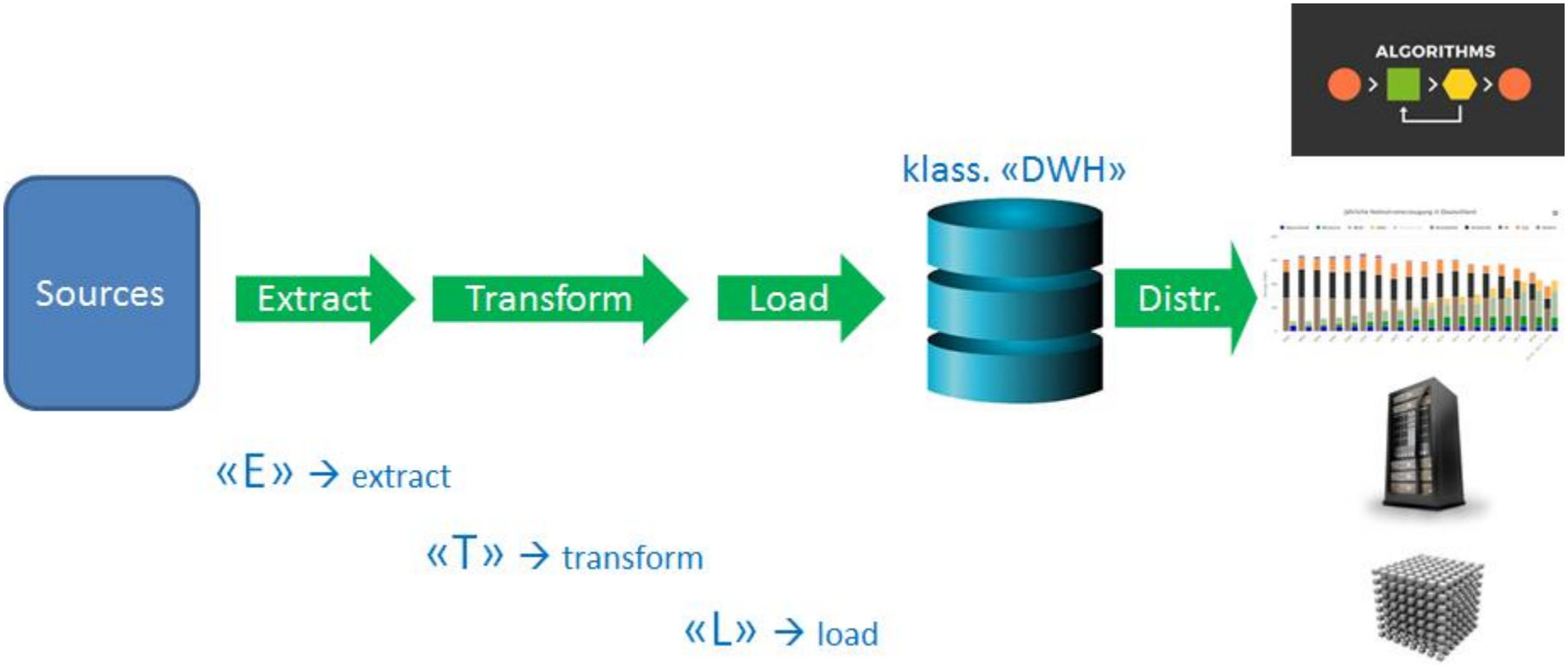
«E» → extract

«L» → load

«T» → transform

ELT is better than ETL if you have unreliable data --> everything is extracted and loaded and then the process is under control of you and you can check if there for example is a data type problem or so

Introduction and Motivation



Introduction and Motivation

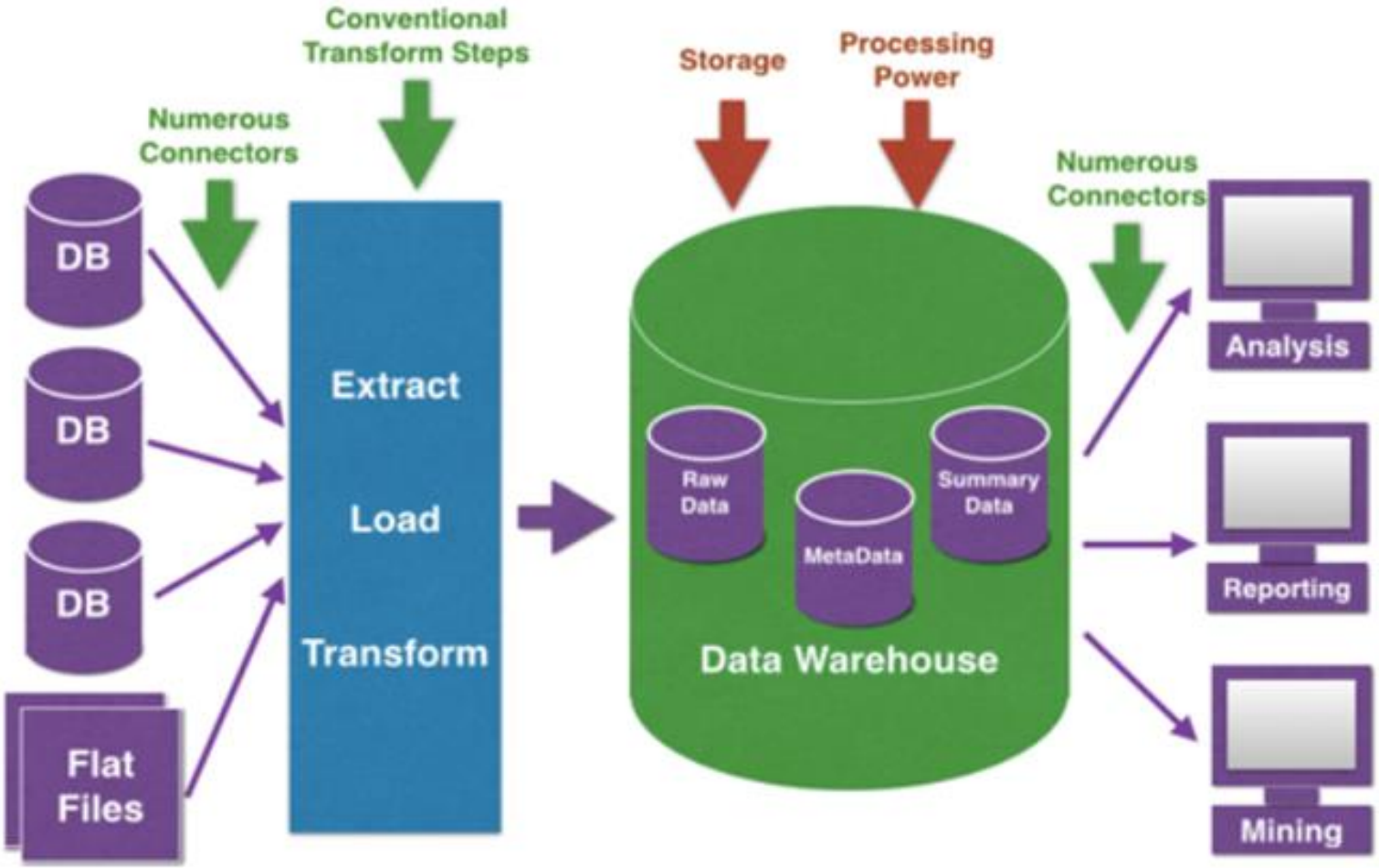
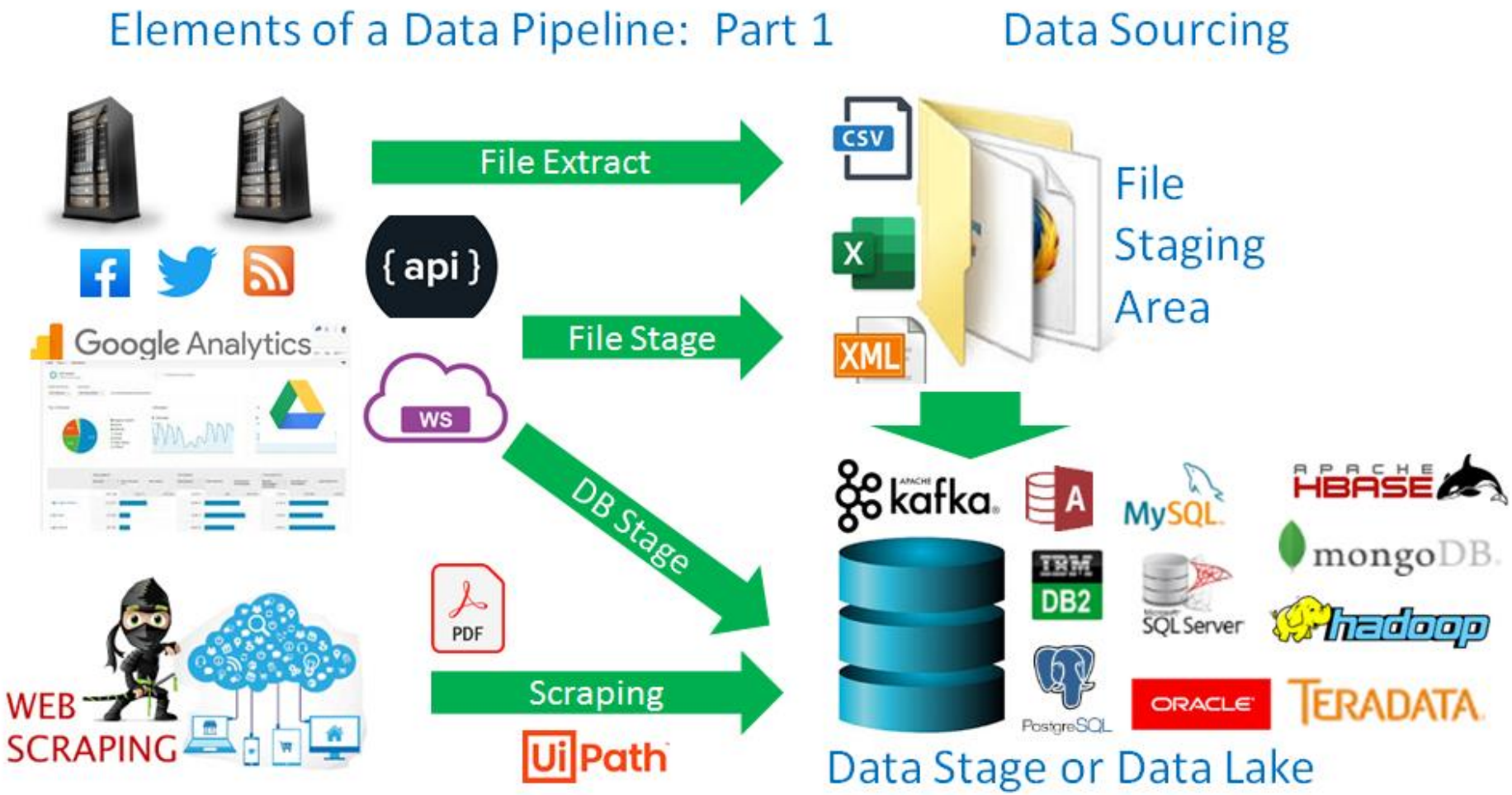


Figure 1 - Traditional Data Integration

Introduction and Motivation



Introduction and Motivation

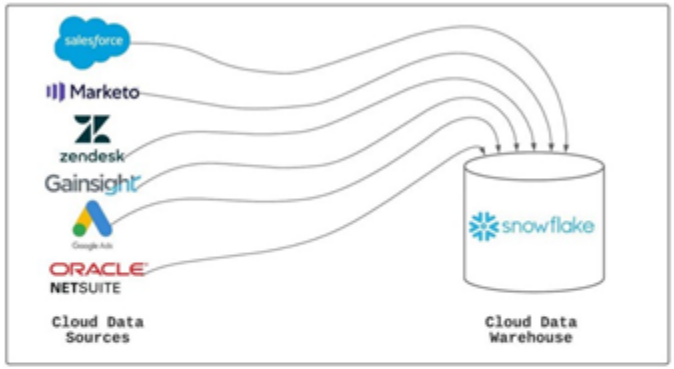
Don't underestimate the managerial and psychological tasks in the Data Sourcing part:

- Information about the elements of KPI's
- Which System holds which data ?
- Who can tell me details in case of missing Data Dictionary ?
- Whow can i access them ?
- Do i really get what i effectively want ? -> testing !

→ don't underestimate the amout of time you use here !

Introduction and Motivation

ETL



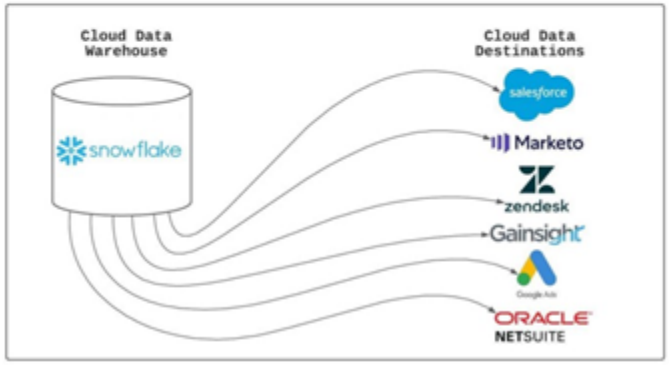
There are 3 primary use cases for Reverse ETL:

1. **Operational analytics** — feeding insights from analytics to business teams in their usual workflow so they can make more data-informed decisions
2. **Data automation —** not all data problems are so glamorous. "Can I get a CSV to issue some invoices?", your finance team asks. Reverse ETL poses a simple solution.
3. **Data infrastructure —** with a growing number of data sources, reverse ETL is emerging as a general-purpose pattern in software engineering.

Definition:

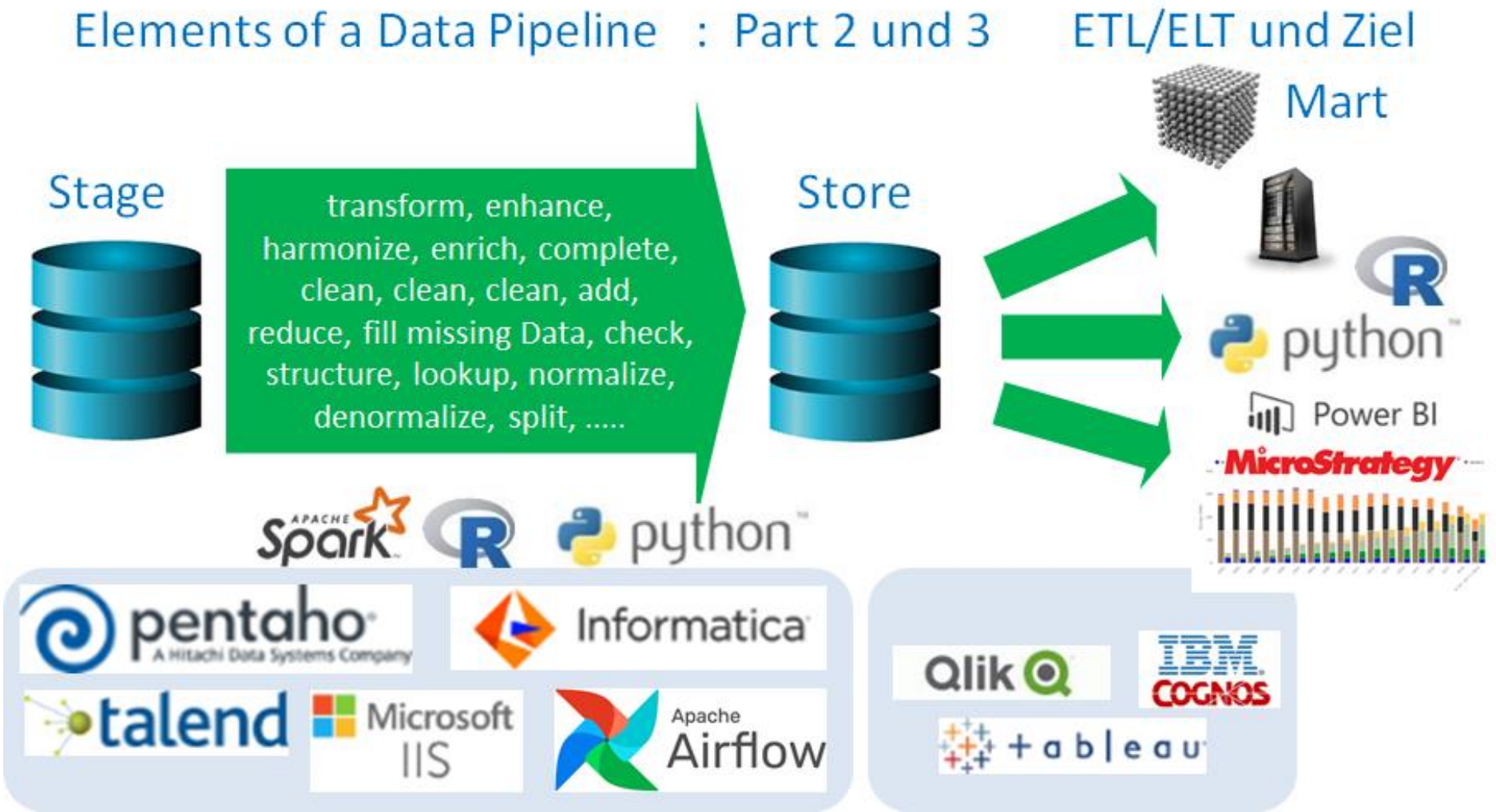
process of copying data from the data warehouse into systems of record across a company.

Reverse ETL



Source: https://tejasmanohar.medium.com/what-is-reverse-etl-6f228a14f6ec

Introduction and Motivation



Introduction and Motivation

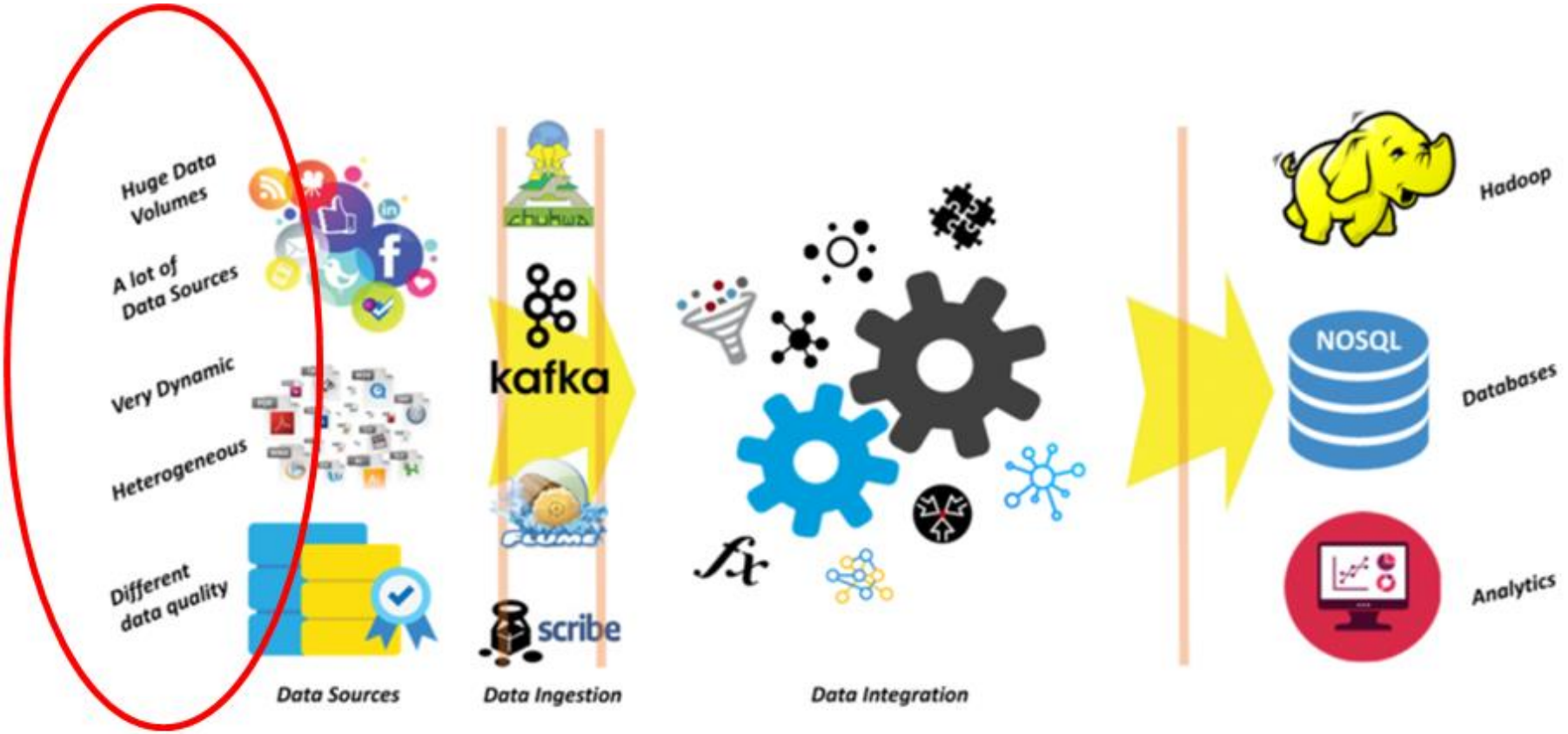
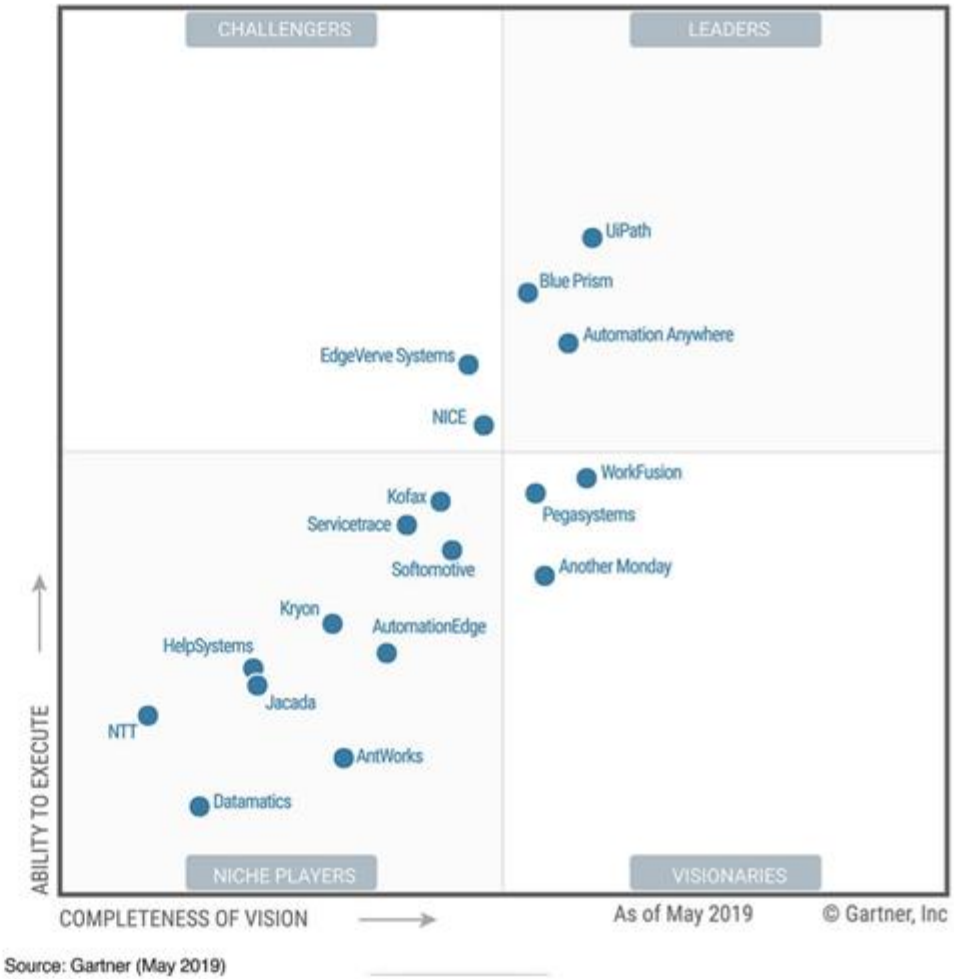


Figure 2 - Big Data Integration

Introduction and Motivation

Robotic Process Automation:

--> enhance ETL tool functionality



Source: Gartner (May 2019)

Introduction and Motivation

DEMO: VIDEO

Example of Web Scraping

https://www.youtube.com/watch?v=m_PkWeHSOrE