Projekte. Beratung. Spezialisten.

IKS
Individuelle Softwarelösungen

Darf es ein bisschen mehr sein?

Konzepte und Strategien zur Bewältigung großer und wachsender Datenmengen

IKS-Thementag

Autor: Christoph Schmidt-Casdorff

14.05.2019

Agenda

- Was ist Big Data?
- Architekturaspekte in Big Data
- Big Data at Rest
- Data Processing – Batch a là Big Data
- Stream Processing – Big Data at Flow
- Big-Data-Architekturen
- *Hadoop Eco System* und Plattform-Anbieter
- Abschluss

Große Datenmenge

IKS
Individuelle Softwarelösungen

Frage	Antwort

Historisch: Warum Big Data?	– Große Datenmengen und massives Datenwachstum
	  – Facebook sammelt mehr als 250 Terabytes pro Tag
	  – Daten kommen aus unterschiedlichsten Quellen/Kanälen
	  – Daten haben unterschiedlichste (oder gar keine) Strukturen

Was charakterisiert Big Data ?	– Hohe Geschwindigkeit des Datenaufkommens – Hohe Schwankung des Datenaufkommens – Große Menge der Daten – Keine Beschränkung der Variabilität der Datenstrukturen – Keine Beschränkung in der Varianz der Datenqualität
Beispiele von Datenquellen	– Sensor Daten (IoT) – Kreditkarten-Transaktionen – Aktienbewegungen – Blog Posts – Network Traffic – Log Entries
Welche technischen Herausforderungen sind zu bewältigen?	– Speicherung der Datenmengen – Lesen und Analyse der Datenmengen – Befüllung der Datenmengen

Daten in Ruhe

https://www.fjordblick.com/tjodnane-see-am-preikestolen-foto-des-monats-dezember-2018/

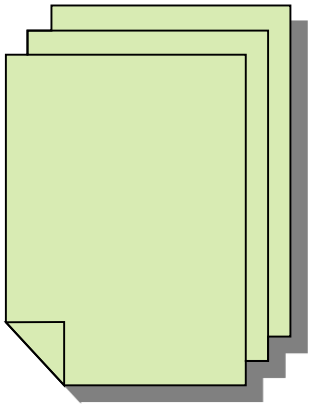
Daten in Ruhe

CD

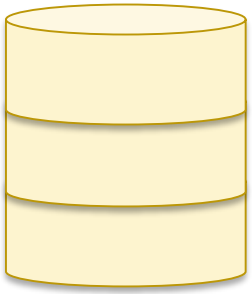
Diskette

Magnetband

Festplatten



Datei



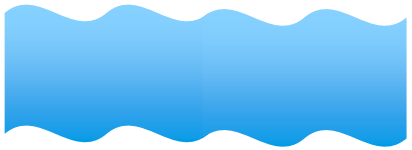
Datenbank

Informationen und Erkenntnisse gewinnt man nur aus Daten im Fluss

Verarbeitung muss mit Daten versorgt werden

https://commons.wikimedia.org/wiki/File:Wechsel_-_M%C3%BChlrad_beim_M%C3%BChlenplatzl_am_Wildwasserweg.jpg

Notation: Daten im Fluss



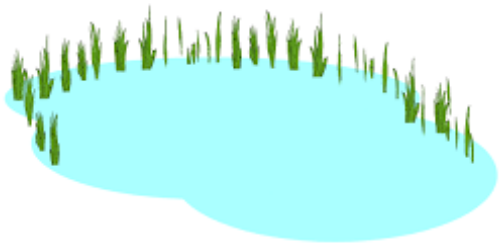
Datenfluss

übergibt Ergebnisse

stellt Daten zu

Datenverarbeitung/
Informationsgewinn

schreibt und liest



Datenspeicher

Big Data ….

https://www.worldatlas.com/articles/the-great-lakes-ranked-by-size.html

https://afrika-junior.de/inhalt/kontinent/regionen/das-suedliche-afrika-die-suempfe-und-die-namib-wueste/der-sambesi-die-lebensader-im-suedlichen-afrika.html

Funktioniert mit hergebrachten Methoden nicht mehr

https://www.n-tv.de/panorama/Land-unter-in-Bayern-und-Sachsen-article10751731.html

Warum reicht es nicht mehr?

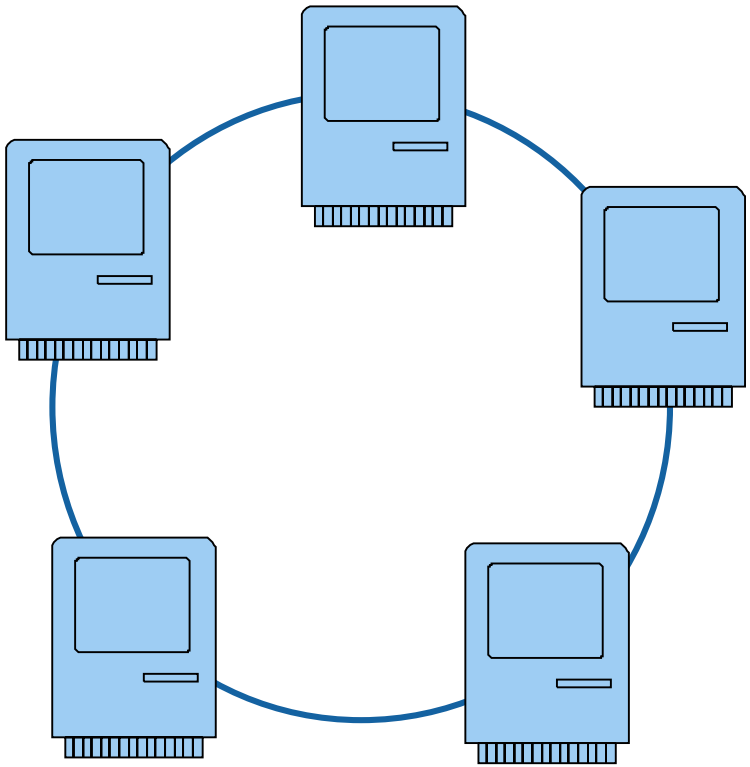
- Physische Kapazitätsbegrenzung der Hardware
  - Plattenplatz
  - Hauptspeicher
- Kapazitätsbegrenzung von Systemen wie relationalen Datenbanken

Horizontaler statt vertikaler Skalierung

Aufteilung des großen Datensees in beliebig viele Datenteiche

https://urlaubsreich.de/wp-content/uploads/2018/08/Sch%C3%B6n-IMG_5643-ge%C3%A4ndert.jpg

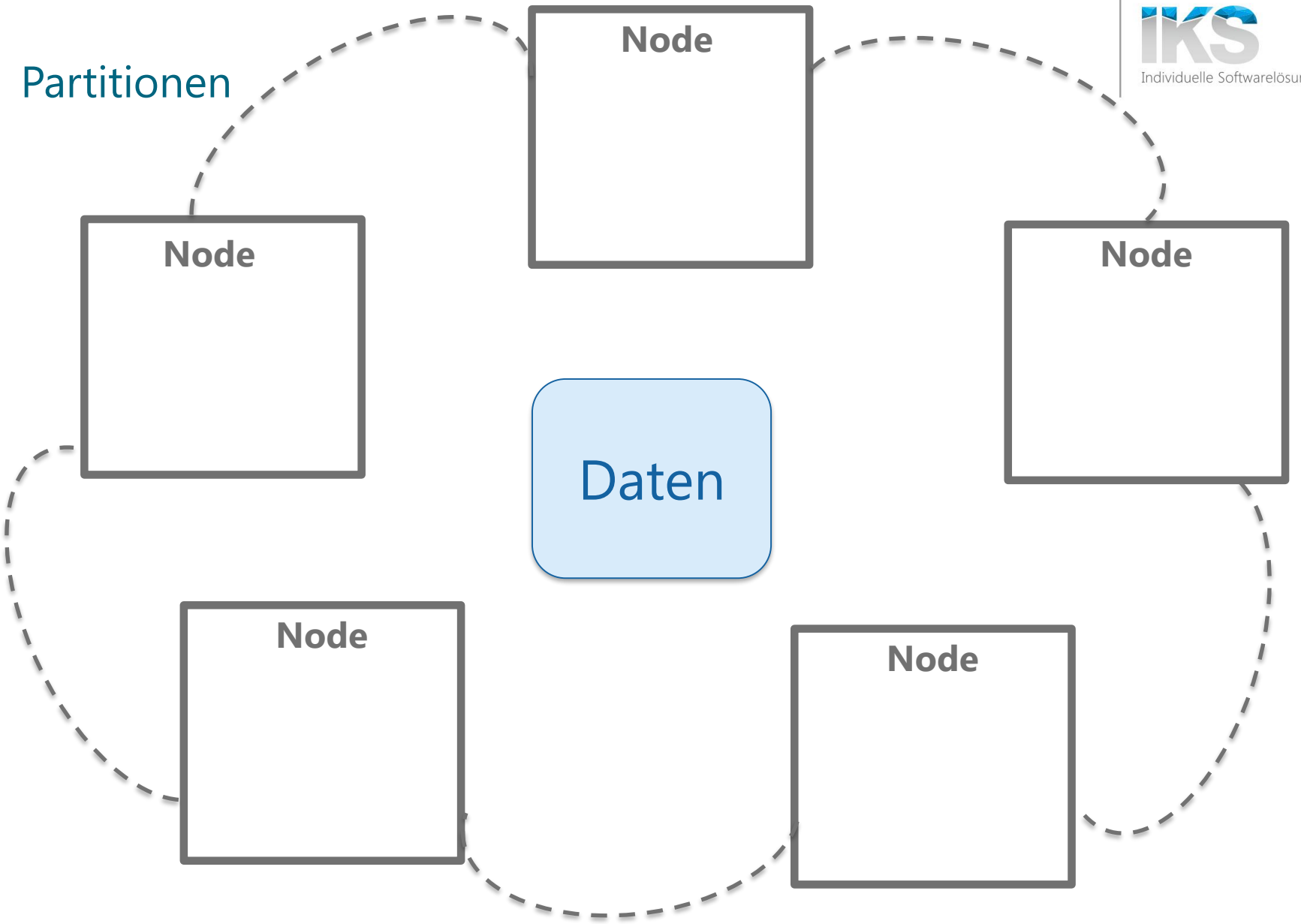
Cluster



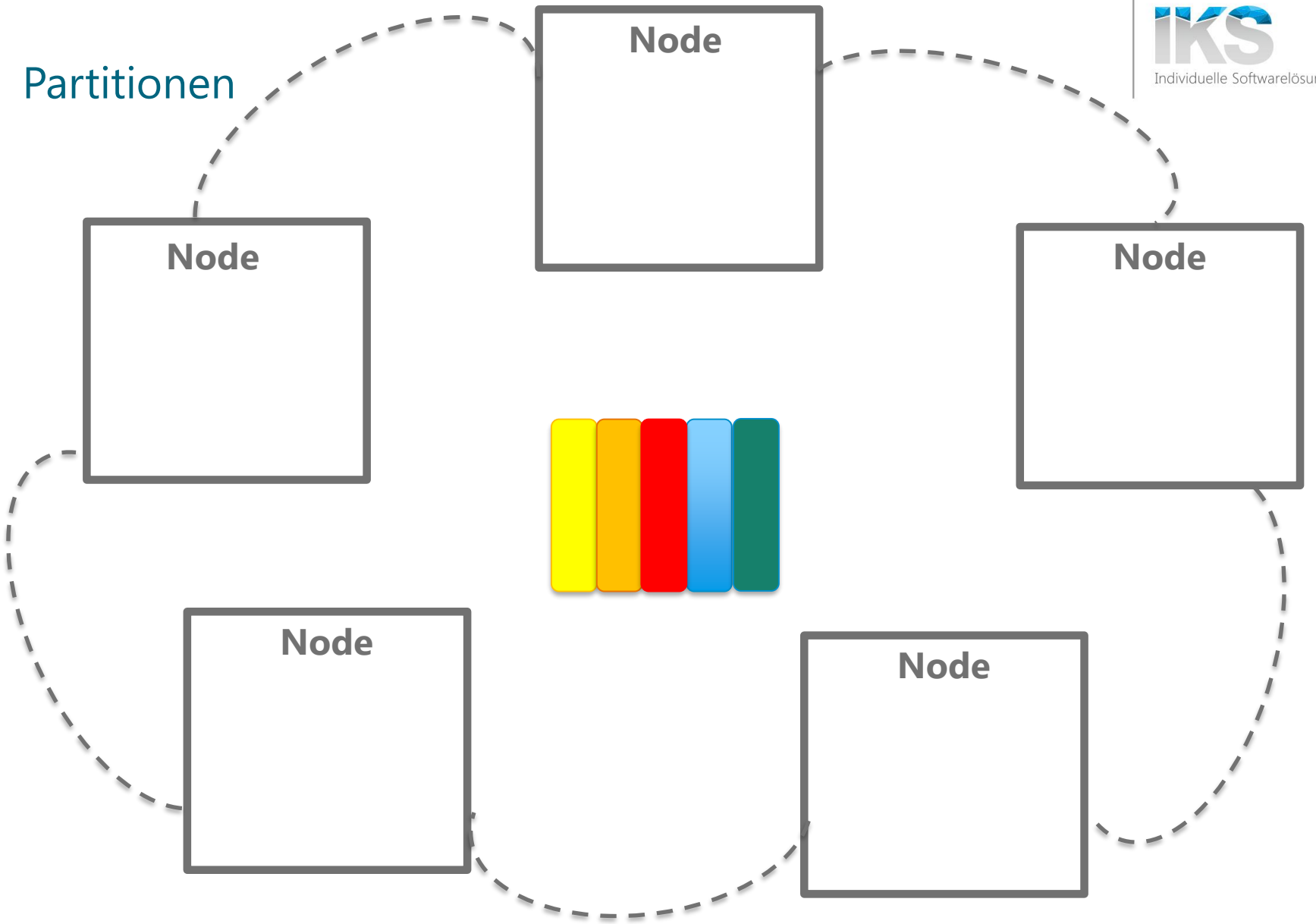
- Ein Cluster repräsentiert eine Menge an Servern/VMs (kurz *Nodes*) als gemeinsames System
  - Die Ressourcen, alle *Nodes*, werden gemeinsam verwaltet
- Cluster besteht i.d.R. aus einem komplexen Zusammenspiel unterschiedlicher Prozesse auf den einzelnen Nodes
- Spezielle Aufgaben sind
  - clusterweites Ressourcenmanagement
  - clusterweite Synchronisation
  - clusterweite Orchestrierung

Partitionen

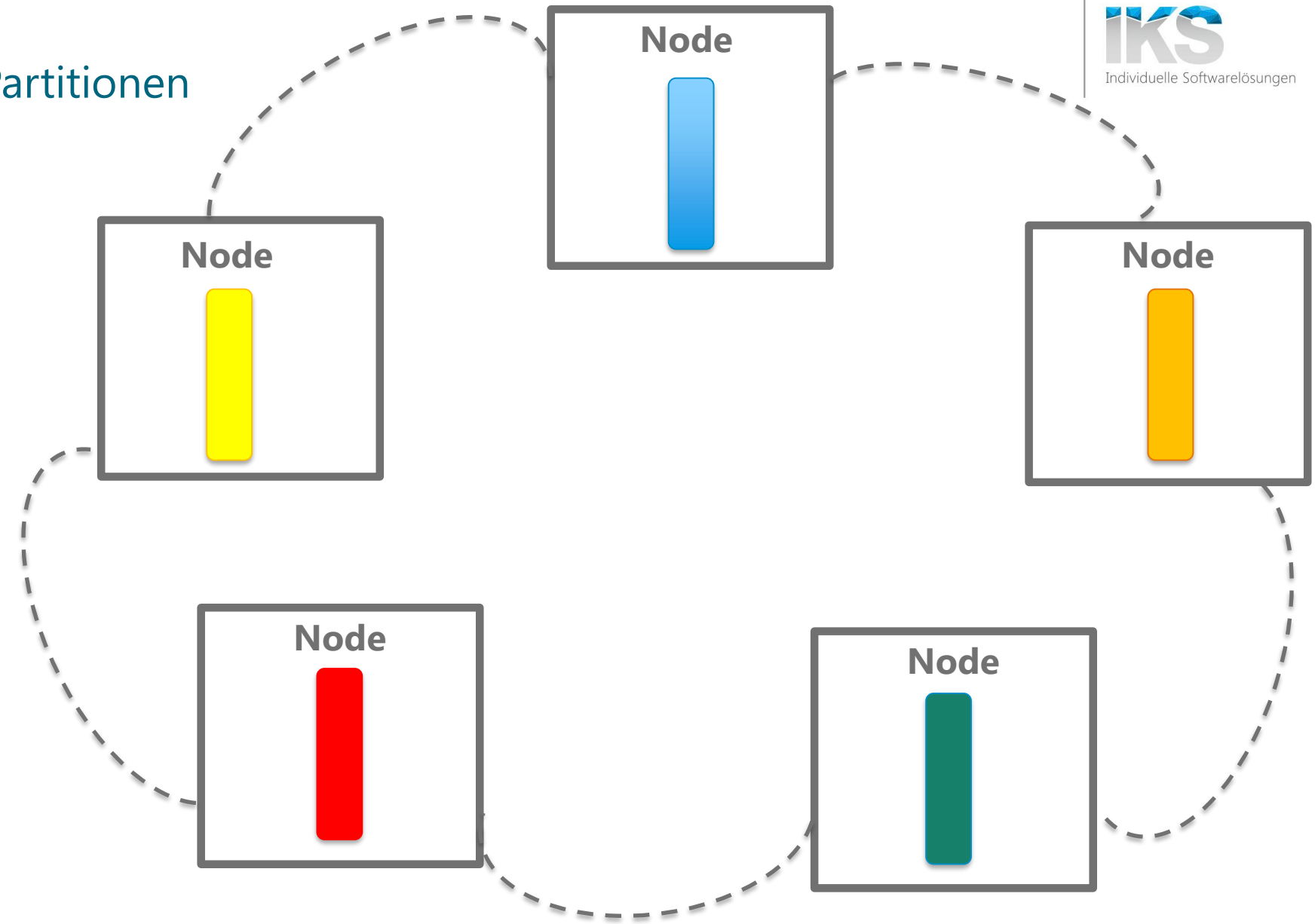
Partitionen



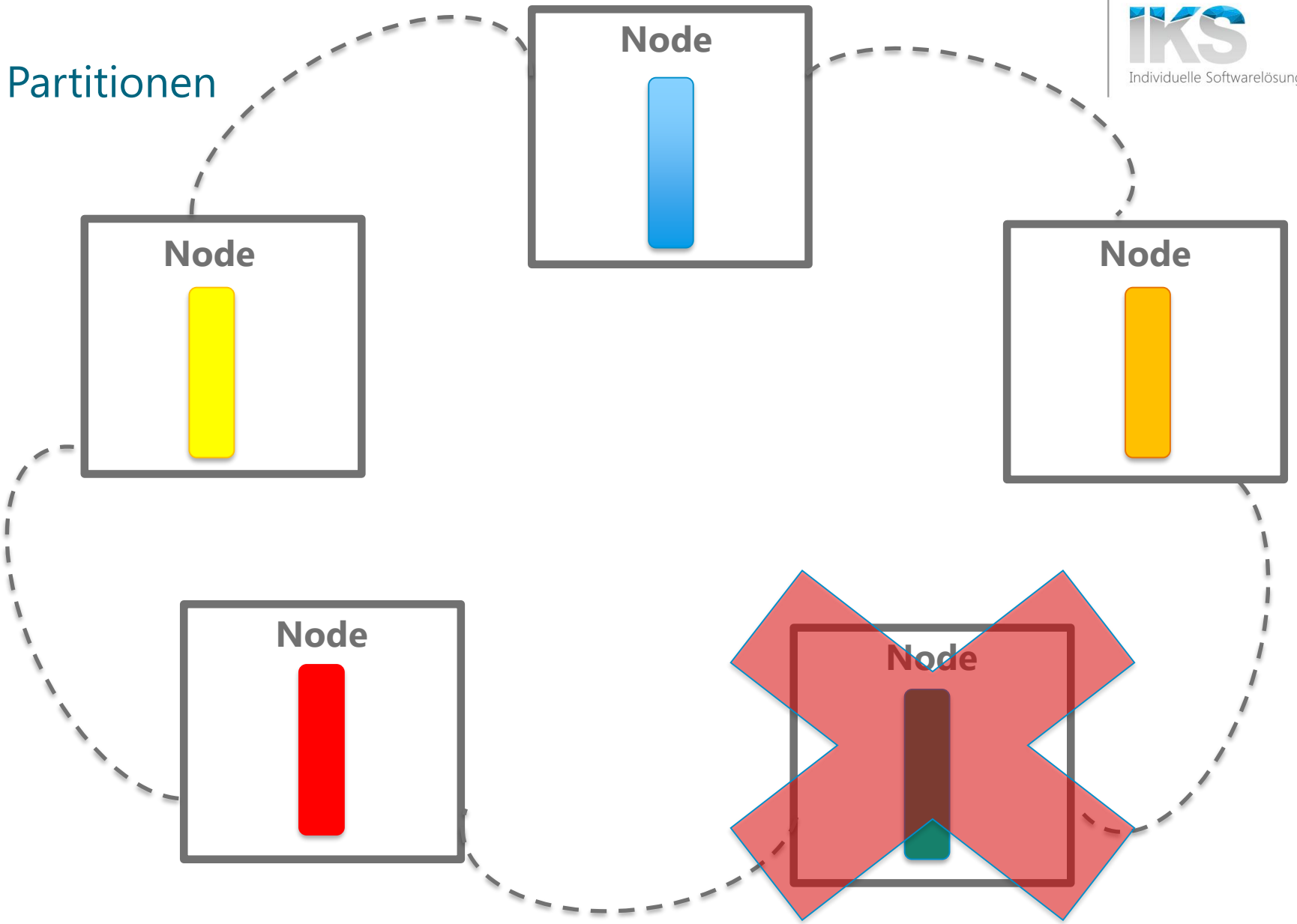
Partitionen



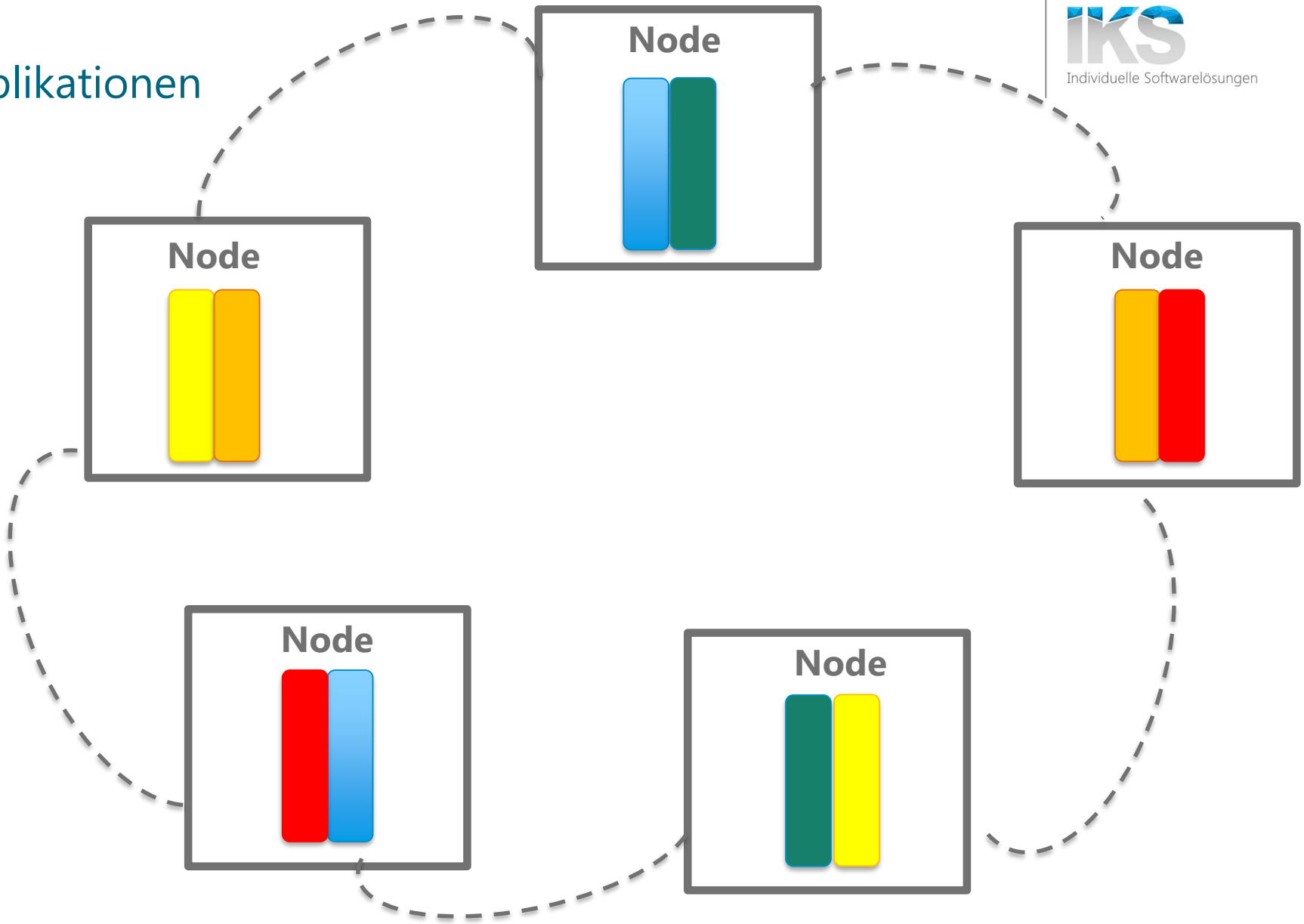
Partitionen



Partitionen



Replikationen



Verteilte Datenhaltungssysteme

- *Replikation* und *Partitionierung* sind grundlegende Konzepte, welche in allen verteilten Datenhaltungssystemen zum Tragen kommen

- Verteilte Datenhaltungssysteme beschäftigen sich damit,
  - wie Daten partitioniert werden
  - wie eine konsistente Sicht auf partitionierte Daten gewährleistet wird
  - wie ein Cluster auszubalancieren ist, wenn einzelne Nodes wegfallen/hinzukommen
  - wie sich ein Cluster verhält, wenn eine große Menge von Nodes wegfallen

Name Dropping

- Verteilte Filesysteme (DFS)
  - Hadoop
- Verteilte NoSql-Datenbanken
  - HBase
  - Cassandra
  - MongoDB
  - Redis
  - Kafka
- Verteiltes Konfigurationsmanagement
  - Zookeeper
  - etcd
  - Consul
- Clustermanagement
  - Yarn
  - Kubernetes
  - Mesosphere

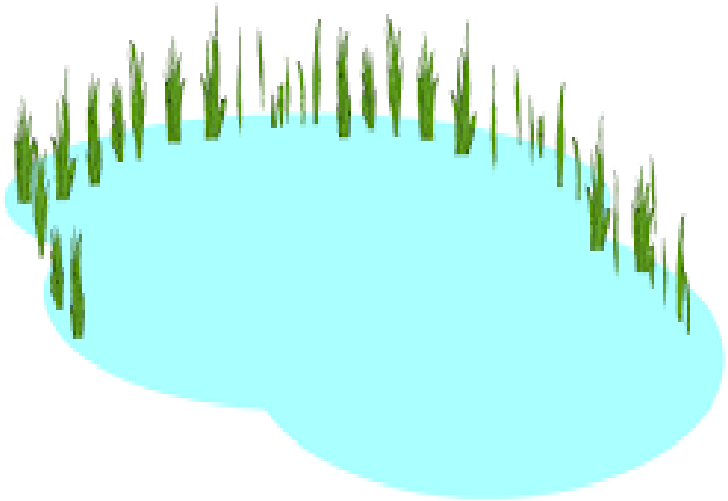
IKS
Individuelle Softwarelösungen

https://urlaubsreich.de/wp-content/uploads/2018/08/Sch%C3%B6n-IMG_5643-ge%C3%A4ndert.jpg

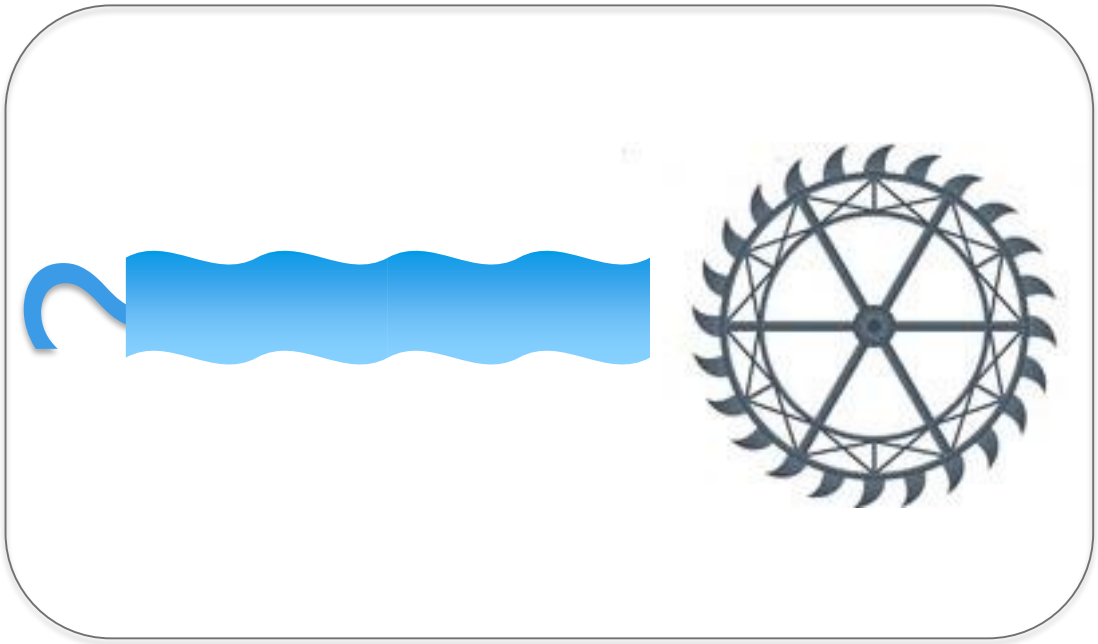
Klassisch: Daten aktiv aus dem Datenspeicher holen

https://www.schwengelpumpe.eu/

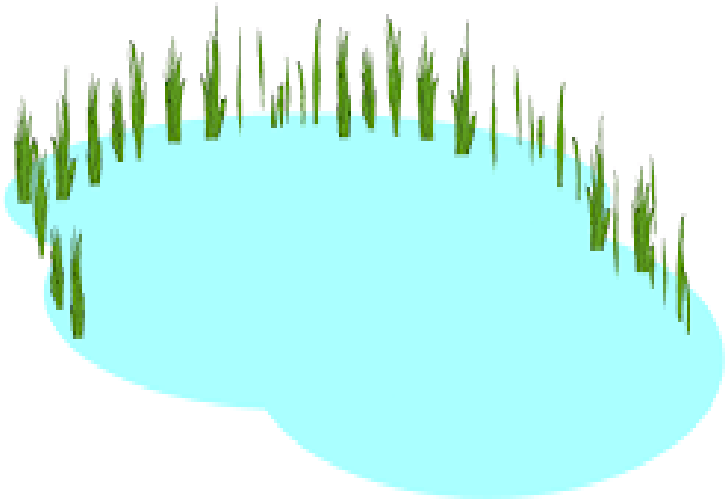
Klassisches Batch Processing



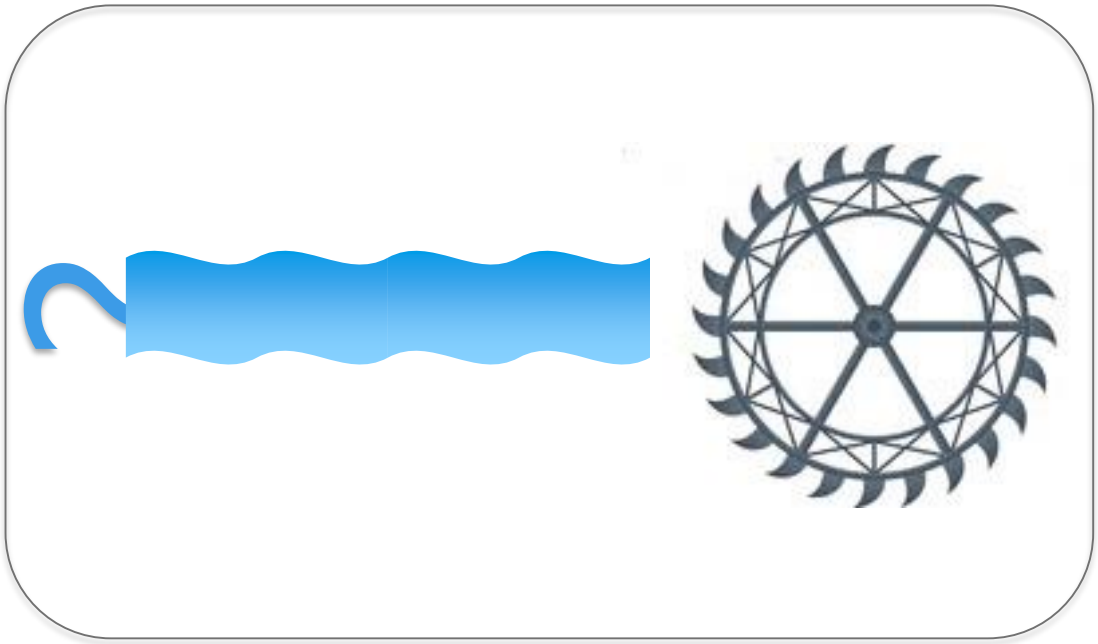
Klassisch:
Datenverarbeitung holt sich ihre Daten



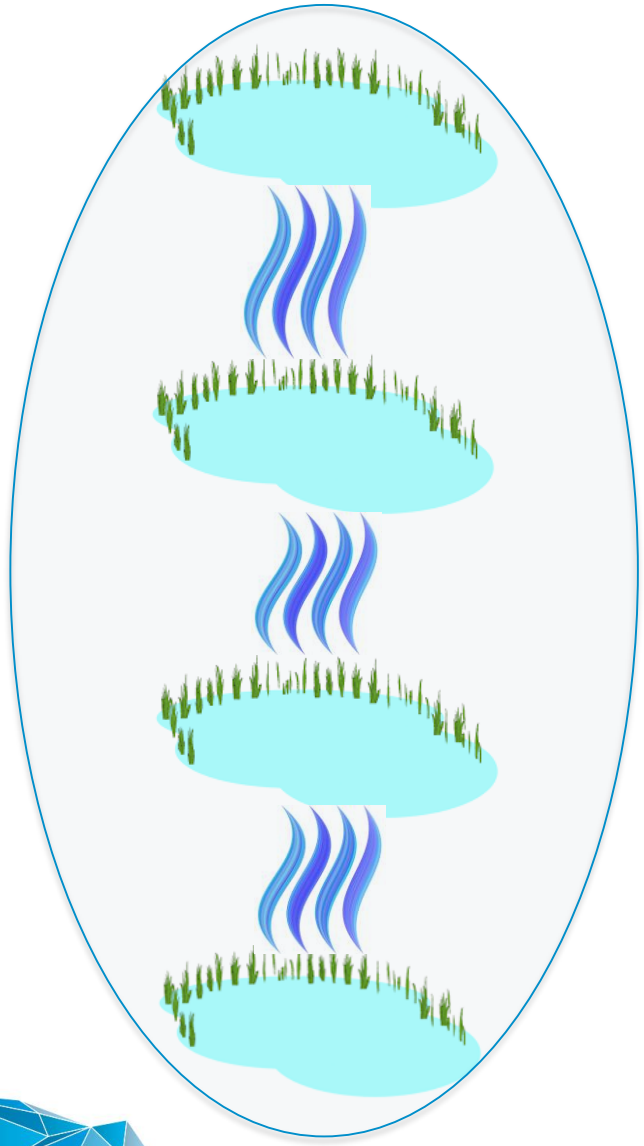
Klassisches Batch Processing



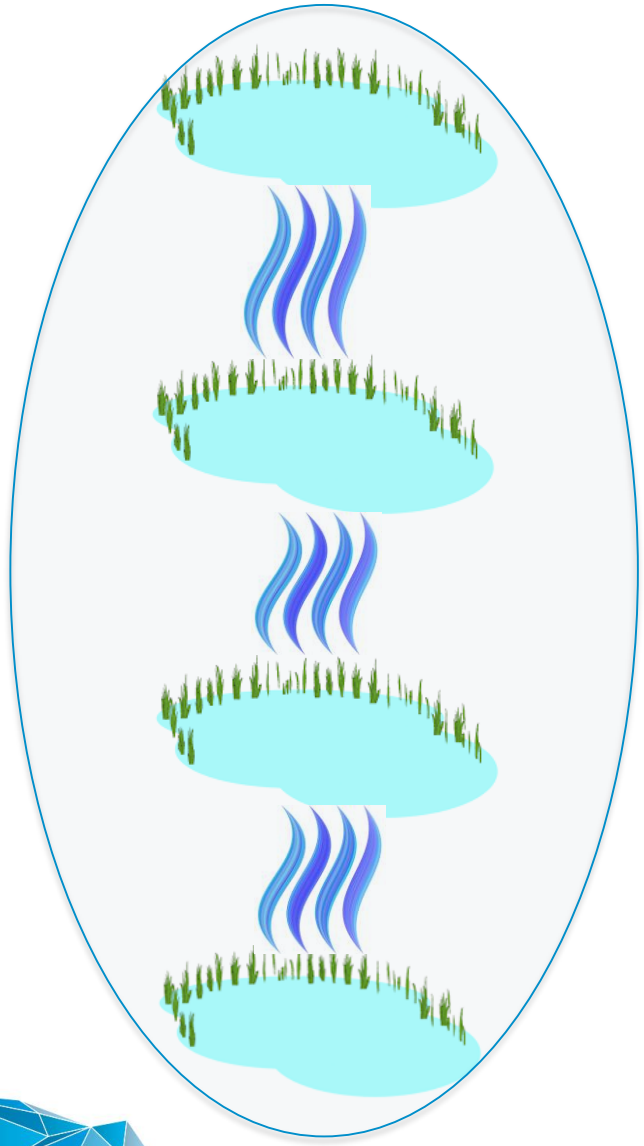
Datenversorgung und Datenverarbeitung in einer Komponente



Batch Processing a là Big Data



Batch Processing a là Big Data (II)



Optimierung: Verteilung der Daten hat Einfluss auf das Programmiermodell

Map Reduce

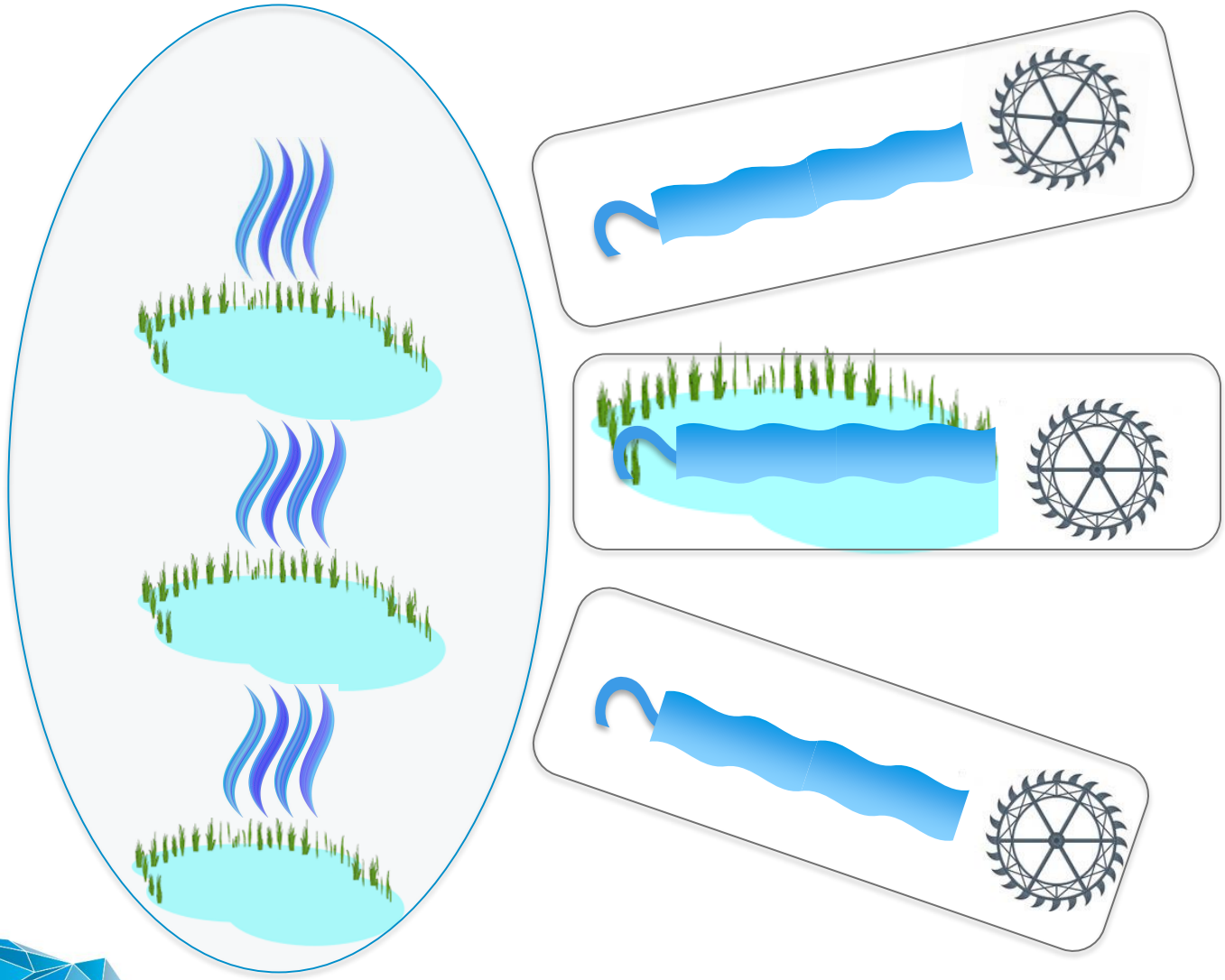
Algorithmus, um Daten aus verteilten Datensystemen zu extrahieren

- *Map Reduce* ist der bekannteste Vertreter dieser Verfahren
- Entwickelt für *Hadoop DFS*
- Eigenes Programmiermodell für verteilte Filesysteme
  - Entwickelt von Google
  - Bekannt als Verfahren über *Hadoop*
- Ist relativ langsam
  - Heutzutage werden andere Verfahren angewendet

Dieser Ansatz funktioniert für beschränkte Verarbeitungsmengen

Warum?

Idee verteilter Verarbeitung: Verteilung auf viele Schultern



Probleme dieser Architektur

- Prämisse *„Verarbeitung holt sich die Daten“* hat folgende Konsequenzen:
  - Jede einzelne Komponente verantwortet, welche Daten zu verarbeiten sind
    - . . ., das kann aber nur im Verbund entschieden werden

    → unverhältnismäßig hoher und nicht linear wachsender Abstimmungsbedarf
  - Daten werden nicht verarbeitet, wenn sie entstehen,
    - sondern wenn sie geholt werden
- Um beliebig große Datenmengen zu verarbeiten,
  - muss diese Prämisse umgestellt werden

Die Verarbeitungskomponente bekommt die zu verarbeitenden Daten zugestellt

Aufteilung der Verarbeitung auf viele Schultern setzt eine dynamische Partitionierung der Daten voraus

Diese Datenmenge muss in Fluss gebracht werden

Trennung von Datenzustellung und Datenverarbeitung

IKS
Individuelle Softwarelösungen

https://afrika-junior.de/inhalt/kontinent/regionen/das-suedliche-afrika-die-suempfe-und-die-namib-wueste/der-sambesi-die-lebensader-im-suedlichen-afrika.html

Stream Processing – Metapher für Daten im Fluss

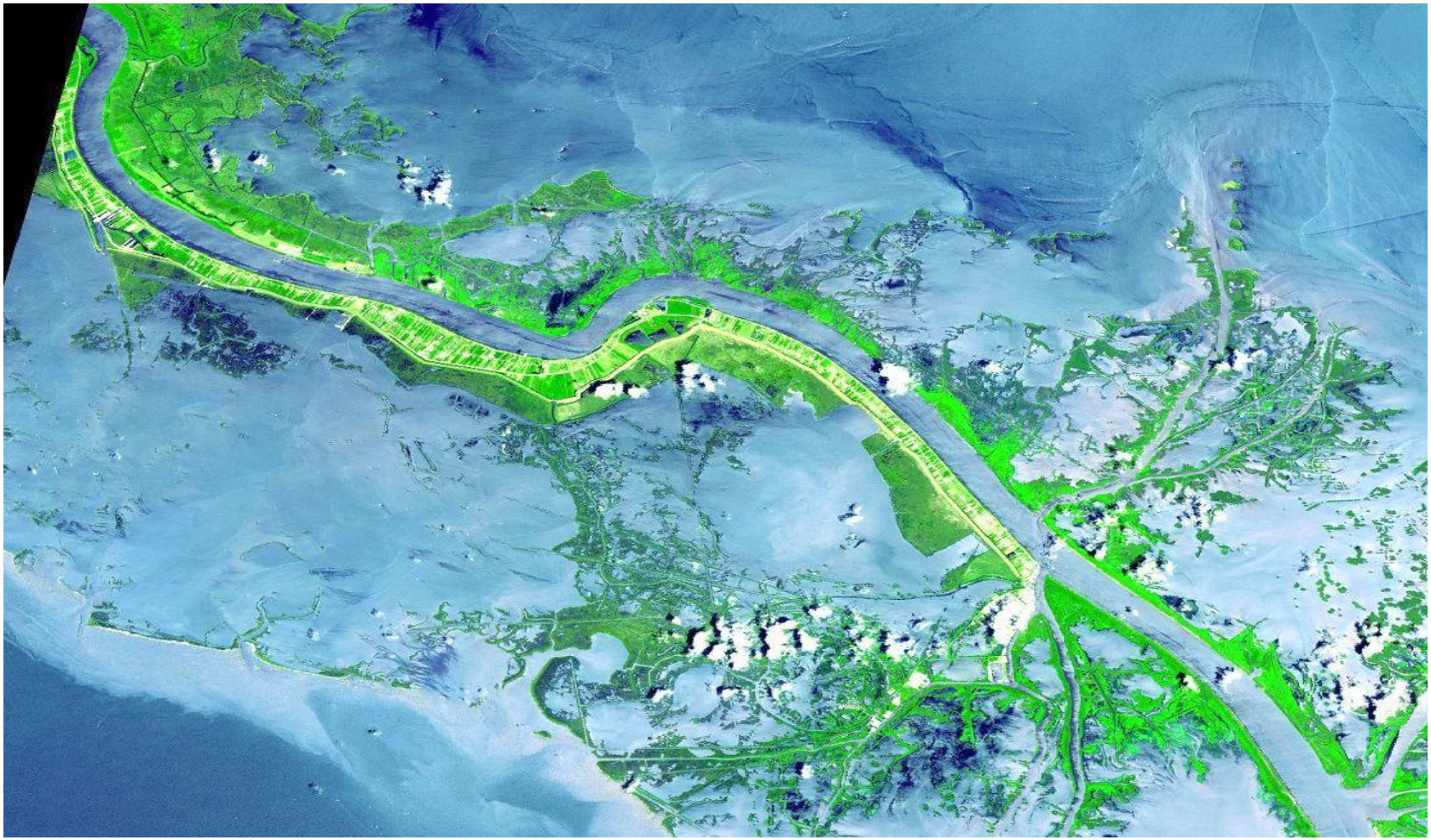
IKS Individuelle Softwarelösungen

- Daten als kontinuierlicher Strom, der der Verarbeitung zugeführt wird
- Datenstrom verästelt sich durch ein System von Leitungen
- Daten wandern/meandern durch dieses System
  - mehr oder weniger gesteuert

“Stream Processing is a type of data processing engine that is designed with infinite data sets in mind”

https://www.oreilly.com/ideas/the-world-beyond-batch-streaming-101

Mississippi Delta – System von Leitungen

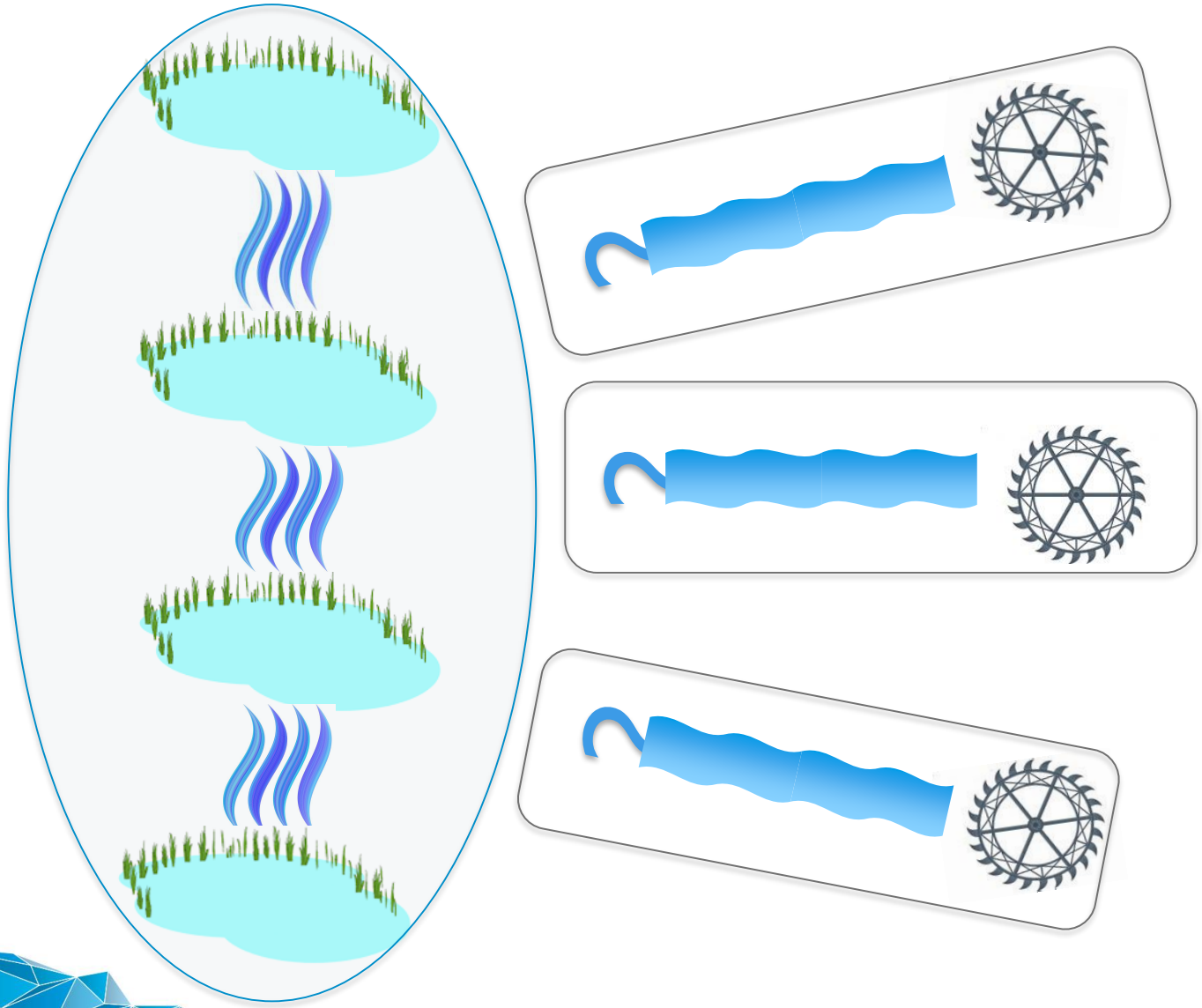


https://commons.wikimedia.org/wiki/File:Mississippi_delta_from_space_detalle.jpg

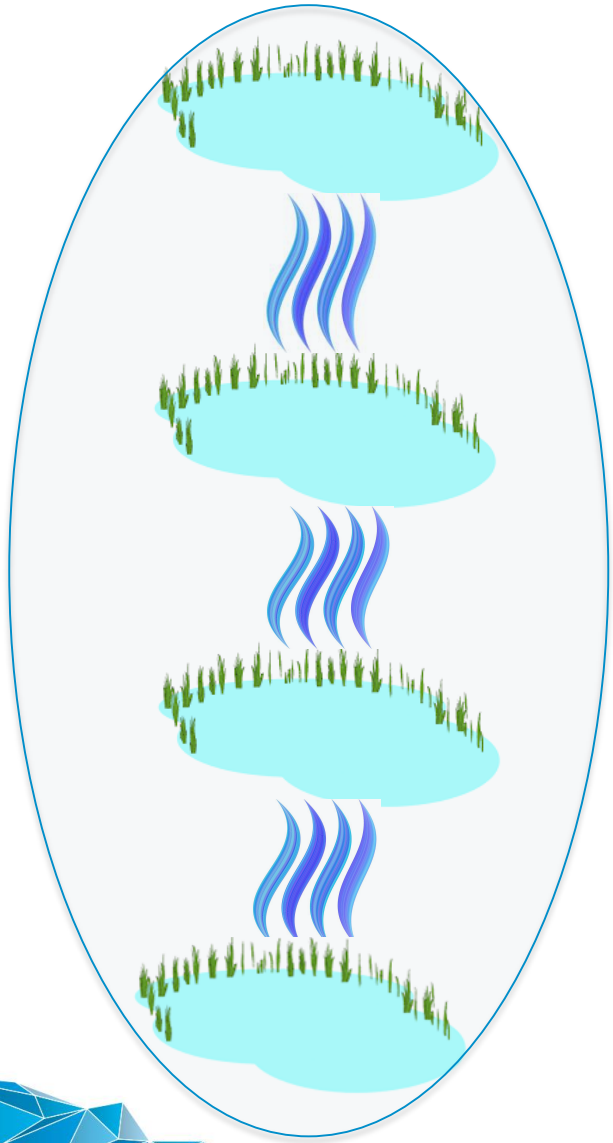
Blattadern – System von Leitungen

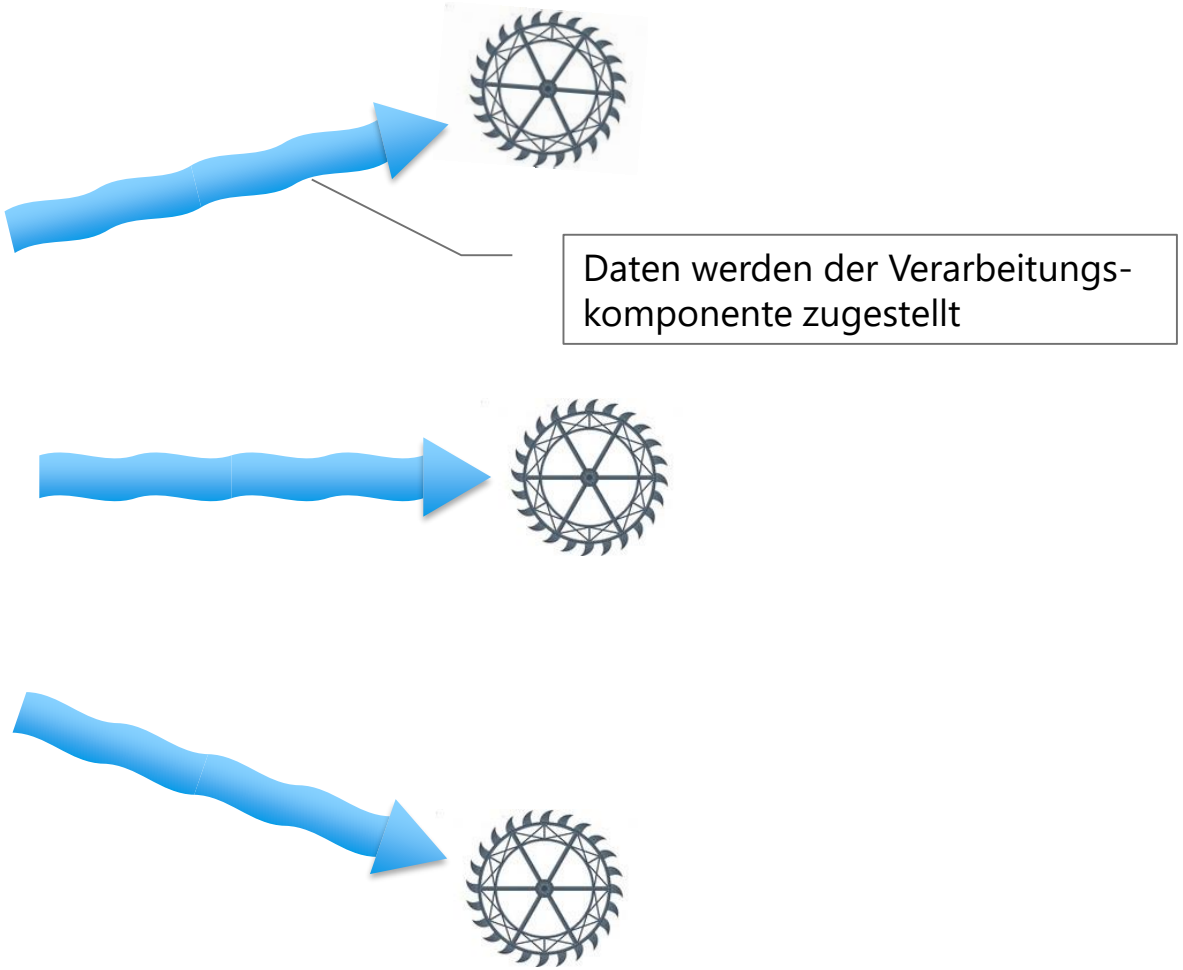
https://blendezwo.de/wp-content/uploads/2015/03/db090510102.jpg

Pull-Prinzip



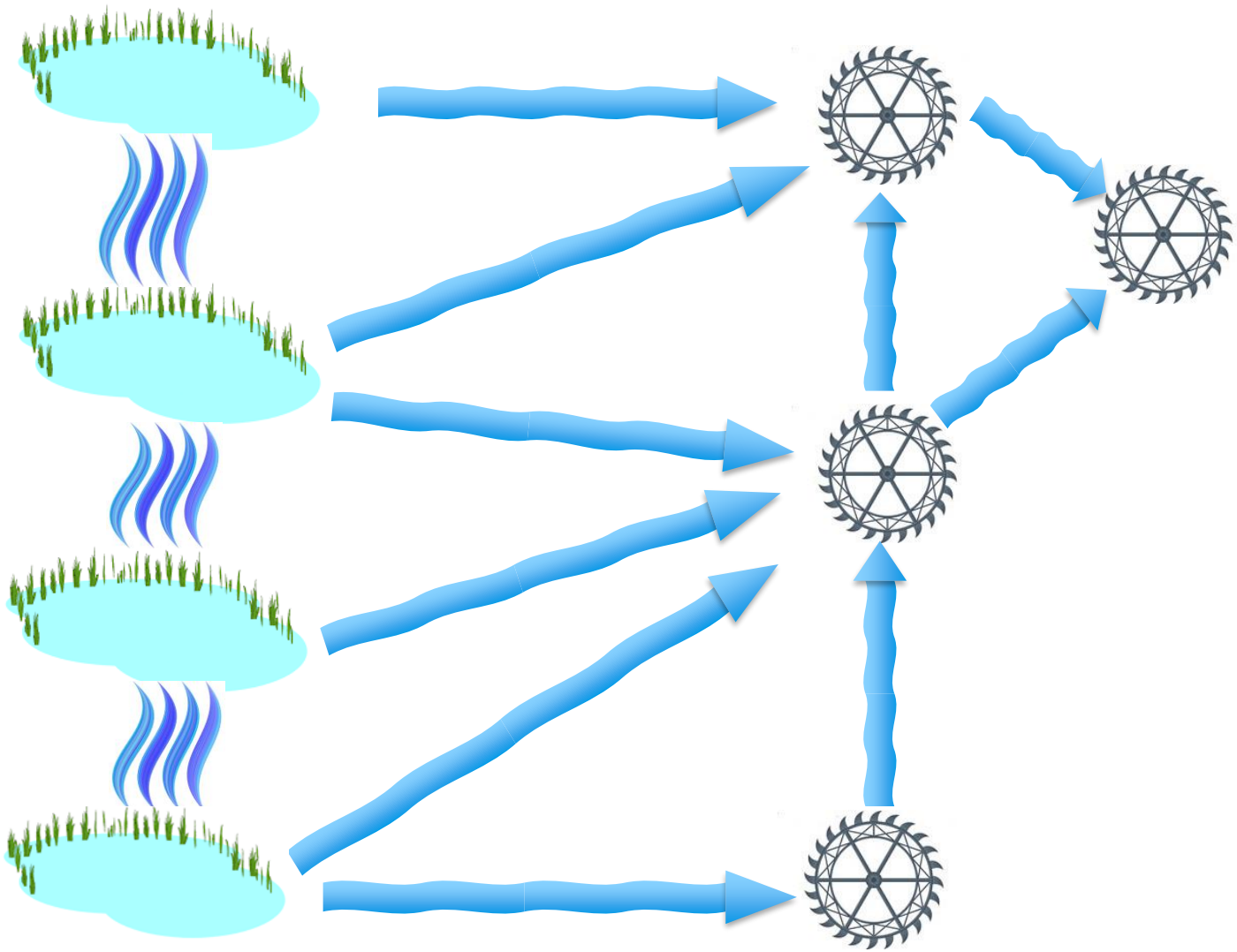
Push-Prinzip





Daten werden der Verarbeitungs-komponente zugestellt

Daten fließen



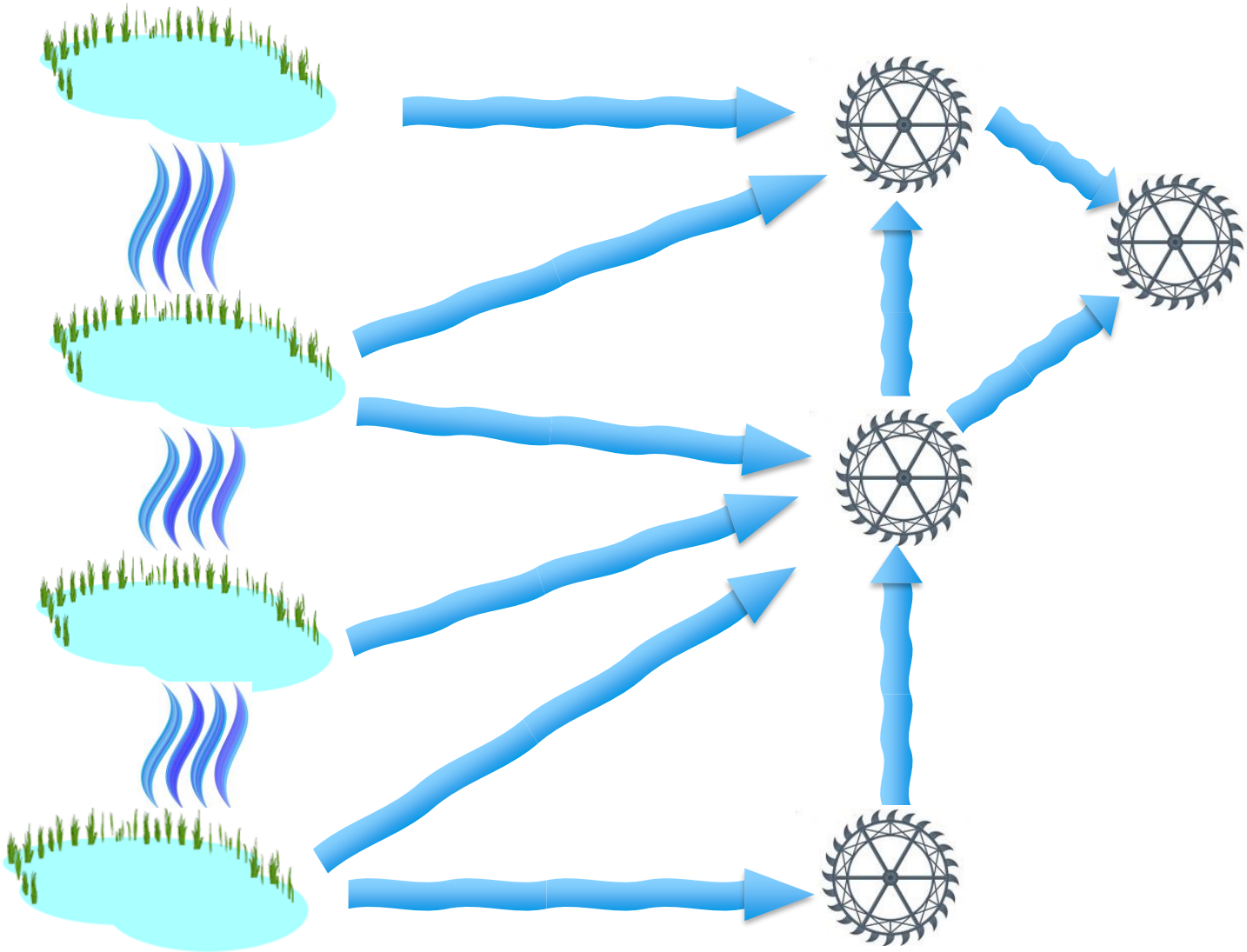
Stream Processing

- Metapher
  - Daten als kontinuierlicher Strom, der der Verarbeitung zugeführt wird
- Es gibt eine Infrastruktur, welche
  - Daten den Verarbeitungskomponenten zustellt (*data driven*)
  - mit schwankenden Mengen von Daten zurechtkommt (*Elastizität*)

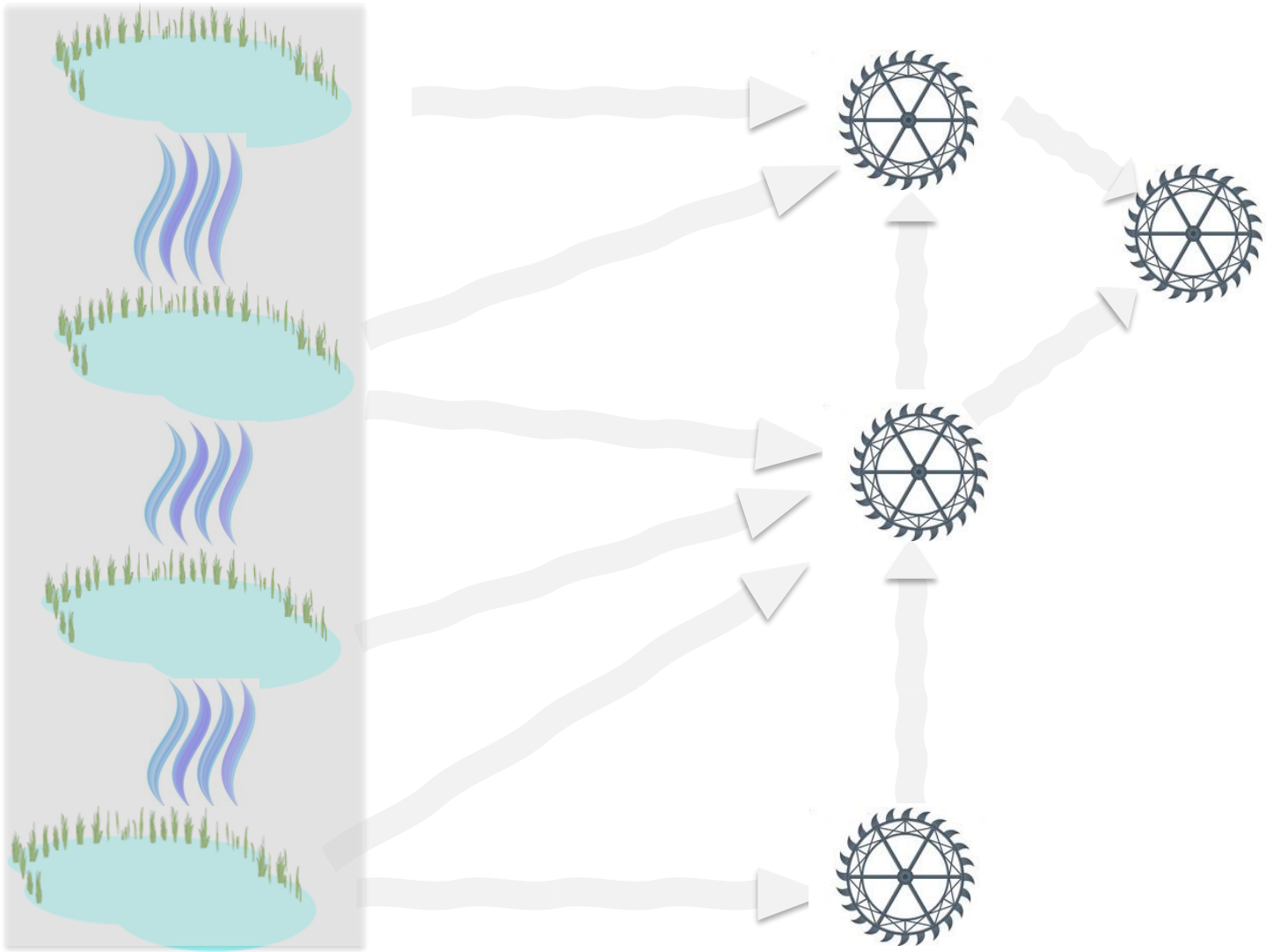
Was heißt dies für die Verarbeitung?

- Datenverarbeitung wird von kleinen, autonomen Einheiten durchgeführt
  - *Processoren* genannt
- *Processoren* beschreiben, für welchen Typ von Daten sie zuständig sind
  - … und bekommen diese Daten durch die Infrastruktur zugestellt
- *Processoren* werden im Cluster verteilt und skaliert
- *Processoren* werden durch die Infrastruktur mit Daten versorgt …
  - … verarbeiten diese …
  - … und geben diese an die Infrastruktur zurück

Wo bleiben Ihre Programme?



Wo bleiben Ihre Programme ?



Name Dropping

- Messaging Systeme
  - Rabbit MQ
  - Active MQ
  - . . .
- Hybride zwischen verteilter Datenhaltung und Messaging
  - Kafka
- Dezidierte *Streaming Engines*
  - Apache Spark
  - Apache Flink
  - Apache Samza

IKS
Individuelle Softwarelösungen

Big-Data-Architektur

https://www.mch-group.com/en-US/news/blog/2016/05/powertage-2016-messe-zuerich.aspx

Es sind Architekturen entstanden, die …

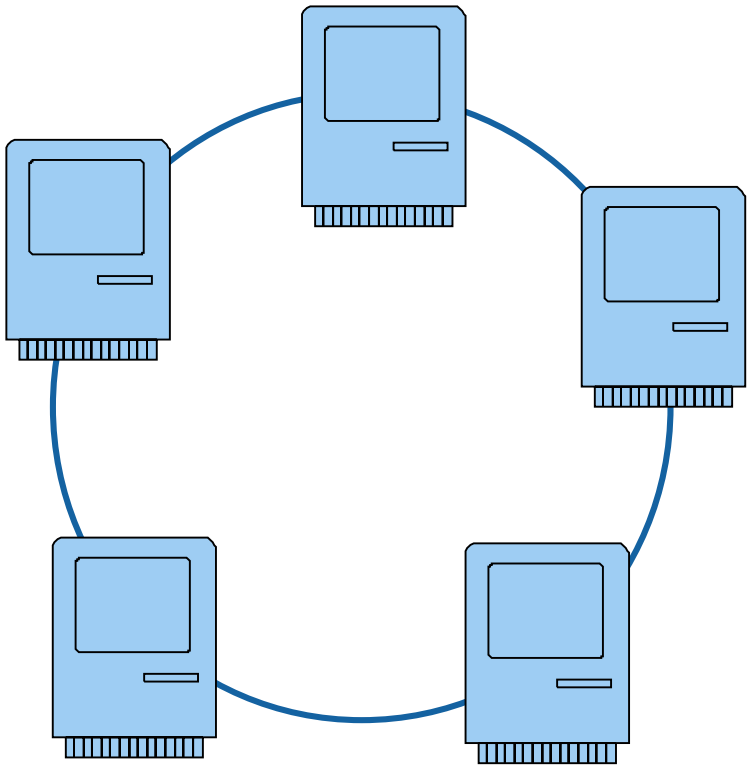
- sehr gut horizontal skalieren
  - bzgl. Datenmengen
  - bzgl. Datendurchsatz
  - bzgl. Ressourcen
- hochgradig parallele Anwendungen ermöglichen
- sehr gut mit schwankenden Datenmengen umgehen können
- Daten verarbeiten, sobald diese im System verfügbar sind
- lose gekoppelte Architekturen propagieren
  - Je enger die Kopplung, desto schwieriger ist die Skalierung
- i.d.R. Open Source sind
  - Es gibt allerdings immer kommerzielle Erweiterungen und Support

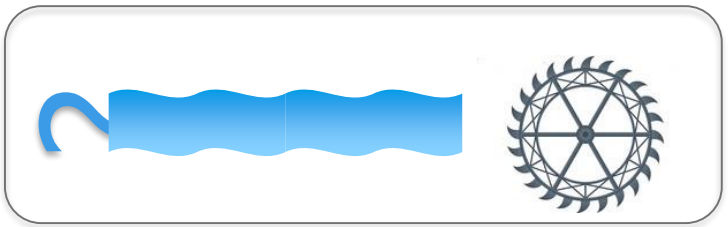
Technologische Lösungsansätze

- *Cluster*
  - Horizontale Skalierbarkeit/Elastizität
- *Data Storage* mit verteilten Datenhaltungssystemen
  - Strukturierte sowie un- und semistrukturierte Daten
  - Gemischte Datenformate gleichzeitig
- *Batch Processing* mit speziellen Programmiermodellen
  - Für Verarbeitung begrenzter Datenmengen
- Stream Processing
  - Datengetriebene Verarbeitung (*data driven*)
    - Daten werden den *Processoren* zugestellt
  - Asynchrone Verarbeitung
    - mit Messagingsystemen
    - mit speziellen *Stream Processing Engines*
  - Parallele und skalierbare Verarbeitung

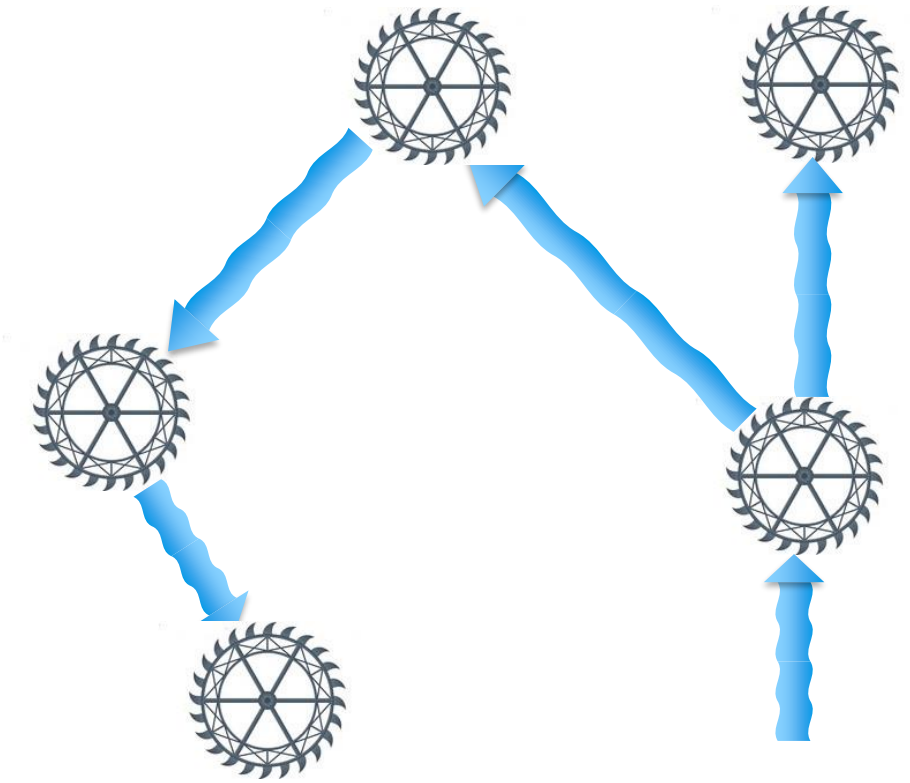
- Diese Architekturen sind historisch durch Big Data entstanden
- Bieten ganz neue Einsatzmöglichkeiten
  - Auch jenseits von „riesigen Datenmengen“

Active Archiv

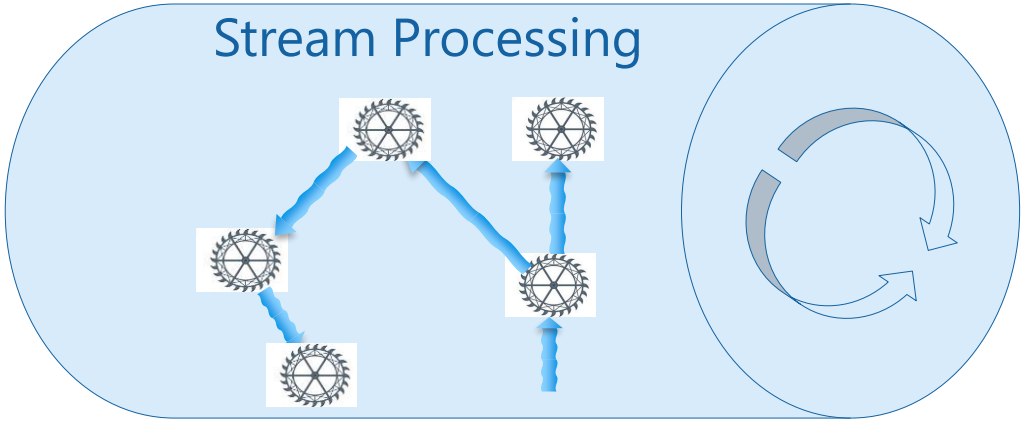




Streaming Architekturen



Streaming Architekturen



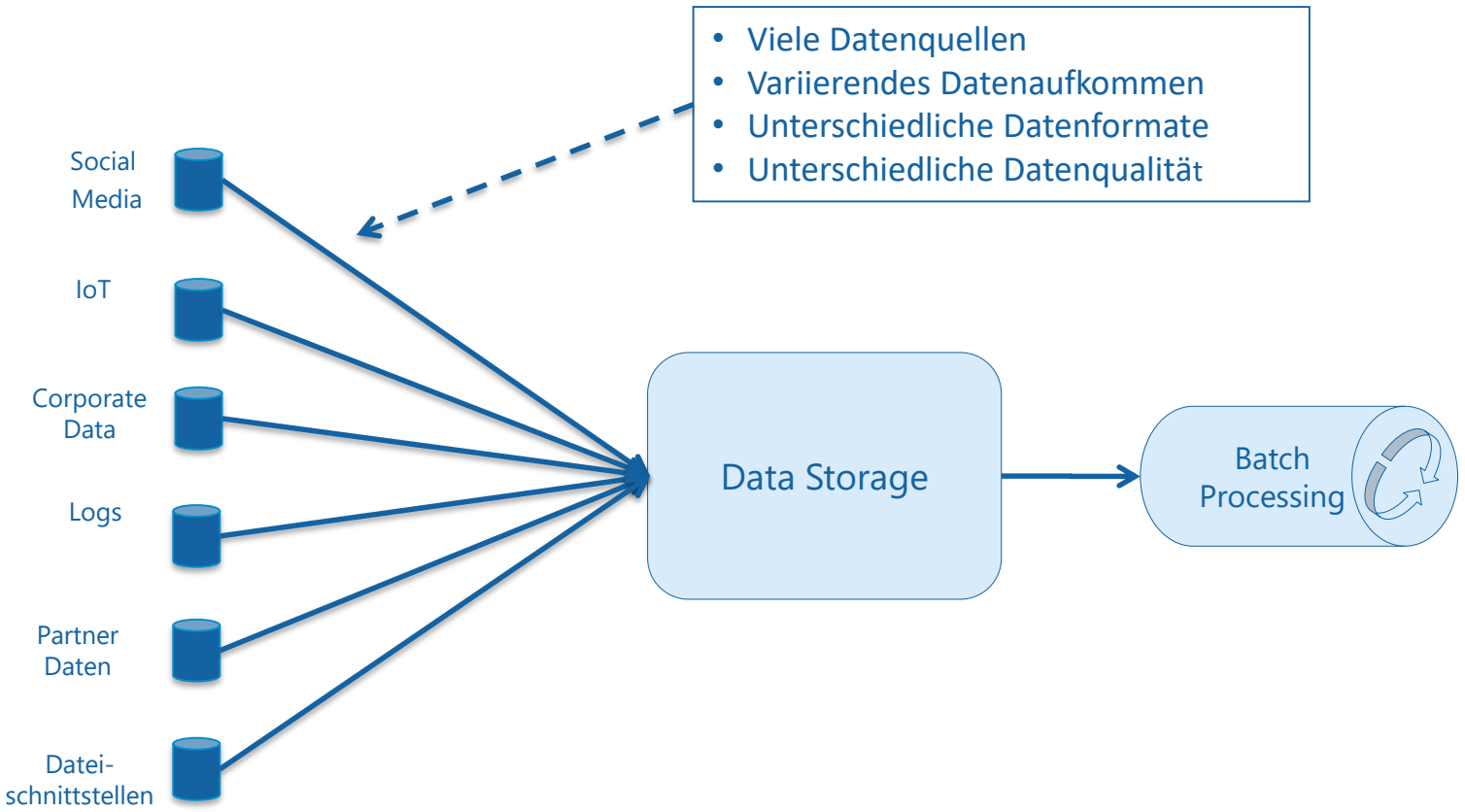
Stream Processing und *Machine Learning*

- *Machine Learning* und *Stream Processing*
  - *Machine Learning* sucht Muster in einem Strom von Daten
  - Ergänzen sich ideal
  - *Machine Learning* ist i.d.R. ressourcen-intensiv
  - Alle *Streaming Engines* besitzen *Machine-Learning*-Funktionalität
- Beispiel: *Fraud Detection*

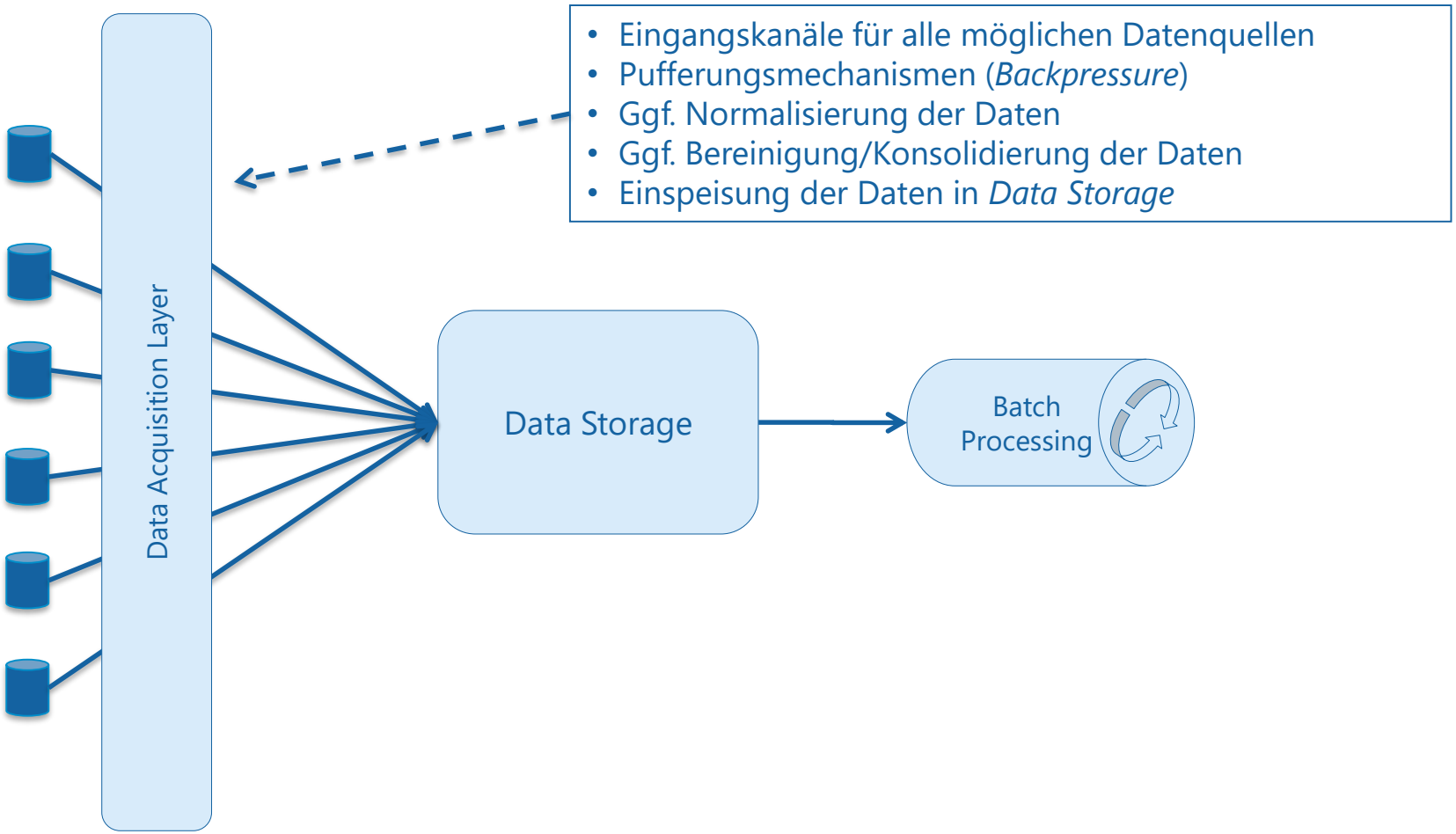
- Spezielle Anforderungen an ML
  - Zustandsverwaltung
  - Fehlerhandling
  - Rekursionen
  - Behandlung von *bounded Windows*

Wie kommen die Daten in das System?

Wie kommen die Daten in das System?

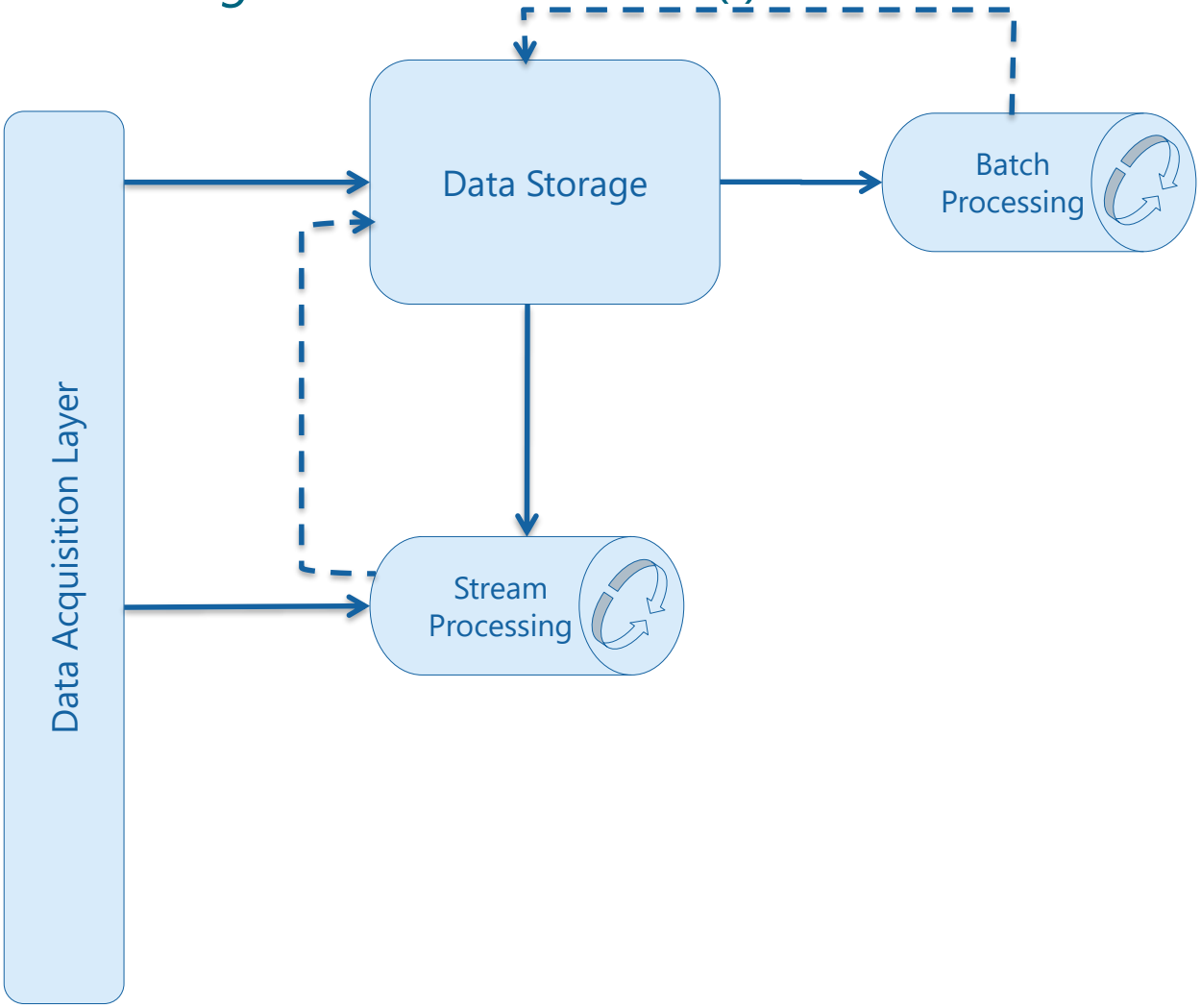


Data Acquisition Layer

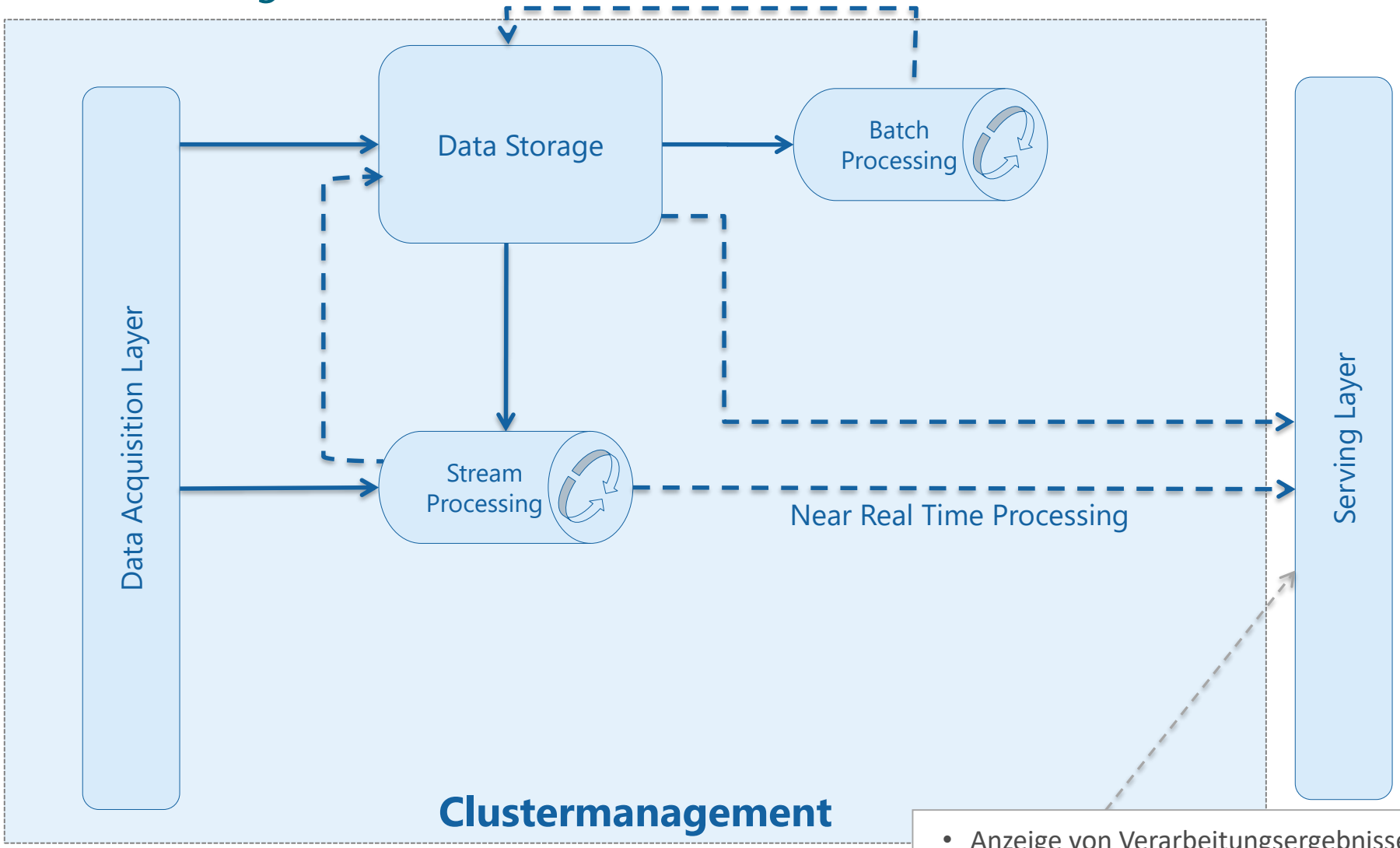


- Eingangskanäle für alle möglichen Datenquellen
- Pufferungsmechanismen (*Backpressure*)
- Ggf. Normalisierung der Daten
- Ggf. Bereinigung/Konsolidierung der Daten
- Einspeisung der Daten in *Data Storage*

Skizze *Big Data Architecture* (I)

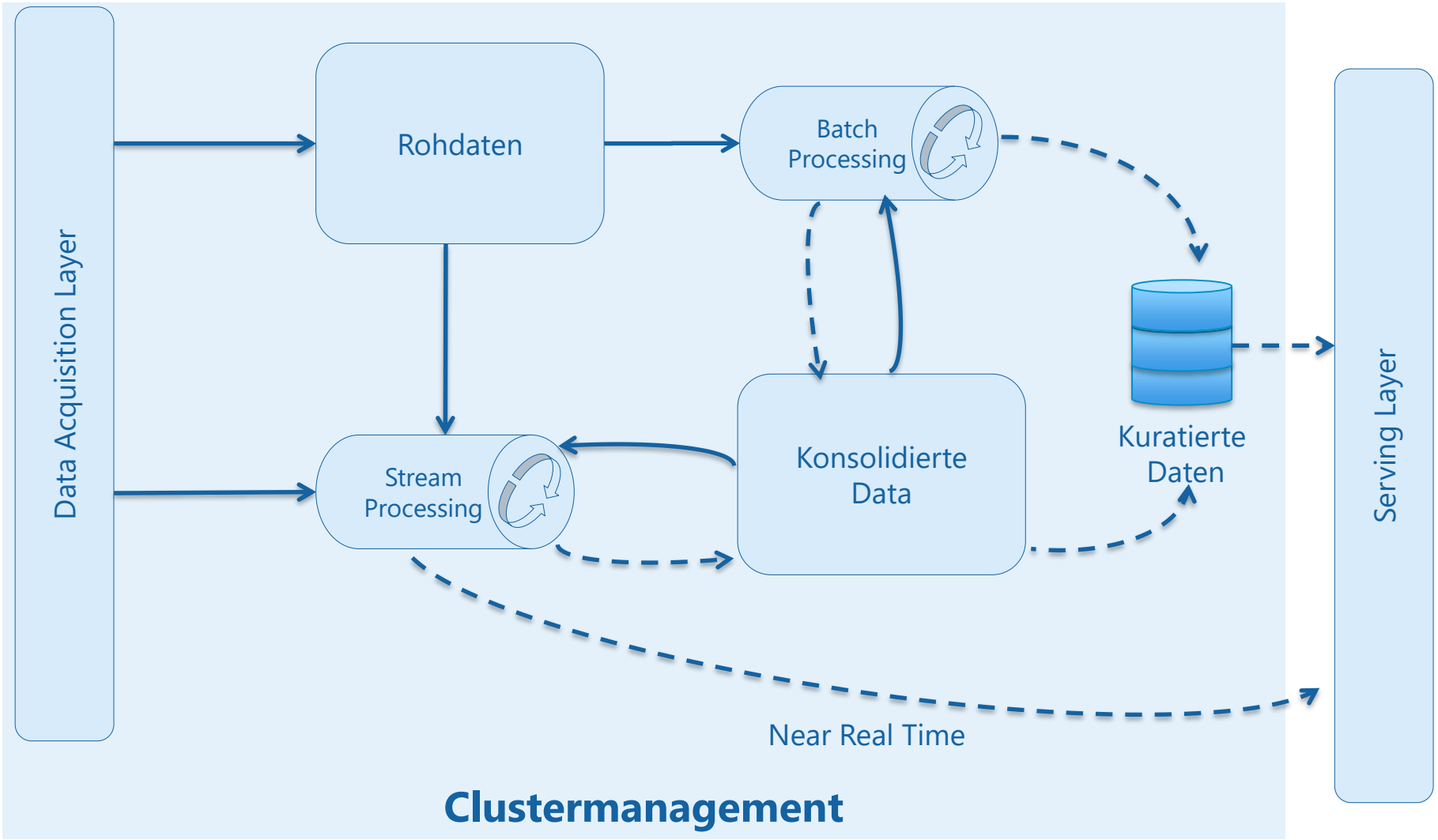


Skizze *Big Data Architecture* (I)

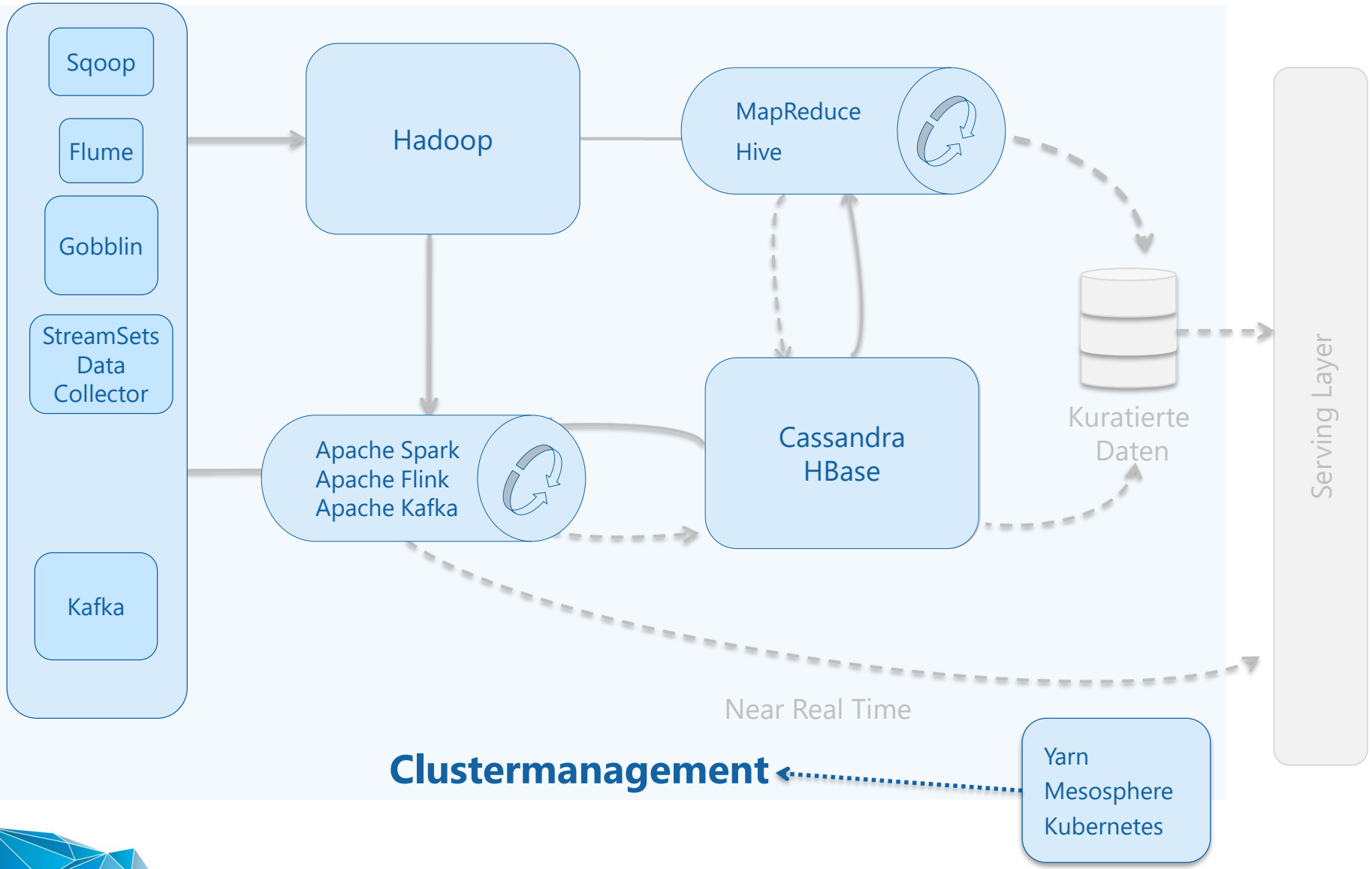


- Anzeige von Verarbeitungsergebnissen
- Schnittstellen für *self-service Data*
- BI

Skizze *Big Data Architecture* (II) – Blick auf die Daten



Name Dropping (Apache/Hadoop Eco System)



Was haben wird noch nicht betrachtet?

- Big Data und Datenserialisierung
  - Serialisierung bei Persistenz von beliebigen Datenstrukturen
  - Problem kleiner Daten bei HDFS
- Big Data und Security/Mehrmandantenfähigkeiten
  - Authentifizierung, Autorisierung, feingranulare Zugriffsberechtigungen, Datenverschlüsselung, Kommunikationsverschlüsselung
- Administration
  - Administration der Plattform
- Auswertungen und *self-service data*
  - Öffnen der Plattform für *data scientists*
- Data Governance
  - Orchestrierung des Zusammenspiels von Menschen, Prozessen und Daten
- Big Data und Hardware
  - Insbesondere Big Data/Distributed Processing in virtualisierten Umgebungen

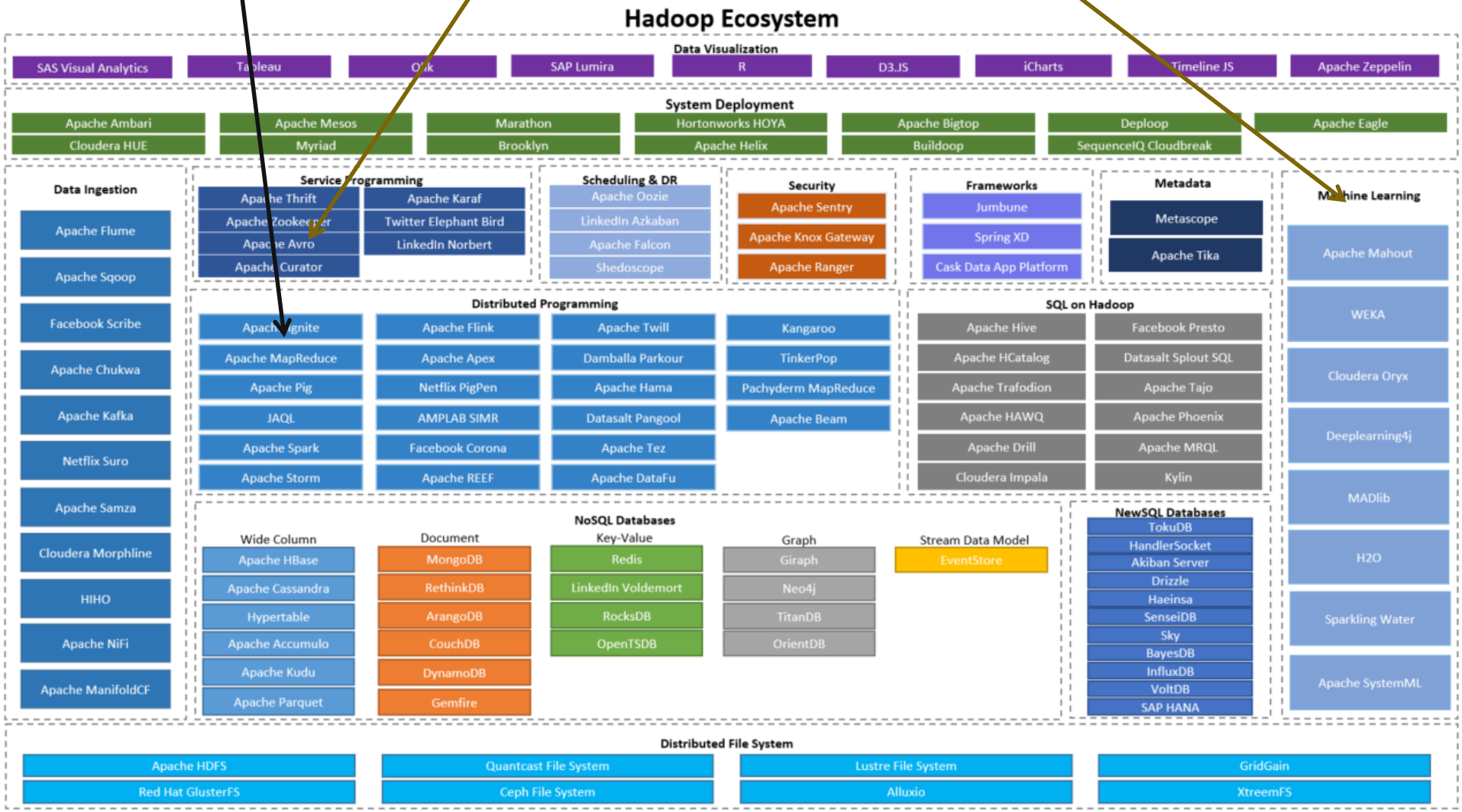
Hadoop Ecosystem

Stream und Batch Processing

Datenserialisierung

Machine Learning
- komplexe Datenanalyse

IKS
Individuelle Softwarelösungen



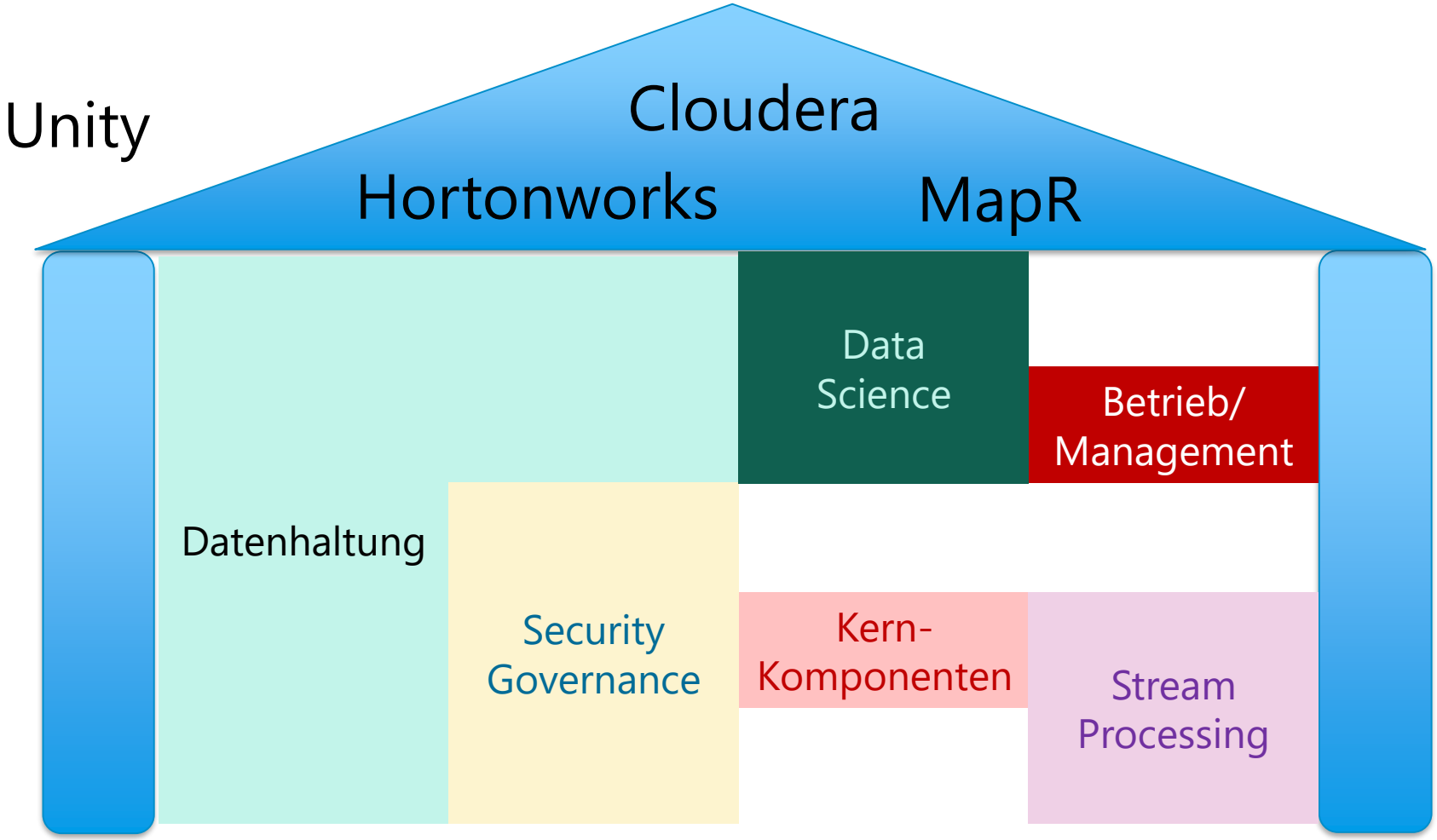
https://autofei.wordpress.com/2017/07/06/hadoop-ecosystem/

Hadoop Eco System

- *Hadoop Eco Systems* ist groß und im Fluss

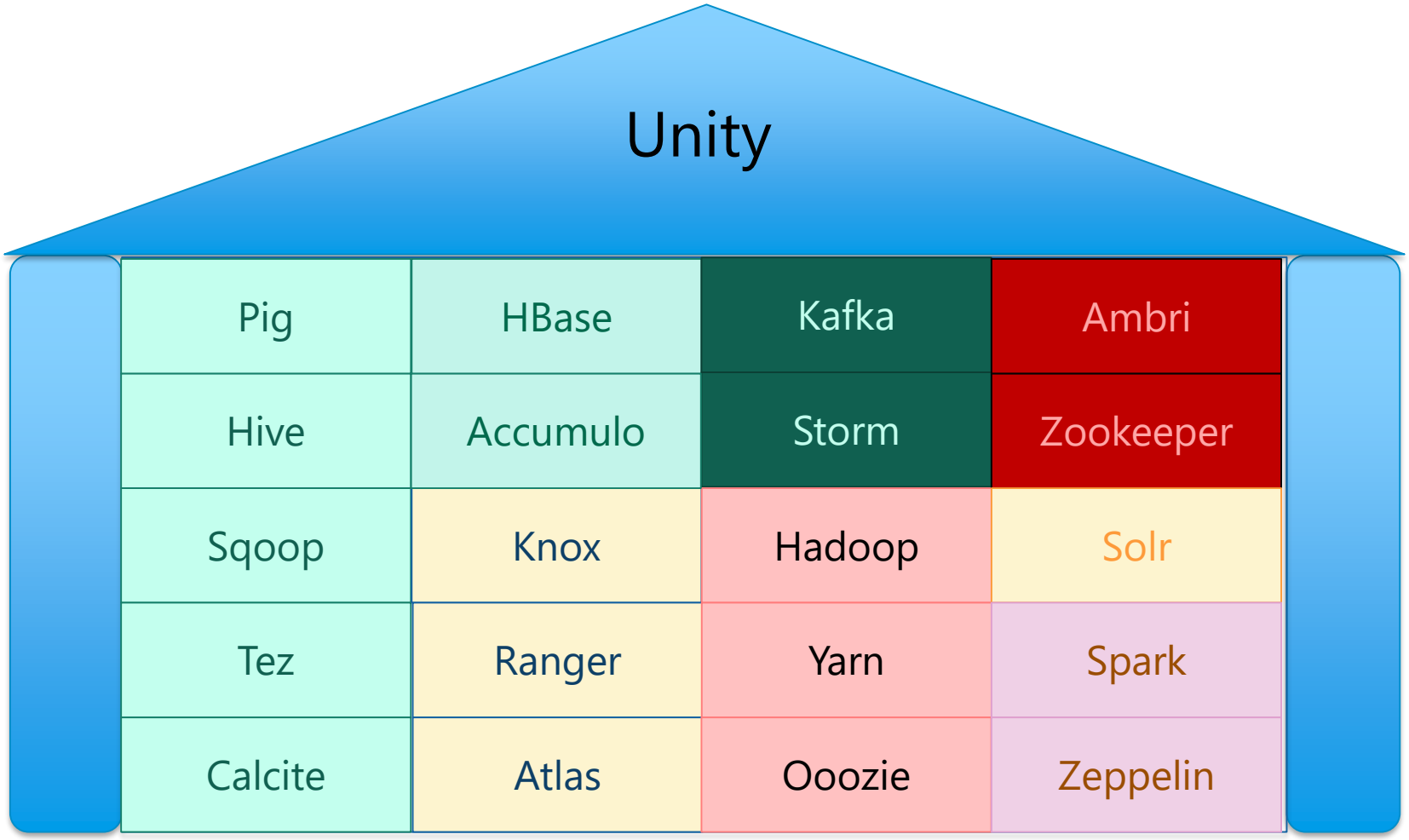
- Anbieter von *Hadoop-Plattformen* bündeln die einzelnen Tools
  - Bieten eine konsistente Plattform an
    - Abgestimmte Funktionalitäten und Versionen
    - Transparente Integration zwischen den einzelnen Tools
  - Entwickeln insbesondere Managementwerkzeuge aktiv weiter

Anbieter von *Hadoop-Plattformen*



https://searchdatamanagement.techtarget.com/news/252455907/Cloudera-and-Hortonworks-combo-to-push-CDP-machine-learning

Hadoop-Plattform am Beispiel von *Unity*



https://de.hortonworks.com/ecosystems/

Big Data in der Cloud

- Bisher haben wir betrachtet:
  Selbst verwaltete Plattformen auf Basis des *Hadoop Eco System*
  - *on premise*
- Infrastruktur für *Elastizität* und *Data Processing* wird an die *Cloud* delegiert
  - Es werden u.U. andere Komponenten zum Processing und zur Datenhaltung genutzt
  - Die *Processoren* werden selbst entwickelt und deployed
- Die Infrastrukturkomponenten mögen andere sein . . .
- . . . aber die Konzepte/Prinzipien bleiben die Gleichen

Abschluss

Big Data – nicht nur für riesige Datenmengen

IKS
Individuelle Softwarelösungen

- Lösungen aus *Big Data* sind nicht nur für riesige Datenmengen zugeschnitten

- Mögliche Architekturtreiber für den Einsatz von *Big-Data*-Lösungen:
  - *Event-* oder *message driven* Systeme/Architekturen
  - hohe Anforderungen an Elastizität/horizontale Skalierbarkeit
  - dynamisches Datenaufkommen
  - un- oder semistrukturierte Daten
  - (dynamische) Auslastung von Systemen
  - (dynamische) Verteilung ressourcen-intensiver Operationen
  - Einsatz von ML

Big Data – technologische Herausforderungen

- Die Frameworks aus HDFS, NoSQL, Stream/Batch Processing
  - sind komplex zu erlernen
  - sind u.U. komplex zu betreiben
  - jedes einzelne und es sind viele
- Technologiewelt wird volatiler
- Big-Data-Lösungen hängen eng mit der Problemstellung zusammen
  - Die speziellen Anforderungen bestimmen die Lösung
- Big-Data-Lösungen haben Einfluss auf die zu wählende Hardware
  - Stichwort: *commodity hardware*
- Big-Data-Technologien sind häufig dissruptiv
  - Passen i.d.R. nicht zu bestehenden Architekturen
  - Passen i.d.R. nicht zu bestehenden Systemumgebungen
  - Ersetzen i.d.R. bestehende Technologien

Fragestellungen an Big-Data-Lösungen

- Welche *Big-Data*-Lösung ist die richtige Lösung ?
  - Welche Geschäftsanforderungen werden mit Big Data adressiert?
- Welche Datenquellen werden importiert?
  - Warum werden diese Daten gesammelt?
  - Welche Schlüsse können aus diesen Daten gezogen werden?
  - Haben diese Datenquellen eine kritisches Volumen?
- Welche neuen Anforderungen an Kompetenzen der Mitarbeiter stellt Big Data?
  - Neue technologische Kompetenzen
  - Neue Berufsbilder wie *Data Scientist*
  - Woher kommen die Kompetenzen der Mitarbeiter?

Wie fangen Sie an?

Vielen Dank!

Fragen?

Referenzen

- Big Data - Entwicklung und Programmierung von Systemen für große Datenmengen und Einsatz der Lambda-Architektur by Nathan Marz (2016)

- http://www.harvardbusinessmanager.de/blogs/a-862657.html
- http://www.harvardbusinessmanager.de/blogs/a-862658-2.html
- http://www.harvardbusinessmanager.de/blogs/a-861011.html

Impulsvorträge für Ihr Unternehmen

IKS
Individuelle Softwarelösungen

- Überblick über das gesamte Angebot an Impulsvorträgen unter: www.iks-gmbh.com/impulsvortraege

- Ihr Nutzen:
  - Unabhängiges, aktuelles Expertenwissen.
  - Individuell auf Ihr Publikum und Ihr Unternehmen zugeschnittene Vorträge.
  - Referenten mit langjähriger und branchenübergreifender Expertise in der IT-Beratung.
  - Praxisnahe Vorträge, die aus Projektarbeit entstanden sind, frei von Produktwerbung.
  - Ideale Ergänzung für Ihre Führungskräftetreffen, Abteilungsmeetings, Hausmessen, Innovation Days, Konferenzen, Open Spaces, Kick-off-Meetings oder Zukunftsworkshops.

IKS
Individuelle Softwarelösungen

WWW.IKS-GMBH.COM

- https://stock.adobe.com/de/search?k=m%C3%BChlenrad&asset_id=50391961

https://pixabay.com/illustrations/graphic-design-wave-element-curves-2822614/