Big Data mit Open Source
Marcel Bernet

Mittwoch: 4. Mai 2017

CH Open Business Lunch

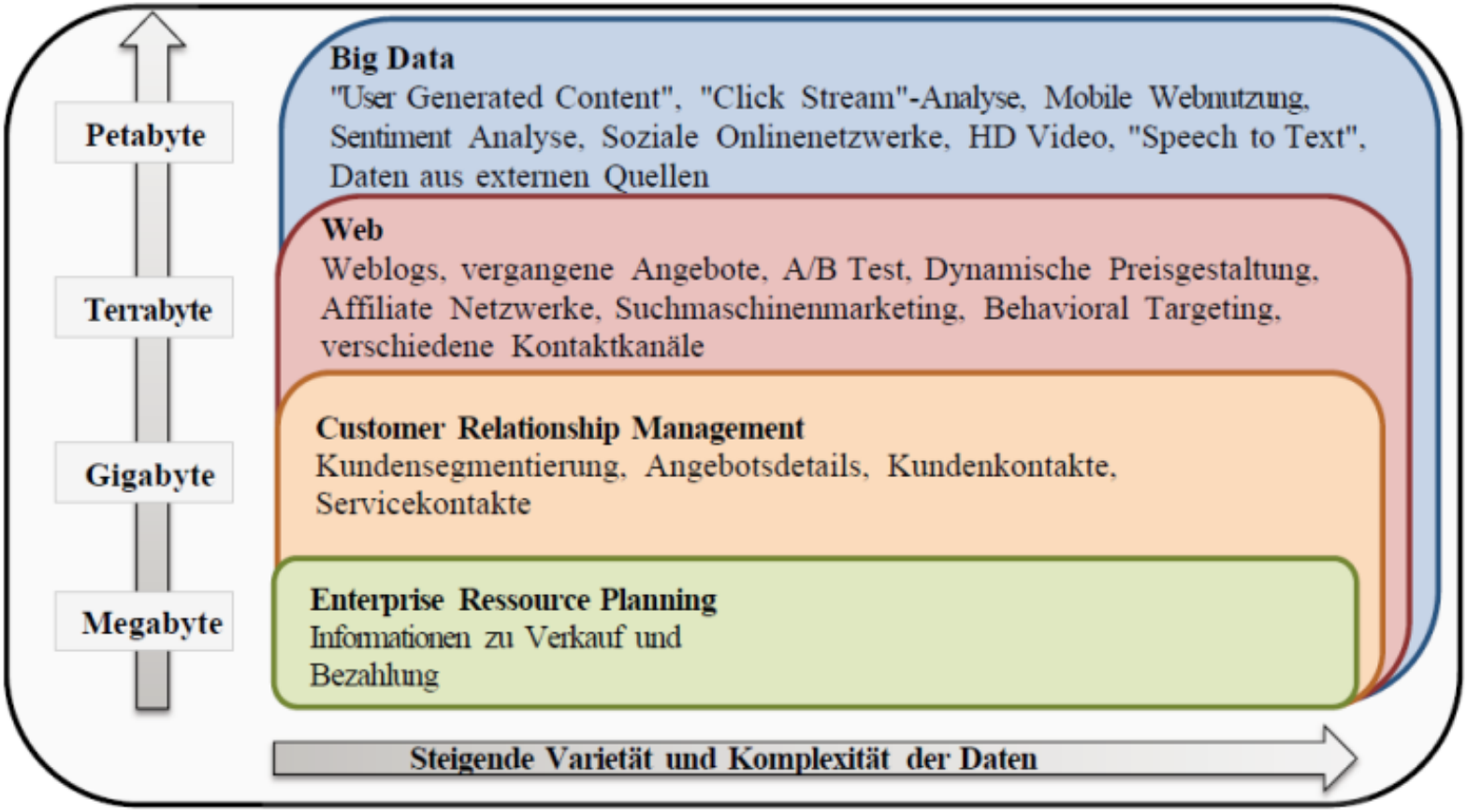
Restaurant Mère Catherine,
Nägelihof 3, 8001 Zürich

Über Marcel Bernet:

- Nach langen Jahren als Technologie-Vordenker für diverse Firmen und der öffentlichen Verwaltung sowie Mitglied in diversen Expertenkommissionen widmet sich Marcel Bernet heute hauptsächlich der Beratung und Schulung. Als ehemaliger CH open-Präsident und heutiges Ehrenmitglied entwickelt und fördert er Open Source und setzt diese in seinen Projekten ein.
  Im Rahmen des Veranstaltungsreihe «Digitalisierung und Gesellschaft» und dem verbunden Kursangebot hat er sich mit dem Thema Big Data befasst und eine Open Source Big Data Umgebung entworfen.
- Kurse:
  - Internet der Dinge – Grundlagen
  - Internet der Dinge – Aufbau 1 – Komplexe Anwendungen und die Cloud
  - Internet der Dinge – Aufbau 2 – Raspberry Pi und Co. als Server
  - Big Data – Überblick
  - Digitale Transformation
  - Infrastructure as Code
  - Docker

Big Data: Datenmenge

1 Petabyte PB = 1’000’000’000’000’000 Bytes (1’000 TB)

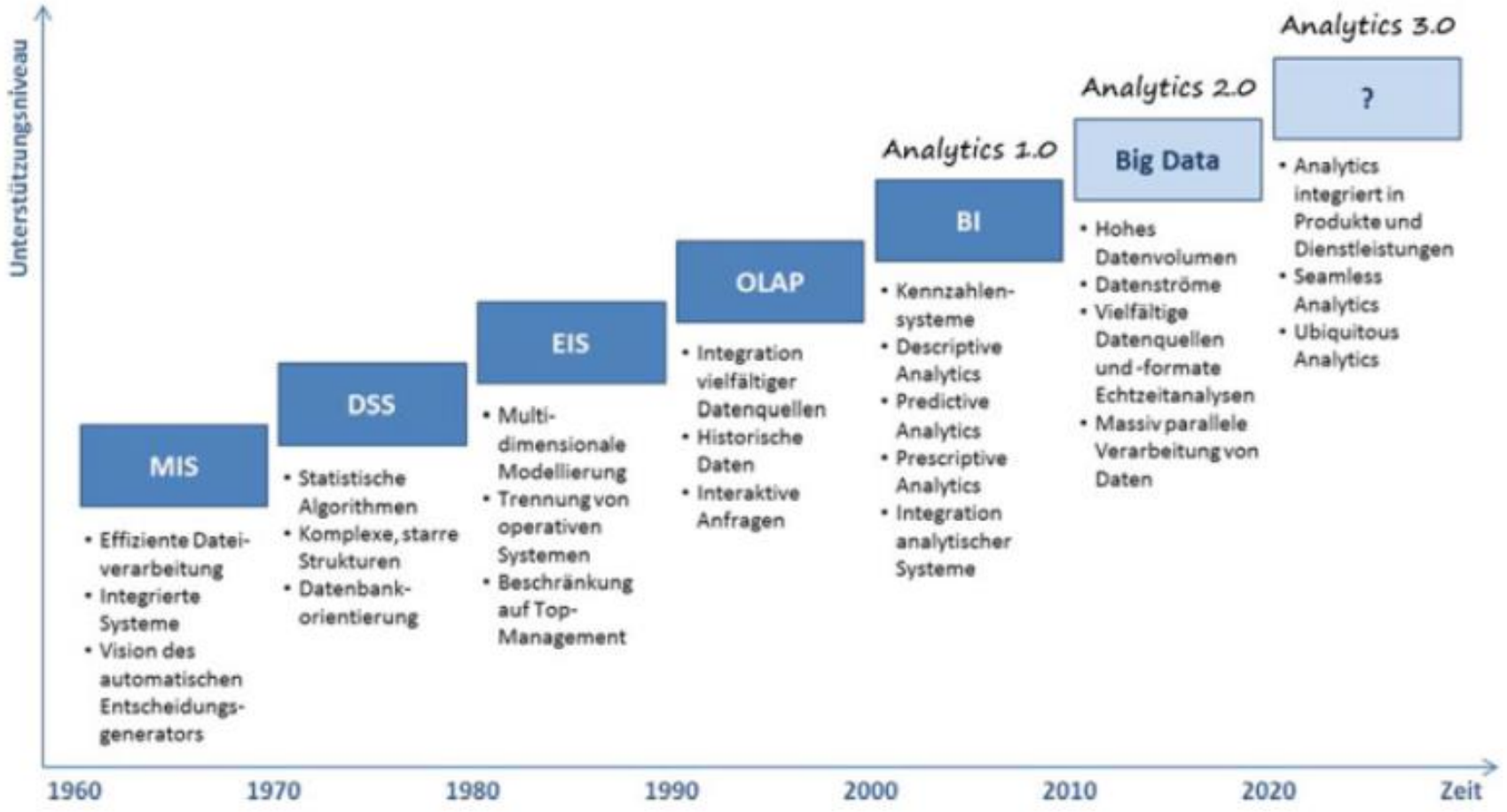


Quelle: Big Data, Potential und Barrieren der Nutzung im Unternehmenskontext

Big Data: Charakteristika

- **Umfang** („Volume“): eine grosse Menge an Daten, die aufgenommen, analysiert und gemanagt werden muss. Der Datenumfang steigt mit der Anzahl der Quellen und der höheren Auflösung bzw. Datentiefe.
- **Varietät** („Variety“): Daten stammen zunehmend aus neuen Quellen innerhalb und ausserhalb der Organisation, ihre Struktur variiert stark, es können auch bisher unbekannte Strukturierungsformen (z.B. Open311, KMZ) auftreten.
- **Schnelllebigkeit** („Velocity“): die Geschwindigkeit, mit der Daten produziert und verändert werden müssen. Dies verlangt eine rasche Analyse und Entscheidungsfindung. Die Schnelllebigkeit wird von der Anzahl der Quellen und der gesteigerten Rechenleistung der datengenerierenden Geräte beeinflusst.
- **Richtigkeit** („Veracity“): die Qualität und Quelle der rezipierten Daten. Erstere wird unter anderem von Inkonsistenz, Unvollständigkeit und Mehrdeutigkeit beeinflusst. Das Fällen von datenbasierten Entscheidungen verlangt Nachvollziehbarkeit und Begründbarkeit

Big Data: Stammbaum



BI-Stammbaum, Quelle: Praxishandbuch Big Data, Wirtschaft – Recht – Technik

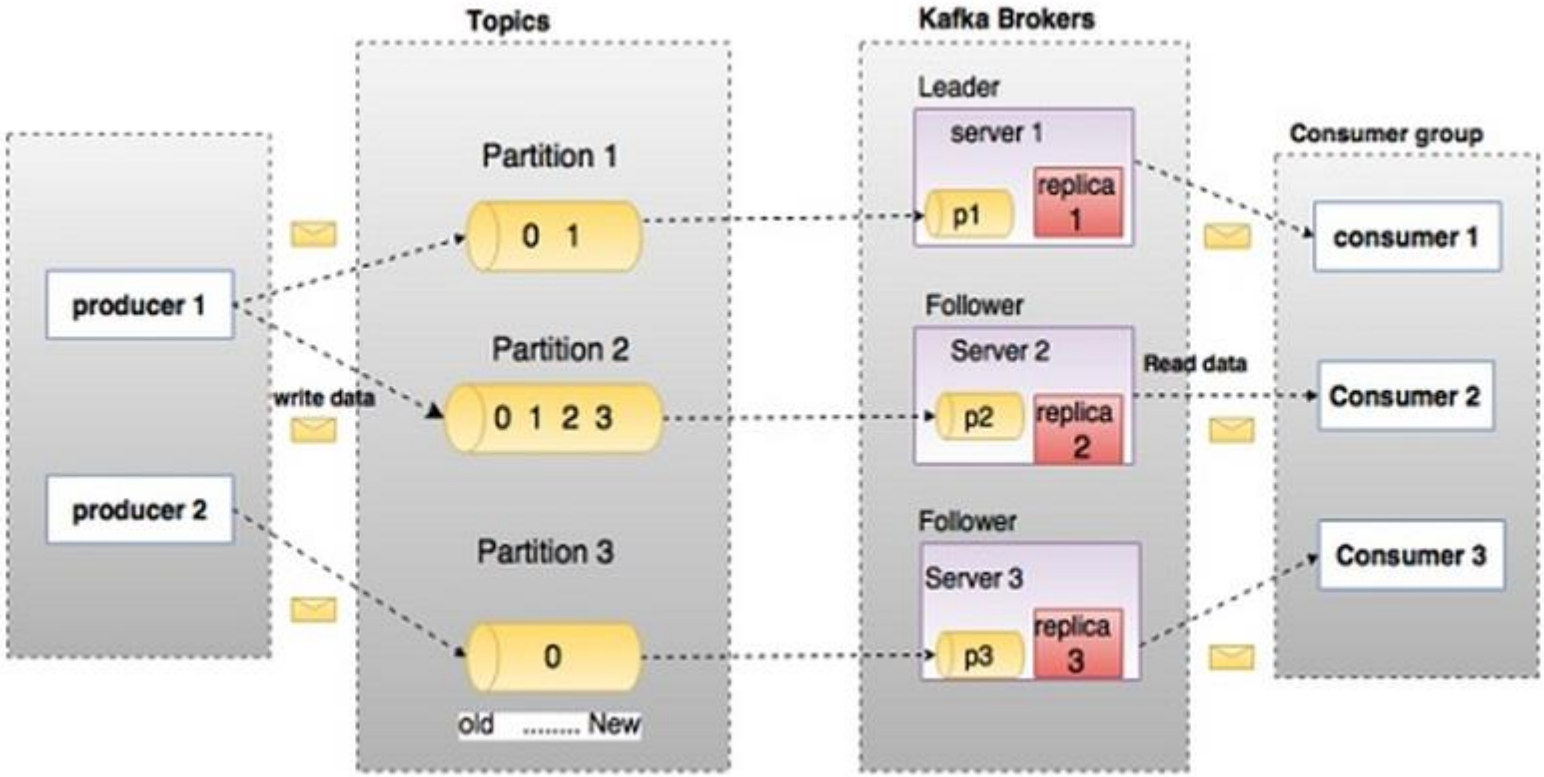
Datenquellen: Open Data

www.opendata.ch

Datenquellen: Links

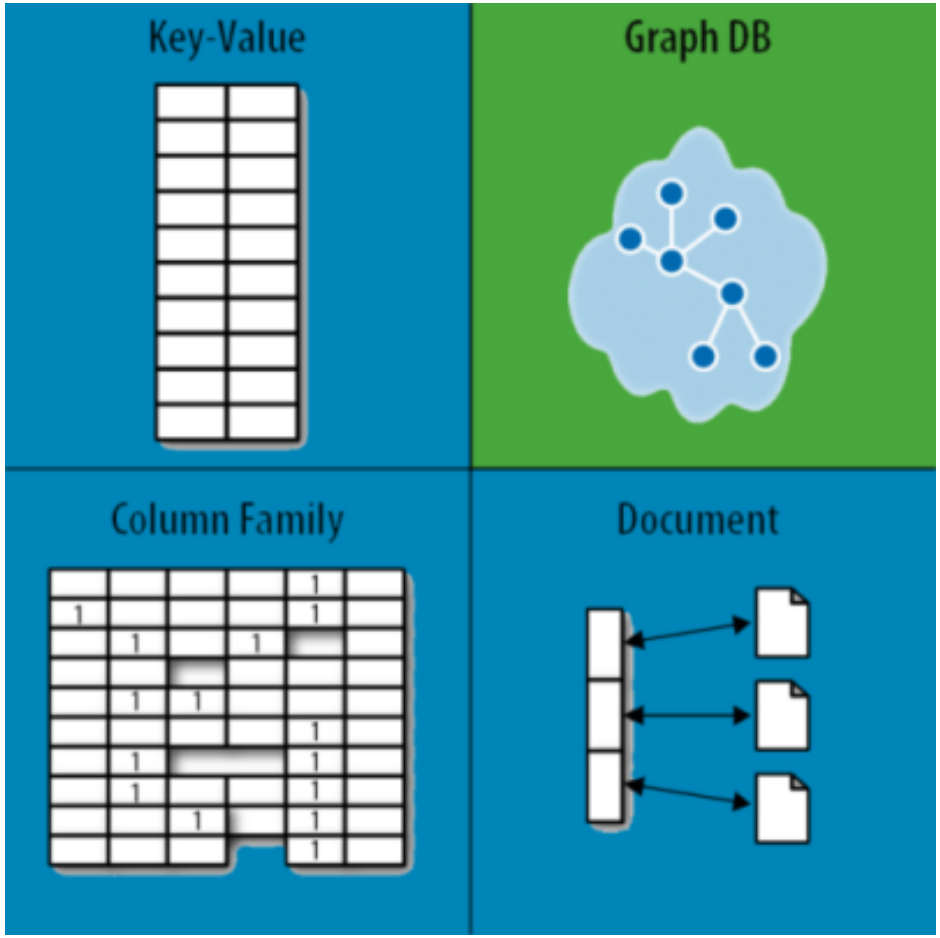
- Stadt Zürich
- Bundesamt für Statistik
- opendata.swiss
- Swiss public transport API
- Programmable Web

Sammeln: Streaming via Topics



Quelle: https://www.tutorialspoint.com/apache_kafka/apache_kafka_quick_guide.htm

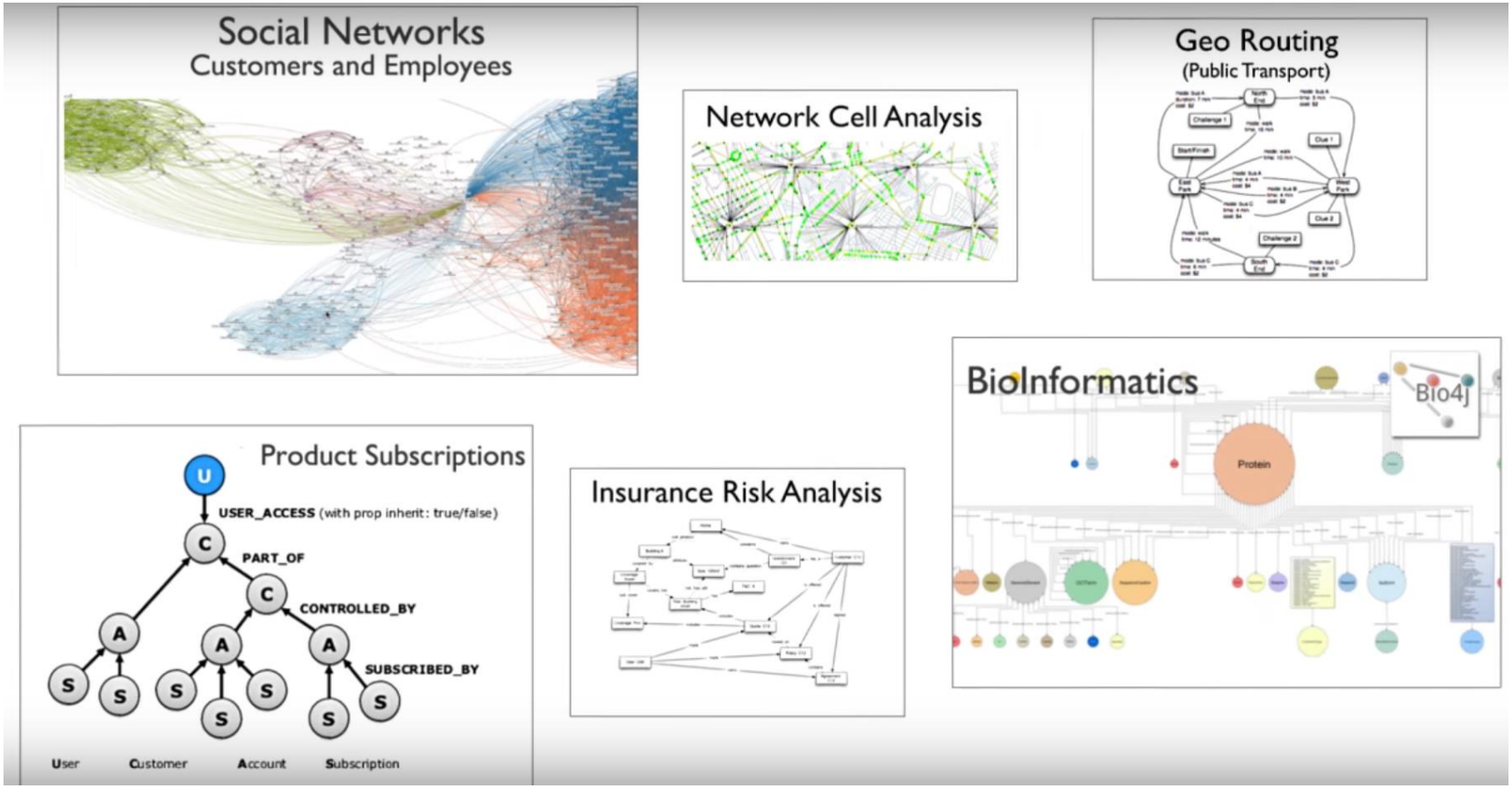
Speichern: NoSQL – Not only SQL



Quelle: Neo4j Blog

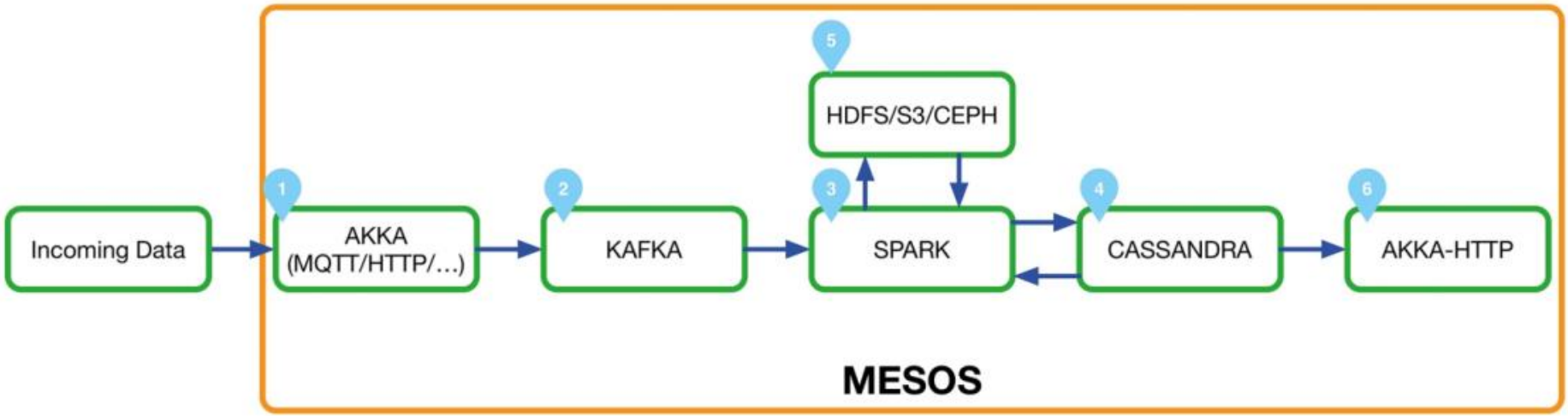
NoSQL: Document Stores

NoSQL: Graph Databases



Quelle: https://neo4j.com/graphacademy/online-training/introduction-graph-databases/

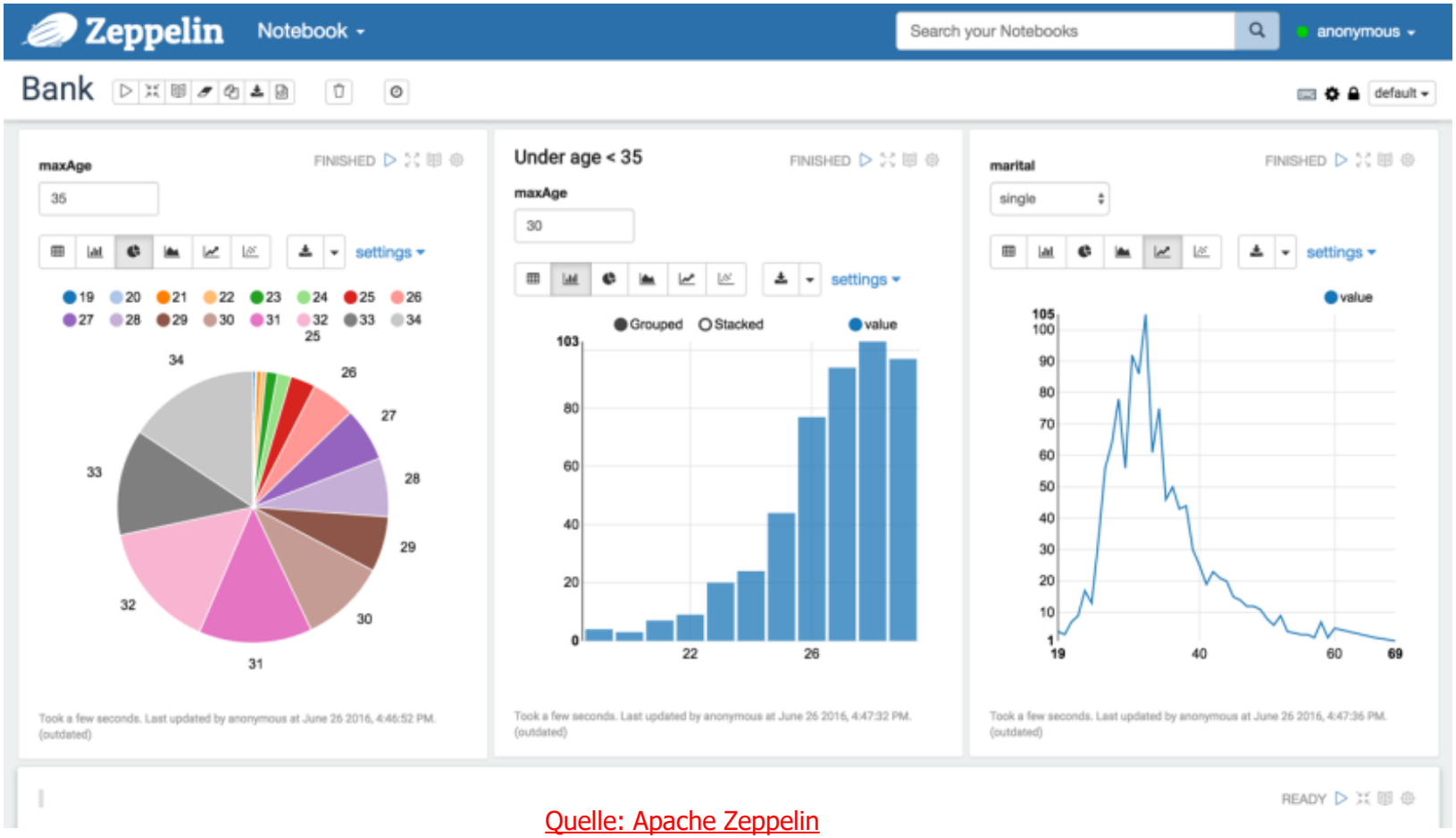
Framework: SMACK



Quelle: https://www.codecentric.de/leistungen/loesungen/technologie-treibt-veraenderung/big-data/smack/

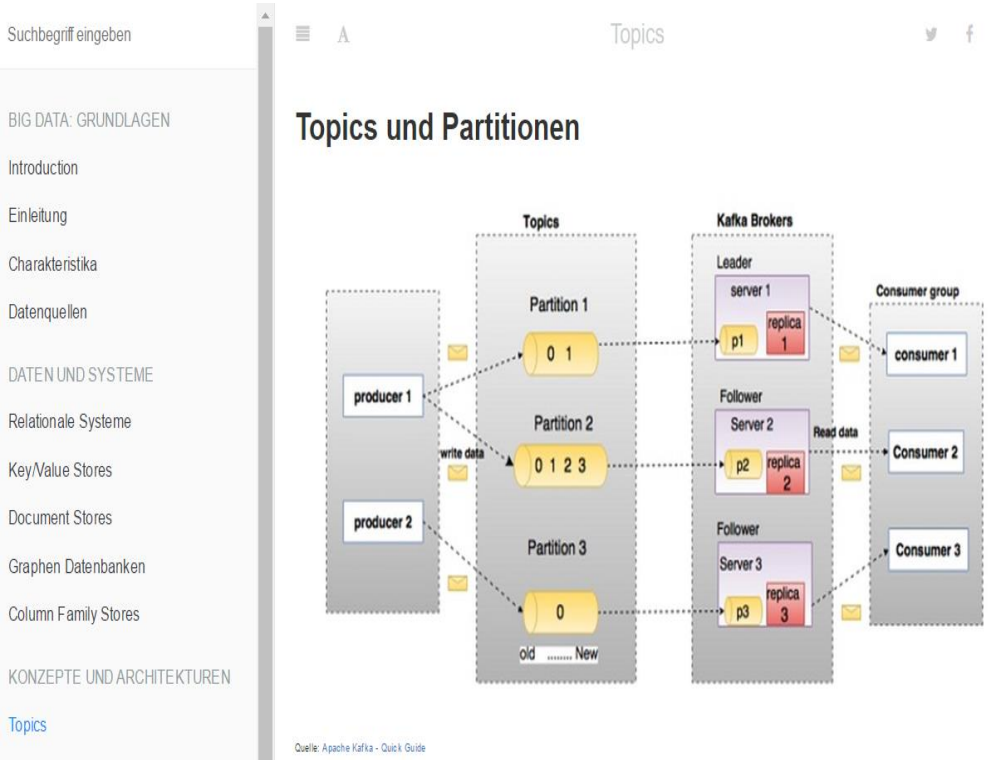
- Akka – Daten entgegennehmen
- Kafka – Message Broker (verteilen)
- Spark – Verarbeitung
- Cassandra – Speicherung
- Akka – Daten zur Verfügung stellen

Auswerten: Apache Zeppelin



Quelle: Apache Zeppelin

Aufbereitet als VM/Docker Umgebung



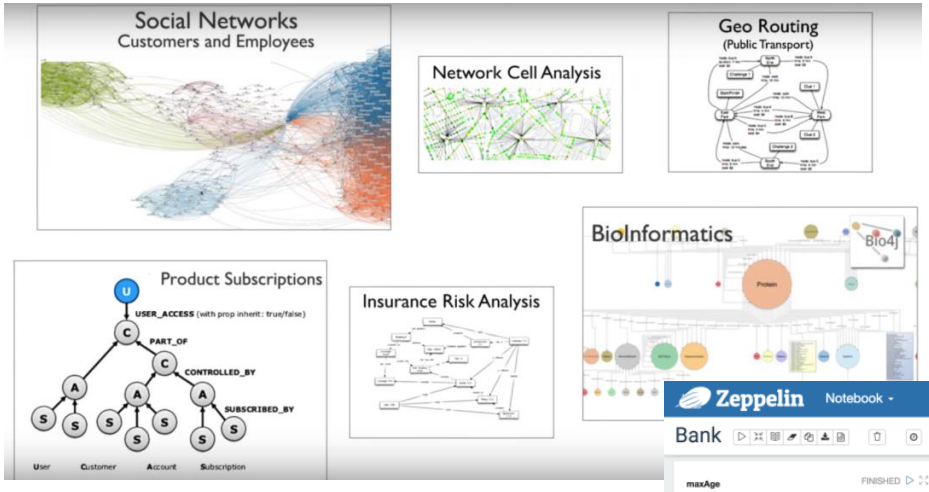

```
Vagrant.configure(2) do |config|
 # Docker Provisioner
  config.vm.provision "docker" do |d|
    d.build_image "/vagrant/mysql", args: "--tag mysql"
    d.build_image "/vagrant/redis", args: "--tag redis"
    d.build_image "/vagrant/mongodb", args: "--tag mongodb"
    d.build_image "/vagrant/spark", args: "--tag spark"
    d.pull_images "neo4j:3.0"
    d.pull_images "aquila/zeppelin"
    d.pull_images "cassandra"
    d.pull_images "jplock/zookeeper"
    d.pull_images "ches/kafka"
    d.build_image "/vagrant/gitbook", args: "--tag gitbook"
    d.run "gitbook", image: "gitbook", args: "-p 4000:4000 -v
/vagrant:/srv/gitbook"
    d.run "zeppelin01", image: "aquila/zeppelin", args: "-p 8080:
    d.run "neo4j01", image: "neo4j:3.0", args: "-p 7474:7474"
  end

end
```

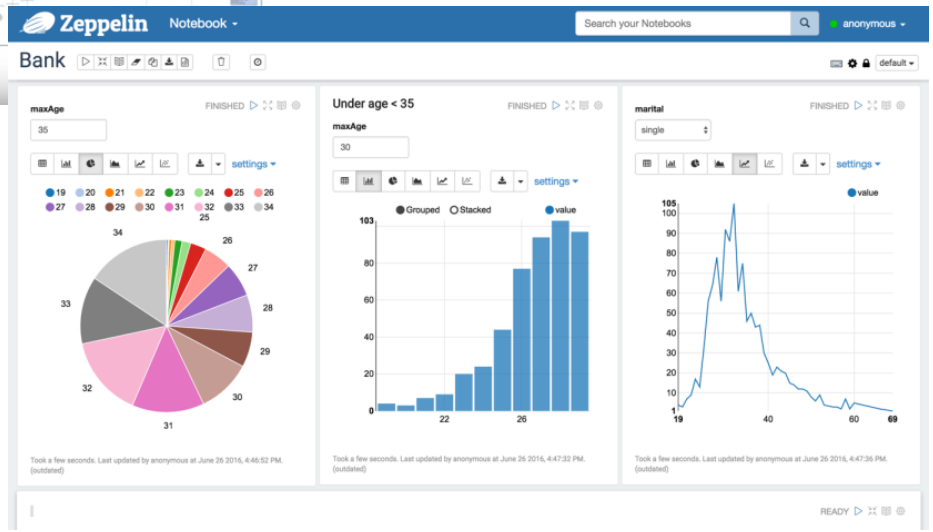
- https://github.com/mc-b/bigdata
- http://iotkit.mc-b.ch/2016-04-06-OBL-IAS/

Open Source Software

- Relationale Datenbanken
  - MySQL, MariaDB, PostgreSQL
- NoSQL
  - Redis, MongoDB, CouchDB, Neo4J, Cassandra, Hbase
- KI, Streaming
  - Spark, Akka, Kafka
- UI
  - Apache Zeppelin
- Programmiersprachen
  - R, Scala
- Datacenter
  - DC/OS, Apache Mesos

Demo


Social Networks
Customers and Employees
Network Cell Analysis
Geo Routing
(Public Transport)
BioInformatics
Bio4j
Protein
Product Subscriptions
USER_ACCESS (with prop inherit: true/false)
PART_OF
CONTROLLED_BY
SUBSCRIBED_BY
User
Customer
Account
Subscription
Insurance Risk Analysis

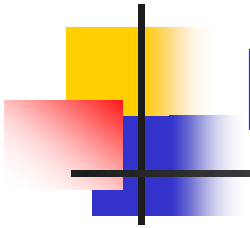

Zeppelin
Notebook
Search your Notebooks
anonymous
Bank
maxAge
35
FINISHED
settings
Under age < 35
maxAge
30
Grouped
Stacked
value
marital
single
READY

Zusammenfassung

- Big Data sind Datenmengen, die **zu gross**, **zu komplex**, **zu schnelllebig** oder zu **schwach strukturiert** sind, um sie mit manuellen und herkömmlichen Methoden der Datenverarbeitung auszuwerten.

- Zur Speicherung und Auswertung werden deshalb neue Tools wie NoSQL Datenspeicher und neue Abfragesprachen wie Scala verwendet.

Fragen ?



Kontakt

Marcel Bernet
Mail: marcel.bernet@ch-open.ch

Big Data Umgebung

- https://github.com/mc-b/bigdata