



Big Data mit Open Source

Marcel Bernet

Mittwoch: 4. Mai 2017

CH Open Business Lunch
Restaurant Mère Catherine,
Nägelihof 3, 8001 Zürich

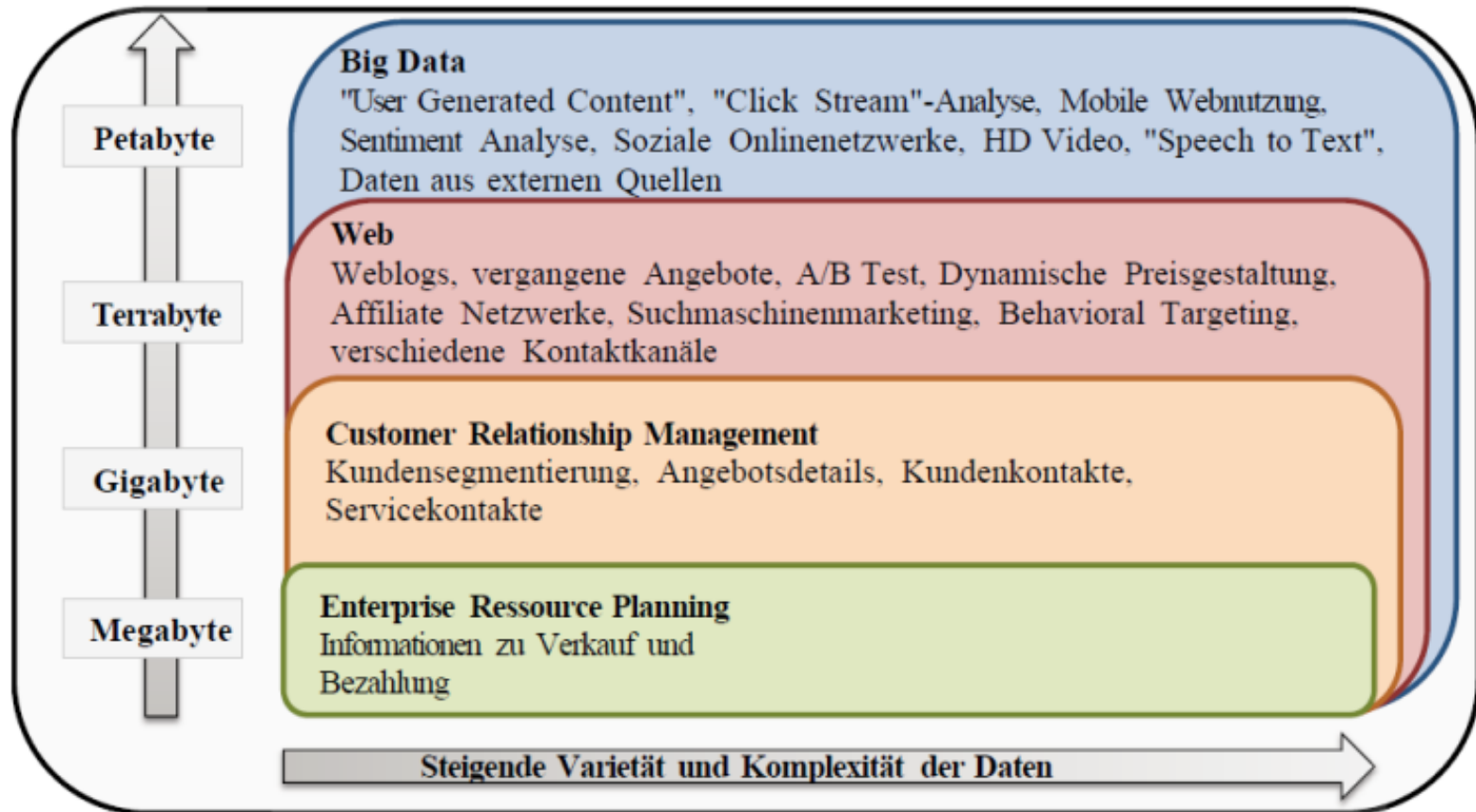


Über Marcel Bernet:

- 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 der Veranstaltungsreihe «Digitalisierung und Gesellschaft» und dem verbundenen 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)



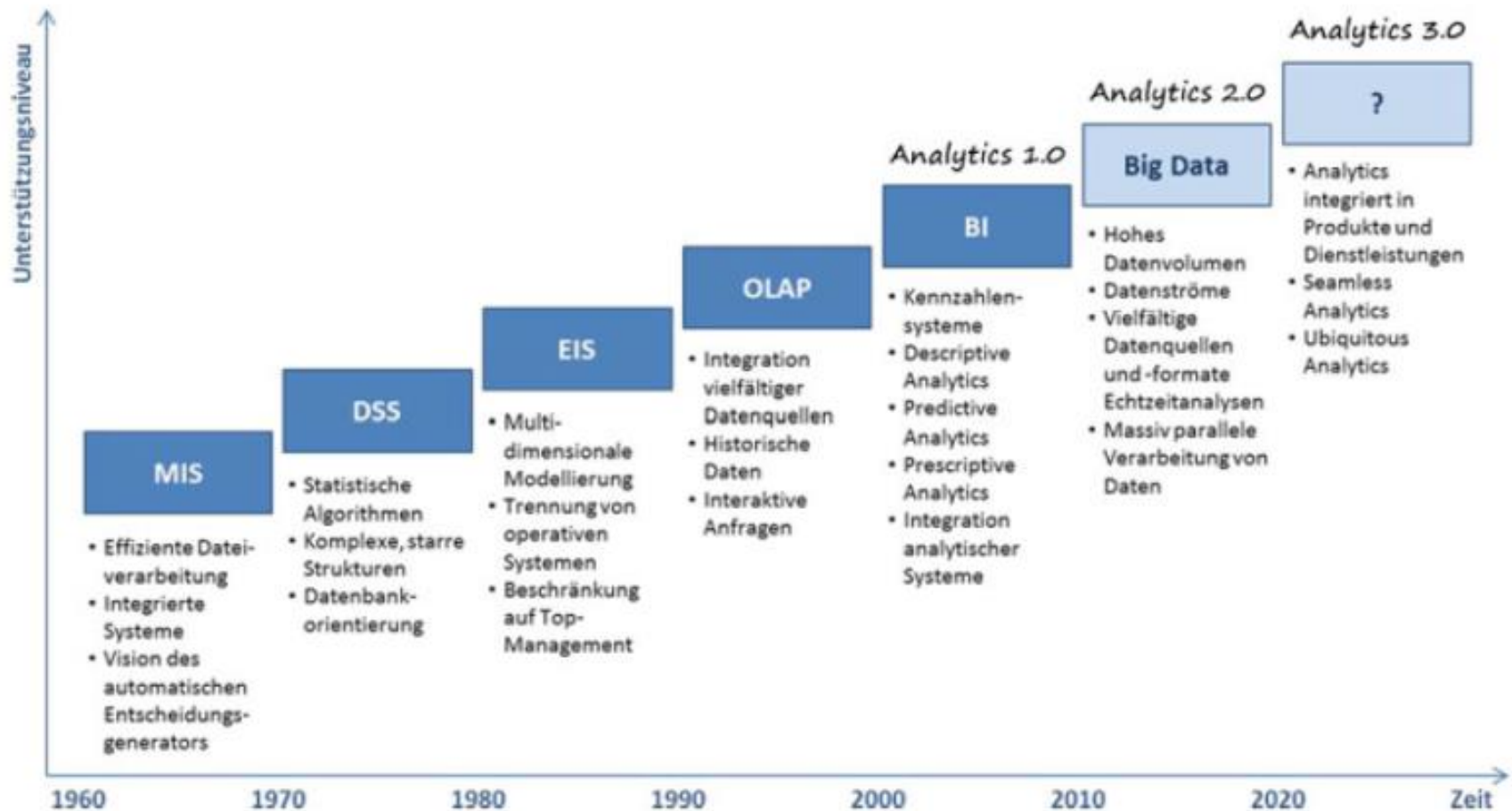
Quelle: Big Data, Potential und Barrieren der Nutzung im Unternehmenskontext



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 Füllen von datenbasierten Entscheidungen verlangt Nachvollziehbarkeit und Begründbarkeit

Big Data: Stammbaum



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

Datenquellen: Open Data



News

Manifest

Anlässe

Organisation

Vorstand

Mitglied werden

Kontakt

FR / DE

Suche



Anlässe

Der Verein Opendata.ch organisiert verschiedene Veranstaltungen. Regelmässige Veranstaltungen sind die [make.opendata.ch](#) Camps sowie die jährlichen Opendata.ch Konferenzen.

Open Cultural Data Hackathon
2017



Opendata.ch/2017



Mitgliederversammlung 2016



Opendata.ch/2016



April 2016: Energy Hackdays



Election Hackdays



Mitgliederversammlung 2015

Opendata.ch/2015 Konferenz

Mitgliederversammlung 2014

www.opendata.ch



Datenquellen: Links

- [Stadt Zürich](#)
- [Bundesamt für Statistik](#)
- [opendata.swiss](#)
- [Swiss public transport API](#)
- [Programmable Web](#)



The screenshot shows the homepage of the 'Stadt Zürich Open Data' portal. The header is dark blue with navigation links: 'Open Data Portal', 'Anwendungen', 'Werkstatt', 'Blog' (highlighted), and social media icons for Twitter and email. Below the header, the 'Stadt Zürich Open Data' logo is on the left, and navigation buttons for 'Startseite', 'Datensätze', and 'Kategorien' are in the center. A search bar with the placeholder 'Suche' and a magnifying glass icon is on the right. The main content area has a white background with the heading 'Willkommen auf dem Open Data Katalog'. Below this, a paragraph states: 'Der Datenkatalog ist Ihr zentraler Einstiegspunkt zur Suche und Nutzur stehen kostenlos und zur freien – auch kommerziellen - Weiterverwend'. To the right of this paragraph, the title 'Fahrzeiten der VBZ im SOLL-IST-Vergleich' is displayed.

Open Data Portal Anwendungen Werkstatt **Blog**  

 **Stadt Zürich**
Open Data

Startseite Datensätze Kategorien

Suche 

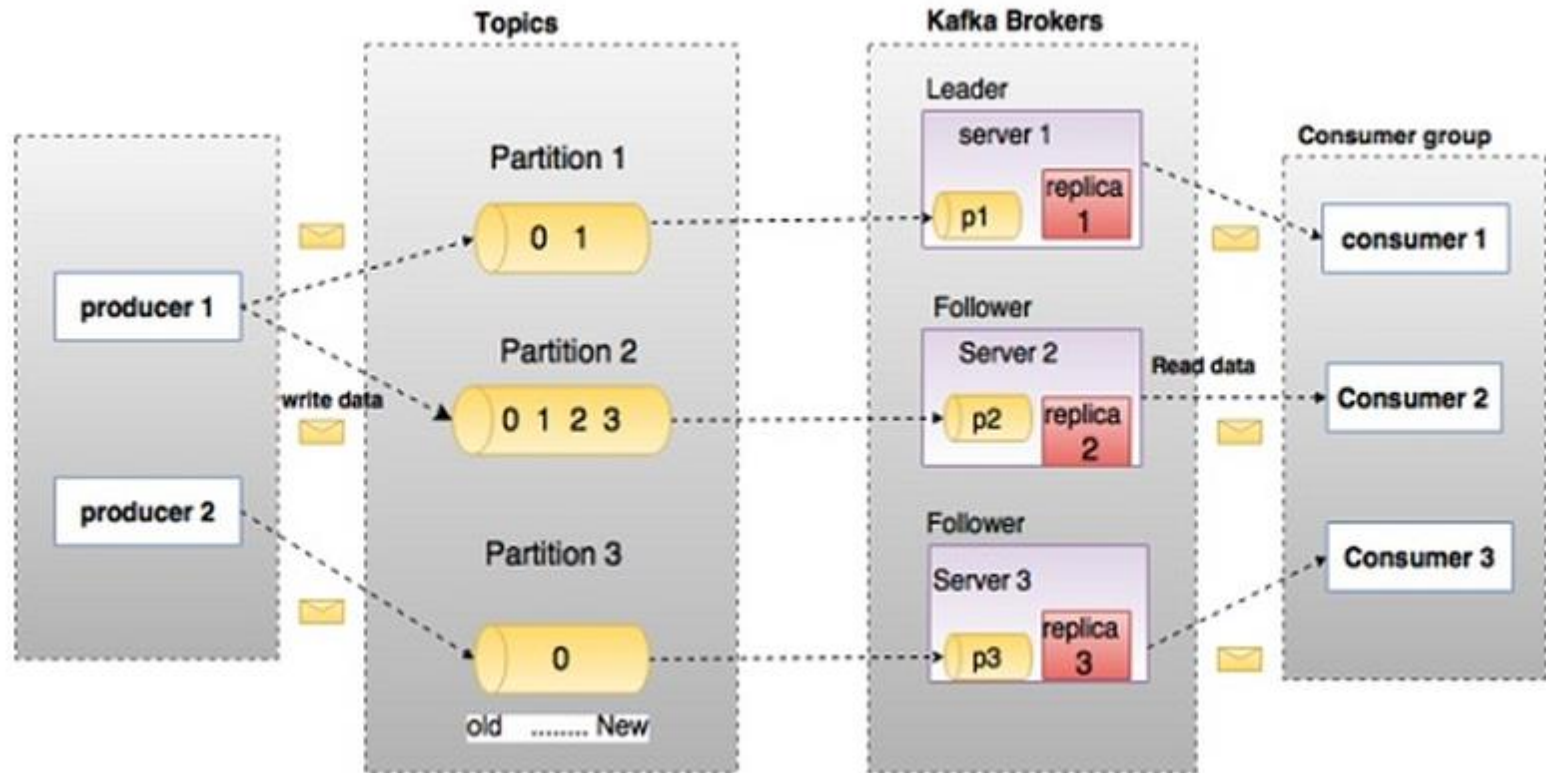
Willkommen auf dem Open Data Katalog

Der Datenkatalog ist Ihr zentraler Einstiegspunkt zur Suche und Nutzur stehen kostenlos und zur freien – auch kommerziellen - Weiterverwend

Fahrzeiten der VBZ im SOLL-IST-Vergleich

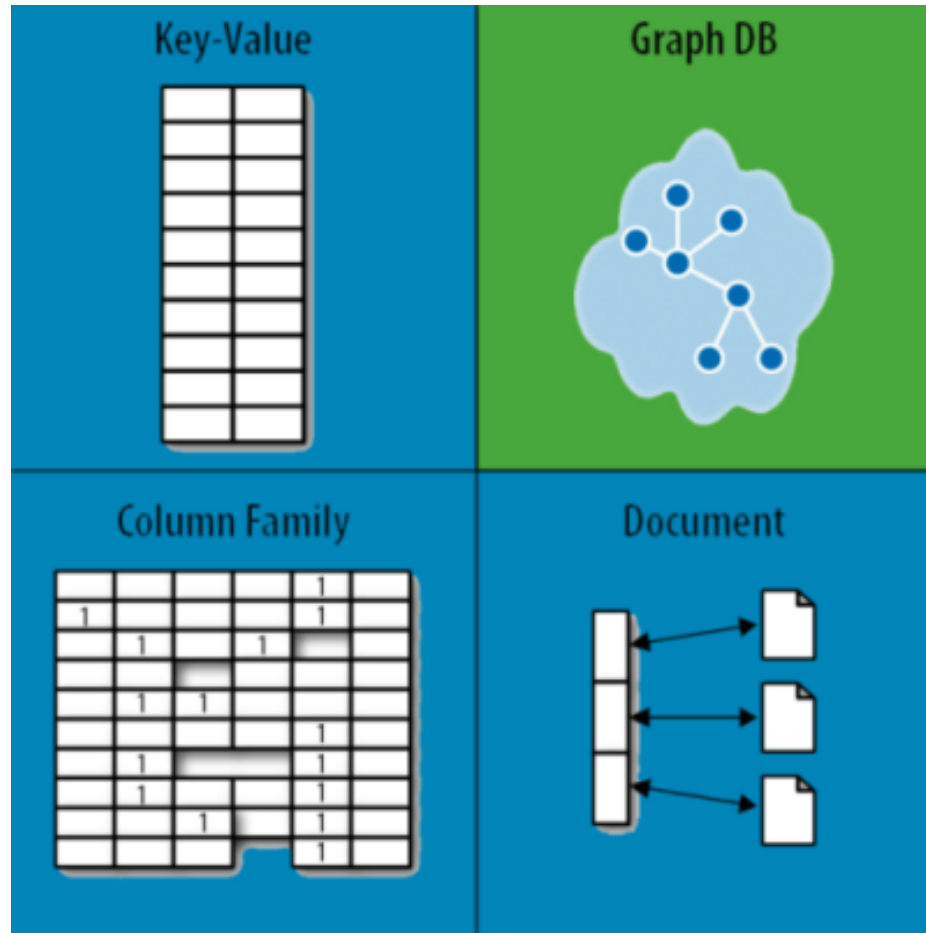
Diese Daten ermöglichen einen sehr genauen Vergleich zwischen den geplanten (SOLL) und den effektiv erfolgten (IST) Fahrzeiten jedes Fahrzeugs der Verkehrsbetriebe Zürich (VBZ). Die Haltestellenankünfte und -abfahrten jedes Fahrzeugs auf seiner Linie wird sekundengenau gemessen.

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

```
{
  _id: <ObjectId>,
  username: "123xyz",
  contact: {
    phone: "123-456-7890",
    email: "xyz@example.com"
  },
  access: {
    level: 5,
    group: "dev"
  }
}
```



Embedded sub-document



Embedded sub-document

NoSQL: Graph Databases

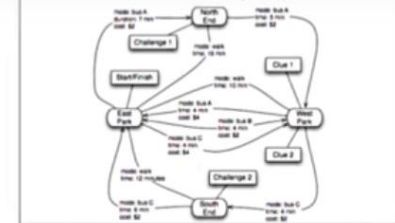
Social Networks
Customers and Employees



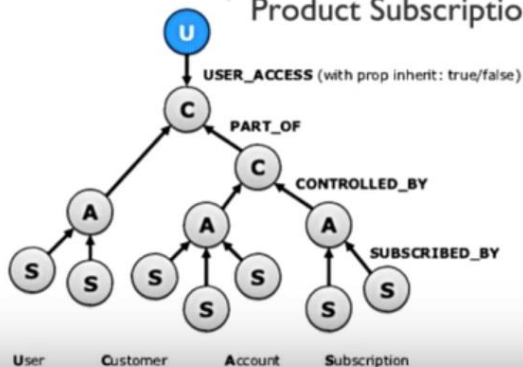
Network Cell Analysis



Geo Routing
(Public Transport)



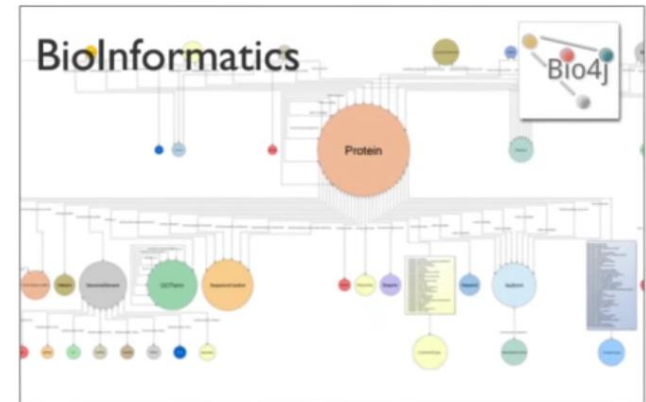
Product Subscriptions



Insurance Risk Analysis

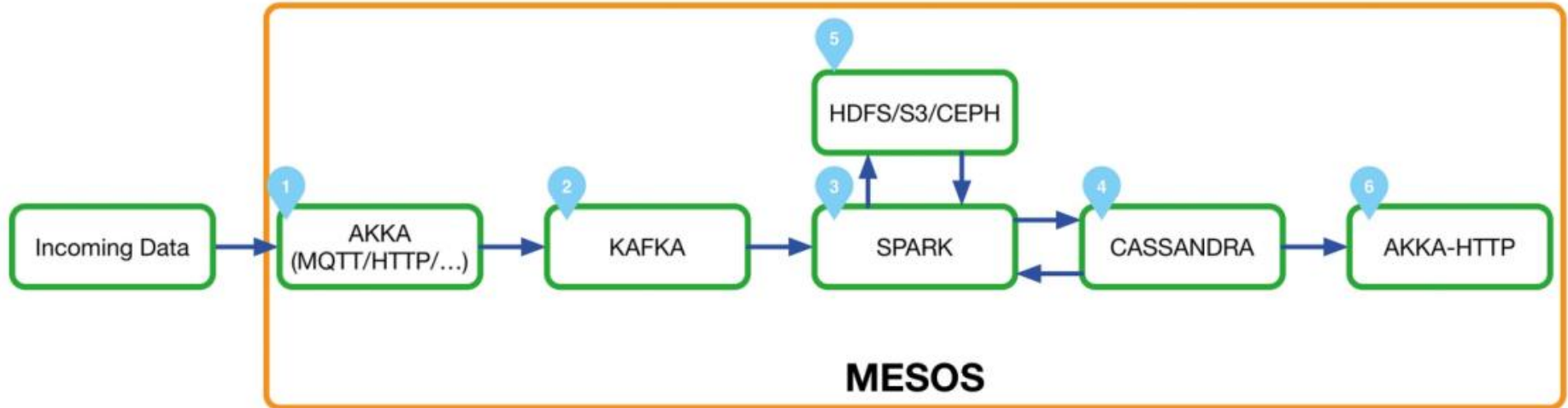


BioInformatics



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

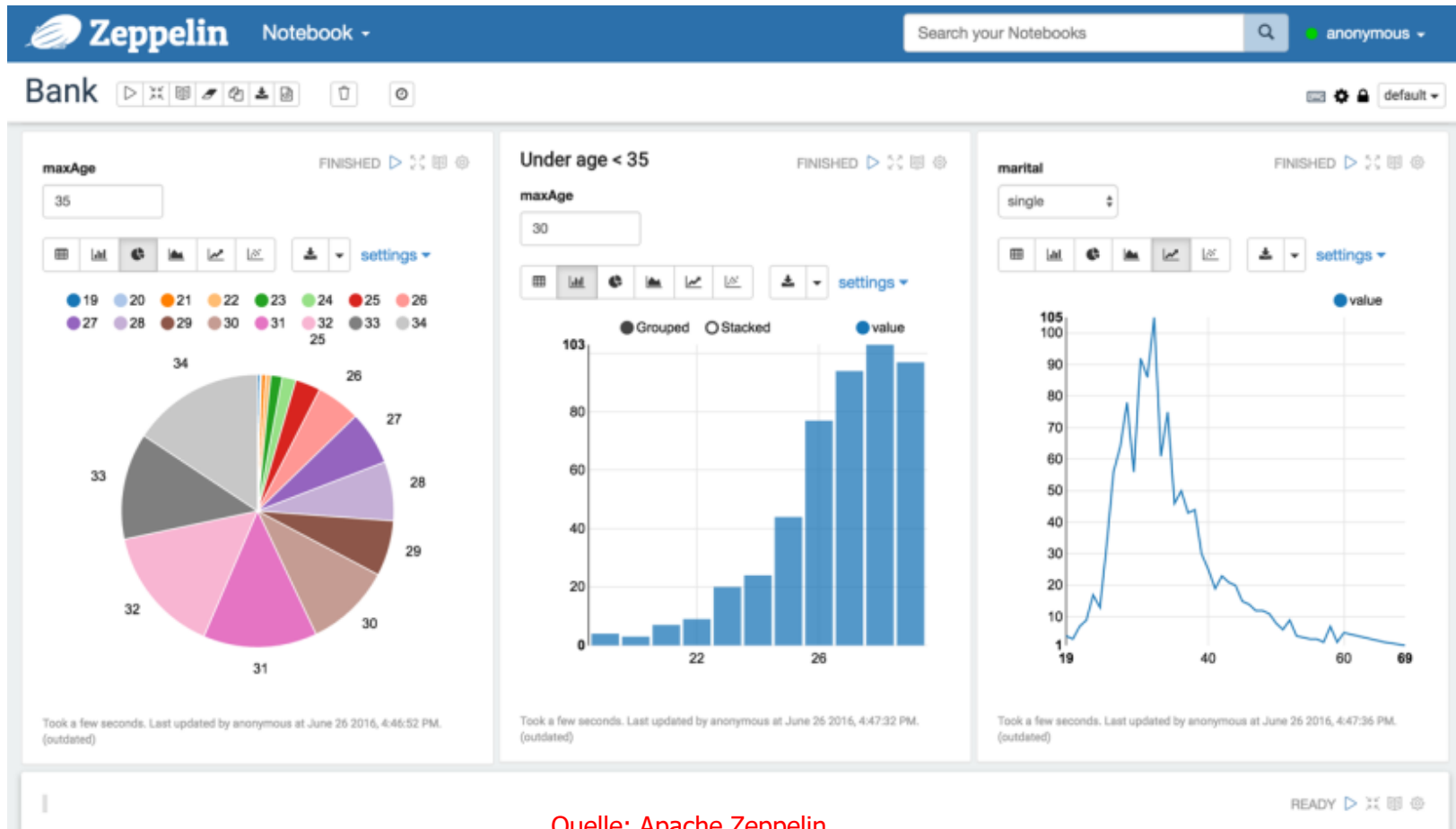
Framework: SMACK



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

Suchbegriff eingeben

BIG DATA: GRUNDLAGEN

Introduction

Einleitung

Charakteristika

Datenquellen

DATEN UND SYSTEME

Relationale Systeme

Key/Value Stores

Document Stores

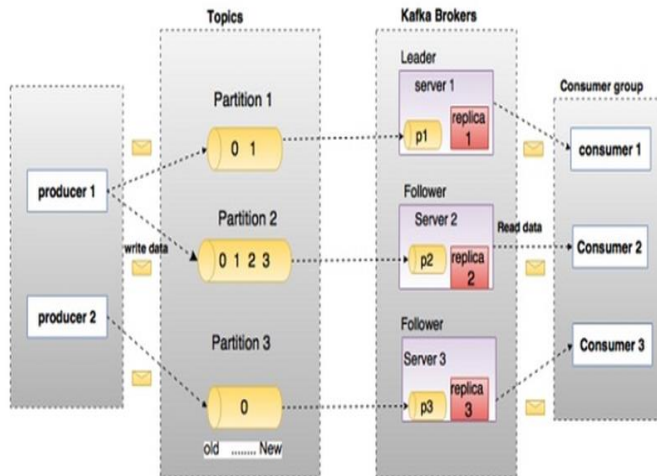
Graphen Datenbanken

Column Family Stores

KONZEPTE UND ARCHITEKTUREN

Topics

Topics und Partitionen



Quelle: Apache Kafka - Quick Guide

```
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 "jplack/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
```

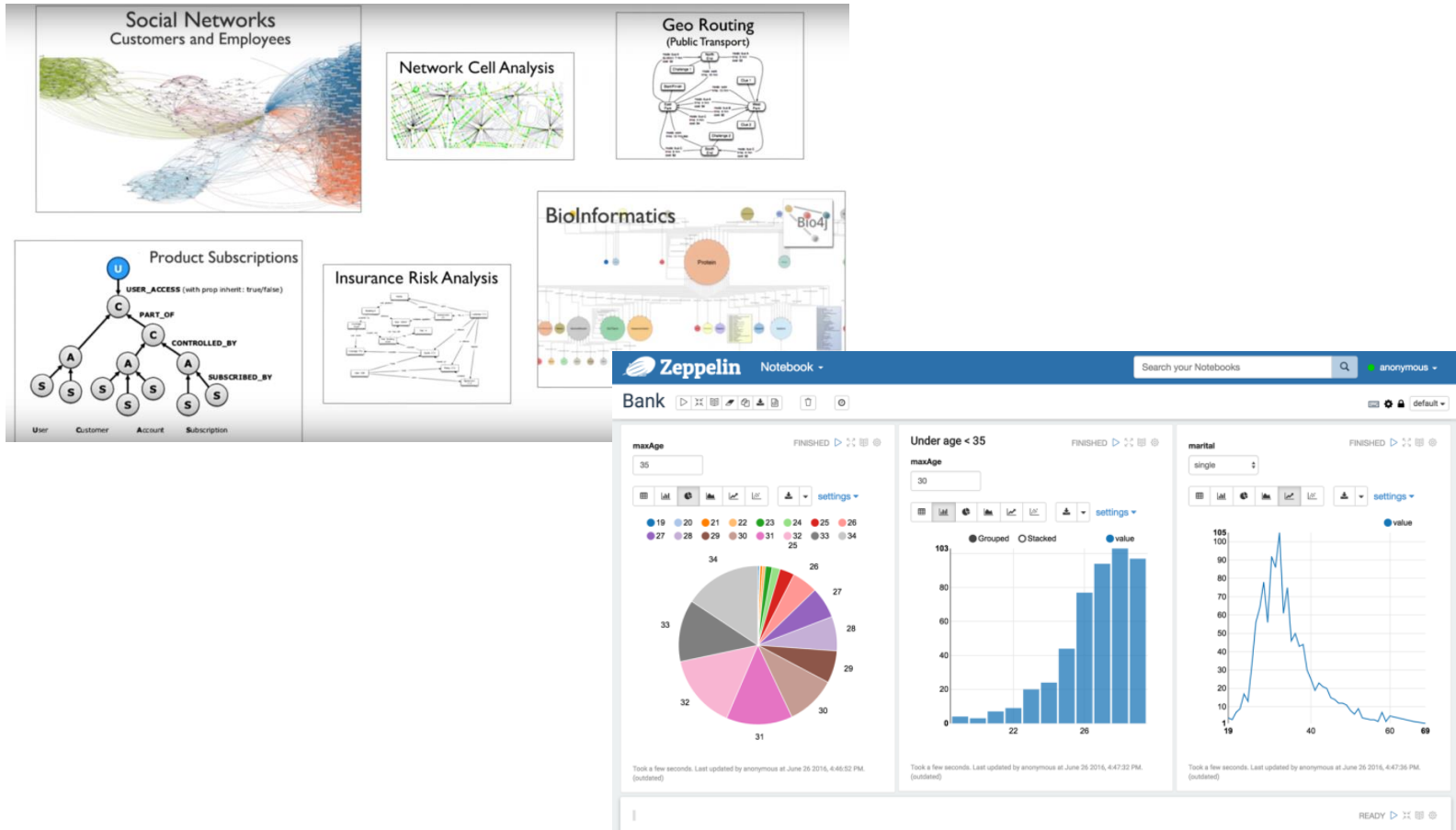
- <https://github.com/mc-b/bigdata>
- <http://iotkit.mc-b.ch/2016-04-06-OBL-IAS/>



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





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>