Benchmarking von Big Data Anwendungen in einer skalierbaren Infrastruktur

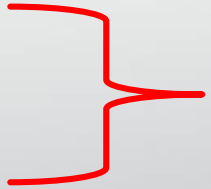
Hauptseminar-Vortrag von Tim Horgas

Gliederung

- Einleitung und Motivation
  - Bezug zu bisherigen Arbeiten
  - Erkenntnisse
- Big Data Architekturen
  - Anwendungsfälle/Beispiele
- Ideen für Umsetzung
  - Konzepte Clustermanager
  - Benchmarking der Implementation
  - Vorläufige Vision Masterarbeit

Gliederung

- Einleitung und Motivation
  - Bezug zu bisherigen Arbeiten
  - Erkenntnisse
- Big Data Architekturen
  - Anwendungsfälle/Beispiele
- Ideen für Umsetzung
  - Konzepte Clustermanager
  - Benchmarking der Implementation
  - Vorläufige Vision Masterarbeit

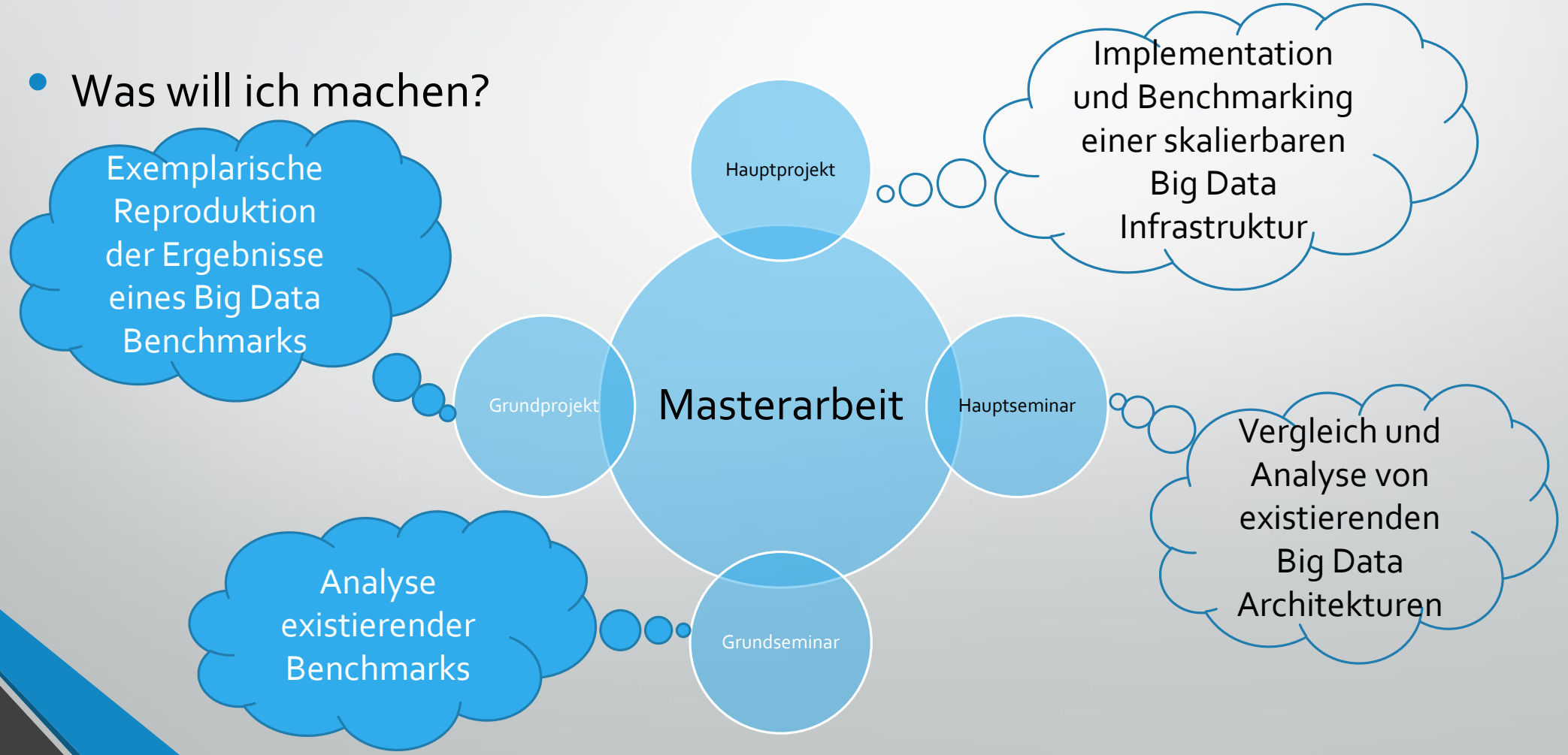


Risiko-Analyse

Einleitung und Motivation

Einleitung und Motivation

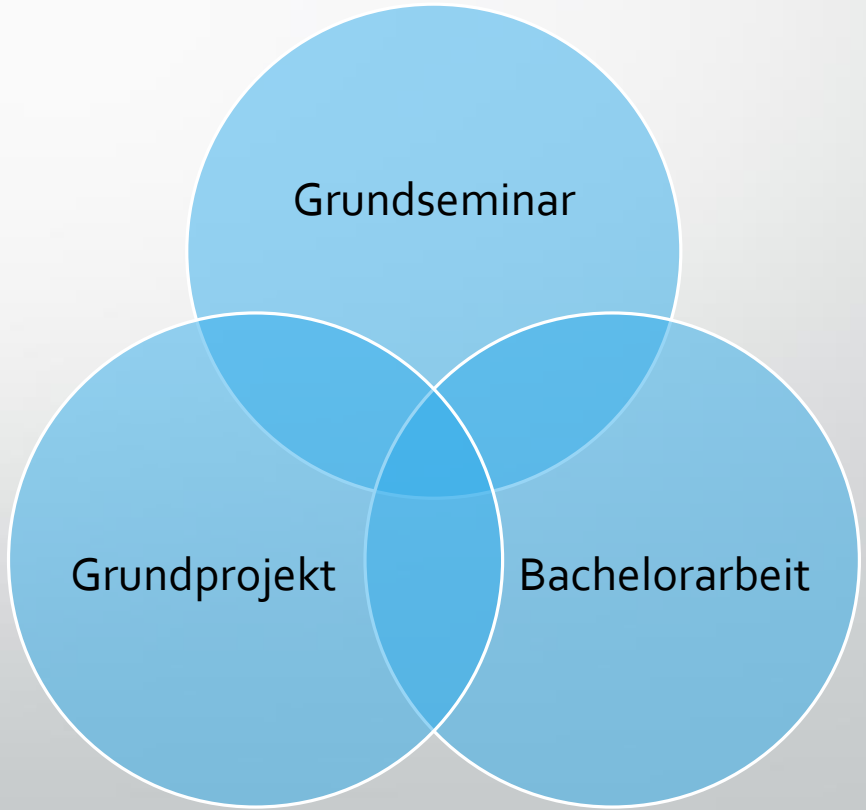
- Was will ich machen?



Einleitung und Motivation

Was habe ich bisher gemacht?

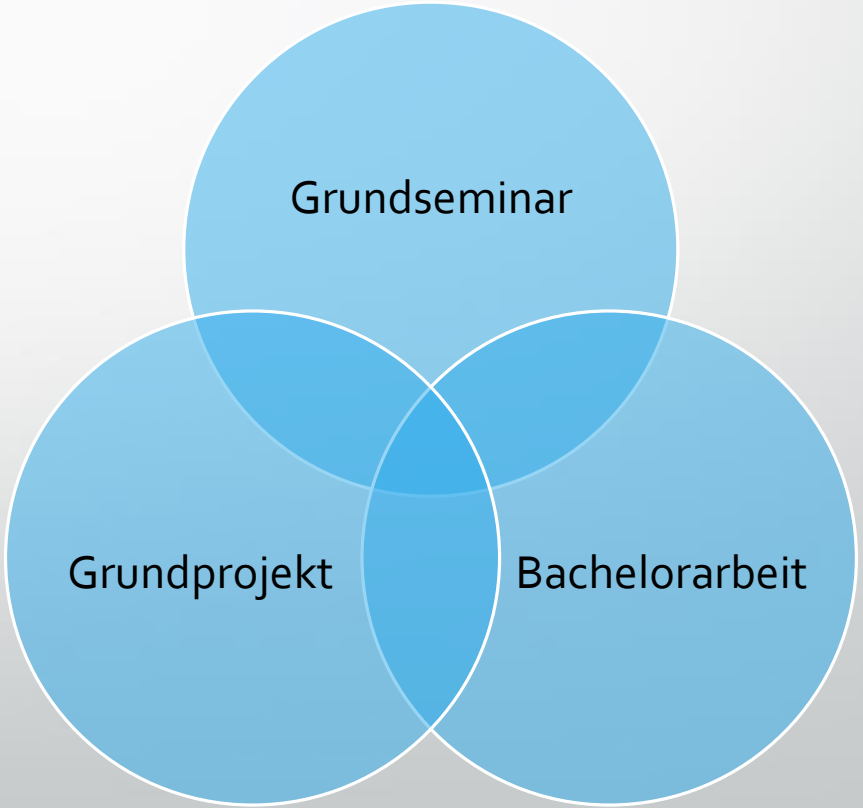
- Bachelorarbeit: eigener Benchmark von Apache Spark und Apache Hadoop
- Grundseminar: Analyse der Thematik Benchmarking und existierender Benchmarks in Kontext von Big Data
- Grundprojekt: Ausführung von BigDataBench



Einleitung und Motivation

Erkenntnisse?

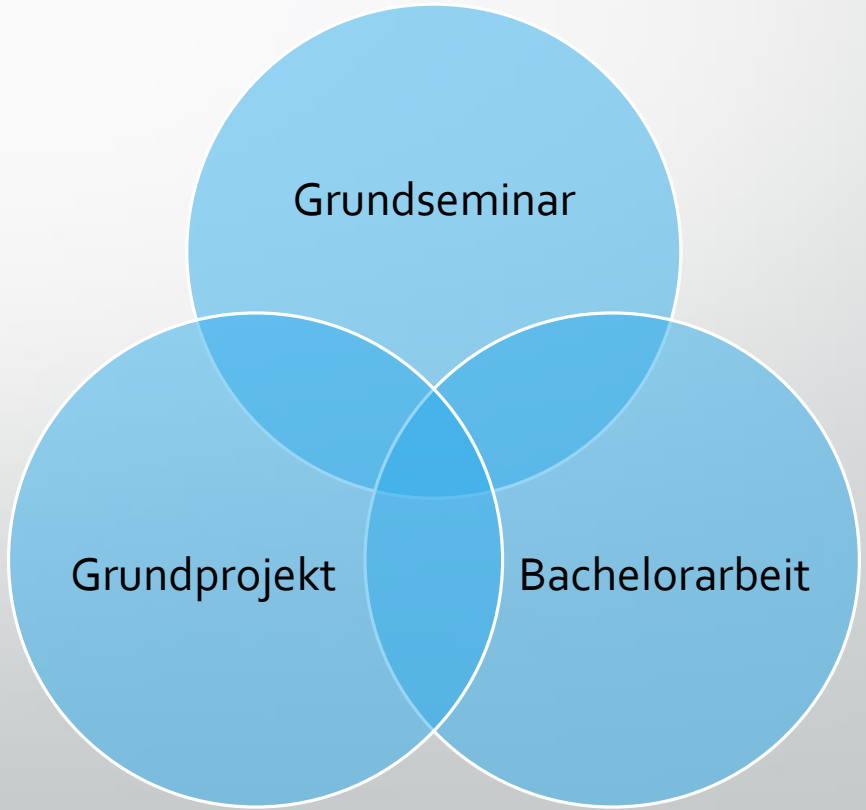
- Bachelorarbeit: eigener Benchmark sinnvoll bei speziellen Anforderungen
  - Flexibilität bei Konzeption und Ausführung
- Grundseminar: sehr große Auswahl existierender Benchmarks vorhanden
  - Viele Anwendungsfälle abgedeckt, allerdings oft veraltet
- Grundprojekt: Ausführung und Konfiguration von BigDataBench sehr aufwendig
  - Aktualität des Softwarestacks problematisch
  - Aussagekraft des Benchmarks teilweise fragwürdig



Einleitung und Motivation

Zentrale Herausforderungen!

- Bachelorarbeit: eigener Benchmark sinnvoll bei speziellen Anforderungen
  - Flexibilität bei Konzeption und Ausführung
- Grundseminar: sehr große Auswahl existierender Benchmarks vorhanden
  - Viele Anwendungsfälle abgedeckt, allerdings oft veraltet
- Grundprojekt: Ausführung und Konfiguration von BigDataBench sehr aufwendig
  - Aktualität des Softwarestacks problematisch
  - Aussagekraft des Benchmarks teilweise fragwürdig



Einleitung und Motivation

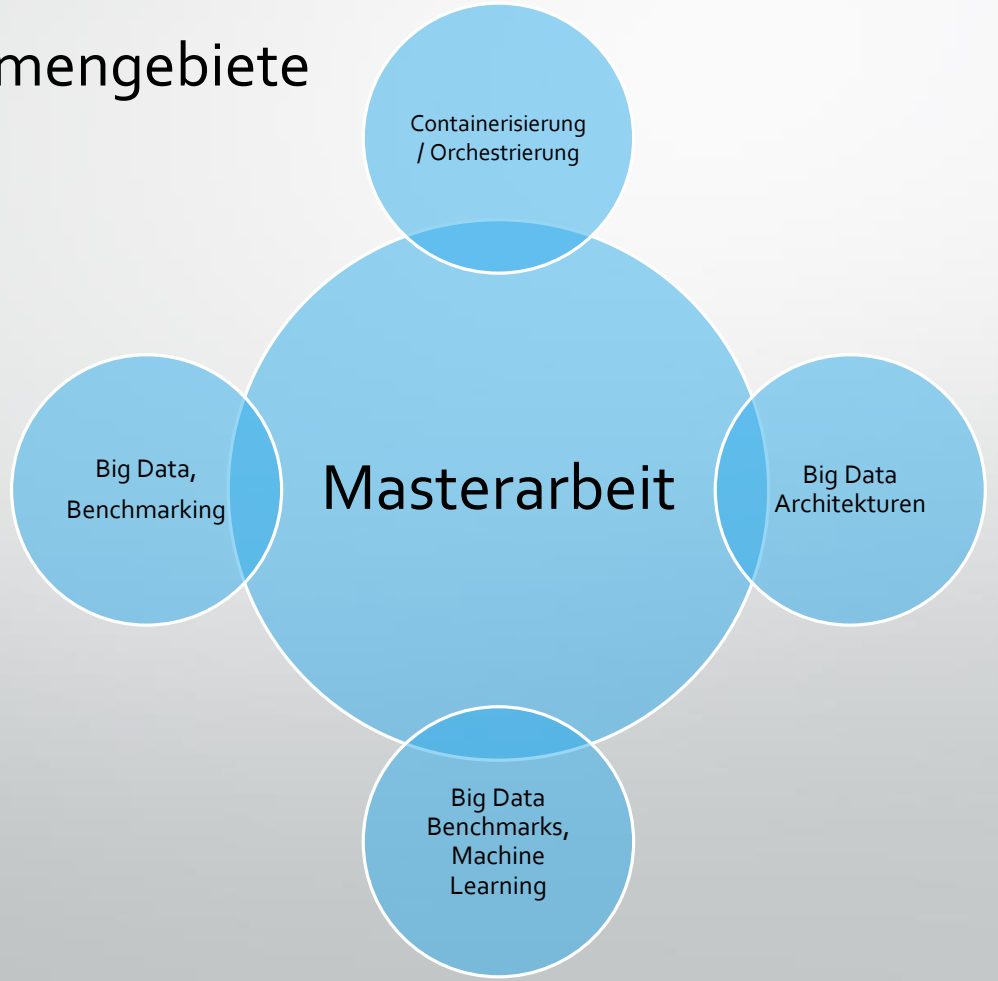
Herausforderungen von Big Data Benchmarks:

- Eindeutige Ergebnisbildung
- Aktualität
- Aussagekraft
- Am Use Case orientiert

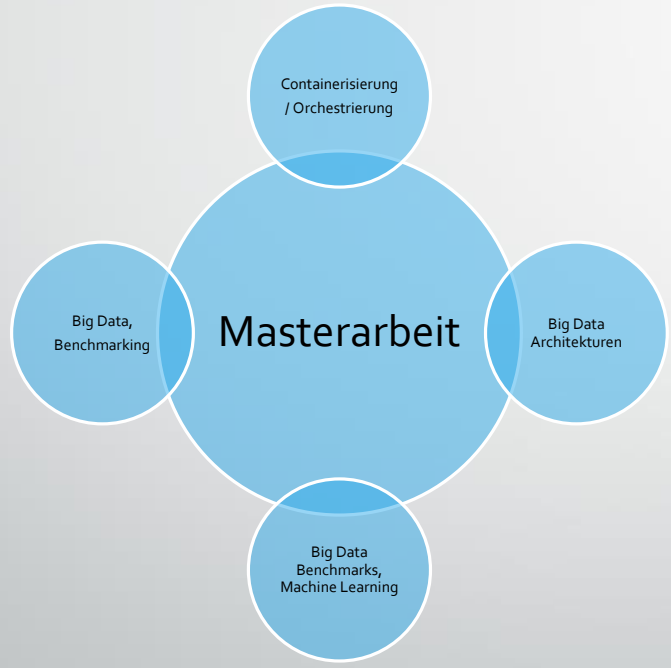
Notwendigkeit einer Big Data Architektur

Einleitung und Motivation

Grundlegende Themengebiete für Masterarbeit



Einleitung und Motivation



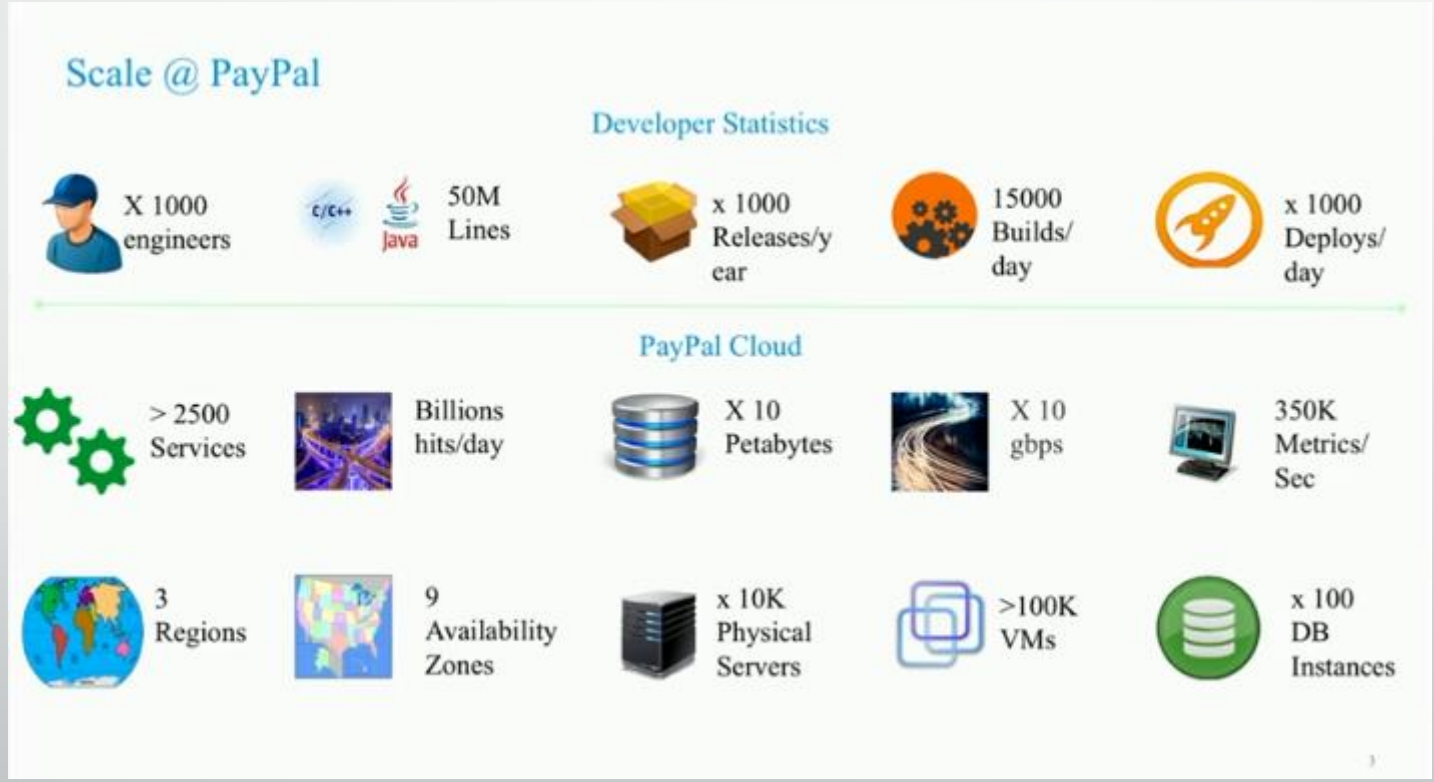
Forschungsgebiete

- Big Data + Machine Learning: Andrew Ng
- Benchmarking: Jim Gray, Michael Stonebraker
- Big Data Benchmarks: Yahoo! Cloud Serving Benchmark, BigDataBench, TPCx-HS
- Big Data Architekturen: Matei Zaharia, Werner Vogels (Amazon), Raul Estrada, Isaac Ruiz
- Containerisierung / Orchestrierung: Matei Zaharia

Big Data Architekturen

Big Data Architekturen

Big Data Anwendungen?



[1]

Big Data Architekturen

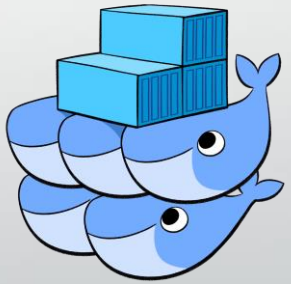
- BD-Architekturen so variabel wie Anwendungsfälle
- Trends: Clustermanager, Serverless

[2]

MESOS

kubernetes

[3]



Docker Swarm

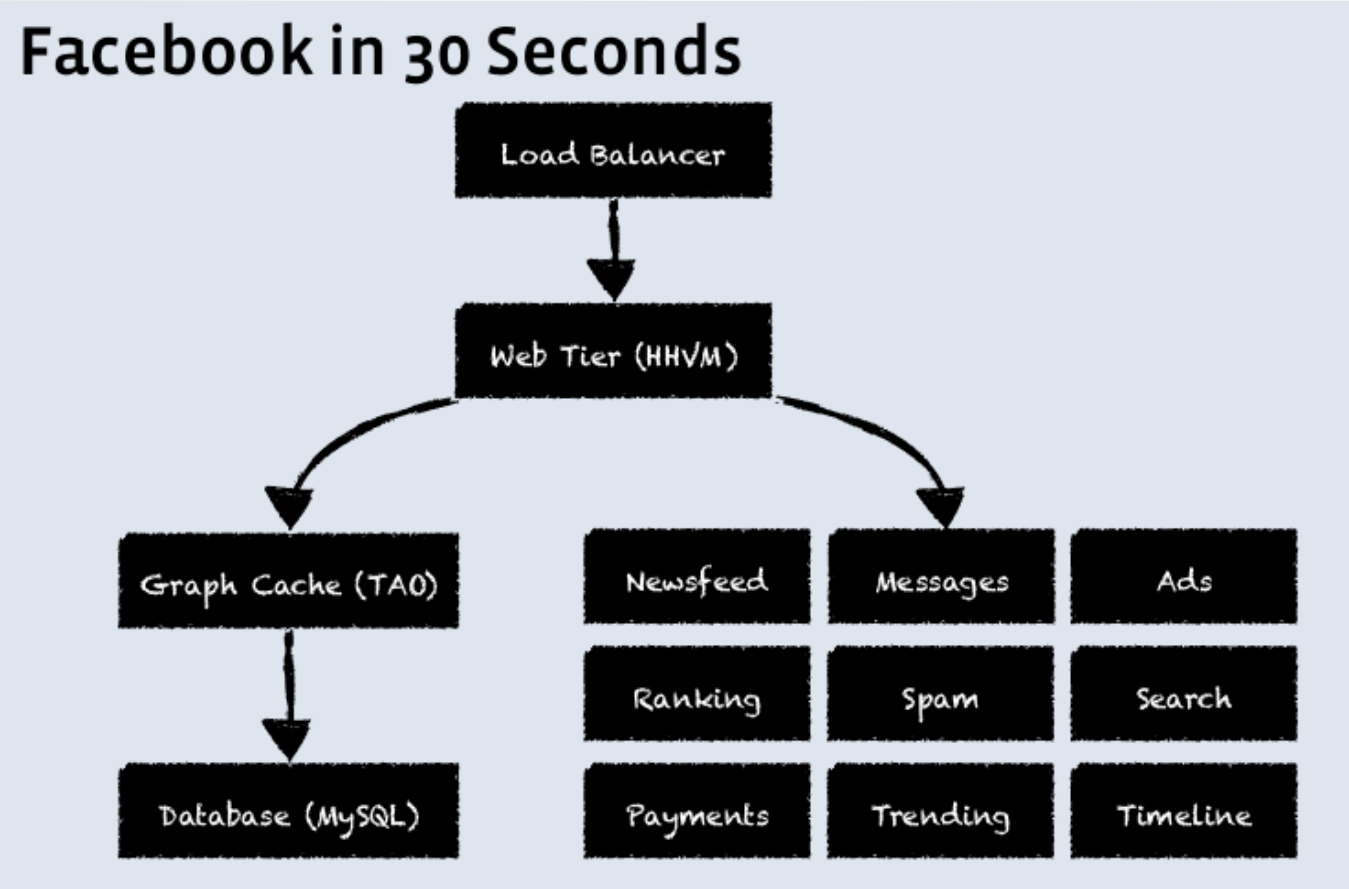
[4]

SERVERLESS

[5]

Big Data Architekturen

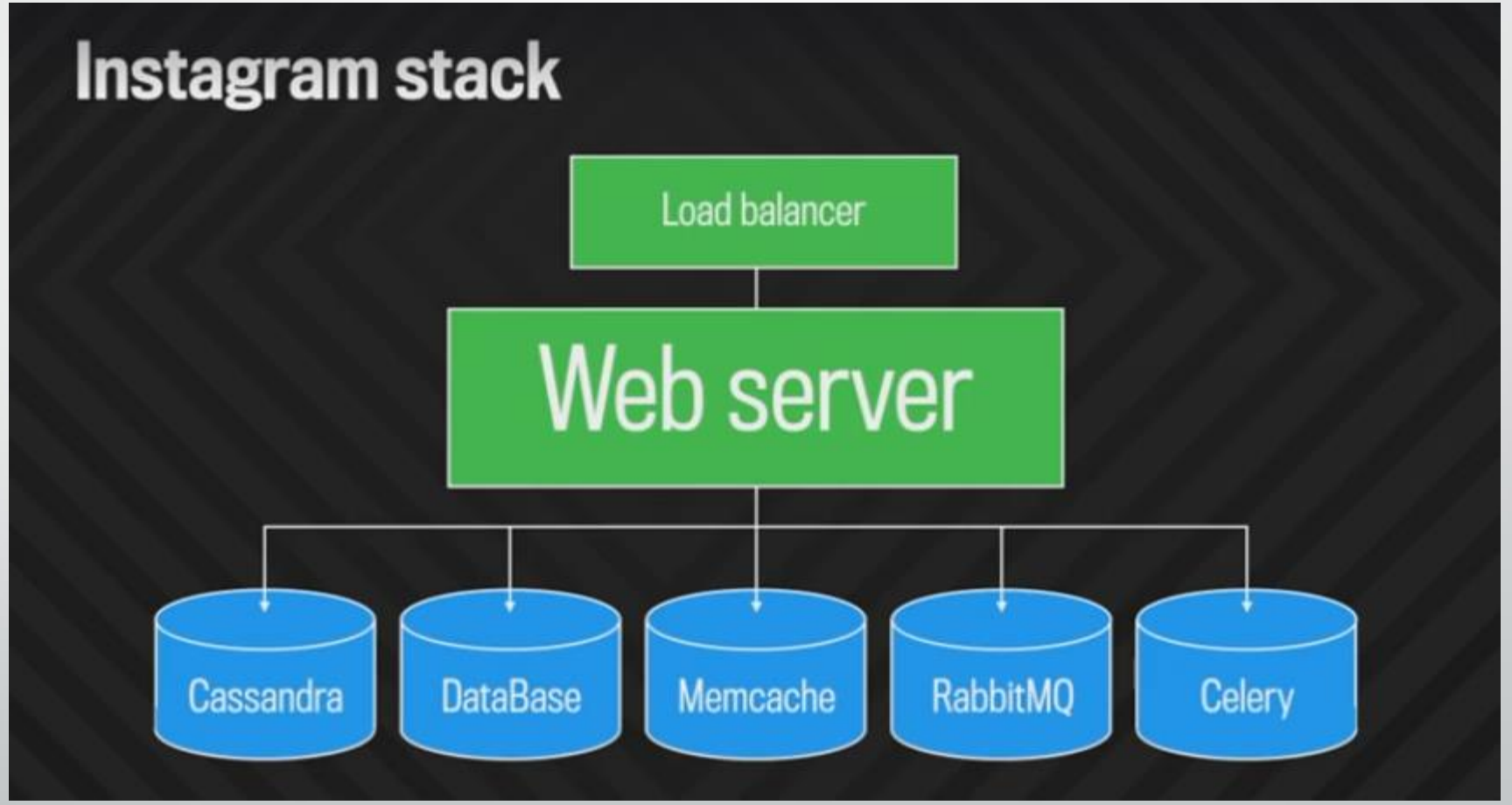
- Anwendungsfall 1: Facebook



[6]

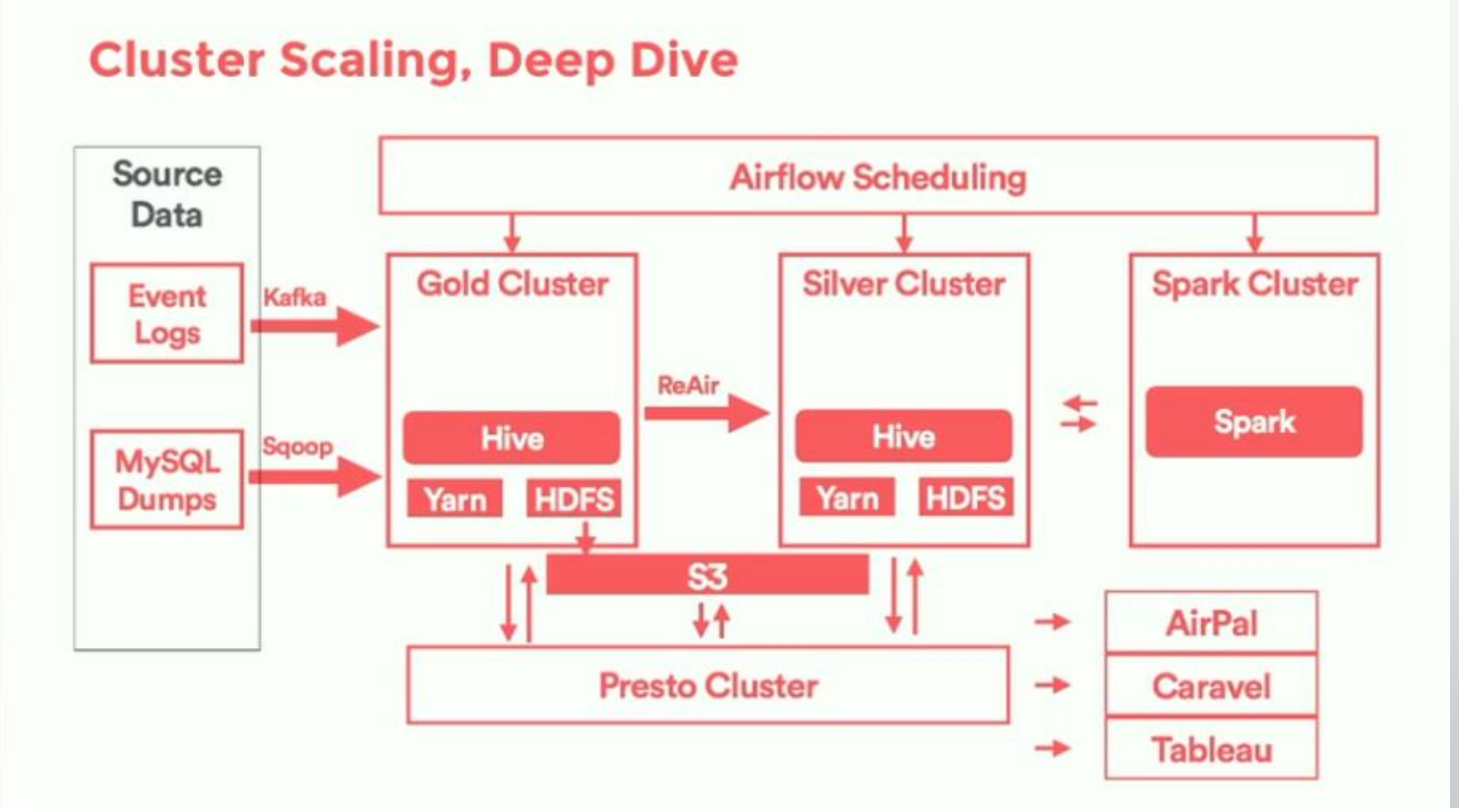
Big Data Architekturen

- Anwendungsfall 2: Instagram



[7]

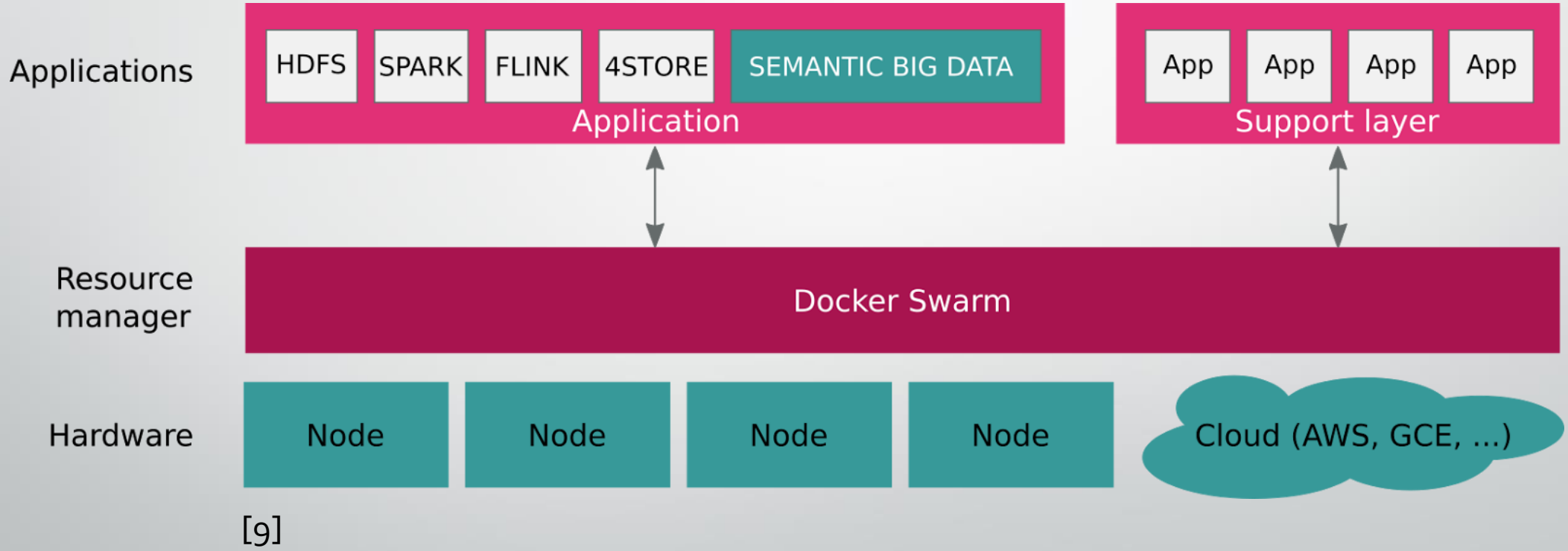
Big Data Architekturen



[8]

- Anwendungsfall 3: AirBnB

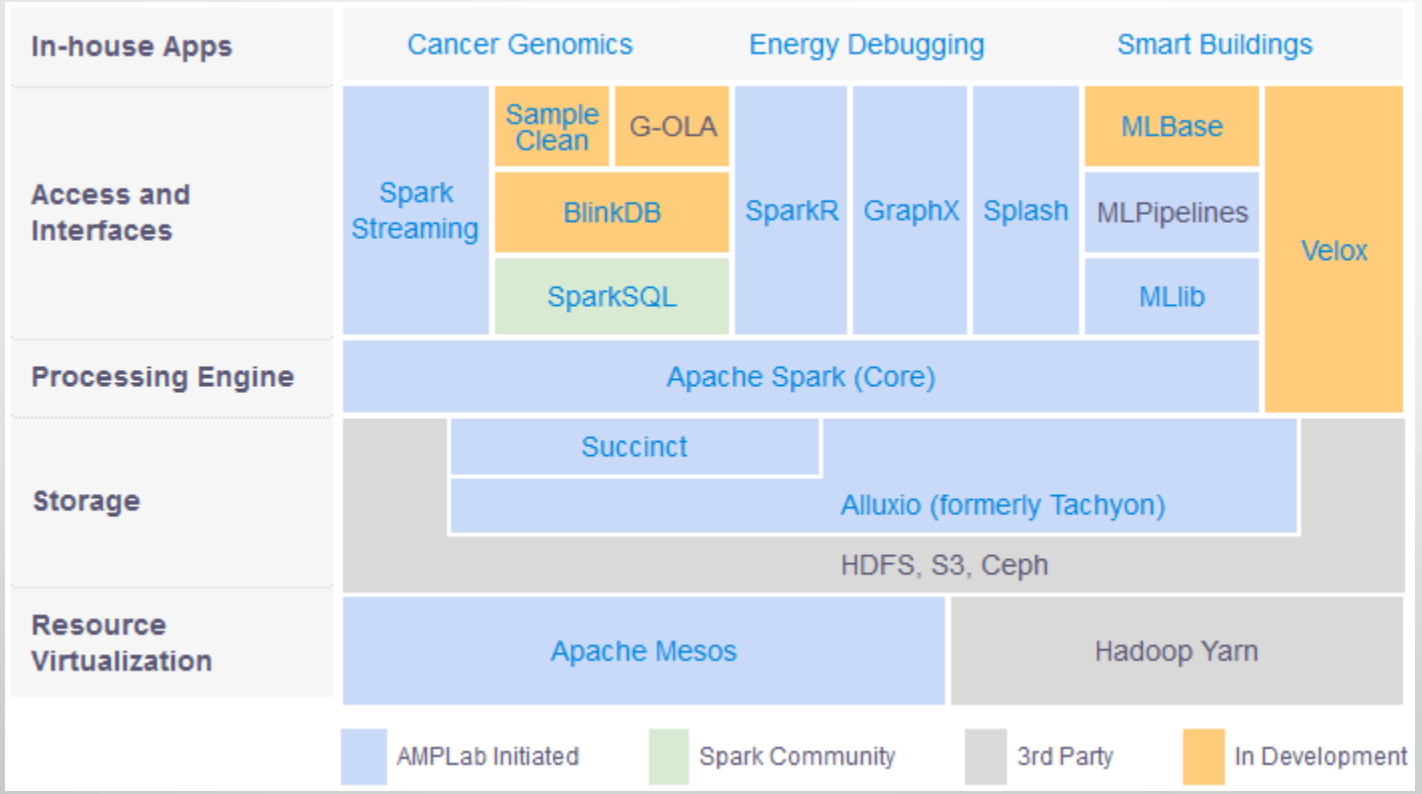
Big Data Architekturen



[9]

- BigDataEurope Architektur

Big Data Architekturen



[10]

- BDAS (Berkeley Data Analytics Stack)

Big Data Architekturen

[11]

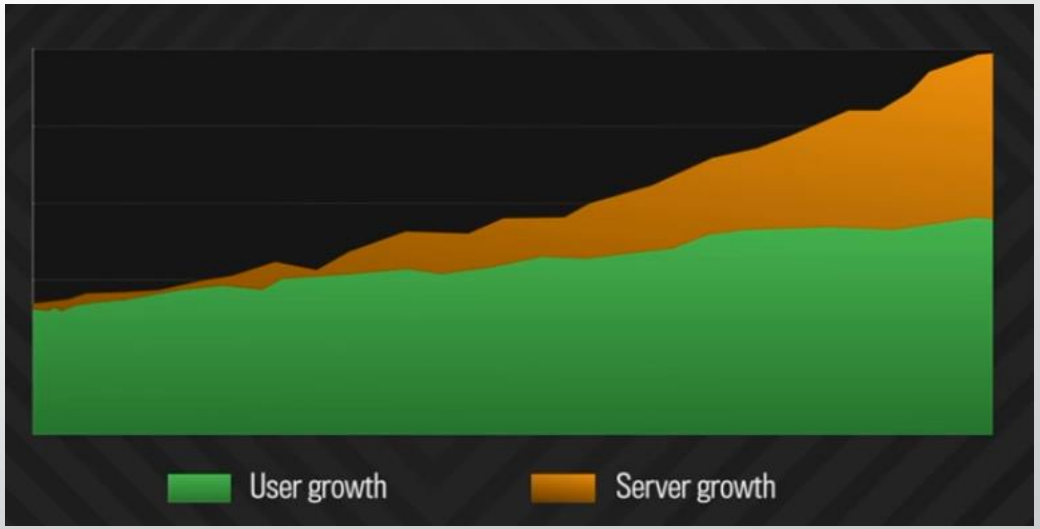
- Grundsätzliche Architektur: Big Data SMACK

Big Data Architekturen

Identifizierte Forschungslücken

1. Benchmarking von Big Data Anwendungen / Use Cases
2. Benchmark der verschiedenen Clustermanager als Grundlage einer Big Data Architektur
3. Benchmarking der Skalierungsfähigkeit einer Big Data Architektur

Instagram „Benchmark“



[12]

Big Data Architekturen

Wichtige wissenschaftliche Konferenzen

- IEEE International Congress on Big Data
- IEEE International Conference on Cloud and Big Data Computing
- IEEE/ACM International Symposium on Cluster, Cloud and Grid
- ACM International Conference on Web Search and Data Mining

[13]

[14]

Praxisorientierte Konferenzen

- https://www.big-data-europe.eu/
- http://acl2016.org/
- https://atscaleconference.com/

ACL 2016
AUGUST 7 – 12 | BERLIN, GERMANY

[15]

BIG DATA EUROPE
Empowering Communities with Data Technologies

[16]

@SCALE

[17]

Ideen für Umsetzung

Konzepte der Clustermanager

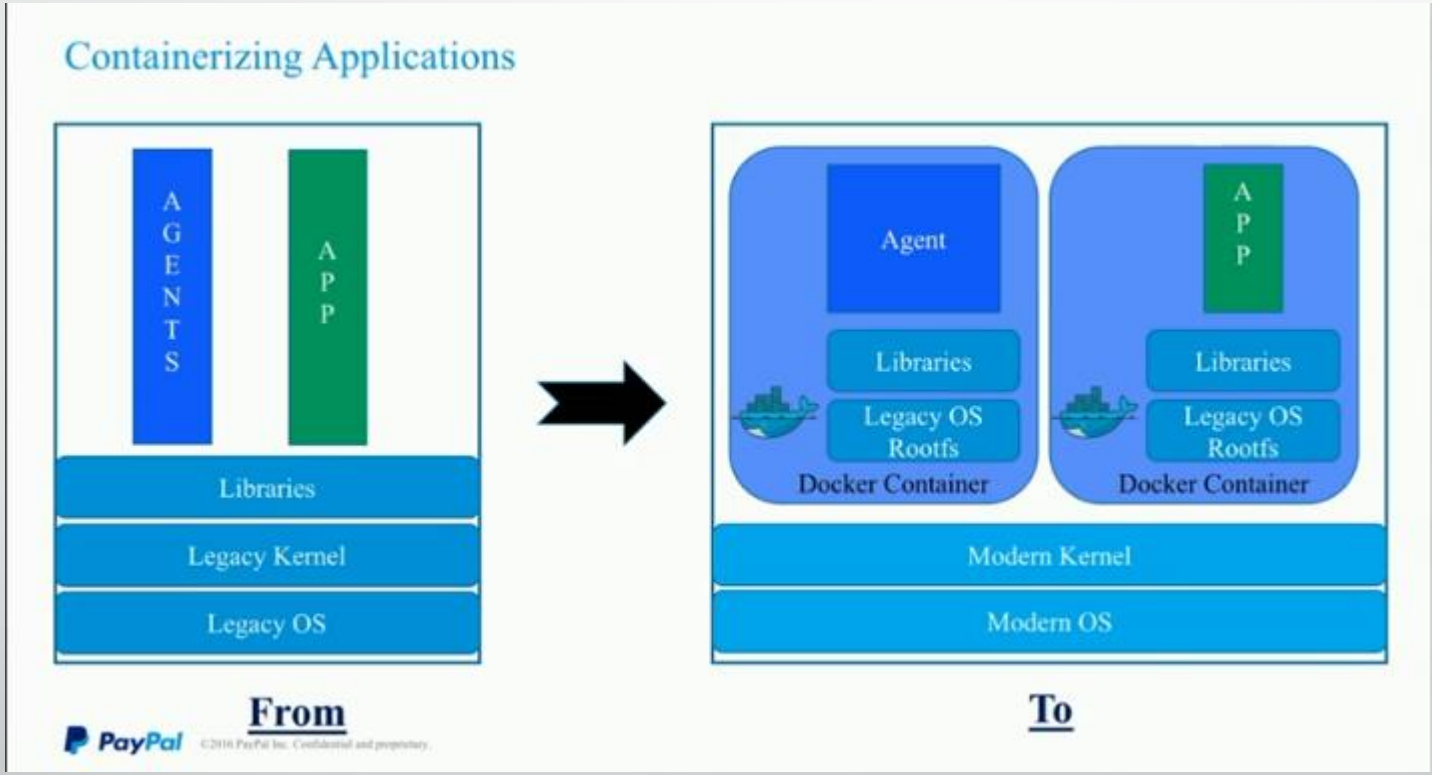
Benchmarking einer Big Data Infrastruktur

Vorläufige Vision Masterarbeit

Konzepte der Clustermanager

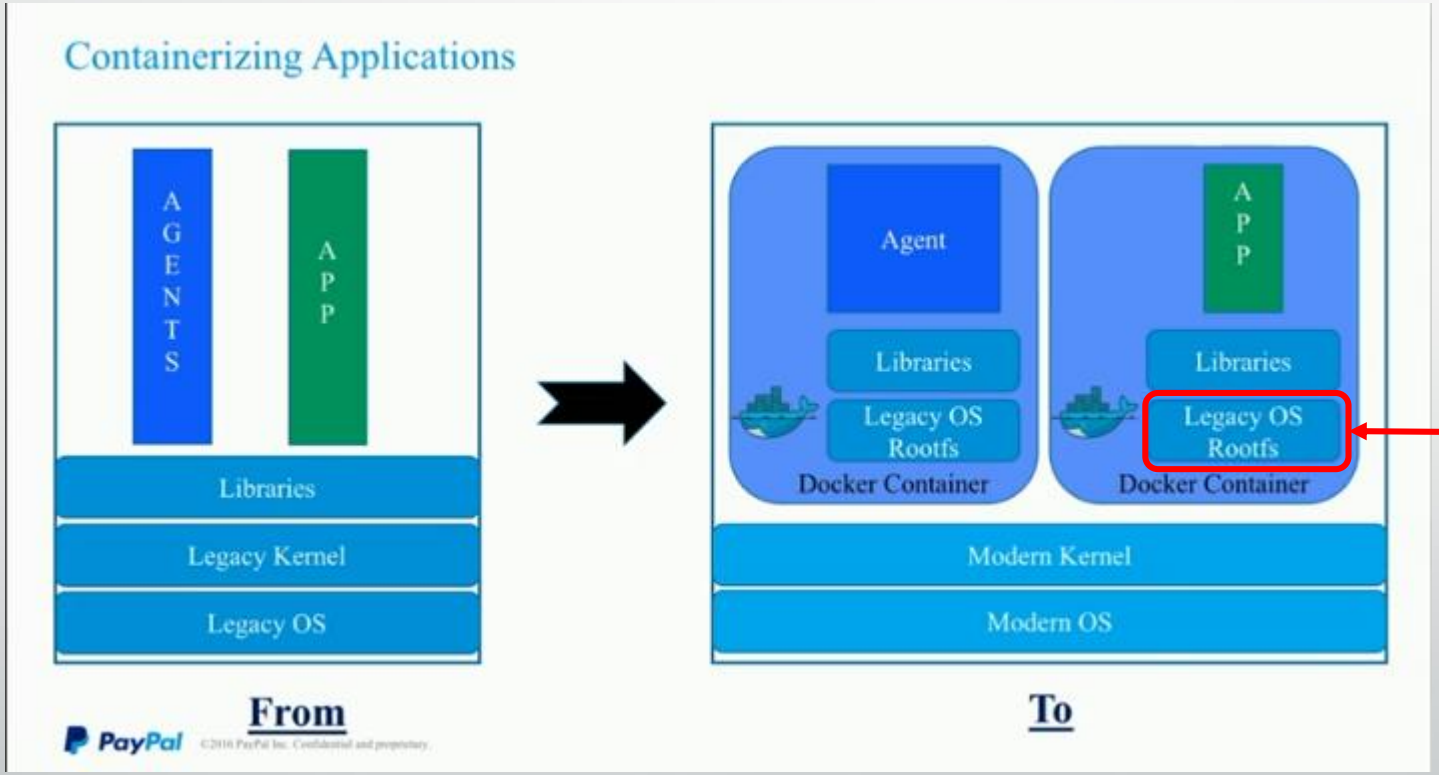
- Clustermanager als zentrale Instanz in Big Data Architekturen vorhanden
- Grundlage für flexibles und schnelles Cluster-Deployment von Anwendungen als Docker-Container -> Containerisierung
- Flexible Entwicklung und Portierung von Benchmarks inklusive passender Infrastruktur

Konzepte der Clustermanager



[18]

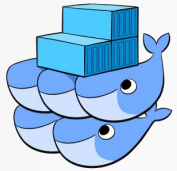
Konzepte der Clustermanager



Problematik veralteter Software und Benchmark-Stacks kann behoben werden -> erleichterte Portierung auf moderne Stacks

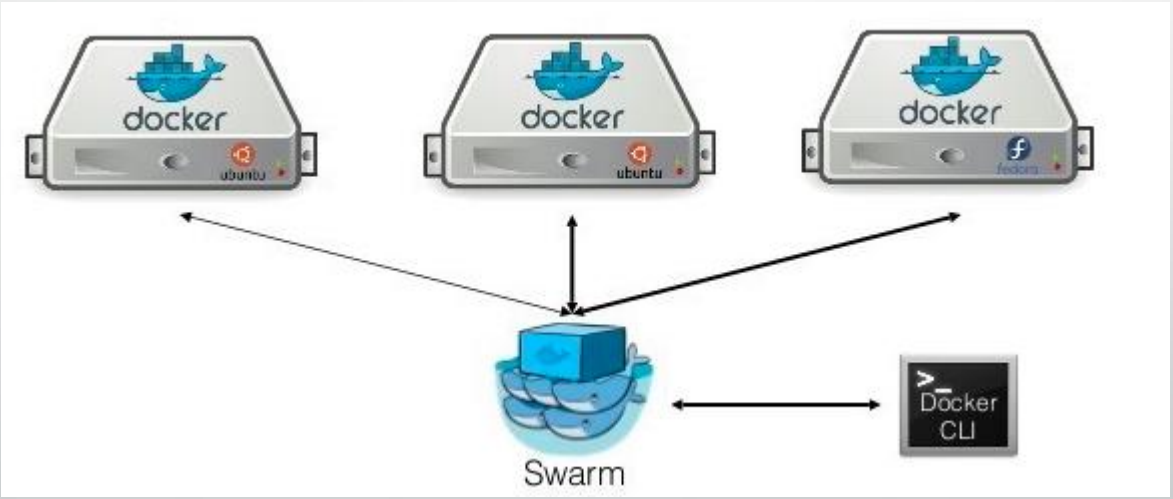
[18]

Konzepte der Clustermanager



[4]

- Docker Swarm

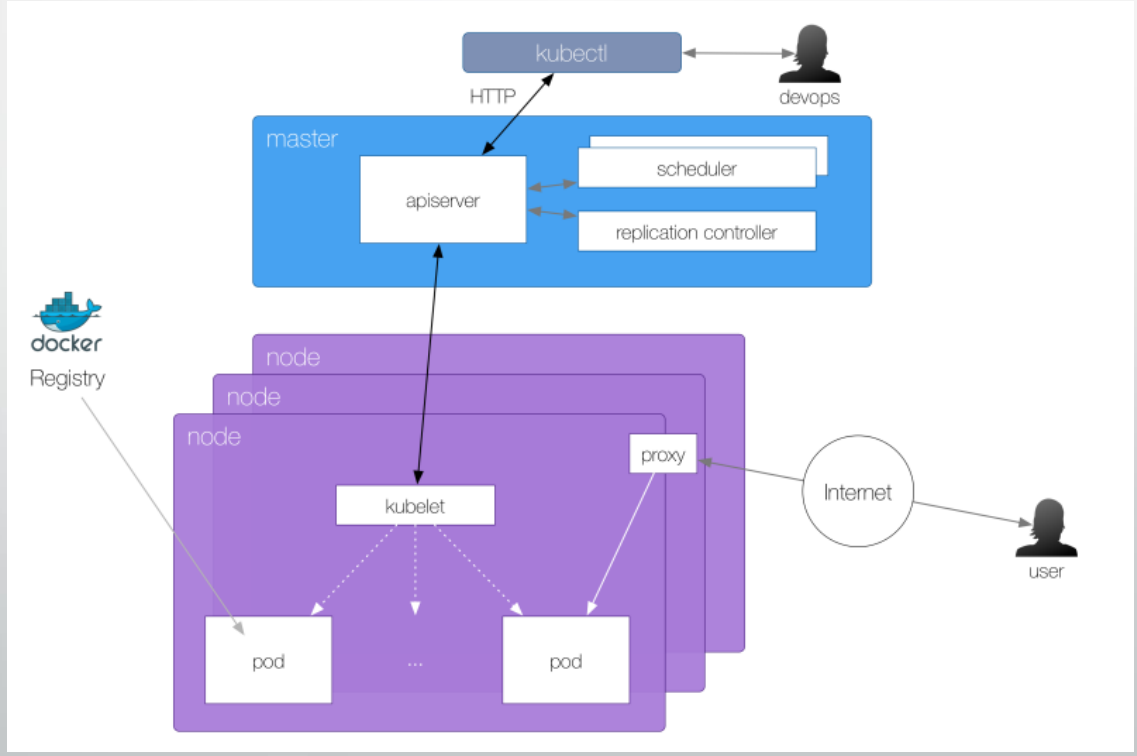


[19]

Konzepte der Clustermanager

kubernetes [3]

- Google Kubernetes

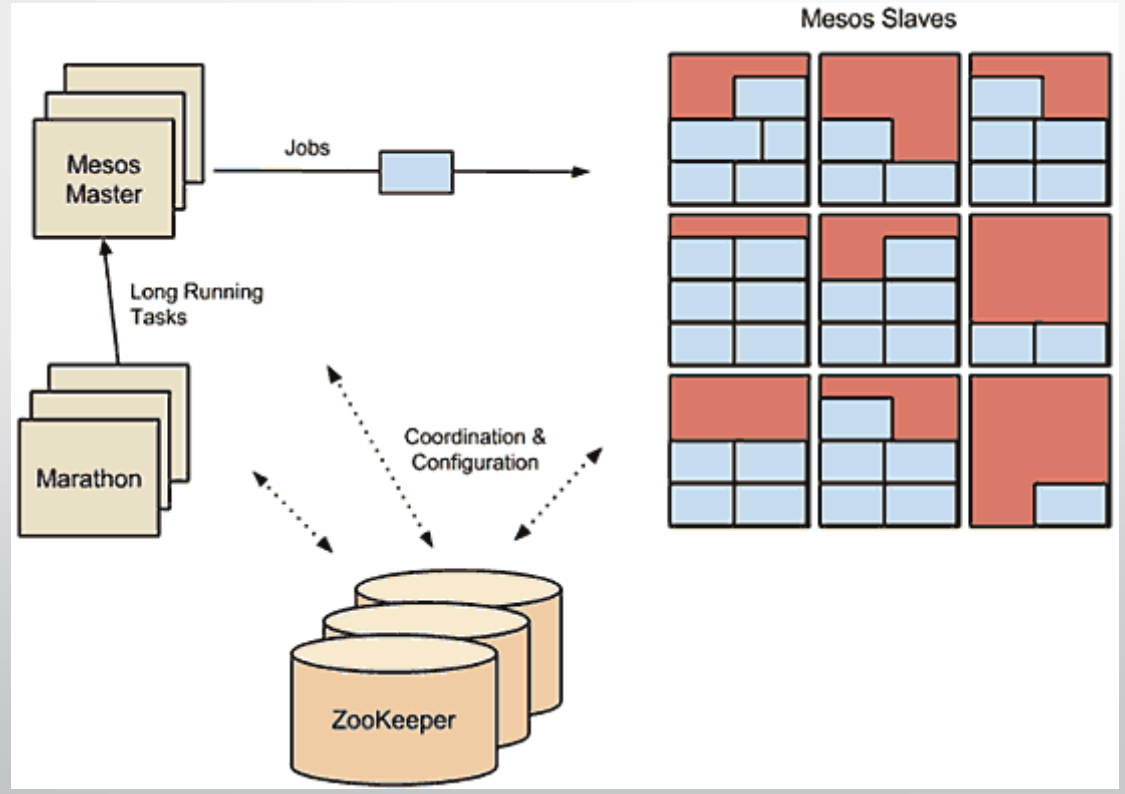


[20]

Konzepte der Clustermanager

MESOS [2]

- Apache Mesos



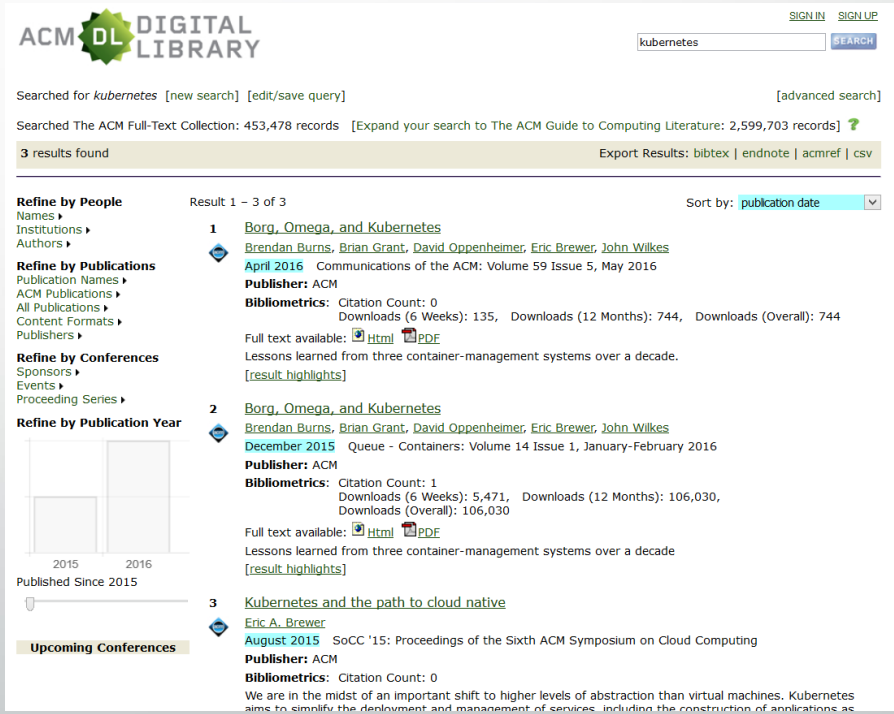
[21]

Konzepte der Clustermanager

Risiken

- Sehr starke Dynamic auf dem Markt der Clustermanager
  - Produktabhängigkeit vermeiden
- Hohe Komplexität individueller Systeme/Frameworks/Benchmarks
  -> Installation schwierig
  - Hypothese: Problem kann durch Containerisierung gelöst werden

    -> Aufwendige Infrastruktur



[22]

Benchmarking einer Big Data Infrastruktur

Zentrale Eigenschaften:

- (Automatische) Skalierung
- Fehlertoleranz (eventuell Benchmarks aus dem Themengebiet Overlay Networking anschauen)
- (Automatisches) Loadbalancing
- Test von heterogener Hardware/Container

Benchmarking einer Big Data Infrastruktur

Herausforderungen
- Metriken
- Benchmark-Monolithen und Big Data Infrastruktur anpassen
  - Containerisierug -> bspw. Docker

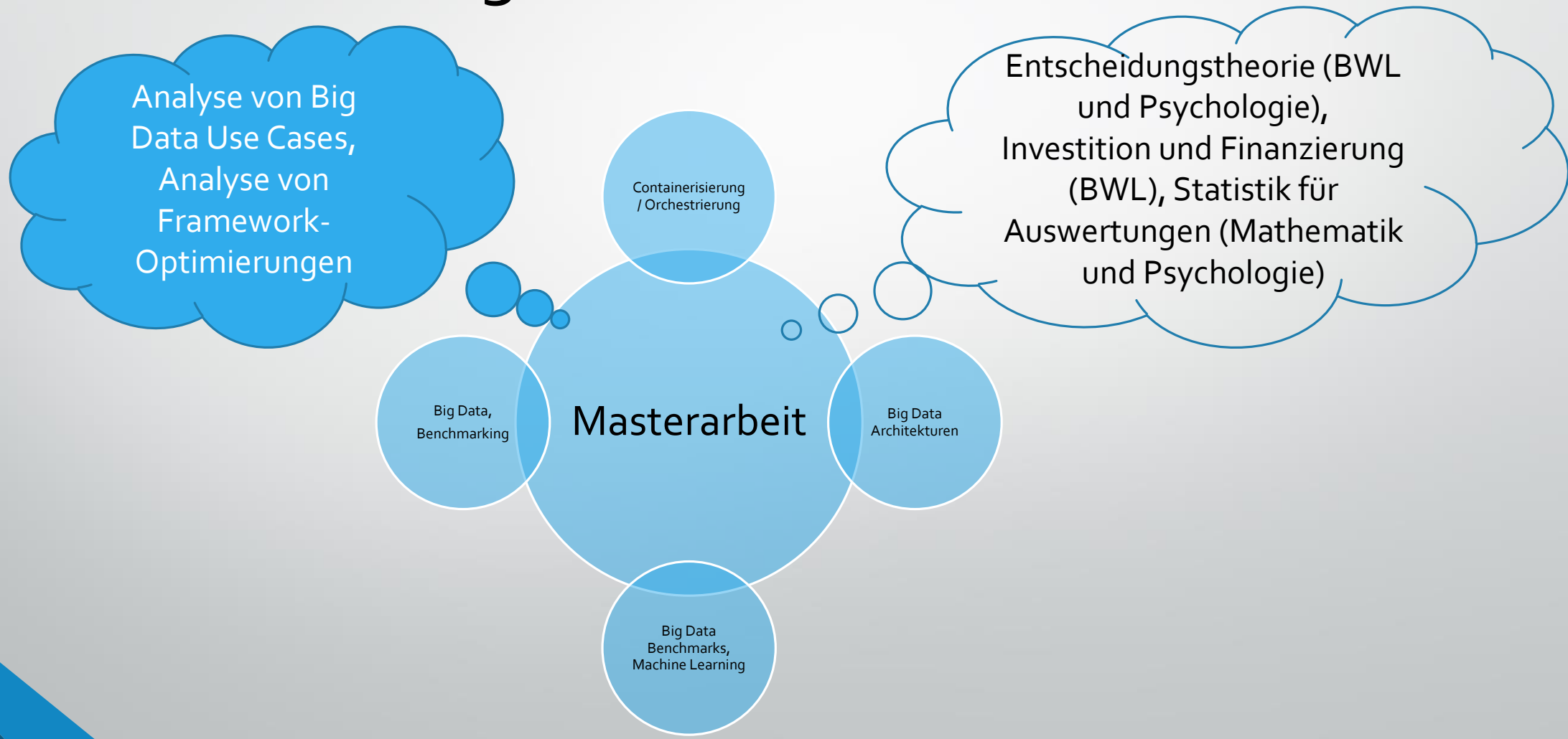
Schwer möglich für KMU

Implementation und Konfiguration einer Big Data Infrastruktur inklusive Benchmarks zum Finden von Use Cases und zur Optimierung

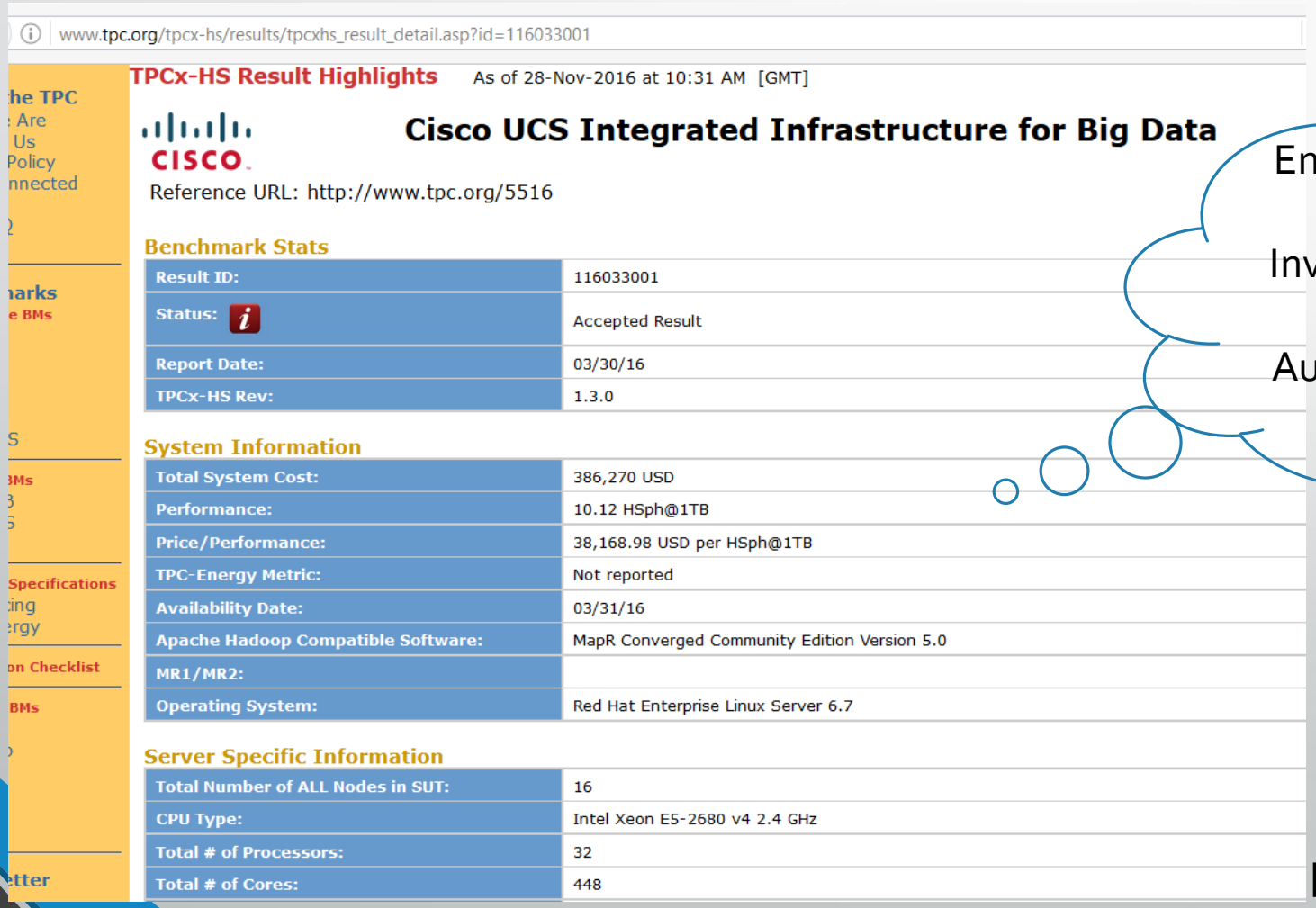
Benchmarking einer Big Data Infrastruktur

- Hauptseminar
  - Welche Big Data Architektur und welcher Softwarestack?
- Hauptprojekt
  - Auswahl einer Big Data Architektur/Softwarestack und Implementation im Cluster
  - Erste Benchmark-Konzepte von Big Data Anwendungen
- Masterarbeit
  - Analyse von Use Cases
  - Anwendung von Use Cases auf Big Data Architektur
  - Spezialisierte Eigenschaften der Frameworks benchmarken
  - Komplexe Analysen (z.B. Machine Learning, Data Mining)

Vorläufige Vision Masterarbeit



Vorläufige Vision Masterarbeit


www.tpc.org/tpcx-hs/results/tpcxhs_result_detail.asp?id=116033001

TPCx-HS Result Highlights As of 28-Nov-2016 at 10:31 AM [GMT]

Cisco UCS Integrated Infrastructure for Big Data

Reference URL: http://www.tpc.org/5516

Benchmark Stats

Result ID:	116033001
Status:	Accepted Result
Report Date:	03/30/16
TPCx-HS Rev:	1.3.0

System Information

Total System Cost:	386,270 USD
Performance:	10.12 HSph@1TB
Price/Performance:	38,168.98 USD per HSph@1TB
TPC-Energy Metric:	Not reported
Availability Date:	03/31/16
Apache Hadoop Compatible Software:	MapR Converged Community Edition Version 5.0
MR1/MR2:	
Operating System:	Red Hat Enterprise Linux Server 6.7

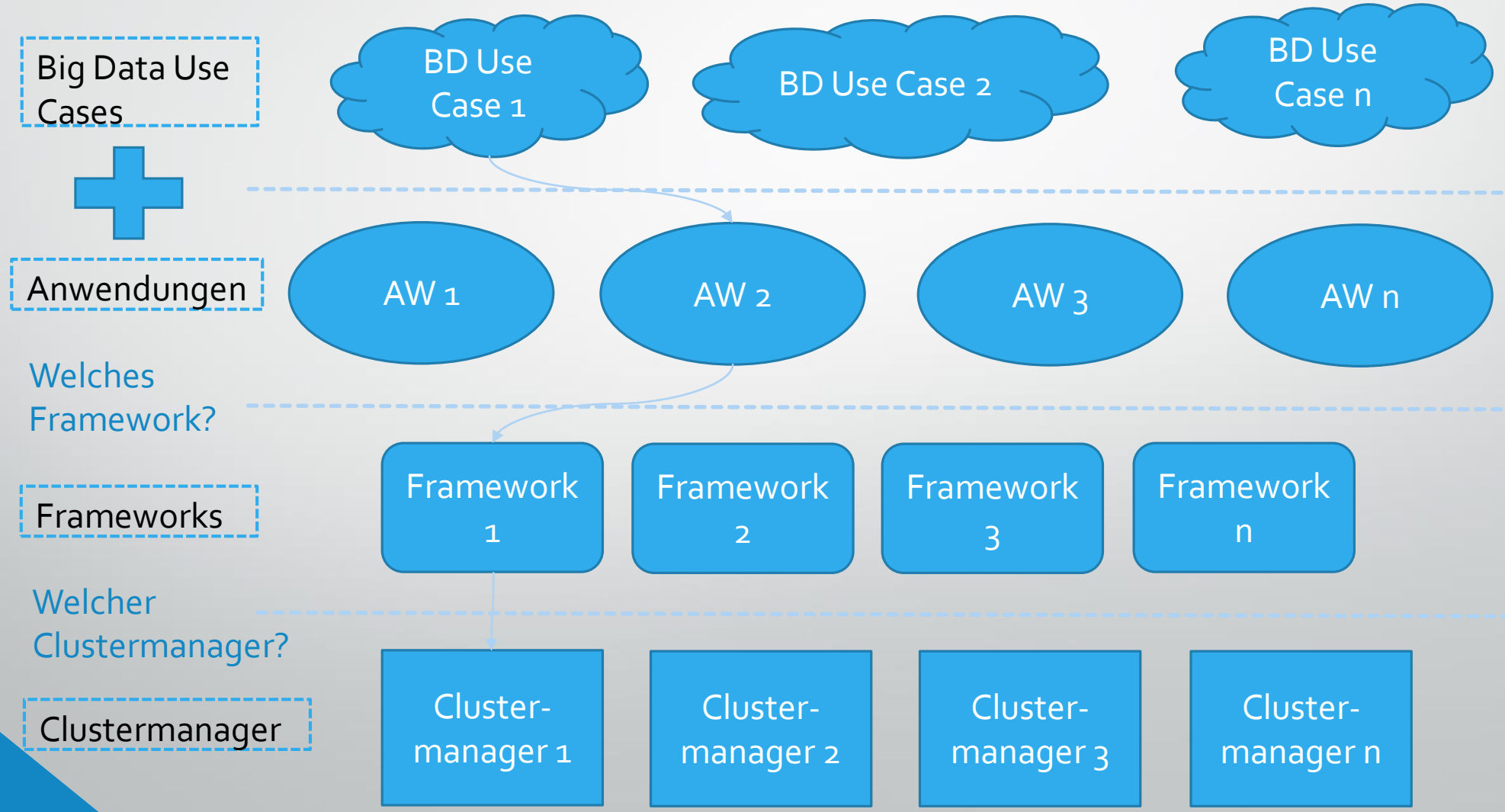
Server Specific Information

Total Number of ALL Nodes in SUT:	16
CPU Type:	Intel Xeon E5-2680 v4 2.4 GHz
Total # of Processors:	32
Total # of Cores:	448

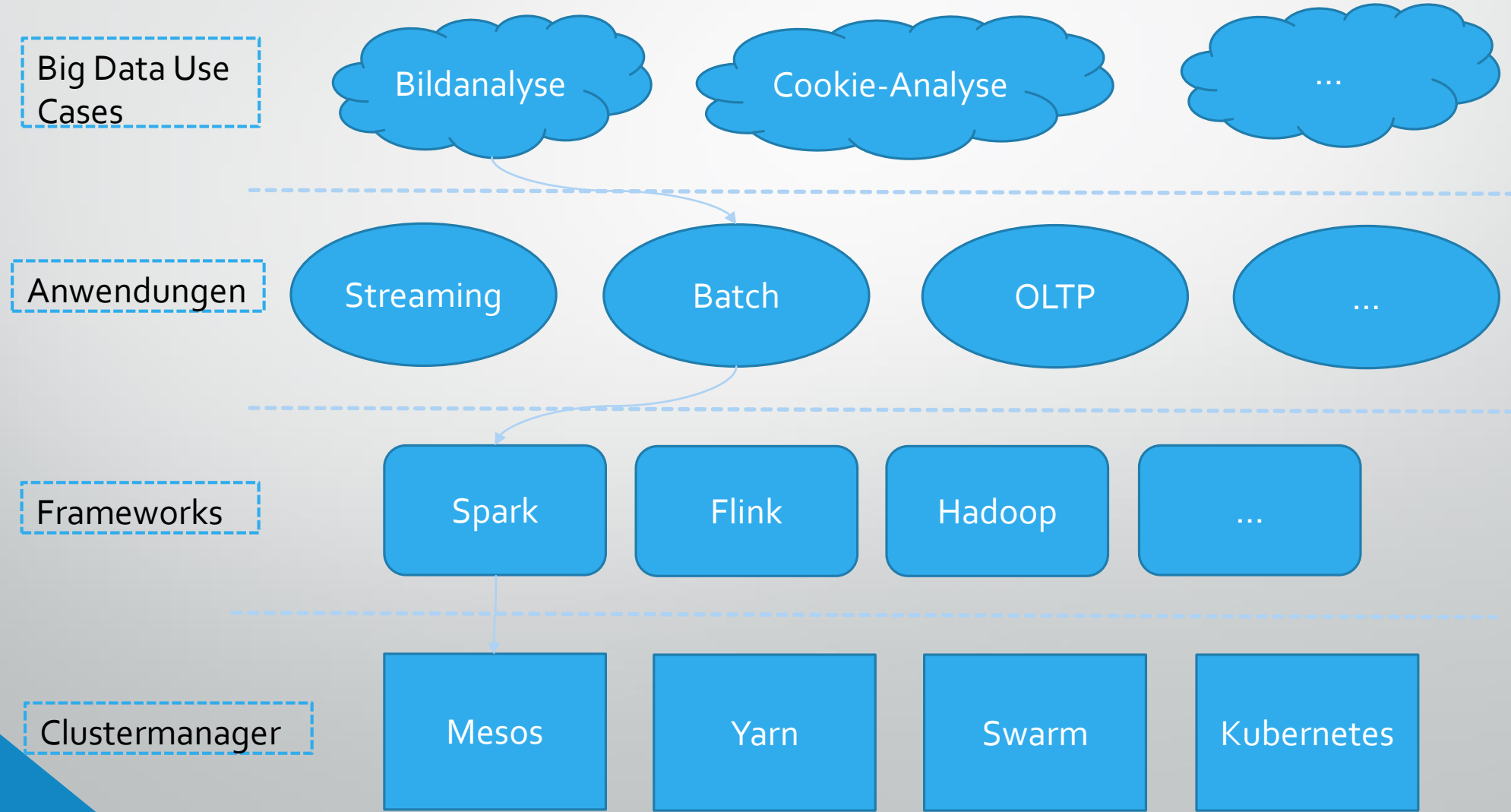
Entscheidungstheorie (BWL und Psychologie), Investition und Finanzierung (BWL), Statistik für Auswertungen (Mathematik und Psychologie)

[23]

Vorläufige Vision Masterarbeit



Vorläufige Vision Masterarbeit



Vielen Dank fürs Zuhören!

Fragen, Kritik und Anregungen?

Bildquellen (I)

[1] Abb. Entnommen aus https://atscaleconference.com/videos/devops-and-containerization-at-scale/, Abruf 14.12.16

[2] Abb. Entnommen aus https://mesosphere.com/why-mesos/, Abruf 14.12.16

[3] Abb. Entnommen aus https://plus.google.com/116512812300813784482/videos, Abruf 14.12.16

[4] Abb. Entnommen aus: https://www.docker.com/products/docker-swarm,Abruf 14.12.16

[5] Abb.. Entnommen von: https://serverless.com/, Abruf 14.12.16

[6] Abb.. Entnommen von: https://queue.acm.org/downloads/applicative/bmaurer.pdf, Abruf 14.12.16

[7] Abb.. Entnommen von: https://atscaleconference.com/videos/running-instagram/, Abruf 14.12.16

[8] Abb. Entnommen aus: https://atscaleconference.com/videos/airbnbs-data-evolution-lessons-from-building-a-world-class-data-ecosystem/, Abruf 14.12.16

Bildquellen (II)

[9] Abb. Entnommen von https://docs.google.com/presentation/d/1U5IEXdjYCRIreuoigtYsAIDciHpx8-6bEydVwlh1ZZ4/edit#slide=id.g17b76a891f_0_117, Abruf 14.12.16

[10] Abb. Entnommen von https://amplab.cs.berkeley.edu/software/, Abruf 14.12.16

[11] Estrada, Raul ; Ruiz, Isaac: Big Data SMACK: A Guide to Apache Spark, Mesos, Akka, Cassandra, and Kafka : Apress, 2016 — ISBN 978-1-4842-2175-4, S. 12

[12] Abb. Entnommen von https://atscaleconference.com/videos/running-instagram/, Abruf 14.12.16

[13] Abb. Entnommen von: https://cbdcom2016.sciencesconf.org/ , Abruf 14.12.16

[14] Abb.. Entnommen von: http://www.ieeebigdata.org/2015/, Abruf 14.12.16

[15] Abb.. Entnommen von: http://acl2016.org/, Abruf 14.12.16

[16] Abb.. Entnommen von: https://https://www.big-data-europe.eu/, Abruf 14.12.16

[17] Abb. Entnommen von: https://atscaleconference.com/, Abruf 14.12.16

Bildquellen (III)

[18] Abb. Entnommen von https://atscaleconference.com/videos/devops-and-containerization-at-scale/, Abruf 14.12.16

[19] Abb. Entnommen von http://www.slideshare.net/Docker/docker-swarm-020, Abruf 14.12.16

[20] Abb. Entnommen von https://mesosphere.com/blog/2015/09/25/kubernetes-and-the-dcos/, Abruf 14.12.16

[21] Abb. Entnommen von: https://www.oreilly.com/ideas/swarm-v-fleet-v-kubernetes-v-mesos, Abruf 14.12.16

[22] Abb.. Entnommen von: http://dl.acm.org/results.cfm?query=+kubernetes&Go.x=0&Go.y=0, Abruf 14.12.16

[23] Abb.. Entnommen von: http://www.tpc.org/tpcx-hs/results/tpcxhs_result_detail.asp?id=116033001, Abruf 14.12.16