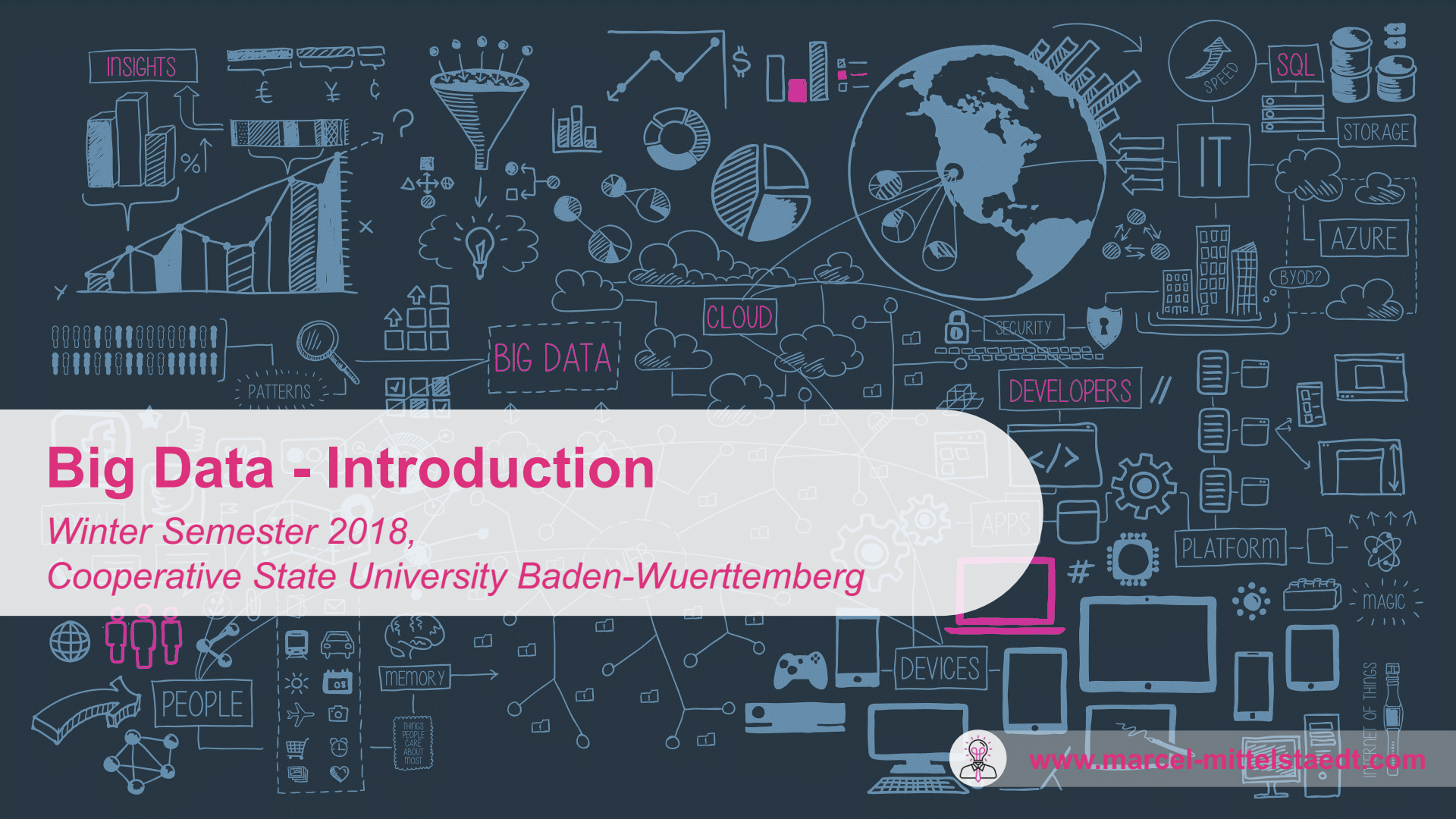


Big Data - Introduction

Winter Semester 2018,
Cooperative State University Baden-Wuerttemberg

www.marcel-mittelstaedt.com

Agenda – 01.10.2018

01

About This Lecture

Audience, Schedule, Materials (Slides, Script, Exercises, Books), Prerequisites, Exam, Feedback, Questions

02

Introduction To Big Data

Distributed Computing, Big Data V's, Big Data and Consistency, Big Data in Business, Definition of Big Data, Big Data Science

03

HandsOn - Hadoop

Quick Introduction To The Hadoop Ecosystem, HDFS, Yarn and MapReduce

04

Exercise

Setup Single Node Hadoop Cluster, Run MapReduce Jobs (Yarn)

About This Lecture

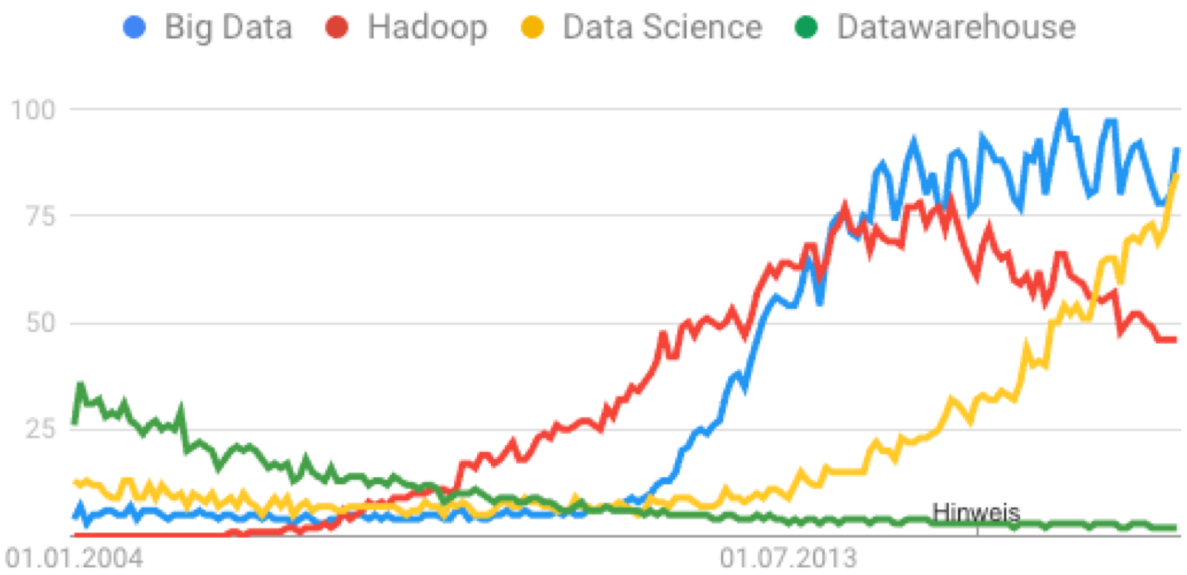
Scope, Schedule, Materials, Prerequisites, Exam, Feedback, Questions

About Big Data

“Big Data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so every one claims they are doing it.”

— Dan Ariely, Professor of Psychology and Behavioral Economics, Duke University

About Big Data



Google Trends, 26.09.2018

About This Lecture

- Sorting the **buzzwords** (Big Data, NoSQL, Scalability, Replication, Sharding…)
- Understand the **business value**
- Develop a basic understanding of **Big Data** and **Distributed Systems**:
  - How and why do they work?
  - Fundamental approaches (for e.g. Scaling, Distributed Storage, Distributed Processing, …)
  - Data Models and Query Languages
  - Consistency guarantees of Distributed Big Data data-systems
  - Advantages and disadvantages of different software and approaches
  - Combination of different technologies

About This Lecture

- **HandsOn Exercises** on several software and frameworks (e.g. Hadoop, Map-Reduce, Hive, Spark etc.)
- **How does a productive environment look like?** (e.g. Partitioning, Replication, ETL Workflow-Management etc.)
- Quick **introduction** to **data science**

- **Not in this Lecture**:
  - Database fundamentals (*previous lectures*)
  - …

About Me

2018
- today

Co-Founder
DataWhizz, www.datawhizz.io

2018
- today

University Lecturer
Cooperative State University Baden-Wuerttemberg,
www.dhbw-stuttgart.de

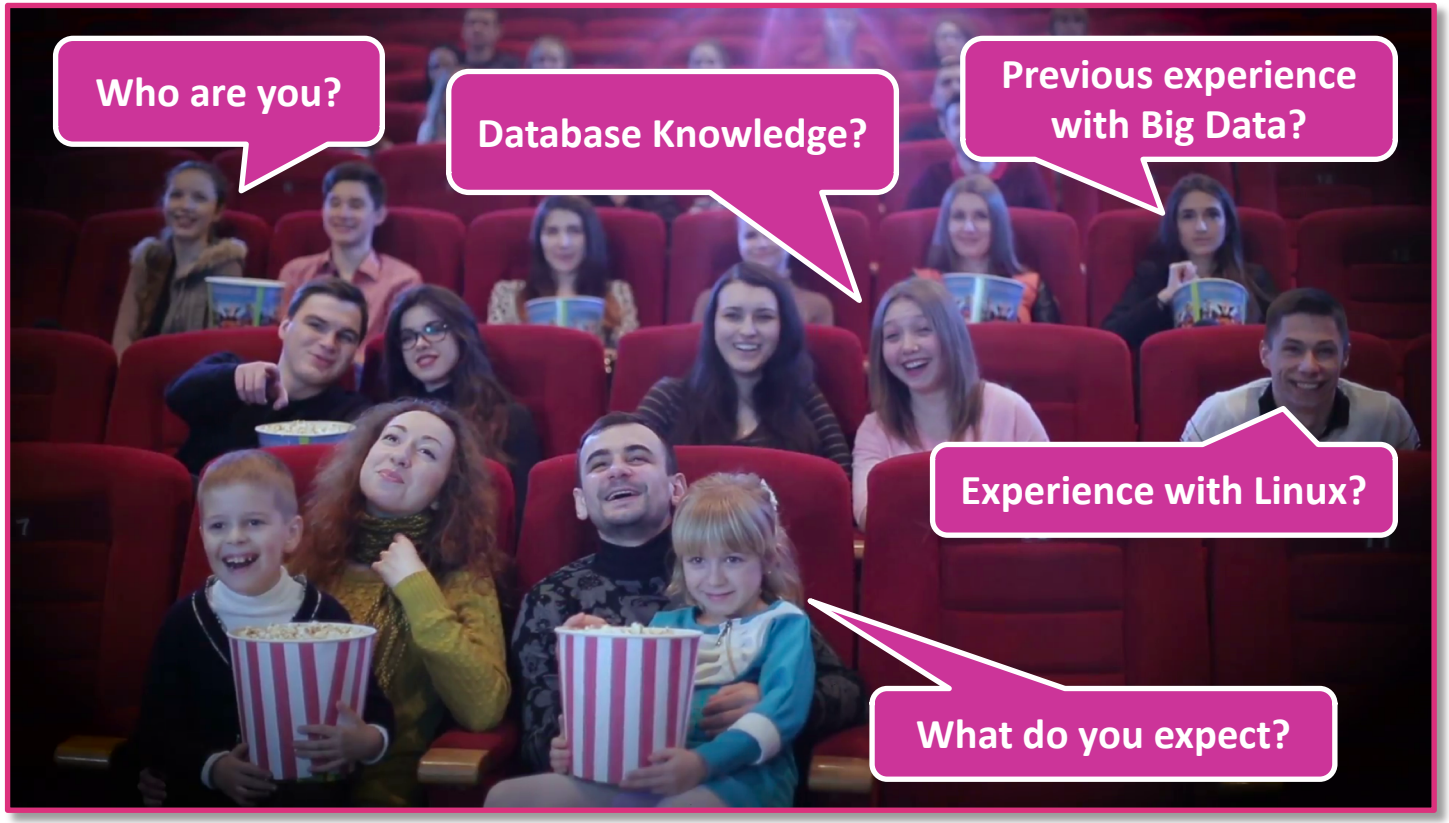
2015 - 2018
sep - nov

Head of Data Architecture and Development
ProSiebenSat.1 Media SE,
www.prosiebensat1.com

2009 - 2013
oct - dec

IT Architect
IBM Global Business Services,
ww.ibm.com

About You


Who are you?
Database Knowledge?
Previous experience with Big Data?
Experience with Linux?
What do you expect?

Schedule

01.10.2018	**16:00-18:30**	**Ro. 0.11**	**About This Lecture, Introduction to Big Data**	HandsOn Hadoop
08.10.2018	16:00-18:30	Ro. 0.11	Non-Functional Requirements Of Distributed Data-Systems	HandsOn MapReduce
15.10.2018	16:00-18:30	Ro. 0.11	Data Models and Access	HandsOn Hive and UDF's
22.10.2018	16:00-18:30	Ro. 0.11	Challenges Of Distributed Data Systems: Replication	HandsOn Spark
29.10.2018	16:00-18:30	Ro. 0.11	Challenges Of Distributed Data Systems: Partitioning	HandsOn MongoDB
05.11.2018	16:00-18:30	Ro. 0.11	Batch Processing	HandsOn Cassandra
12.11.2018	16:00-18:30	Ro. 0.11	Stream Processing	HandsOn Spark Streaming
19.11.2018	16:00-18:30	Ro. 0.11	ETL Workflow And Automation	HandsOn PDI/Airflow
26.11.2018	16:00-18:30	Ro. 0.11	Software And Frameworks	HandsOn tbd
03.12.2018	16:00-18:30	Ro. 0.11	Exam Preparation / Data Science	HandsOn PySpark
10.12.2018	tbd	tbd	**Exam**	

Schedule – Lecture Composition

01.10.2018	**16:00-18:30**	**Ro. 0.11**
08.10.2018	16:00-18:30	Ro. 0.11
15.10.2018	16:00-18:30	Ro. 0.11
22.10.2018	16:00-18:30	Ro. 0.11
29.10.2018	16:00-18:30	Ro. 0.11
05.11.2018	16:00-18:30	Ro. 0.11
12.11.2018	16:00-18:30	Ro. 0.11
19.11.2018	16:00-18:30	Ro. 0.11
26.11.2018	16:00-18:30	Ro. 0.11
03.12.2018	16:00-18:30	Ro. 0.11
10.12.2018	tbd	tbd

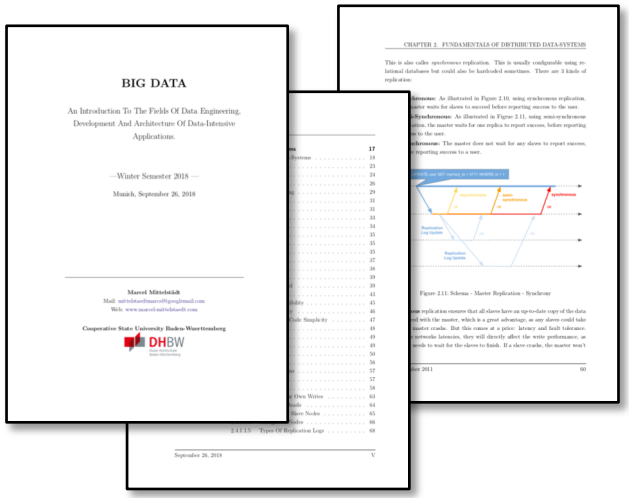
1. **Presentation, Review** and **Discussion** of **exercises** from previous lecture
2. **Lecture** of current topic
3. **HandsOn** to Software&Frameworks
4. **Exercise** execution

Materials

Script

https://github.com/marcelmittelstaedt/BigData/tree/master/script

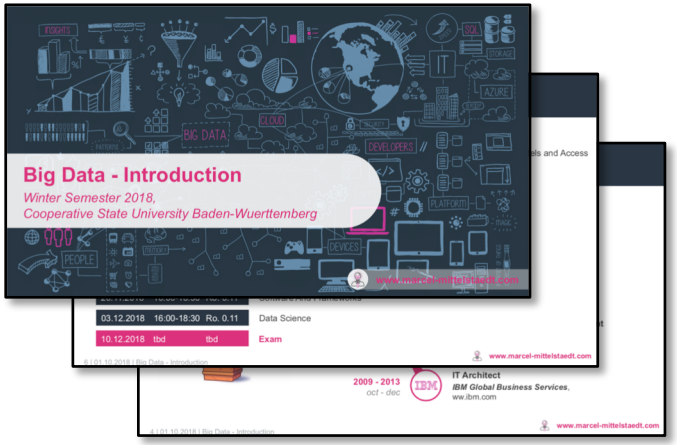
- citations and references are only in the script



Slides

https://github.com/marcelmittelstaedt/BigData/tree/master/slides

- no citations and references (see script)



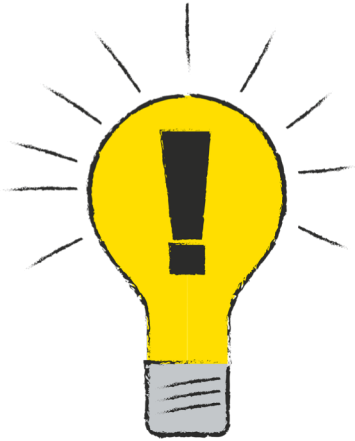
Materials

Exercises

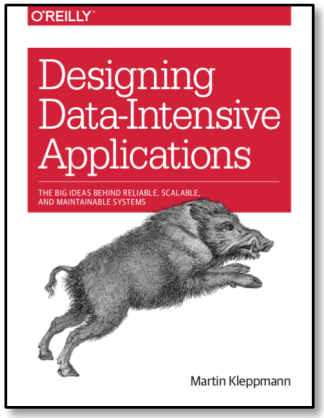
https://github.com/marcelmittelstaedt/BigData/tree/master/exercises

Solutions

https://github.com/marcelmittelstaedt/BigData/tree/master/solutions

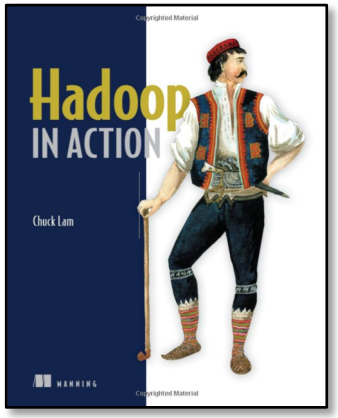


Literature / Course Books



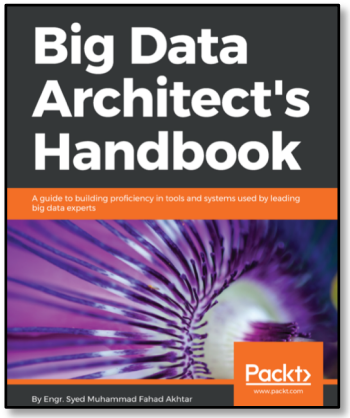
Designing Data-Intensive Applications
Martin Kleppmann

Amazon Link



Hadoop In Action
Chuck Lam

Amazon Link



Big Data Architect's Handbook
Syed Muhammad Fahad Akhtar

Amazon Link

PacktPub Link

Tech Ebooks from developer for developer

www.packtpub.com

Prerequisites

To efficiently participate:

- Background on and interest in **databases**
- **Basic knowledge** about and HandsOn **experience** with **Linux** (ideally Ubuntu)

For exam:

- **Attendance** on **lectures**
- **Doing** and **completion** of **exercises**

Feedback

Questions: anytime

This is a new and my first lecture ;)

Feedback: after lecture or via mail

If you find any mistakes or misspellings in the script or slides:

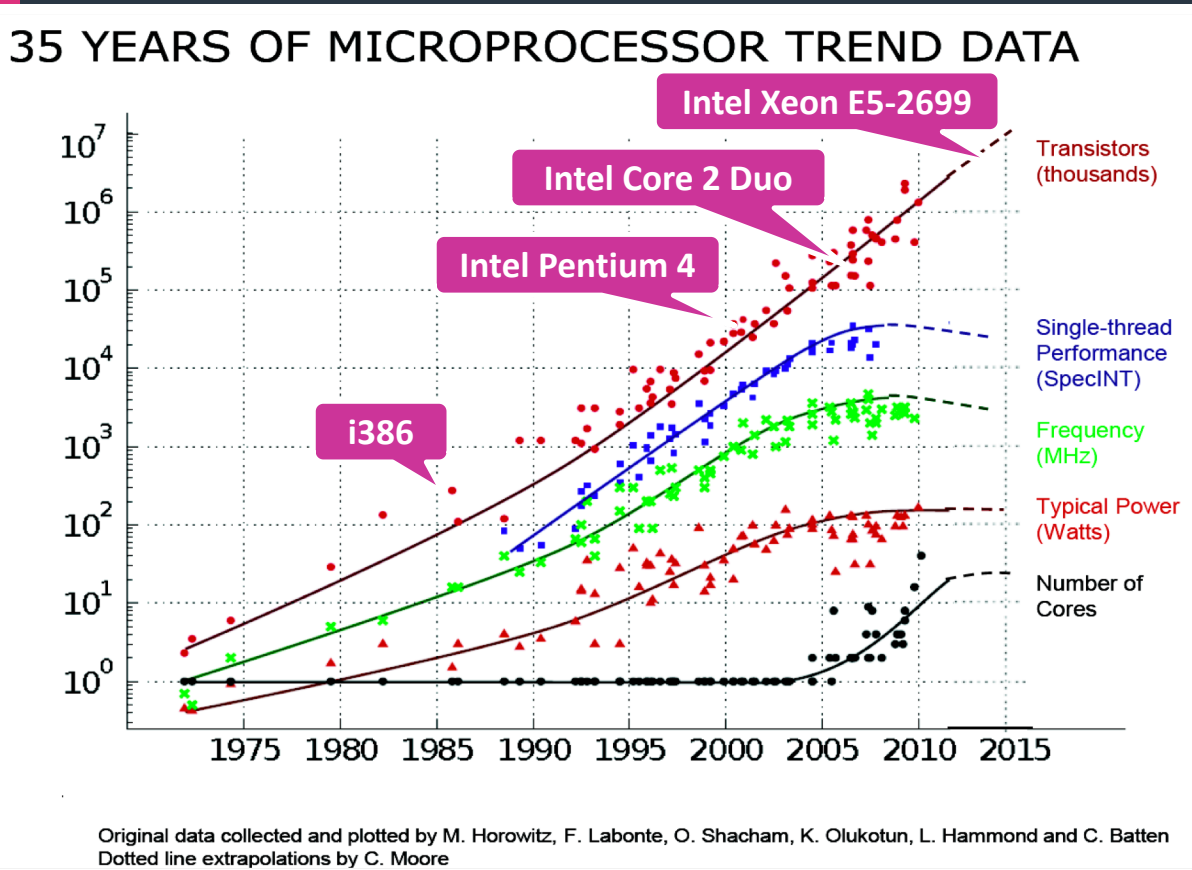
- Mail: contact@marcel-mittelstaedt.com
- commit a push request (git)

Questions

Introduction To Big Data

Distributed Computing, Big Data V's, Big Data and Consistency, Big Data in Business, Definition of Big Data, Big Data Science

Motivation - Paradigm Shift In Computing



- Frequency does not increase significantly anymore
- Watt does not increase anymore → ***power wall*** (temperature and power consumption)
- Transistor count still increases (***Moore's Law***)

Paradigm Shift:

- **Back-In-Time**: → *scale vertical* – optimize code for a **single thread**
- **Today**: *scale horizontal* – run code **in parallel**

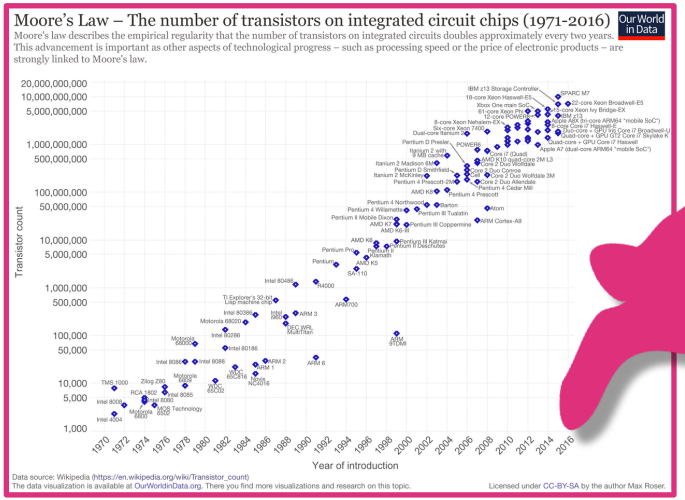
www.marcel-mittelstaedt.com

Motivation – **Definition Distributed Computing**

"***Distributed computing*** *is a field of computer science that studies distributed systems.*
A ***distributed system*** *is a system whose components are located on different networked computers, which then communicate and coordinate their tasks by passing messages to one other. The components interact with one other in order to achieve a common goal/task.*"

— https://en.wikipedia.org/wiki/Distributed_computing

Distributed Computing - **Bypassing Moore's Law**

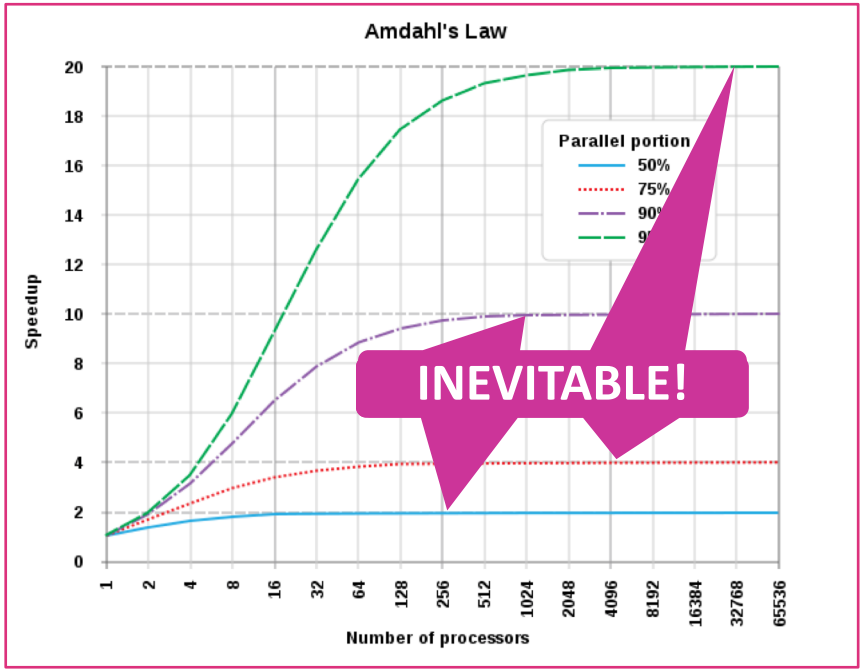


Moore's Law: *is the observation that the number of transistors in a dense integrated circuit doubles about every two years.*

Distributed systems: allow us to build data-systems with **any count of transistors**, just by adding further nodes to a cluster.

Distributed Computing – Amdahl's Law

Amdahl's Law: *is a formula to predict the theoretical speed-up when using multiple processors for distributed/parallel computing. The speedup is limited by the sequential part of the program.*



Formula:

$$S_{\text{latency}}(s) = \frac{1}{(1-p) + \frac{p}{s}}$$

$$\lim_{s \to \infty} S_{\text{latency}}(s) = \frac{1}{1-p}$$

s *= #Cores/Degree of parallelization*

p *= percentage of part of the code, which benefits from distributed/parallel execution*

Example Program:

→ takes **20 hours** on a **single core**
→ **95%** of the code can be **executed in parallel**
→ **8 cores** available

p = 0,95 *s = 8*

$S_{latency}(s) = 1 / ((1 - 0,95) + (0,95 / 8)) = \mathbf{5,9}$

$\lim S_{latency}(s) = 1 / (1 - 0,95) = \mathbf{20}$

→ *execution time can **never** be **less than one hour***

Distributed Technologies – **Small Scale**

Size: 0 - 10 Nodes
Hardware: “Commodity Hardware”

Space: < ~500 TB
Cores: < ~100 CPUs
Costs: cheap

Low-Cost
e.g. Raspberry Pi

Costs Per Node:
~30€

Desktop PC’s
e.g. anyone

Costs Per Node:
~1.000-3.000€

Distributed Technologies – **Medium Scale**

Size: 10 - 100 Nodes
Hardware: “Commodity Hardware” in terms of racks, usually with tuned parts (e.g. SAS instead of SATA drives, 8GBit Rack Uplink, 64-512GB RAM instead of 16-32GB etc.)

Space: < ~7,2 PB
Cores: < ~1.600-3.200 CPUs
Costs: inexpensive (compared to ***scale-up***)

Racks
e.g. HP ProLiant DL 380 Gen 10, 2RU

Costs Per Node:
~8.000-12.000€

Desktop PC’s
e.g. anyone

Costs Per Node:
~1.000-3.000€

Distributed Technologies – **Large Scale**

Size: 100 - X Nodes (e.g. 4.000 Nodes within Yahoo Hadoop Cluster in 2008)

Hardware: "Commodity Hardware" in terms of racks, usually with tuned parts (e.g. SAS instead of SATA drives, 8GBit Rack Uplink, 64-512GB RAM instead of 16-32GB etc.)

Space: > ~10 PB
Cores: > ~3.200 CPUs
Costs: still rather cheap than pricey (compared to ***scale-up***)

Yahooo Hadoop Cluster, 4.000 nodes, 2008

Racks
40 Nodes / Rack
Rack Uplink 8Gbit

Costs Per Node:
~8.000-15.000€

Distributed Technologies – **Supercomputer**

Name	Standort	TeraFLOPS	Konfiguration	Energiebedarf	Zweck
Summit	Oak Ridge National Laboratory (Tennessee, USA)	122.300,00	9.216 POWER9 CPUs (22 Kerne, 3,1 GHz), 27,648 Nvidia Tesla V100 GPUs	15.000 kW	Physikalische Berechnungen
Sunway TaihuLight	National Supercomputing Center, Wuxi, Jiangsu	93.014,60	40.960 Sunway SW26010 (260 Kerne, 1,45 GHz), 1,31 PB RAM, 40 Serverschränke mit jeweils 4 × 256 Nodes, insgesamt 10.649.600 Kerne	15.370 kW	Wissenschaftliche und kommerzielle Anwendungen
Sierra[6]	Lawrence Livermore National Laboratory (Kalifornien, USA)	71.600,00	IBM Power9 (22 Kerne, 3,1 GHz) 1,5 PB RAM	~12.000 kW	physikalische Berechnungen (z. B. Simulation von Kernwaffentests)
Tianhe-2[7]	National University for Defense Technology, Changsha, China finaler Standort: National Supercomputer Center (Guangzhou, China)	33.862,70 aufgerüstet auf 61.400,00	32.000 Intel Xeon E5-2692 CPUs (Ivy Bridge, 12 Kerne, 2,2 GHz) + 48.000 Intel Xeon Phi 31S1P Co-Prozessoren (57 Kerne, 1,1 GHz), 1,4 PB RAM	17.808 kW	Chemische und physikalische Berechnungen (z. B. Untersuchungen von Erdöl und Flugzeugentwicklung)
Piz Daint	Swiss National Supercomputing Centre (CSCS) (Schweiz)	19.590,00	Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect, NVIDIA Tesla P100, Cray Inc. (361.760 Kerne)	2.272 kW	wissenschaftliche und kommerzielle Anwendungen
Titan	Oak Ridge National Laboratory (Tennessee, USA)	17.590,00	Cray XK7, 18.688 AMD Opteron 6274 CPUs (16 Kerne, 2,20 GHz) + 18.688 Nvidia Tesla K20 GPGPUs, 693,5 TB RAM	8.209 kW	Physikalische Berechnungen
Sequoia[8]	Lawrence Livermore National Laboratory (Kalifornien, USA)	17.173,20	IBM BlueGene/Q, 98.304 Power BQC-Prozessoren (16 Kerne, 1,60 GHz), 1,6 PB RAM	7.890 kW	Simulation von Kernwaffentests
	Advanced Institute for		88.128 SPARC64-VIII 8-		

https://de.wikipedia.org/wiki/Supercomputer

"Summit",
Oak Ridge National Laboratory, Tennessee, USA

Cores: 202.752
TeraFLOPS: 122.300,00
Energy: 15.000 kW

Summit Supercomputer, Oak Ridge National Laboratory

Distributed Technologies – **Large Cloud Cluster**

Size: 200 Nodes

Cores: 12.800 vCPUs / 6.400 CPUs

** **DISCLAIMER:** Completely ignoring storage, network, virtualization and sharing of CPUs etc.*

AWS:

Instance: *m4.16xlarge*
vCPUs *per Instance:* 64
RAM *per Instance:* 256
Costs per hour *per Instance:* *3,20 USD*

Costs for ***Large Cluster*** *per hour:*
*200 * 3,20 USD =* ***640,00 USD***

https://aws.amazon.com/de/ec2/pricing/on-demand/

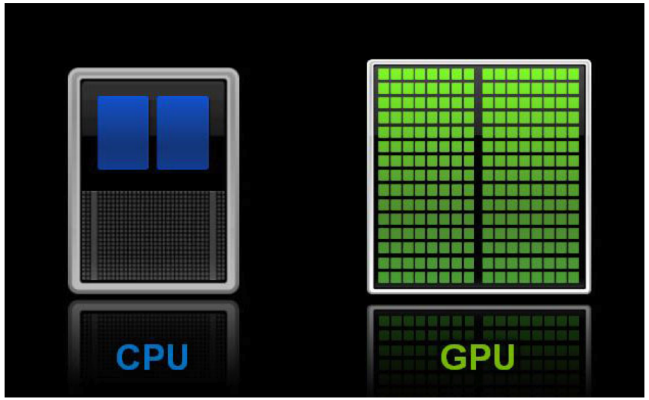
Google Cloud:

Instance: *n1.standard-64*
vCPUs *per Instance:* 64
RAM *per Instance:* 240
Costs per hour *per Instance:* *3,04 USD*

Costs for ***Large Cluster*** *per hour:*
*200 * 3,04 USD =* ***608,00 USD***

https://cloud.google.com/compute/pricing

Distributed Technologies – **GPU Computing**

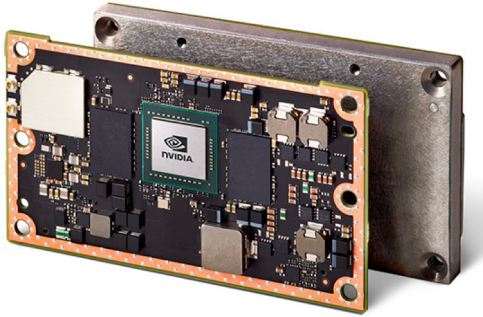


- advanced GPUs **initially developed for** (3D) **gaming**

→ **now also used for** (distributable) **computational workloads** (e.g. Big Data, Machine Learning)

Cores:	**2-18** *(e.g. 12 Cores for Intel Xeon E5-2690)*	**1.000-4.000** *(e.g. 3.584 CUDA for GTX 1080 Ti*
Freq.:	**2-4 Ghz** *(e.g. 2,6-3,5 Ghz for Intel Xeon E5-2690)*	**1.000-2.000 MHz** *(e.g. 1.582MHz for GTX 1080 Ti)*

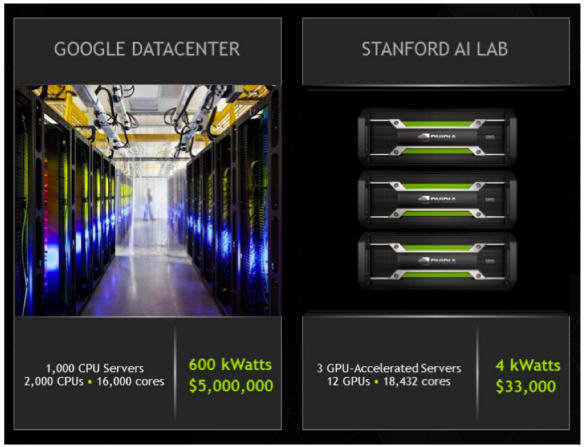
Nvidia Jetson TX2
Cores: *256 (CUDA)*
RAM *per Instance:* *8*



https://www.nvidia.com/de-de/autonomous-machines/embedded-systems-dev-kits-modules/

Distributed Technologies – GPU Cluster

On-Premise:



https://cs.stanford.edu/csdcf/sail-compute-cluster

https://www.slideshare.net/papisdotio/introduction-to-multi-gpu-deep-learning-with-digits-2-mike-wang

Cloud:

AWS:

Amazon EC2 Elastic GPUs – Preise

Mit Amazon EC2 Elastic GPUs zahlen Sie ausschließlich für die tatsächliche Nutzung. Die Preise für Elastic GPUs sind unten aufgeführt.

USA Ost (Ohio) und USA Ost (Nord Virginia)

Größe der Elastic GPU	GPU-Arbeitsspeicher	Preis
eg1.medium	1 GiB	0,050 USD/Stunde
eg1.large	2 GiB	0,100 USD/Stunde
eg1.xlarge	4 GiB	0,200 USD/Stunde
eg1.2xlarge	8 GiB	0,400 USD/Stunde

https://aws.amazon.com/de/ec2/elastic-gpus/pricing/

Google Cloud:

GPU-Modelle für Compute Engine

GPU-Modell	GPUs	GPU-Speicher	Verfügbare vCPUs	Verfügbarer Speicher
NVIDIA® Tesla® V100	1 GPU	16 GB HBM2	1–12 vCPUs	1–78 GB
	8 GPUs	128 GB HBM2	1–96 vCPUs	1–624 GB
NVIDIA® Tesla® P100	1 GPU	16 GB HBM2	1–16 vCPUs	1–104 GB
	2 GPUs	32 GB HBM2	1–32 vCPUs	1–208 GB
	4 GPUs	64 GB HBM2	1–64 vCPUs (us-east1-c, europe-west1-d, europe-west1-b)	1–208 GB (us-east1-c, europe-west1-d, europe-west1-b)
			1–96 vCPUs	1–624 GB

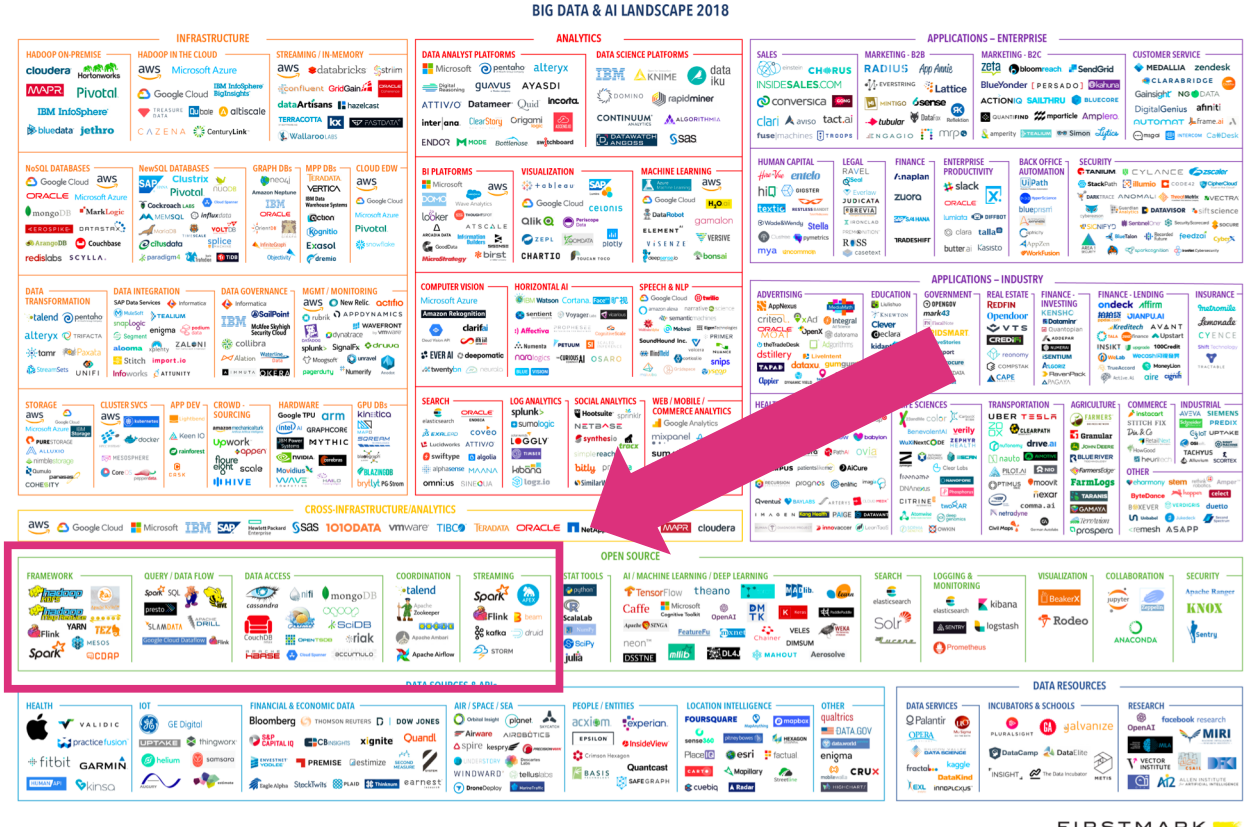
https://cloud.google.com/compute/docs/gpus/

Distributed Technologies – Big Data

Distributed Processing (Computing)

Distributed Data (Storaging)

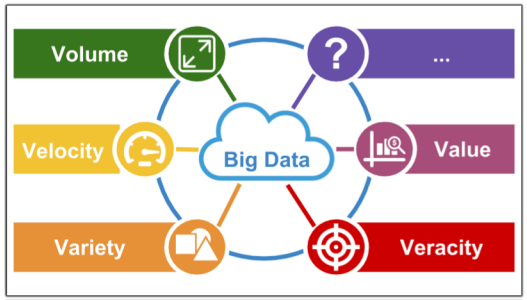
Distributed Technologies – Big Data Landscape



Big Data V's

- Introduced by **Gartner** in **2011**
- **Key-Characteristics** of Big Data
- **More V's** have established over time

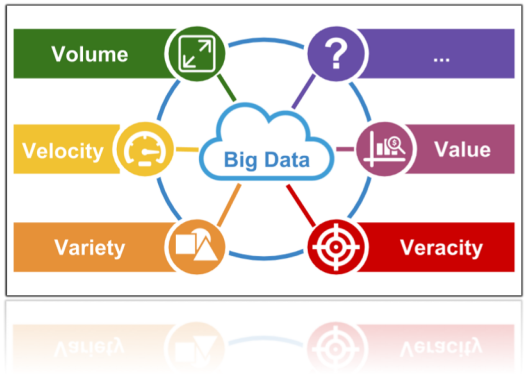
Big Data V's - **Volume**



- **Wikipedia:** in 2014 the english part of Wikipedia and Wikipedia Commons had a size of around **23TB**
- **Spotify:** in 2013 data generated by users of spotify was about **1,5TB/day** → **534TB/year**
- **Google:** search index size currently about **100PB**

- **Rule Of Thumb:** data that can easily be handled (e.g. processed, saved or queried) within a traditional database (e.g. PostgreSQL, Informix, DB2, Oracle etc.) and belonging datawarehouse, like a few gigabytes per week or month, is not Big Data.

Big Data V's - **Velocity**

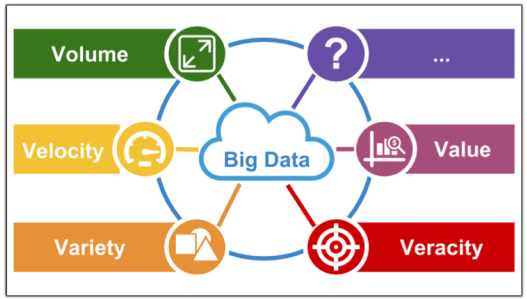


- **Facebook:** in 2012 **every 60 seconds:**
  510.000 comments,
  293.000 status updates and
  136.000 photos
  had been posted on average.

- **Twitter:** in 2013 usually **500 million tweets** were posted each day on Twitter, which means **5.7000 Tweets per second** on average.

- **Google:** According to Internet Live Stats, Google is handling around **6 billion search requests** per day right now, which means **70.000 per second**.

Big Data V's - **Variety**



Structured Data:

- **plain** data structures with **fixed** attributes/columns
- **specific data types** and **defined domain**
- related but different kinds of data records can easily be linked (primary/foreign keys)
- can easily (almost directly) be stored into a traditio nal relational database

Semi-Structured Data:

- **do not** (easily) **fit** into the **constraints** of the **relational data model**
- possible (with additional effort) to transfer into relational world, but violating relational data model constraints
- no need to worry about *object-relational impedance mismatch*
- vulnerable to *garbage in – garbage out*

Big Data V's - **Variety**

Semi-Structured Data:

- **highly-nested** and **hierarchical** data structures
- Examples:
  - **JSON** documents (e.g. provided by the Facebook Graph API or the Twitter API)
  - **XML** documents (e.g. provided by the Google Maps Geocoding API or the OpenWeatherMap Weather API)
  - **HTML** documents, for instance if your developing a web crawler (e.g using Scrapy for data gathering or Selenium for testing purposes)

Big Data V's - **Variety**

Unstructured Data:

- everything that is **neither structured** or **semi-structured**
- Examples:
  - **Natural Language:**
    - Social Media Posts (e.g. Facebook and Twitter),
    - Ratings (e.g. on Amazon, IMDB or other platforms) or
    - plain text (e.g. Wikipedia articles, product information on Amazon or any e-commerce shop)

    → you may want to analyze by *natural language processing* with the purpose of calculating a **sentiment** (*What do customers think about your product?*) or the meaning and truthfulness of a rating (*Is the revi ew positive or negative*? Is it fake?).

Big Data V's - **Variety**

Unstructured Data:

- Examples:
  - **Geographical Data**:
    - GPS,
    - Radar,
    - sonar data or images

    → with the purpose of analyzing meteorological, vecicular or seismic behaviour, e.g. for the purpose of predicting the future.

  - **Photos and Videos**:

    → for instance of traffic and surveillance cameras to analyze and optimize traffic flow or for the sake of security.

Big Data V's - **Variety**

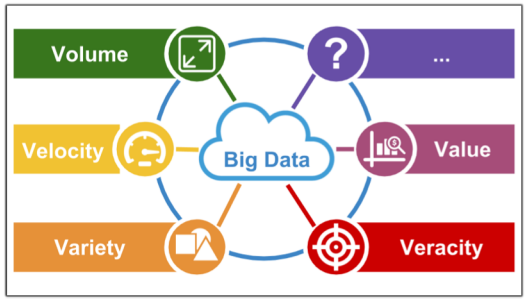
Unstructured Data:

- Examples:
  - **Scientific Data**:
    - physic measuring and sensor data,
    - seismic imagery or
    - athmospheric data

    → to analyze and optimize a physics experiment.

    For instance CERN makes use of Hadoop for unstructured sensor data, to unlock insights from machine data.

Big Data V's - **Veracity**



- ***veracity*** is about the **trustworthiness** and **accuracy** of a certain dataset, as data usually ivolves some uncertainty and ambiguities
- not just about the quality of the data itself, but also tasks of **quality assurance** and data cleansing, like removing **bias**, **superfluous**, **redundant** or just **duplicate records** or attributes.

- **trustworthiness** is highly vulnerable to the **volatility of the data** (which you usually cannot control), as a lot of data sources and related records frequently change in their lifetime

→ an account balance quickly changes over time (due to ongoing transactions)

→ people which have liked a topic, post or site on a social media platform may dislike it the next day

Big Data V's - **Veracity**

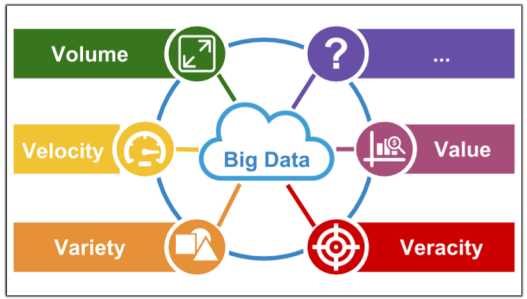
- even if you have a fully-fletched environment which **ensures data quality in its own limits**, you cannot be sure the data you are processing is 100% accurate, as their could already be a **mistake in the source system**

- Big Data data-system should always try to **achieve** the **best data quality possible**, even it is seen a lot: *Garbage In - Garbage Out* is not acceptable.

→ Garbage Data + Perfect ETL or Data Science Model = **Garbage Results**

→ Perfect Data + Messy ETL or Data Science Model = **Garbage Results**

Big Data V's - **Value**



- *volume*, *velocity*, *variety* and *veracity* - all of them are meaningless if you dont derive **business value** from your data, but they are also significant **enabler** and **success factors** for the ***value*** **of your data**
- **value** can probably be found in any kind of data, but the challenge is to **pick the right data** (business case) and **use it the right way**

- Volume and **Value**: The more data you collect, the more insights you can probably gather and the more informed your insights and decisions may be.

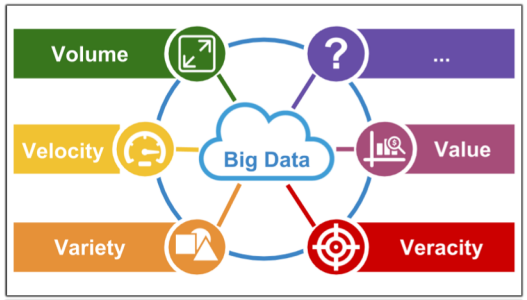
- Velocity and **Value:** The faster data is collected and available for analysis processes, the faster it can be used for decision-making processes.

Big Data V's - **Value**

- Variety and **Value**: The more data sources you connect to your data-system, the more insights you can potentially create.
  You will no longer rely an a single data source for a certain business object (like a client), which overall increases data quality and depth. Enabling you, for instance to build customer journeys, calculate a CLV, and in this way improve engagement and retention. At the end the value of your data increases.

- Veracity and **Value:** The more trustworthy and accurate your data is, more business critical desicisions and applications will rely on it and in this way the revenue of your company will probably increase as well as the value of your data.

Big Data V's – Other V's



- **Volatility:** How long data is valid and how long it should be stored. Retention requirements? Data Protection requirements?

- **Variability:** Format, schema or semantic changes.

- **Validity:** Like veracity, is the data correct and accurate for the intended use? Clearly valid data is key to making the right decisions.

- **Venue:** Different location require different access and work methods.

- **…**

Big Data Challenges – **ACID**

- Gives following **4 guarantees**:

- **Atomic:** all operations in a transaction succeed or every operation is rolled back

- **Consistent:** Before and after a transaction, the data-system is structurally sound.

- **Isolated:** Transactions do not contend with one another. Contentious access to data is moderated by the database so that transactions appear to run sequentially.

- **Durable:** The results of applying a transaction are permanent, even in the presence of failures.

→ **Ensures safe and reliable data storage and processing**
→ **Requires a lot of additonal workload like locks, moderated data access, failover protection**

Big Data Challenges – **BASE**

- Not (fully) ACID/ soft version of ACID

Basic Avaiability: the data-system should work an be available most of the time.

→ availability less than 100%

Soft-State: data stores do not have to be write-consistent, nor do different replicas have to be mutually consistent all the time.

→ Stored data may be inconsistent, but data store can derive consistent states.

Eventual Consistency: Stores exhibit consistency at some point later in time. For instance at at read time.

→ usually consistent within seconds/milliseconds

→ eventual consistency does not mean no-consistency

Big Data Challenges – **ACID** vs **BASE**

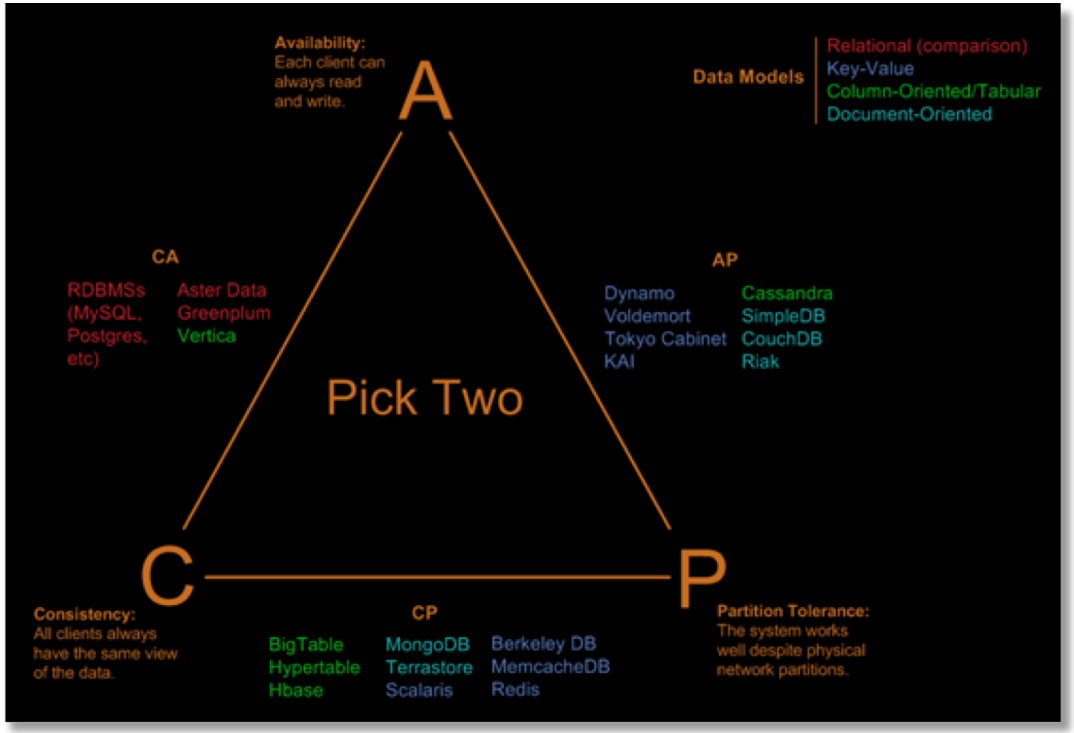
ACID	BASE
Strong consistency	Weak consistency
Isolation	Availability first/Last write wins
Focus on „commit“	„Best effort“
Trnsactions	Programmer managed
Nested transaction	Approximate answer
Conservative (pessimistic)	Aggressive (optimistic)
Robust database	Simple database
Simple application code	Complex application code

Big Data Challenges – **CAP** (Eric Brewer)

- **Availability**:
  Every request receives a (*non-error*) **response** - without guarantee that it contains the most recent write

  → e.g. **high load**,
  → server **failures** or
  → query **congestion**
  may **deny service availability**

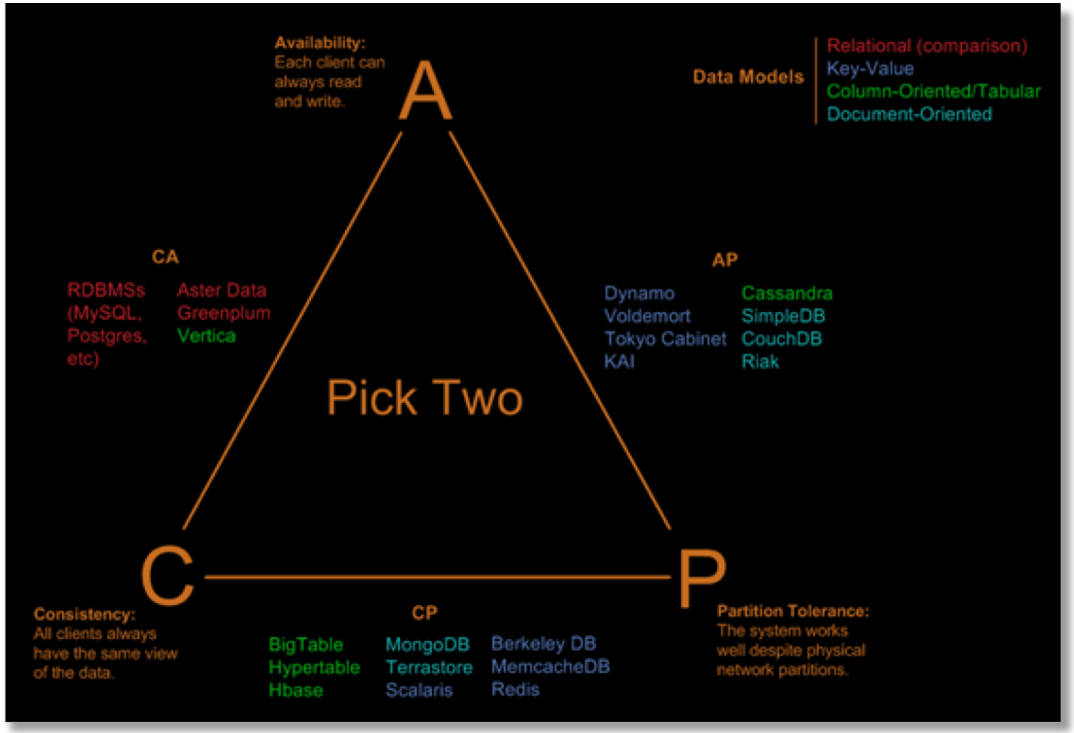
Big Data Challenges - CAP



- **Consistency**:
  Every read/all clients receive the **most recent write** or an error.

- e.g. ***ACID consistency*** (consistent transactions)
- but on a **cluster-wide** **not node-wi de** consistency

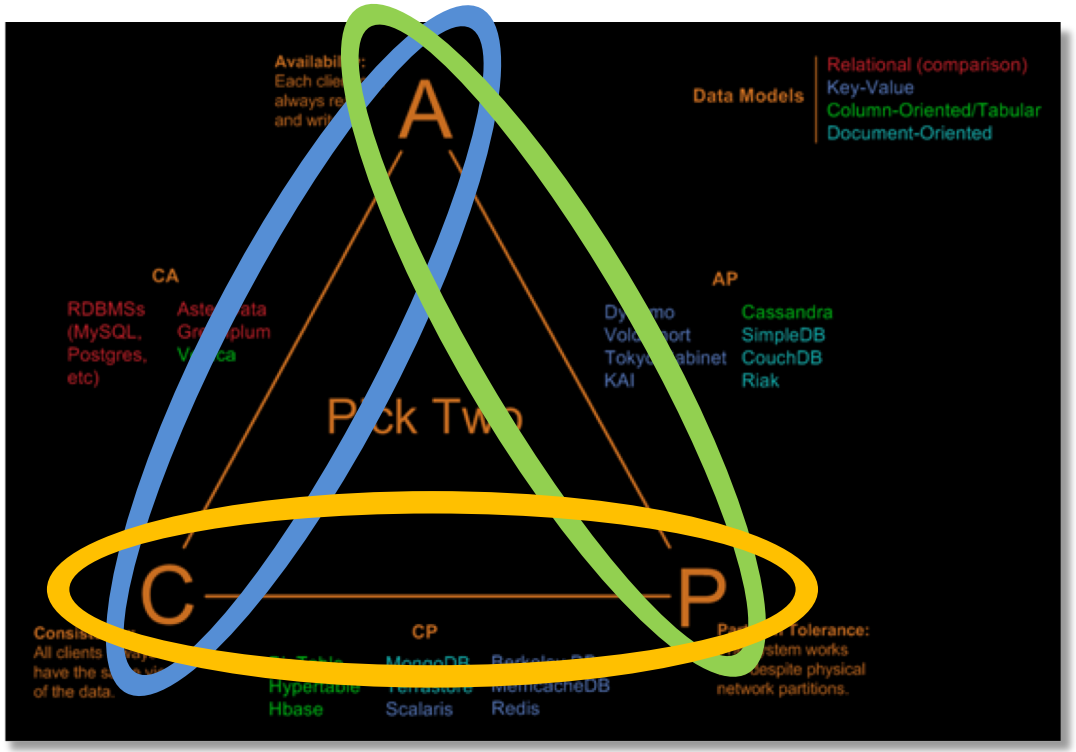
Big Data Challenges - CAP



- **Partition Tolerance**:
  The **system continues to operate** despite an **arbitrary number of messages being dropped** (or delayed) by the network in between nodes.

→ only total network failure might cause the data-system to respond incorrectly.

Big Data Challenges – CAP Examples



- Usually achieved by **not sharding**.
- If **server** or **network failures** occure, return **whatever is possible**.
- If **server** or **network failures** occure, try to recover and **deny availability** until state is consis tent again.
- **Traditional databases** usually achieve all 3 as they are not distributed.

Big Data Roles - **Data Engineer / Architect**

Study Direction: Computer Science (databases, software engineering, distributed systems and processing, real-time processing)

(Programming) Languages: Java, Scala, Python, Bash, if necessary ETL Tools (e.g. Pentaho Data Integration, Informatica, DataStage, Talend)

Key-Skills:

- data modeling
- ETL and dataflow architecture and design
- data pipeline/workflow design and management
- stream- and batch-processing
- understand the concept, performance factors and limitations of distributed data-systems

Big Data Roles - **Data Engineer / Architect**

Key-Skills:

– alignment with larger organizations goals and strategy as well as corporate guidelines (e.g. data protection, information security, lice nse and toolset policies)

Tasks:

– develop ETL jobs, data- and workflows
– develop and define architecture of a data-system,
– integration of data-sources
– alignment with larger organizations goals and strategy as well as corporate guidelines (e.g. data protection, information security, license and toolset policies)

Big Data Roles - **Data Ops**

Study Direction: Computer Science (databases, software engineering, distributed systems and processing, real-time processing)

(Programming) Languages: Python, Bash

Key-Skills:

– strong communication and collaboration skills
– broad knowledge of Ops Tools, best practices and automation
– frequently refreshs his knowledge as Ops tools crop up continuosly
– Continuos Integration and Deployment Automation (e.g. Jenkins, Gitlab CI),

Big Data Roles - **Data Ops**

Key-Skills:

- Infrastructure Automation (e.g. Foreman, Puppet, Ansible),
- Containerization (e.g. Docker, LXD etc.),
- Orchestration (e.g. Kubernetes, Mesos, Swarm),
- Cloud (e.g. AWS, Google Cloud Platform, Azure, OpenStack)
- Source Control (e.g. Git, Bitbucket, SVN)

Tasks:

- installs, configures, upgrades, scales and operates infra-structure and applications
- monitoring and management of infrastructure and applications
- provides and manages tools for continues delivery
- incident and problem management
- interface to developers (requirements for infrastructure, support developers needs and productive application operation)

Big Data Roles - **Data Scientist**

Study Direction: Statistics, Data Science, Mathematics, Computer Science (Java, Python, R, SQL, Machine Learning)

(Programming) Languages: Python, R, Matlab, Bash, Java, Scala

Key-Skills:

- strong communication and collaboration skills
- business domain expertise
- deep understanding of the data he is working with
- broad experience in using programming languages and tools for statistical purposes (Python, R, SQL, MapReduce, Spark, SparkML, Pandas, dataframes)

Big Data Roles - **Data Scientist**

Key-Skills:

– knowledge of a variety of machine learning techniques (e.g. clustering, regression, decision trees, neural networks, properties of distr ibution)
– strong skills in data visualization (e.g. seaborn, Matplotlib, ggplot)

Tasks:

– identify and evaluate data-analytics problems that offer the greatest opportunities and/or business value
– determining the right data sets and variables
– cleansing and validation of data to ensure completeness, accuracy and uniformity
– developing and applying models and algorithms on data
– analyze data to identify patterns, risks and trends
– commmunication with and visualization for (business) stakeholder

Big Data Roles - **Salaries**

Big Data Ops

€€€ ?

Big Data Engineer

€€€ ?

Data Scientist

€€€ ?

Big Data Roles - **Salaries**

Big Data Ops

45.000€ - 70.000€

Big Data Engineer

50.000€ - 80.000€

Data Scientist

50.000€ - 100.000€

https://www.glassdoor.de/Gehälter

Big Data Definition

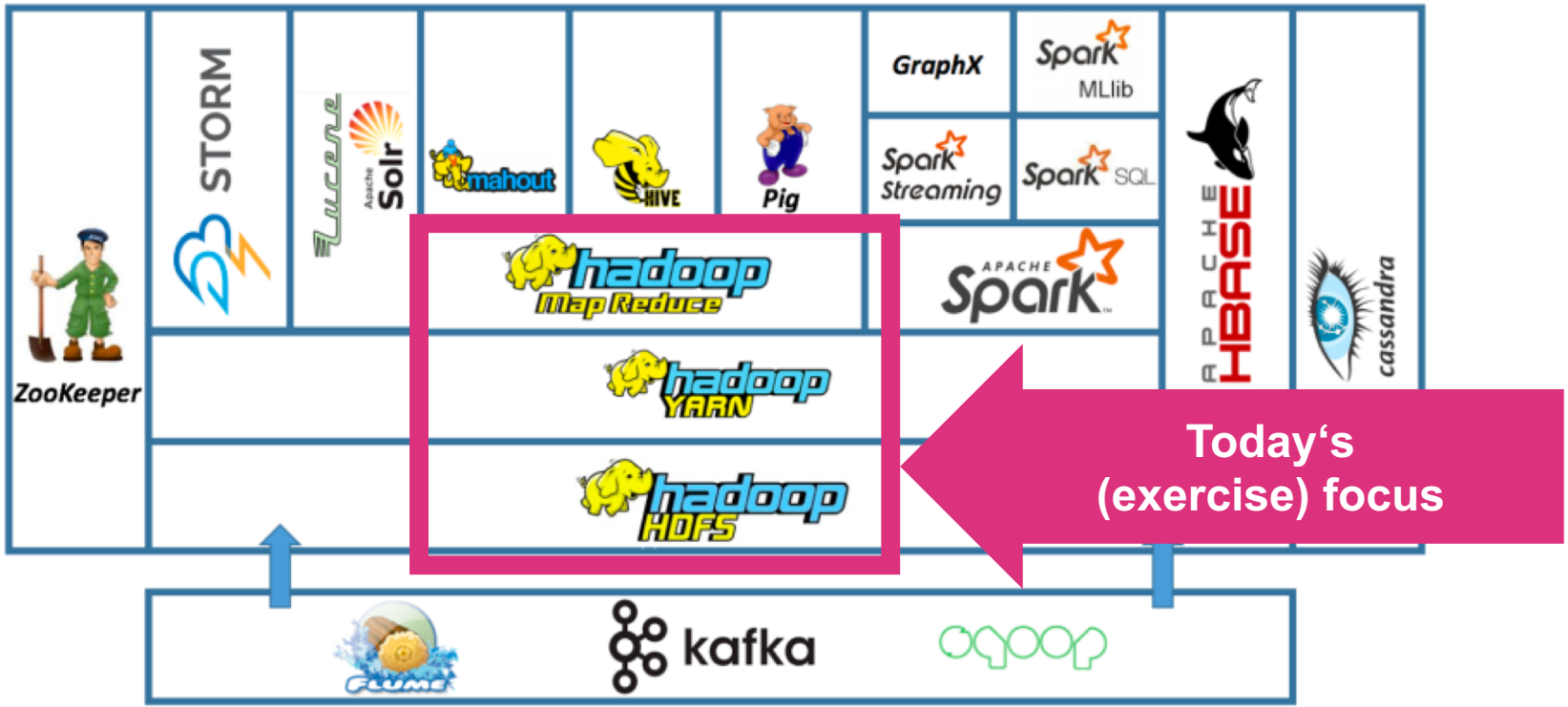
*"**Big Data** is a term for **datasets** that are **too large**, **too fast** and/or **too complex** for traditional databases and ETL tools to handle.*

*The term **Big Data** also refers to all related **algorithms**, **software**, **infrastructure** and **tools** used to be able to cope with such datasets."*

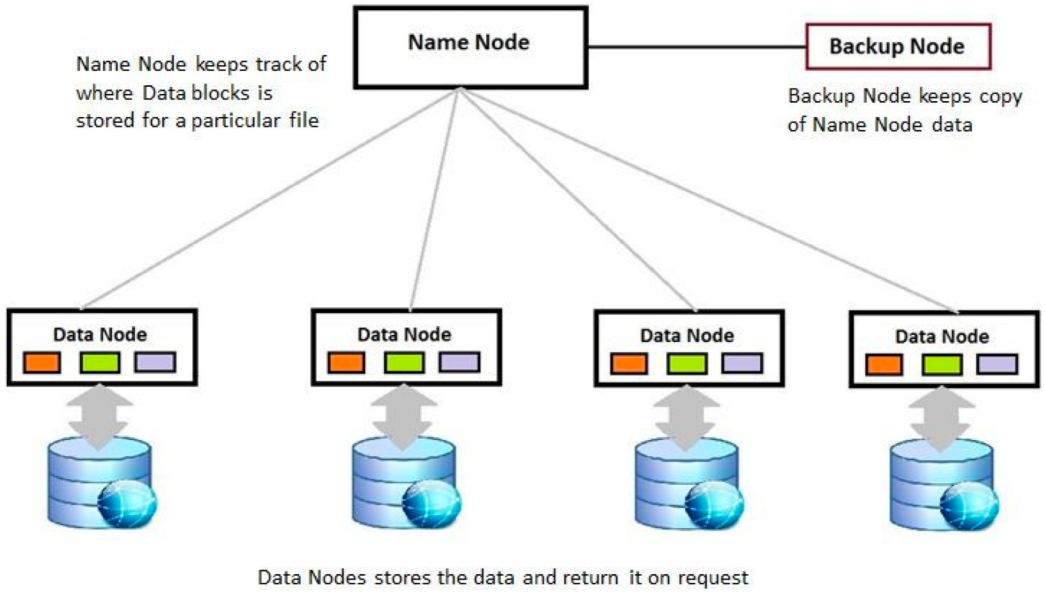
HandsOn - Hadoop

Quick Introduction To The Hadoop Ecosystem,
HDFS, Yarn and MapReduce

The Hadoop Ecosystem



The HDFS

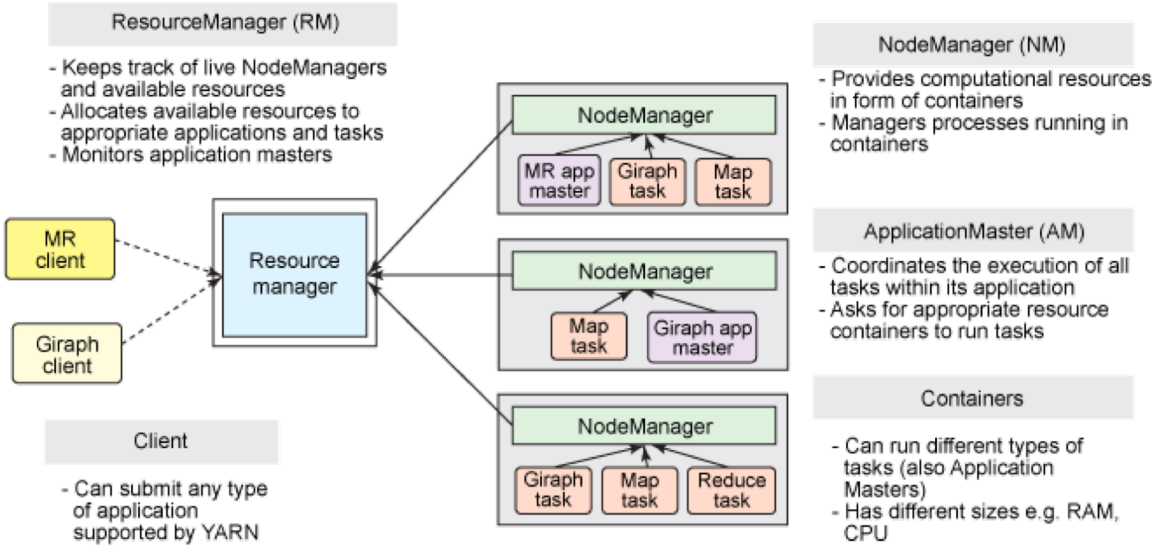


- ***H**adoop **D**istributed **F**iles**s**ystem*
- Highly available and distributed filesystem
- Can easily be scaled to a cluster of hundreds or even thousands of nodes and in this way **Petabytes of data**

YARN

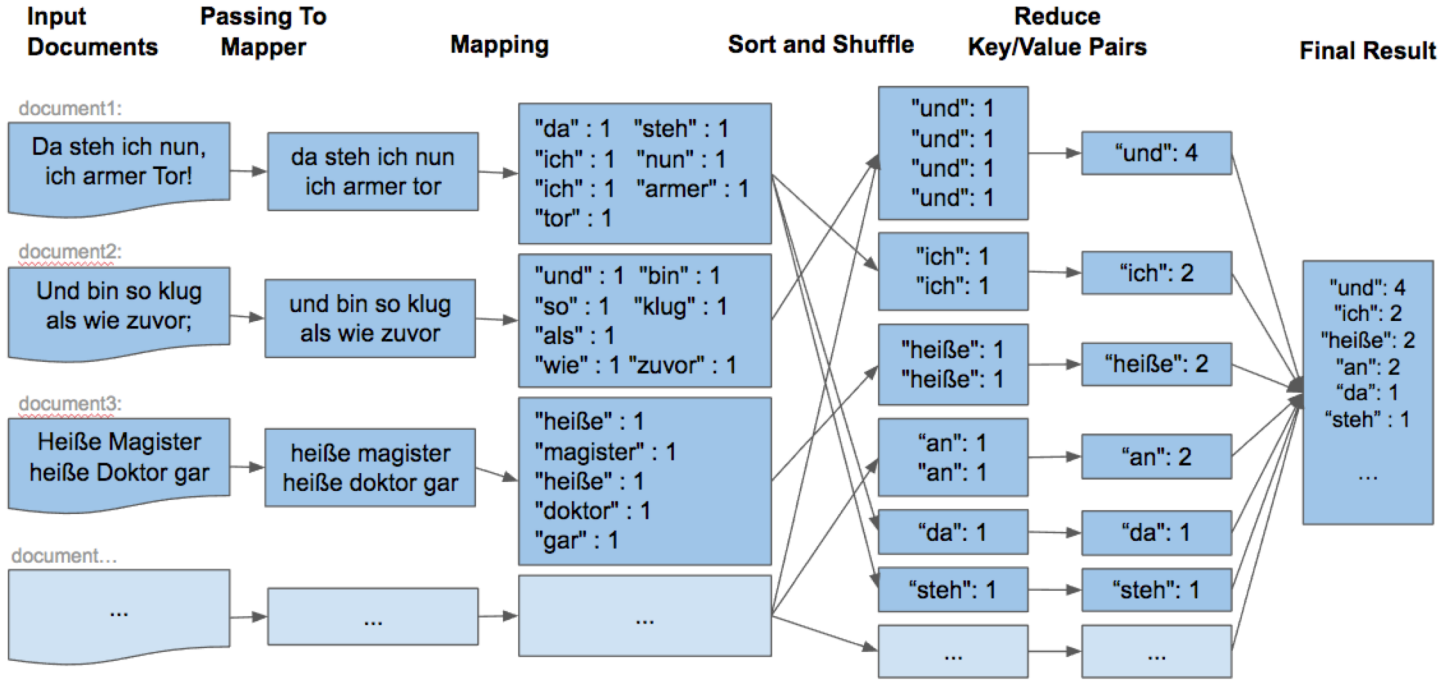
- ***Y**et **a**nother **R**essource **N**egotiator*

- Hadoop Ressource Manager (e.g. CPU, Memory, Job Scheduling, Preparation, Execution and Priorization)



MapReduce

- Programming paradigm for processing large datasets in parallel on a distributed cluster



Exercises Preparation

Setup Hadoop, HDFS, Yarn

Get Ubuntu VM ready

1. Get Virtual Box:

Mac OSX:

```
wget https://download.virtualbox.org/virtualbox/5.2.18/VirtualBox-5.2.18-124319-OSX.dmg
```

Windows:

```
wget https://download.virtualbox.org/virtualbox/5.2.18/VirtualBox-5.2.18-124319-Win.exe
```

Linux: No need, you‘re lucky

2. Get Latest Ubuntu LTS Image File (18.04.):

```
wget http://releases.ubuntu.com/18.04.1/ubuntu-18.04.1-desktop-amd64.iso
```

Get Ubuntu VM ready

3. Install Virtual Box.

4. Create and install Ubuntu VM:

CPU Cores:	2++
RAM:	4096MB++
Disk Space:	20GB++ *(VDI, dynamic allocation)*

5. Insert and Install *Vbox_Guest_Additions.iso*

6. Restart VM.

Install and Setup Java

1. Install OpenJDK (JDK 8):

```
sudo apt-get install openjdk-8-jdk
```

2. Verify installation:

```
marcel@VirtualBox:~$ java -version
openjdk version "1.8.0_181"
OpenJDK Runtime Environment (build 1.8.0_181-8u181-b13-0ubuntu0.18.04.1-b13)
OpenJDK 64-Bit Server VM (build 25.181-b13, mixed mode)
```

2. SET *JAVA_HOME* and *JRE_HOME*:

```
sudo vi /etc/environment
```

```
JAVA_HOME="/usr/lib/jvm/java-8-openjdk-amd64"
JRE_HOME="/usr/lib/jvm/java-8-openjdk-amd64/jre"
```


Setup Hadoop User

1. Create User:

```
adduser hadoop
passwd hadoop
```

2. Switch To User:

```
su - hadoop
```


Setup SSH (needed by Hadoop components)

1. Install SSH and PDSH:

```
sudo apt-get install ssh
sudo apt-get install pdsh
```

2. Create Private/Public Keypair (*without passphrase*):

```
ssh-keygen -t rsa -b 4096
```

3. Add Public Key To Authorized Keys file (to enable passwordless ssh login)

```
cat .ssh/id_rsa.pub >> .ssh/authorized_keys
chmod 0600 ~/.ssh/authorized_keys
```


Setup SSH (needed by Hadoop components)

4. Check If SSH Is Working

```
hadoop@marcel-VirtualBox:~$ ssh localhost
Welcome to Ubuntu 18.04.1 LTS (GNU/Linux 4.15.0-29-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage


 * Canonical Livepatch is available for installation.
   - Reduce system reboots and improve kernel security. Activate at:
     https://ubuntu.com/livepatch

175 packages can be updated.
65 updates are security updates.

Last login: Sat Sep 29 22:56:15 2018 from 127.0.0.1
hadoop@marcel-VirtualBox:~$ exit
logout
Connection to localhost closed.
hadoop@marcel-VirtualBox:~$
```


Install Hadoop

1. Download Hadoop (v3.1.1):

```
wget http://apache.cs.utah.edu/hadoop/common/hadoop-3.1.1/hadoop-3.1.1.tar.gz
```

2. Extract Binaries:

```
tar -xzf hadoop-3.1.1.tar.gz
```

3. Move Binaries:

```
sudo mv hadoop-3.1.1 /home/hadoop/hadoop
```


Configure Hadoop

1. Set Up **UNIX** Environment Variables

```
vi .bashrc
```

```
export HADOOP_HOME=/home/hadoop/hadoop
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
```

```
source .bashrc
```

Configure Hadoop

2. Add **Hadoop** Environment Variables *(hadoop-env.sh)*

```
vi /home/hadoop/hadoop/etc/hadoop/hadoop-env.sh
```

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
```

Configure Hadoop

3. Set Up **CORE** Variables *(core-site.xml)*

```
vi /home/hadoop/hadoop/etc/hadoop/core-site.xml
```

```
<configuration>
     <property>
          <name>fs.default.name</name>
          <value>hdfs://localhost:9000</value>
     </property>
</configuration>
```


Configure Hadoop

4. Set Up **HDFS** Variables *(hdfs-site.xml)*

```
vi /home/hadoop/hadoop/etc/hadoop/hdfs-site.xml
```

```
<configuration>
      <property>
             <name>dfs.replication</name>
             <value>1</value>
      </property>

      <property>
             <name>dfs.name.dir</name>
             <value>file:///home/hadoop/hadoopdata/hdfs/namenode</value>
      </property>

      <property>
            <name>dfs.data.dir</name>
            <value>file:///home/hadoop/hadoopdata/hdfs/datanode</value>
      </property>
</configuration>
```

Configure Hadoop

5. Set Up **MapReduce** Variables *(mapred-site.xml)*

```
vi /home/hadoop/hadoop/etc/hadoop/mapred-site.xml
```

```
<configuration>
      <property>
            <name>mapreduce.framework.name</name>
            <value>yarn</value>
      </property>
      <property>
            <name>yarn.app.mapreduce.am.env</name>
            <value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>
      </property>
      <property>
            <name>mapreduce.map.env</name>
            <value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>
      </property>
      <property>
            <name>mapreduce.reduce.env</name>
            <value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>
      </property>
</configuration>
```


Configure Hadoop

6. Set Up **YARN** Variables *(yarn-site.xml)*

```
vi /home/hadoop/hadoop/etc/hadoop/yarn-site.xml
```

```
<configuration>
          <property>
                    <name>yarn.nodemanager.aux-services</name>
                    <value>mapreduce_shuffle</value>
          </property>
</configuration>
```

Configure Hadoop

7. Clear HDFS

```
hdfs namenode -format
```

8. Start HDFS:

```
start-dfs.sh
Maybe requires: export PDSH_RCMD_TYPE=ssh
```

9. Start YARN:

```
start-yarn.sh
```


Check Hadoop/HDFS

10. Run Admin Status Report

```
hdfs dfsadmin -report
```

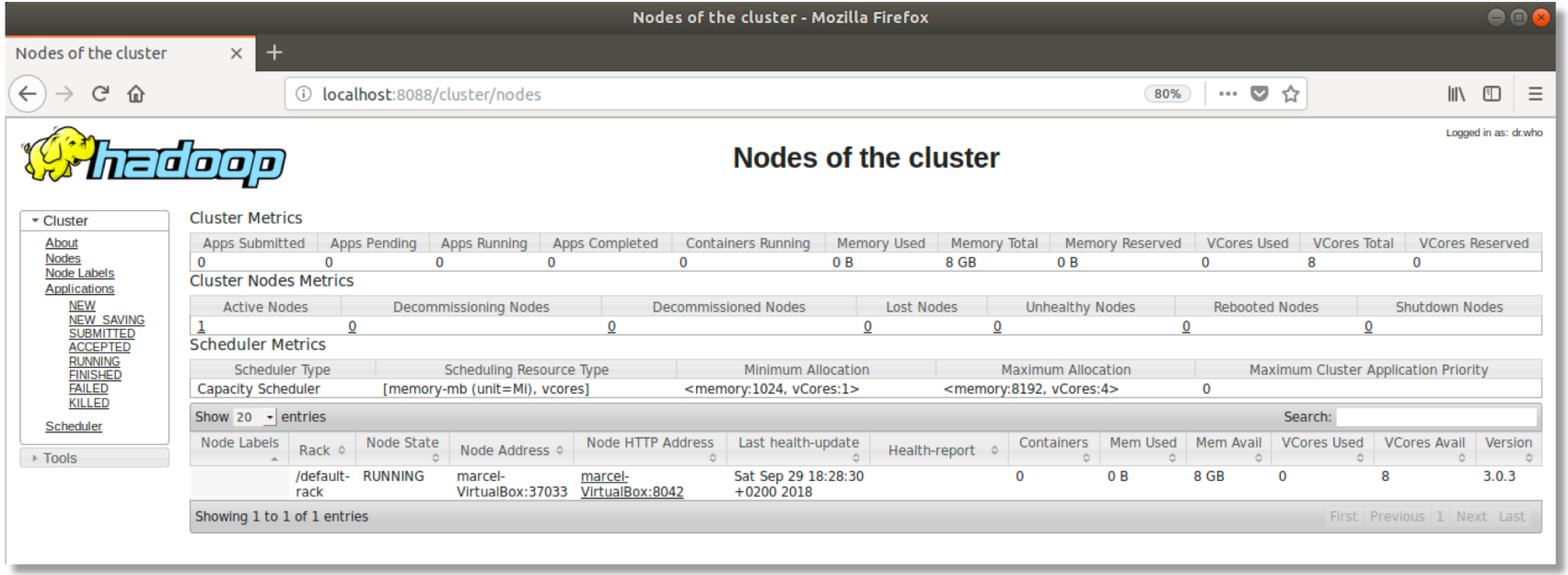
```
Configured Capacity: 20852596736 (19.42 GB)
Present Capacity: 12480020480 (11.62 GB)
DFS Remaining: 12479995904 (11.62 GB)
DFS Used: 24576 (24 KB)
DFS Used%: 0.00%
Replicated Blocks:
                        Under replicated blocks: 0
                        Blocks with corrupt replicas: 0
                        Missing blocks: 0
                        Missing blocks (with replication factor 1): 0
                        Pending deletion blocks: 0
Erasure Coded Block Groups:
                        Low redundancy block groups: 0
                        Block groups with corrupt internal blocks: 0
                        Missing block groups: 0
                        Pending deletion blocks: 0

-------------------------------------------------
Live datanodes (1):

Name: 127.0.0.1:9866 (localhost)
Hostname: marcel-VirtualBox
Decommission Status : Normal
Configured Capacity: 20852596736 (19.42 GB)
DFS Used: 24576 (24 KB)
Non DFS Used: 7289712640 (6.79 GB)
DFS Remaining: 12479995904 (11.62 GB)
DFS Used%: 0.00%
DFS Remaining%: 59.85%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Sat Sep 29 18:04:04 CEST 2018
Last Block Report: Sat Sep 29 18:01:52 CEST 2018
```

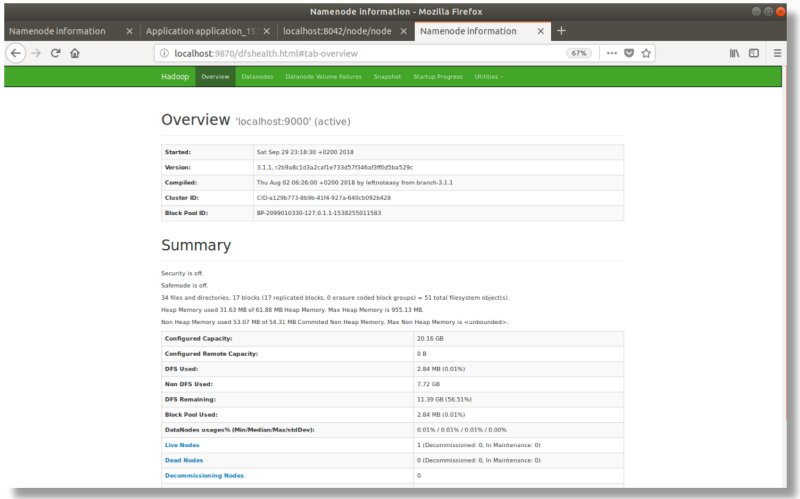
Check Hadoop/HDFS

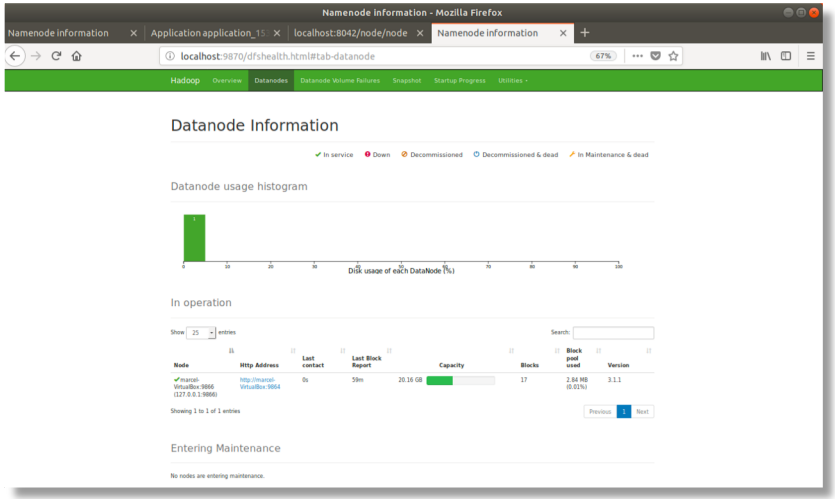
11. Check Ressource Manager Landing Page (http://localhost:8088/cluster):



Check Hadoop/HDFS

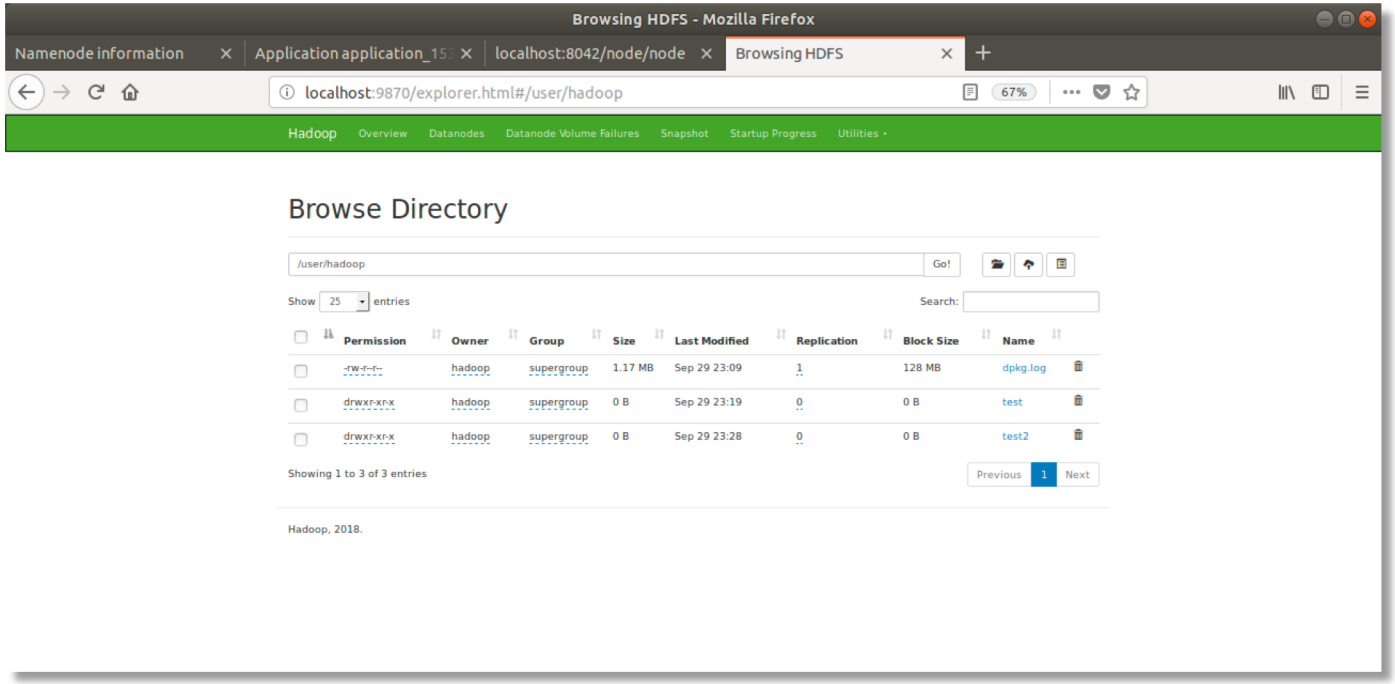
12. Check NameNode Landing and Status Page (http://localhost:9870*):*





Check Hadoop/HDFS

13. Check HDFS File Browser (http://localhost:9870/explorer.html#/):



Working with HDFS

1. Create User Directory (*on HDFS*):

```
hadoop fs -mkdir /user
hadoop fs -mkdir /user/hadoop
```

2. List Directories (*on HDFS*):

```
hadoop@marcel-VirtualBox:~$ hadoop fs -ls /
Found 2 items
drwx------   - hadoop supergroup          0 2018-09-29 23:11 /tmp
drwxr-xr-x   - hadoop supergroup          0 2018-09-29 23:09 /user
hadoop@marcel-VirtualBox:~$
```


Working with HDFS

3. Copy File (just a *random log file*) from local directory to HDFS:

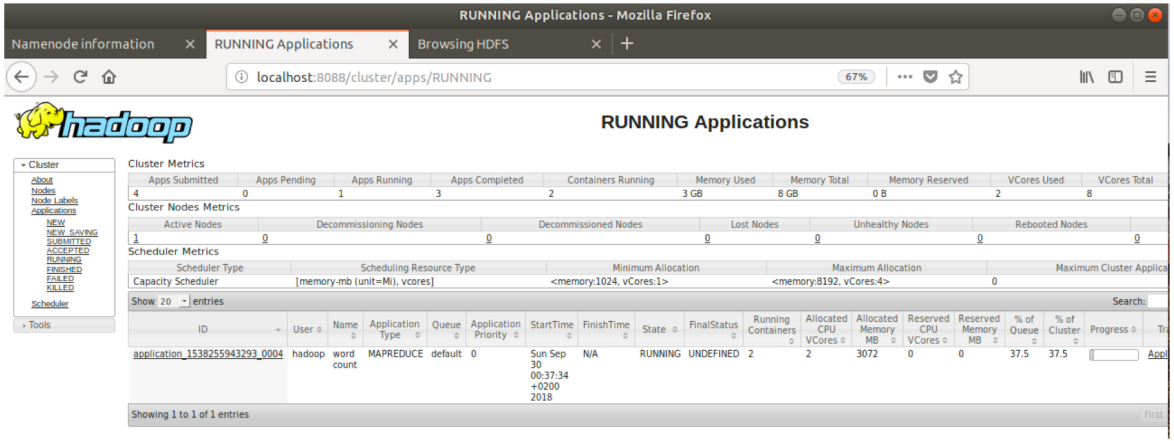
```
hadoop fs -put /var/log/dpkg.log /user/hadoop/dpkg.log
```


Run Example MapReduce Job

1. Using MapReduce WordCount Jar provided by Hadoop

```
hadoop jar hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.1.1.jar
wordcount /user/hadoop/dpkg.log /user/hadoop/test_output
```

2. View Running MapReduce Job:



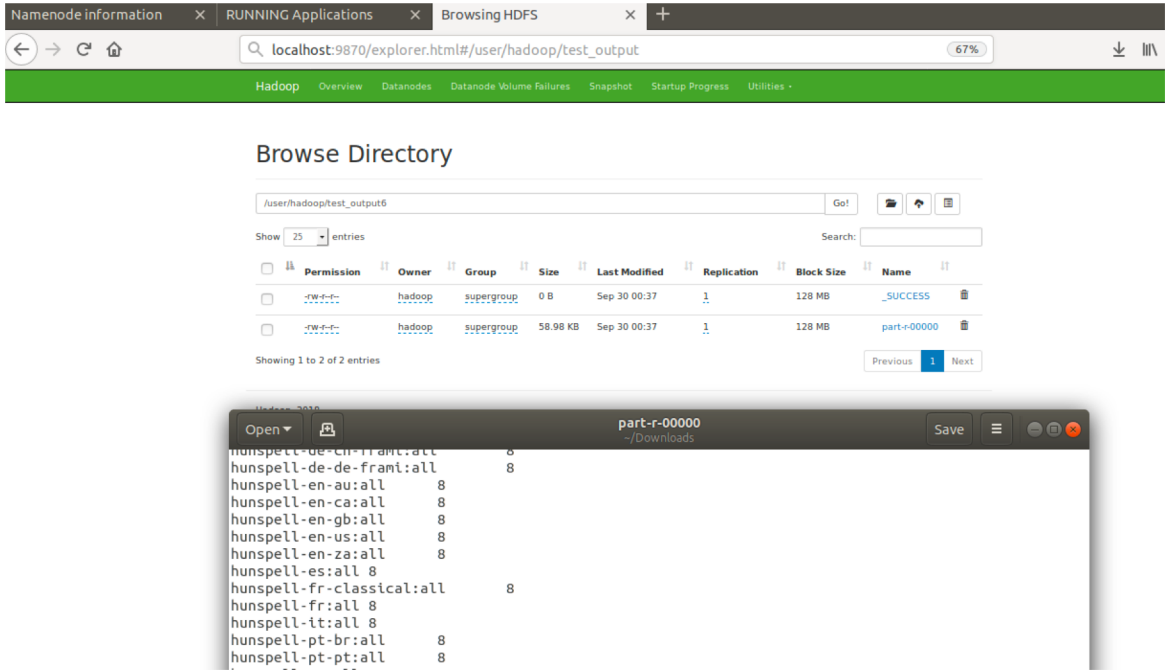
Run Example MapReduce Job

3. Take A Look At The Output/Result (*via Bash*):

```
hadoop@marcel-VirtualBox:~$ hadoop fs -cat /user/hadoop/test_output/part-r-00
000
hunspell-de-de-frami:all      8
hunspell-en-au:all  8
hunspell-en-ca:all  8
hunspell-en-gb:all  8
hunspell-en-us:all  8
hunspell-en-za:all  8
hunspell-es:all     8
hunspell-fr-classical:all     8
hunspell-fr:all     8
hunspell-it:all     8
hunspell-pt-br:all  8
hunspell-pt-pt:all  8
…
```


Run Example MapReduce Job

4. Take A Look At The Output/Result (*via Web HDFS File Browser*):



Exercises

Hadoop, HDFS, Yarn

Exercises

1. **Clone git repo** (to get sample data):

```
git clone https://github.com/marcelmittelstaedt/BigData.git
```

2.

- **Copy sample file** (*/BigData/exercises/01_hadoop/sample_data/Faust_1.txt*) from Git Repo **to HDFS**.
- Use and **run** default **MapReduce** Jar (*hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.1.1.jar*) **to calculate wordcount** for text file.
- **Copy result** of MapReduce job **back to local** ubuntu **filesystem**.

3.

- Use and **run** default **MapReduce** Jar (*hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.1.1.jar*) to **get the count of occurences** of the exact string ***‚Faust‘*** within text file.
- **Copy result** of MapReduce job **back to local** ubuntu **filesystem**.
- **Tip:** don‘t use *wordcount* part of jar.

Help

MapReduce Examples within *hadoop-mapreduce-examples-3.1.1.jar*:

aggregatewordcount:	An Aggregate based mapreduce program that counts the words in the input files.
aggregatewordhist:	An Aggregate based mapreduce program that computes the histogram of the words in the input files.
bbp:	A mapreduce program that uses Bailey-Borwein-Plouffe to compute exact digits of Pi.
dbcount:	An example job that counts the pageview logs stored in a database.
distbbp:	A mapreduce program that uses a BBP-type formula to compute exact bits of Pi.
grep:	A mapreduce program that counts the matches of a regex in the input.
join:	A job that performs a join over sorted, equally partitioned datasets.
multifilewc:	A job that counts words from several files.
pentomino:	A mapreduce tile laying program to find solutions to pentomino problems.
pi:	A mapreduce program that estimates Pi using a quasi-Monte Carlo method.
randomtextwriter:	A mapreduce program that writes 10 GB of random textual data per node.
randomwriter:	A mapreduce program that writes 10 GB of random data per node.
secondarysort:	An example defining a secondary sort to the reduce phase.
sort:	A mapreduce program that sorts the data written by the random writer.
sudoku:	A sudoku solver.
teragen:	Generate data for the terasort.
terasort:	Run the terasort.
teravalidate:	Checking results of terasort.
wordcount:	A mapreduce program that counts the words in the input files.
wordmean:	A mapreduce program that counts the average length of the words in the input files.
wordmedian:	A mapreduce program that counts the median length of the words in the input files.
wordstandarddeviation:	A mapreduce program that counts the standard deviation of the length of the words in the input files.