Distributed Data Management

Introduction

Thorsten Papenbrock
Felix Naumann
F-2.03/F-2.04, Campus II
Hasso Plattner Institut
Distributed Data Management

Introduction

Data Profiling
- Data Profiling
- Web Data
- Entity Recognition
- project DataChEx

Data Fusion
- Service-Oriented Systems
- project DuDe

Data Integration
- Information Integration
- Duplicate Detection
- Entity Search
- Data Scrubbing
- project Stratosphere

Information Quality
- Data Cleansing
- Dependency Detection
- Linked Open Data
- Opinion Mining
- RDF Data Mining
- ETL Management

Web Science
- Data as a Service
- Text Mining
- Linked Open Data
- Change Exploration
- Data Change

Data Preparation
- ETL Management
- Opinion Mining

Dr. Thorsten Papenbrock
Dr. Ralf Krestel
Tim Repke
Julian Risch

Diana Stephan
Prof. Felix Naumann

Hazan Harmouch

John Koumarelas
Tobias Bleifuß
Leon Bornemann
Lan Jiang
Konstantina Lazariduo
Michael Loster

Text Mining

Opinion Mining

Dr. Ralf Krestel
Tim Repke
Julian Risch

Diana Stephan
Prof. Felix Naumann

Hazan Harmouch

John Koumarelas
Tobias Bleifuß
Leon Bornemann
Lan Jiang
Konstantina Lazariduo
Michael Loster
Introduction: Audience

- English?
- Which semester?
- ITSE, DE, DH?
- Database knowledge?
- HPI or Guest?
- Other related lectures?
- Distributed experience?
Lectures
- Datenbanksysteme II (Bachelor)
- Deep Learning for Text Mining (Master)
- Distributed Data Management (Master)

Seminars
- Methoden der Forschung (Master)
- Data Preparation for Science (Master)
- Recommender Systems (Master)
- Social Network Analysis in Practice (Master)
Domain Knowledge

Data Science

Domain Expertise (e.g., Industry 4.0, Medicine, Physics, Engineering, Energy, Logistics)

Mathematical Programming
- Linear Algebra
- Statistics
- Text Mining
- Graph Mining
- Signal Processing
- Stochastic Gradient Descent
- Machine Learning
- Error Estimation
- Active Sampling
- Monte Carlo
- Regression
- Predictive Analytics
- Sketches
- Data Obfuscation
- Convergence
- Decoupling
- Iterative Algorithms
- Curse of Dimensionality

Relational Algebra / SQL
- Data Warehouse/OLAP
- NF² / XQuery
- RDF / SparQL
- Information Integration
- Information Extraction
- Visual Analytics
- Privacy
- Memory Management
- Parallelization
- Scalability
- Memory Hierarchy
- Fault Tolerance
- Security
- Data Analysis Languages
- Query Optimization
- Indexing
- Control Flow
- Real-Time

Domain Analysis (DA)
- Machine Learning
- Statistics
- Data Analysis

Data Integration (DI)
- Scalable Data Management
- Business Models
- Legal Aspects
- Fault Tolerance
- Memory Management
- Data Flow
Distributed Data Management

This Lecture

Lecture
- For master students
  (IT-Systems Engineering, Digital Health, Data Engineering)
- 6 credit points, 4 SWS
- Mondays 13:30 – 15:00
  Tuesdays 9:15 – 10:45

Exercises
- Interleaved with lectures

Slides
- On website

Prerequisites
- To participate:
  A little background and interest in databases (e.g. DBS I lecture)
- For exam:
  Attending lectures, participation in exercises, and completion of exercise homework tasks

Antirequisite
- Distributed Data Analytics

Exam
- Written exam
- Probably first week after lectures

Website
Distributed Data Management

Question any time please!

- During lectures
- Visit us: Campus II, Rooms F-2.03 and F-2.04
- Email:
  - thorsten.papenbrock@hpi.de
  - felix.naumann@hpi.de

Also: Give feedback about ...

- improving lectures
- informational material
- organization

Official evaluation

- At the end of this semester
- ... too late for important feedback!
Distributed Data Management
Distributed Data Management
Motivation: “Distributed” Paradigm Shift in Software-Writing

The free lunch is over!
- Clock speeds stall
- Transistor numbers still increase
  - Cores in CPUs/GPUs
    - CPUs/GPUs in compute nodes, compute nodes in clusters

Paradigm Shift:
- Earlier: optimize code for a single thread
- Now: solve tasks in parallel

Distributed computing
“Distribution of work on (potentially) physically isolated compute nodes”

http://www.gotw.ca/publications/concurrency-ddj.htm
Motivation: “Distributed” Surpassing Moor’s Law

Moore’s Law (Observation)
“The number of transistors in integrated circuits doubles approximately every two years”

With **distributed machines**, we can already build systems with any number of transistors! (don’t even need to wait for a new processors)
Motivation: “Distributed”
A Rule to Acknowledge

Amdahl’s Law
“The speedup of a program using multiple processors for parallel computing is limited by the sequential fraction of the program”

\[ Speedup(s) = \frac{1}{(1 - p) + \frac{p}{s}} \]

s: degree of parallelization (e.g. #cores)
Amdahl’s Law cannot work around this law!

Even *distributed parallelization* cannot work around this law!
Motivation: “Distributed”
New Technologies

Distributed Computing
- Hadoop
- akka
- Spark
- Flink
- STORM

Distributed Storage
- Hadoop
- Redis
- CouchDB
- Cassandra
- MongoDB
Distributed Data Management
Introduction

Thorsten Papenbrock
Slide 16
Motivation: “Distributed”  
Small and Medium Scale

Low-cost and low energy cluster of Cubieboards running Hadoop

A cluster of commodity hardware running Hadoop
Motivation: “Distributed”
Large Scale

A cluster of machines running Hadoop at Yahoo!
Motivation: "Distributed" Super Large Scale

Top 10 Super Computers 2017

1. Sunway TaihuLight - Sunway MPP, Sunway SW26010 260C 1.45GHz. Sunway, NRPC, National Supercomputing Center in Wuxi, China
   - Cores: 10,649,600
   - Rmax (TFlop/s): 93,014.6
   - Rpeak (TFlop/s): 125,435.9
   - Power (KW): 15,371

2. Tianhe-2 (MilkyWay-2) - TH/IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.20GHz, TH Express-2, Intel Xeon Phi 315P, NUDT National Super Computer Center in Guangzhou, China
   - Cores: 3,120,000
   - Rmax (TFlop/s): 33,862.7
   - Rpeak (TFlop/s): 54,902.4
   - Power (KW): 17,808

3. Piz Daint - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect, NVIDIA Tesla P100, Cray Inc., Swiss National Supercomputing Centre (CSCS), Switzerland
   - Cores: 361,760
   - Rmax (TFlop/s): 19,590.0
   - Rpeak (TFlop/s): 25,326.3
   - Power (KW): 2.272

4. Titan - Cray XK7, Opteron 6274 16C 2.20GHz, Cray Gemini interconnect, NVIDIA K20x, Cray Inc., DOE/SC/Oak Ridge National Laboratory, United States
   - Cores: 560,640
   - Rmax (TFlop/s): 17,590.0
   - Rpeak (TFlop/s): 27,112.5
   - Power (KW): 8,209

5. Sequoia - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom, IBM DOE/NNSA/LLNL, United States
   - Cores: 1,572,864
   - Rmax (TFlop/s): 17,173.2
   - Rpeak (TFlop/s): 20,132.7
   - Power (KW): 7,890

6. Cori - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect, Cray Inc., DOE/SC/LBNL/NERSC, United States
   - Cores: 622,336
   - Rmax (TFlop/s): 14,014.7
   - Rpeak (TFlop/s): 27,880.7
   - Power (KW): 3,939

7. Oakforest-PACS - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path, Fujitsu Joint Center for Advanced High Performance Computing, Japan
   - Cores: 556,104
   - Rmax (TFlop/s): 13,554.6
   - Rpeak (TFlop/s): 24,913.5
   - Power (KW): 2,719

8. K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect, Fujitsu RIKEN Advanced Institute for Computational Science (AICS), Japan
   - Cores: 705,024
   - Rmax (TFlop/s): 10,510.0
   - Rpeak (TFlop/s): 11,280.4
   - Power (KW): 12,660

9. Mira - BlueGene/Q, Power BQC 16C 1.60GHz, Custom, IBM DOE/SC/Argonne National Laboratory, United States
   - Cores: 786,432
   - Rmax (TFlop/s): 8,586.6
   - Rpeak (TFlop/s): 10,066.3
   - Power (KW): 3,045

10. Trinity - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect, Cray Inc., DOE/NNSA/LANL/SLAC, United States
    - Cores: 301,056
    - Rmax (TFlop/s): 8,100.9
    - Rpeak (TFlop/s): 11,078.9
    - Power (KW): 4,233

https://www.top500.org/lists/2017/06/
Motivation: “Distributed” Super Large Scale

Top 10 Super Computers 2017

<table>
<thead>
<tr>
<th>Rank</th>
<th>System</th>
<th>Cores</th>
<th>Rmax (TFlop/s)</th>
<th>Rpeak (TFlop/s)</th>
<th>Power (KW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sunway TaihuLight - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCC, National Supercomputing Center in Wuxi, China</td>
<td>10,649,600</td>
<td>93,014.6</td>
<td>125,435.9</td>
<td>15,371</td>
</tr>
<tr>
<td>2</td>
<td>Tianhe-2 (MilkyWay-2) - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2,200GHz, TH Express-2, Intel Xeon Phi 31S1P, NUDT National Super Computer Center in Guangzhou, China</td>
<td>3,120,000</td>
<td>33,862.7</td>
<td>54,902.4</td>
<td>17,808</td>
</tr>
<tr>
<td>3</td>
<td>Piz Daint - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect, NVIDIA Tesla P100, Cray Inc., Swiss National Supercomputing Centre (CSCS), Switzerland</td>
<td>361,760</td>
<td>19,590.0</td>
<td>25,326.3</td>
<td>2,272</td>
</tr>
<tr>
<td>4</td>
<td>Titan - Cray XK7, Opteron 6274 16C 2,200GHz, Cray Gemini interconnect, NVIDIA K20x, Cray Inc., DOE/SC/Oak Ridge National Laboratory, United States</td>
<td>560,640</td>
<td>17,590.0</td>
<td>27,112.5</td>
<td>8,209</td>
</tr>
<tr>
<td>5</td>
<td>Sequoia - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom, IBM DOE/NSA/LLNL, United States</td>
<td>1,572,864</td>
<td>17,173.2</td>
<td>20,132.7</td>
<td>7,890</td>
</tr>
<tr>
<td>6</td>
<td>Cori - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect, Cray Inc., DOE/SC/LBNL/NERSC, United States</td>
<td>622,336</td>
<td>14,014.7</td>
<td>27,880.7</td>
<td>3,939</td>
</tr>
<tr>
<td>7</td>
<td>Oakforest-PACS - PRIMERGY CX640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path, Fujitsu Joint Center for Advanced High Performance Computing, Japan</td>
<td>556,104</td>
<td>13,554.6</td>
<td>24,913.5</td>
<td>2,719</td>
</tr>
<tr>
<td>8</td>
<td>K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect, Fujitsu RIKEN Advanced Institute for Computational Science (AICS), Japan</td>
<td>705,024</td>
<td>10,510.0</td>
<td>11,280.4</td>
<td>12,660</td>
</tr>
<tr>
<td>9</td>
<td>Mira - BlueGene/Q, Power BQC 16C 1.60GHz, Custom, IBM DOE/SC/Argonne National Laboratory, United States</td>
<td>786,432</td>
<td>8,586.6</td>
<td>10,066.3</td>
<td>3,945</td>
</tr>
<tr>
<td>10</td>
<td>Trinity - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect, Cray Inc., DOE/NSA/LANL/SLNL, United States</td>
<td>301,056</td>
<td>8,100.9</td>
<td>11,078.9</td>
<td>4,233</td>
</tr>
</tbody>
</table>

https://www.top500.org/lists/2017/06/
Motivation: “Distributed” Super Large Scale

Top 10 Super Computers 2017

<table>
<thead>
<tr>
<th>Rank</th>
<th>System</th>
<th>Cores</th>
<th>Rmax (TFlop/s)</th>
<th>Rpeak (TFlop/s)</th>
<th>Power (KW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sunway TaihuLight - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRPC, National Supercomputing Center in Wuxi, China</td>
<td>10,649,600</td>
<td>93,014.6</td>
<td>125,435.9</td>
<td>15,371</td>
</tr>
<tr>
<td>2</td>
<td>Tianhe-2 (MilkyWay-2) - TH/IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.20GHz, TH Express-2, Intel Xeon Phi 31S1P, NUDT National Super Computer Center in Guangzhou, China</td>
<td>3,120,000</td>
<td>33,862.7</td>
<td>54,902.4</td>
<td>17,808</td>
</tr>
<tr>
<td>3</td>
<td>Piz Daint - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect, NVIDIA Tesla P100, Cray Inc., Swiss National Supercomputing Centre (CSCS), Switzerland</td>
<td>361,760</td>
<td>19,590.0</td>
<td>25,326.3</td>
<td>2,272</td>
</tr>
<tr>
<td>4</td>
<td>Titan - Cray XK7, Opteron 6274 16C 2.20GHz, Cray Gemini interconnect, NVIDIA K20x, Cray Inc., DOE/SC/Oak Ridge National Laboratory, United States</td>
<td>560,640</td>
<td>17,590.0</td>
<td>27,112.5</td>
<td>8,209</td>
</tr>
<tr>
<td>5</td>
<td>Sequoia - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom, IBM, DOD/NNSA/LBNL, United States</td>
<td>1,572,864</td>
<td>17,173.2</td>
<td>20,132.7</td>
<td>7,890</td>
</tr>
<tr>
<td>6</td>
<td>Cori - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect, Cray Inc., DOE/SCL/NERSC, United States</td>
<td>622,336</td>
<td>14,014.7</td>
<td>27,880.7</td>
<td>3,939</td>
</tr>
<tr>
<td>7</td>
<td>Oakforest-PACS - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path, Fujitsu, Joint Center for Advanced High Performance Computing, Japan</td>
<td>556,104</td>
<td>13,554.6</td>
<td>24,913.5</td>
<td>2,719</td>
</tr>
<tr>
<td>8</td>
<td>K computer, SPARC64 VIII fx 2.0GHz, Tofu interconnect, Fujitsu, RIKEN Advanced Institute for Computational Science (AICS), Japan</td>
<td>705,024</td>
<td>10,510.0</td>
<td>11,280.4</td>
<td>12,660</td>
</tr>
<tr>
<td>9</td>
<td>Mira - BlueGen/Q, Power BQC 16C 1.60GHz, Custom, IBM, DOE/S/Argonne National Laboratory, United States</td>
<td>796,432</td>
<td>8,586.6</td>
<td>10,066.3</td>
<td>3,045</td>
</tr>
<tr>
<td>10</td>
<td>Trinity - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect, Cray Inc., DOE/NNSA/LANL/SNL, United States</td>
<td>301,056</td>
<td>8,100.9</td>
<td>11,078.9</td>
<td>4,233</td>
</tr>
</tbody>
</table>

https://www.top500.org/lists/2017/06/
Distributed Data Management

“The ability of extracting different kinds of information from data!”

- Structural information
- Explicit information
- Implicit/derived information
A market worth $122 billion in 2016 with a growth of 11.3% per year!

Excellent job opportunities in many companies!

For a world that created an entire zettabyte (which is exactly $10^{12}$ GB) of data in the 2010 alone!

https://www.idc.com/getdoc.jsp?containerId=prUS41826116
<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>29 Aug (Tue)</td>
<td>8:30-10:00</td>
<td>Welcome Messages + Wolfgang Lehner keynote</td>
</tr>
<tr>
<td></td>
<td>10:00-10:30</td>
<td>break</td>
</tr>
<tr>
<td></td>
<td>10:30-12:00</td>
<td>Panel</td>
</tr>
<tr>
<td></td>
<td>12:00-13:30</td>
<td>lunch</td>
</tr>
<tr>
<td></td>
<td>13:30-15:00</td>
<td>Query Processing and Optimization</td>
</tr>
<tr>
<td></td>
<td>15:00-15:30</td>
<td>break</td>
</tr>
<tr>
<td></td>
<td>16:00-18:00</td>
<td>Transactions</td>
</tr>
<tr>
<td></td>
<td>poster reception</td>
<td></td>
</tr>
<tr>
<td>(Wed)</td>
<td>8:30-10:00</td>
<td>Endowment Update + VLDB Awards + Michael Franklin keynote</td>
</tr>
<tr>
<td></td>
<td>10:00-10:30</td>
<td>break</td>
</tr>
<tr>
<td></td>
<td>10:30-12:00</td>
<td>High Performance Query Processing</td>
</tr>
<tr>
<td></td>
<td>12:00-13:30</td>
<td>lunch</td>
</tr>
<tr>
<td></td>
<td>13:30-15:00</td>
<td>Scalable Storage</td>
</tr>
<tr>
<td></td>
<td>15:00-15:30</td>
<td>break</td>
</tr>
<tr>
<td></td>
<td>15:30-17:00</td>
<td>Concurrency Control</td>
</tr>
<tr>
<td></td>
<td>17:00-18:00</td>
<td>poster reception</td>
</tr>
<tr>
<td></td>
<td>19:00 - …</td>
<td>“VLDB Oktoberfest” banquet at the Hofbräuhaus</td>
</tr>
<tr>
<td>31 Aug (Thu)</td>
<td>8:30-10:30</td>
<td>Endowment Awards + Demo Award + Jens Dittrich plenary</td>
</tr>
<tr>
<td></td>
<td>10:30-10:55</td>
<td>break</td>
</tr>
<tr>
<td></td>
<td>10:55-12:00</td>
<td>Transactions and Persistence</td>
</tr>
<tr>
<td></td>
<td>12:00-13:30</td>
<td>lunch</td>
</tr>
<tr>
<td></td>
<td>13:30-15:00</td>
<td>Data Partitioning</td>
</tr>
<tr>
<td></td>
<td>15:00-15:30</td>
<td>break</td>
</tr>
<tr>
<td></td>
<td>15:30-17:00</td>
<td>DB Engines 2</td>
</tr>
<tr>
<td></td>
<td>17:00-18:00</td>
<td>poster reception</td>
</tr>
</tbody>
</table>

International conference “Very Large Data Bases”
### VLDB 2017 Program

**With strong relation to data analytics and/or distributed processing**
Motivation: “Data Analytics”
Back to Super Computers

For what tasks are these computers used?

- Weather forecasting
- Market analysis
- Crash simulation
- Disaster simulation
- Brute force decryption
- Molecular dynamics modeling
- …

Data Analytics tasks!
Distributed Data Management

“The ability to efficiently read, transform, and store large amounts of data!”

- Static (block) data
- Volatile (streaming) data

Processing
Motivation: “Data Processing”

Data Processing in Practice

Startup

- Everyone starts small but be prepared for the hype!
Motivation: “Data Processing”
Data Processing in Practice

Example: Mobile Motion GmbH

- An HPI-Startup of 2013
- Founders:
  - Jonas Drüppel, Roland Grenke, Daniel Taschik

November 19, 2014:
  - Launch of the Dubsmash app

November 26, 2014:
  - Dubsmash reached the number one downloaded app in Germany

June 1, 2015:
  - Dubsmash had been downloaded over 50 million times in 192 countries
Motivation: “Data Processing”

Data Processing in Practice

Many further HPI Startups!
Motivation: “Data Processing”

Data Processing in Practice

Successful IT-Startups in recent years are masters of data:

1. AirBnB
2. Instagram
3. Pinterest
4. Angry Birds
5. Linkedin
6. Uber
7. Snapchat
8. WhatsApp
9. Twitter
10. Facebook
11....

Peta- to Exabytes of...
- profile data (names, addresses, friends, ...)
- content data (images, videos, messages, ...)
- event data (logins, interactions, games, ...)
- ...

Challenged with...
- streaming
- persistence
- analytics
- load-balancing
- ...

Thursten Papenbrock
Distributed Data Management
Introduction
Motivation: “Data Processing”

Driving Forces

- **Data volumes increase:**
  business data, sensor data, social media data, ...

- **Data analytics gains importance:**
  downtime-less, real-time, predictive

- **Parallelization paradigm shifts:**
  multi-core and network speeds increase while CPU clock speeds stall

- **Computation resources become more available:**
  IaaS, PaaS, SaaS

- **Free and open source software gains popularity:**
  setting standards, utilizing external development resources, improving software quality, avoiding vendor locks ...
Distributed Data Management

Related Topics

- Database Systems
- Software Architecture
- Parallel Computing
- Data Mining

Distributed Data Management

Introduction

Thorsten Papenbrock
Slide 33
Distributed Data Management

Related Topics

- Database Systems
- Software Architecture
- Parallel Computing
- Data Mining

Distributed Data Management

Introduction

ThorstenPapenbrock
Slide 34
Touch points

- Data models, query languages, and consistency guarantees
- Distributed storage and retrieval of data
- Index structures

Not in this lecture

- Physical data storage
- In-depth presentations language constructs and transactions
- Core database technology, e.g., query optimizer

More focused lectures

- Database Systems I + II (Prof. Naumann)
- Trends and Concepts in Software Industry (Prof. Plattner)
Distributed Data Management

Related Topics

- Database Systems
- Software Architecture
- Parallel Computing
- Data Mining

Distributed Data Management

Introduction

ThorstenPapenbrock
Slide 36
Distributed Data Management
Software Architectures

Touch points

- Requirements, design, and architecture of distributed systems
- Pros and cons of different technologies for distributed systems

Not in this lecture

- Non-distributed systems
- Agile software development techniques
- Software patterns

More focused lectures

- Software Architecture (Dr. Uflacker)
- Software Technique (Dr. Uflacker)
Distributed Data Management

Related Topics

- Database Systems
- Software Architecture
- Data Mining
- Parallel Computing

Distributed Data Management

Introduction

Thorsten Papenbrock
Slide 38
Touch points
- Distributed data storage concepts
- Distributed programming models, e.g., actor programming and MapReduce

Not in this lecture
- Parallel, non-distributed programming languages, e.g., CUDA or OpenMP
- Core parallel computing concepts, e.g., scheduling or shared memory
- Processor architectures, cache hierarchies, GPU programming, ...

More focused lectures
- Parallel Programming (Dr. Tröger)
- Programmierung paralleler und verteilter Systeme (Dr. Feinbube)
Distributed Data Management

Related Topics

- Database Systems
- Software Architecture
- Parallel Computing
- Data Mining

Distributed Data Management

Introduction

ThorstenPapenbrock
Slide 40
Touch points

- Data analytics: aggregation queries and basic data mining algorithms

Not in this lecture

- Detailed introduction to machine learning, e.g., neuronal networks, (un)supervised learning, or Bayesian classification
- Statistics, linear algebra, and most sophisticated mining algorithms

More focused lectures

- Big Data Analytics (Prof. Müller)
- Graph Mining (Prof. Müller)
- Data Profiling (Prof. Naumann)
- Data Mining and Probabilistic Reasoning (Dr. Krestel)
Distributed Data Management

Related Topics

Database Systems
Software Architecture
Parallel Computing
Data Mining

Distributed Data Management

Introduction

Thorsten Papenbrock
Slide 42
Sorting the buzzwords
- NoSQL, Big Data, OLAP, Web-scale, ACID, Sharding, MapReduce, ...

Understanding the systems
- How and why do distributed systems work?
- What are their individual core technologies?
- Which advantages and disadvantages do the technologies have?
- How can different technologies be combined?

Exercising in data analytics
- Distributed processing of data-parallel tasks
- Data mining on streams
Distributed Data Management
Lecture Outline

1. Introduction
2. Foundations
3. OLAP and OLTP
4. Encoding and Evolution
5. Hands-On: Akka
6. Data Models and Query Languages
7. Storage and Retrieval
8. Replication
9. Partitioning
10. Batch Processing
11. Hands-On: Spark
12. Distributed Systems
13. Consistency and Consensus
14. Transactions
15. Stream Processing
16. Hands on: Flink
17. Mining Data Streams
18. Distributed Algorithms
19. Services and Containerization
20. Cloud-based Data Systems
21. Lecture Summary and Exam Preparation

On-site and by dataArtisans!

Distributed Data Management
Introduction

ThorstenPapenbrock
Slide 44
Designing Data-Intensive Applications

- Author: Martin Klappmann
- Date: March 2017
- Publisher: O'Reilly Media, Inc
- ISBN: 978-1-449-37332-0
- References: [https://github.com/ept/ddia-references](https://github.com/ept/ddia-references)

Scope for this lecture

- Distributed and parallel systems for data analytics and data mining
- Big data storage
- Batch and stream processing
Web-links are given on the slides during the lecture.

Distributed Data Management
Introduction

Thorsten Papenbrock
Slide 47
### Distributed Data Analytics (Winter term 2017/2018)

<table>
<thead>
<tr>
<th>Degree</th>
<th>Type</th>
<th>Responsible</th>
<th>Voters</th>
<th>Distribution and average grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA ITSE</td>
<td>Lecture</td>
<td>Prof. Dr. Felix Naumann</td>
<td>18/28</td>
<td>64%</td>
</tr>
</tbody>
</table>

- I learned a lot in the course.  
  - Score: 1.7
- ... gave the course in an interesting way.  
  - Score: 1.9
- I was satisfied how the exercise submission worked.  
  - Score: 1.8
<table>
<thead>
<tr>
<th>What did you like?</th>
<th></th>
<th>What could be improved?</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Ich fand gut, dass du im Laufe der Veranstaltung nicht mehr so nervös warst wie am Anfang.</td>
<td>• Maybe try a more project based approach as well instead of the two hands ons? Or a project seminar that fits the lecture</td>
<td></td>
</tr>
<tr>
<td>• Thorsten hat eine sehr ruhige Art, was es möglich macht ihm gut zuzuhören. Man hat auch das Gefühl, dass er wirklich Herzblut in die Veranstaltung gesteckt hat. Dankeschön!</td>
<td>• Klarer skizzieren wieso bestimmte Vorlesungspunkte wichtig sind. Wie würde man Spark o.ä. in einer kompletten Pipeline verwenden? Also angenommen man erhält kontinuierlich Daten und soll diese analysieren und das Ergebnis ausgeben - wie kann man dies erreichen?</td>
<td></td>
</tr>
<tr>
<td>• Du hast eine sehr angenehme und ruhige Art zu präsentieren. Man merkt, dass dir die Studenten und die Lehre wichtig sind.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>