



# BA-Praktikum WS2019/2020

## February 18, 2020



Universität Hamburg  
DER FORSCHUNG | DER LEHRE | DER BILDUNG

base.camp – Eugen Ruppert

---

**BA-PRAKTIKUM WS2019/2020**  
**BIG DATA**

Big Data

Themen / Datensets

Hadoop

Software Engineering

Beispielprojekte

Lösungsansätze

Weiterer Ablauf

Qualifizierte Fragen?

# Ablauf heute



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

10:15 Einleitung, Vorstellung der Themen

10:45 Einführung Hadoop

11:30 Mittagspause

12:15 Gruppen- und Themenzuteilung, Raumaufteilung, Zugänge

# Big Data

- Datenmengen wachsen ständig
  - Text, Web
  - Photos
  - Videos
  - Sensoren, Logs, Web of Things
- Speicherkosten sinken kontinuierlich
- **Wir brauchen Methoden, um mit den Daten zurecht zu kommen!**

# Polaroid



- 10 Bilder
- 1 Euro pro Foto

# Kleinbild



- 36 Bilder
- 20 Cent pro Foto

# Smartphone



- 1.000.000 Bilder
- 0 Cent pro Foto?



- 5.000.000 Hamburger Bürger
- Aufnahmen von Sensordaten (Position + ID + Werte)
- 1 Jahr lang, alle 5 Minuten
- etwa 15 TB an Daten

# Themen / Datensets

# Twitter – Social Media



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

- eine der größten Social Media Sites
- Trendanalysen (Sentiment, Hashtags, Retweets)
- Maschinelles Lernen: Klassifizieren auf einem Cluster

- Social Media mit ”“freier Meinungsäußerung” – aber auch viel Hatespeech
- Trendanalysen (Sentiment, Hashtags, Retweets)
- Maschinelles Lernen: Klassifizieren auf einem Cluster

- große Sammlung von Wissen
- Extraktion von explizitem Wissen für Taxonomien  
*Hunde, Katzen und andere Tiere – Hund isA Tier, Katze isA Tier*
- NLP-Bezug

- Hamburger Transparenzportal
- Logs mit Suchen und Fehlern
- Erkennung von Trends

# Eigene Themen



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Verarbeitung großer Datenmengen
- visualisierbares Problem
- eigene Ideen willkommen

Hadoop



# Hello Wor(l)dCount



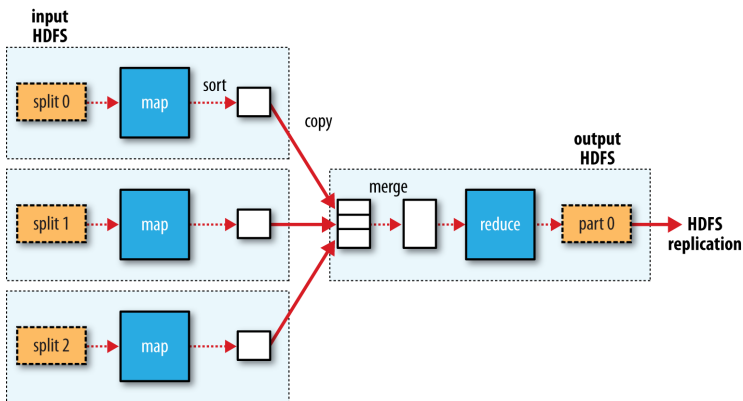
Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

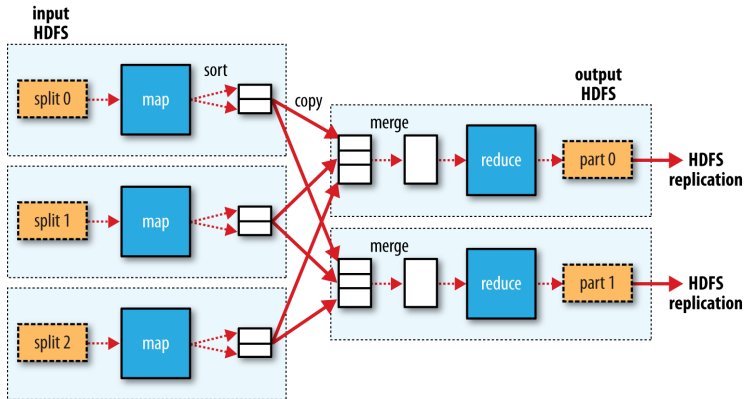
- Wir haben ein großes Textkorpus
- Wie bestimmen wir, wie häufig jedes Wort vorkommt?

- Aufteilung des Tasks in eine Map- und eine Reduce-Phase
- Map: generelle Verarbeitung, filtering, Annotation, Klassifikation
- Reduce: Summe, Max-/Min-Werte, Durchschnitt

# MapReduce



# MapReduce – parallel



# Hadoop FS basics

Task	Command
read directory	<code>hadoop fs -ls</code> <code>hadoop fs -du [-h]</code>
create directory	<code>hadoop fs -mkdir folder</code>
delete directory	<code>hadoop fs -rm -r folder</code>
copy file to HDFS	<code>hadoop fs -put FILE folder</code>
copy STDIN stream to HDFS	<code>hadoop fs -put - folder</code>
delete file	<code>hadoop fs -rm FILE</code>
read contents of a folder	<code>hadoop fs -text folder/*</code>

<http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html>

# Hadoop FS basics – Practice



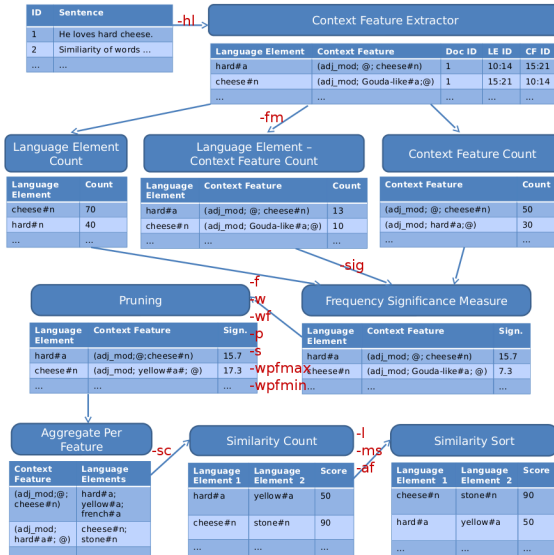
Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

- create folder on HDFS  
`hadoop fs -mkdir DATASET`
- upload file from client directly to HDFS  
`cat FILE | ssh ltheadnode "hadoop fs -put - DATASET/corpus.txt"`
- read text  
`hadoop fs -text DATASET/*`

- Tom White: *Hadoop: The Definitive Guide*. 2009. O'Reilly Media  
<http://shop.oreilly.com/product/9780596521981.do>
- Jeffrey Dean and Sanjay Ghemawat: *MapReduce: Simplified Data Processing on Large Clusters*. 2004  
[https://www.usenix.org/legacy/publications/library/proceedings/osdi04/tech/full\\_papers/dean/dean\\_html/index.html](https://www.usenix.org/legacy/publications/library/proceedings/osdi04/tech/full_papers/dean/dean_html/index.html)
- Jimmy Lin and Chris Dyer: *Data-Intensive Text Processing with MapReduce*. 2010. Morgan & Claypool Publishers  
<http://lintool.github.io/MapReduceAlgorithms/index.html>

# Pipeline of MapReduce Tasks



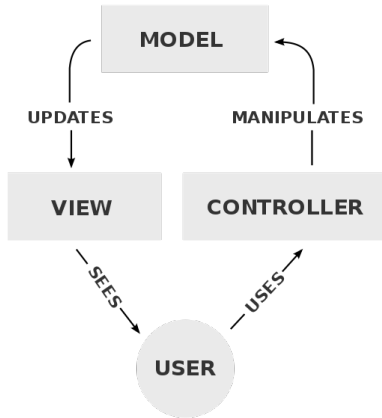


# Software Engineering

- Unterteilung des Programms in 3 unabhängige Teile:
  - Model** Das Datenmodell; alle Datenoperationen werden vom Modell durchgeführt
  - View** Representation des Modells und das User Interface
  - Controller** Verarbeitung von Inputs, Schnittstelle zwischen Model und View
- Spring Boot ist ein MVC-Framework
- MVC ermöglicht wiederverwendbaren Code ohne große Abhängigkeiten

# Software Engineering

## MVC



<https://en.wikipedia.org/wiki/Model-view-controller>

- IntelliJ IDEA  
<https://www.jetbrains.com/idea/>
- moderne Entwicklungsumgebung
- Linux, Mac und Windows
- unterstützt Git, GitHub und GitLab
- Projektmanagement mit Maven  
<https://maven.apache.org/>
- IDEs für verschiedene Programmiersprachen verfügbar

- Projektmanagement  
<https://maven.apache.org/>
- Konzept: Project Object Model (POM)
- definiert Abhängigkeiten sehr genau
- definiert den Build-Prozess
- schneller Einstieg über POM-File
- Maven Central als zentrale Bibliothek von Java Libraries  
<https://search.maven.org/>

# Software Engineering

## Beispiel POM

```
<project>
  <groupId>basecamp</groupId>
  <artifactId>ba19-service-backend</artifactId>
  <version>0.0.1-SNAPSHOT</version>

  <parent>
    <groupId>org.springframework.boot</groupId>
    <artifactId>spring-boot-starter-parent</artifactId>
    <version>1.3.3.RELEASE</version>
  </parent>

  <dependencies>
    <dependency>
      <groupId>org.springframework.boot</groupId>
      <artifactId>spring-boot-starter</artifactId>
    </dependency>
  </dependencies>

  <build>
    <plugins>
      <plugin>
        <groupId>org.springframework.boot</groupId>
        <artifactId>spring-boot-maven-plugin</artifactId>
      </plugin>
    </plugins>
  </build>
</project>
```

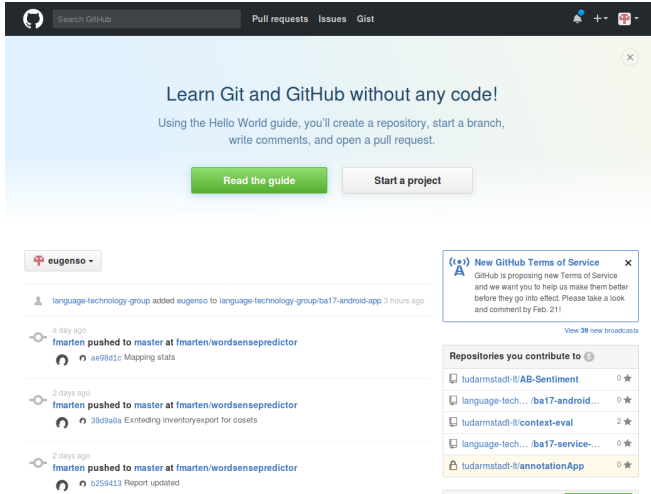
- Git ist ein modernes Versionsverwaltungssystem (version control system; <https://git-scm.com/>)
- ermöglicht leichte Collaboration an Projekten (der Linux Kernel wird mit Git gemanagt)
- Workflow:
  - anderes Projekt "forken"
  - "check out" auf den eigenen Rechner
  - Änderungen durchführen
  - Änderungen "committen", Änderung ist nur lokal
  - Änderungen "pushen", alle bisherigen Änderungen werden transferiert
  - wenn es ein Bugfix war, kann man den Besitzer des Projekts mit einem "pull request" informieren

- GitHub ist eine Webseite für Projektverwaltung mit Git  
<https://github.com/>
- Projekte können von anderen Nutzern "geforkt" werden
- GitHub ist kostenlos nutzbar
- andere Nutzer können zu Projekten hinzugefügt werden
- GitHub hat Eigenschaften eines Social Networks
  - Projekte können geteilt werden
  - man kann Sterne vergeben
  - man kann Leuten folgen
- Recruiter schauen sich auch GitHub-Profile an, ein gutes Profil kann bei Bewerbungen einen guten Eindruck machen



# Software Engineering

## Git, GitHub und GitLab



The screenshot shows the GitHub homepage. At the top is a dark navigation bar with the GitHub logo, a search bar labeled "Search GitHub", and links for "Pull requests", "Issues", and "Gist". On the right of the bar are notification, user, and organization icons. Below the navigation bar is a large light blue banner with the text "Learn Git and GitHub without any code!" and a subtext "Using the Hello World guide, you'll create a repository, start a branch, write comments, and open a pull request." There are two buttons: a green "Read the guide" button and a white "Start a project" button. Below the banner is a section for the user "eugenso". It shows a list of recent activity: "language-technology-group added eugenso to language-technology-group/ba17-android-app 3 hours ago", "fmarten pushed to master at fmarten/wordsensepredictor" (with a commit "ae98d1c Mapping stats" from "a day ago"), "fmarten pushed to master at fmarten/wordsensepredictor" (with a commit "39d9a9a Extending inventory export for cosets" from "2 days ago"), and "fmarten pushed to master at fmarten/wordsensepredictor" (with a commit "b259413 Report updated" from "2 days ago"). On the right side, there is a notification box for "New GitHub Terms of Service" and a section titled "Repositories you contribute to" which lists: "tudarmstadt-It/AB-Sentiment" (0 stars), "language-tech... /ba17-android..." (0 stars), "tudarmstadt-It/context-eval" (2 stars), "language-tech... /ba17-service..." (0 stars), and "tudarmstadt-It/annotationApp" (0 stars).

- GitLab ist eine Plattform ähnlich wie GitHub  
<https://about.gitlab.com/>
- kann selbst gehostet werden
- Continuous Integration ist eingebaut
- private Informatik-Instanz:  
<https://git.informatik.uni-hamburg.de/>
- privates GitLab eignet sich besser für Projekte mit Geschäftszielen als GitHub
- (obwohl GitHub seit ca. 1 Jahr auch private Projekte erlaubt)

# Beispielprojekte

- WordCount MapReduce Anwendung
  - Java  
`https://github.com/basecamp-uhh/Java-MapReduce`
  - Python  
`https://github.com/basecamp-uhh/py-mapreduce`
- nächste Schritte
  - Login prüfen
  - Projekt "forken" (gerne auch 1x pro Gruppe; Mitglieder können zum Projekt hinzugefügt werden)
  - Projekt auschecken und bearbeiten

# Beispielprojekte

## Git Checkout

- IntelliJ IDEA öffnen
- VCS -> Checkout from Version Control -> Git
- GitLab-Link einfügen (vom eigenen Projekt)
- Account-Daten eingeben und speichern
- Clone -> Open -> This Window
- Mit Maven builden
- auf Server laufen lassen

Java:

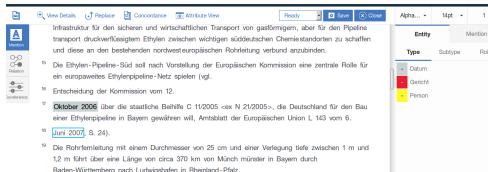
```
hadoop jar mapreduce.wordcount-0.0.1-SNAPSHOT.jar INPUT  
OUTPUT
```

Python:

```
hadoop jar /opt/cloudera/parcels/  
CDH/lib/hadoop-mapreduce/hadoop-streaming.jar -file  
tokenmapper.py -file sumreducer.py -mapper tokenmapper.py  
-reducer sumreducer.py -input INPUT -output OUTPUT
```

# Lösungsansätze

- Daten sind immer verrauscht!
- auf korrekte Kodierung achten: UTF-8
- Daten im System anschauen (Umlaute, Satztrennung, ...)
- auf gleiche Verarbeitung aller Daten achten (z.B. Input/Output)
- prüfen, ob die Sprache vernünftig erkannt wird



The screenshot shows a document viewer interface. The main text area contains a paragraph about infrastructure for gas transport and a list of references. The sidebar on the left shows a search bar and a list of documents. The right-hand panel shows a table with columns for Type, Subtype, and Role, and a list of entities: Datum, Gericht, and Person.

Infrastruktur für den sicheren und wirtschaftlichen Transport von gasförmigen, aber für den Pipeline transport druckverfüssigem Ethylen zwischen wichtigen süddeutschen Chemiestandorten zu schaffen und diese an den bestehenden nordwesteuropäischen Rohrleitung verbund anzubinden.

10 Die Ethylen-Pipeline-Süd soll nach Vorstellung der Europäischen Kommission eine zentrale Rolle für ein europaweites Ethylenpipeline-Netz spielen (vgl.

11 Entscheidung der Kommission vom 12.

12 Oktober 2006 über die staatliche Beihilfe C 11/2005 -ex N 21/2005-, die Deutschland für den Bau einer Ethylenpipeline in Bayern gewähren will, Amtsblatt der Europäischen Union L 163 vom 6.

13 Juni 2007, S. 24).

14 Die Rohrleitung mit einem Durchmesser von 25 cm und einer Verlegung tiefe zwischen 1 m und 1,2 m führt über eine Länge von circa 370 km von Münch münster in Bayern durch Baden-Württemberg nach Ludwigshafen in Rheinland-Pfalz.



# Lösungsansätze

## Best Practices



Universität Hamburg  
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- keine Daten in Git veröffentlichen!
- keine Service-Login Daten veröffentlichen!  
besser: Logins aus Dateien auslesen
- nice-to-have: korrekte Lizenzen ausweisen

# Lösungsansätze

## Best Practices MapReduce



Universität Hamburg  
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- keine Daten auf ltheadnode speichern
- Mapper sollten nur wenige Minuten laufen
  - Der Task kann optimal auf die Nodes aufgeteilt werden
  - Andere User können parallel arbeiten
  - Bei großen Daten ist es OK, wenn ein paar Mapper crashen; Toleranzen können eingestellt werden

# Lösungsansätze

## Weitere Ressourcen



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

- JSON Lint

<http://jshint.com/>

- JSON Validator/Formatter

<https://jsonformatter.curiousconcept.com/>

Weiterer Ablauf

# Weiterer Ablauf

## Scrum

- agile Softwareentwicklung  
<https://de.wikipedia.org/wiki/Scrum>
- schnelle Iterationen
- Kundenwünsche entstehen durch die Interaktion
- wöchentliche kurze Meetings:
  - was wurde gemacht?
  - was mache ich heute?
  - gibt es aktuell Blocker?
- ER als Scrum-Master, immer auf dem aktuellen Stand

# Weiterer Ablauf

## Lernziele



Universität Hamburg  
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Arbeit mit großen Daten
- MapReduce
- Darstellung von Daten
- Präsentationen
- Projektarbeit

# Weiterer Ablauf

## Bewertung



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Qualität der Programme (Funktionalität, Benutzbarkeit, Code, Dokumentation)
- Teilnahme an Meetings
- Präsentationen
- Abschlussbericht (Teilnehmer kennzeichnen, welcher Teil jeweils von ihnen gemacht wurde)

# Weiterer Ablauf

## Hilfe



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Mattermost:

`https:`

`//mattermost.informatik.uni-hamburg.de/signup_user_complete/?id=t6dnzjjs1tbkfnco8y7d43swre`

- per Mail so gut wie immer erreichbar

- qualifizierte Fragen:

`https://blog.codinghorror.com/rubber-duck-problem-solving/`



Qualifizierte Fragen?