Wetterdatenanalyse mit SparkR

Max Rossbach
Ying-Chi Lin
Dan Häberlein

Universität Leipzig

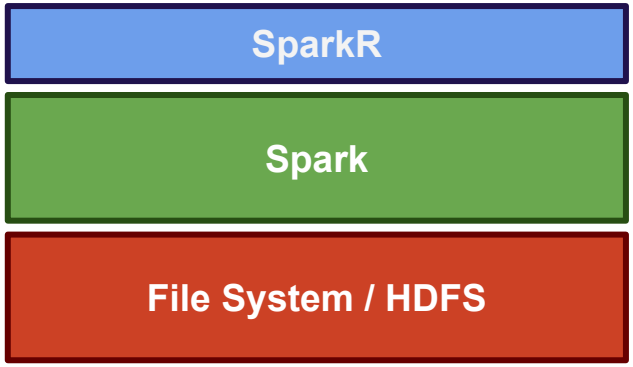
Datensatz-Analyse

- **Deutscher Wetterdienst (DWD)**:
  - Wetterdaten von hiesigen Wetterstationen
  - Zeitraum: 1781 bis heute
  - Verschiedene Datensätze mit unterschiedlicher, vorqualifizierter DQ:
    - verschiedene Messzeiträume (stündlich, täglich, monatlich, jährlich)
    - historische (1781-2015) / aktuelle Datensätze (2014-02/2016!)
    - Niederschlags-, Bodentemperatur-, Strahlungs-, Stationsdaten

- **Werkzeuge / Artefakte:**
  - Groovy Scripting Language → Datenvorbereinigung
  - Spark → Ausführung der Anlyseprogramme
  - SparkR API / R → statistische Analyse / Visulalisierung
  - Webanwendung (RIA) → interaktive Visulalisierung der Ergebnisse

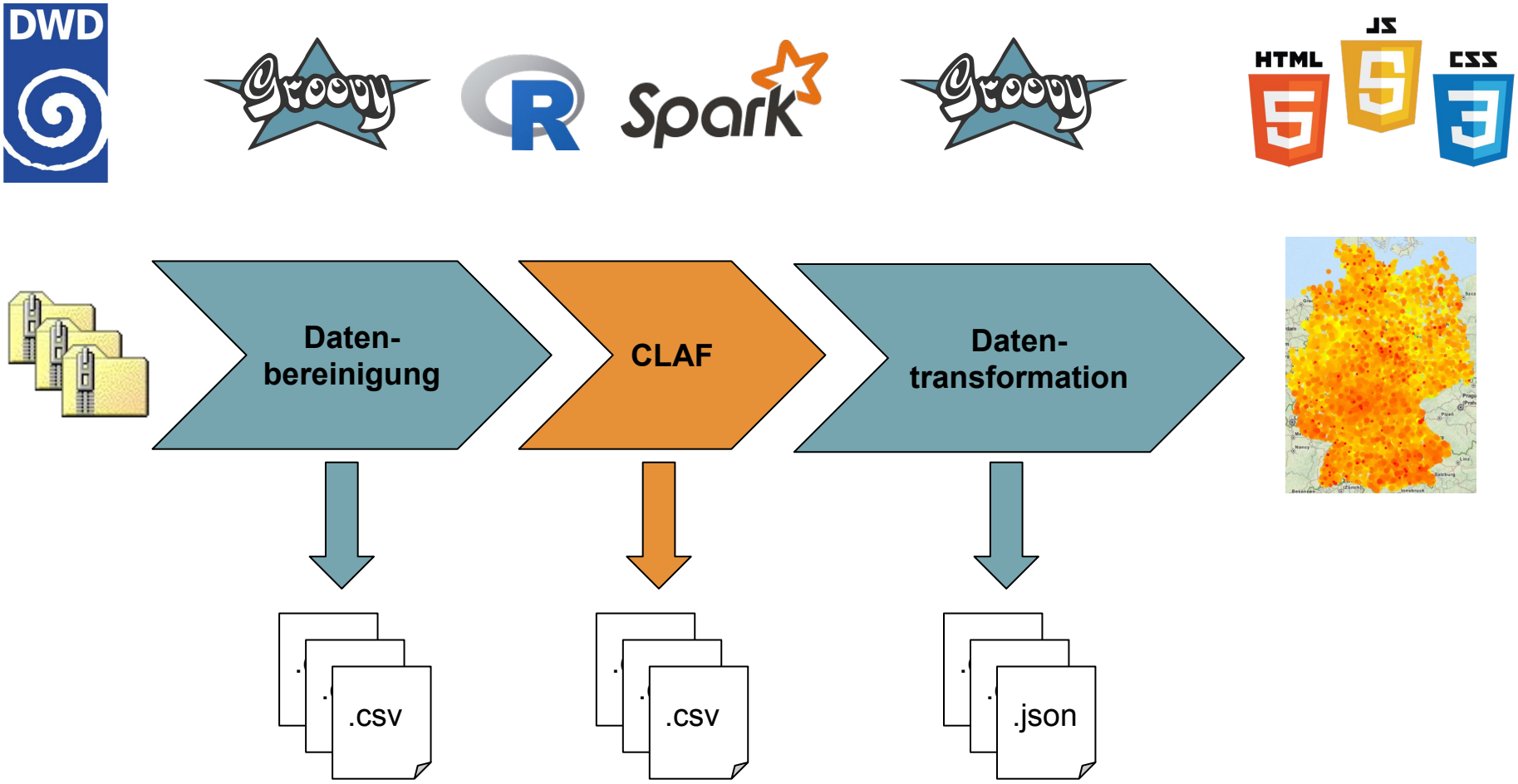
SparkR API

- **Spark**
  - alternative Cloudcompunting-Umgebung zu Hadoop Map/Reduce
  - Nutzung einer Abstraktion (RDD) für verteilten Speicher
  - funktionale Operatoren / SQL basierte APIs zur Manipulation
  - Ziel: Ausnutzung des Cluster-Hauptspeichers zur Beschleunigung der verteilten Programme
- **R**
  - Programmiersprache zur Beantwortung statistischer Fragestellungen
  - Diagramme zur Visualisierung



- **SparkR**
  - API für R um Spark Programme zu formulieren
  - aktuell in Version 1.6.0
  - Nutzung von Version 1.5.1 im Projekt
  - Fokus auf Datenanalysten und Fachanwender

Analyseprozess



Analyse der Bauernregeln

Für Ja/Nein-Regeln

- **Ziel**: Gültigkeit in Prozent pro Station
- **Grundlage**: Durchschnitt der Variablen über jeweiligen Zeitraum
- **Beispiele**:
  - Gefriert in der ersten November-dekade schon das Wasser, wird der Januar umso nasser.

01.11-10.11	01.01.-31.01.
Lufttemperatur am Erdboden	Niederschlagshöhe

Region	Prozentsatz	Total Jahre / Region	Regel Erfüllt Jahre / Region
De	56,02	9922	5559
N	58,72	1548	909
O	61,50	1603	986
S	55,37	4231	2343
W	52,00	2540	1321

Analyse der Bauernregeln

Für Korrelations-Regeln

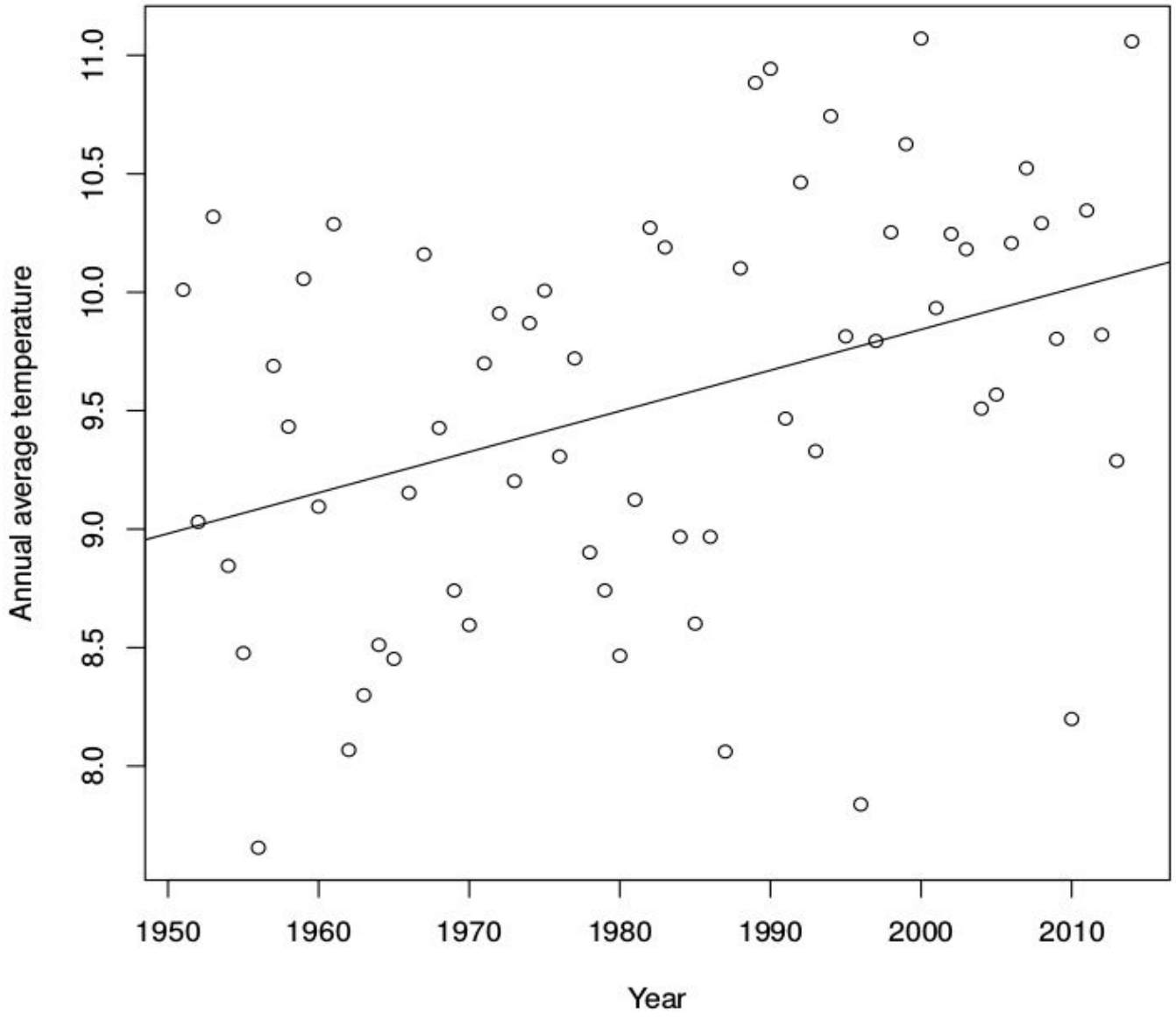
- **Ziel**: Korrelationskoeffizient pro Station
- **Grundlage**: Durchschnitt der Variablen über jeweiligen Zeitraum

- **Beispiele**:
  - Wie's Wetter am Siebenschläfertag, so bleibt es sieben Wochen danach.

26.06 - 28.06	29.06 - 17.08
Lufttemperatur Sonnenscheindauer Niederschlagshöhe	Lufttemperatur Sonnenscheindauer Niederschlagshöhe

Region	Sonne				Regen				Temperatur					
	max	min	> 0.5	> 0.4	max	min	> 0.5	> 0.4	max	min	> 0.5	> 0.4	%	# Station
De	0.599	-0.070	4	14	0.442	-0.319	0	5	0.594	-0.005	31	101	44.3	228
N	0.472	0.086	0	4	0.282	-0.097	0	0	0.573	0.324	17	39	81.2	48
O	0.409	-0.053	0	1	0.151	-0.164	0	0	0.533	0.170	2	18	38.3	47
S	0.447	-0.070	0	1	0.442	-0.319	0	2	0.525	-0.005	2	12	15.2	79
W	0.599	-0.006	4	8	0.435	-0.278	0	3	0.594	0.288	10	32	59.3	54

Regionale Analyse



Referenzen

[1] **Bauernregeln von wetter.de**
http://www.wetter.de/bauernregeln
[2] **Spark + R**
http://deepsense.io/spark-r-sparkr/
[3] **Die beste Bauenregel für jeden Tag**
Jürik Müller, BLV-Verlag München, 2014
[4] **SparkR API**
https://spark.apache.org/docs/1.5.2/sparkr.html
[5] **Deutscher Wetterdienst FTP**
http://www.dwd.de/DE/klimaumwelt/cdc/cdc_node.html
[6] **R Projekt**
https://www.r-project.org/