24. SAS CLUB

WIEN, 18. NOVEMBER 2014
25HOURS-HOTEL

AGENDA

- **Mehr als linear oder logistisch? – Quantils Regressionen und Adaptive Splines in SAS**
  Dr. Mihai Paunescu

- **Moderne SAS Analytik Architekturen**
  Mag. Gernot Engel, Ing. Phillip Manschek,
  DI Rainer Sternecker, Dr. Gerhard Svolba

- **Eine Liebeserklärung an den SAS Enterprise Guide**
  Mag. Bernadette Fabits

- **SAS Analytics 4 U – Neue Möglichkeiten für Studierende und Dozenten**
  DI (FH) Bettina Brandl

sas club

Der Business Analytics Club für SAS User

MEHR ALS LINEAR ODER LOGISTISCH?

QUANTILS REGRESSIONEN UND ADAPTIVE SPLINES IN SAS

MIHAI PAUNESCU

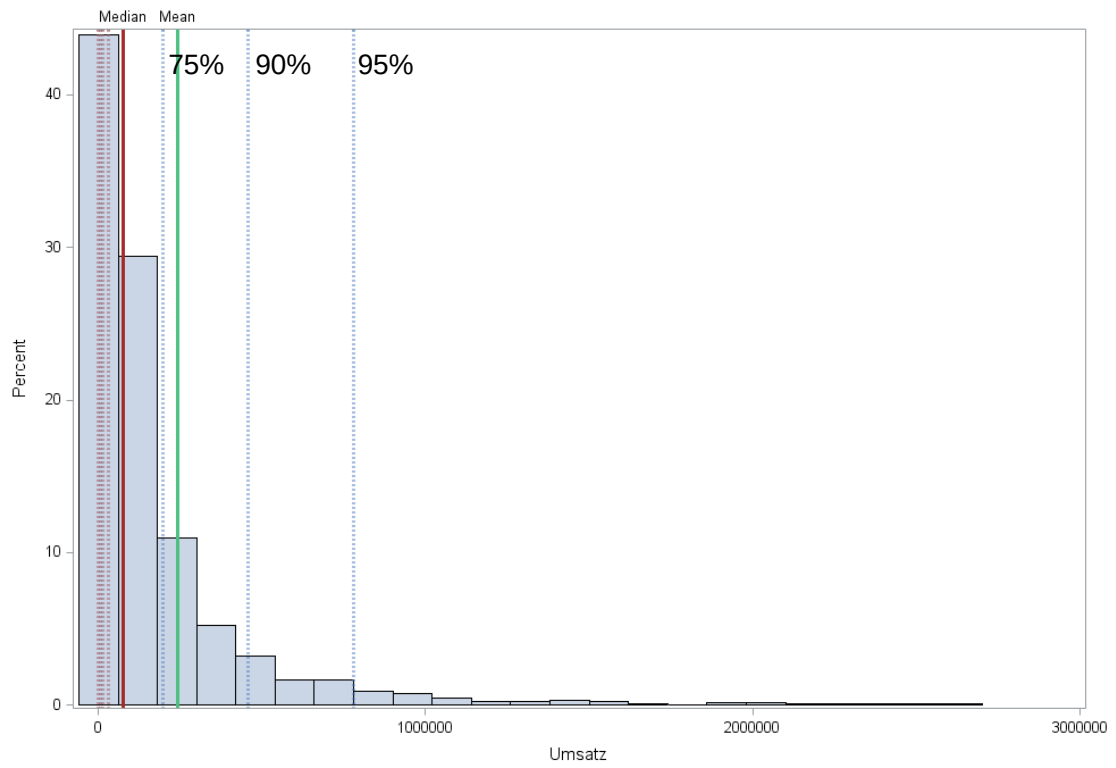
QUANTILE


```
proc univariate data=dat;
 ods select moments quantiles;
 var sales;
run;
```

Basic Statistical Measures			
Location		**Variability**	
Mean	241825.8	**Std Deviation**	823581
Median	**75875.1**	**Variance**	6,78E+16
Mode	.	**Range**	18605198
		Interquartile Range	168416

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	18605230.40
99%	2699730.30
95%	779877.70
90%	458327.90
75% Q3	198564.00
50% Median	**75875.10**
25% Q1	30147.60
10%	13607.10
5%	5598.60
1%	500.75
0% Min	32.55

QUANTILE IM HISTOGRAMM



Quantiles (Definition 5)	
Level	**Quantile**
100% Max	18605230.40
99%	2699730.30
95%	779877.70
90%	458327.90
75% Q3	198564.00
50% Median	75875.10
25% Q1	30147.60
10%	13607.10
5%	5598.60
1%	500.75
0% Min	32.55

```
ods graphics on / width=1000px height=700px;
proc sgplot data=dat;
 where sales < 2699730.30;
 histogram sales;
 refline 779877.70 458327.90 198564.00 / axis=x lineattrs=(color=bigb pattern=dot thickness=3);
 refline 30147.60 13607.10 5598.60 500.75 /axis=x lineattrs=(color=dapk pattern=dot thickness=3);
 refline 75875.10 /axis=x label='Median' lineattrs=(color=brown thickness=3);
 refline 241825.8 /axis=x label='Mean' lineattrs=(color=big thickness=3);
run;
```

PROC QUANTREG UND MEDIAN


```
proc quantreg data=dat;
  model Sales= / quantile=(0.5);
run;
```

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	75875.10	2.654.220	70.670.691	81.079.509	28.59	<.0001

Quantiles (Definition 5)

Level	Quantile
100% Max	18605230.40
99%	2699730.30
95%	779877.70
90%	458327.90
75% Q3	198564.00
50% Median	75875.10
25% Q1	30147.60
10%	13607.10
5%	5598.60
1%	500.75
0% Min	32.55

KLASSIFIKATIONSVARIABLEN


```
proc quantreg data=dat ci=resampling ;
 class cust_grp;
 model Sales=cust_grp / quantile=(0.5) seed=12345;
run;
```

Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept		1	47335.60	3.071.371	41.313.237	53.357.963	15.41	<.0001
Cust_grp	I	1	88078.60	6.903.957	74.541.278	101615.92	12.76	<.0001
Cust_grp	O	1	21870.35	5.324.410	11.430.214	32.310.486	4.11	<.0001
Cust_grp	N	0	0.0000	0.0000	0.0000	0.0000	.	.

```
proc means data=dat P1 median P90;
 var sales;
 class Cust_grp;
run;
```

Analysis Variable : Sales

Cust_grp	N Obs	Median
I	793	135414.2
O	1024	69062.08
N	1004	47323.3

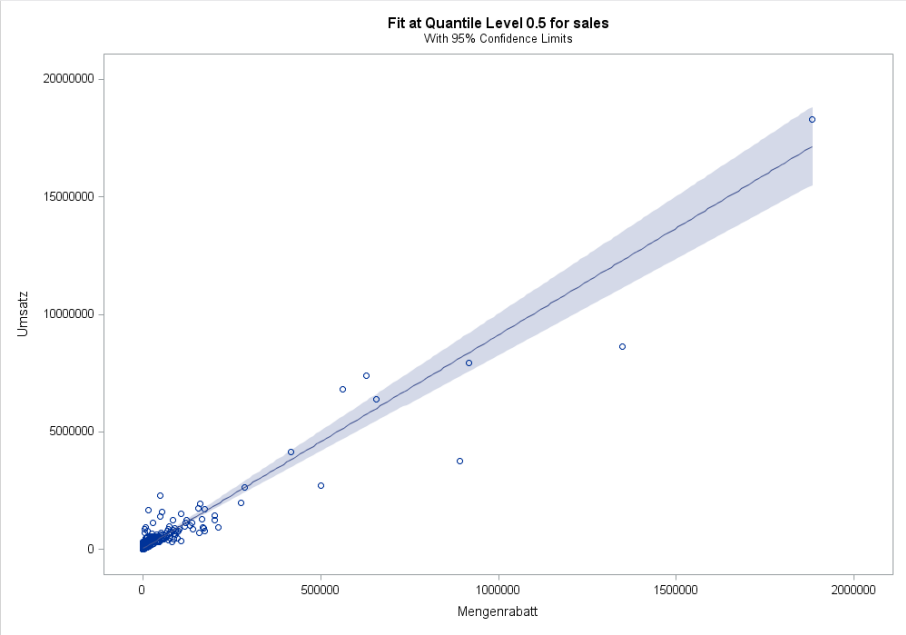
47335.6 + 88078.6 = 135414.2

MEDIAN SCHÄTZUNG FÜR INTERVALVARIABLE

```
proc quantreg data=dat ci=resampling plots=(fitplot);
 where cust_grp='n';
 model Sales=Rabatt_Menge / quantile=(0.5) seed=12345;
run;
```

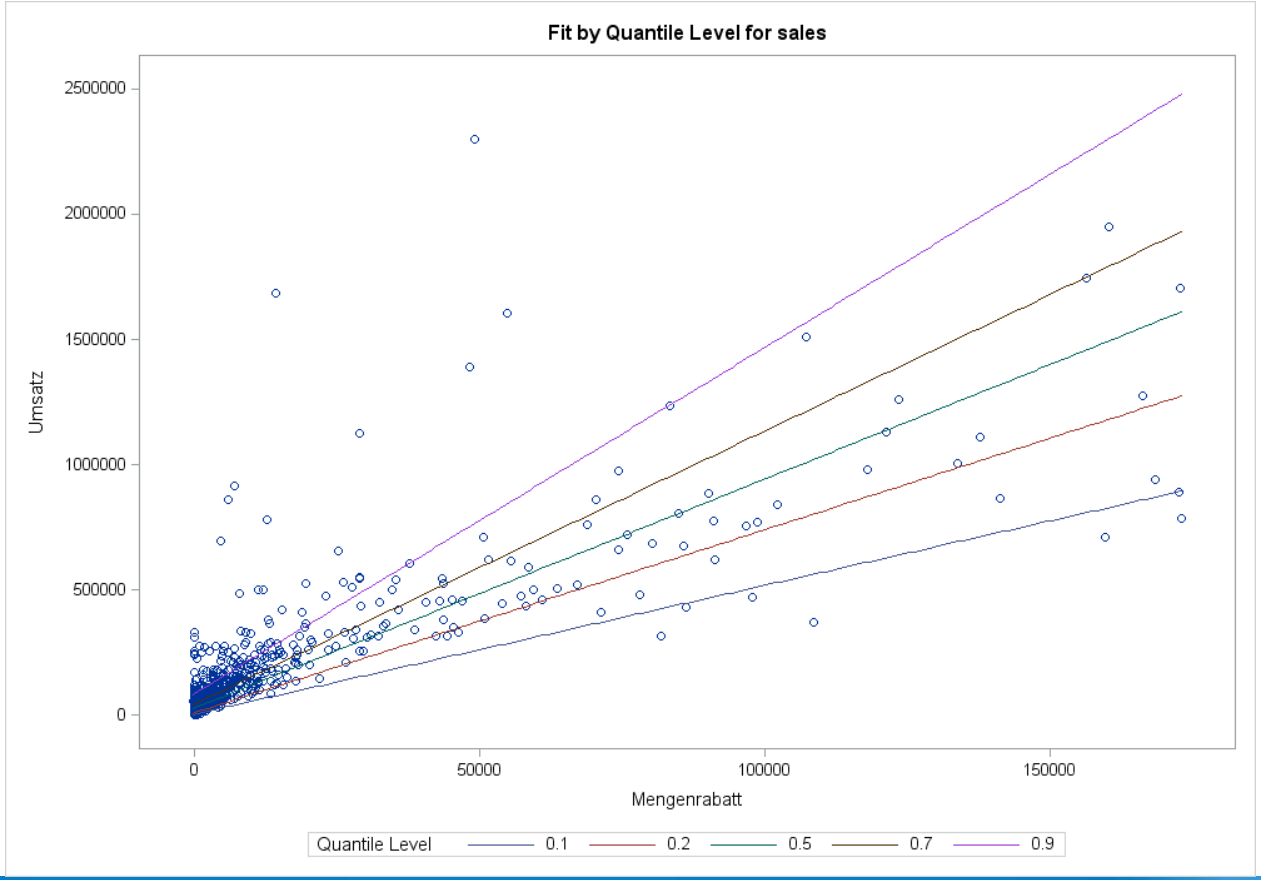
Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	28552.1	5873.371	17026.581	40077.615	4.86	<.0001
Rabatt_Menge	1	9.1	0.4522	8.2082	9.9829	20.12	<.0001



QUANTILSCHÄTZUNG FÜR INTERVALVARIABLE

```
proc quantreg data=dat ci=resampling plots=(fitplot);
 ods select fitplot;
 where cust_grp='n' and Rabatt_Menge < 200000;
 model Sales=Rabatt_Menge / quantile=(0.1 0.2 0.5 0.7 0.9)  seed=12345;
run;
```



13.83

10.9

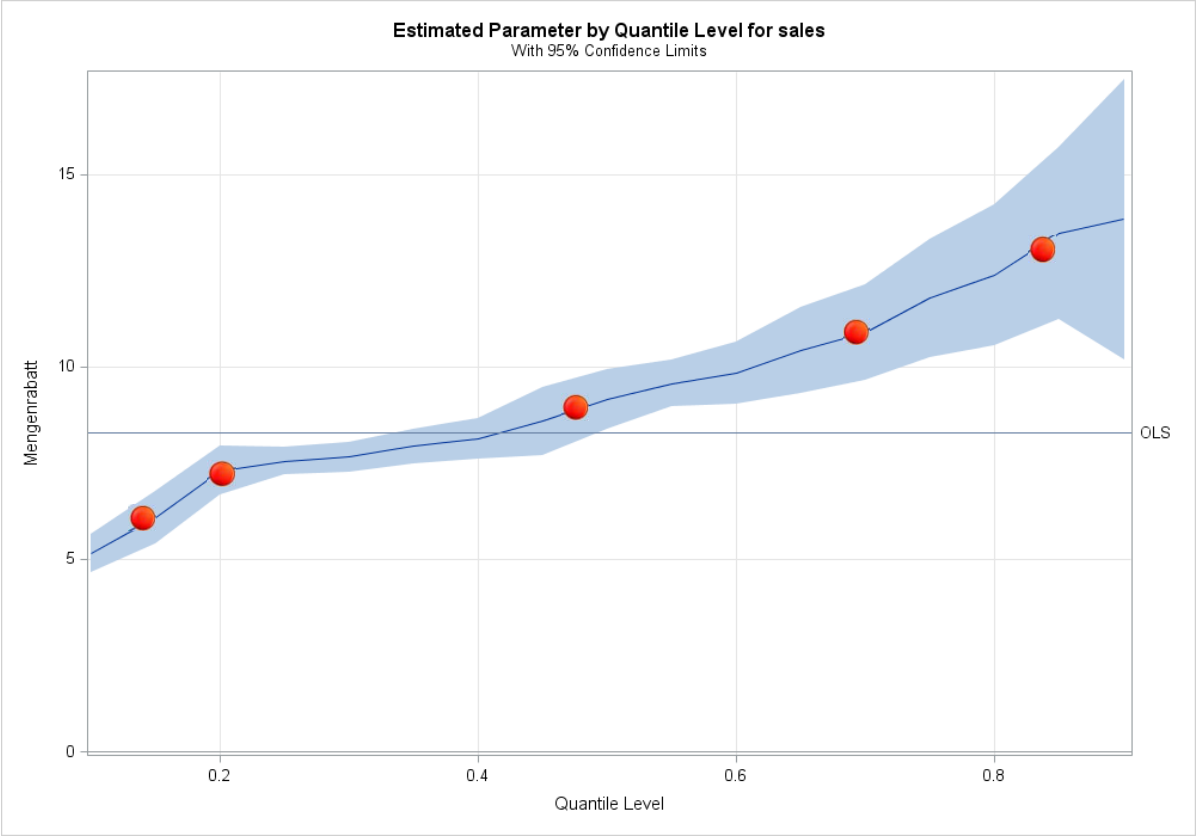
9.14

7.31

5.15

QUANATILE PROZESS PLOT

```
proc quantreg data=dat ci=resampling;
ods select quantplot;
where cust_grp='n' and Rabatt_Menge < 200000;
model Sales=Rabatt_Menge
/quantile=(0.1 to 0.9 by 0.05)  plot=(quantplot /unpack ols) seed=1268 ;
run;
```



13.83

10.9

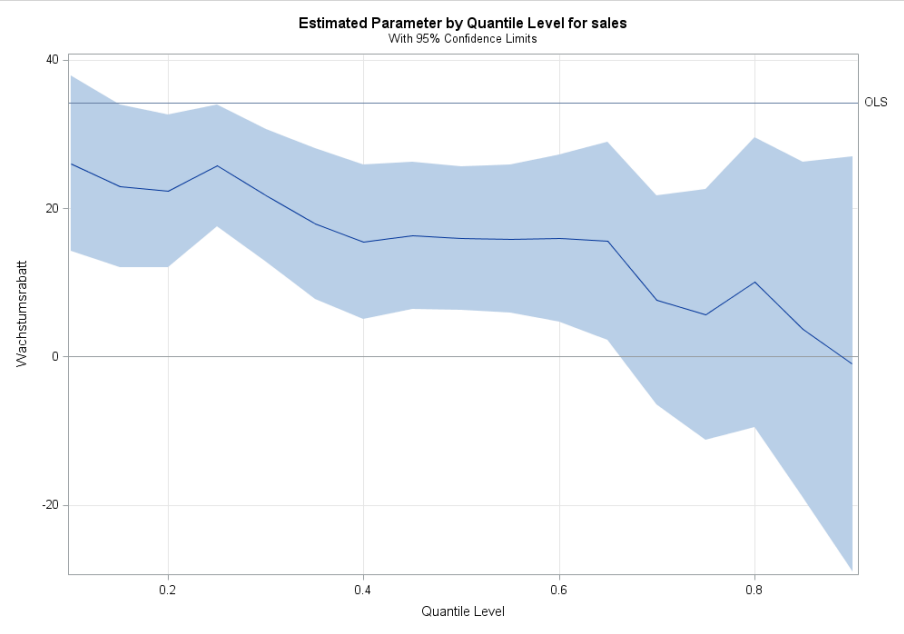
9.14

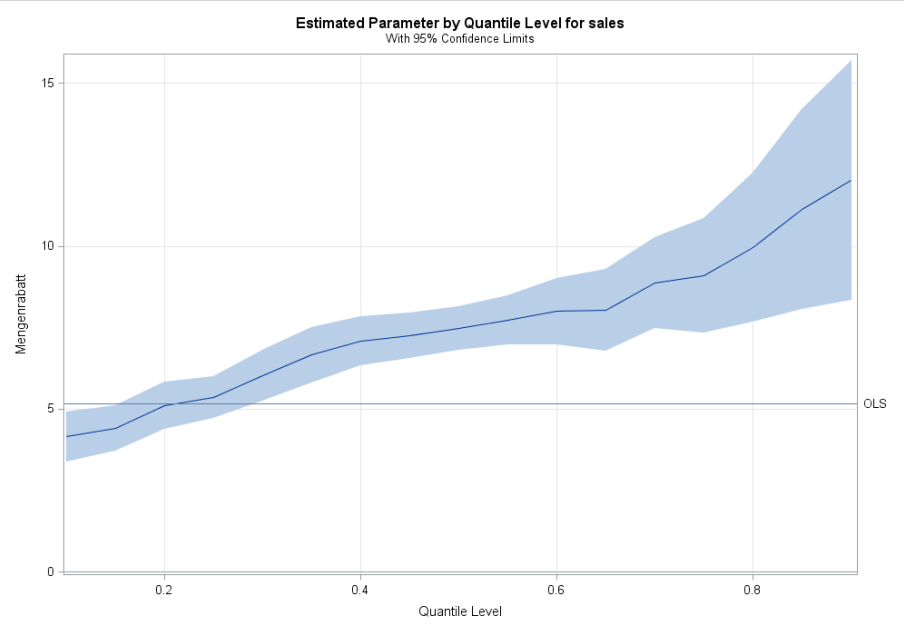
7.31

5.15

MULTIVARIATE QUANTILSREGRESSION

```
proc quantreg data=dat ci=resampling;
where cust_grp='n';
 model Sales=Rabatt_Wachstum Rabatt_Menge  /
 quantile=(0.1 to 0.9 by 0.05)  plot=(quantplot /unpack ols) seed=1268 ;
run;
```



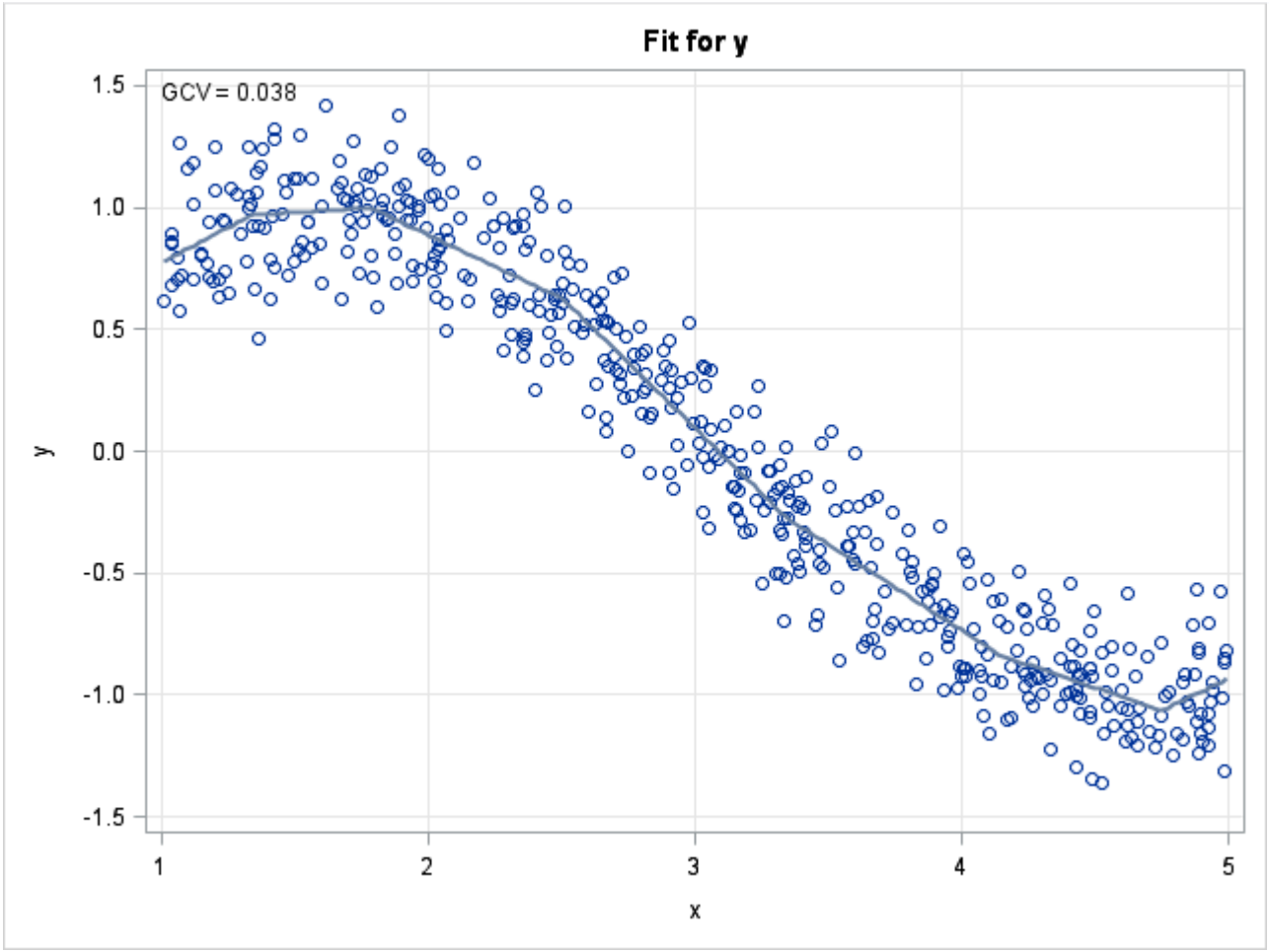


FAZIT

- Verwendung:
  - Zusammenhänge entdecken für extreme Bereiche der Zielvariable
  - Robuste Medianschätzung gegenüber Ausreißer, ohne Verteilungsannahmen
- Hinweise:
  - Ist nicht äquivalent zu linearen Regressionen für Segmente von Beobachtungen
  - Schneller: HPQUANTSELECT ab SAS/STAT 13.2

ADAPTIVEREG

MULTIVARIATE ADAPTIVE REGRESSION SPLINES



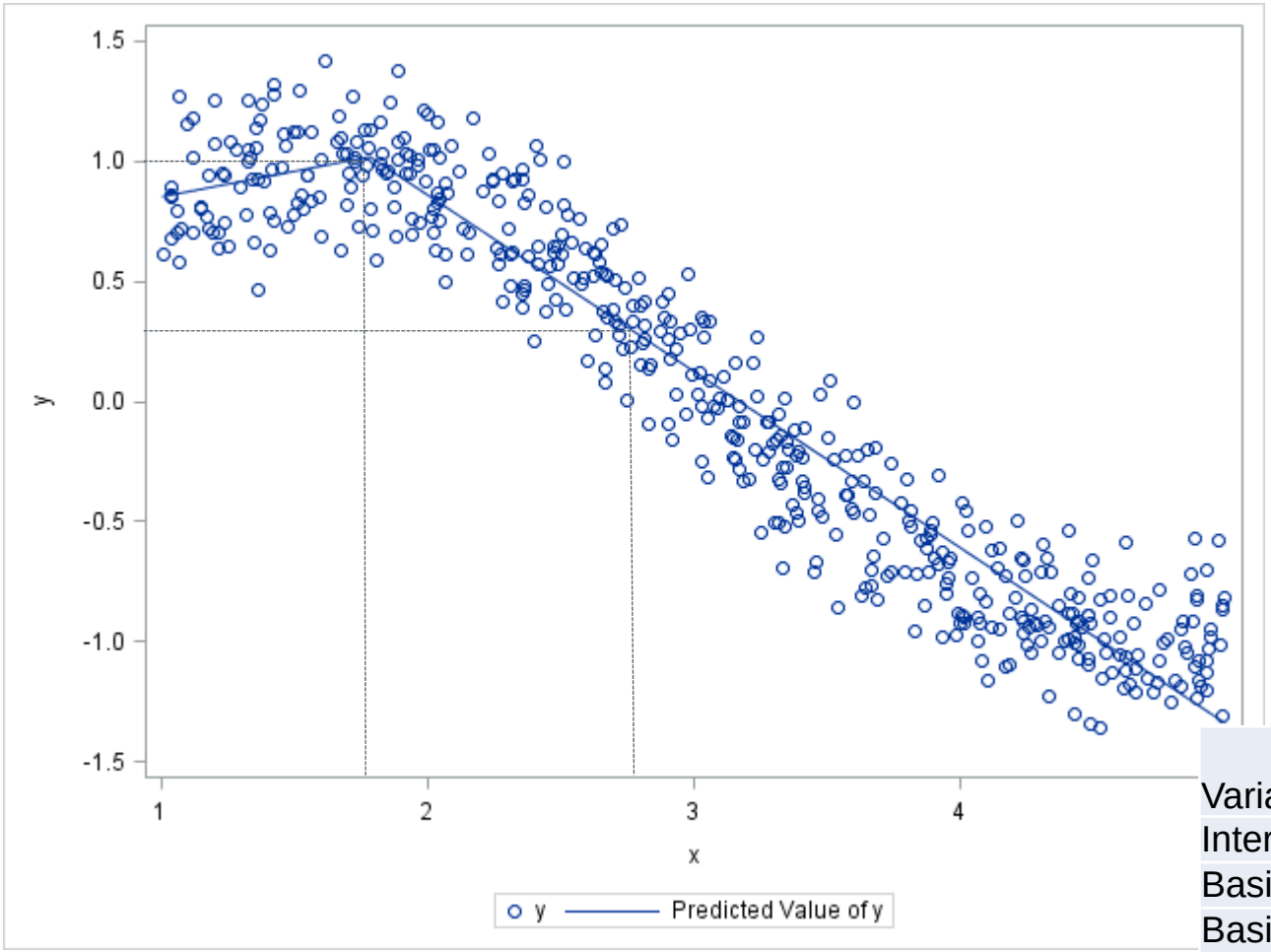

```
proc adaptivereg
plots=all
details=bases;

model y = x;

run;
```

ADAPTIVEREG

SELEKTION VON GEEIGNETEN KNOTEN WERTEN



Basis1	MAX(x - 1.8,0)
Basis2	MAX(1.8 - x,0)

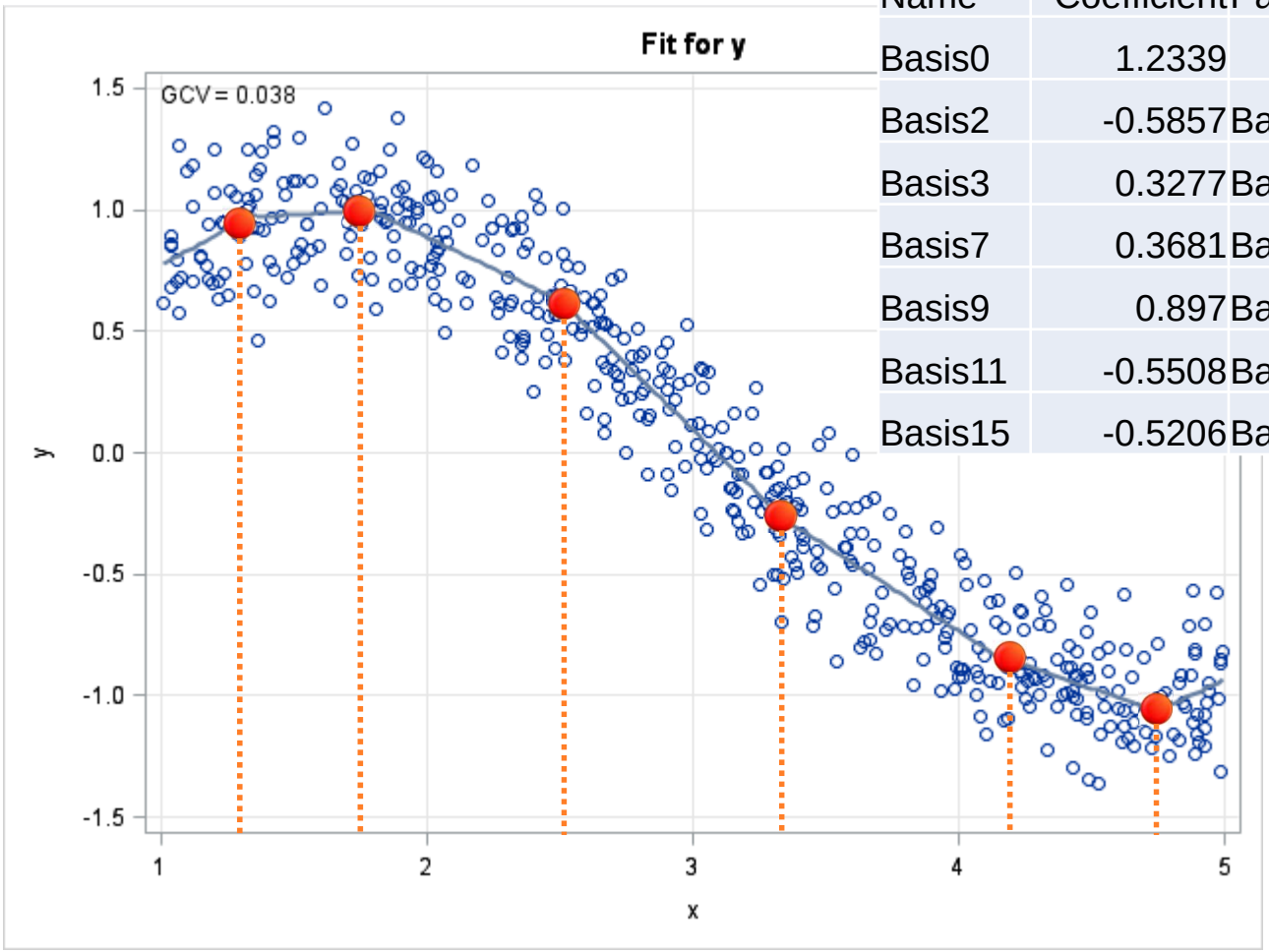
```
data ds2;
set ds;
Basis1 = max(x-1.8, 0);
Basis2 = max(1.8 - x, 0);
run;

proc reg data=ds2;
 model y = basis1 basis2;
run;
```

Variable	Parameter Estimate	t Value	Pr > |t|
Intercept	1.02	52.04	<.0001
Basis1	-0.73	-67.82	<.0001
Basis2	-0.21	-3.31	0.001

ADAPTIVEREG

REDUKTION DES MODELS DURCH SELEKTIONSMETHODEN



Regression Spline Model after Backward Selection				
Name	Coefficient	Parent	Variable	Knot
Basis0	1.2339		Intercept	
Basis2	-0.5857	Basis0	x	1.7865
Basis3	0.3277	Basis0	x	4.1447
Basis7	0.3681	Basis0	x	3.3424
Basis9	0.897	Basis0	x	4.7489
Basis11	-0.5508	Basis0	x	2.5061
Basis15	-0.5206	Basis0	x	1.3345

ADAPTIVEREG

GEEIGNET FÜR KLASSIFIKATIONS-VARIABLEN UND MISSING VALUES

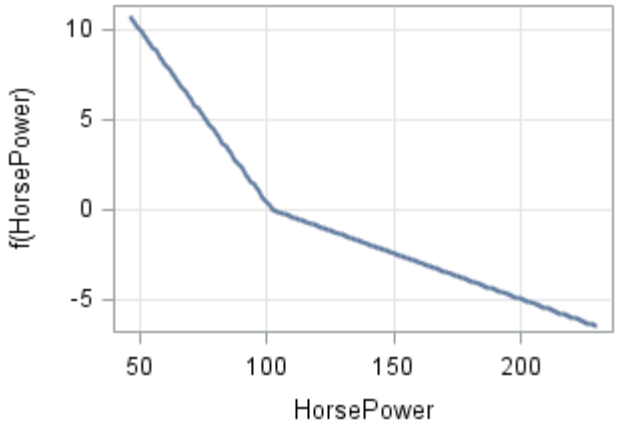
```
proc adaptivereg data=autompg plots=all details=bases;
class cylinders year origin;
model mpg = cylinders displacement horsepower
weight acceleration year origin / additive;
run;
```

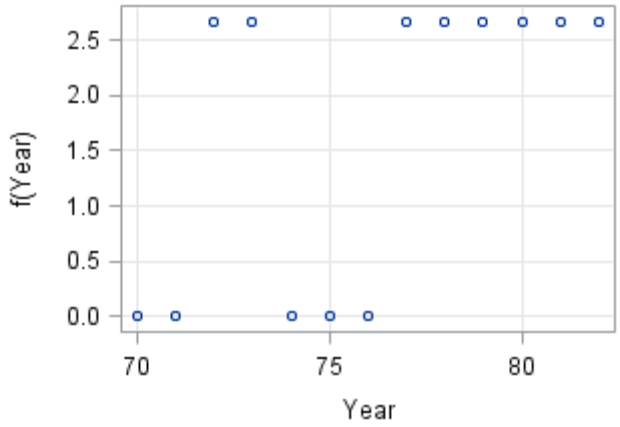
Basis Information

Name	Transformation
Basis0	1
Basis1	Basis0*MAX(Weight - 3139,0)
Basis2	Basis0*MAX(3139 - Weight,0)
Basis3	**Basis0*NOT(MISSING(HorsePower))**
Basis4	**Basis0*MISSING(HorsePower)**
Basis5	**Basis3*MAX(HorsePower - 102,0)**
Basis6	**Basis3*MAX(102 - HorsePower,0)**
Basis7	Basis0*(Year = 80 OR Year = 82 OR Year = 81 OR Year = 79 OR Year = 78 OR Year = 77 OR Year = 73 OR Year = 72)
Basis8	Basis0*NOT(Year = 80 OR Year = 82 OR Year = 81 OR Year = 79 OR Year = 78 OR Year = 77 OR Year = 73 OR Year = 72)
Basis9	Basis0*MAX(Displacement - 85,0)
Basis10	Basis0*MAX(85 - Displacement,0)
Basis11	Basis0*MAX(Displacement - 97,0)
Basis12	Basis0*MAX(97 - Displacement,0)
Basis13	Basis0*MAX(Acceleration - 21,0)
Basis14	Basis0*MAX(21 - Acceleration,0)
Basis15	Basis3*MAX(Displacement - 105,0)
Basis16	Basis3*MAX(105 - Displacement,0)

VISUALISIERUNG DER MODEL-KOMPONENTEN

Basis Information		
Name	Coefficient	Transformation
Basis3	-4.03	Basis0*NOT(MISSING(HorsePower))
Basis4	0	Basis0*MISSING(HorsePower)
Basis5	-0.05	Basis3*MAX(HorsePower - 102,0)
Basis6	0.19	Basis3*MAX(102 - HorsePower,0)
Basis7	2.67	Basis0*(Year = 80 OR Year = 82 OR Year = 81 OR Year = 79 OR Year = 78 OR Year = 77 OR Year = 73 OR Year = 72)
Basis8	0	Basis0*NOT(Year = 80 OR Year = 82 OR Year = 81 OR Year = 79 OR Year = 78 OR Year = 77 OR Year = 73 OR Year = 72)





PROC Adaptivereg
http://support.sas.com/resources/papers/proceedings13/457-2013.pdf

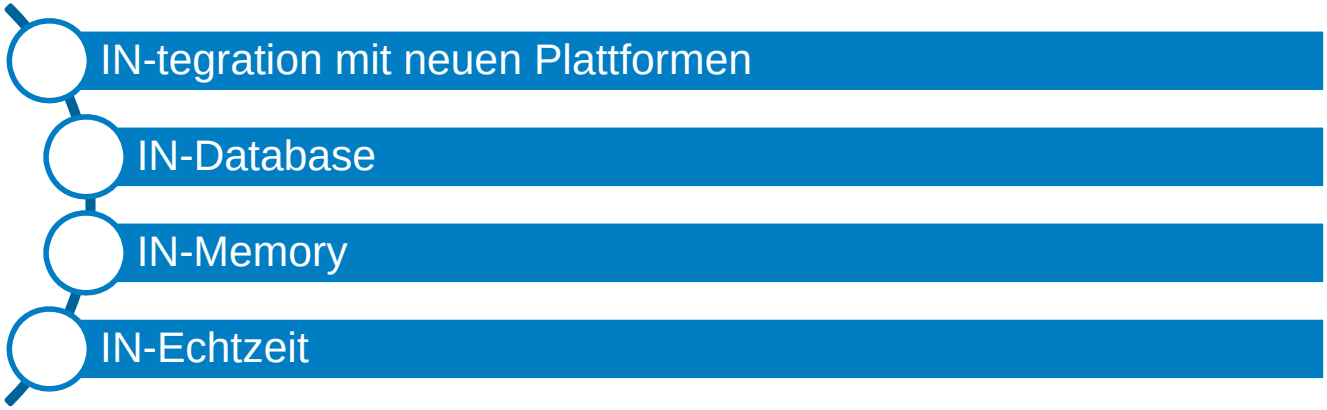
Weitere Nicht-Parametrische Regressionen: PROC GAM, PROC LOESS

sas club

Der Business Analytics Club für SAS User

MODERNE SAS ANALYTIK ARCHITEKTUREN

SAS IST “IN”

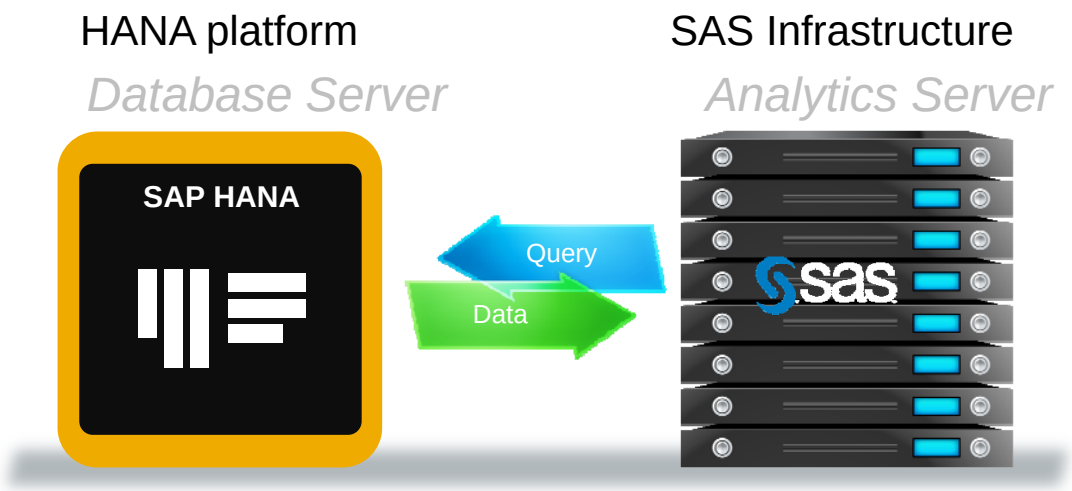


ACCESS SCHNITTSTELLEN

- SAS/ACCESS Interface to
  - PostGreSQL
  - Vertica
  - PI System
  - Impala
  - SAP Hana
  - Hadoop

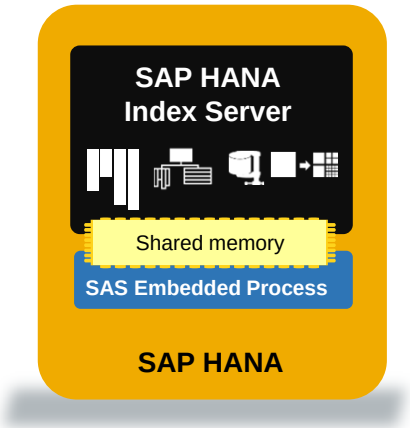
sas club

- SAS/ACCESS Interface to SAP Hana



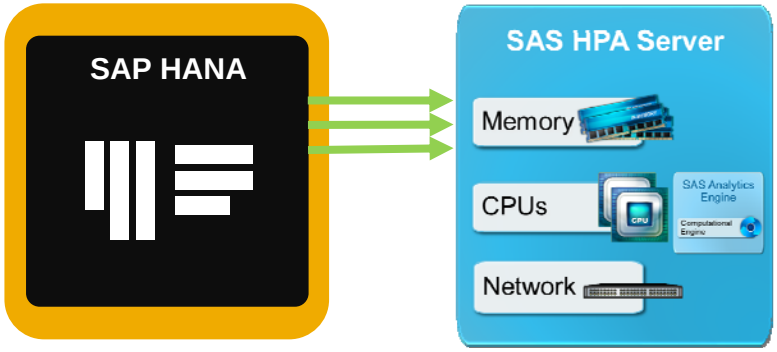
sas club

- SAS Scoring Accelerator for SAP HANA



sas club

- SAS High-Performance Predictive Modeling Workbench for SAP HANA



sas club

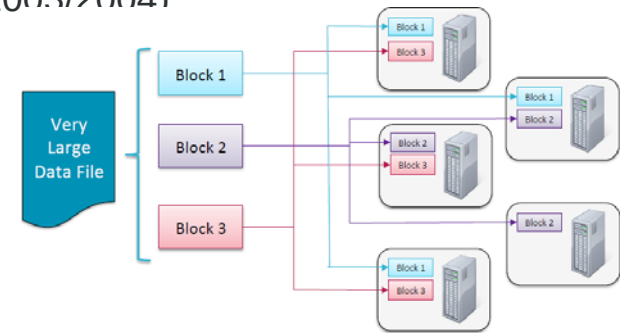
Der Business Analytics Club für SAS User

HADOOP UND SAS – STATUS UND AUSBLICK

WIEN, OKT 2014
GERNOT ENGEL , SAS AUSTRIA

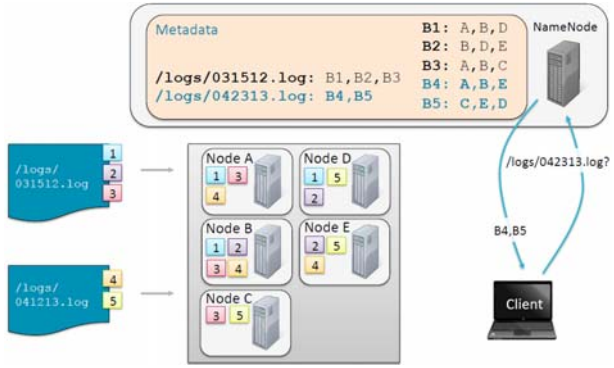
Historie:

- *basiert auf **Google** Forschungspapieren für ein verteiltes Dateisystem(2003/2004)
- *von **Yahoo** für die Entwicklung einer WWW-Suchmaschine aufgegriffen
- *an die **Apache** Foundation übergeben, seitdem open-source (2009)



Technologie:

- Prinzip der **horizontalen Skalierung** auf kostengünstiger Hardware („Scale out“)
- Prinzip der **Datenlokalität**: der Programmcode wird auf die Cluster-Nodes mit den zugehörigen Daten verteilt, dort verteilt ausgeführt und (Teil-)Ergebnisse wieder zusammengeführt
- Structure on read / late binding



- **HDFS**: Hadoop Distributed File System
- **Map/Reduce**: ein Verarbeitungsverfahren nach dem „Teile und herrsche“-Ansatz

Mehr Daten

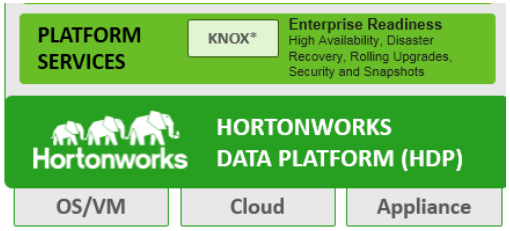
Andere Daten

Kostengünstig

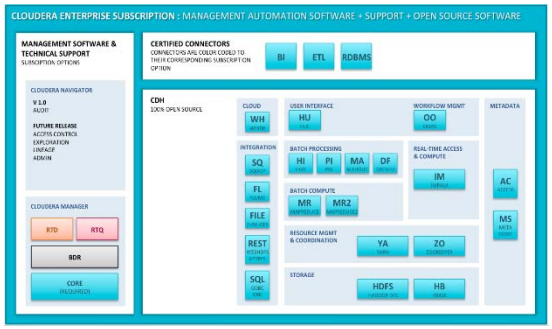
DAS HADOOP ÖKOSYSTEM

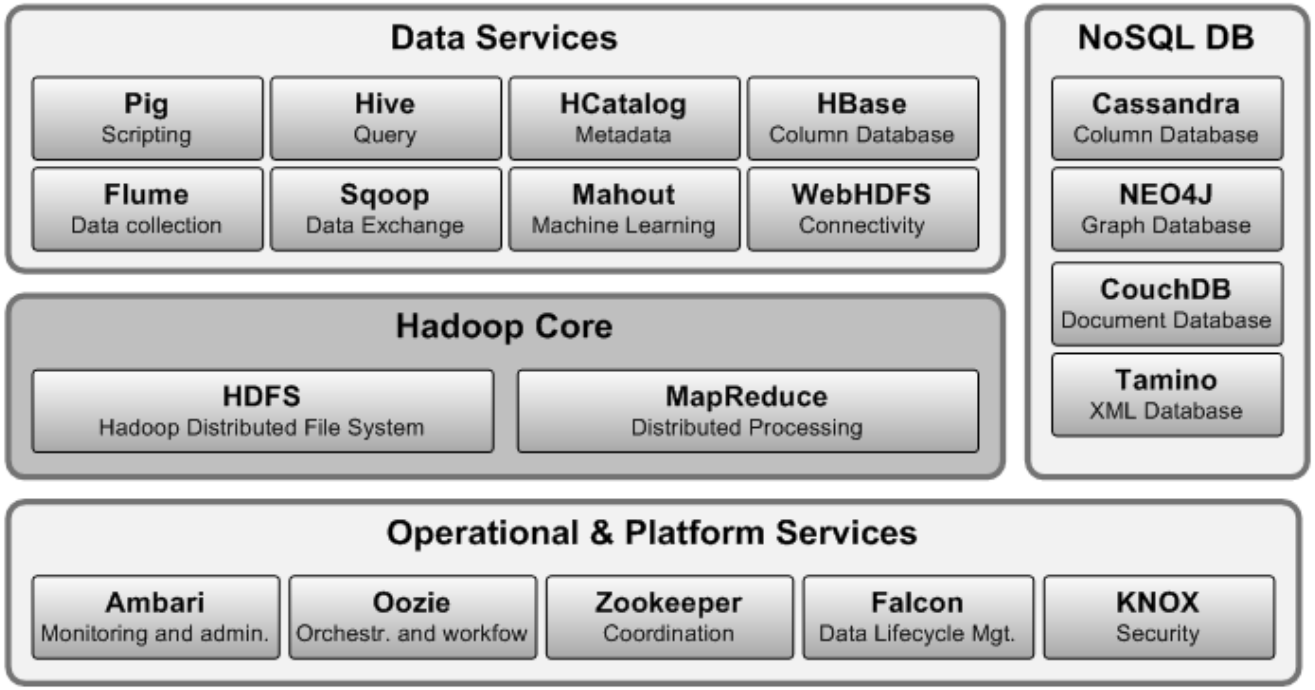
HADOOP IST EINE PROJEKTPLATTFORM MIT HDFS UND MAP/REDUCE ALS KERNMODULEN

Hortonworks HDP



Cloudera CDH

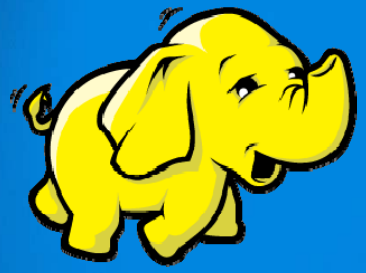




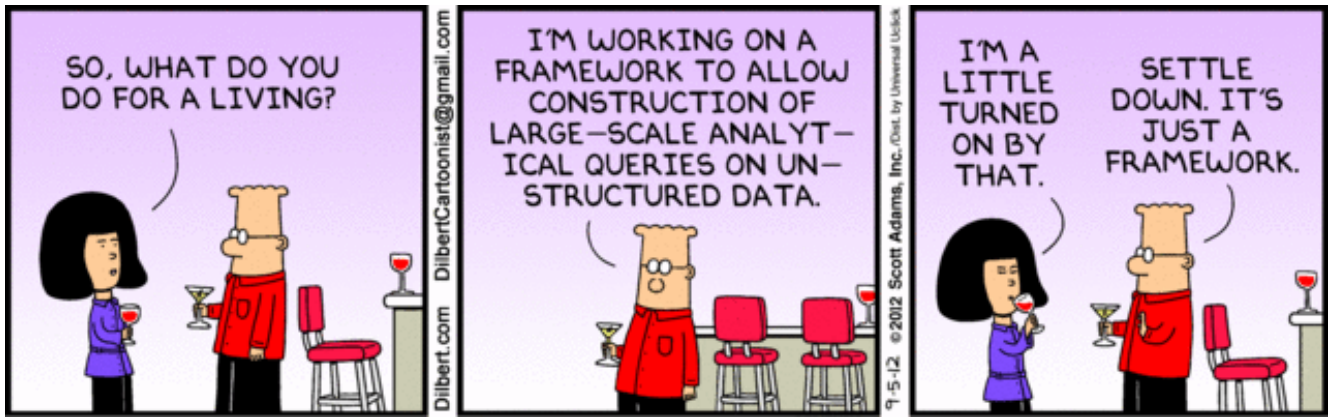
SAS Support
Tier1 : Cloudera, Hortonworks
Tier2 : MapR, Pivotal, IBM ..

SAS FOR HADOOP

VISION - WARUM UNTERSTÜTZT SAS HADOOP?



To be the Analytic and Data Management solution of choice for Hadoop.



sas club

SAS/Access to Hadoop Push *some* SAS processing from Hadoop into SAS

Hive | Impala

SAS/Access to Hadoop
SAS/Access to Cloudera Impala
SAS DI Server

Embedded Process - Push SAS data processing to Hadoop with Map Reduce

Score A | Code A

SAS/Scoring Accelerator for Hadoop
*SAS Code Accelerator for Hadoop **
*SAS Data Quality Accelerator for Hadoop**
*SAS Data Loader(*EA @ 9.4M2)*

In-Memory Analytics - Use Hadoop for Storage persistence and commodity computing.

HPA | LASR

SAS Visual Analytics
SAS Visual Statistics
SAS in memory Statistics for HADOOP
SAS HPA Produkt bundles

HADOOP DATA INTEGRATION PLATFORM

FUNKTIONEN IM DATENMANAGEMENT FÜR HADOOP

Base SAS

- Map Reduce + Pig Scripting + HDFS Kommandos

SAS Access to Hadoop

- Hive, Hive2 + eigene Metadaten („Information Maps für HDFS Dateien“)
- Pushdown von Procs: `FREQ`, `RANK`, `REPORT`, `SORT`, `SUMMARY/MEANS` & `TABULATE`

SAS Access to Impala (Cloudera)

SAS Data Integration Studio Transformationen

- Read/Write HDFS files
- Submit HiveQL code
- Execute Map/Reduce code
- Submit Pig Latin
- Transfer data to/from Hadoop using Hadoop utilities
- SQL transforms pushed down with Access to Hadoop engine
- Submit DQ and DS2 code with Code Accelerator

SAS Federation Server

- Datenvirtualisierung und Zugriffsschutz für Hadoop und andere Datenquellen

SAS Event Stream Processing Engine

- Hadoop Adapter für SAS ESP, um Data-Streams im HDFS zu speichern

```
/* SUBMIT SQL QUERY TO HIVE */
LIBNAME MYHDP HADOOP PORT=10000 SERVER=HDPSRV02;
PROC SQL;
  INSERT INTO MYHDP.CARS_HIVE
    SELECT MAKE, MODEL, MSRP FROM SASHELP.CARS;
QUIT;

/* SUBMIT PIG SCRIPT, HDFS COMMAND, MR JOB */
FILENAME CFG "C:\SAMPLE_DATA\HADOOP_CONFIG.XML";
FILENAME PIGCODE1 "C:\SAMPLE_DATA\PIG_CD.TXT";
PROC HADOOP OPTIONS=CFG;
    PIG CODE=PIGCODE1;
     HDFS DELETE="/USER/HADOOP/OUTPUT_MR1";
     MAPREDUCE INPUT="/.." OUTPUT="/.."
          JAR=".." MAP=".." REDUCE="..";
RUN;

/* COPY FILE FROM HDFS TO LOCAL SAS */
FILENAME CFG "C:\SAMPLE_DATA\HADOOP_CONFIG.XML";
PROC HADOOP OPTIONS=CFG USERNAME="HADOOP"
PASSWORD="XXXX";
HDFS COPYTOLOCAL="/USER/HADOOP/TESTFOLDER"
OUT="C:\SAMPLE_DATA\" ;
RUN;
```

HADOOP DATA INTEGRATION PLATFORM

SAS PRODUKTE FÜR ALLE DREI EBENEN

FROM Hadoop

IN Hadoop

WITH Hadoop

HADOOP DATA INTEGRATION PLATFORM

FUNKTIONEN IM DATENMANAGEMENT IN HADOOP

sas club

IN Hadoop

SAS Scoring Accelerator for Hadoop
Ausführen von Scoring Modellen aus EM und STAT Projekten

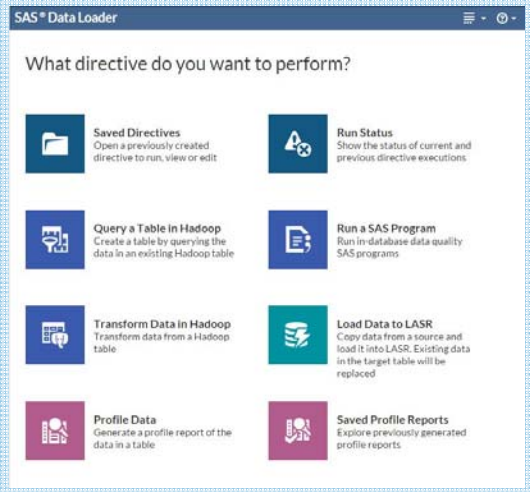
SAS Data Quality Accelerator for Hadoop (EA in 9.4M2)
Ausführen von DQ Routinen (Parse, Standardize, Gender Analysis, Identification, Match Code…)

SAS Code Accelerator for Hadoop
Ausführen von DataStep2 Code

SAS Data Loader for Hadoop (EA in 9.4M2)
Web-basiertes In-Database Tool für Endanwender
(z.B. aus den Fachbereichen)

Steckbrief SAS Data Loader 2.1

- Point & Click Datenmanagement-Routinen, die in Hadoop ausgeführt werden (Abfragen, Tabellen anlegen, SAS Programme)
- In-Memory Beladung: Transfer von Hadoop-Daten in den LASR Server
- “Profile Data” für Hadoop-Daten
- HTML5 basiertes Interface
- Wird als vApp ausgeliefert (läuft in VMWare Player)



Roadmap
- Joins in Hadoop
- Sqoop Support
- Data Quality Verfahren
- Metadaten / Lineage
- Security
- Monitoring

sas club

SAS/Access to Hadoop Push *some* SAS processing from Hadoop into SAS

Hive | Impala

SAS/Access to Hadoop
SAS/Access to Cloudera Impala
SAS DI Server

Embedded Process - Push SAS data processing to Hadoop with Map Reduce

Score A | Code A

SAS/Scoring Accelerator for Hadoop
*SAS Code Accelerator for Hadoop **
*SAS Data Quality Accelerator for Hadoop**
*SAS Data Loader(*EA @ 9.4M2)*

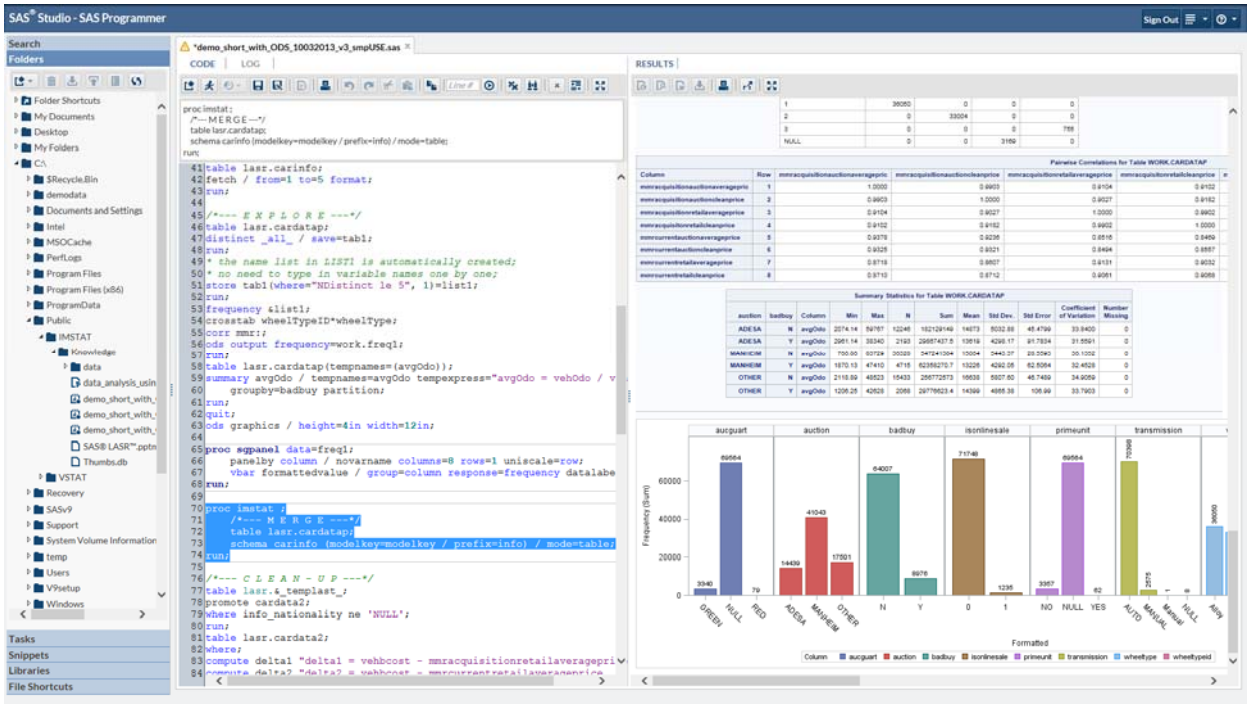
In-Memory Analytics - Use Hadoop for Storage persistence and commodity computing.

HPA | LASR

SAS Visual Analytics
SAS Visual Statistics
SAS in memory Statistics for HADOOP
SAS HPA Produkt bundles

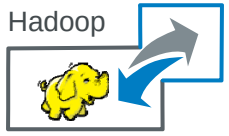
SAS® IN MEMORY
STATISTICS FOR HADOOP

ANALYTICS IN MEMORY AUF HADOOP DATEN

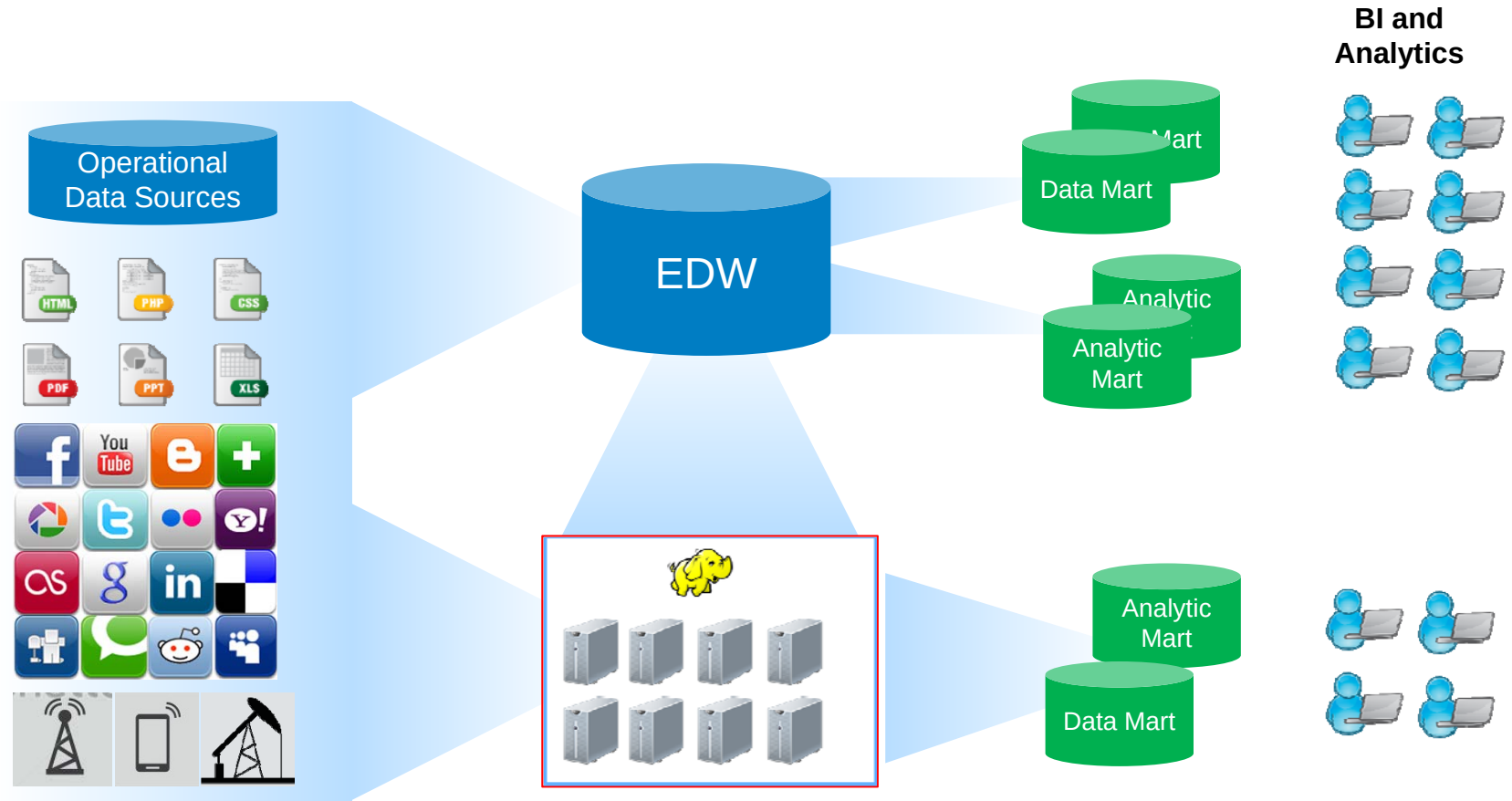


- Linear Regression
- Multiple Regression
- Logistic Regression
- Analysis of Variance
- K-means Clustering
- Decision Trees
- Random Woods
- Generalized Linear Models
- Density-based Spatial Clustering (DBSCAN)
- Text pre-processing
- Text parsing
- Singular Value Decomposition
- Topic analysis
- Associations
- Implicit and explicit recommendations
- Group By processing

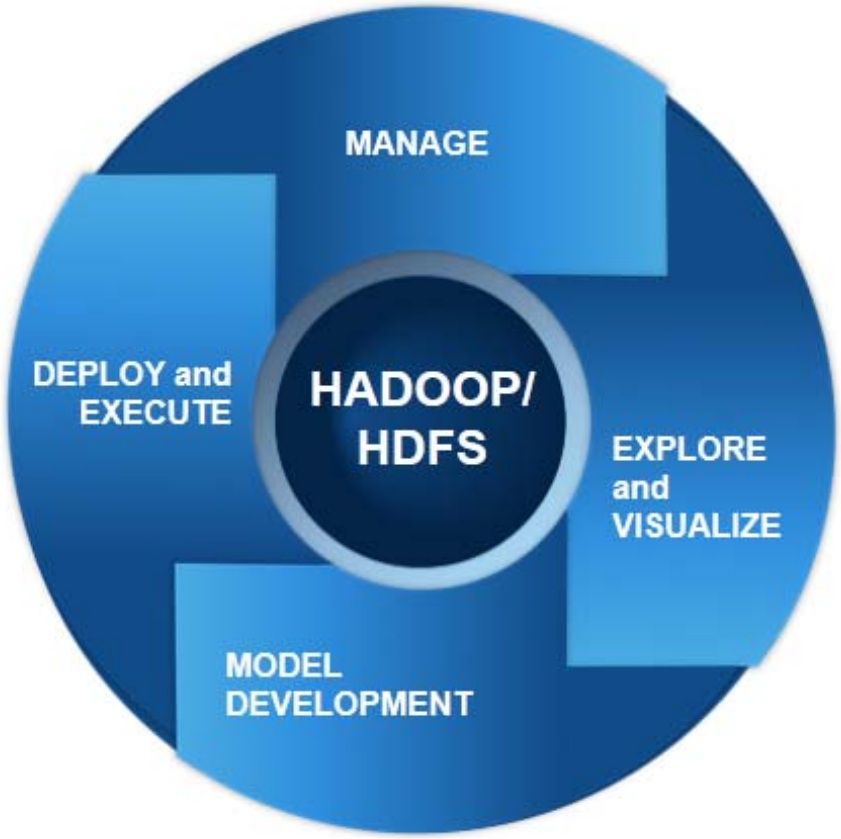
WITH Hadoop



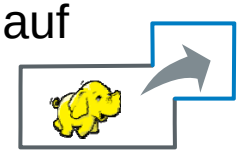
HADOOP ANALYTICS FÜR SPEZIALTHEMEN
INPUT / ANREICHERUNG EINES EDW



VIELFÄLTIGE UNTERSTÜTZUNG, MORE TO COME!

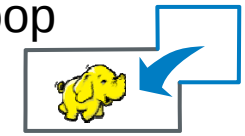


1. **Data Management**: SAS optimiert und erleichtert den Zugriff auf Daten in Hadoop



2. **In-Memory Analytics**: SAS erweitert und beschleunigt Analytik auf Hadoop-Daten.

3. **In-Database Processing**: SAS verlagert (analytische) SAS Funktionalität in das Hadoop Cluster.



WEITERE INFORMATIONEN

RUND UM DAS THEMA SAS UND HADOOP

sas club

- SAS und HADOOP Informationen:
  http://www.sas.com/de_de/software/sas-hadoop.html - http://www.sas.com/en_us/software/sas-hadoop.html
  Code Beispiele:
  http://support.sas.com/resources/papers/proceedings14/SAS033-2014.pdf
- Interessante White papers:
  http://www.sas.com/en_us/whitepapers/big-data-analytics-hadoop-107049.html
  http://www.sas.com/en_us/whitepapers/bringing-power-of-sas-to-hadoop-105776.html
  BARC: Big data analytics in der DACH region:
  http://www.sas.com/de_de/whitepapers/ba-wp-barc-big-data-analytics-2014-2298353.html

- Tom White: Hadoop : The Definitive Guide (O'Reilly)
  http://shop.oreilly.com/product/0636920021773.do
- Edward Capriolo: Programming Hive (O'Reilly)
  http://shop.oreilly.com/product/0636920023555.do

- Next Steps, wie Starten?
  Test SAS&HADOOP - SAS Testlizenzen, Testumgebungen, Erfahrungsaustausch…
  Kontakt : gernot.engel@sas.com, rainer.sternecker@sas.com

IN-MEMORY ANALYTICS

SAS VISUAL ANALYTICS

WITH Hadoop

Central Entry Point | Integration | Role-based Views

DATA BUILDER

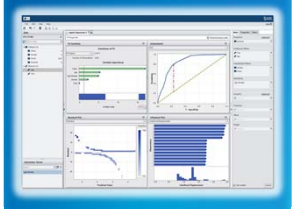
- Join data from multiple sources
- Create calculated and derived columns
- Load data

ADMINISTRATOR

- Monitor SAS® LASR™ Analytic server
- Load/unload data
- Manage security

EXPLORER

- Perform ad-hoc analysis and data discovery
- Apply advanced analytics



VISUAL STATISTICS

- Take VA analytics one step further
- Perform statistical modeling and classification

DESIGNER

- Create dashboard style reports for web or mobile

MOBILE BI

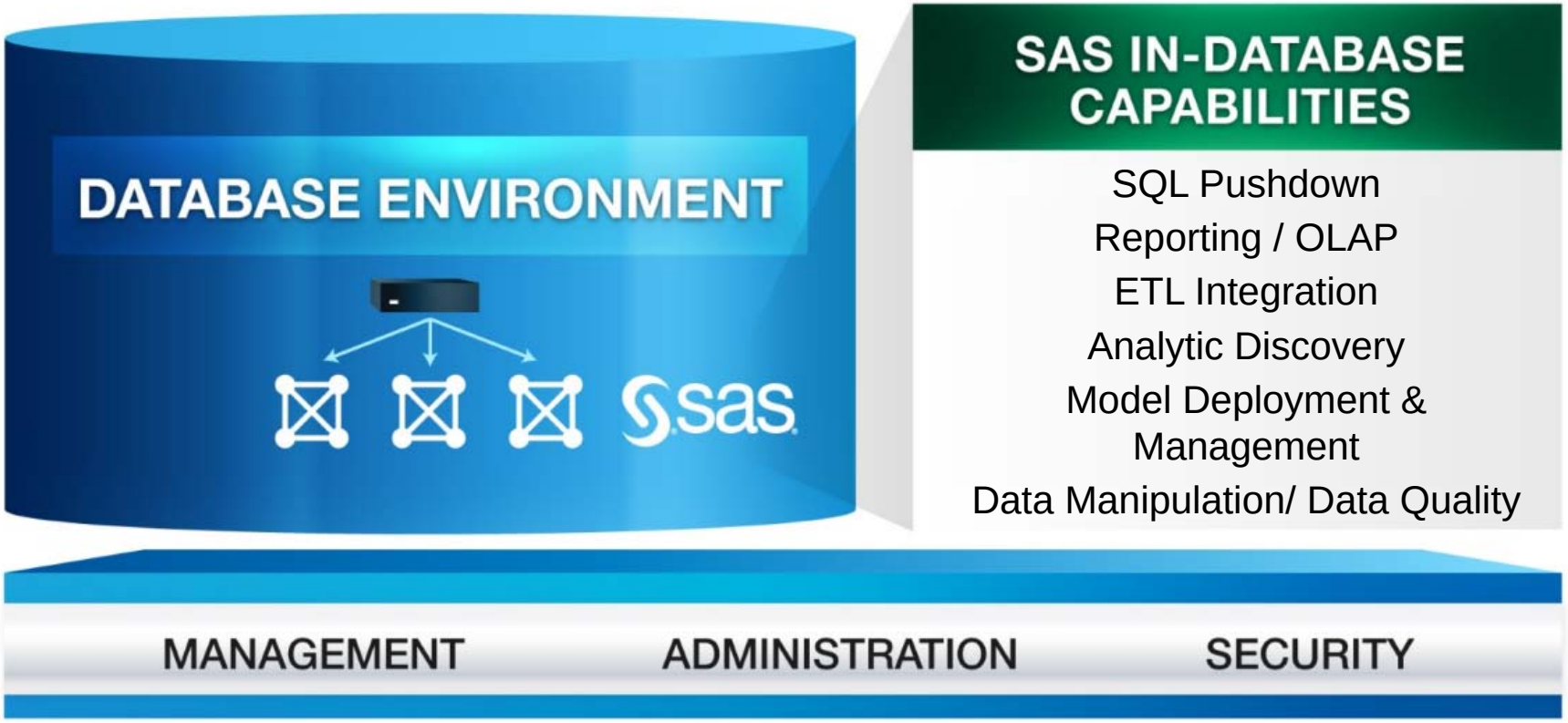
- Native iOS and Android applications that deliver interactive reports

SAS® LASR™ ANALYTIC SERVER

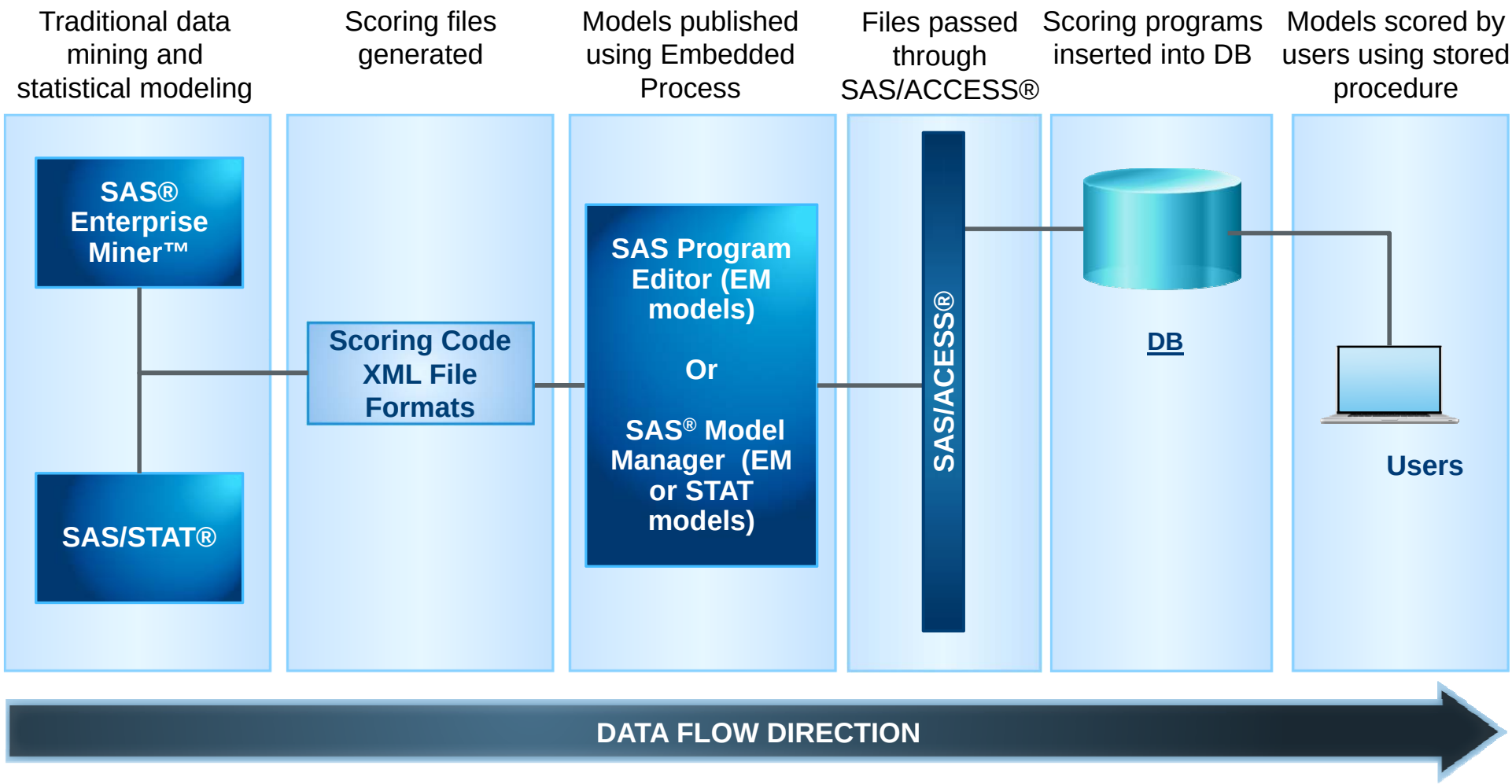
sas club

- Demo Visual Analytics / Visual Statistics

sas club



IN-DATABASE | SCORING ACCELERATOR



IN-MEMORY | ÜBERBLICK

sas club

High-Performance Statistics
- HPLOGISTIC
- HPREG
- HPLMIXED
- HPNLMOD
- HPSPLIT
- HPGENSELECT

High-Performance Data Mining
- HPREDUCE
- HPNEURAL
- HPFOREST
- HP4SCORE
- HPDECIDE

High-Performance Text Mining
- HPTMINE
- HPTMSCORE

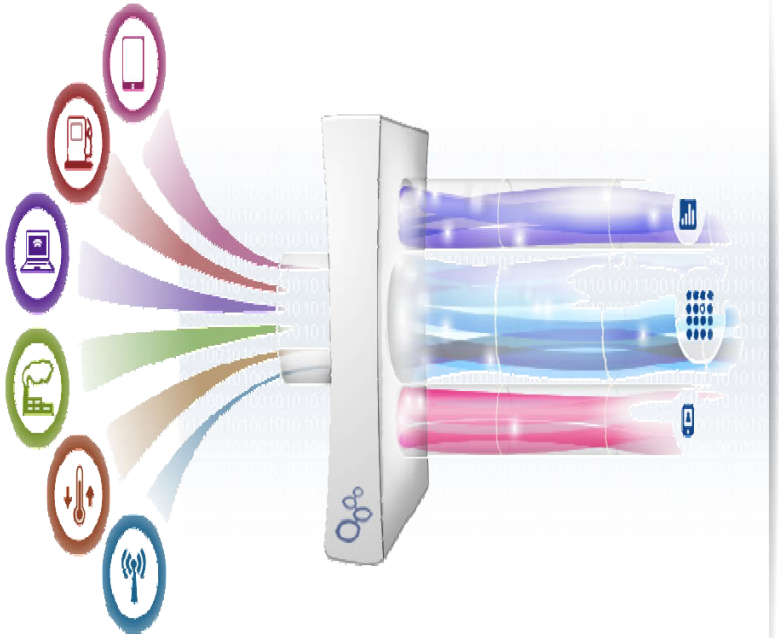
High-Performance Optimization
- OPTLSO
- Select features in
  - OPTMILP
  - OPTLP
  - OPTMODEL

High-Performance Econometrics
- HPCOUNTREG
- HPSEVERITY
- HPQLIM

High-Performance Forecasting
- HPFORECAST

Common procedures (HPDS2, HPDMDB, HPSAMPLE, HPSUMMARY, HPIMPUTE, HPBIN, HPCORR)

IN-ECHTZEIT | SAS EVENT STREAM PROCESSING



SAS® ESP

- ermöglicht und erleichtert die Verarbeitung von Massendaten (Sensoren, Telematik, Netzwerkverkehr, …)
- kann Daten in Bewegung verarbeiten, ohne sie speichern zu müssen
- kann als Frontend für beliebige Systeme eingesetzt werden
- skaliert sehr gut (mehrere Mio. Events/Sek), bei extrem niedrigen Antwortzeiten (Millisekunden)
- arbeitet mit Datenflüssen (gerichteten Graphen) deren Knoten u.a. SQL-Operationen (wie aggregate, join, union, compute, filter, copy) und pattern matching repräsentieren

sas club

Event Sources

Streams

events

events

Applications

events

Databases

events

Telemetry

Event Stream Processing

- Publish / subscribe
- Inserts/updates/deletes
- Continuous queries
  - Aggregate (group by)
  - Correlate (join)
  - Compute
  - Filter
  - Procedural
  - User defined functions
  - Retention windows
  - Pattern matching
- Ad-hoc queries
- Command & control
- Security (Auth, Encrypt, AC)
- Persistence / recovery
- Fail-over
- Distributed services
- Model management

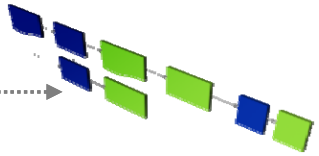
Processed Event Consumers

events

events

Mobile BI / BAM Alerts

events



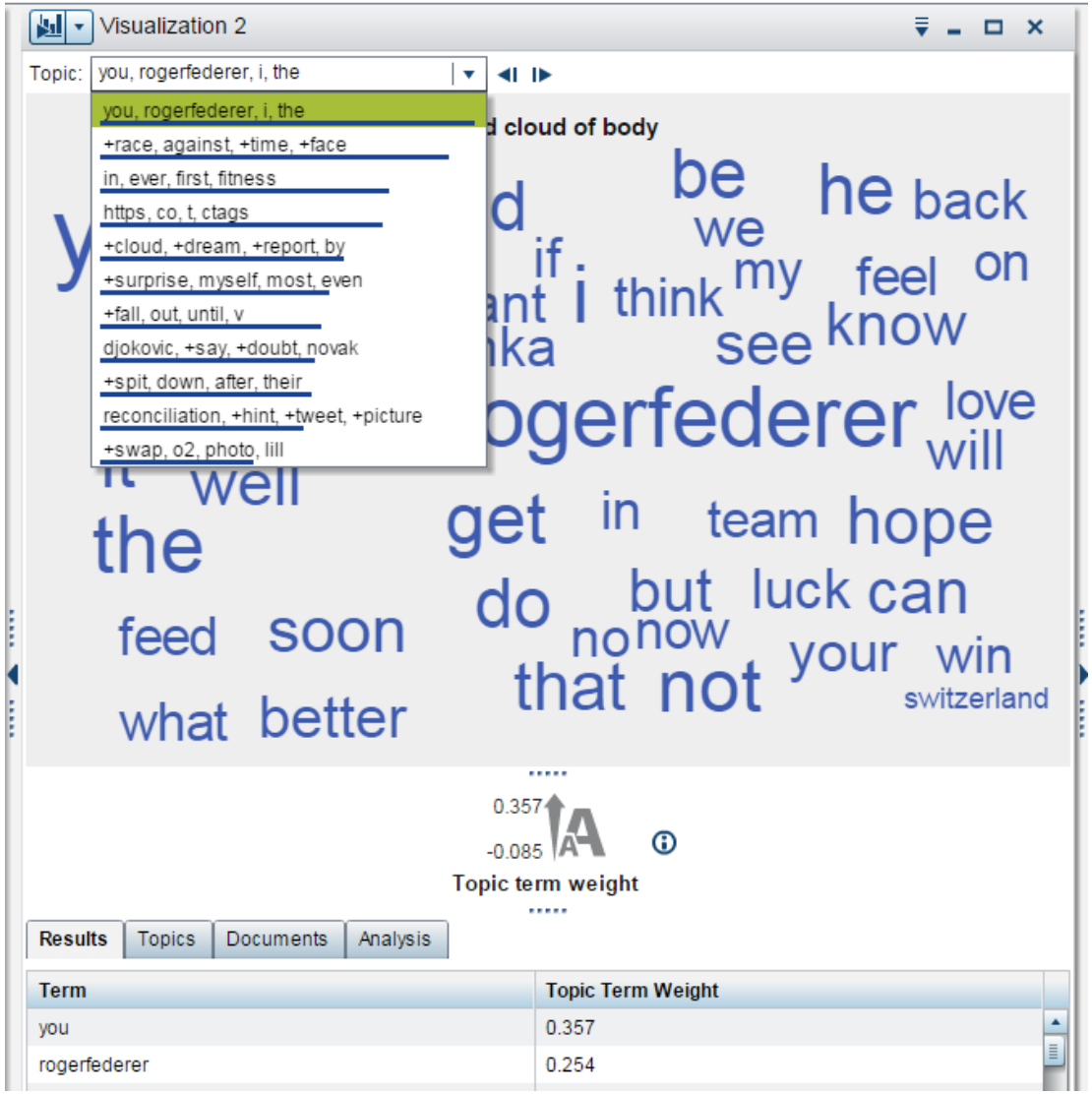
Workflows

AUCH IN SAS VISUAL ANALYTICS

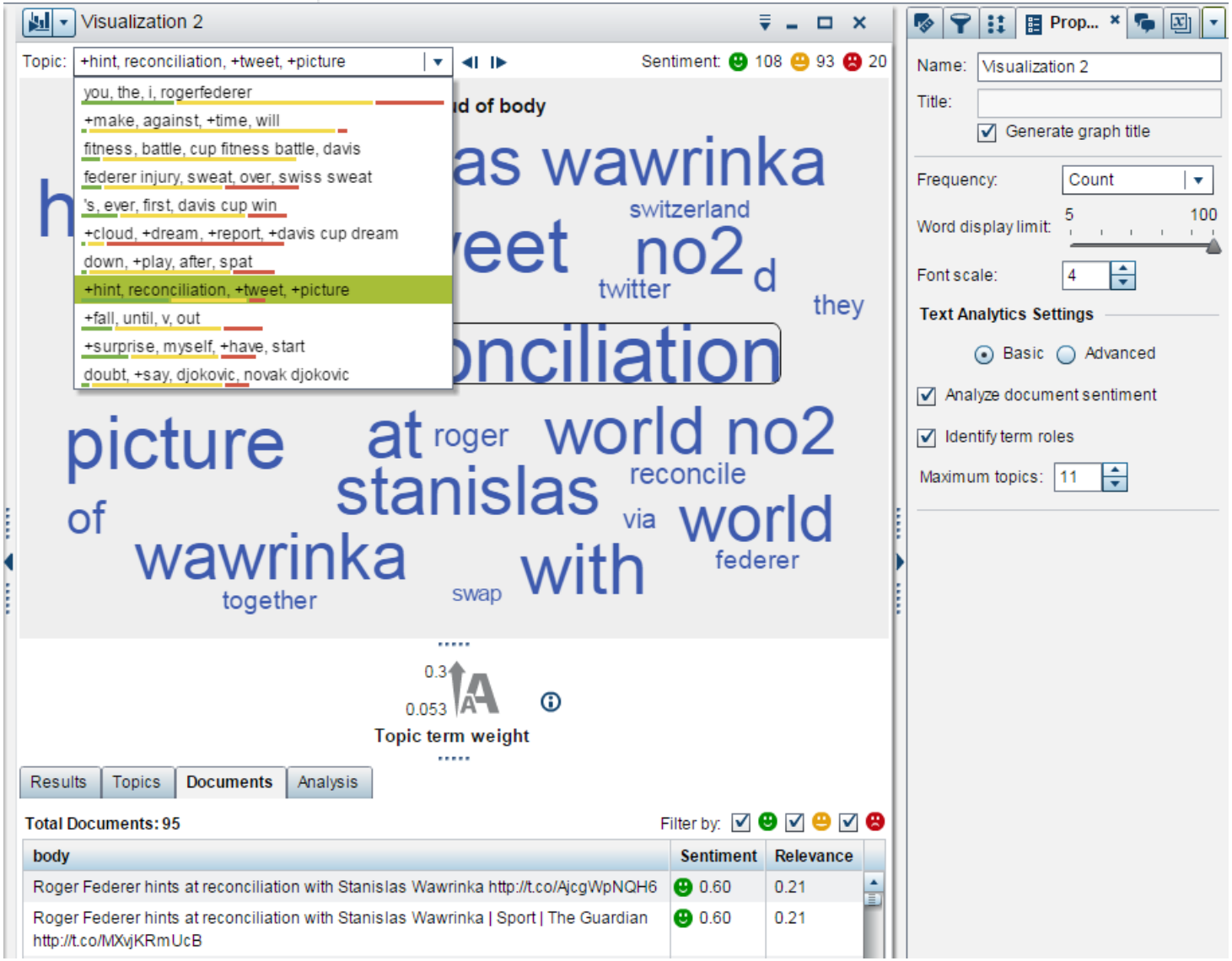
- Demo VA in Echtzeit

SAS VISUAL ANALYTICS

WORD CLOUDS – TOPICS & TERMS



WORD CLOUDS – SENTIMENT



WORD CLOUDS – ROLE RECOGNITION

Results | Topics | Documents | Analysis

Term	Topic Term Weight	Role
twitter	0.083	Proper noun with ambiguous classi...
via	0.081	Preposition
they	0.078	Pronoun
together	0.077	Adverb
federer	0.076	Proper noun with ambiguous classi...
switzerland	0.072	Geographical place
swap	0.065	Verb

SAS VISUAL ANALYTICS

TWITTER API – GEO COORDINATES