



SAS® & HADOOP®

HANS-JOACHIM EDERT, SAS® HIGH PERFORMANCE ANALYTICS



AGENDA

- Hadoop® 101
 - Grundlagen, Distributionen, Architekturen
- SAS® Data Management für Hadoop®
- SAS® In-Database und In-Memory Analytics für Hadoop®
- Zusammenfassung: SAS® & Hadoop®
 - Support Matrix
 - Road Map



HADOOP®

GRUNDLAGEN UND EINORDNUNG



Hadoop® ist „in“

SUCHANFRAGEN WELTWEIT

"big data"

Suchbegriff

Hadoop

Suchbegriff

Teradata

Suchbegriff

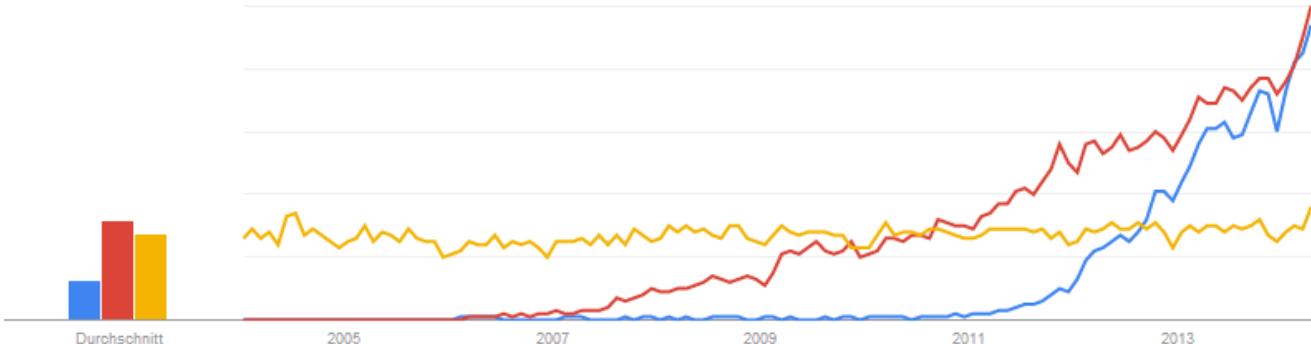
+ Suchbegriff hinzufügen

Teilen ▾

Interesse im zeitlichen Verlauf ?

Nachrichtenschlagzeilen

Prognose ?

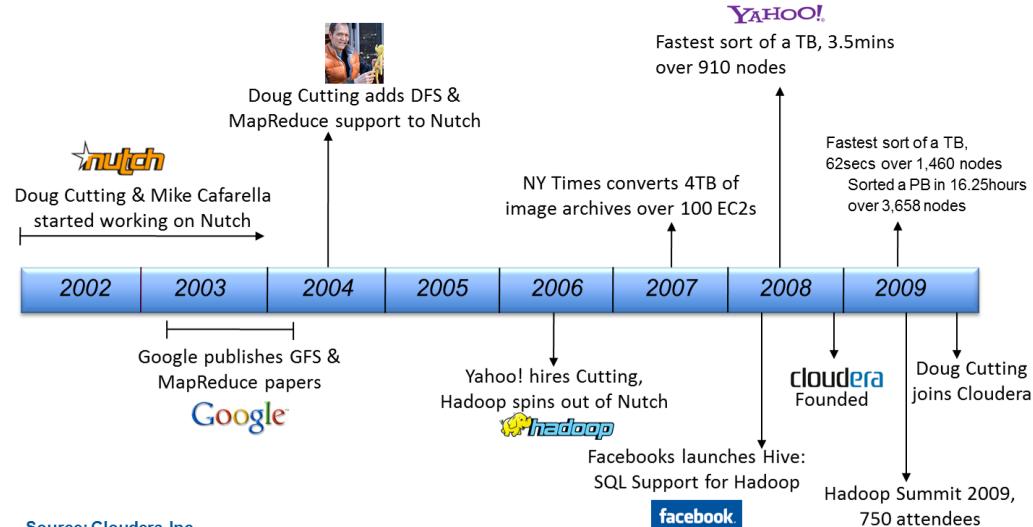


„Teradata“ wurde als Vergleichsbegriff gewählt, da er recht scharf umgrenzt für etablierte Enterprise DWHs steht.

Open-Source Framework zum **verteilten Speichern und parallelen Verarbeiten** von großen, (un-)strukturierten Datenmengen.

- Welche Vorteile ergeben sich durch den Einsatz von Hadoop®?
 - Hadoop® erweitert die Handlungsmöglichkeiten von IT und Analytikern: Dank Hadoop® können nun **mehr** und **andere** Daten effizient analysiert werden.
 - Hadoop® hilft **Kosten zu reduzieren**, da es auf kostengünstiger Hardware aufbaut.
-  Hadoop® hat sich zum **de-facto Standard** von „Big Data“-Anwendungen entwickelt.
- Hadoop® ist **branchenübergreifend** überall dort relevant, wo die anfallenden Datenmengen die etablierten Ansätze in Frage stellen.

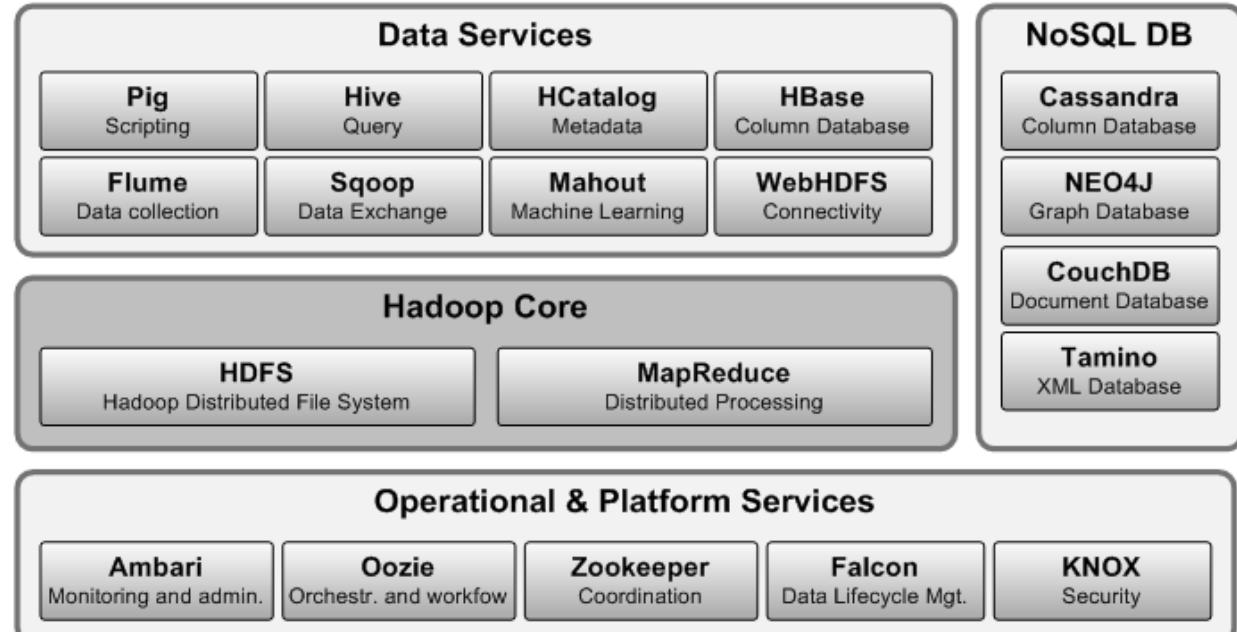
- Basiert auf **Google** Forschungspapieren für ein verteiltes Dateisystem und für ein Datenverarbeitungsverfahren in Clustern (2003/2004).
- Von **Yahoo** für die Entwicklung einer WWW-Suchmaschine aufgegriffen (bis 2008).
- An die **Apache** Foundation übergeben, seitdem open-source (2009).
- Heute vorangetrieben als **Apache** Projekt und von kommerziellen Distributionen (z.B. **Cloudera**, **Hortonworks**).



HADOOP® GRUNDLAGEN

Hadoop® ist eine
Projektplattform
mit HDFS und
Map/Reduce als
Kernmodulen

DAS HADOOP® ÖKOSYSTEM



Fundamentale Hadoop®-Prinzipien:

- **Horizontales Skalieren** → **HDFS**: Hadoop® Distributed File System
- **Datenlokalität (Code to Data)** → **Map/Reduce**: ein Verarbeitungsverfahren nach dem „Teile und herrsche“-Ansatz

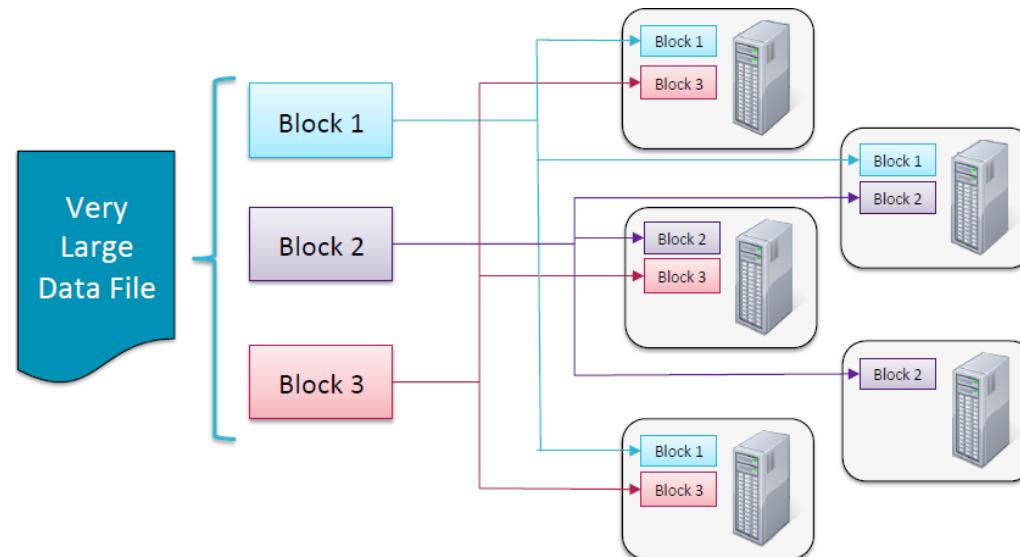
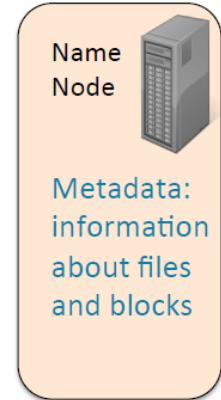
HADOOP® GRUNDLAGEN

Dateien werden in Blöcke gesplittet und auf das Dateisystem verteilt.

Jeder Block wird auf mehrere Nodes repliziert (im Standard 3x).

Der NameNode, ein zusätzlicher Rechner, speichert Metadaten über die Verteilung der Daten.

HADOOP® DISTRIBUTED FILE SYSTEM



(Source: Cloudera Training)

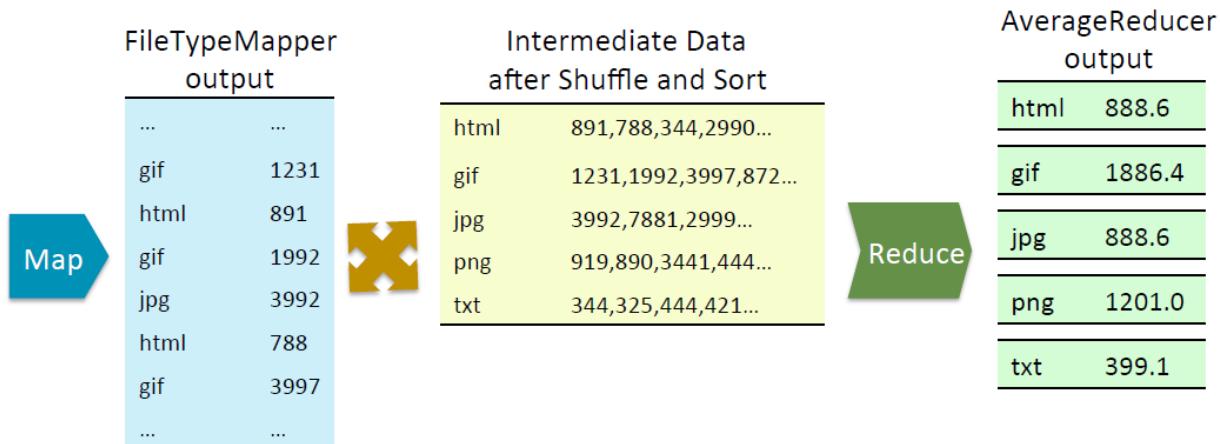
HADOOP® GRUNDLAGEN

Beispiel:
Analysieren von
Webserver Logs
mit Map/Reduce.

Ziel:
Berechnen der
durchschnittlichen
Antwortzeiten nach
Dateityp.

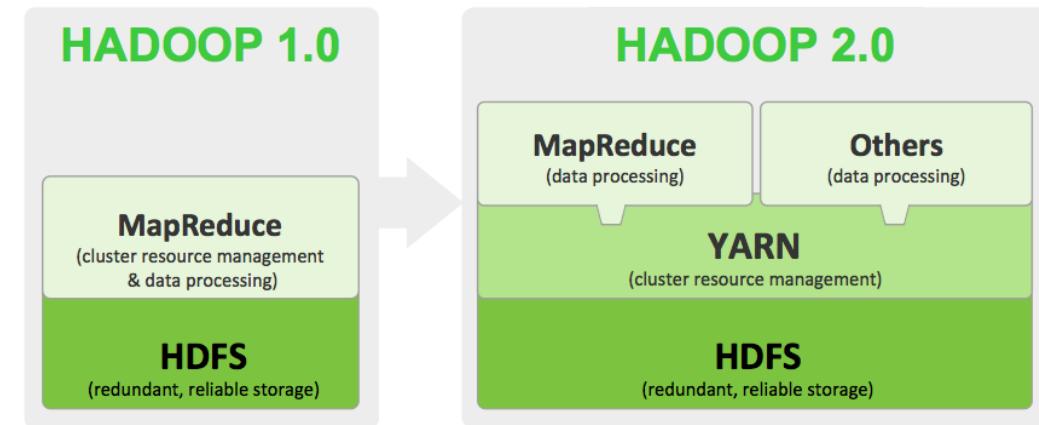
MAP/REDUCE ALGORITHMUS

```
...
2013-03-15 12:39 - 74.125.226.230 /common/logo.gif 1231ms - 2326
2013-03-15 12:39 - 157.166.255.18 /catalog/cat1.html 891ms - 1211
2013-03-15 12:40 - 65.50.196.141 /common/logo.gif 1992ms - 1198
2013-03-15 12:41 - 64.69.4.150 /common/promoex.jpg 3992ms - 2326
...
```



(Source: Cloudera Training)

- YARN (=Yet Another Resource Negotiator) ist der Nachfolger des Map/Reduce Frameworks.
 - Umstellung auf YARN ist eine der wichtigsten Neuerungen der aktuellen Hadoop® Releases.
- YARN verwaltet die Cluster-Resourcen und weist Hadoop® Jobs dynamisch Verarbeitungskapazitäten zu.
 - Bisher war Map/Reduce das einzige unterstützte Verarbeitungsverfahren.
 - Unter YARN ist Map/Reduce nur eines unter vielen Verfahren.
 - SAS® als YARN Service!



HIVE



- Deklarative, SQL-ähnliche Abfragesprache
~ Hive QL (Query Language).
- METASTORE: Metadaten für Dateien im HDFS, in externem RDBMS gespeichert.
- Hive Engine kompiliert Hive QL in Map/Reduce Code.

IMPALA



- Eine verteilte arbeitende SQL Query Engine, die nicht auf Map/Reduce beruht.
- Verwendet den Metastore, den ODBC Treiber und die SQL Syntax von Hive.
- Nutzt die Security, Datenformate und Resourcenverwaltung des Hadoop® Clusters.

HBASE (NON-RELATIONAL)



- „A column-oriented database management system that runs on top of HDFS“.
- Arbeitet nach dem Key-Value Prinzip, Vorbild war Google's BigTable.
- Gestattet zufällige Lese-/Schreibzugriffe in Echtzeit.
- Kann „unter“ Hive betrieben werden.

✓ Was ist Hadoop®?

- Historie: Anhaltende Erfolgsstory seit Veröffentlichung, „Quasi-Standard“ für Big Data.
- Rundprinzipien: **horizontale Skalierung (scale out)** und **Datenlokalität**.
- Kein Produkt, sondern „**Ökosystem**“: Open Source Plattform zur verteilten Datenhaltung und -verarbeitung.
- Wie kann ich es einsetzen? Wieso sollte ich es einsetzen?
- Welche Rolle kann SAS® dabei spielen?



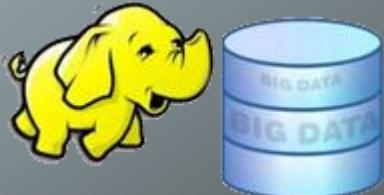
HADOOP GRUNDLAGEN

KEINE GEGENSÄTZE
... ABER OFT NICHT GEMEINSAM BETRACHTET!

Hadoop als “Data Platform”

SAS® Data Management

- SAS®/Access to Hadoop/Impala
- SAS® Data Loader



SAS® In-Database Analytics

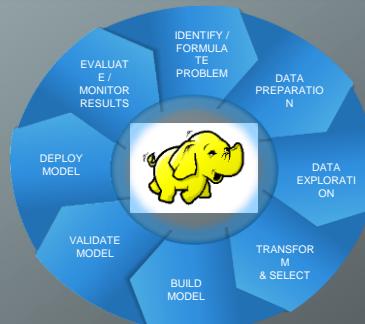
- SAS® Scoring Accelerator
- SAS® Code Accelerator
- SAS® DQ Accelerator

...ist Baustein einer Transformation der IT Landschaft

Hadoop als Kernkomponente einer “next gen” BI und Analytics Strategie

SAS® In-Memory Analytics

- SAS® Visual Analytics
- SAS® Visual Statistics
- SAS® In-Memory Statistics



...dient zur Unterstützung neuer Fragestellungen in den Fachbereichen



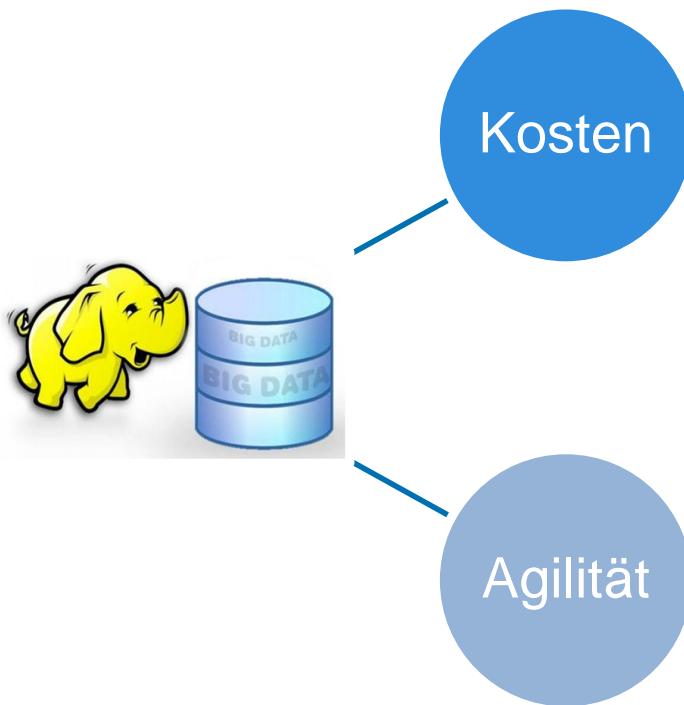
HADOOP® DATA PLATFORM

ALS NEUER MITSPIELER IN BESTEHENDEN IT LANDSCHAFTEN



- Der Einzug von Hadoop® in die Rechenzentren leitet eine IT Transformation ein:
 - Erweiterung der bestehenden Landschaft mit Auswirkungen auf:
 - Architektur: Datenströme werden umgeleitet.
 - Prozesse: neue Werkzeuge, neue Verarbeitungsketten.
- Die zentralen Treiber einer solchen Umstellung:
 - Erwartete **Kosteneinsparungen**.
 - Erwartete höhere **Agilität** im Umgang mit (neuen) Datenquellen.

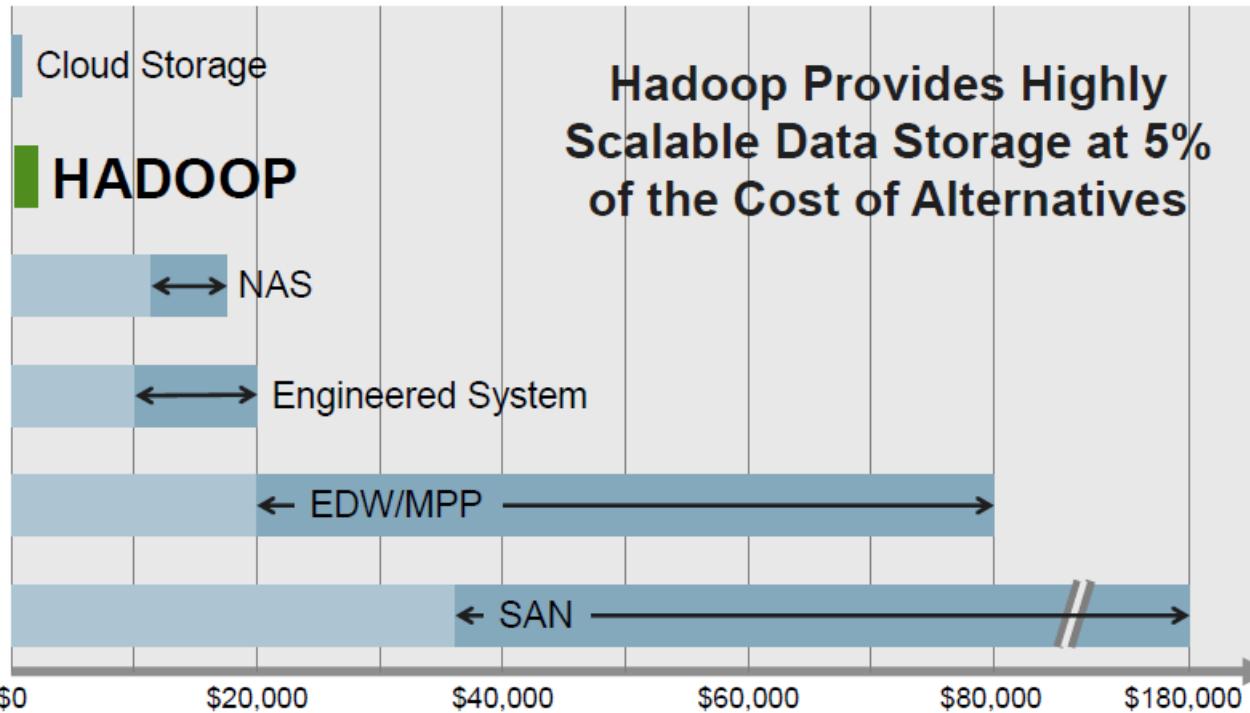
WARUM SETZEN UNTERNEHMEN AUF HADOOP®?



- **Günstiger** Datenspeicher (günstiger Einstieg, skaliert linear).
- **Einfache Beladung** ("schema on read" = keine Verzögerungen durch Datenmodellierung).
- **Ablage** für historische Daten und Verlagerung des ETL aus dem "teuren" EDW .

- Zugang zu **neuen Datenquellen**, ermöglichen innovative Usecases.
- Fachanwender erhalten **Zugang zu allen Daten** (zu jedem Zeitpunkt).
- **High Performance** durch parallele Verarbeitung
- Kann neben einem bestehenden EDW **koexistieren** (falls gewünscht).

Fully Loaded Cost per Raw TB of Data (min – max cost)



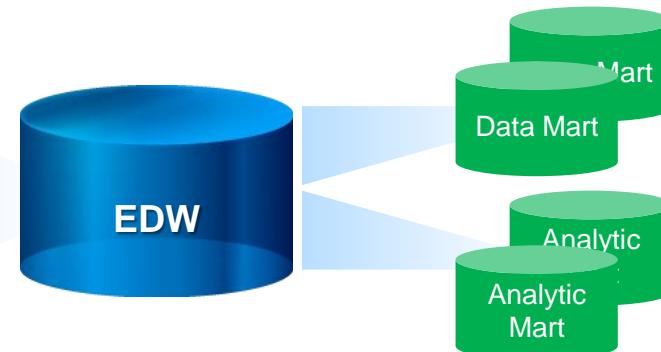
(Source: Hortonworks Training)

HADOOP® DATA PLATFORM

Operational Data Sources



DER AUSGANGSPUNKT HEUTE



Unstrukturierte und semi-strukturierte Daten, Datenströme (z.B. Sensordaten) laufen oft am DWH vorbei oder werden nicht gesammelt.

BI und
Analytics

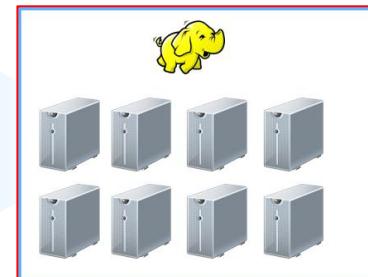
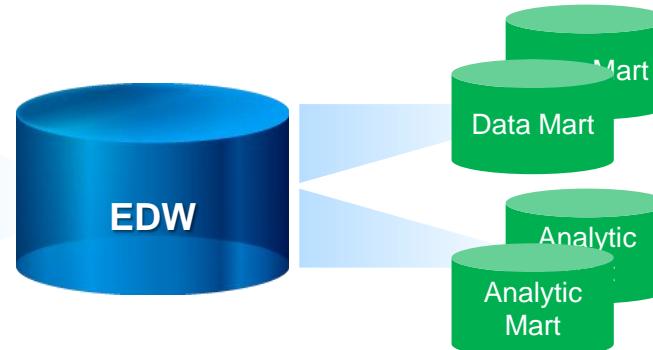


HADOOP® DATA PLATFORM

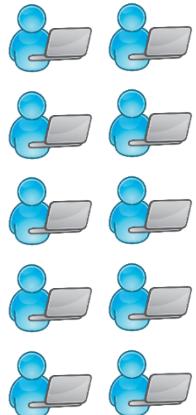
Operational Data Sources



HADOOP® ALS WEITERE DATENSENKE



BI und
Analytics

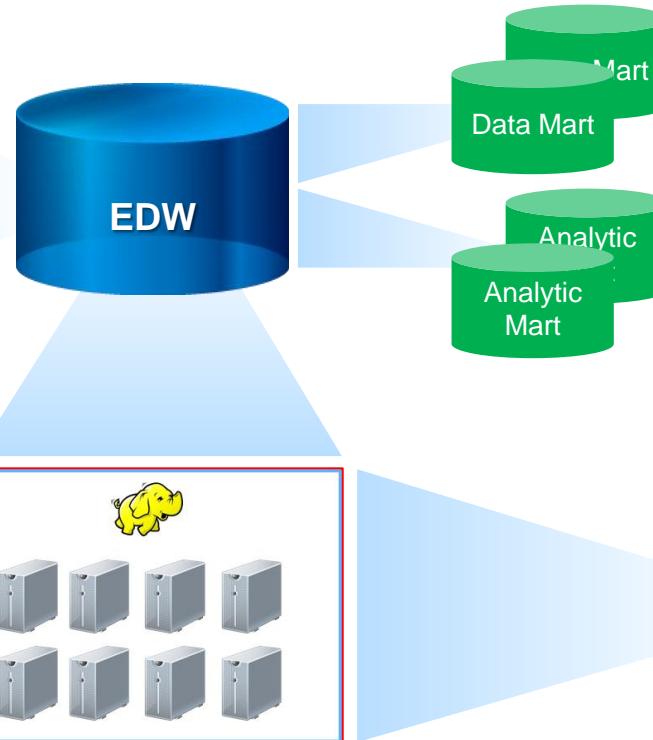


HADOOP® DATA PLATFORM

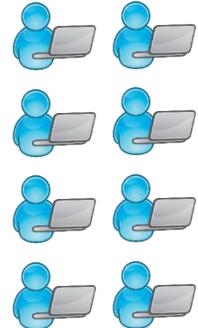
Operational Data Sources



HADOOP® ALS DATENQUELLE FÜR DAS EDW



BI und
Analytics



Hadoop® ist (zumeist) kein Ersatz für ein RDBMS.

- Designed für große Datenmengen im read-only Zugriff.
 - Keine Random Reads/Writes, keine In-Place Modifikation von Daten möglich: nur SQL Append, kein Update, kein Delete.
- Nicht designed für strukturierte, relational organisierte Daten.
 - Verwendet “Schema on Read” (vs. “Schema on Write” bei RDBMS).
- Designed für Batchverarbeitung.
 - Hohe Antwortzeiten machen Hadoop® ungeeignet für interaktives Arbeiten.
 - Keine Indexierung für Daten im HDFS.
- Hadoop® ist kein Produkt, sondern ein Ökosystem.
 - Viele Projekte in unterschiedlichen Reifegraden und Entwicklungsgeschwindigkeiten.
 - Unterschiede in den Distributionen, Inkompatibilitäten (Vendor Lock-in).
 - Sind die benötigten Enterprise Features alle bereits vorhanden (z.B. Security)?



TIERED STORAGE ARCHITEKTUR

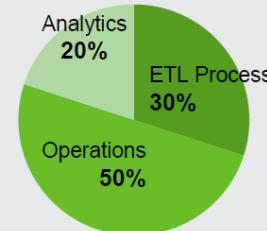
Hadoop® als Erweiterung und Ergänzung

für das Enterprise DWH einsetzen:

- Einfach skalierbarer, kostengünstiger Massenspeicher.
- Räumt Verarbeitungs-Kapazitäten im DWH frei durch Auslagern von ETL Prozessen.
- Auch „archivierte“ Daten stehen weiterhin für Analysen zur Verfügung.

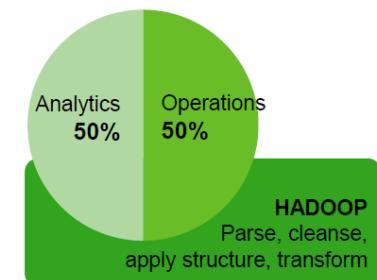
CHALLENGE

- The EDW is at capacity
- Older transformed data archived, not available for exploration
- Strict Schema



SOLUTION

- Free up EDW resources
- Keep 100% of source data
- Mine data for value after loading it because of schema-on-read
- Reduce incremental EDW spend



HADOOP® DATA PLATFORM

BASE SAS®

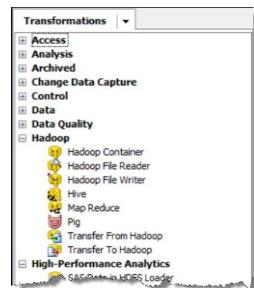
- Map Reduce + Pig Scripting + HDFS Kommandos.

SAS®/Access to Hadoop®

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

SAS®/Access to Impala (Cloudera only)

SAS® Data Integration Studio (Transformationen) im Paket Data Management Standard / Advanced:



- 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

SAS® Federation Server

- Datenvirtualisierung und Zugriffsschutz für Hadoop® und andere Datenquellen.

SAS® Event Stream Processing Engine

- Hadoop® Adapter für SAS® ESP, um Data Streams in HDFS zu speichern.

SAS® PRODUKTE FÜR DATA MANAGEMENT IN HADOOP®

```
/* 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;

/* RUN PIG SCRIPT */
FILENAME CFG "C:\SAMPLE_DATA\HADOOP_CONFIG.XML";
FILENAME PIGCODE1 "C:\SAMPLE_DATA\PIG_CD.TXT";
PROC HADOOP OPTIONS=CFG;
  PIG CODE=PIGCODE1;
RUN;

/* SUBMIT HDFS COMMAND, RUN MR JOB */
PROC HADOOP OPTIONS=CFG;
  HDFS DELETE="/USER/HADOOP/OUTPUT_MR1";
  MAPREDUCE INPUT="/" OUTPUT="/"
    JAR=".." MAP=".." REDUCE="..";
RUN;

/* CREATE METADATA ON HDFS FILE */
PROC HDMD NAME=HDLIB.CUSTOMERS2014
  DATA_FILE="CUSTOMERS_2014.JSON"
  FILE_FORMAT=JSON;
  COLUMN KUNDEN_ID      CHAR(9) TAG="KNR";
  COLUMN DATUM          DATE     TAG="DT";
RUN;
```

HADOOP® DATA PLATFORM

SAS® PRODUKTE FÜR DATA MANAGEMENT IN HADOOP®

SAS® Data Quality Accelerator for Hadoop (EA in 9.4M2):

- Ausführen von Data Quality Routinen (Parse, Standardize, Gender analysis, Match Code, Identification Analysis, Casing).

SAS® Code Accelerator for Hadoop®:

- Ausführen von DataStep2 Code in Hadoop®.

SAS® Data Loader for Hadoop® (EA in 9.4M2):

- Web-basiertes In-Database Tool für Endanwender (z.B. aus den Fachbereichen).



- Point & Click Datenmanagement Routinen, die in Hadoop® ausgeführt werden.
- HTML5 basiertes Interface
- Wird als vApp ausgeliefert (läuft in VMWare Player).

Data Loader 2.1 Capabilities

- Query a Table in Hadoop
- Run a SAS® Program
- Transform Data in Hadoop®
- Load Data to LASR
- Profile Data in a table

Data Loader Roadmap

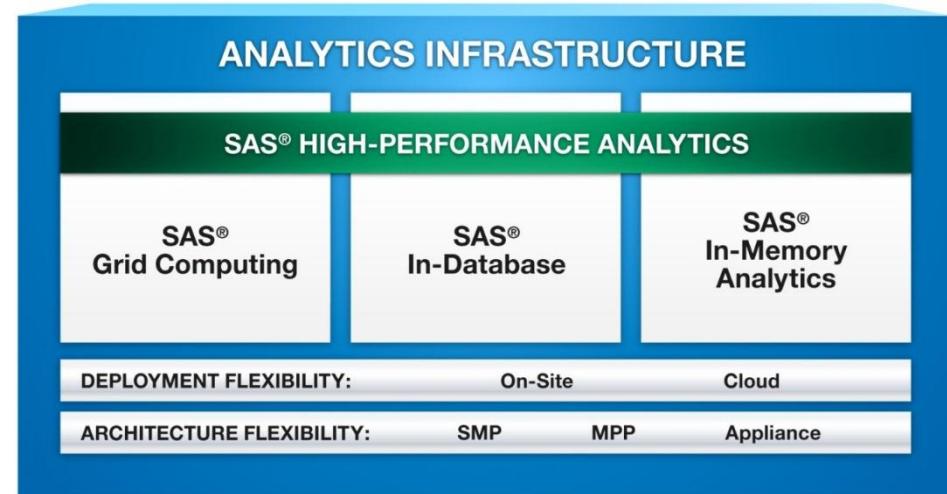
- Joins in Hadoop®
- Sqoop
- Data Quality Routines
- Metadaten / Lineage
- Security
- Monitoring

HADOOP® FÜR REPORTING UND ANALYTICS

IN DER SAS® WELT



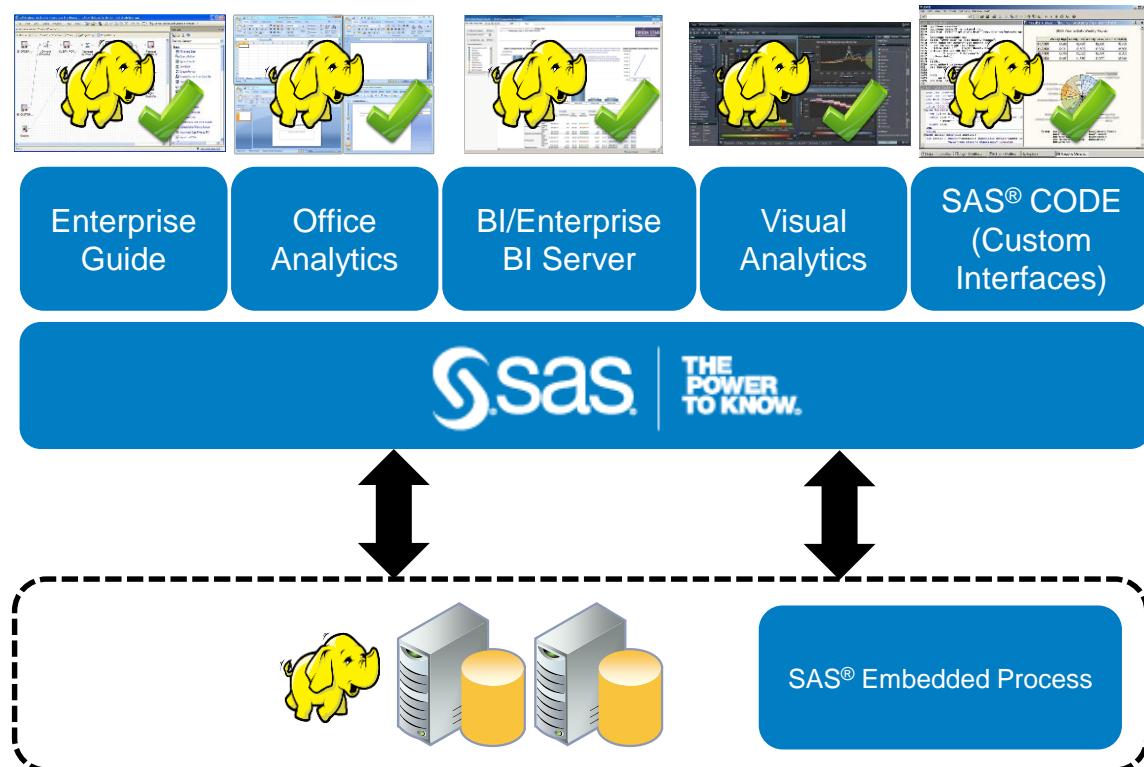
- Batch-Architektur: Hadoop® eignet sich nur bedingt für interaktives Reporting und Analytik.
 - Impala, HBase ...
- Konsequenz: Hadoop® Daten müssen verdichtet und entladen werden.
 - Der Anspruch, dem Anwender alle Daten jederzeit zur Verfügung zu stellen, wird nicht erfüllt.
- SAS® In-Database und In-Memory Technologien erlauben neue Auswertungsansätze, die aus den Hadoop® -Daten erst einen unternehmerischen Wert generieren.



HADOOP® FÜR BI UND ANALYTICS

REPORTING UND VISUALISIERUNG AUF HADOOP® MIT SAS®

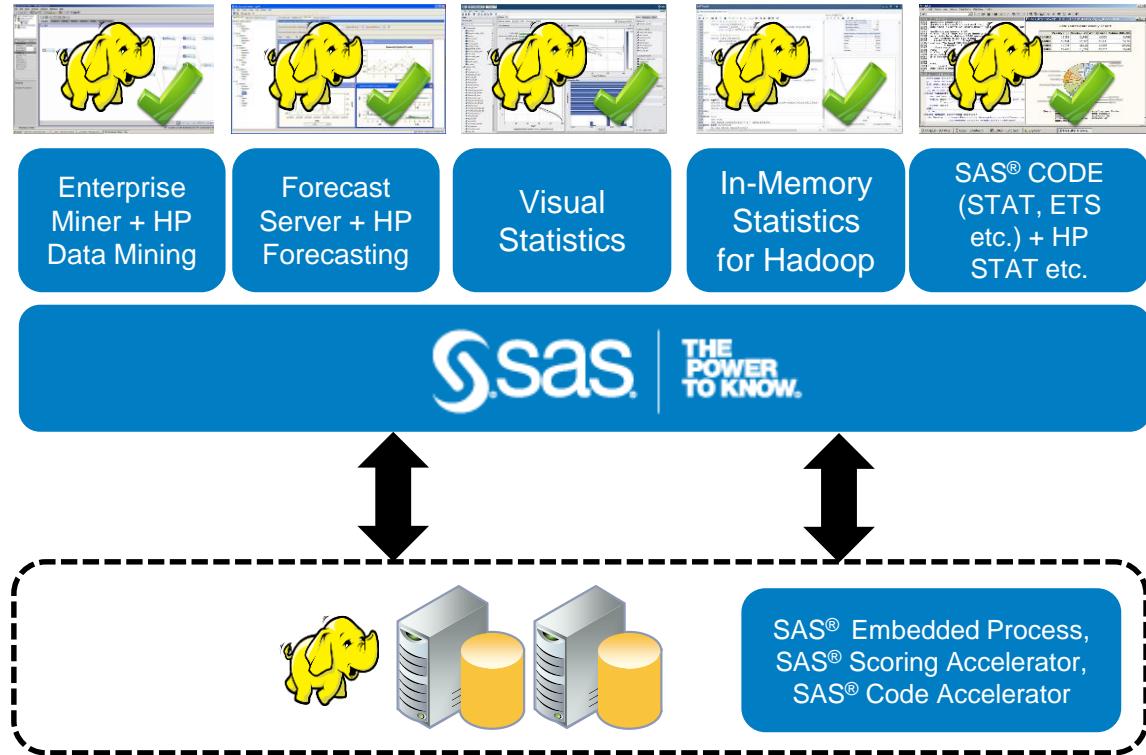
- SAS®/Access to Hadoop®/Impala erlaubt auch „klassischen“ SAS® Clients den Zugriff auf Hadoop® - Daten:
 - Zugriff über Hive / SQL.
 - Latenzen!
- „Sonderfall“ Visual Analytics:
 - performance-optimierte Hadoop® Schnittstellen (HDAT, EP).



HADOOP® FÜR BI UND ANALYTICS

ANALYTICS AUF HADOOP® MIT SAS®

- SAS®/Access to Hadoop®/Impala erlaubt auch „klassischen“ SAS® Clients den Zugriff auf Hadoop®-Daten:
 - Zugriff über Hive / SQL.
 - Data Extraktion / Sampling.
- In-Memory Analytics:
 - performance-optimierte Hadoop® Schnittstellen (HDAT, EP).
- In-Database Analytics:
 - SAS® Modelle und Codes werden in Hadoop® ausgeführt.



SAS® Accelerators verlagern die Datenverarbeitung in die Datenbank:

- Nutzen die MPP Architektur der Datenbank aus.
- Vermeiden das Bewegen von Daten.
- Optimale Ausnutzung der vorhandenen Hardware: "Score ALL your data".
- Verwenden die SAS® Embedded Process (EP) Technologie.

SAS® Embedded Process for Hadoop® : Funktionsbaustein, der auf allen Nodes des Hadoop® Clusters installiert und ausgeführt wird:

- SAS® Code wird über Map/Reduce übertragen.
- Liest und transformiert lokale Daten
- Integration mit MapReduce1 und Yarn.

SAS® SCORING ACCELERATOR

SCORING VERFAHREN AUF HADOOP® DATEN

Create a model



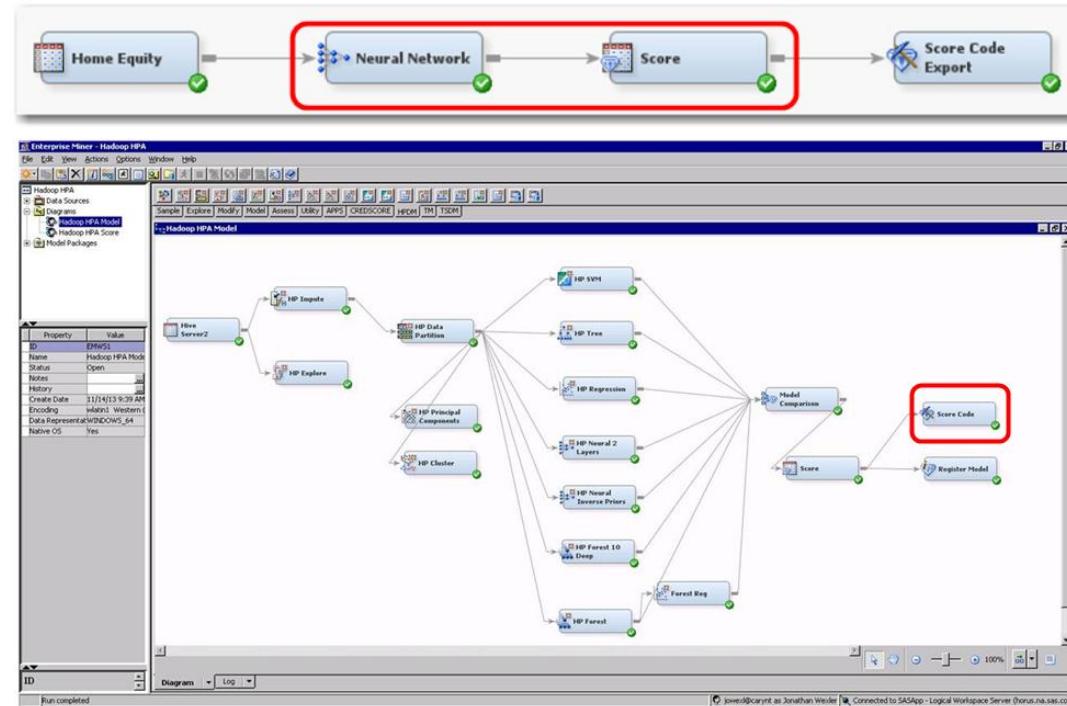
Export the model



Publish the model
to Hadoop® *



Run the model in
Hadoop® *



* über SAS® Makros

SAS® In-Memory Analytics: Produktfamilie auf einer gemeinsamen technischen Infrastruktur für unterschiedliche Zielgruppen und Anwendungsbereiche:

- **SAS® Visual Analytics:**
„Approachable Analytics“, Visualisierung und Analytik für Fachanwender.
- **SAS® Visual Statistics:**
Add-on für Visual Analytics mit Fokus auf weiterführende Analysen.
- **SAS® In-Memory Statistics for Hadoop®:**
SAS®/BASE Schnittstelle für Analytik und Datenmanagement auf In-Memory Daten.
- **SAS® High-Performance Analytics Prozeduren:**
Übertragung der analytischen SAS® Prozeduren auf die In-Memory Architektur, Fokus auf Operationalisierung (Modellgenerierung).

- SAS® In-Memory Produkte verwenden eine verteilte In-Memory Architektur:
 - Jedes Cluster besitzt 1 Head Node und n Worker Nodes.
 - Scale-Out Ansatz: Cluster skaliert über zusätzliche Hardware.
 - Daten werden vor der Verarbeitung in den Arbeitsspeicher der Nodes geladen.
- Bei großen Datenmengen ist ein performantes Verfahren zur Datenbeladung entscheidend:
 - Parallel Beladung der Cluster Nodes mit Daten wäre optimal.
 - Setzt eine Datenquelle voraus, die ihre Daten ebenfalls verteilt speichert.

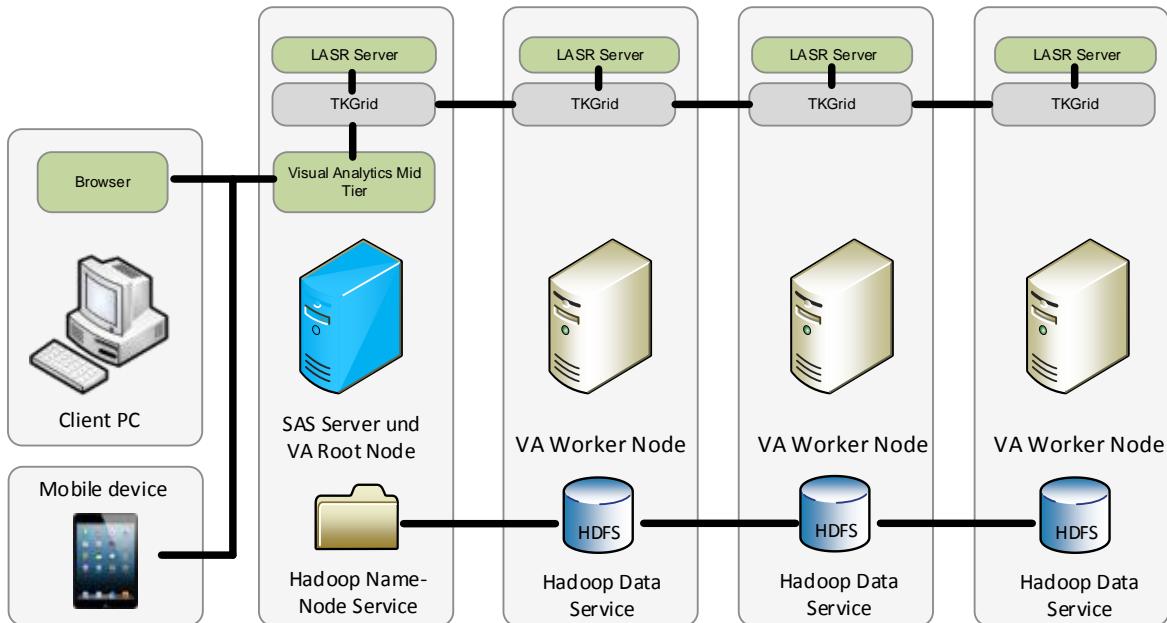


Hadoop® ist die ideale „Partner“-Technologie.

IN-MEMORY ANALYTICS

- Typische Topologie für Visual Analytics:
 - a.k.a. „co-located“: Hadoop® und LASR nutzen gemeinsam die verfügbaren Nodes.
 - Nur für Hadoop® : Memory Mapping von HDFS Daten durch SAS® eigenes Datenformat (SASHDAT).
- **Vorteil:** Performance-Gewinn bei großen Datenmengen!
- **Nachteil:** Proprietäres Datenformat.

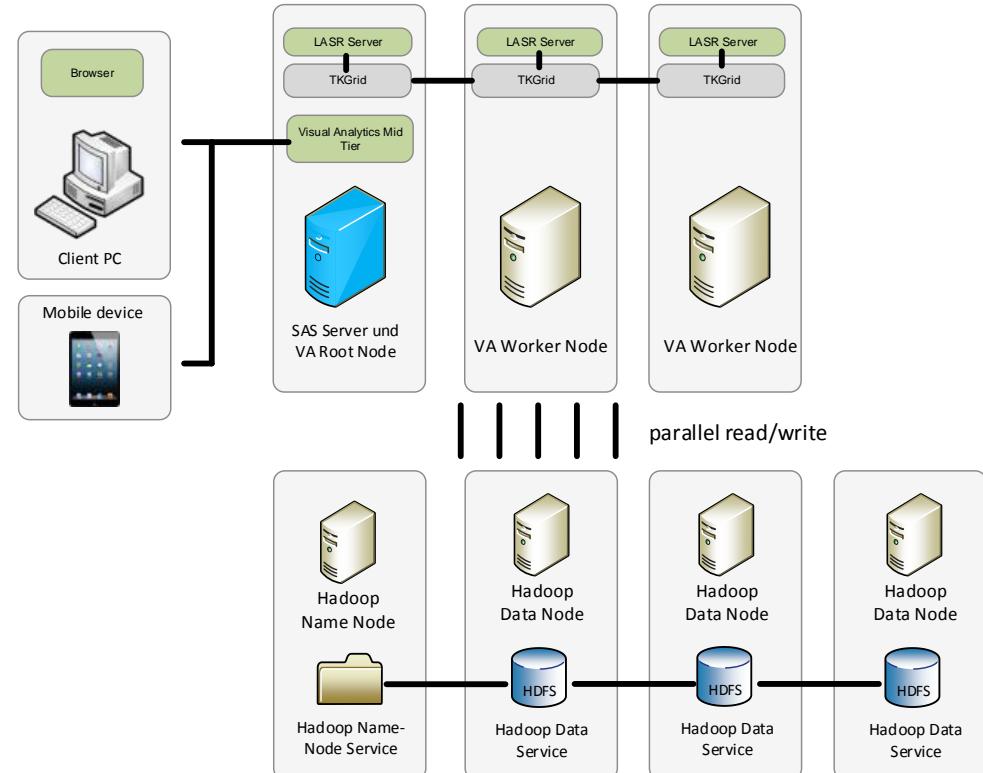
TYPISCHE TOPOLOGIE AM BEISPIEL VISUAL ANALYTICS



IN-MEMORY ANALYTICS

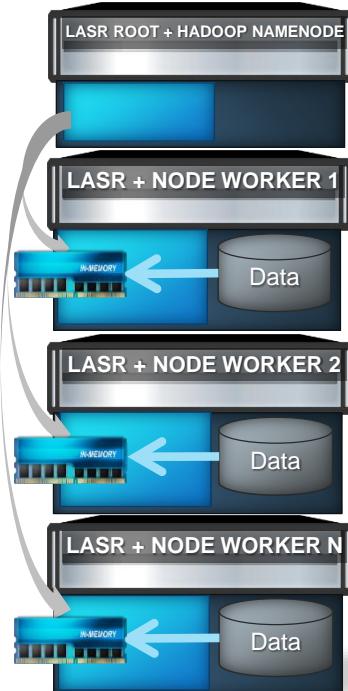
ERWEITERTE TOPOLOGIE AUF VORHANDENEM HADOOP®

- Dediziertes SAS® Compute Server Modell (logisch und physisch getrennt von Hadoop® Datenhaltung):
 - Erlaubt einen „dual use“ für ein bereits vorhandenes Hadoop® Cluster.
 - Embedded Process als „parallel data feeder“ im Hadoop® extrahiert die Daten parallel in den Arbeitsspeicher aller SAS® Nodes.
- Vorteil:** Beide Layer können voneinander unabhängig skalieren.
- Vorteil:** Keine Resourcenkonflikte zwischen SAS® und Hadoop®.
- Nachteil:** Höhere Anschaffungs- und Betriebskosten (mehr Hardware).



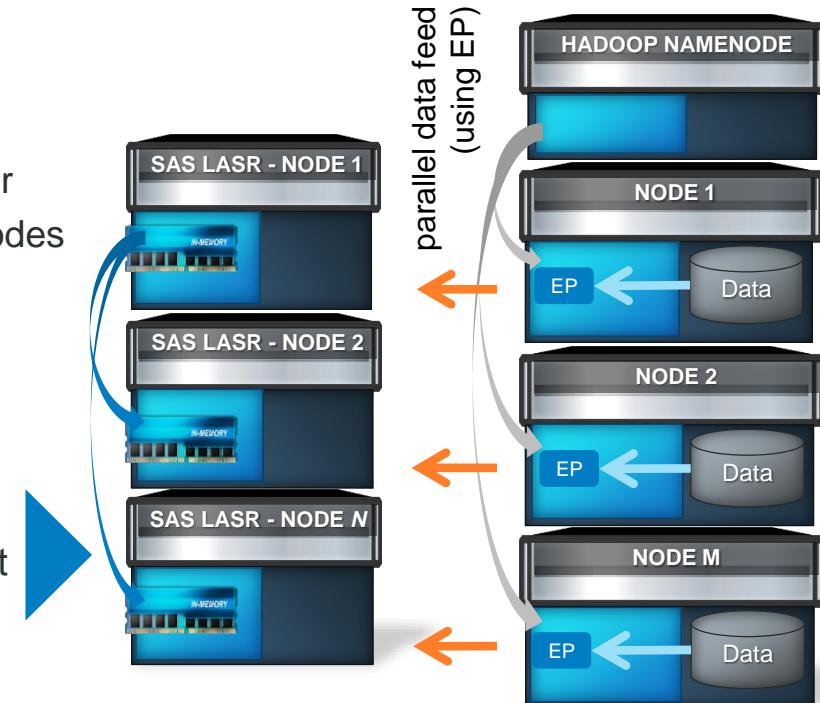
IN-MEMORY ANALYTICS

VISUAL ANALYTICS TOPOLOGIEN



Co-located (LASR Server läuft auf den gleichen Nodes wie auch Hadoop®).

Asymmetric oder „remote LASR“ (LASR Cluster läuft auf separaten Nodes).



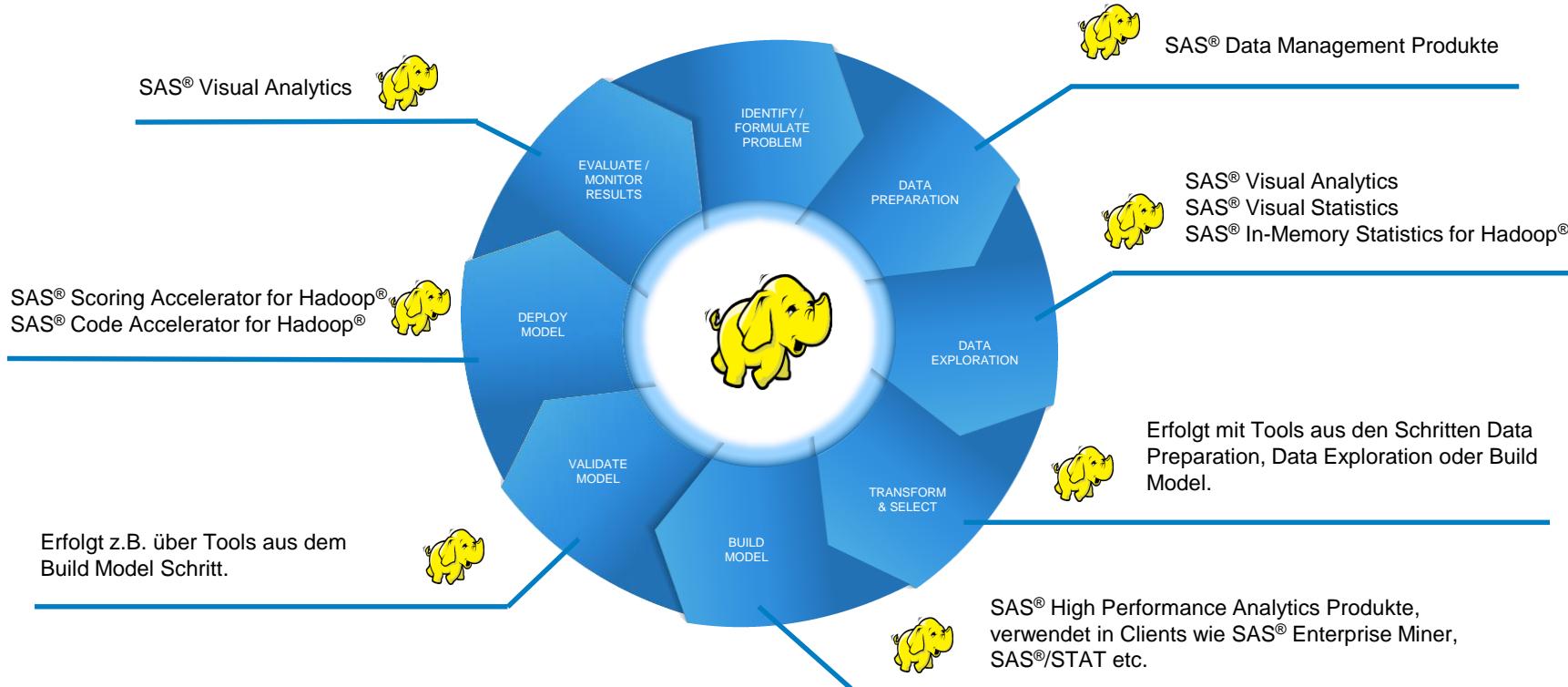


SAS® & HADOOP®

ZUSAMMENFASSUNG, SUPPORT MATRIX UND ROAD MAP



UNSER ZIEL: DEN ANALYTICAL LIFECYCLE FÜR HADOOP® MÖGLICH MACHEN



SAS® & HADOOP®

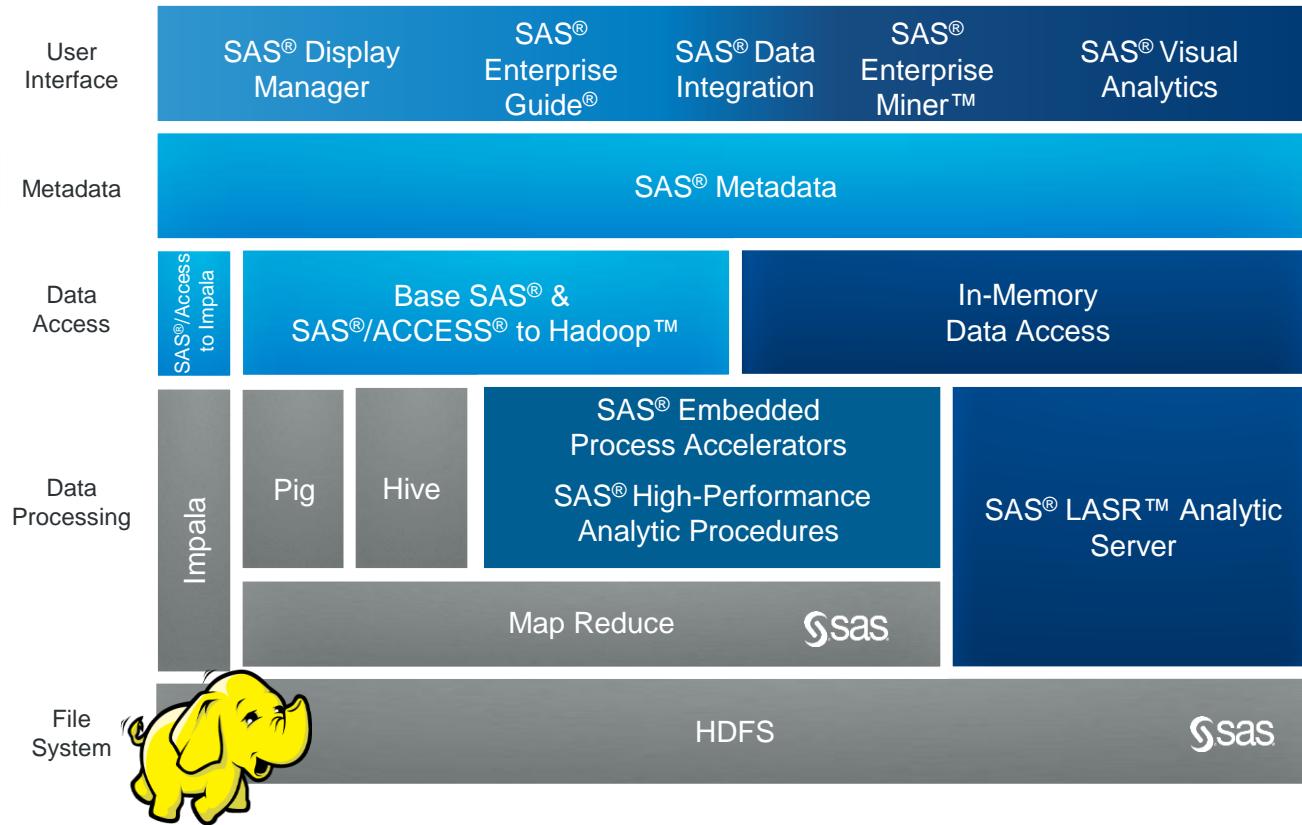
SAS® IM HADOOP® ÖKOSYSTEM



SAS® User



**Next-Generation
SAS® User**



SAS® & HADOOP® WOHIN GEHT DIE REISE?

- SAS® Produkt-Releases für Hadoop® 2014:
 - SAS®/Access Interface to Hadoop® / Impala (Update).
 - SAS® Scoring Accelerator for Hadoop® (Update).
 - SAS® Data Quality Accelerator for Hadoop® (Neu).
 - SAS® Code Accelerator for Hadoop® (Neu).
 - SAS® In-Memory Statistics for Hadoop® (Neu).
 - SAS® Data Loader for Hadoop® (Neu).
- Ziele für die zukünftige Weiterentwicklung:
 - Verbesserung des Supports für Hive Server2 & Impala.
 - Unterstützung für weitere Hadoop® Distributionen und SQL Engines:
 - MapR, BigInsights, Pivotal HD, Hawq.
 - Vereinfachte Installation, Verbesserungen in der Performance.
 - Neue SAS® PROCs mit Pushdown Support:
 - Freq*, Means*, Tabulate*, Summary*, Report*, Sort, SQL*, Transpose, Corr, Univariate.

- Voller Support für die Distributionen von **Cloudera** (CDH 4.5/5.x) und **Hortonworks** (HDP 1.3/2.x) (Baseline, Hive, EP, HPA/LASR).
- Eingeschränkter Support for **Apache Hadoop®**, **Pivotal HD** (PHD 1.1), **IBM Big Insights** (2.1), **MapR** (V3) (Hive).

“And later” / “Alternate distributions” position:

SAS® will work with customers to support a SAS® deployment with later versions or alternative distributions on a reasonable-effort basis.

SAS® & HADOOP® ZUSAMMENFASSUNG

- ✓ Was ist Hadoop®?
 - **Horizontale Skalierung** (scale out) und **Datenlokalität**.
 - Kein Produkt, sondern „**Ökosystem**“.
- ✓ Wie kann ich es einsetzen? Und wieso?
 - Als **Erweiterung** des EDW.
 - **Kostenvorteile** und höhere **Agilität**.
- ✓ Welche Rolle kann SAS® dabei spielen?
 - SAS® stellt Hadoop® in den Mittelpunkt des Analytical Lifecycles: Datenintegration, Datenexploration und In-Database Analytics auf Hadoop® Big Data.



SAS® & HADOOP® LINKS UND REFERENZEN

- White Paper "Bringing the Power of SAS® to Hadoop®"
http://sas-competence-network.com/mediacenter/abstract/client/business-analytics/target_content_dokumente_white_paper_e2360_views2361_display_ger.html
- Globale Startseite „SAS® Solutions for Hadoop®“
http://www.sas.com/en_us/software/sas-hadoop.html
- Hortonworks: “Apache Hadoop® YARN: Yet Another Resource Negotiator”
<http://www.socc2013.org/home/program/a5-vavilapalli.pdf>
- 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>
- John Russell: "Cloudera Impala" (O'Reilly, free eBook)
<http://www.oreilly.com/data/free/cloudera-impala.csp>

4. November, Frankfurt am Main

Big Data Analytics Forum 2014



Wie profitieren Unternehmen von Big Data Analytics? SAS Highlight-Event mit Vorträgen zu Technologie, Skills & Rahmenbedingungen, dazu Marktübersicht, Visionen und die SAS-Strategie zu Big Data. Mit Beiträgen von Fraport, Adidas, Roland Berger, BARC, Fraunhofer, Claas u.a.

JETZT ANMELDEN: www.sas.de/bda-forum2014



VIELEN DANK !!



THE
POWER
TO KNOW.[®]