Big Data Projekte / Services

Übersicht aktueller Projekte &
Leistungsspektrum
Web Computing GmbH

Münster, 25.02.2016

Services / Leistungen

Servicekompetenzen und Leistungsspektrum

Services / Leistungen

Was wir bieten

01 Frontend-Systeme

Frontend-Systeme sind das visuelle Aushängeschild der zugrundeliegenden Datenschicht. In unseren Eigenentwicklungen setzen wir auf etablierte Technologien zur grafischen Aufbereitung der Inhalte (z. B. Highcharts.js) oder individualisieren vorhandene Lösungen (z. B. Logstash + Kibana).

02 (Near) Real Time

“Fast Data Strategien” sind die disziplinierte Herangehensweise an die Nutzung agiler, real-time, self-service Daten-Technologien, wie z. B. “Data virtualization”. Mit Hilfe dieser Services lassen sich Informationen schneller an Entscheider verteilen und erzielen somit einen positiven Effekt auf effizientere Entscheidungsfindung.
Branchenführende Unternehmen nutzen „Fast Data Strategien“ zur Erzielung umgehender time-to-value Resultate in ihren BI, Big Data und Data Services Projekten.

03 Big OLAP Systeme

Big OLAP Systeme setzen dort an, wo die klassischen analytischen Informationssysteme an ihre Grenzen stoßen. Neuartige Cluster-Services ermöglichen die Verwendung von Faktentabellen mit mehreren Milliarden Zeilen und bieten bei Abfragen, je nach Komplexität, sogar Antwortzeiten im Sub-Second Bereich. Die Befüllung der Daten-Cubes erfolgt in diesen Systemen entweder per Batch oder als kontinuierlicher Stream.

04 Data Lake

Je größer das Unternehmen, desto mehr Anwendungen, Geräte und andere Quellen erzeugen Log-, Clickstream oder ähnliche Formen von Informationen. Die Bewältigung dieser Massen an Daten wird durch die Verwendung von Data Lakes als Shared Service zwischen den unterschiedlichen Systemen ermöglicht. Flexible „Schema on Read“-Ansätze erlauben die nachträgliche Strukturierung und den geordneten Zugriff auf die in den Daten enthaltenen Informationen.

05 Data Mining

Heute benutzen Unternehmen aller Art maschinen-generierte Daten, um Informationen zur präzisieren und analysieren. Mittels Data Mining finden Sie z. B. Antworten auf die Fragen „Welche Produkte kaufen Besucher tendenziell zusammen und was werden sie am wahrscheinlichsten in Zukunft kaufen?“ oder „Wo sollte ich investieren, um die Nutzererfahrung meiner Website zu korrigieren oder zu verbessern?“

06 In-Memory Computing

Durch die ständige und explosionsartige Ansammlung von unstrukturierten Massen-Daten & Informationen werden immer größere Herausforderungen an die Verarbeitung, Analyse und Strukturierung gestellt. Die Ergebnisse sollen möglichst in Echtzeit ausgegeben werden und zum Abruf zuverlässig zur Verfügung stehen. In-Memory Verarbeitung zielt dabei nicht nur auf die Hardware ab, sondern bietet mit dem entsprechenden Softwaregegenstück ganz neue Möglichkeiten durch enorme Geschwindigkeitsvorteile gegenüber der klassischen Auswertung von Big-Data.

Agenda

Big Data – Anwendungen

Big Data – Technologien (Auszug)

Big Data – Projekte (auf Anfrage)

Anwendungen

Exkurs: Big Data - Definitionsaspekte

Die drei „V“

Kriterium	Business Intelligence	Big Data
Datenmenge („Volume“)	Über klassische relationale Datenbanktechnologien darstellbar	Nicht bzw. nur mit erhöhtem Aufwand über klassische relationale Datenbanktechnologien darstellbar
Datenart („Variety“)	Strukturiert (vorrangig)	Strukturiert, semistrukturiert, unstrukturiert
Datenverarbeitungs-geschwindigkeit („Velocity“)	Batch (vorrangig)	Batch (asynchron), Near-Time (semi-synchron), (Near) Real-Time (pseudo-synchron)

Exkurs: Dataspace

Was? Datenintegration ohne feste Zielstruktur

Warum? Reduktion und Variablisierung des Integrationsaufwandes

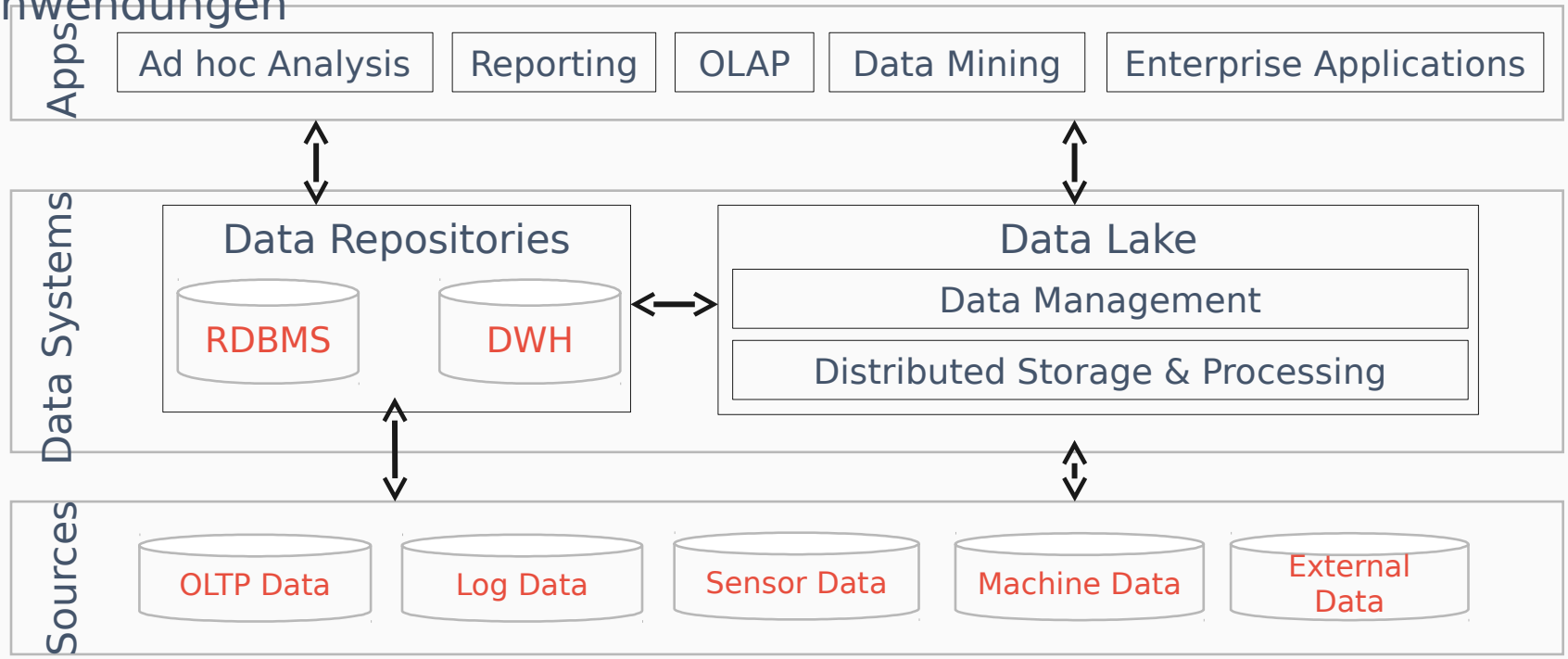
Kriterium	Database	Dataspace
Schemadefinition	Ex ante („Schema on write“)	Ex nunc („Schema on read“)
Datenobjekt	Tabelle (logisch)	Datei (physisch)
Datenart	Strukturiert	Alle Datenarten
Datenumfang	Auswahl von Daten	Alle Daten
Datenabfrage	Query	Search, Query („Schema on Read“)
Datenherkunft	Nach ETL i.d.R. nicht rekonstruierbar	„Data lineage“

Vgl.: Franklin, M. und Halevy, A. und Maier, D..: From Databases to Dataspaces, http://www.eecs.berkeley.edu/~franklin/Papers/dataspaceSR.pdf, Abruf: 24.10.2014

Data Lake

Was? Sammlung, Speicherung, Verwaltung von Rohdaten und verarbeiteten Daten

Warum? Skalierbarer Datenspeicher für Ad-hoc-Analysen und Anwendungen

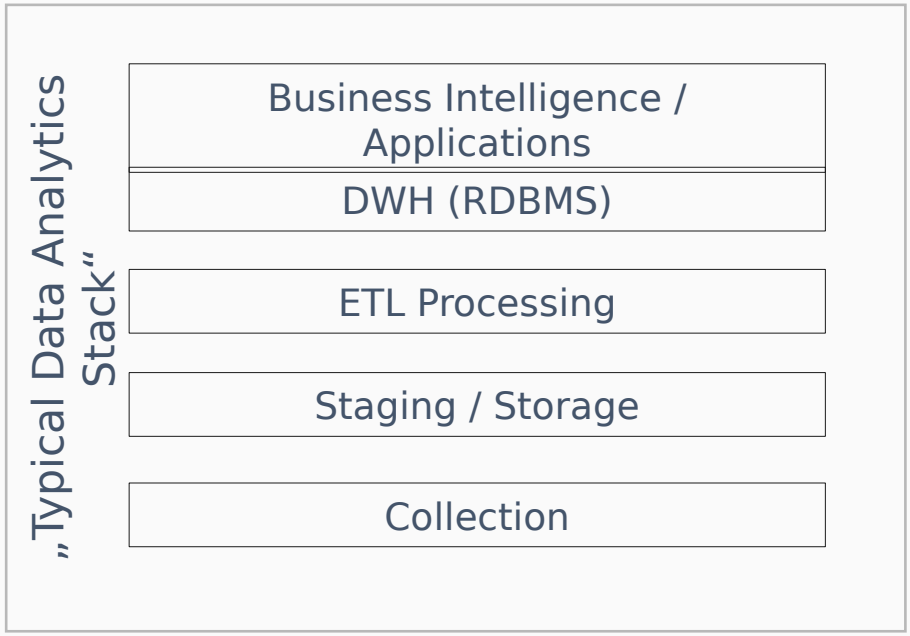


1) Abbildung in Anlehnung an: Connolly, S.: Enterprise Hadoop and the Journey to a Data Lake, http://hortonworks.com/blog/enterprise-hadoop-journey-data-lake/, Abruf: 24.10.2014

Enterprise Data Hub (EDH)

Was? Universale skalierbare Datenspeicherung und -verarbeitung

Warum? Kosteneffizienz und neue Auswertungsmöglichkeiten



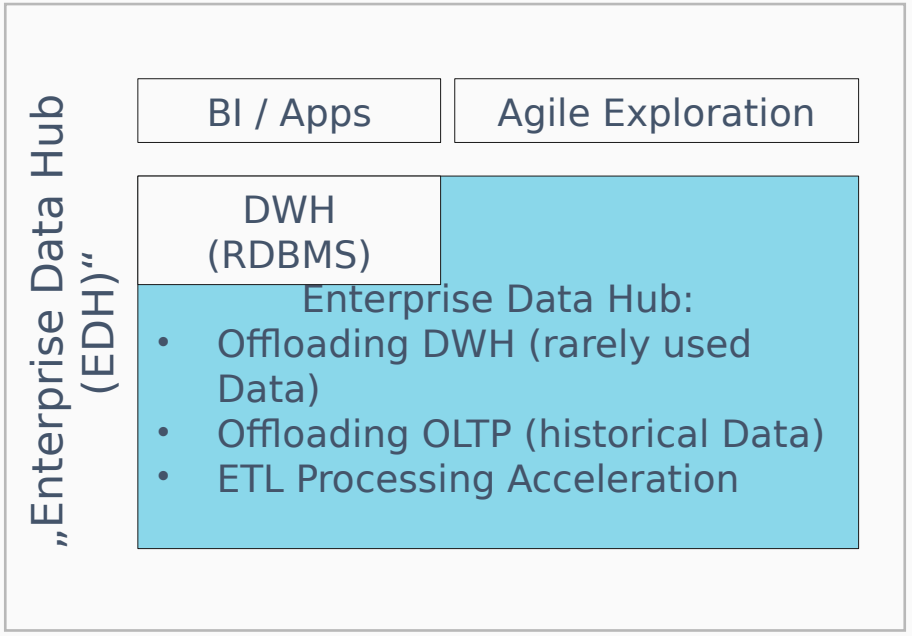


Abbildung in Anlehnung an: Patterson, C: The Future of Data Management [...], http://de.share.net/cloudera/keynote-future-of-big-data-32682662, Abruf: 24.10.2014

(Near) Real-time Scalable Data Warehousing

Was? Verteiltes DWH-System, das u. g. Anforderungen realisiert

Warum? Multidimensionale interaktive Auswertung für Billing, Reporting, ...

- **Atomare Aktualisierungen**: Einzel-Transaktionen (Fakten) müssen dem DWH während des laufenden Betriebs hinzugefügt werden können
- **Datenkonsistenz und -korrektheit**: Gleichzeitig müssen die Daten jederzeit streng konsistent sein (erfordert ACID-Eigenschaften)
- **Verfügbarkeit**: Kein „single point of failure“, keine „downtime“ für geplante oder ungeplante Wartung (inkl. eines Ausfalls eines ganzen Rechenzentrums)
- **Nahzeit-Aktualisierungs-Durchsatz**: Kontinuierliche Datenaktualisierung mit Millionen von Zeilen je Sekunde; müssen in Minuten für Abfragen bereitstehen
- **Abfragegeschwindigkeit**: Abfragen müssen eine Latenz von weniger als Hunderte-Millisek. aufweisen; Abfragedurchsatz: Billionen von Zeilen/Tag
- **Skalierbarkeit**: Skalierbarkeit für den Daten- und Abfrageumfang (Petabytes)
- **Online-Datenschemaänderungen**: Datenschema-Änderungen im laufenden Betrieb ohne Update- u. Query-Beeinträchtigungen

In Anlehnung an.: Gupta, A. et. al.: Mesa: Geo-Reeplicated, Near Real-Time, Scalable Data Warehousing, http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/42851.pdf, Abruf: 15.11.2014

Real-time Applications

Was? Echtzeit-Auswertung von zeitkritischen Ereignisdaten
Warum? Prävention und Optimierung von Geschäftsvorgängen

	„Prevent“ Use Cases	„Optimize“ Use Cases
Banking	▪ Fraud Detection ▪ Compliance Violations	▪ Algorithmic Trading ▪ Dynamic Pricing
Telecom	▪ Network Monitoring ▪ Security Intelligence	▪ Bandwith Allocation ▪ Customer Service
E-Commerce	▪ Recommendations ▪ Fraud Detection	▪ Customer Scoring ▪ Dynamic Pricing
Web of Things	▪ Sensor Data Monitoring ▪ Control Devices	▪ Event Analytics ▪ Control Devices

Abb. in Anlehnung an: Hortonworks: Apache Storm: A System for processing streaming data in real time, http://hortonworks.com/hadoop/storm/, Abruf: 25.10.2014

Data Mining

Was? Mustererkennung in Massendaten zur Entscheidungsvorbereitung
Warum? Generierung von Wissen (Muster) aus Erfahrung (Daten)

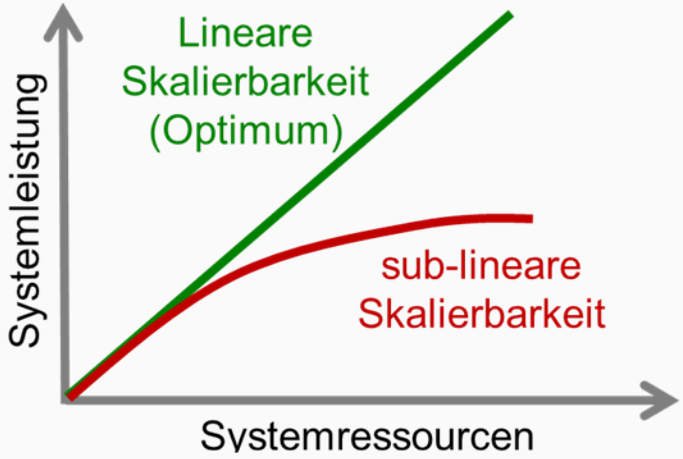
Untersuchungsfragestellungen:	Exemplarische Anwendung
Assoziation (strukturierte Daten)	☾ Warenkorb-Analyse („Cross-/Up-Selling Analysis“) ☾ Online-Empfehlungen („Recommendation System“)
Klassifikation (strukturierte Daten)	☾ Kreditwürdigkeitsanalyse („Credit Scoring“) ☾ Kundenabwanderungsanalyse („Churn Analysis“) ☾ Betrugserkennung („Fraud Detection“)
Segmentierung (strukturierte Daten)	☾ Kundensegmentierung („Customer Segmentation“) ☾ Kundenwertanalyse („Customer Value Analysis“)
„Text Mining“ (unstrukturierte Daten)	☾ Stimmungsanalyse („Customer Sentiment Detection“) ☾ Wettbewerbsanalyse („Competitive Intelligence“)

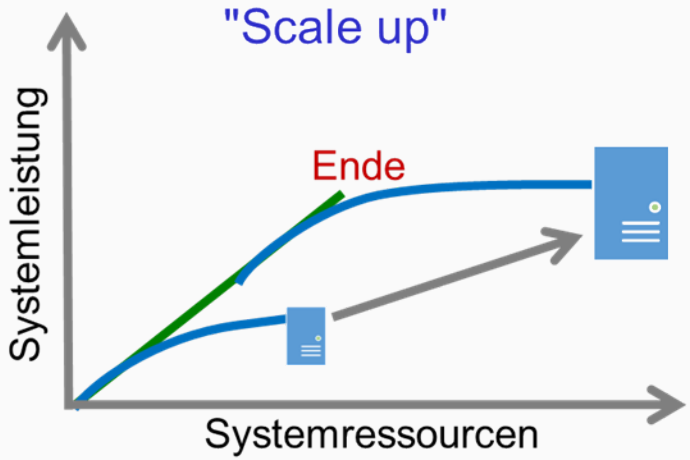
In Anlehnung an.: Gupta, A. et. al.: Mesa: Geo-Reeplicated, Near Real-Time, Scalable Data Warehousing, http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/42851.pdf, Abruf: 15.11.2014

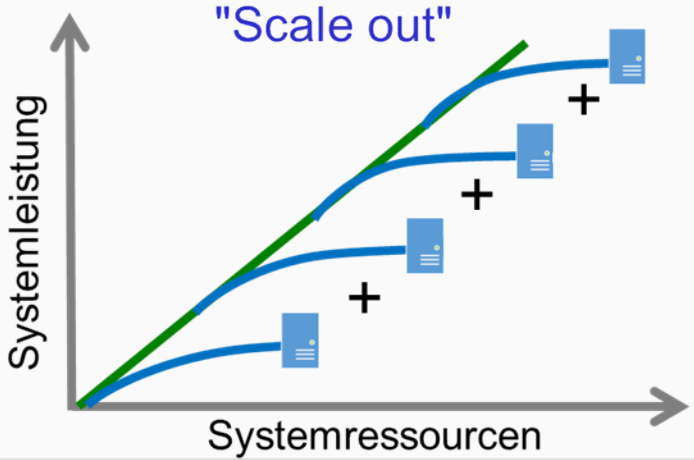
Technologien

Exkurs: System Scalability

Was? Fähigkeit eines Systems, die Verarbeitungsleistung durch Erweiterung von Ressourcen linear steigern zu können

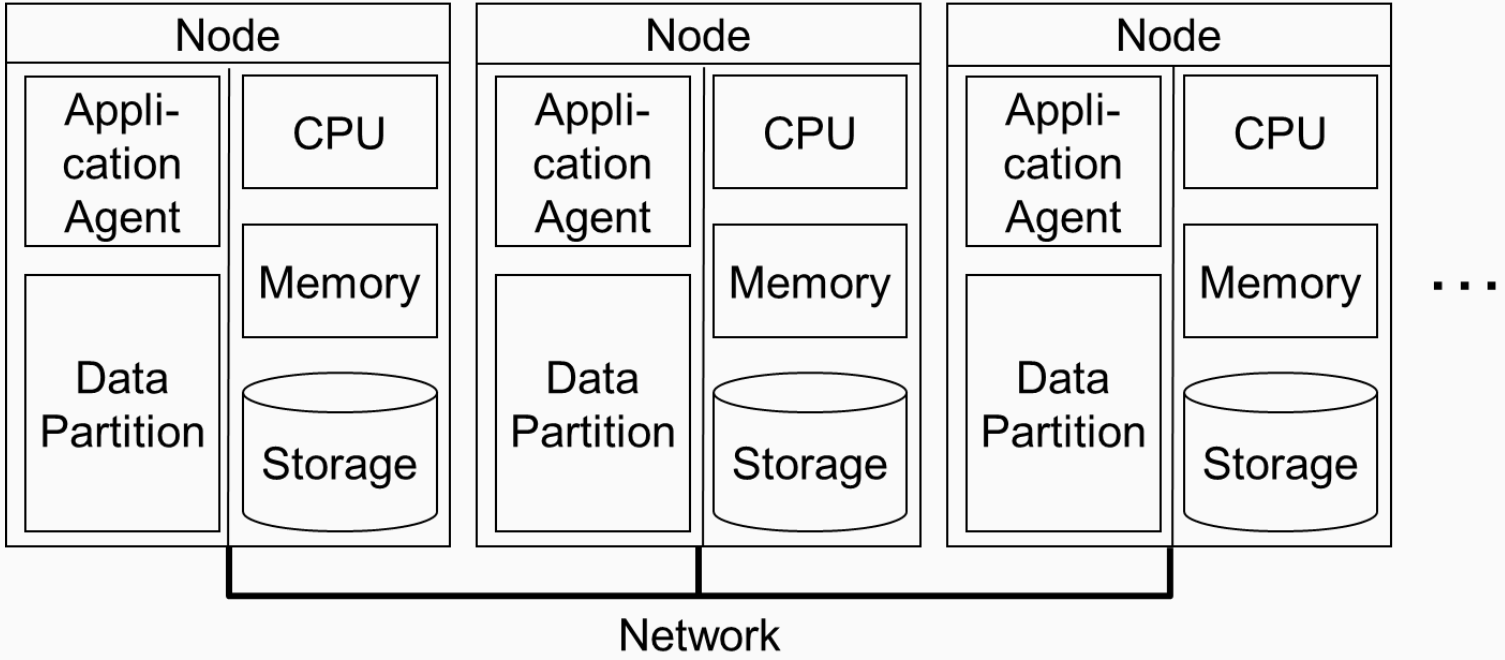






Exkurs: Shared Nothing Architecture

Was? Jede Station (syn.: Knoten, node) ist unabhängig von anderen Stationen im verteilten System, d. h. ist zuständig für eine bestimmte Datenpartition und verfügt über eigene Applikationsagenten sowie Ressourcen (CPU, Memory, Storage)



Distributed Data Storage

Was? Ausfalltolerante verteilte Datenspeicherung und -verarbeitung im Verbund (unzuverlässiger) Standard-Rechner

Warum? Kosteneffiziente und skalierbare Massendatenspeicherung

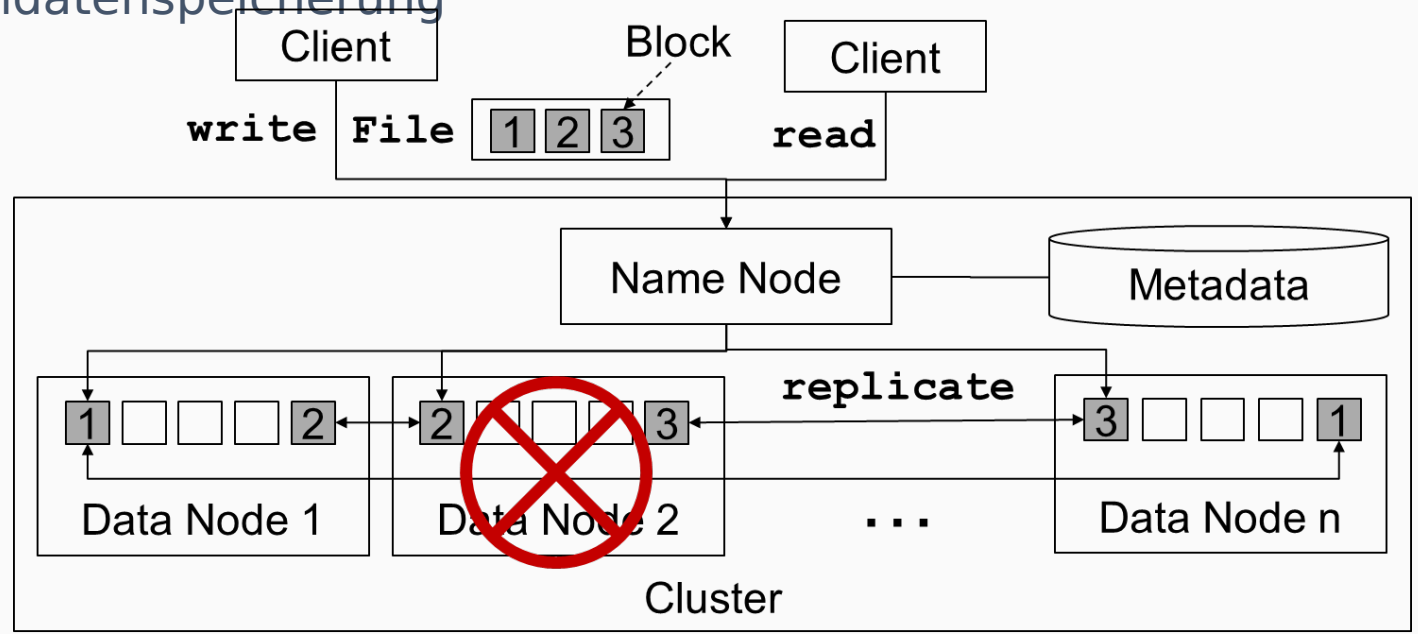


Abbildung in Anlehnung an: http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html, Abruf: 24.10.2014

Distributed Data Processing - Cluster Resource & Job Management

Was? Aufgabenverwaltung und Ressourcenzuteilung im Rechnerverbund

Warum? Abstraktion für verteilte Datenverarbeitungsalgorithmen

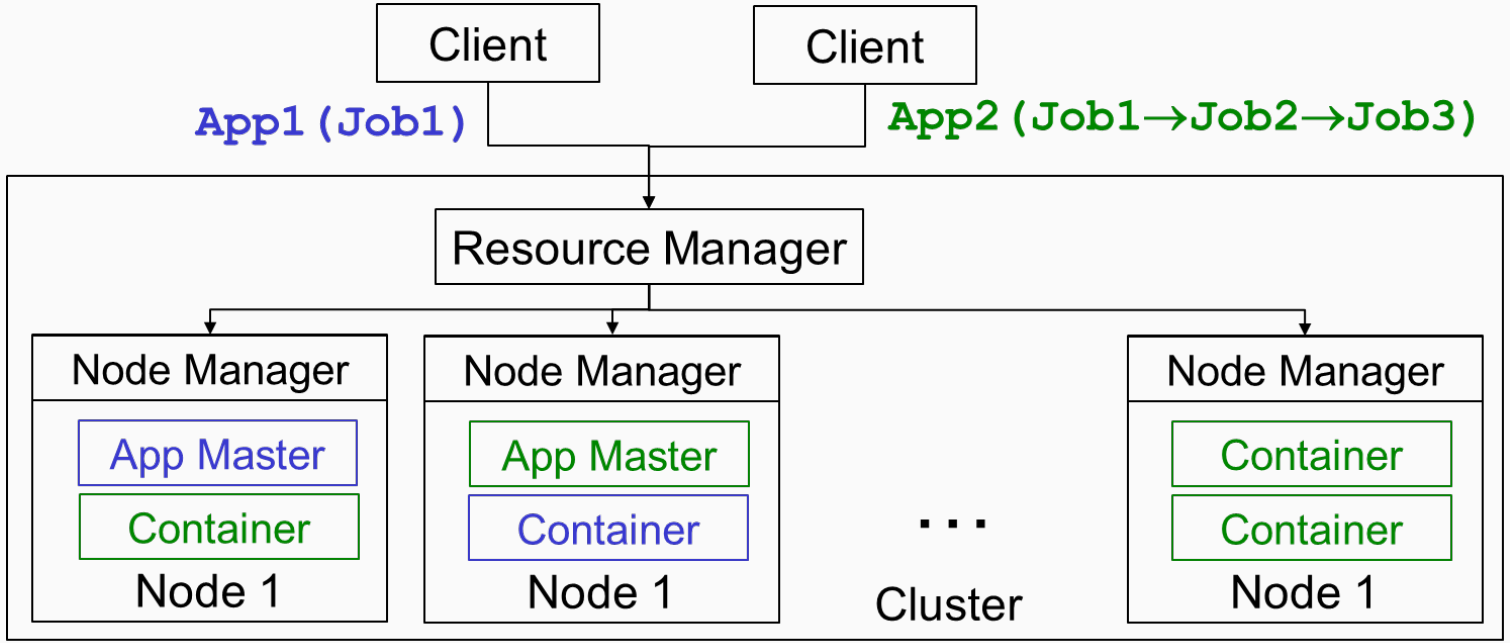
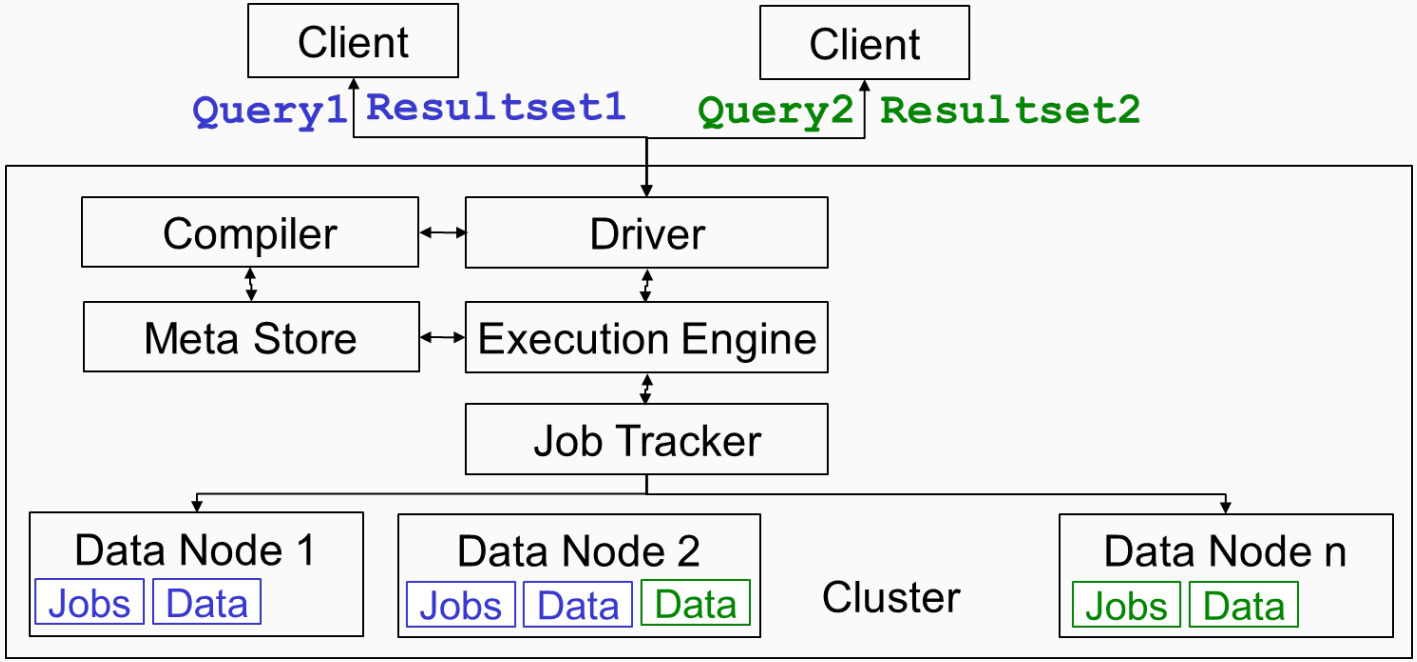


Abbildung in Anlehnung an: http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html, Abruf: 24.10.2014

Data Warehouse Infrastructure based on Distributed Data Storage & Processing

Was? DWH-Metadatenverwaltung und Datenabfrageschnittstelle

Warum? Ermöglichung SQL-Abfragen auf relationale Schemata

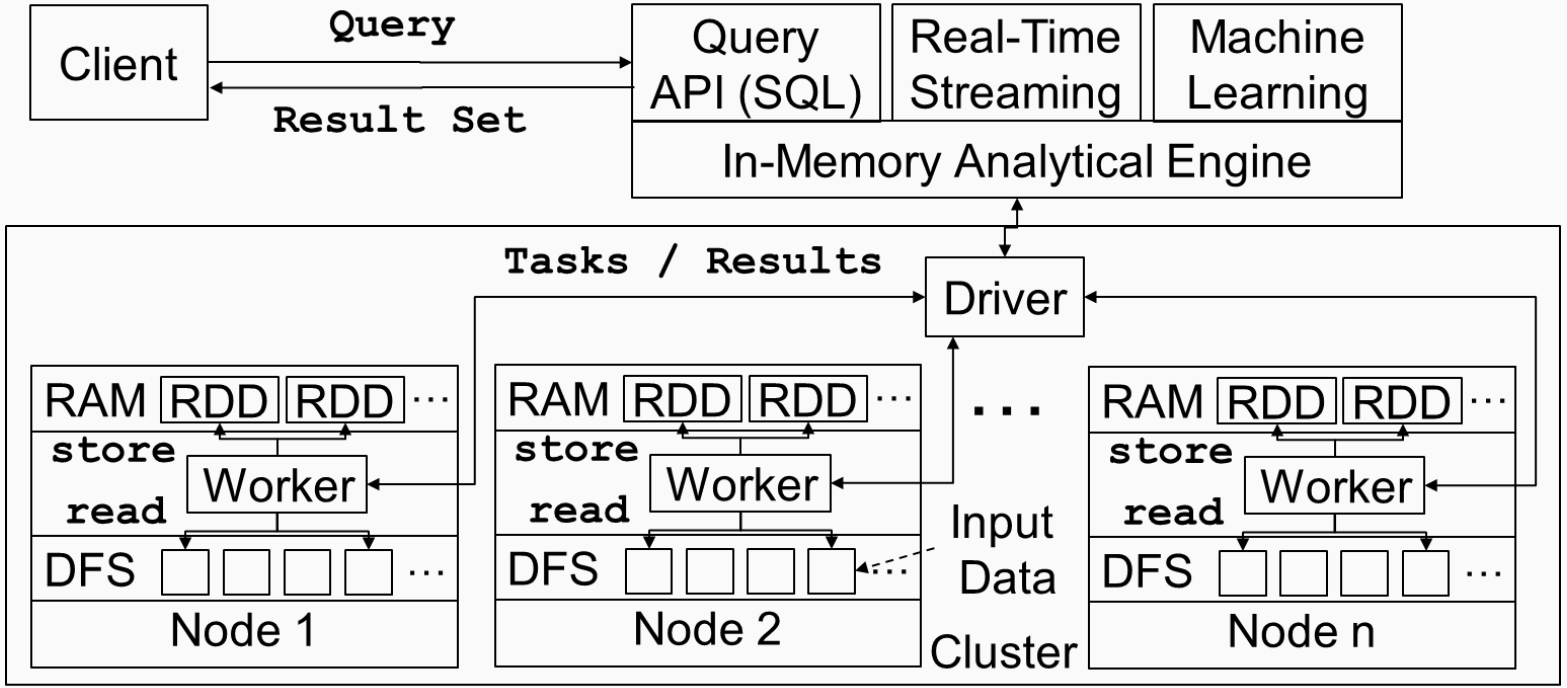


Vgl. hierzu: https://cwiki.apache.org/confluence/display/Hive/Design/, Abruf: 20.11.2014

Data Analytics In-Memory Cluster Computing

Was? Verteilte & parallele Datenverarbeitung im Arbeitsspeicher

Warum? Voraussetzung für interaktive und Echtzeit-Analysen

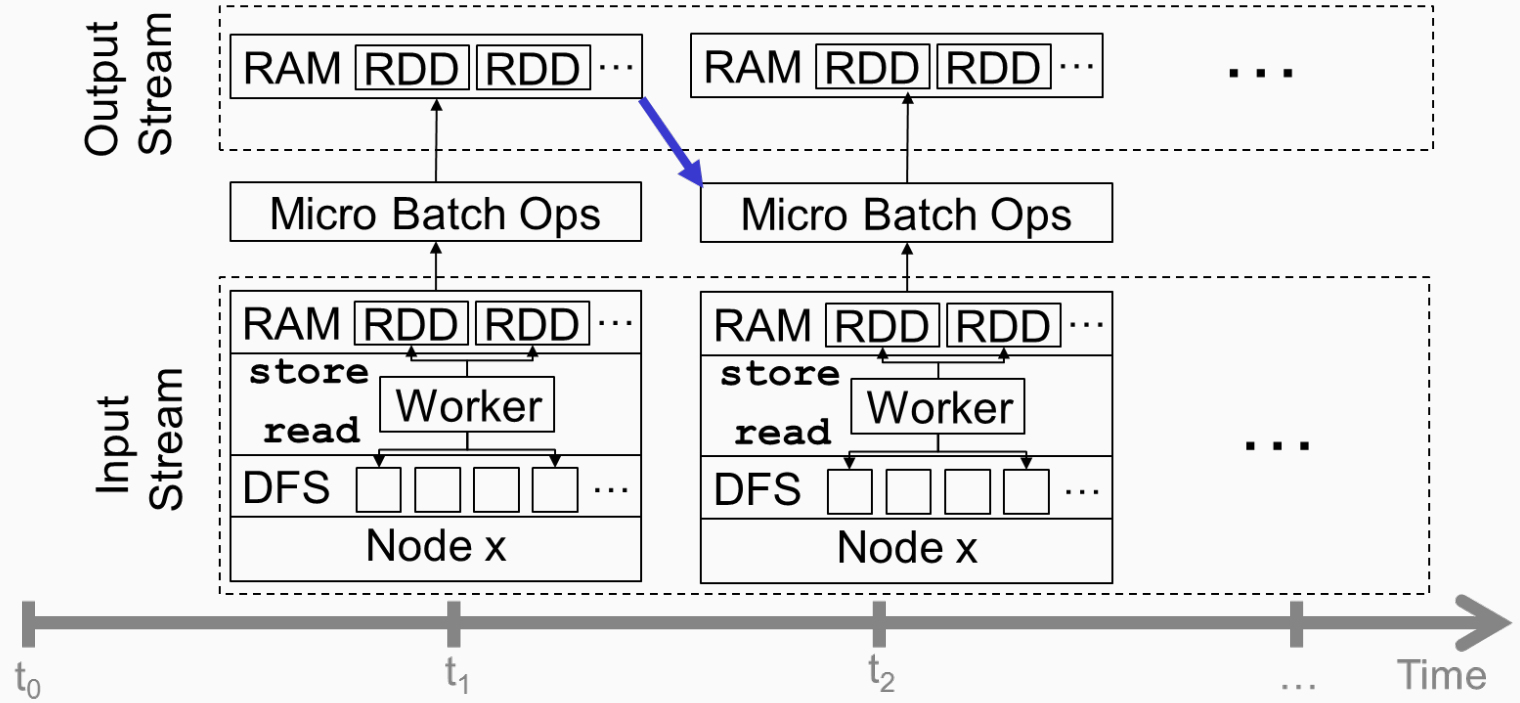


Vgl.: Zaharia, M. et al.: Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing, https://www.cs.berkeley.edu/~matei/papers/2012/nsdi_spark.pdf, Abruf: 25.10.2014
Abkürzungen: Resilient Distributed Dataset (RDD), Distributed File System (DFS), Structured Query Language (SQL), Application Programming Interface (API)

Real-Time In-Memory Data Processing

Was? „Micro Batch“-Verarbeitung im Arbeitsspeicher

Warum? Auswertungen von zeitkritischen Ereignisdaten

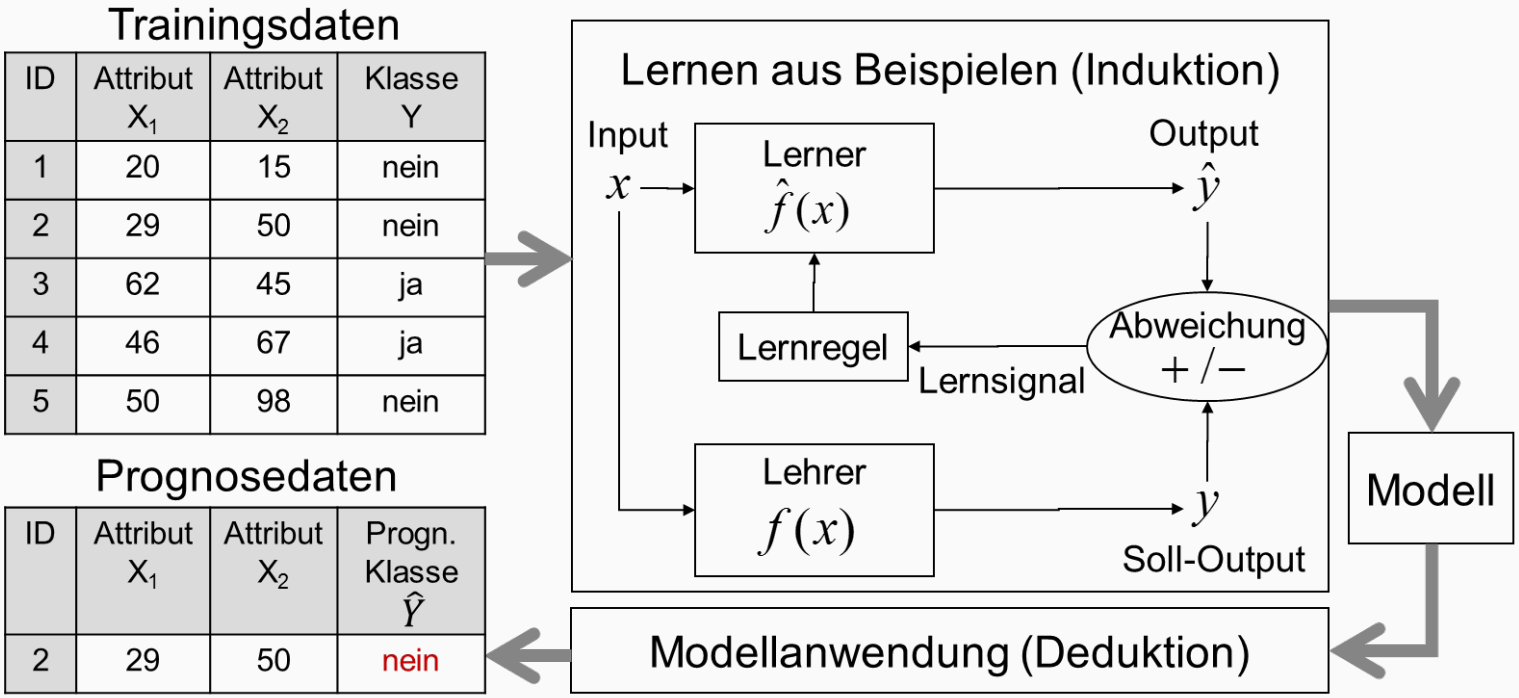


Vgl.: Zaharia, M. et al.: Discretized Streams: An Efficient and Fault-Toleratn Model for Stream Processing on Large Clusters, https://www.usenix.org/system/files/conference/hotcloud12/hotcloud12-final28.pdf, Abruf: 24.10.2014

Machine Learning

Was? Rechner-basierte mathem. Verfahren zur Mustererkennung

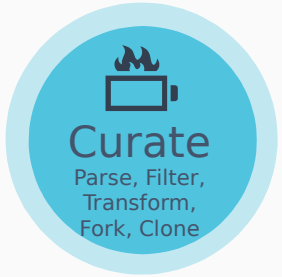
Warum? Automatische Abstraktion (Modell) aus Einzelfällen (Daten)



Apache NiFi

- *Daten verbinden*:
  Aggregation der Daten zahlreicher IoAT Systeme: Sensoren, Geo-location Geräte, Maschinen, Logs, Dateien und Feeds über abgesicherte Kanäle

- *Steuerung des Data Flows*:
  Verbinden von point-to-point und bi-direktionalen Data Flows. Zuverlässige Zustellung von Informationen an real-time Applikationen und Storage-Plattformen



- *Einblicke gewinnen*:
  Parsen, Filtern, Verbinden, Transformieren, Forken und Klonen von beweglichen Daten schafft die Möglichkeit zur Analyse kurzlebiger Informationen

Apache Hadoop

The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing

- Freies, in Java geschriebenes Framework für skalierbare, verteilt arbeitende Software
- Basiert auf dem MapReduce-Algorithmus von Google Inc.
- Ermöglicht es, intensive Rechenprozesse mit großen Datenmengen (Big Data, Petabyte-Bereich) auf Computerclustern durchzuführen
- Besteht aus mehreren Komponenten:
  - **Hadoop Common**: bietet ein Toolset aus Grundfunktionen, das alle anderen Bausteine benötigen (u. a. Schnittstelle für die Remote-Procedure-Call-(RPC-)Kommunikation innerhalb des Clusters)
  - **Hadoop Distributed File System (HDFS)**: hochverfügbares Dateisystem zur Speicherung sehr großer Datenmengen auf den Dateisystemen mehrerer Rechner (Knoten); Dateien werden in Datenblöcke mit fester Länge zerlegt und redundant auf die teilnehmenden Knoten verteilt.
  - **Yarn**: ermöglicht es, die Ressourcen eines Clusters dynamisch für verschiedene Jobs zu verwalten
  - **MapReduce**: MapReduce-Algorithmus mit konfigurierbaren Klassen für Map, Reduce und Kombinationsphasen

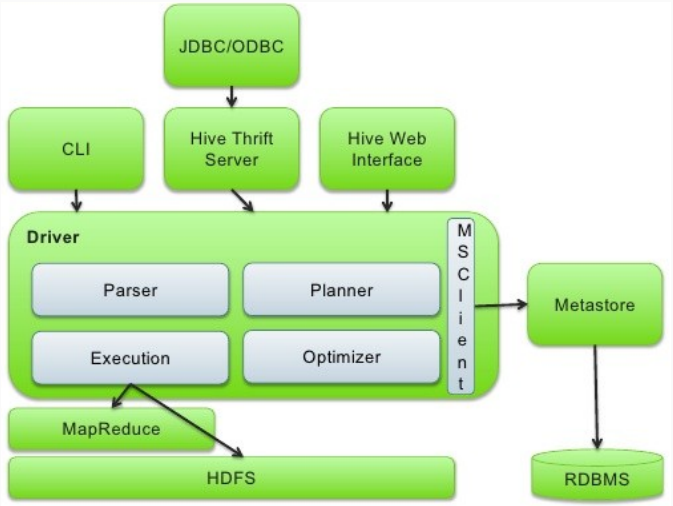
Apache Hive

The de facto standard for SQL queries in Hadoop

- Ziel: Abfrage von riesigen Datenmengen auf verteilten Systemen mit SQL
- Data Warehouse – Schicht für HDFS-Daten

- Nutzt HiveQL, eine an SQL2011 angepasste SQL-Variante
- SQL-Abfragen werden in MapReduce oder Tez-Jobs zerlegt
- Greift den Hadoop-typischen „Schema-on-read“ – Ansatz auf
- Schemata werden in HiveQL geschrieben
- Mehrere Schemata auf den selben Daten sind möglich
- Verarbeitung als Batch-Job ebenso wie interaktive Abfragen

- JOINs ermöglichen Verbindungen zwischen verschiedenen Dateien
- INSERT, UPDATE und DELETE Statements können durchgeführt werden
- Spaltenorientierte Speicherformate wie ORC werden unterstützt

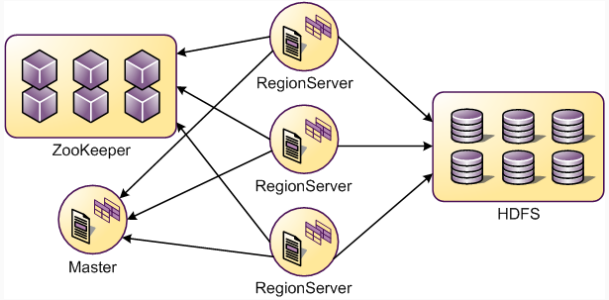


Bildquelle: https://hive.apache.org/images/hive_logo_medium.jpg

Apache HBase

Apache HBase – The Hadoop database, a distributed, scalable, big data store.

- Ziel: Bereitstellung von Tabellen mit Milliarden an Zeilen und Millionen an Spalten
- NoSQL-Datastore auf Basis von Googles BigTable



- Unterstützung von Datenversionierung
- Multidimensional (Zeitbezug automatisch vorhanden)
- Automatisches Sharding
- Automatisches Failover
- Schnittstellen: Java-API, Thrift Gateway und REST-ful Web-Service
- Strikte Konsistenz (CP-Einordnung beim CAP-Theorem)

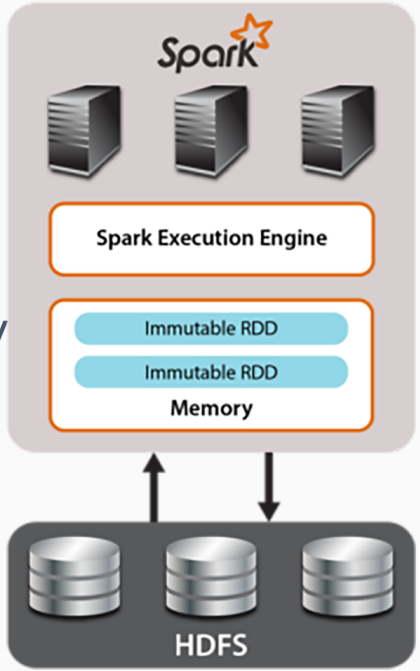
- Clusterbasiert auf Basis von Apache Hadoop und Apache HDFS
  - Horizontale Skalierbarkeit
  - Vorteile von HDFS, etwa Datenreplikation
  - Vollständige Integration in den Hadoop-Stack ☾ Kombinierbar z.B. mit Apache Hive und Apache Ranger

Bildquelle: https://hbase.apache.org/images/hbase_logo_with_orca_large.png

Apache Spark

Apache Spark – In-Memory Cluster Computing Framework

- In-Memory Cluster Computing Framework
- Daten werden In-Memory auf dezentralen Cluster-Nodes gespeichert, verwaltet und verarbeitet
- Ausgelegt für interaktive Datenanalysen
- Datenmenge kann vertikal skalieren (Scale-out Ansatz)
- Abfragen per SQL (Spark SQL) möglich
- Enge Kopplung mit Apache Hadoop und dem Hadoop Ökosy
  - Apache YARN
  - Apache Hive
  - Apache HDFS

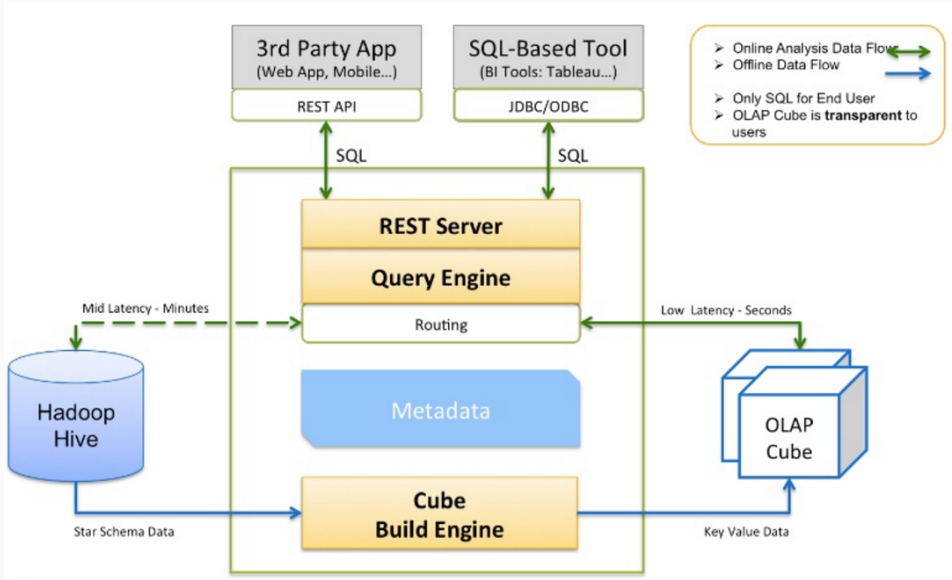


Apache Kylin

Apache Kylin – Extreme OLAP Engine for Big Data

- Ziel: OLAP-Abfragen mit SQL auf Basis von Apache Hadoop
- Multidimensional Online Analytical Processing (MOLAP)-System auf Basis von Apache HBase

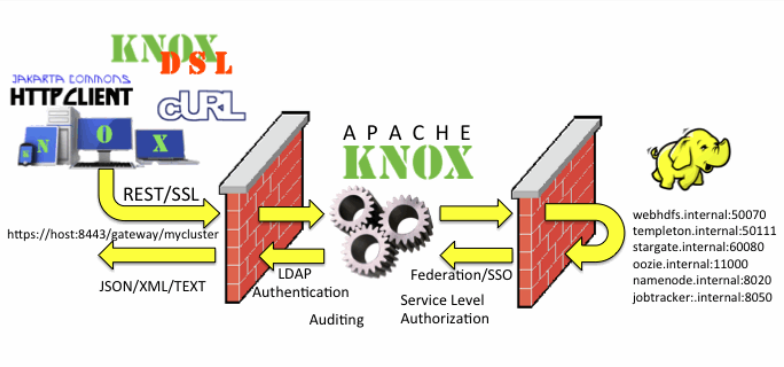
- Definition von OLAP-Cubes im Star-Schema
- Datenbasis aus Apache Hive
- Ausgelegt für OLAP-Sub-Second – Querys auf Milliarden von Zeilen
- Abfragen erfolgen in Form von ANSI-SQL
- Schnittstellen: JDBC, REST-ful Web-Service
- Interaktion mit BI-Tools möglich, z. B. Tableau, Microstrategy, Excel, ...
- Job-Management / Monitoring



Apache Knox / Apache Ranger

Apache Knox – Secure entry point for Hadoop clusters / Apache Ranger – Comprehensive security for Enterprise Hadoop

- Ziel: Bereitstellung eines Single Access Points für REST-ful Kommunikation
- REST API Gateway für den Hadoop Cluster



- Knox arbeitet als zustandsloser Reverse Proxy
- Berücksichtigung der REST und HTTP-Kommunikation
- Verbirgt Cluster-Host-Adressen und Ports
- Nutzung von LDAP oder Active Directory (Authentifizier
- Aufgreifen extern Authentifizierungs-Events (Federatior
- Unterstützung von Kerberos
- Kombinierbar mit Apache Ranger (Autorisierung)
- Auditing-Funktionalitäten
- Unterstützte Services: HDFS, HBase, Oozie, Hive, Yarn, Storm

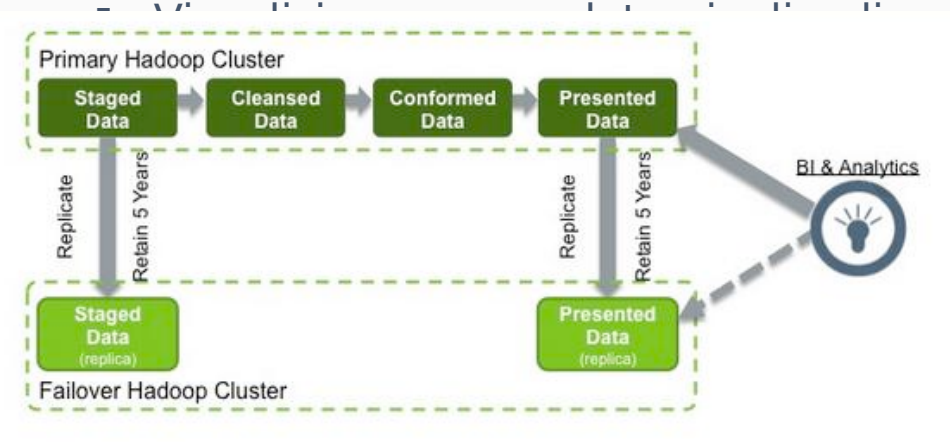
- Durchsetzung der Policies durch leichtgewichtete Java-Plugins
- Auch bei Ausfall des Policy-Servers werden Policies durchgesetzt
- Monitoring-Daten werden von den Plugins gesammelt
- Zentrale Speicherung der Audit-Logs im Audit Server

Apache Falcon

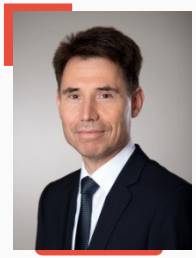
Apache Falcon – Feed processing and feed management system

- Zentralisiertes Data-lifecycle-management
  - Definition & Management von Abläufen für Dateneinspeisung, -verarbeitung und –export
  - Sicherstellung von Disaster-readiness
  - Policies für Datenreplikation (Retention, etc.)
  - End-to-End Monitoring von Abläufen

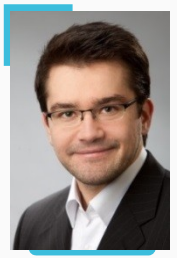
- Compliance und Auditing Funktionalitäten
  - Visualisierung von Daten-Lineage‘ / Überwachung von ‚data pipeline audit
  - Daten-Replikation und –Archivierung
    - Data lineage unterstützender Dokumentation
    - Heterogenes ‚Storage tiering‘ in HDFS
    - Definition von hot/cold storage tiers im Cluster
    - HDFS-Snapshot Support



Kontakt



Prof. Dr. Wolfgang Wicht
Partner
Mail ww@web-computing.de
Mobil +49 177 9721441



Sebastian Zimmermann
CEO
Mail sz@web-computing.de
Telefon +49 251 39655243
Mobil +49 177 6563083