Beratung

Results,
no Excuses.

Lösungen

Grown from
Experience.

Ventum Consulting

SQL auf Hadoop

Oliver Gehlert

© 2014 Ventum Consulting

1 Ventum Consulting – Daten und Fakten

Results, no excuses

Fachwissen

- Strategisches Businessverständnis
- Fachliche und methodische Prozessexpertise
- Breites IT KnowHow

Branchenkenntnis

- Automotive & Manufacturing
- Telco & Media
- Financial Services
- Professional Services

Umsetzungsstärke

- DE: >70 MA Int: ~100 MA
- DE|AT|CH|CN
- Fachliches Netzwerk

Unser Versprechen

Results, no excuses.

© 2014 Ventum Consulting

Agenda

1 Einführung

2 Hive, Impala, Tajo and Co

3 Performance

4 Ausblick

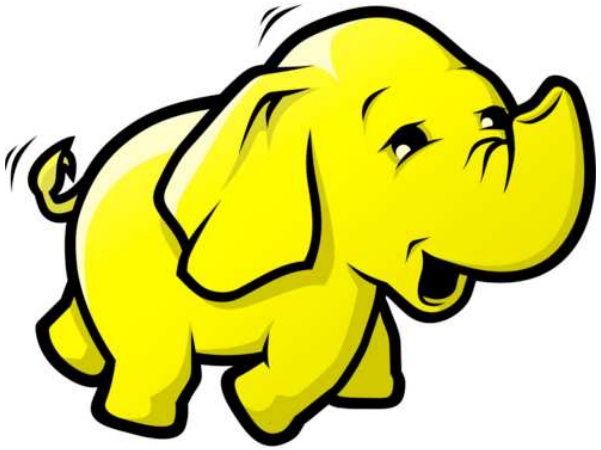
© 2013 Ventum Consulting

SQL auf Hadoop

Muss man für Hadoop Java programmieren?

ventum

Java Programmierung war die Basis für den Zugriff auf Hadoop



© 2014 Ventum Consulting

SQL auf Hadoop

Es gibt zahlreiche Ansätze für SQL auf Hadoop

ventum

Ausgewählte SQL Frameworks

© 2014 Ventum Consulting

SQL auf Hadoop

SQL auf Hadoop verlangt strukturierte Daten

Images office.com

© 2014 Ventum Consulting

SQL Frameworks

Wie haben wir getestet

Testsetup

- **Sandboxserver mit**
  - Cloudera 5 Beta 2
  - 16 GB RAM
  - Testdatengröße 20 GB
- **Sandboxserver mit**
  - Hortonworks HDP 2.0
  - 16 GB RAM
  - Testdatengröße 20 GB
- **Cluster**
  - Amazon Elastic Map Reduce
  - 5 Nodes m2.4xl
  - Testdatengröße 100 GB

Testdaten

- **Hive-Testbench basierend auf TPC-DS**
  - Download unter https://github.com/cartershanklin/hive-testbench
  - Starschema und ausgewählte Abfragen

© 2014 Ventum Consulting

Agenda

1 Einführung

2 Hive, Impala, Tajo and Co

3 Performance

4 Ausblick

© 2013 Ventum Consulting

SQL Frameworks
Hive

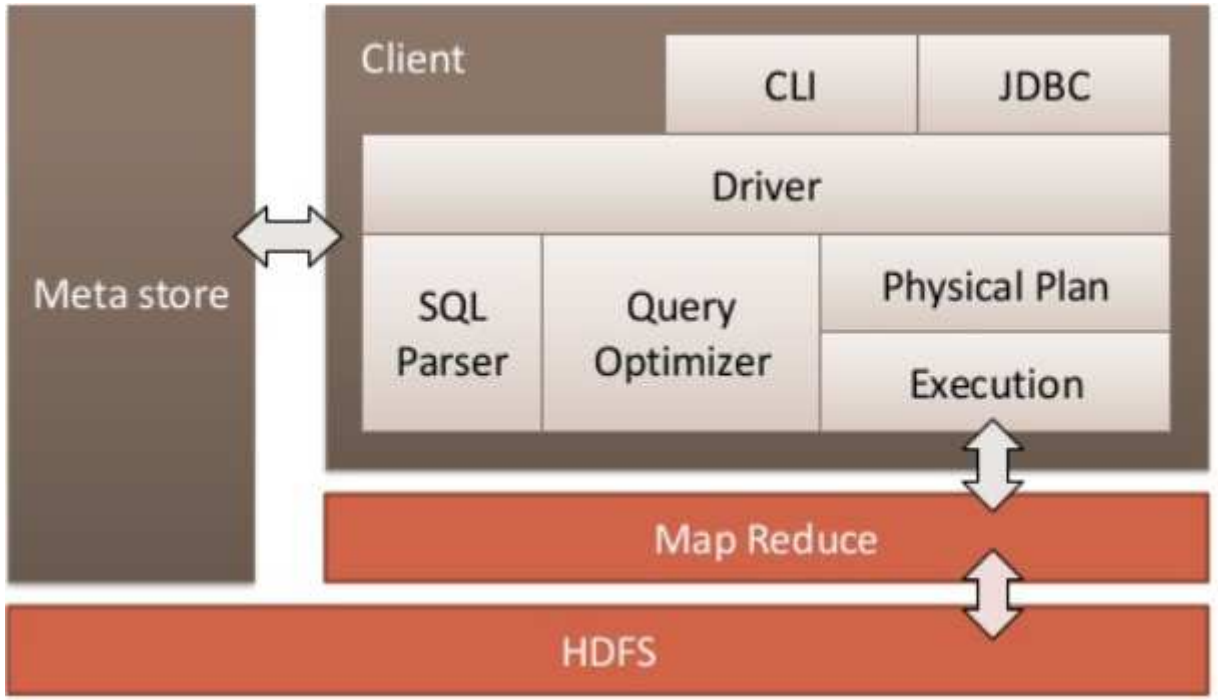
Überblick

- Ein Datawarehouse System um strukturierte Daten im HDFS zu speichern
- Unterstützt zahlreiche Dateiformate
- Strukturiert die Daten analog zu Datenbankkonzepten in
  - Tabellen
  - Partitionen
  - Spalten
  - Zeilen
- Ermöglicht einen einfachen SQL basierten Zugriff
- Hive Metadata Store wird von vielen Frameworks verwendet

© 2014 Ventum Consulting

SQL Frameworks
Hive Architektur

Architektur



Images apache.org

© 2014 Ventum Consulting

SQL Frameworks

Hive – Stärken und Schwächen

Stärken

- Stabil
- Intensive Weiterentwicklung
- Memoryanforderungen
- Batchprocessing
- Umfangreiche SQL Unterstützung
- Hive Metastore
- UDF

Schwächen

- Interaktive Abfragen
- Keine Einzelsatzverarbeitung
- Map-Reduce

© 2014 Ventum Consulting

SQL Frameworks

Hive – SQL Unterstützung

Unterstützung

- Analytische Funktionen
- Subselects
- Group by
- Rollup
- UNION ALL

Nicht unterstützt:

- UNION
- MINUS
- INTERSECT

© 2014 Ventum Consulting

SQL Frameworks

Stinger bündelt Performanceoptimierungen

ventum

Überblick

- **Die Stinger Initiative vereint mehrere Performance-optimierungen für Hive**
- **Optimiertes Fileformat orcFile**
  - Komprimierung um Faktor 4 im Vergleich zu Text
  - Predicate Pushdown
- **Analytische Funktionen**
- **YARN**

© 2014 Ventum Consulting

SQL Frameworks

Stinger – Stärken und Schwächen

Stärken

- Stabil
- Intensive Weiterentwicklung
- Memoryanforderungen
- Batchprocessing
- Umfangreiche SQL Unterstützung
- Hive Metastore

Schwächen

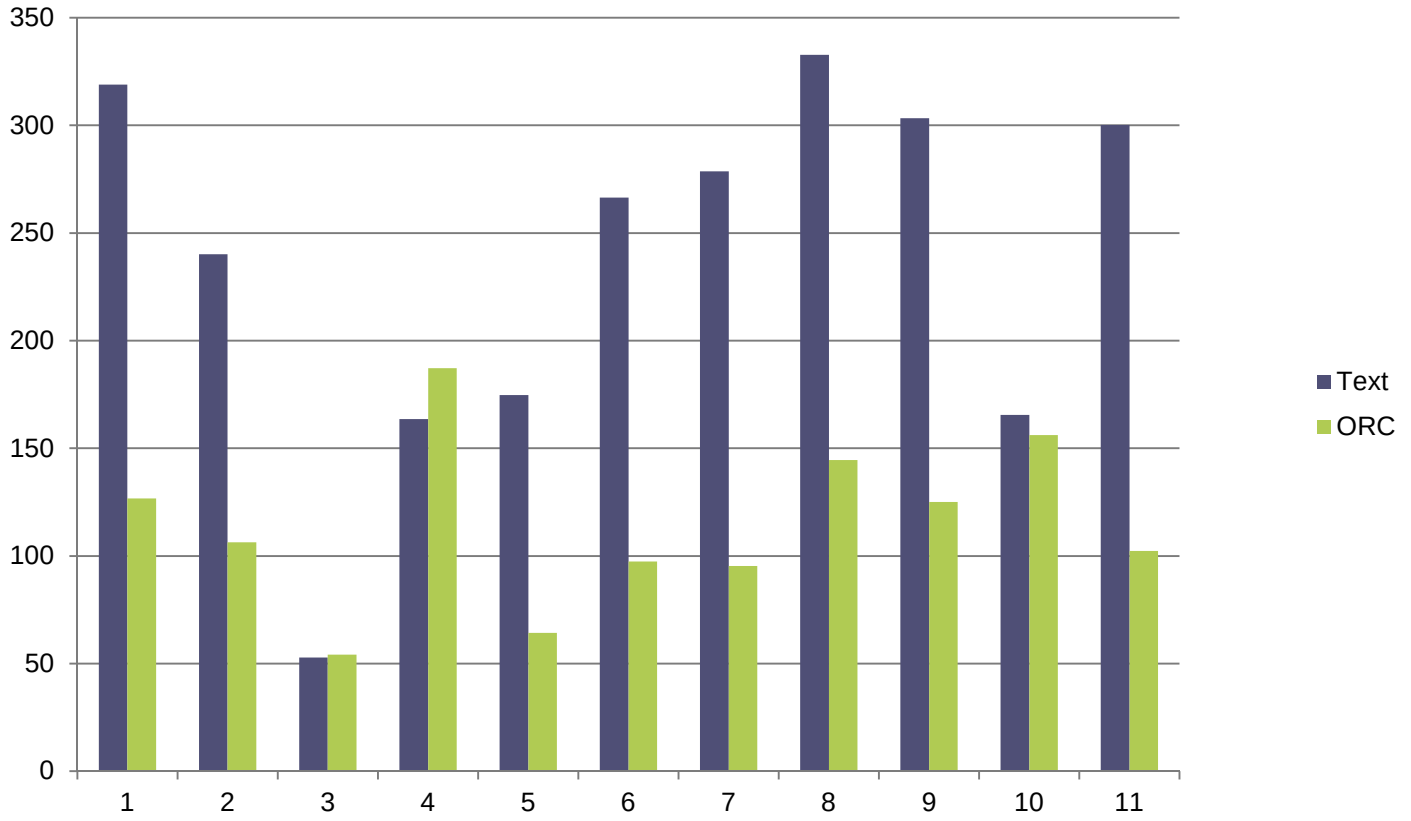
- Interaktive Abfragen
- Map-Reduce
- Keine Einzelsatzverarbeitung

© 2014 Ventum Consulting

SQL Frameworks

Stinger – SQL Performance

Performance Vergleich Speichertypen



© 2014 Ventum Consulting

SQL Frameworks

Tajo

ventum

Überblick

- Ein Datawarehouse System um strukturierte Daten im HDFS zu speichern
- Toplevel Projekt der Apache Foundation
- Unterstützt zahlreiche Dateiformate
- Ermöglicht einen einfachen SQL basierten Zugriff
- Kann Hive Metadata Store benutzen
- Fully distributed SQL processing

© 2014 Ventum Consulting

SQL Frameworks
Tajo Architecture

Architektur

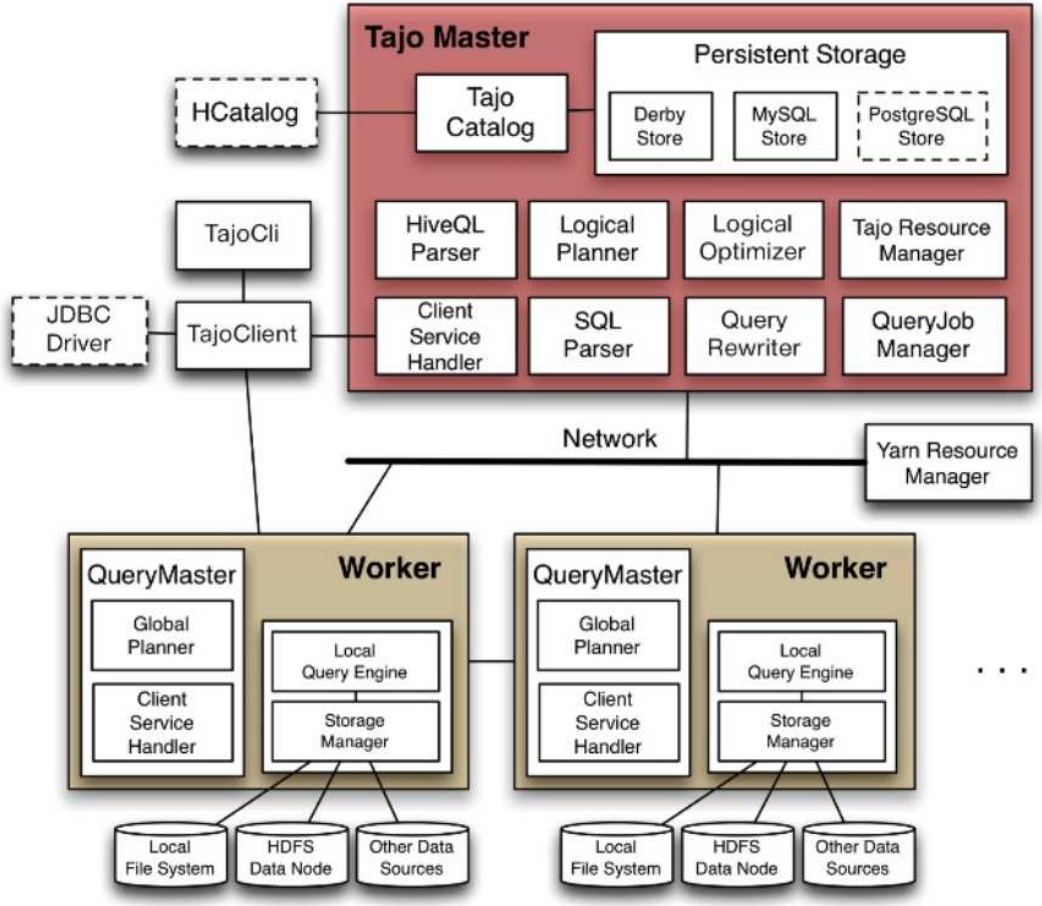


Image from apache.org

© 2014 Ventum Consulting

SQL Frameworks

Tajo – Stärken und Schwächen

Stärken

- Hive Metastore möglich
- Performance
- Aufstieg zum Apache Toplevel Project

Schwächen

- Installation auf aktuellen Cloudera bzw. Hortonworks
- Dokumentation

© 2014 Ventum Consulting

SQL Frameworks

Tajo – SQL Unterstützung

SQL Unterstützung

- Basis SQL
- Union (all)
- Group by

Nicht unterstützt

- Analytische Funktionen
- Subselect nur mit Alias
- substr
- Minus
- Intersect

© 2014 Ventum Consulting

SQL Frameworks
Impala

Überblick

- MPP SQL engine
- Nicht Map-Reduce basiert
- Für interaktive SQL-Abfragen
- Unterstützt zahlreiche Dateiformate
- Verwendet Hive Metadata Store
- Daten im HDFS oder Hbase gespeichert

© 2014 Ventum Consulting

SQL Frameworks
Impala Architecture

Architektur

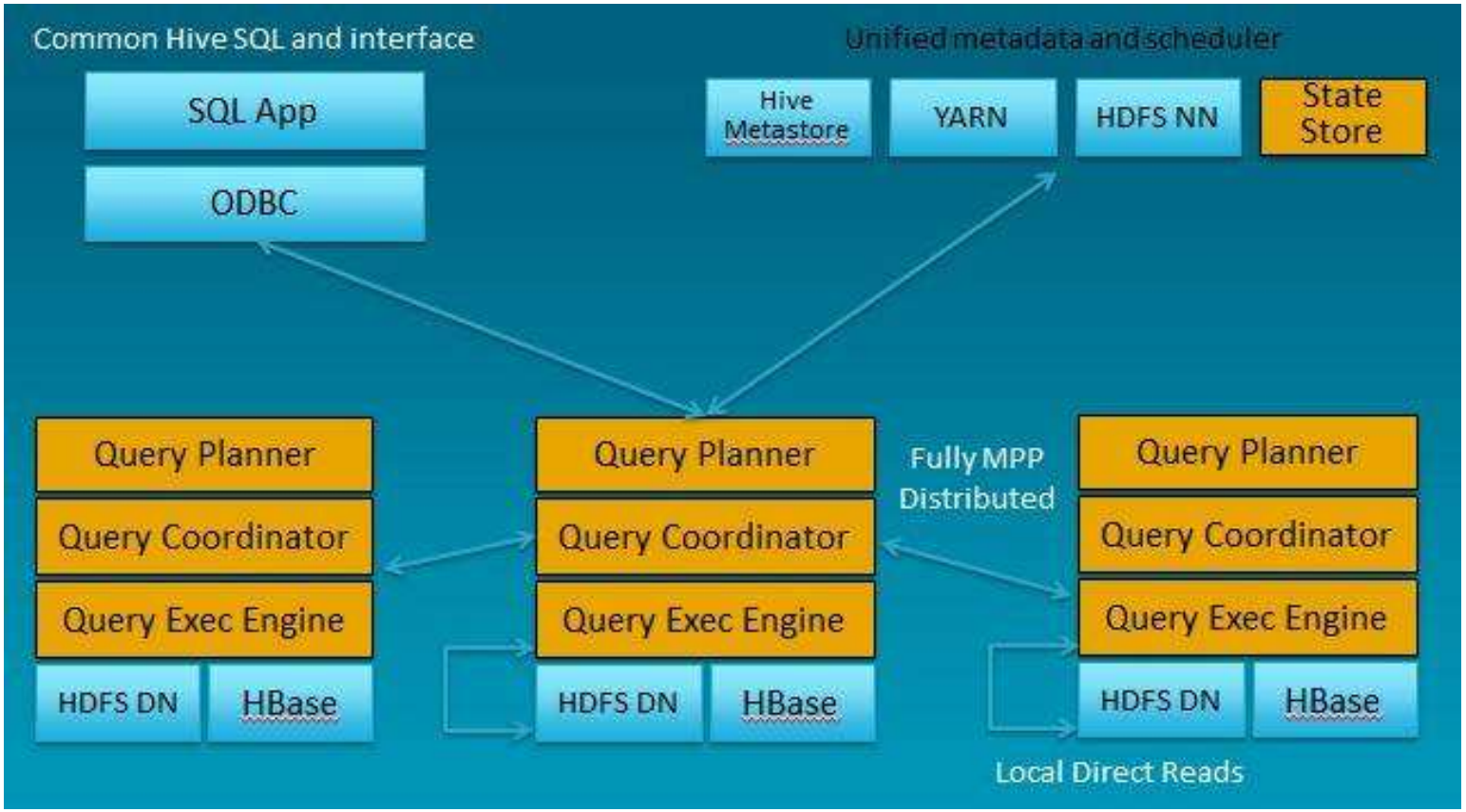


Image from Cloudera.com

© 2014 Ventum Consulting

SQL Frameworks

Impala – Stärken und Schwächen

Stärken

- Performance
- Nutzung von Hive-Komponenten
- Unterstützung zahlreicher Dateiformate, Hbase und Hive

Schwächen

- RAM-Anforderungen 128 GB empfohlen
- Abbrüche durch Out-of-Memory
- SQL Unterstützung geringer als bei Hive
- Keine User Defined Functions (UDF)

© 2014 Ventum Consulting

SQL Frameworks

Impala– SQL Unterstützung

SQL Unterstützung

- Basis SQL
- UNION

Nicht unterstützt

- Analytische Funktionen
- Group by ohne Limit
- Rollup
- Subselects ohne Alias
- MINUS
- INTERSECT

© 2014 Ventum Consulting

SQL Frameworks

Shark

Überblick

- Fully distributed SQL processing
- Nicht Map-Reduce basiert
- In-Memory oder disk basiert
- Unterstützt zahlreiche Dateiformate
- Verwendet Hive Metadata Store
- Verwendet HIVE Frontend
- Spaltenbasierte Tabellen in-Memory
- Setzt auf Apache Spark auf

© 2014 Ventum Consulting

ventum

Architektur

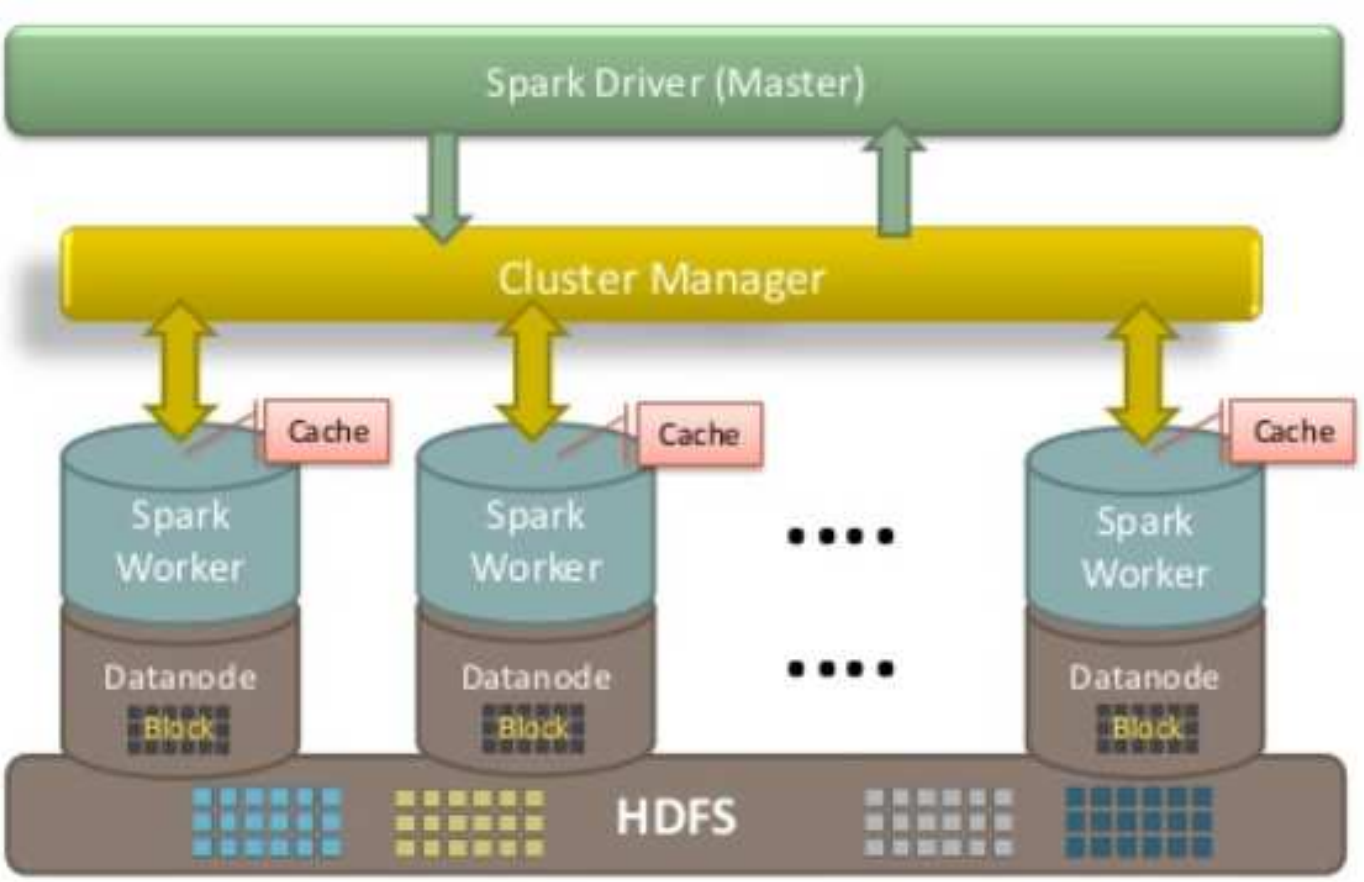


Image from apache.org

SQL Frameworks
Shark Architecture

Architektur

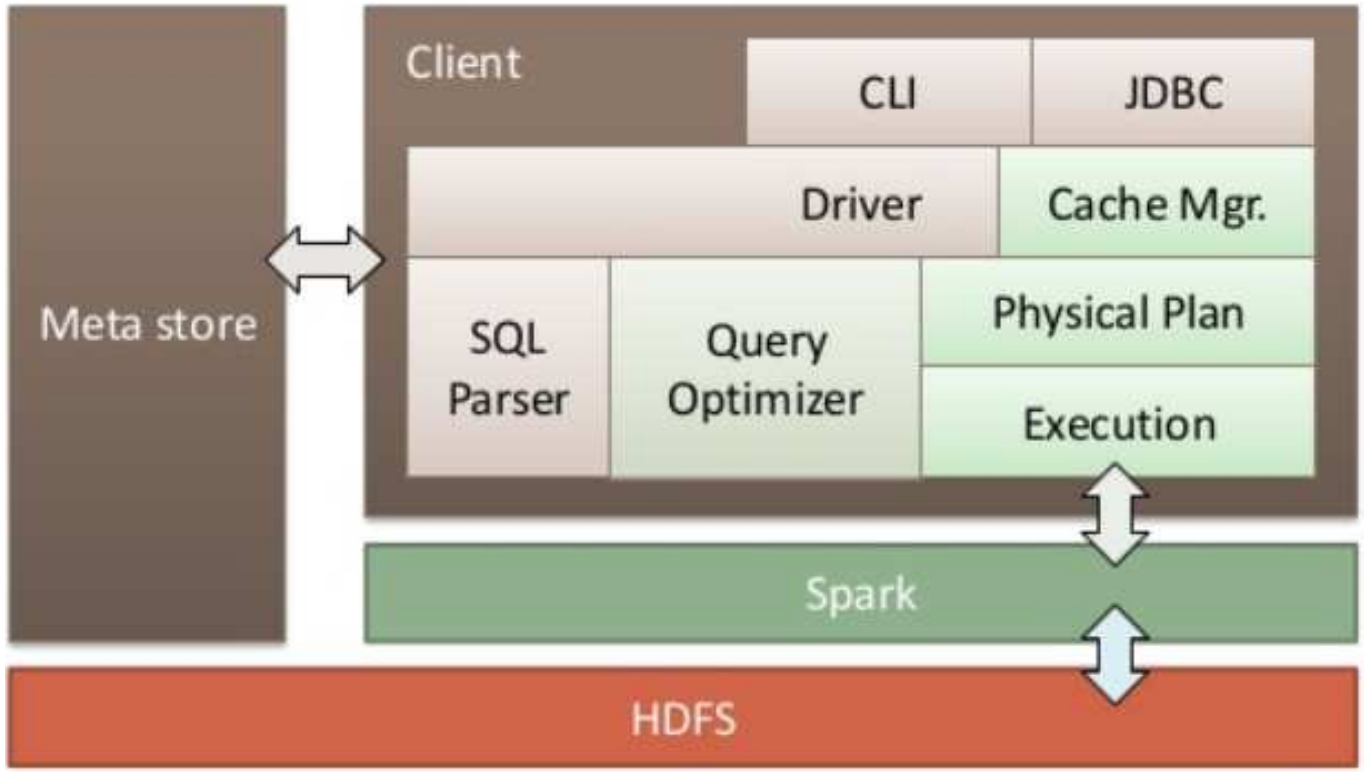


Image from apache.org

© 2014 Ventum Consulting

SQL Frameworks

Shark – Stärken und Schwächen

Stärken

- Performance
- Abbruchtolerant
- Nutzung von Hive-Komponenten
- Unterstützung zahlreicher Dateiformate und Hive

Schwächen

- Benötigt passende Hive Version
- Benötigt Spark
- Memorybedarf für hohe Performanceanforderungen

© 2014 Ventum Consulting

SQL Frameworks

Shark – SQL Unterstützung

Unterstützung

- Analytische Funktionen
- Subselects
- Group by
- Rollup
- UNION ALL

Nicht unterstützt:

- UNION
- MINUS
- INTERSECT

© 2014 Ventum Consulting

Agenda

1 Einführung

2 Hive, Impala, Tajo and Co

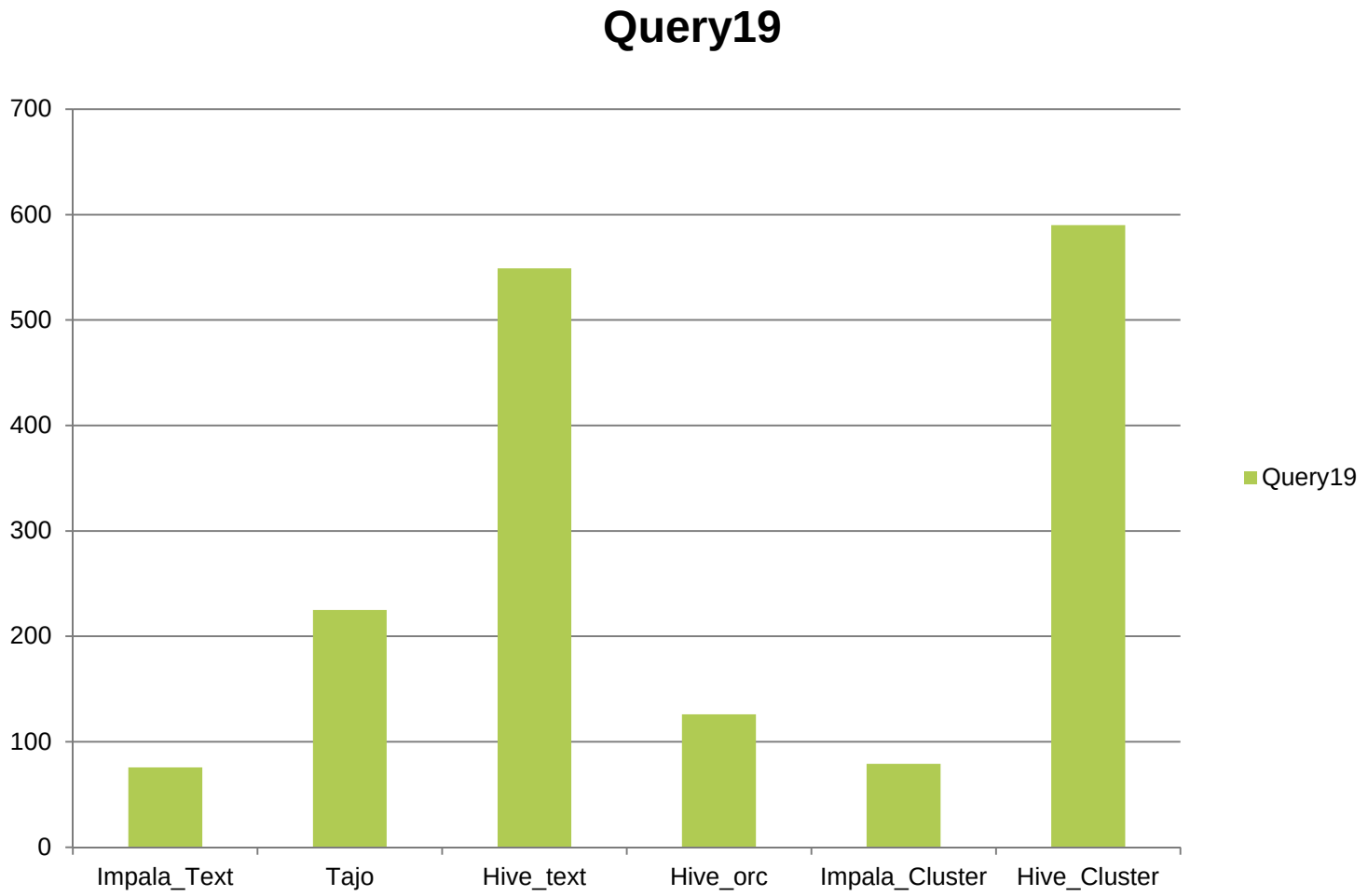
3 Performance

4 Ausblick

© 2013 Ventum Consulting

Performance

Deutliche Unterschiede zwischen den Frameworks



Agenda

1	Einführung
2	Hive, Impala, Tajo and Co
3	Performance
4	Ausblick

© 2013 Ventum Consulting

Fazit
SQL Frameworks demokratisieren den Zugriff

Fazit

- Alle vorgestellten Frameworks erleichtern den Zugriff auf Big Data
- Hive Metastore als übergreifendes Element
- Tradeoff zwischen Hardwareanforderungen und Performance
- Die Entwicklung ist sehr dynamisch
- Unterschiedliche Reifegrade

© 2014 Ventum Consulting

ventum

Consulting

Beratung

Results,
no Excuses.

Lösungen

Grown from
Experience.

Ventum Consulting

Office München

Infanteriestr. 11a
D-80797 München

Telefon +49 89 1222 1964 2
Fax +49 89 1222 1964 25

Office Wien

Ernst Melchior Gasse 24
A-1020 Wien

Telefon +43 (1) 535 34 22 - 0
Fax +43 (1) 535 34 22 – 99

www.ventum-consulting.com

© 2014 Ventum Consulting