FRANKFURT UNIVERSITY OF APPLIED SCIENCES

DATENBANKSYSTEME

5B | APACHE HADOOP UND SPARK - ÜBERBLICK

VERSION 2023

Inhalt, Lernziele, Fragen

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Inhalt

- Technologien im Datenmanagement
- Der Hadoop Technology Stack
- Map-Reduce
- Einsatzfelder von Hadoop
- Vergleich mit Spark

Lernziel

Sie sind vertraut mit dem Hadoop-Ökosystem und verstehen in groben Zügen den Map-Reduce-Algorithmus.

Fragen

? Für welche Lösungen eignet sich Hadoop, bzw. die Hadoop-Produkte

? Welche Probleme haben diese heute

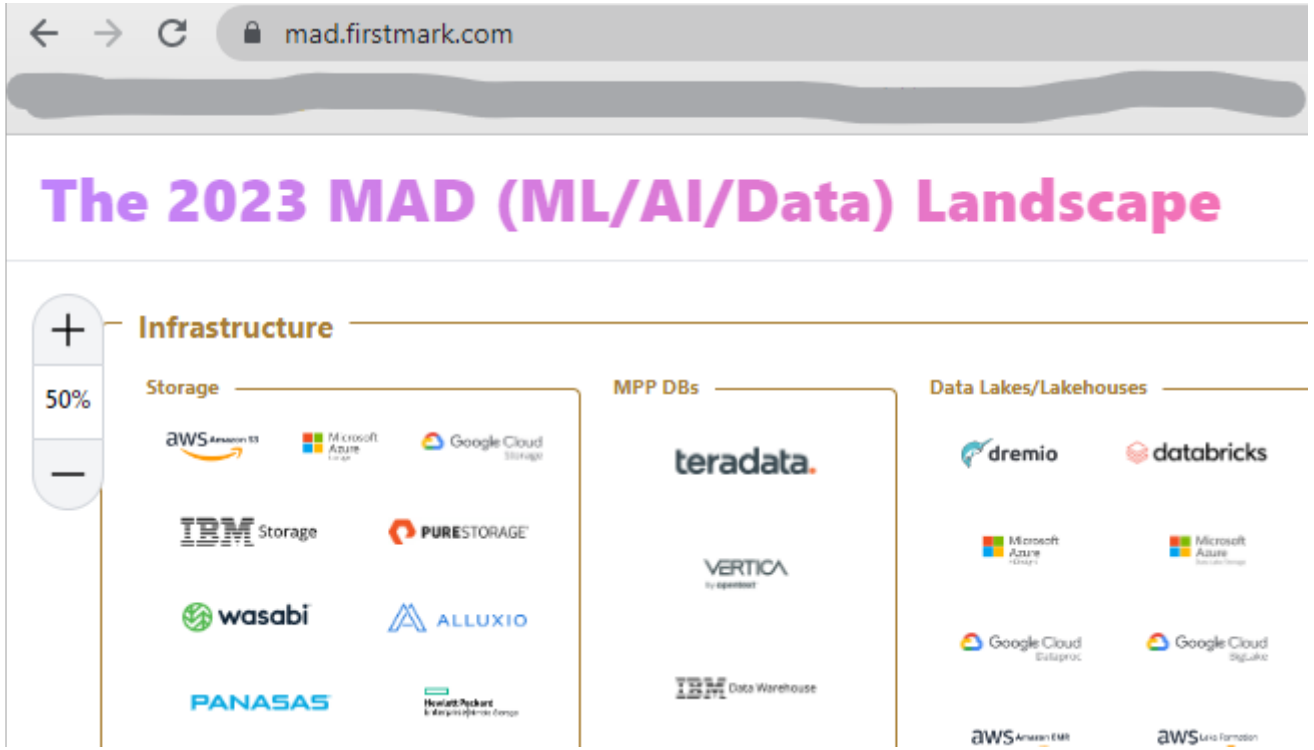
? Wie sollten Sie eine Datenverwaltungs-Architektur aufsetzen

Intro - Data Landscape

- Die im Bereich Data Management verwendeten Systeme sind von Zahl und inhaltlicher Ausrichtung kaum mehr zu überblicken
- Eine Übersicht finden Sie z.B. unter:

https://mad.firstmark.com/

- Häufig ist Hadoop eine der Basistechnologien der kommerziellen Produkte



FRANKFURT UNIVERSITY OF APPLIED SCIENCES

HADOOP

Überblick

- Hadoop Common
  - ist die Basis für weitere Produkte
  - grundlegende Dienste und Prozesse
  - Abstraktionsschicht über dem Betriebssystem und Dateisystem
  - Java-Pakete zum Start der Plattform
  - Zur Verwaltung sind grundlegende Kenntnisse von Betriebssystemen notwendig

- HDFS – Hadoop Distributed File System
  - Ein Dateisystem mit hohen Datenverarbeitungsraten
  - Kann auf Server-Hardware und einfacher Hardware (Laptop) ausgeführt werden, z.B. auch in einer VM
  - Daten können über tausende Server verteilt werden
  - Ausfallsicherheit bei Hardware-Problemen

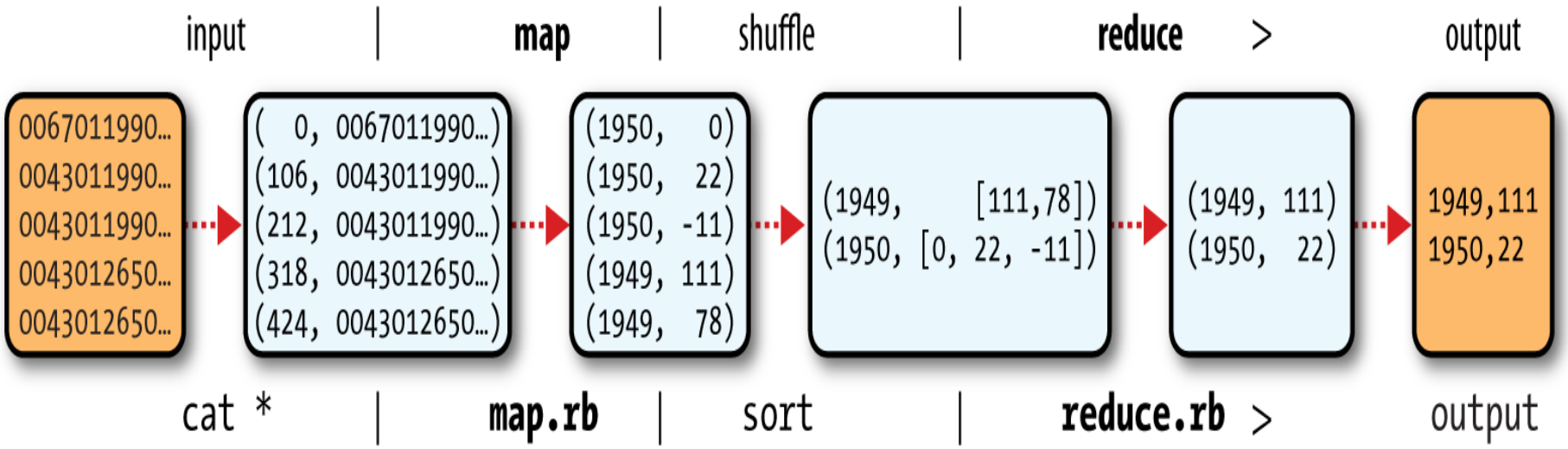
Überblick

- Hadoop MapReduce
  - Algorithmus für parallele Datenverarbeitung und Verdichtung der Daten in "managebare" Portitionen, die zu Analysezwecken benötigt werden → vgl. DWH-Folien
  - Programmierkomponente von Hadoop
  - Verarbeitung großer Datenmengen
  - Batch-Processing möglich
  - Verarbeitung der Daten aus dem HDFS über parallele, aufgeteilte Workloads

- Hadoop YARN
  - Yet Another Resource Negotiator
  - Verteilt Ressourcen (CPU, Speicher, …) zwischen mehreren Prozessen, z.B. MapReduce-Prozessen und/oder Frameworks wie MapReduce, Impala und Spark

MapReduce

- Beispiel-Verarbeitung mittels MapReduce



Quelle: White, T. (2015): Hadoop – The definitive Guide, 4th ed., O'Reilly

Überblick

- Hadoop Zookeeper
  - wird für das technische / organisatorische Management von Hadoop benötigt
  - Koordination der Bestandteile eines Hadoop-Systems
  - verwaltet "Konfigurationen", Namen(sräume) und Gruppen
  - koordiniert Systembestandteile

- Apache Hive
  - Data Warehouse auf Basis von Hadoop
  - HiveQL als Abfragesprache stark an SQL angelehnt

- Apache Spark
  - schnelle In-Memory-Data Processing Engine
  - Unterstützung von Programmiersprachen (Java, Python, Scala, R) und SQL
  - Verarbeitung von "DataFrames" möglich
  - Komponenten für Machine Learning (Mllib) und Graphdaten-Verarbeitung (GraphX)

Layers der Hadoop Architecture

- Verteilte Speicherung
- Cluster Resource Management
- Processing Frameworks
- Application Programming Interfaces

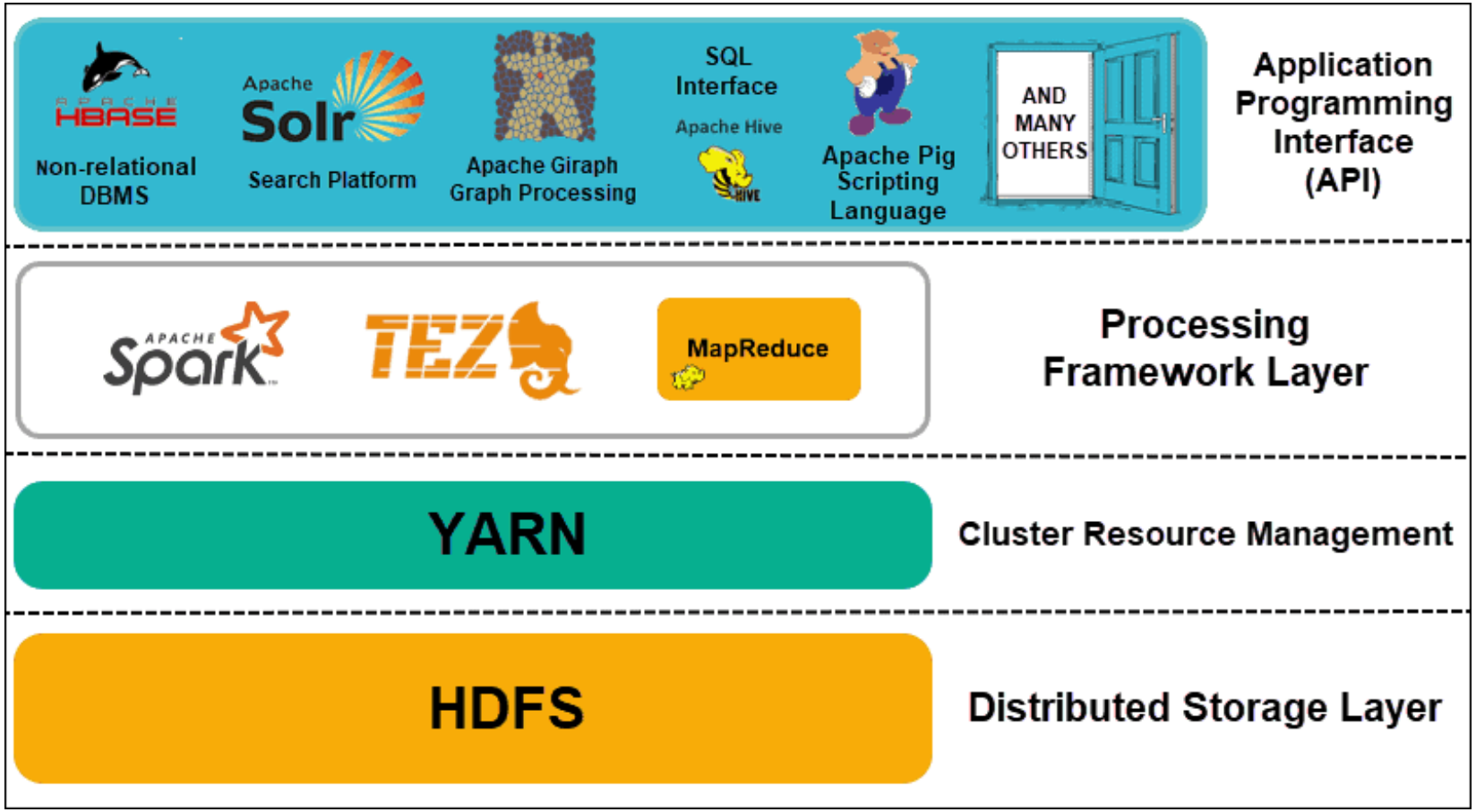
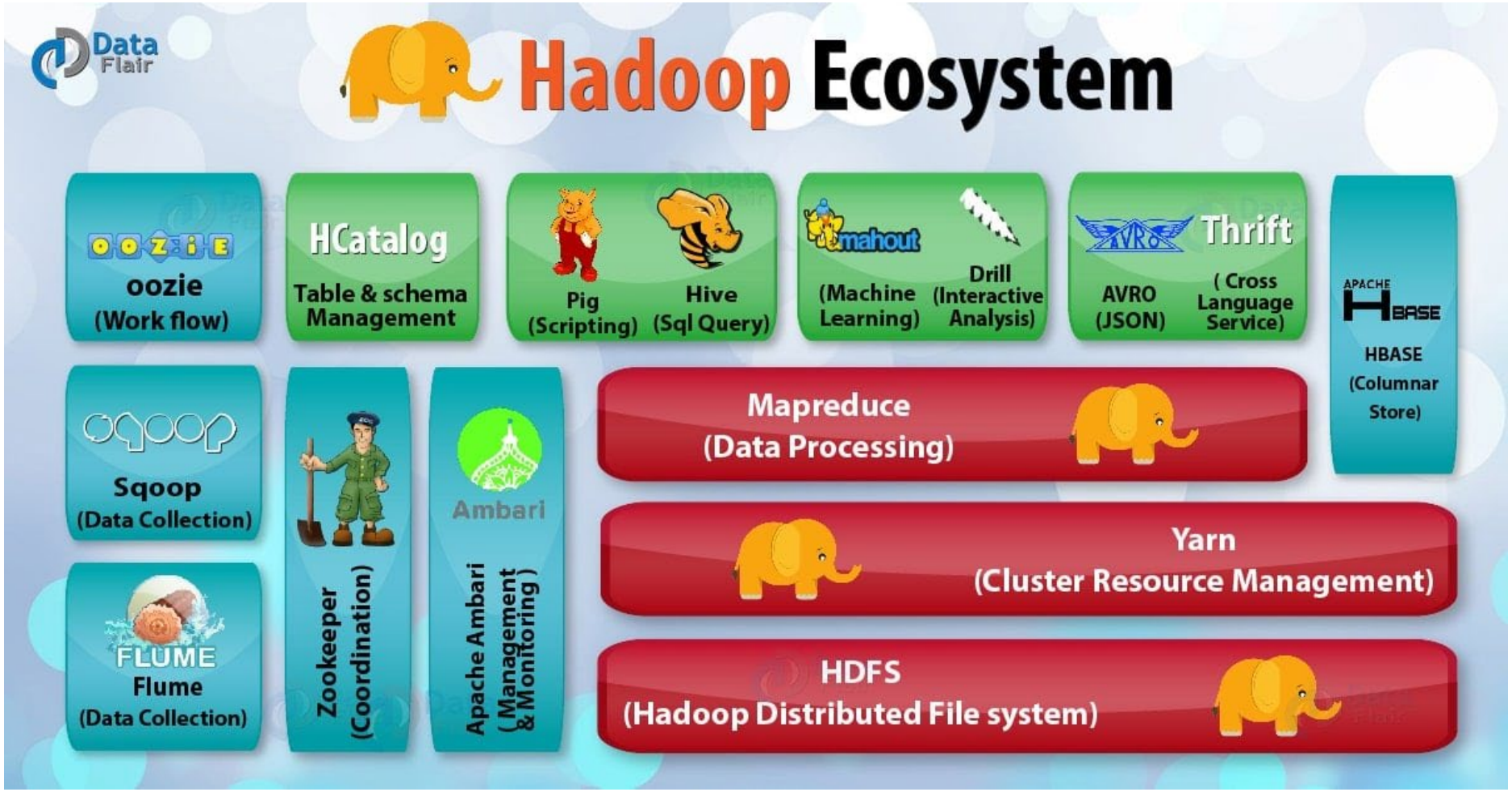


Bild-Quelle: https://phoenixnap.com/kb/wp-content/uploads/2021/04/hadoop-ecosystem-layers.png

Das Hadoop-Ökosystem – weitere Komponenten



Quelle: https://data-flair.training/blogs/hadoop-ecosystem-components/

Aufgabe (25-30 Minuten)

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

1. Ergänzen Sie Beschreibungen für die noch nicht vorgestellten Technologiekomponenten auf Folie 10
   - HBase
   - Pig und
   - Tez

2. Ergänzen Sie die Beschreibungen um eine der folgenden Technologien:
   - Sqoob
   - Flume

3. Beschreiben Sie den MapReduce-Algorithmus
   - einmal hinsichtlich der technischen Verarbeitung (s. unter "Map Phase" auf https://phoenixnap.com/kb/apache-hadoop-architecture-explained)
   - einmal anhand eines "fachlichen" Beispiels

Einsatzfelder

- Hadoop
  - kann auf einfacher Hardware in einem Cluster aufgesetzt werden
  - unterstützt das redundante Speichern von Daten im Cluster auf einem Filesystem
  - kann mit seinen Komponenten bekannte Konzepte unterstützten: Datenbanken, SQL, DWH, JSON-Storage..
  - ist dann geeignet, wenn Daten schnell geschrieben, gelesen und verarbeitet werden müssen

- Nachteile
  - die Installation und Wartung erfordert tiefgreifende technologische Kenntnisse
  - daher häufig Installation von Apache Spark ohne den Hadoop-Stack (HDFS…)
  - Hadoop ist nicht "paketiert" → häufig Einkauf von Hadoop-Lösungen erforderlich (Cloudera u.a.)

Vergleich zwischen Hadoop und Spark

Category	Hadoop	Spark
Performance	Langsamer, Disk Storage	Schnell, in-memory Performance mit reduzierten Disk Operationen
Kosten	Open-source, relativ günstig. Consumer Hardware. Experten am Markt verfügbar	Open-source, in-memory-Verarbeitung → relative hohe Kosten.
Data Processing	Eher batch, MapReduce	Iterative und Live-Stream-Verarbeitung, z.B. zur Datenanalyse
Ausfallsicherheit	Replikation → hohe Sicherheit	Kann mittels RDD block creation process datasets rekonstruieren im Fehlerfall
Skalierbarkeit	gegeben	gegeben
Sicherheit	Sehr sicher, LDAP, ACLs, …	Nicht sicher, in Verbindung mit Hadoop kann Security Level erreicht warden.
Sprachen	Java, Python für MapReduce, eingeschränkt	sehr viele Sprachen und interaktiver Shell-Modus
Machine Learning	langsamer, Datenfragment-Größe kann zu Bottlenecks führen	Sehr schnell, unterstützt durch eigene Bibliothek
Scheduling und Resource Management	Extern, durch YARN oder Oozie für Workflows	Eingebaut

Übersetzt und angepast, Quelle: https://phoenixnap.com/kb/hadoop-vs-spark

Aufgabe (10 Minuten)

1. Forschen Sie im Internet, welche Technologie sich hinter Data Bricks verbirgt und beschreiben Sie diese kurz.

2. Welche Vorteile und Nachteile würden für Sie im Unternehmen entstehen, wenn Sie Data Bricks für die Verarbeitung von Massendaten verwenden?

Lernziel / Fragen

Sie haben einen Überblick über das Apache Hadoop-Ökosystem und können grob dessen Komponenten beschreiben. Mit Spark haben Sie eine Alternative zu Hadoop kennen gelernt.

- ✓ NoSQL-Systeme eignen sich zur Verwaltung großer Datenmengen in Netzwerkstrukturen.
- ✓ Die Systeme bedienen unterschiedliche Bedarfe. Die Auswahl eines Systems muss sich (wie immer) am Anwendungsfall orientieren.
- ✓ Die Auswahl der Systeme muss auch nach dem CAP-Theorem erfolgen, da nicht alle Bedingungen des CAP zugleich erfüllt sein können.
- ✓ …. relationale Datenbanken sind weiterhin für einige Anwendungsfälle notwendig.

Literatur

- Antony, B. et al. (2016): Professional Hadoop, Wrox Press
- White, T. (2015): Hadoop – The definitive Guide, 4th ed., O'Reilly