THM TECHNISCHE HOCHSCHULE MITTELHESSEN | CAMPUS GIESSEN | MNI Mathematik, Naturwissenschaften und Informatik

Apache Hadoop

Hauptseminar "Cloud-Plattformen und Big Data"
Dozent Steffen Rupp

von

René Gentzen

rene.gentzen@mni.thm.de

im WS22/23

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Listings

1 Hadoop Fundamentals

1.1 Anforderungen von Big Data

1.1.1 Die Limits von Vertical Scalability

1.1.2 Die Vorteile und Challenges von Horizontal Scalability

1.2 Hadoop Historie

1.2.1 GFS und MapReduce

1.3 Hadoop 1.0 Aufbau

1.3.1 HDFS in depth theoretisch

1.3.2 MapReduce in depth theoretisch

1.4 Hadoop 2.x mit YARN

1.4.1 Neue Möglichkeiten mit YARN

1.4.2 Neuer Aufbau, neue Verpflichtungen

1.4.3 Code einer YARN Application

1.5 Hadoop 3.x anreißen

2 Das Hadoop Ecosystem

2.1 Storage

2.1.1 HDFS

2.1.2 HBase

2.2 Management / Configuration

2.2.1 YARN

2.2.2 Oozie

2.2.3 ZooKeper

2.2.4 Ambari

2.3 Datentransfer

2.3.1 Sqoop

2.3.2 Kafka

2.3.3 AVRO

2.4 Processing

2.4.1 MapReduce

2.4.2 Pig

2.4.3 Hive

2.4.4 Flume

2.4.5 Mahout

2.4.6 Spark

2.4.7 Solr

3 Hadoop Operations

3.1 Hadoop Setup

3.1.1 Single Node Setup

3.1.2 Pseudo-Distributed Setup

3.1.3 Fully-distributed Cluster

3.1.4 Docker Images

3.1.5 VM Distributionen

3.1.6 Hadoop in der Cloud

HDFS oder Cloud FS

3.2 Praxis

3.2.1 Start eines Single Node Clusters lokal

3.2.2 Erste Übung zum Umgang mit dem HDFS

3.2.3 MapReduce Workflow: Erstellung eines MapReduce Jobs, Kopieren auf den Name Node und Ausführung

3.2.4 YARN Application auf einem Cluster in der Cloud laufen lassen

4 ETL mit Pig

Auch wenn es vermehrt von Spark verdrängt wird

4.1 Anwendungsfälle

Welche neuen Dinge ermöglicht dieses Tool Eine Zeile Pig Latin entspricht vielen Zeilen MapReduce

4.2 Architektur

4.3 Pig Latin

4.4 Praxis

4.4.1 Hinzufügen zum Cluster

4.4.2 Anwendung auf dem Cluster

5 Data Ingestion

5.1 Sqoop

5.2 Flume

6 Datawarehousing mit Hive

6.1 Anwendungsfälle

Welche neuen Dinge ermöglicht dieses Tool

6.1.1 Unterschiede zu Pig

6.2 Architektur

6.3 Interaktion

6.3.1 HiveQL

6.3.2 CLI

6.3.3 Java API

6.4 Praxis

6.4.1 Hinzufügen zum Cluster

6.4.2 Einrichtung einer Datenbank

6.4.3 Einlesen von Daten im CLI

6.4.4 Einlesen von Daten mit Sqoop

6.4.5 Absetzen einer Query

7 NoSQL mit HBase

7.1 Anwendungsfälle

Welche neuen Dinge ermöglicht dieses Tool

7.1.1 CAP-Theorem

7.1.2 ACID und BASE

7.2 Architektur

7.3 Interaktion

7.3.1 HBase Shell

7.3.2 Java API

7.4 Praxis

7.4.1 Hinzufügen zum Cluster

7.4.2 Einrichtung einer Datenbank

7.4.3 Einlesen von Daten

7.4.4 Datenmigration aus einem RDBMS

7.4.5 Absetzen einer Query

8 Streaming mit Kafka

8.1 Anwendungsfälle

Welche neuen Dinge ermöglicht dieses Tool Ersetzt durch Spark Streaming

8.2 Architektur

8.3 Interaktion

8.3.1 Who knows

8.4 Praxis

8.4.1 Hinzufügen zum Cluster

8.4.2 Maybe, vielleicht kann man ja was zeigen

9 Hadoop heute

9.1 Aktuelle Anwendungsbeispiele zu Hadoop

9.1.1 AirBnB

9.2 Apache Spark als Gold Standard

9.2.1 Kann eh alles besser