Big Data Processing mit Apache Spark

Sascha P. Lorenz, Email: sascha.lorenz@contexagon.com
Hochschule Emden-Leer, Fachbereich Technik

Ausgangslage:

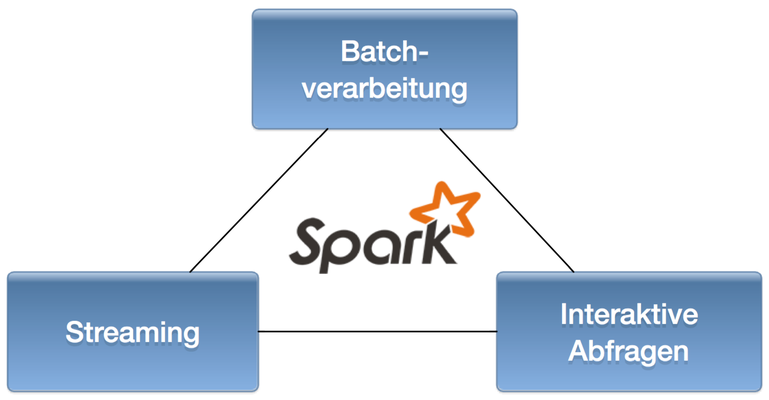
Das Buzzword „Big Data“ begegnet uns mittlerweile überall. Daten aller Art werden in astronomischer Geschwindigkeit generiert und es ist kein Ende des Wachstums in Sicht. Doch wieso sind diese Daten für uns wichtig? Daten sind wertvoll aufgrund der Entscheidungen, die sie ermöglichen.

Als Yahoo! im Jahr 2009 den Quellcode ihrer Big-Data-Anwendung „Hadoop“ freigegeben hat, hatte dies zur Folge, dass Big Data nun auch für andere Industriezweige und Anwendungsgebiete verfügbar ist und massive Vorteile bringt.

Hadoop ist mittlerweile mit seinem Map/Reduce-Paradigma die Standardtechnologie für Big Data. Da die Zwischenergebnisse hier auf Festplatten persistiert werden, ergeben sich für einige Anwendungsbereiche Unzulänglichkeiten in der Ausführung. Deshalb wurde rund um Hadoop ein ganzes Ökosystem von Anwendungen entwickelt. Dennoch ist die Nutzung in einigen Kontexten umständlich und für einige Anwendungen (z.B. für den schnellen Datenaustausch bei Parallelverarbeitung) gar unbrauchbar.

Diese Herausforderungen löst Apache Spark. Dieses Framework ist eine hochgeschwindigkeits Cluster-Plattform mit einem einheitlichen Ansatz für Batch, Streaming und interaktive Nutzung.

Ein einzelnes Framework für alle Aufgaben



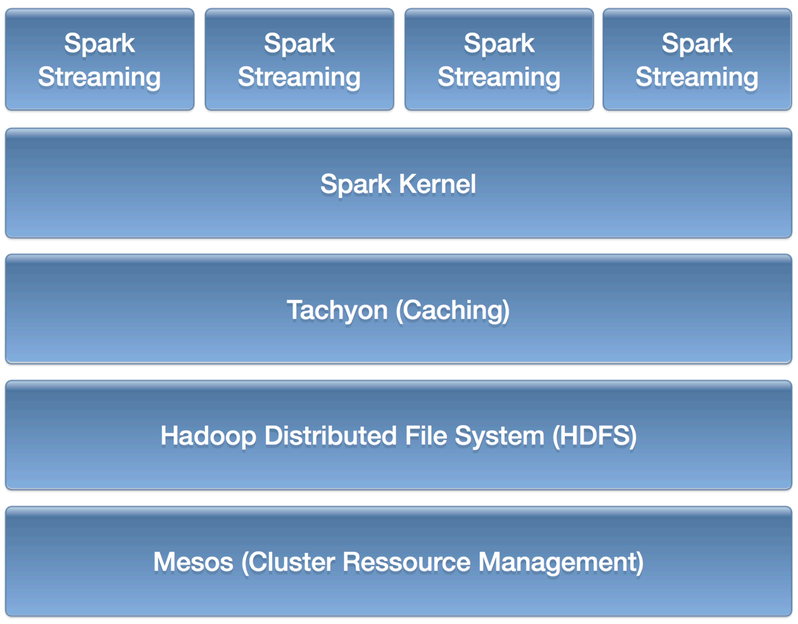
Zielsetzung:

Diese Masterarbeit soll einen Überblick und Erkenntnisse über das gesamte Ökosystem rund um Apache Spark liefern – den sogenannten Berkeley Data Analytics Stack. Dieser besteht aus verschiedenen Schichten, begonnen mit dem Dateisystem, einer Caching-Schicht, dem Spark-Kern und verschiedener Anwendungsbibliotheken und APIs für Machine-Learning, Graphenanwendungen, Streaming und Datenbankabfragen. Innerhalb dieser Arbeit werden verschiedene Ansätze für die Nutzung geprüft, die Anwendung der APIs gezeigt und Vergleiche zu alternativen Implementierungen gezogen.

Vorgehen:

Zunächst wurden verschiedene Cluster für die Versuchsanordnung definiert. Zum Einsatz kamen sowohl lokale Installationen, als auch ein Cluster mit 16 Prozessoren und einigen Terrabyte Gesamthauptspeicher. Auf diesen Maschinen wurden verschiedene Konfigurationen von Spark getestet.

Der Berkeley Data Analytics Stack



Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.

At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.

At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Ergebnisse:

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.

At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.

Ausblick:

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.

At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.

At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.

At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.

At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.