ToiletPaper #86

SMACK Stack

Autor: Andreas Scharf / Senior Software Architect / Business Division New Business

Definition SMACK Stack:

Ein SMACK Stack ist eine Sammlung von Technologien zum Aufbau einer robusten und verteilten Datenverarbeitungsarchitektur, die eine Echtzeit-Datenanalyse und eine schnelle Bereitstellung ermöglicht. Das Akronym **SMACK** steht dabei für die Spark-Engine, den Mesos-Manager, das Akka-Toolkit und die -Runtime, die Cassandra-Datenbank und den Kafka-Message-Broker. Alle Komponenten außer Akka sind Apache-Projekte. Die Software ist Open Source und hat sich in der Praxis bewährt. Durch den Einsatz dieser lose gekoppelten Toolchain von Technologien ist es möglich, eine private Cloud-Plattform zu schaffen, die große Datenmengen verarbeitet und gleichzeitig die Bindung von Cloud-Anbietern verhindert.

✓ Spark

Spark ist ein Engine für großangelegte Datenverarbeitung und macht es einfach, parallele Anwendungen oder Batch-Jobs zu erstellen. Ein Vorteil gegenüber Hadoop MapReduce ist eine überlegene Leistung. Als Benutzer hat man die Möglichkeit, strukturierte Daten mit Spark SQL abzufragen. Darüber hinaus befasst sich Spark Streaming mit Echtzeit-Anwendungsfällen: Eingehende Daten werden in Mikrochargen zerlegt und separat verarbeitet.

✓ Mesos

Mesos ist ein Scheduling-Framework zur Verwaltung von Clustern. Es stellt Ressourcen für Anwendungen, Dienste und Jobs zur Verfügung und abstrahiert die zugrundeliegende Hardware. Die Arbeitslast wird auf dem Cluster verteilt. Das verteilte Betriebssystem DC/OS, das auf Mesos aufbaut, vereinfacht die Bereitstellung und Skalierung von containerisierten Anwendungen.

✓ Akka

Akka ist eine Implementierung des Aktorenmodells und ermöglicht die Erstellung von nachrichtengesteuerten Anwendungen in Scala oder Java. Das Toolkit ist von der Sprache Erlang inspiriert und seine Fehlertoleranz beruht auf einer aktorbasierten Parallelität. Aktoren können den lokalen Zustand ändern, werden ihn aber nicht sichtbar machen. Stattdessen verwenden sie asynchrone Nachrichten, um mit anderen Aktoren zu interagieren.

✓ Cassandra

Cassandra ist eine NoSQL-Datenbank. Es handelt sich um einen Wide Column Store, der als zweidimensionaler Key-Value Store betrachtet werden kann. Die Abfrageoperationen von Cassandra sind begrenzt, um Leistung und lineare, horizontale Skalierung zu gewährleisten. Es gibt keinen Single Point of Failure und es ist möglich, einen Cassandra-Cluster über mehrere Rechenzentren hinweg zu verteilen.

✓ Kafka

Kafka ist ein verteiltes Messaging-System, das für niedrige Latenzzeiten und hohe Verfügbarkeit sowie hohen Datendurchlauf bekannt ist. Kafka verwendet intern ein gemeinsames Commit-Log. Kafkas USP ist die Möglichkeit, das Log erneut einzuspielen. Producer können Nachrichten in Publish-Subscribe Warteschlangen veröffentlichen. Kafka partitioniert die Daten innerhalb eines Topics. Partitionen können auf Clusterknoten verteilt werden und Consumer erhalten die gewünschten Nachrichten.