Sandro Lehmann, mimacom ag
Sandro Lehmann ist als Senior Software Engineer bei der mimacom ag tätig. Seine Schwerpunkte sind Projekte in der Entwicklung von Individualsoftware. Seit mehreren Jahren arbeitet er mit diversen Spring-Projekten und Webtechnologien.

Open Source Big Data mit Apache Hadoop

Apache Hadoop, mit seinem schnell wachsenden Ökosystem, geniesst als Lösung für Big Data seit einiger Zeit viel Aufmerksamkeit. Dieser Beitrag zeigt, wie die beiden Projekte Spring for Apache Hadoop und Spring XD den Einstieg in die Welt von Hadoop erleichtern.

Hadoop ist nach wie vor DIE Plattform für «Big Data» und verteiltes Rechnen. Jedoch besitzt Hadoop ein Low-Level-Programmiermodell, welches die Programmierung aufwändig macht. Man kommt häufig nicht darum herum, viel Infrastruktur-Code zu schreiben. Bereits für das Konfigurieren eines einfachen Jobs müssen einige Zeilen Java programmiert werden, ganz abgesehen von der Programmierung der dazugehörenden Funktionen selbst. Deswegen gibt es eine Reihe von Erweiterungen, die auf Hadoop aufbauen, auf einer höheren Abstraktionsebene sind und somit die Komplexität vermindern.

Spring for Apache Hadoop

Aus diesen Gründen wurde auch das Spring-Projekt «Spring for Apache Hadoop» ins Leben gerufen. Das Ziel von «Spring for Apache Hadoop» ist die Vereinfachung der Entwicklung von Hadoop-Applikationen. Es bietet ein bekanntes und konsistentes Programmierungs- und Konfigurationsmodell an. Das Spring-Projekt baut zudem auf existierenden Service-Layer-Abstraktionen auf. Das Spring-Framework und die Erweiterungen Spring Batch, Spring Integration und Spring Data sind integrierte Bestandteile. Die Vorteile liegen auf der Hand - mit all den Komponenten können grosse und komplexe Applikationen gebaut werden.

Spring XD

Für die Validation, die Verarbeitung und den Import (Data-Ingestion) von Daten in Hadoop drängt sich die Verwendung von Spring XD auf, da die Vereinfachung dieser Funktionen eines der Hauptziele des Projekts ist. Spring XD kann als verteilter und erweiterbarer Service für Data-Ingestion, Echtzeitanalyse, Batch-Jobs und Datenexport genutzt werden. Ein sogenannter Stream definiert in der Welt von Spring XD die ereignisgesteuerte Datenaufnahme von einer Source (Datenquelle) zu einem Sink (Datenausgabe), mit einer beliebige Anzahl von Prozessoren (z.B. Filter und Transformatoren) dazwischen. Streams haben eine lange Lebensdauer und müssen manuell gestoppt werden, falls man sie nicht mehr braucht. Die Module (Sourcen, Sinks und Prozessoren) werden mit einem Pipe-Symbol verbunden (analog dem Verbinden von einzelnen Befehlen auf einer Unix-Shell). Das vereinfacht den Einstieg für gewohnte Unix-Benutzer. Die einzelnen Module basieren wiederum auf Spring-Programmen und lassen sich auch gut selbst modifizieren oder neu entwickeln.

Spring ermöglicht die Einsparung eines erheblichen Entwicklungsaufwands von Big-Data-Applikationen basierend auf Apache Hadoop.

Fazit

Die beiden vorgestellten Spring-Projekte sind für gewohnte Anwender von Spring schnell erschliessbar und bieten viele attraktive Funktionalitäten. Spring vereinfacht und beschleunigt die Entwicklung erheblich. Komplexität wird verborgen und der Entwickler kann sich mehr den fachlichen Aufgaben widmen. Mit geringem Aufwand lassen sich bereits vielschichtige Applikationen ableiten, für deren Entwicklung man ohne Spring viel mehr Zeit investieren müsste.