Westfälische Hochschule

Wissen.
Was praktisch zählt.

Big Data Analyse mit Apache Spark

Modul: Masterprojekt Wirtschaftsinformatik

Team: Martha Janka, René Kanzenbach, Nils Nordmann, Carsten Schober, Yakup Öztürk

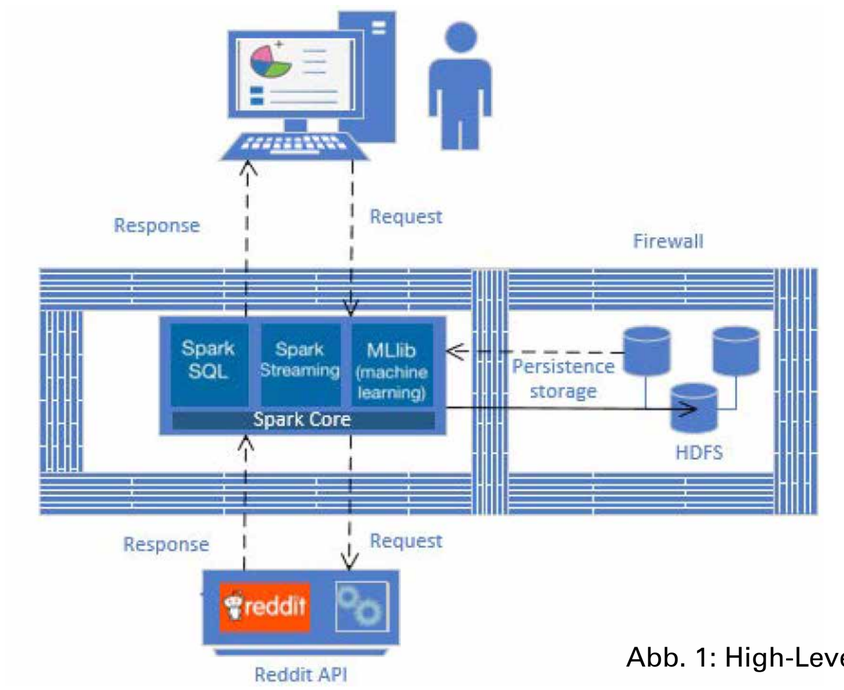

Abb. 1: High-Level-Architektur

Problemstellung / Aufgabenstellung

- Wie intensiv werden ein Unternehmen und dessen Produkte auf Reddit diskutiert?
- Welche Haltung (positiv vs. negativ) haben Reddit-Nutzer gegenüber dem Unternehmen und/oder seiner Produkte?
- Wie entwickelt sich die Diskussionsintensität und die Haltung über einen Zeitraum hinweg?
- Welche Themen stehen im Zusammenhang mit dem Unternehmen sowie den Produkten?

Idee und Konzept

- Massendaten aus dem Reddit müssen mit einem für Big Data geeignetem Tool verarbeitet werden. –> Apache Spark
- Erstellung eines Analyse-Tools, welches Daten zu einem Suchbegriff aus der Reddit-Plattform herauszieht, aufbereitet und mittels Apache Spark verarbeitet.
- Statistische Analyse zur Untersuchung der Diskussionsintensität.
- Sentiment-Analyse zur Veranschaulichung des Stimmungsbildes.
- Semantische Analyse zur Identifikation von thematischen Zusammenhängen.

Technische Umsetzung

Die Aufbereitung und Auswertung der Reddit-Daten besteht aus drei grundsätzlichen Schritten:

1. Daten aus der Reddit-Plattform laden und mittels eines JSON-Parsers extrahieren.
2. Die extrahierten Daten mittels Apache Spark verarbeiten. Hier werden Analyse-Aufgaben in einzelnen Tasks auf den Clustern verteilt bearbeitet und die Ergebnisse abschließend zusammengeführt.

   Spark Streaming-API: Für die Verarbeitung der Reddit-Daten in Echtzeit.

   Spark MLlib: Um sowohl die Stimmung von Nutzern herauszukristallisieren als auch zusammenhängende Themen aufzudecken.
3. Die Eingabe der Suchanfrage sowie Darstellung der Ergebnisse erfolgt über eine mit JavaFX erstellte Benutzungsoberfläche.

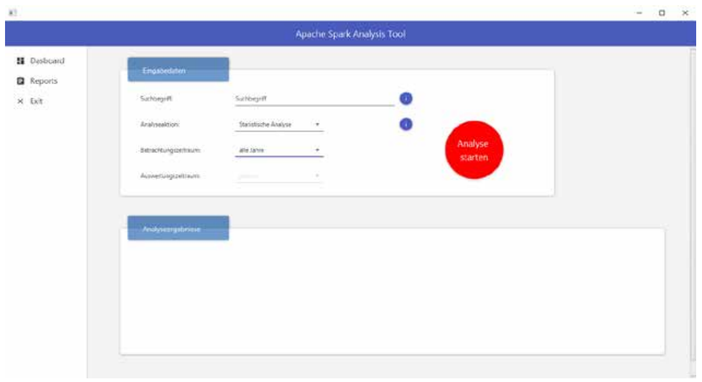

Abb. 2: Eingabeformular des Analyse-Tools

Abb. 3: Auswertungs-Diagramme im Analyse-Tool

Team

Martha.Janka@studmail.w-hs.de
Rene.Kanzenbach@studmail.w-hs.de
Nils.Nordmann@studmail.w-hs.de
Carsten.Schober@studmail.w-hs.de
Yakup.Oeztuerk@studmail.w-hs.de

Betreuung

Prof. Dr. Siegbert Kern siegbert.kern@w-hs.de
Prof. Dr. Henning Ahlf henning.ahlf@w-hs.de
Kolja Dunkel MBA kolja.dunkel@w-hs.de
Fachgebiet Wirtschaftsinformatik

Westfälische Hochschule
Fachbereich Informatik und Kommunikation
Neidenburger Straße 43
45897 Gelsenkirchen
www.w-hs.de