Westfälische Hochschule

Gelsenkirchen Bocholt Recklinghausen

Shoppers-Challenge

Projektdokumentation

---

im Masterstudiengang Verteilte Systeme

vorgelegt von: Daniel Hardes, Dennis Miller,
Fabian Paus, Christian Schlütter,
Lutz Kalkofen, Marvin Weck,
Johannes Döing

Modul: Fortgeschrittene Datenbanktechniken (FDB)

Gutachter: Prof. Dr. Benrhard Convent

Abgabetermin: 21. Mai 2015

Inhaltsverzeichnis

1. Einleitung

Im Zeitalter von Google, Amazon und Facebook ist das Thema „Big-Data“ allgegenwärtig. So auch im Modul Fortgeschrittene Datenbanktechniken im zweiten Semester des Masterstudiengangs Verteilte Systeme. Im Rahmen eines Projektes sollte das Thema unter einer beliebigen Aufgabenstellung selbstständig untersucht und erarbeitet werden. Dabei haben wir uns für das Projekt „Shoppers-Challenge“ von Kaggle, das im nächsten Abschnitt genauer beschrieben wird, entschieden. Bei der Umsetzung haben wir basierend auf dem Hadoop-Framework verschiedene Technologien wie beispielsweise MapReduce-Algorithmen kennengelernt.

1.1. Aufgabenstellung

Ziel dieses Projekts ist es, ein möglichst hohes Rating bei der „Shoppers Challange“ auf kaggle.com zu erreichen.

Die Plattform kaggle.com bietet Wettbewerbe im Bereich „Big-data“ an, wobei hier größtenteils Vorhersagemodelle und Analysen erstellt werden sollen. Firmen und Wissenschaftler können auf dieser Plattform Aufgaben und zugehörige Daten zur Verfügung stellen, die von Statistikern und Data-Minern aus der ganzen Welt bearbeitet werden. Die Lösungen werden dabei nach ihrer Qualität, bezogen auf die Genauigkeit der Vorhersage, in einem Ranking aufgeführt und teilweise sogar prämiert. Damit bietet die Plattform kaggle.com für die Aufgabensteller den enormen Vorteil, dass sie aus vielen Strategien die beste auswählen können und die Bearbeiter gleichzeitig die Qualität ihrer Lösung mit denen der Konkurrenz vergleichen können.

Die „Shoppers Challenge“ beschäftigt sich mit der Frage, ob ein Kunde der einen Gutschein genutzt hat, zu einem „treuen“ Kunden wird und in Zukunft noch weitere Produkte kauft. Zur Lösung des Problems muss ein Vorhersagemodell erstellt werden, das mit Hilfe der vom Aufgabensteller bereitgestellten Daten, die Wahrscheinlichkeit für einen erneuten Kauf für jeden Kunden vorhersagt.

1.2. Projektmanagement

- Phasen - Einarbeitung - Implementierung (iterativ) - Projektabschluss

- Meilensteine - Technologieentscheidung treffen - Ërste Submission Wechsel von Lokal VM nach AWS - "Finale Submission Dokumentation abgeschlossen - Präsentation abgeschlossen

TODO: terminieren

- ggf. verwendete Werkzeuge (Git, Drive)

2. Einarbeitung

- Hadoop (Hortonworks VM) - Map / Reduce

2.1. Datenmodell

2.2. Data-Mining-Verfahren

2.3. Technologieentscheidung

3. Implementierung

3.1. Iteration 1

3.2. Amazon Web Services

3.3. Iteration 2

4. Fazit

A. Anhang 1