Bachelorprojekt für 2010/2011
Fachgebiet Informationssysteme
Prof. Dr. Felix Naumann

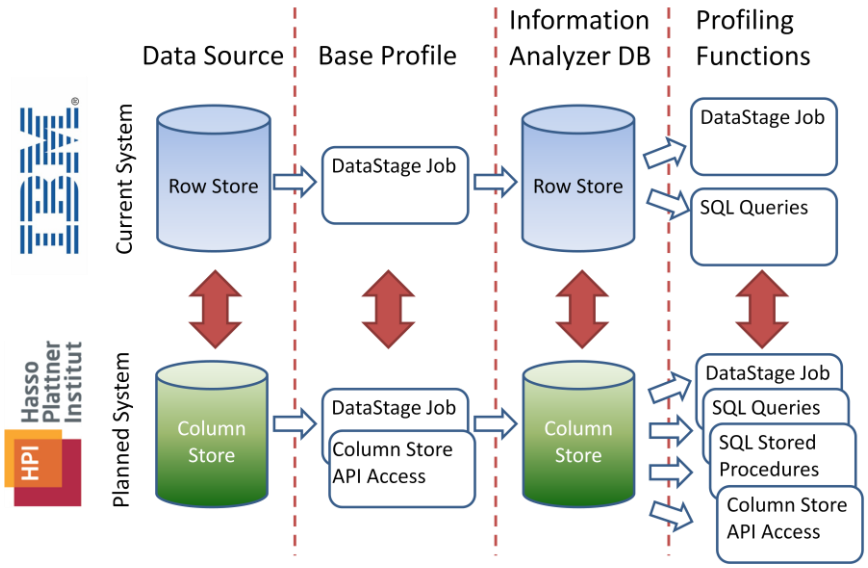
ProCSIA: Profiling Column Stores with IBM's Information Analyzer

Data Profiling ist die systematische Analyse unbekannter Datenbestände um Metadaten zu ermitteln. Diese Metadaten reichen von einfachen Statistiken über Spalten (Anzahl verschiedener Werte) oder einfacher Musteranalyse für Datenwerte (reguläre Ausdrücke, Datentypen) bis zu komplexen Zusammenhängen wie Fremdschlüsselbeziehungen oder funktionalen Abhängigkeiten. Unter der Vielzahl kommerzieller Werkzeuge ist der IBM InfoSphere Information Analyzer („IA") ein prominentes Beispiel.

Data Profiling Werkzeuge basieren in aller Regel auf der klassischen zeilenbasierten Speicherstruktur relationaler Datenbanken. Ein in den letzten Jahren immer beliebter werdendes Speicher-Modell ist jedoch die spaltenorientierte Ablage relationaler Daten in sogenannten *Column Stores*, wie etwa Vertica, C-Store, MonetDB, IBMs ISAO oder SAPs TREX. Da viele der Data Profiling Ergebnisse sich auf Spalten beziehen, liegt es nahe, entsprechende Verfahren speziell für Column Stores zu entwickeln und zu prüfen, unter welchen Umständen und inwieweit diese Datenstruktur effizienteres Data Profiling ermöglicht.

Projektbeschreibung

Im Information Analyzer wird von einem traditionellen relationalen Datenbanksystem (Row Store) als Datenquelle ausgegangen. Vorliegende Daten werden mittels DataStage, eine IBM Parallelisierungsengine, exportiert und analysiert. Die entstehenden Zwischenergebnisse, das *base profile*, werden ebenfalls in einem Row Store abgelegt, wo sie wiederum mittels DataStage und SQL weiterverarbeitet werden.



Das ProCSIA Projekt hat mehrere Ziele (siehe Abbildung):

- Untersuchen von bestehenden Column Stores auf Eignung für den Einsatz für den Information Analyzer (IA).
- Ein oder mehrere Column Stores nach Wahl als Grundlage für den IA zur Verfügung stellen.

- Implementierung/Anpassung von Profiling-Verfahren des IAs unter Verwendung von Column Stores in verschiedenen Varianten (DataStage jobs, SQL-Anfragen, stored procedures, direkter Zugriff auf Column Store API).
- Entwicklung effizienter Verfahren auf Column Stores für neue Profiling –Aufgaben.
- Strategien zur adaptiven Verwendung von Daten aus Row- bzw. Column Stores.
- Entwicklung von Testdaten.

Rahmenbedingungen

- Das Projekt wird vom Hasso Plattner Institut, vertreten durch das Fachgebiet Informationssysteme, in **Zusammenarbeit mit IBM Deutschland** durchgeführt.
- Die Teilnehmerzahl ist auf 8 Mitglieder beschränkt.
- Voraussichtlicher Projektbeginn: 11.10.2010

Für weiterführende Informationen stehen Prof. Dr. Felix Naumann und Christoph Böhm zur Verfügung. Eine Terminabsprache ist über das Sekretariat von Prof. Naumann möglich: office-naumann@hpi.uni-potsdam.de

Vergangene Bachelorprojekte:
http://www.hpi.uni-potsdam.de/naumann/teaching/bachelorprojekte.html

IBM Deutschland GmbH: http://www.ibm.com/de/entwicklung/

Information Analyzer: http://www.ibm.com/software/data/infosphere/information-analyzer/