Expose

für die Diplomarbeit

Implementierung des TPC-DI Benchmark für Talend Open Studio

Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät
Institut für Informatik

Einleitung

Was sind Daten? Was ist Datenintegration? Wie kann man die Integration von Daten bewerten? Nach dem Duden werden dem Begriff Daten vier Bedeutungen beigemessen.

„1. *Plural von Datum*

2. *(durch Beobachtungen, Messungen, statistische Erhebungen u.a. gewonnene) [Zahlen]werte, (auf Beobachtungen, Messungen, statistischen Erhebungen u.a. beruhende) Angaben, formulierbare Befunde*

3. *(EDV) elektronisch gespeicherte Zeichen, Angaben, Informationen*

4. *(Mathematik) zur Lösung oder Durchrechnung einer Aufgabe vorgegebene Zahlenwerte, Größen*“[1]

Unternehmen erfassen Billionen von Bytes an Daten über ihre Kunden, ihre Zulieferer und ihren operativen Betrieb. Millionen von Sensoren sind eingebaut in zum Beispiel Mobiltelefonen und Autos. Sowohl jene Telefone als auch Autos messen, erzeugen und bewegen Daten innerhalb von Netzwerken. Hinzu kommen soziale Netzwerke, Privatpersonen und jegliche multimedialen Plattformen, die das exponentielle Datenwachstum befeuern. Ein großer Berg an Daten, ist in allen Sektoren der Wirtschaft verfügbar zum Erfassen, Kommunizieren, Aggregieren, Speichern und Analysieren [2].

Datenintegration befasst sich mit dem Problem der Zusammenführung von Daten, die sich in verschiedenen Quellsystemen befinden sowie der Bereitstellung einer einheitlichen Sicht auf jene Daten [3]. Die Frage der Datenintegration ist eine beständige Herausforderung für Anwendungen die ihre Daten von autonomen und zugleich heterogenen Systemen beziehen. In großen Unternehmen, die eine Vielzahl an Datenquellen besitzen, ist die Integration dieser Daten ein entscheidender Faktor. Ebenso ist dieser Punkt wichtig für den Fortschritt groß angelegter Forschungsprojekte, bei denen Daten unabhängig voneinander durch mehrere Forscher produziert werden. Die Datenintegration ist auch für Regierungseinrichtungen bedeutend, die jeweils ihre eigenen Datenquellen besitzen. Ein weiterer Rahmen ist das World Wide Web. Hier möchte man eine hohe Suchqualität bei einer Suche verteilt auf Millionen von Datenquellen gewährleisten [4], wobei gesetzliche Vorgaben zum Datenschutz und zum Datenaustausch zu beachten sind.

[1] Duden [1]

Der Begriff ETL

Der Begriff ETL ist eine Abkürzung für Extraktion, Transformation, Laden [5]. Der ETL-Prozess spiegelt das Beziehen von Daten aus verschiedenen Quellen, deren Modifikation [6] und Speichern an einem oder mehreren Zielorten wieder [7]. Im Allgemeinen wird unter ETL das Extrahieren von Daten aus Datenquellen, die Ausführung von Funktionen zur Datenintegration auf ihnen und das Laden in ein Data Warehouse verstanden [7]. ETL-Prozesse beinhalten komplexe Arbeitsschritte zur Verarbeitung von Daten, die für die Pflege eines Data Warehouses verantwortlich sind [8]. Für ETL wird oft auch der Begriff Datenintegration (DI) verwendet [9].

Rellevanz von ETL- bzw. DI-Tools

Der multimillionengroße Markt für ETL-Tools zeigt auf, welche praktische Bedeutung ETL-Prozesse haben. Diese werden mit Hilfe von selbstentwickelten Programmen bzw. Datenbankprozeduren aufgebaut oder mittels ETL-Tools. Die meist verbreitete Begründung für die Verwendung von ETL-Tools ist, die damit verbundene Produktivitätssteigerung und die Einfachheit der Pflege von ETL-Prozessen die mit Hilfe solcher Tools erstellt wurden. ETL-Tools sind in der Lage Arbeitsschritte zur Datenintegration vereinfacht zu repräsentieren. Außerdem stellen ETL-Tools Rüstzeug bereit, die eine höhere Geschwindigkeit erreichen lassen [7].

Bedeutung eines DI Benchmarks

Es ist eine kritische Aufgabenstellung eine hoch performante, skalierbare, und einfach wartbare Datenintegrationslösung zu finden. Mit dem beständigen Wachsen von Komplexität, Vielfalt und Masse der Daten wird der Aspekt der Performance von Datenintegrationslösungen zu einer immer wichtiger werdenden Frage. Trotz der Bedeutung von hoch performanten Datenintegrationslösungen gab es bis vor kurzem keinen industriellen Standard zur Messung und zum Vergleich jener Lösungen. Das Transaction Processing Performance Council (TPC) füllte diese Leere mit dem Benchmark TPC-DI für Datenintegration [9].

TPC-DI

Der TPC-DI Benchmark ist ein Performance Benchmark für DI- bzw. ETL-Tools. Durch den mitgelieferten Datengenerator werden Daten in Form von Extrakten generiert und zur Verfügung gestellt. Der Benchmark definiert Arbeitsschritte zur Integration und Vorbereitung der generierten Daten, für die Nutzung im Data Warehouse. Die generierten Daten repräsentieren synthetische, aber realitätsnahe Extrakte aus hypothetischen OLTP-Systemen. Definitionen der Schemas (im Quell- und Zielbereich), Transformationen und deren Implementationsregeln sind angelehnt an moderne Datenintegrationsanforderungen. Als Grundlage für das Umgebungsmodell dient eine Wertpapierhandelsgesellschaft [10].

Ziel der Diplomarbeit

Das Ziel dieser Diplomarbeit ist, den TPC-DI Benchmark entsprechend der Vorgaben in der Spezifikation [10] für das ETL-Tool Talend Open Studio for Big Data 6.3.1 zu implementieren. Dabei ist die Hauptmotivation, die Auswirkung von paralleler Datenverarbeitung im Transformationsprozess zu beobachten. Die Parallelisierung soll mittels Apache Hadoop erreicht werden. Die Transformationsvorgaben selbst werden mit Hilfe von Apache Pig umgesetzt. Somit ist das Kernstück der Staging Area Apache Hadoop mit dem darauf aufgesetzten Apache Pig. Da es Ziel ist alle Operationen weitestgehend mit Funktionalitäten durchzuführen, die durch Talend Open Studio for Big Data zur Verfügung gestellt werden, wird versucht möglichst nicht auf User Defined Functions zurückzugreifen. Die Hadoop Distribution, die im Rahmen des Benchmarks zum Einsatz kommen soll, ist HDP 2.5 von Hortonworks. Die Installation notwendiger Komponenten geschieht mittels Apache Ambari. Apache Ambari ist ein Apache Projekt zur Vereinfachung des Hadoop-Managements. Ambari bietet die Möglichkeit des Monitoring und des Konfigurierens über ein Web UI [11]. Eine weitere wichtige Komponente aus der Hortonworks Data Platform HDP 2.5, die zum Einsatz kommt, ist Apache Sqoop. Apache Sqoop ist ein ein Tool für den Bulk-Transfer zwischen dem Hadoop Distributed File System und relationalen Datenbanken [12]. Apache Sqoop soll den Datentransfer zwischen der Staging Area und dem Data Warehouse gewährleisten. Bevor jedoch HDP 2.5 installiert wird, ist die Größe des Hadoop Clusters festzulegen. Danach werden auf den jeweiligen Knoten des Clusters, zuerst eine Linux Distribution installiert die kompatibel mit HDP 2.5 ist. Im Rahmen dieser Diplomarbeit wird auf jedem Knoten CentOS 7 installiert.

Da der Parallelisierung das Hauptaugenmerk gilt, werden im Rahmen der Diplomarbeit zwei Clustergrößen zum Einsatz kommen. Dabei soll anhand von Clustern mit zwei und vier Knoten der Performance Unterschied beobachtet werden. Die jeweiligen Knoten sind auf einer virtuellen Maschine zu konfigurieren. Das Data Warehouse, in das die Daten aus den Quellsystemen hineinfließen sollen, wird sich in einer Postgres Datenbank befinden. Die Postgres Instanz wird auf einer eigenen virtuellen Machine laufen. Ein weiterer wichtiger Punkt ist der Datenverkehr zwischen dem Hadoop Distributed File System und dem Data Warehouse. Dieser Datenverkehr soll durch Apache Sqoop gewährleistet werden.

Die Auswertung

Nachdem überprüft wurde, dass die Implementation des TPC-DI Benchmarks, der Spezifikation des TPC genügt kommt es zur Ausführung. Nach Ausführung des Benchmark werden alle Informationen für den Full Disclosure Report gemäß Spezifikation zusammengetragen. Außerdem wird eine Gegenüberstellung der Werte für die jeweiligen Clustergrößen aufgestellt. Danach ist der Benchmark vollständig.

Verwandte Arbeiten

In [7] wird die Notwendigkeit von ETL Benchmarks für reale Szenarien zum Ausdruck gebracht und der damalig in Entwicklung befindliche TPC-DI Benchmark umrissen.

Zu Zeiten als der TPC-DI Benchmark angekündigt aber noch nicht veröffentlicht war entschied man sich im Rahmen der Arbeit [6] für einen eigenen Benchmark mit Hauptaugenmerk auf Performanz. Den Benchmark führte man für Talend Open Studio und Pentaho durch. Dabei besteht die Testumgebung aus einer OLTP-Komponente, dem ETL-Prozess selbst und dem Data Warehouse System. Der ETL-Prozess sebst besteht aus zwei Phasen. Zum einen dem Historical Load und

zum anderen dem Incremental Load. Die Performance wird unter den Gesichtspunkten Laufzeit, Inanspruchnahme von Rechenkapazität, Inanspruchnahme von Hauptspeicherkapazitäten und der Anzahl von Select Statements gemessen. Die Quellkomponente in der Testumgebung ist ein OLTP-System einer Bücherei. Die Generierung der Daten oder die Inhalte des Transformationsprozesses werden nicht offengelegt.

In [8] wird ein allgemeinerer und eher modellbasierter Ansatz für ETL-Benchmarks beschrieben. Dabei versucht man hier die Haupteigenschaften und Besonderheiten von ETL-Prozessen aufzudecken und dafür einheitliche Testszenarien zu finden. Gemessen werden sollen die ETL-Prozesse in den Punkten Effektivität und Effizienz. Hierfür werden Messgrößen definiert. Ausgehend von der Annahme, dass typische ETL-Prozesse einem Butterflymuster folgen, werden Einzelheiten anhand dieses Musters diskutiert. Der linke Flügel des Schmetterlings soll die Quellsysteme und Transformationsprozesse inklusive der Hilfsspeicher darstellen. Das Zentrum bzw. der Körper des Schmetterlings steht für den zentralen und persistenden Speicher, in dem die Daten die von den Quellsystemen produziert werden, nach Transformationen abgelegt werden. Dies könnte zum Beispiel eine Faktentabelle zusammen mit Dimensionstabellen sein. Der rechte Flügel stellt die Bereitstellung von Daten aus dem Körper für Reporting- und Analysezwecke dar.

In der Masterarbeit [13] wurde der TPC-DI Benchmark für das ETL-Tool Pentaho Kettle implementiert. Der Benchmark wurde dabei auf einem persönlichen Rechner mit dem Betriebssystem Windows 10 ausgeführt. Für das Data Warehouse wurde eine Postgres Instanz der Version PostgreSQL 9.4 verwendet.

[14] befasst sich mit dem Vergleich der ETL-Tools CloverETL, Talend und Pentaho. Hierfür wird der Benchmark TPC-H des Transaction Processing Performance Council in einer modifizierten Variante für ETL-Tools verwendet.

Zu einem Vergleich der ETL-Tools Data Stage Server 7.5, Data Stage PX 7.5, Talend Open Studio 2.4.1, Informatica 8.1.1, Pentaho Data Integrator 3.0.0 kommt es in [15]. Dafür werden die Tools auf Basis von elf selbstdefinierten Testszenarien verglichen.

Außerdem veröffentlichte Len Wyatt im Jahr 2008, in dem Microsoft Entwickler Netzwerk einen Beitrag unter dem Titel "ETL World Record! "[16] bei dem mehr als 1 TB Daten mit SSIS innerhalb von weniger als dreißig Minuten geladen seien. Die Daten wurden mittels dem Datengenerator der für den Benchmark TPC-H, ein weiterer Benchmark des Transaction Processing Performance Council, mitgeliefert wird generiert.

Des Weiteren wird in [17] ein Vergleich zwischen den ETL-Tools Talend Open Studio und SSIS gemacht, wobei es hier um keinen Benchmark handelt sondern um einen allgemeinen Vergleich der Tools.

Literaturverzeichnis

[1] Duden. Aufrufbar unter: `http://www.duden.de/rechtschreibung/Daten`.

[2] McKinsey Global Institute. Big data: The next frontier for innovation, competition,and productivity. Aufrufbar unter: `https://bigdatawg.nist.gov/pdf/MGI_big_data_full_report.pdf`, 2011.

[3] Maurizio Lenzerini. Data Integration: A Theoretical Perspective. In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, pages 233–246, New York, NY, USA, 2002. ACM.

[4] Alon Halevy, Anand Rajaraman, and Joann Ordille. Data Integration: The Teenage Years. In *Proceedings of the 32Nd International Conference on Very Large Data Bases*, VLDB '06, pages 9–16. VLDB Endowment, 2006.

[5] Wikipedia. ETL-Prozess — Wikipedia, Die freie EnzyklopÃ¤die, 2017. [Online; Stand 1. Mai 2017].

[6] Tim A. Majchrzak, Tobias Jansen, and Herbert Kuchen. Efficiency Evaluation of Open Source ETL Tools. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, SAC '11, pages 287–294, New York, NY, USA, 2011. ACM.

[7] Len Wyatt, Brian Caufield, and Daniel Pol. *Principles for an ETL Benchmark*, pages 183–198. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[8] Panos Vassiliadis, Anastasios Karagiannis, Vasiliki Tziovara, and Alkis Simitsis. Towards a Benchmark for ETL Workflows. In *QDB*, 2007.

[9] Meikel Poess, Tilmann Rabl, Hans-Arno Jacobsen, and Brian Caufield. TPC-DI: The First Industry Benchmark for Data Integration. *Proc. VLDB Endow.*, 7(13):1367–1378, August 2014.

[10] TPC Benchmark DI - Standard Specification. Aufrufbar unter: `http://www.tpc.org/tpc_documents_current_versions/pdf/tpc-di_v1.1.0.pdf`, 2014.

[11] Apache Ambari Website. Aufrufbar unter: `https://ambari.apache.org/`.

[12] Apache Sqoop Website. Aufrufbar unter: `http://sqoop.apache.org/`.

[13] Maurice Bleuel. Implementation and Evaluation of the TPC-DI Benchmark for Data Integration Systems. Master's thesis, Humboldt-Universität zu Berlin, 2016.

[14] Petr Uher. Comparison CloverETL vs. Talend, Pentaho. Aufrufbar unter: `https://is.muni.cz/th/373858/fi_b/Comparison_CloverETL_vs_Talend_Pentaho.pdf`, 2009.

[15] ETL Benchmarks. Aufrufbar unter: https://marcrussel.files.wordpress.com/2008/10/etlbenchmarks_manappsc221008.pdf, 2008.

[16] Len Wyatt. ETL World Record! Aufrufbar unter: https://blogs.msdn.microsoft.com/sqlperf/2008/02/27/etl-world-record/, 2008.

[17] R. Katragadda, S.S. Tirumala, and D. Nandigam. ETL tools for Data Warehousing: An empirical study of Open Source Talend Studio versus Microsoft SSIS, 2015.