



Diplomhausarbeit

**Open Source Business Intelligence -
Evaluation der Effizienz von ETL-Werkzeugen**

**Open Source Business Intelligence -
Performance Evaluation of ETL-Tools**

Tobias Jansen

In Kooperation mit der



Themensteller: Prof. Dr. Herbert Kuchen
Betreuer: Dr. Frank Köhne (viadee) / Tim A. Majchrzak, MScIS
Institut für Wirtschaftsinformatik
Praktische Informatik in der Wirtschaft

Inhaltsverzeichnis

Abkürzungsverzeichnis	IV
1 Einleitung	1
2 Grundlagen und kontextbezogene Begriffe	4
2.1 Open Source Software	4
2.1.1 Definition und Entstehung	4
2.1.2 Vor- und Nachteile für den Einsatz von Open Source Software.....	6
2.1.3 Rechtliche Rahmenbedingungen und Lizenztypen	8
2.2 Business Intelligence	10
2.2.1 Klassische Management Support Systeme.....	10
2.2.2 Begriffsabgrenzung und Definitionsansätze	14
2.2.3 Architekturmodell	16
2.3 Data Warehouse.....	18
2.3.1 Definition und Architekturen	18
2.3.2 Multidimensionale Datenmodellierung.....	21
2.4 ETL	25
2.4.1 Einordnung und Bestandteile	25
2.4.2 Slowly Changing Dimensions.....	27
2.4.3 ETL-Werkzeuge	28
3 Auswahl und Vorstellung der zu evaluierenden ETL-Werkzeuge	29
3.1 Auswahlprozess	29
3.1.1 Vorstellung	29
3.1.2 Durchführung	30
3.2 Talend Open Studio	35
3.3 Pentaho Data Integration	36
4 Konzeption und Umsetzung der Performance-Messung	38
4.1 Entwicklung eines ETL-Benchmarks	38
4.2 Testszenario	40
4.3 Konzeption und Umsetzung des ETL-Prozesses.....	46
4.3.1 Konzeption	46
4.3.2 Umsetzung mit Talend Open Studio	49
4.3.3 Umsetzung mit Pentaho Data Integration	49

4.4 Testkonzeption und -umsetzung	51
4.4.1 Definition der Performance-Indikatoren	51
4.4.2 Konzeption des Testlaufs	52
4.4.3 Konfiguration der Testfälle	54
4.4.4 Testdurchführung	58
5 Analyse und Bewertung der Performancetests	62
5.1 Streuungsanalyse	62
5.1.1 Übersicht	62
5.1.2 Ausführungszeit	63
5.1.3 CPU-Auslastung	65
5.1.4 Speicherauslastung	68
5.2 Performanceanalyse und Diskussion	71
5.2.1 Übersicht und allgemeine Beobachtungen	71
5.2.2 Datenmenge (Bücheranzahl)	73
5.2.3 Arbeitsspeicher	76
5.2.4 JDK	78
5.2.5 ETL-Organisation	79
5.3 Gründe für Performance-Unterschiede	81
5.4 Bewertung	82
6 Schlussbetrachtung und Ausblick	84
Literaturverzeichnis	86
Anhang	93