

**ETL4NoSQL: Um Framework Programável para Extração, Transformação e Carga de Banco de Dados NoSQL**

**Por**

***Carine Calixto Aguena***

**Dissertação de Mestrado**



Universidade Federal de Pernambuco

posgraduacao@cin.ufpe.br

www.cin.ufpe.br/~posgraduacao

RECIFE/2017

Carine Calixto Aguena

**ETL4NoSQL: Um Framework Programável para Extração, Transformação e Carga de Banco de Dados NoSQL**

*Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.*

Orientadora: Prof. Valéria Cesário Times

RECIFE 2017

*Eu dedico este trabalho à toda minha família, amigos e professores que me deram todo apoio necessário para alcançar meus objetivos.*

Agradecimentos

À minha família, meus pais que sempre foram meu alicerce e estrutura de vida, ao meu irmão e cunhada pelos estímulos e conselhos, à minha sobrinha pelo carinho, ao meu amor por todo suporte e compreensão nos momentos difíceis e desafiantes que este trabalho proporcionou.

À Prof. Valéria, o meu reconhecimento pela oportunidade de realizar este trabalho ao lado de alguém paciente e que transpira sabedoria; meu respeito e admiração pela sua serenidade, capacidade de análise do perfil de seus alunos, e pelo seu dom no ensino da Ciência, inibindo sempre a vaidade em prol da simplicidade e eficiência.

Aos amigos, e todos que de alguma forma colaboraram nessa grande jornada, os precurso- res de tudo, que exemplificam a ética e competência profissionais, a dedicação e o aprimoramento contínuos, pelo incentivo e oportunidade de convívio.

E, finalmente, dedico este trabalho à Deus, que sempre me guia e impulsiona pelos caminhos que acalmam meu coração. Agradeço pela maravilhosa oportunidade de obter novos conhecimentos e encontrar pessoas maravilhosas.

*Sempre que houver alternativas, tenha cuidado. Não opte pelo conveniente, pelo confortável, pelo respeitável, pelo socialmente aceitável, pelo honroso. Opte pelo que faz o seu coração vibrar. Opte pelo que gostaria de fazer, apesar de todas as consequências.*

—OSHO

Resumo

A integração de dados e os processos de extração, transformação e carga de dados (ETL) são procedimentos cruciais para a criação de *data warehouses* (DW). Porém, os processos de ETL e integração de dados são habitualmente desenvolvidos para dados estruturados por modelos relacionais, que representam apenas uma pequena parte dos dados mantidos por muitas empresas. Dessa forma, existe uma demanda crescente para extrair, transformar e carregar dados estruturados por modelos de dados não relacionais e carregá-los em um repositório de dados unificado. Porém, devido a complexidade desses modelos de dados, existem vários desafios para a realização da extração, transformação e carga de dados organizados por modelos não relacionais que precisam considerar características específicas, como por exemplo, a heterogeneidade e distribuição dos dados, em um ambiente de extração, transformação e carga de dados.

Além disso, muitas empresas encontram dificuldades ao lidar com as ferramentas de ETL disponíveis no mercado, por causa muitas vezes da necessidade de integração destas ferramentas de ETL com sistemas legados. Aprender a lidar com essas ferramentas pode ser muito custoso em termos financeiro e de tempo, e por isso, muitas empresas acabam optando por desenvolver os seus processos utilizando uma linguagem de programação de propósito geral.

Portanto, neste trabalho propomos ETL4NoSQL, um frameworkprogramável, flexível e integrado para auxiliar a modelagem e execução de processos de ETL, que possibilita a extração, transformação e carga de dados estruturados em modelos de dados não relacionais. Apresentamos os componentes do frameworkETL4NoSQL, bem como suas interfaces e funcionalidades. Ademais, realizamos um estudo experimental de software, que teve como objetivo verificar se o ETL4NoSQL é adequado para auxiliar no desenvolvimento de processos de ETL. O estudo consistiu na análise das ferramentas de ETL estudadas, com o propósito de caracterizá-las por meio de suas funcionalidades no contexto comparativo entre elas. Finalmente, propomos um ambiente de implementação de ETL que permite o reuso e a extensão de interfaces de programação de ETL4NoSQL para desenvolver aplicações de ETL utilizando dois tipos diferentes de SGBD NoSQL.

Palavras-chave: ETL, Frameworks, NoSQL, Data Warehouse, Estudo Experimental de Software.

Abstract

Data integration and data extraction, transformation, and loading (ETL) processes are procedures for creating data warehouses (DW). However, ETL processes and data integration are usually developed for data structured by relational models, which represent only a small part of the data maintained by many companies. In this way, there is a growing demand to extract, transform, and load structured data from non relational data models and load them into a unified data repository. However, due to the complexity of the data models, there are several challenges for an extraction, transformation and loading of data organized by non relational models, requirements, specific characteristics, such as heterogeneity and distribution of data, in an environment of extraction, transformation and data loading. In addition, many companies are struggling to deal with ETL tools available in the market, often because of the need to integrate ETL tools with legacy systems. Learning how to handle these tools can be very costly in terms of time and financially, and so many companies end up opting to develop their processes using a general purpose programming language. Therefore, we propose ETL4NoSQL, a programmable, flexible framework and integration to help, model and execute ETL processes, which enables structured data extraction, transformation and loading in non relational data models. We present the components of the ETL4NoSQL framework, as well as its interfaces and functionalities. In addition, we conducted an experimental software study, which aimed to verify if ETL4NoSQL is suitable to assist in any development of ETL processes. The study consisted of the analysis of ETL tools found in the literature, with the purpose of characterizing them through the intersection of their functionalities in the comparative context between them. Finally, we propose an ETL implementation environment that allows the reuse and extension of ETL4NoSQL programming interfaces for ETL development applications, using two different types of NoSQL DBMSs.

Keywords: ETL, Frameworks, NoSQL, Data Warehouse, Experimental Software Study.

Lista de Figuras

2.1 Inversão de controle em *framework*. (Adaptado de SOMMERVILLE (2013))................ 15

2.2 Exemplo da Ferramenta ARKTOS II em uso (Adaptado de VASSILIADIS et al. (2005)) ...................................................................................................................................................... 17

2.3 Fluxo de dados ETL no framework MapReduce (Adaptado de LIU; THOMSEN;  PEDERSEN(2011)) ................................................................................................................. 19

2.4 Arquitetura do CloudETL (Adaptado de LIU; THOMSEN; PEDERSEN (2013)) ....... 20

2.5 Interface de configuração do P-ETL (Adaptado de BALA (2014)) .............................. 22

2.6 Quadro comparativo entre frameworks para desenvolvimento de sistemas de ETL ..... 25

3.1 ModeloconceitualdoETL4NoSQL ....................... 30

3.2 DiagramadeEstadodoETL4NoSQL ...................... 33

3.3 Definição de interfaces do modelo de negócio do ETL4NoSQL . . . . . . . . . 38

3.4 Especificação da arquitetura do componente de ETL4NoSQL . . . . . . . . . . 39

3.5 Diagrama de colaboração para conectar à base de fonte de dados . . . . . . . . 40

3.6 Diagrama de colaboração para verificar a estrutura de dados . . . . . . . . . . 40

3.7 Diagrama de colaboração para verificar se existe permissão de escrita na base de dados destino ................................... 40

3.8 Diagrama de colaboração para leitura dos dados da fonte . . . . . . . . . . . . 40

3.9 Diagrama de colaboração criação das operações de ETL . . . . . . . . . . . . 41

3.10 Diagrama de colaboração modelar as operações criadas . . . . . . . . . . . . . 41

3.11 Diagrama de colaboração para gerenciar as operações . . . . . . . . . . . . . . 41

3.12 Diagrama de colaboração para processar as operações . . . . . . . . . . . . . . 41

3.13 Diagrama de especificação das interfaces de ETL4NoSQL . . . . . . . . . . . 42

3.14 Tela do IDE LiClipse com a implementação da interface ETL4NoSQLMgr . . . 46

3.15 Tela do IDE LiClipse com a implementação da interface IDataMgr . . . . . . . 47

3.16 Tela do IDE LiClipse com a implementação da interface IModelMgr . . . . . . 47

3.17 Tela do IDE LiClipse com a implementação da interface IOpMgr . . . . . . . . 48

3.18 Tela do IDE LiClipse com a implementação da interface IProcMgr . . . . . . . 48

4.1 Questionário de Funcionalidades......................... 60

4.2 Quantidade de Presença para cada funcionalidade . . . . . . . . . . . . . . . . 61

4.3 Níveis de utilidade para cada funcionalidade. . . . . . . . . . . . . . . . . . . 61

4.4 Necessidade de melhoria para cada funcionalidade . . . . . . . . . . . . . . . 62

5.1 Modelo Multidimensional da aplicação ETL4NoSQLMongoStar . . . . . . . . 74

5.2 TeladaaplicaçãoETL4NoSQLMongoStar.................... 75

5.3 Modelo Multidimensional da aplicação ETL4NoSQLCassandraStar . . . . . . 76

5.4 TeladaaplicaçãoETL4NoSQLCassandraStar .................. 77

Lista de Quadros

3.1 Requisitos do ETL4NoSQL............................ 29

3.2 Modelo de Casos de Uso do ETL4NoSQL.................... 31

4.1 Descrição da Instrumentação........................... 54

4.2 Métricas...................................... 55

4.3 Questionário do Perfil da FerramentadeETL .................. 58

4.4 Instrumentação para aplicar o questionário.................... 59

4.5 Resultado do Perfil dos participantes....................... 62

4.6 Legenda...................................... 63

4.7 Estatística Descritiva da Presença de Funcionalidades . . . . . . . . . . . . . . 63

4.8 Estatística Descritiva da Melhoria de Funcionalidades . . . . . . . . . . . . . . 64

4.9 Estatística Descritiva da Utilidade de Funcionalidades . . . . . . . . . . . . . . 64

4.10 Funcionalidade presente e parcialmente útil ................... 65

4.11 Funcionalidade presente e é útil ......................... 65

4.12 Funcionalidade presente enecessita melhorar . . . . . . . . . . . . . . . . . . 66

4.13 Funcionalidade presente e não necessita melhoria . . . . . . . . . . . . . . . . 66

4.14 Quadro de resultado do valor observado de existência da funcionalidade . . . . 67

4.15 Quadro do resultado de χ2 ............................ 68

4.16 Funcionalidades não oferecidas pelo ETL4NoSQL . . . . . . . . . . . . . . . 70

Lista de Acrônimos

APD Área de Processamento de Dados

BDs NoSQL Banco de Dados NoSQL

BI Business Intelligence

DW Data Warehouse

ETL Extract, Transform and Load

JSON JavaScript Object Notation

RDBMS Relational Database Management System

SGBD Sistema Gerenciador de Bancos de Dados

SQL Structured Query Language

OCL Object Constraint Language

UML Unified Modeling Language

XML eXtensible Markup Language

YAML Yet Another Markup Language

Sumário

1 Introdução .....................................................................................................................................1

1.1 Contextualização ........................................................................................................... 2

1.2 Motivação....................................................................................................................... 3

1.3 Objetivos ....................................................................................................................... 5

1.3.1 Objetivos Específicos...................................................................................... 6

1.4 Contribuições.................................................................................................................. 6

1.5 Organização do Trabalho................................................................................................ 7

2 Fundamentação Teórica ................................................................................................................8

2.1 ConceitosBásicos............................................................................................................ 9

2.1.1 ETL.................................................................................................................. 9

2.1.2 Sistemas de Bancos de Dados NoSQL........................................................... 10

2.1.2.1 Sistemas de Banco de dados Orientados à Documentos .................10

2.1.2.2 Sistemas de Banco de dados Famílias de Colunas ..........................11

2.1.2.3 Sistemas de Banco de dados Baseado em Grafos ...........................11

2.1.2.4 Sistemas de Banco de dados Chave-Valor ......................................12

2.1.3 Desenvolvimento Baseado em Componentes e Frameworks ........................12

2.1.4 Estudo Experimental de Software .................... 15

2.2 Trabalhos Correlatos ............................... 17

2.2.1 ARKTOSII................................ 17

2.2.2 PygramETL................................ 18

2.2.3 ETLMR.................................. 18

2.2.4 CloudETL ................................ 20

2.2.5 P-ETL................................... 21

2.2.6 Big-ETL ................................. 22

2.2.7 Outras Ferramentas............................ 23

2.2.7.1 Pentaho ............................ 23

2.2.7.2 TalendStudio ......................... 23

2.2.7.3 CloverETL........................... 23

2.2.7.4 Oracle Data Integrator (ODI) ................. 24

2.2.7.5 FramETL................................. 24

2.2.8 Comparativo das Ferramentas de ETL....24

2.3 Considerações Finais ............................... 25

3 O Framework ETL4NoSQL.. 27

3.1 Requisitos de Software do ETL4NoSQL..................... 28

3.2 ModelagemdoDomíniodeETL4NoSQL .................... 30

3.2.1 Modelo Conceitual............................ 30

3.2.2 Modelo de Casos de Uso......................... 30

3.2.3 Modelo Comportamental......................... 32

3.3 Modelagem da Especificação do ETL4NoSQL.................. 32

3.3.1 Identificação de Componentes ...................... 32

3.3.2 Interfaces de Sistemas .......................... 33

3.3.3 Interfaces de Negócio........................... 34

3.3.4 Especificação da Arquitetura do Componente . . . . . . . . . . . . . . 38

3.4 Interação entre Componentes........................... 39

3.4.1 Operações da interface de negócio.................... 39

3.5 Especificação de Componentes.......................... 41

3.6 Ambiente de Implementação ........................... 44

3.7 Interfaces de Programação ............................ 45

3.8 Considerações Finais ............................... 49

4 Estudo Experimental de Software ... 50

4.1 Objetivos do experimento............................. 51

4.1.1 Objetivo da Medição ........................... 51

4.1.2 Objetivos do Estudo ........................... 51

4.1.3 Questões ................................. 52

4.2 Planejamento ................................... 52

4.2.1 Definição das Hipóteses ......................... 52

4.2.2 Descrição da instrumentação....................... 54

4.2.3 Métricas.................................. 55

4.2.4 Seleção do contexto ........................... 55

4.2.5 Seleção dos indivíduos.......................... 56

4.2.6 Variáveis ................................. 56

4.2.7 Análise Qualitativa............................ 57

4.2.8 Validade.................................. 57

4.3 Operação ..................................... 58

4.3.1 Questionário do Perfil da Ferramenta de ETL . . . . . . . . . . . . . . 58

4.3.2 Questionário de Funcionalidades..................... 59

4.3.3 Resultado do Estudo ........................... 59

4.4 Análise e interpretação dos resultados ...................... 63

4.4.1 Estatística descritiva ........................... 63

4.4.2 Aplicação do teste estatístico....................... 67

4.4.2.1 Análise quantitativa ...................... 684.4.3 Análise qualitativa ............................ 694.4.4 Verificação das hipóteses......................... 70

4.5 Considerações Finais ............................... 71

5 Aplicações de ETL4NoSQL.. 72

5.1 Aplicações de ETL4NoSQL...................... 73

5.2 Aplicação ETL4NoSQLMongoStar ....................... 73

5.3 Aplicação ETL4NoSQLCassandraStar...................... 75

5.4 Considerações Finais ............................... 78

6 Conclusão 79

6.1 Principais Contribuições ............................. 80

6.2 Discussão Final..................................... 81

6.3 Trabalhos Futuros................................. 81

Referências 83