Universidade Federal de Pernambuco
Cin - Centro de Informática

Pós graduação em Ciência da Computação

ETL4NoSQL: Um framework de ETL para BDs NoSQL

Carine Calixto Aguena

Dissertação de Mestrado

Recife
<DATA DA DEFESA>

Universidade Federal de Pernambuco
Cin - Centro de Informática

Carine Calixto Aguena

ETL4NoSQL: Um framework de ETL para BDs NoSQL

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

Orientador: *Valéria Cesário Times*

Recife
<DATA DA DEFESA>

<DIGITE A DEDICATÒRIA AQUI>

Agradecimentos

<DIGITE OS AGRADECIMENTOS AQUI>

<*DIGITE AQUI A CITAÇÂO*>
—<AUTOR> (<NOTA>)

Resumo

<DIGITE O RESUMO AQUI>

Palavras-chave: <DIGITE AS PALAVRAS-CHAVE AQUI>

Abstract

Keywords: <DIGITE AS PALAVRAS-CHAVE AQUI>

Sumário

Lista de Figuras

Lista de Quadros

CAPÍTULO 1

Introdução

Este capítulo contextualiza os principais assuntos abordados neste trabalho de dissertação, apresenta as motivações que levaram à escolha do tema, os objetivos gerais e específicos da proposta desta pesquisa, bem como a justificativa para conduzir uma investigação no assunto debatido.

1.1 Contextualização

Desde a década de 1970, com a criação do modelo relacional por Edgar Frank Codd, a estrutura de armazenamento adotada por muitos desenvolvedores de sistemas da área de tecnologia da informação tem se baseado no conceito de entidade e relação proposto por Codd. A maioria dos sistemas gerenciadores de banco de dados que possui aceitação no mercado fazem uso desse modelo, por exemplo o MySQL, Oracle e Microsoft SQL Server. Porém, os requisitos para o desenvolvimento de ferramentas de software modernas têm mudado significativamente, especialmente com o aumento das aplicações Web [13]. Este segmento de aplicações exige requisitos com alta escalabilidade e vazão, onde sistemas que utilizam um armazenamento com esquema relacional não conseguem atender satisfatoriamente. Em resposta a isso, novas abordagens de armazenamentos de dados utilizando o termo de NoSQL tornaram-se popular.

O termo NoSQL é constantemente interpretado como "*Not Only SQL*", cujo SQL refere-se a linguagem de manipulação de dados dos gerenciadores de armazenamento de dados relacionais (RDBMS - Relational Database Management System) - Structure Query Language [13]. O grande propósito das abordagens NoSQL é oferecer alternativas onde os esquemas relacionais não apresentam um bom desempenho. Esse termo abrange diferentes tipos de sistemas. Em geral, banco de dados NoSQL usam modelo de dados não-relacionais, com poucas definições de esquema, são executados em clusters e aplicados a alguns bancos de dados recentes como o Cassandra, o Mongo, o Neo4J e o Riak [8].

Muitas empresas coletam e armazenam milhares de gigabytes de dados por dia, no qual a análise desses dados torna-se uma vantagem competitiva no mercado. Por isso, há uma grande necessidade de uma nova arquitetura para o gerenciamento de suporte à decisão que possa alcançar melhor escalabilidade e eficiência [12]. Para auxiliar no processo de gerenciamento de suporte à decisão uma das formas mais utilizadas é a criação de um ambiente data warehousing que é responsável por providenciar informações estratégicas e esquematizadas a respeito do negócio [1].

Segundo a definição de [11], data warehouse (DW) é uma coleção de dados para o processo de gerenciamento de suporte à decisão orientado a assunto, integrado, variante no tempo e não volátil. Os dados de diferentes fontes de sistemas são processados em um data warehouse central através da Extração, Transformação e Carga (ETL) de maneira periódica.

Ferramentas de ETL são sistemas de software responsáveis por extrair dados de diversas fontes, transformar e customizar os dados e inseri-los no data warehouse. Comumente, esses processos são executados periodicamente, onde a otimização do seu tempo de execução torna-se importante [14].

O projeto de ETL consome cerca de 70% dos recursos de implantação de um DW, pois desenvolver esse projeto é crítico e custoso, tendo em vista que gerar dados incorretos pode acarretar em más decisões. Porém, por algum tempo pouca importância foi dada ao processo de ETL pelo fato de ser visto somente como uma atividade de suporte aos projetos de DW. Apenas a partir do ano 2000, a comunidade acadêmica passou a dar mais importância ao tema [5].

Tradicionalmente, o DW é implementado em uma base de dados relacional, onde o dado é armazenado nas tabelas fato e tabelas dimensões, na qual forma um esquema em estrela

[11]. Por isso, é comum que as ferramentas de ETL utilizadas no mercado atualmente só dêem suporte aos esquemas relacionais. Para oferecer suporte aos sistemas que necessitem utilizar um esquema não relacional de BDs NoSQL em DW, a proposta desse trabalho é especificar um framework programável, flexível e integrado para modelagem e execução de processos ETL em BDs NoSQL.

1.2 Motivação

A integração de dados e os processos de ETL são procedimentos cruciais para a criação de data warehouses e sistemas BI (business intelligence). Porém, os sistemas para ETL e integração de dados são tradicionalmente desenvolvidos para dados estruturados em modelos relacionais que representam apenas uma pequena parte dos dados mantidos por muitas empresas [3] [Russom 2007, Pedersen 2009]. Dessa forma, existe uma demanda crescente para integrar os dados não estruturados e semi estruturados em um repositório unificado. Devido a complexidade desses dados, novos desafios estão surgindo quando lidamos com dados heterogêneos e distribuídos no ambiente de integração [Salem, 2012].

Além disso, muitas empresas encontram dificuldades para lidar com as ferramentas ETL disponíveis no mercado. Aprender a lidar com essas ferramentas pode ser muito custoso em termos financeiros e de tempo, e por isso, acabam optando desenvolver os seus processos por meio de uma linguagem de programação de propósito geral [Awad et al., 2011; Muñoz et al., 2009].

Portanto, este trabalho propõe um framework programável para desenvolvimento de sistemas de ETL que possibilita a integração de dados estruturados, não estruturados e semi estruturados armazenados em bases relacionais ou NoSQL. O framework possui um ambiente integrado para a importação e mapeamento dos dados, além da modelagem e customização dos processos de ETL. Os processos de importação e mapeamento do framework integram dados estruturados, não estruturados e semi estruturados. Esses processos possibilitam a leitura e manipulação de dados de bases NoSQL, e também o armazenamento desses dados em bases deste tipo, oferecendo uma alternativa não relacional para a construção de DWs.

Uma alternativa para organizar e manipular grandes volumes de dados sem utilizar um modelo relacional e ainda processá-los e armazená-los de maneira distribuída é fazer o uso de BDs NoSQL [4]. Com isso, surge a necessidade de se promover meios para o uso desses BDs em DWs.

As pesquisas presentes na literatura sobre extração de dados em BDs NoSQL mostram que não há uma ferramenta que seja integrada para o uso de BDs NoSQL, as ferramentas existentes no mercado apenas oferecem a possibilidade para alguns SGBDs NoSQL, ficando a cargo da equipe de implantação do projeto de DW todo o trabalho de modelagem e programação ao se utilizar BDs NoSQL [colocar ref das pesquisas].

[5] aponta em sua pesquisa que muitas empresas evitam ferramentas de ETL disponíveis no mercado, e adotam o desenvolvimento dos processos a partir de uma linguagem de programação de propósito geral, pelo fato dessas ferramentas terem uma longa curva de aprendizagem e grande complexidade no seu uso.

O aumento do uso de banco de dados com esquemas não relacionais baseados no para-

digma NoSQL e a falta de uma ferramenta programável, flexível e integrada, independente de plataforma que dê suporte à extração, transformação e carga em data warehouses para esses esquemas é a grande motivação deste trabalho.

Dessa forma, encontrar uma solução que seja programável, flexível e integrada para extração, transformação e carga dos dados em BDs NoSQL é a proposta deste trabalho.

1.3 Objetivos

O objetivo principal desta pesquisa é especificar um framework programável, flexível e integrado para modelagem e execução de processos ETL de banco de dados estruturados, não estruturados e semi estruturados sob os modelos relacionais e NoSQL. Os objetivos específicos são detalhados a seguir.

1.3.1 Objetivo Específico

Este trabalho de dissertação tem como um dos objetivos específicos apresentar os componentes do framework ETL4NoSQL, bem como suas funcionalidades. Outro objetivo deste trabalho é realizar um estudo experimental de software a fim de caracterizar as principais funcionalidades das ferramentas de ETL na manipulação de dados estruturados, semi estruturados e não estruturados. O estudo experimental poderá comparar o framework proposto, suas vantagens e desvantagens, em relação às ferramentas de ETL encontradas na literatura.

1.4 Contribuições

Uma das contribuições deste trabalho é fornecer um framework programável, flexível e integrado que auxilia na modelagem e execução dos processos de ETL em bases de dados estruturadas, semi estruturadas e não estruturadas, denominado ETL4NoSQL. Assim, é possível extrair, integrar e carregar dados que estão armazenados em diversas estruturas como é o caso dos bancos de dados NoSQL, ou até mesmo, repositórios de dados textuais e banco de dados relacionais em um único repositório. O ETL4NoSQL é um recurso valioso, principalmente para os desenvolvedores responsáveis pela fase de ETL, onde muitos encontram dificuldades para lidar com as ferramentas ETL disponíveis no mercado.

Outra contribuição desta pesquisa é apresentar, por meio de um estudo experimental, as principais características, de acordo com algumas ferramentas de ETL presentes na literatura, bem como possíveis melhorias, vantagens e desvantagens, em suas funcionalidades.

1.5 Organização do Trabalho

Estae trabalho está organizado de acordo com a seguinte estrutura:

- **Capítulo 2 (Fundamentação Teórica):** apresenta uma revisão da literatura dos principais assuntos abordados neste trabalho. Serão tratados temas a respeito de ETL, banco

de dados NoSQL, Frameworks e estudo experimental de software.

- **Capítulo 3 (O Framework ETL4NoSQL):** descreve os requisitos, arquitetura e componentes do framework exposto neste trabalho.
- **Capítulo 4 (Estudo Experimental de Software):** expõe o roteiro da experimentação de software para ferramentas de ETL. Define o objetivo, planejamento, operação e resultado do estudo.
- **Capítulo 5 (Considerações Finais):** expressa as limitações e ameaças à validade do trabalho, considerações finais e sugere de trabalhos futuros.

CAPÍTULO 2

Fundamentação Teórica

Neste capítulo são apresentados os conceitos relacionados ao desenvolvimento desta pesquisa, bem como o embasamento teórico necessário para o entendimento do estudo. Os temas abordados são: ETL, Banco de Dados NoSQL, Frameworks, Estudo Experimental de Software e trabalhos correlatos ao tema deste trabalho.

2.1 ETL

ETL sigla para *Extraction, Transform and Load* (Extração, Limpeza/Transformação e Carga) é conhecido na literatura por definir processos que permitem a integração de dados, centralizando-os numa base destino facilitando o gerenciamento e análise dos dados (Kimball and Caserta, 2004). O fluxo do processo de ETL inicia-se com extração dos dados a partir de uma fonte, que podem ser arquivos textuais, banco de dados relacionais ou banco de dados NoSQL. Os dados são propagados para uma Área de Processamento de Dados onde são executadas a limpeza e transformação por meio de mecanismos de ETL definidos como agregação, junção, filtro, união, entre outros. Finalmente, os dados são carregados em estruturas que podem ser data warehouses ou repositórios analíticos.

Kimball and Caserta (2006) definem ETL em quatro macroprocessos, com 34 subsistemas. O quadro 2.1 mostra os subsistemas do processo de ETL. Os quatro macroprocessos são:

- Extração: Recolhe os dados dos sistemas de origem e grava na área de processamento de dados antes de qualquer reestruturação significativa. Esta etapa possui 3 subsistemas.
- Limpeza e Transformação: Envia os dados de origem, por meio de várias etapas de processamento no sistema ETL. Melhora a qualidade dos dados recebidos da fonte, mescla dados de duas ou mais fontes para criar e aplica dimensões e métricas. Esta etapa possui 5 subsistemas.
- Entrega ou Carga: Estrutura fisicamente e carrega os dados conforme desejado em DWs ou repositórios analíticos. Esta etapa possui 13 subsistemas.
- Gerenciamento: Gerencia os sistemas e processos relacionados ao ambiente ETL de forma coerente. Esta etapa possui 13 subsistemas.

Quadro 2.1: Subsistemas do processo de ETL

Etapa	Descrição
Extração	**Data Profiling:** Explora uma origem de dados para determinar seu ajuste para inclusão como uma fonte associado à limpeza e ajuste de requisitos. **Change Data Capture:** Isola as mudanças ocorridas nos sistemas de origem, de forma a reduzir os processos de ETL - Carga Incremental. **Sistema de Extração:** Extração e movimentação dos dados de origem para dentro do DW, para processamento futuro.
Limpeza e Transformação	**Data Cleasing System - Sistema de Limpeza de Dados:** Implementa processos de qualidade de dados para identificar violações de qualidade. **Error Event Tracking - Acompanhamento de erro:** Captura todos os ?eventos de erro?, que serão as entradas vitais para a melhoria da qualidade dos dados.

	Criação de Dimensão de auditoria: Junta Metadados para cada Tabela Fato, como uma dimensão. Este Metadados estará disponível para a geração de aplicações de BI que visualizem a qualidade dos dados. **Deduplication - Tirar a duplicidade de dados:** Elimina dados redundantes de dimensões, como clientes ou produtos. Pode requerer integração cruzada entre multiplas origens e a aplicação de regras para identificar qual a versão mais correta de uma linha duplicada. **Data Conformance - conformidade de dados:** Força o uso de atributos comuns entre as principais Conformed Dimensions versus as métricas comuns nas Tabelas Fato relacionadas.
Entrega ou Carga	**Slowly Changing Dimension (SCD) Manager:** Implementa a lógica para os atributos SCD. **Surrogate Key Generator:** Cria as chaves substitutas (chaves de negócio) - surrogate keys independentes para cada dimensão. **Hierarchy Manager:** Entrega multipla e simultânea de estruturas hierarquicas na dimensão. **Special Dimensions Manager:** Cria locais - placeholders na estrutura de ETL para sustentar os processos repetitivos específicos da organização, no desenho de dimensões específicas coma as Junk Dimensions, Mini Dimensions e indicadores de comportamento. **Fact Table Builders:** Construção dos três tipos básicos de tabela fato: Transacional, Periódico e Cumulativo (transaction grain, periodic snapshot e accumulating snapshot). **Surrogate Key Pipeline:** Substitui, nas dimensões, a chave natural operacional das tabelas de origem pelas chaves substitutas (Surrogate Key) que serão utilizadas para o relacionamento com as tabelas fato. **Multi-Valued Bridge Table Builder:** Construção e Manutenção das tabelas ponte (bridge tables) para suportar os relacionamentos multi-valorados. **Late Arriving Data Handler:** Aplica modificações especiais nas procedures do processo padrão para lidar com tabelas fato recém definidas (late-arriving) e dimensões. **Dimension Manager:** Centraliza a autoridade para preparar e divulgar as dimensões conforme (conformed dimensions) para a comunidade do Data Warehouse. **Fact Table Provider:** Detém a administração de uma ou mais tabelas fato, e a responsabilidade de criação, manutenção e uso.

	Aggregate Builder: Construção e manutenção de agregações que serão usadas de forma continua com tecnologias de navegação agregada para melhorar a performance das consultas. **OLAP Cube Builder:** Seleciona os dados do esquema dimensional para popular os cubos OLAP. **Data Propagation Manager:** Prepara dados conformados e integrados no servidor de apresentação do Data Warehouse, para entrega em outros ambientes, para propósitos especiais.
Gerenciamento	**Job Scheduler:** A estratégia de gerenciamento da execução dos ETLs deve ser confiável, incluindo os relacionamentos e dependências entre os ETLs. **Backup System:** Mantem cópia do ambiente de ETL para propósito de recuperação, restart e arquivamento. **Recovery and Restart:** Processos para recuperação do ambiente de ETL ou processo de reinicio, em caso de eventuais falhas. **Version Control:** Mantem arquivadas versões dos ETLs, para eventual recuperação das lógicas e metadados do 'ETL pipeline'. **Version Migration:** Migração de uma versão completa do 'ETL pipeline' a partir do ambiente de desenvolvimento para um ambiente de testes e, finalmente, para o ambiente de produção. **Workflow Monitor:** Garante que os processos de ETL estão sendo eficientemente executados e que as cargas iniciem precisamente nas janelas de tempo estipuladas. **Sorting:** Garante a fundamental alta performance nos grupos de processos de ETLs. **Lineage and Dependency:** Identifica a origem dos dados, as localizações intermediarias, as transformações e o dado final, permitindo acompanhar de forma estruturada, a trajetória dos dados até a sua carga no Data Warehouse. **Problem Escalation:** Estrutura de suporte que encaminha os problemas encontrados nos processos de ETLs (erros) para o nível de solução apropriado. **Paralleling and Pipelining:** Habilita ao sistema de ETL a potencializar automaticamente o uso de recursos como múltiplos processadores ou computação em grade (grid computing) para entregas dentro dos prazos restritos. **Security:** Garante o acesso autorizado aos ETLs e Metadados, de forma individual ou em grupos, mantendo um histórico dos acessos. **Compliance Manager:** Suporta os requerimentos organizacionais de conformidade, através, tipicamente, da manutenção da custódia da cadeia de dados e do acompanhamento dos acessos aos dados (quem teve o acesso autorizado ao dado).

	Metadata Repository: Captura os metadados do ETL, incluindo os metadados de processo, metadados técnicos e metadados do negócio que significam todos os metadados do ambiente de DW/BI.

2.2 Bancos de Dados NoSQL

Consistem em bancos de dados não relacionais projetados para gerenciar grandes volumes de dados e que disponibilizam estruturas e interfaces de acesso simples (Lima; Mello, 2015). Cada SGBD (Sistema Gerenciador de Banco de Dados) NoSQL possui um esquema de modelagem diferente, nos quais são divididas pela literatura em quatro categorias amplamente usadas: Chave-Valor, Orientado a Documentos, Famílias de Colunas e Baseado em Grafos ([Fowler, 2013], [Kaur; Rani, 2013]).

As principais características dos banco de dados NoSQL são: distribuído, escalabilidade horizontal, construído para grande volume de dados, BASE ao invés de ACID, modelo de dados não relacional, não suporta SQL[Fowler, 2013]. [13]

2.2.1 Banco de dados Orientados à Documentos

Banco de dados orientados a documentos são capazes de armazenar documentos como dado. Esses documentos podem ser em qualquer formato como XML (eXtensible Markup Language), YAML (Yet Another Markup Language), JSON (JavaScript Object Notation), entre outros. Os documentos são agrupados na forma de coleções. Comparando com banco de dados relacional, as coleções são como tabelas e os documentos como os registros. Porém, a diferença entre eles é que cada registro na tabela do banco relacional tem o mesmo número de campos, enquanto que na coleção do banco de dados orientado a documentos, podem ter campos completamente diferentes (Kaur; Rani, 2013).

Existem mais de 15 banco de dados orientados a documentos disponíveis e os mais utilizados são MongoDB, CouchDB e o RavenDB (Kaur; Rani, 2013).

2.2.2 Banco de dados Famílias de Colunas

Banco de dados baseados em Famílias de Colunas são desenvolvidos para abranger três áreas: número enorme de colunas, a natureza esparsa dos dados e frequentes mudanças no esquema. Os dados em Famílias de colunas são armazenados em colunas de forma contínua, enquanto que em bancos de dados relacionais as linhas é que são contínuas. Essa mudança faz com que operações como agregação, suporte para ad-hoc e consultas dinâmicas se tornem mais eficientes (Kaur; Rani, 2013).

A maioria dos bancos de dados baseados em Famílias de Colunas são também compatíveis com o framework MapReduce, no qual acelera o processamento de enorme volume de dados pela distribuição do problema em um grande número de sistemas. Os bancos de dados de Família de Colunas open-source mais populares são Hypertable, HBase e Cassandra (Kaur; Rani, 2013).

2.2.3 Banco de dados Baseado em Grafos

Bancos de dados baseado em Grafos são como uma estrutura de rede contendo nós e arestas, onde as arestas interligam os nós representando a relação entre eles. Comparando com o modelo Entidade-Relacionamento, o nó corresponde à entidade, a propriedade do nó à um atributo, a relação entre as entidades ao relacionamento entre os nós. Nos bancos de dados relacionais as consultas requerem atributos de mais de uma tabela resultando numa operação de junção, por outro lado, bancos de dados baseado em Grafos são desenvolvidos para encontrar relações dentro de uma enorme quantidade de dados rapidamente, tendo em vista que não é preciso fazer junções, ao invés disso, ele fornece indexação livre de adjacência (Kaur; Rani, 2013).

2.2.4 Banco de dados Chave-Valor

Em Bancos de dados Chave-Valor os dados são organizados como uma associação de vetores de entrada consistindo em pares de chave-valor. Cada chave é única e é usada para recuperar os valores associados a ele. Esses bancos de dados podem ser visualizados como um banco de dados relacional contendo múltiplas linhas e apenas duas colunas: chave e valor. Buscas baseadas em chaves resultam num baixo tempo de execução, além disso, os valores podem ser qualquer coisa como objetos, hashes, entre outros (Kaur; Rani, 2013).

Os bancos de dados Chave-Valor mais populares são Riak, Voldemort e Redis (Kaur; Rani, 2013).

2.3 Frameworks

Frameworks podem ser considerados aglomerados de softwares, onde estes são capazes de serem estendidos e adaptados para utilidades específicas (Taligent, 1994). Pree and Sikora (1997), consideram que *frameworks* são aplicações semi-completas e que podem ser reutilizadas para especializar produtos de software customizados. Sommerville (2013), ressalta que *framework* é uma estrutura genérica estendida com o intuito de criar uma aplicação mais específica e Schimidt et al. (2004 [livro sommerville pg300]) define como sendo um conjunto de artefatos de software (como classes, objetos e componentes) que colaboram para fornecer uma arquitetura reusável.

Os *frameworks* possibilitam a reusabilidade de projeto, bem como ao reúso de classes específicas, pois fornecem uma arquitetura de esqueleto para a aplicação, que é definida por classes de objetos e suas interações. As classes são reusadas diretamente e podem ser estendidas usando-se recursos, como a herança (Sommerville, 2013).

Fayad e Schmidt (1997), separam os *frameworks* em três principais classes: de infraestrutura de sistema, de integração de *middleware* e de aplicações corporativas. *Frameworks* de infraestrutura de sistema apoiam o desenvolvimento de infraestruturas, como comunicações, interfaces de usuários e compiladores. Já os *frameworks* de integração de *middleware* são um conjunto de normas e classes de objetos associados que suportam componentes de comunicação e troca de informações. E finalmente, os *frameworks* de aplicações corporativas estão

relacionados com domínios de aplicação específicos, como sistemas financeiros. Eles incorporam conhecimentos sobre o domínios de aplicações e apoiam o desenvolvimento para o usuário final.

Muitas vezes, os *frameworks* são implementações de padrões de projeto, como por exemplo o *framework* MVC. A natureza geral dos padrões e o uso de classes abstratas e concretas permitem a extensibilidade (Sommerville, 2013).

Para estender um *framework*

2.4 Estudo Experimental de Software

Segundo Travassos (2002), a experimentação é o centro do processo científico, por meio dos experimentos que é possível verificar teorias, explorar fatores críticos e formular novas teorias. O autor reforça ainda a necessidade de avaliar novas invenções e sugestões em comparação com as existentes.

Para Wohlin00, existem quatro métodos relevantes para experimentação em Engenharia de Software: científico, de engenharia, experimental e analítico.

O paradigma indutivo, ou método científico, observa o mundo, pode ser utilizado quando se quer entender o processo, produto de software, ambiente. Ele mede e analisa, verifica as hipóteses do modelo ou teoria. Já o método de engenharia observa as soluções existentes, é uma abordagem baseada na melhoria evolutiva, modifica modelos de processos ou produtos de softwares existentes com propósito de melhorar os objetos de estudo. O método experimental é uma abordagem baseada na melhoria revolucionária. Ela sugere um modelo, não necessariamente baseado em um existente, aplica o método qualitativo e/ou quantitativo, faz a experimentação, analisa e repete o processo. Por fim, o método analítico sugere uma teoria formal, é um método dedutivo que oferece uma base analítica para o desenvolvimento de modelos (Travassos, 2002).

Travassos (2002) sugere que a abordagem mais apropriada para a experimentação na área de Engenharia de Software seja o método experimental, pois considera a proposição e avaliação do modelo com os estudos experimentais.

Os principais objetivos relacionados à execução de um estudo experimental de software são: caracterização, avaliação, previsão, controle e melhoria a respeito de produtos, processos, recursos, modelos e teorias.

Os elementos principais do experimento são: as variáveis, os objetos, os participantes, o contexto do experimento, hipóteses e o tipo de projeto do experimento.

Esta pesquisa de dissertação considera a execução do estudo experimental de software para caracterizar, avaliar e propor melhorias ao framework ETL4NoSQL. O objetivo principal da aplicação do experimento é definir se o framework proposto é uma ferramenta adequada para auxiliar no desenvolvimento de processos de ETL em dados estruturados, semi estruturados e não estruturados. Os participantes escolhidos foram as principais ferramentas de ETL encontradas na literatura. Os questionários utilizados para a coleta de dados são baseadas nos requisitos mínimos considerados pela literatura para ferramentas de ETL.

2.5 Trabalhos Correlatos

Esta seção aborda os trabalhos que são correlatos a esta pesquisa, bem como descreve como estes trabalhos diferem do realizado por esta pesquisa.

ARKTOS II: modela os processos de ETL
ETLMR: lida com os processos de ETL utilizando MapReduce
PygramETL:
CloudETL:
P-ETL:
Big-ETL: Foca na paralelização e distribuição.
FramETL
Pentaho
Talend Studio for Data Integration
CloverETL
Oracle Data Integrator (ODI)

CAPÍTULO 3

O Framework ETL4NoSQL

Neste capítulo são apresentados os conceitos do framework ETL4NoSQL, que consiste numa plataforma de software para desenvolvimento de sistemas de ETL, mais especificamente uma ferramenta que auxilia a construção de processos de ETL buscando apoiar a modelagem e o desempenho dos processos.

O ETL4NoSQL oferece um ambiente integrado para modelar processos de ETL e implementar funcionalidades utilizando uma linguagem de programação independente de uma GUI (*Graphical User Interface* - Interface Gráfica do Usuário).

Para a especificação do framework proposto foram definidas as estruturas de dados dos ambientes de origem, destino e da área de processamento de dados e suas respectivas linguagens de manipulação de dados, e também, as principais funcionalidades dos sistemas de ETL, chamados mecanismos de ETL. Para realizar os processos de ETL, por meio de seus mecanismos, foi definido um controlador de operações que é capaz de se comunicar com os ambientes e os mecanismos de ETL.

A seguir, são detalhados os requisitos de software, a arquitetura do sistema e a estrutura dos componentes utilizados no desenvolvimento do framework.

3.1 Requisitos de software do ETL4NoSQL

Requisitos de software são descrições de como o sistema deve se comportar, definidos durante as fases iniciais do desenvolvimento do sistema como uma especificação do que deveria ser implementado (SOMMERVILLE, 1997). Os requisitos podem ser divididos em funcionais e não funcionais, onde o primeiro descrevem o que o sistema deve fazer, ou seja, as transformações a serem realizadas nas entradas de um sistema, a fim de que se produzam saídas, já o outro expressa as características que este software vai apresentar.(SOMMERVILLE e SAWYER, 1997).

O ETL4NoSQL é um framework que tem como principal objetivo auxiliar na criação de processos de ETL ao se utilizar diversas estruturas de armazenamento de dados. Um sistema de software pode ter seus dados armazenados em bases relacionais, que seguem o modelo entidade e relacionamento, ou não relacionais, onde esta possui pouca definição de esquema, não segue um modelo específico e são regularmente chamados de NoSQL. As bases NoSQL possuem quatro paradigmas frequentemente utilizados: Chave-Valor, Família de Colunas, Documentos e Grafo.

As bases de dados relacionais utilizam uma linguagem de gerenciamento de dados padrão conhecida por SQL (Structure Query Language), porém as bases de dados NoSQL não possuem uma linguagem em comum, como as relacionais, cada estrutura de armazenamento possui sua própria linguagem de gerenciamento de dados. Por isso, é essencial que haja um mecanismo que integre a leitura e escrita dos diversos SGBDs NoSQL.

Outra importante características são os processos de ETL que possuem quatro etapas básicas: extração, limpeza/transformação e carga (Kimball and Caserta, 2004). O fluxo do processo de ETL inicia-se com a extração dos dados a partir de uma fonte, que podem ser bases de dados relacionais, bases NoSQL ou arquivos textuais. A partir da extração, os dados passam para uma Área de Processamento de Dados (APD), onde é possível executar processos de limpeza e transformação por meio de mecanismos de junção, filtro, união, agregação e outros. Finalmente, os dados podem ser carregados em estrutura de dados como repositórios analíticos, data warehouses, ou até mesmo em arquivos Linguagem de Marcação Flexível (XML).

Dessa forma, o ETL4NoSQL possui um ambiente que importa os dados dos diversos SGBDs NoSQL, de arquivos textuais, além dos SGBDs relacionais, e que faz a leitura e escrita dos dados permitindo a execução dos processos de ETL. No quadro 3.1 é apresentado os principais requisitos elencados do ETL4NoSQL. Foi definido como importante as prioridades que são imprescindíveis para o desenvolvimento e funcionamento do framework, e desejável as funcionalidades que aprimoram o uso do framework, porém não interferem no seu principal objetivo.

O modelo de processo do funcionamento da ferramenta ETL4NoSQL, baseado nas notações da UML 2.0, é representado na figura 3.1. Esse modelo descreve o processamento dos dados nas atividades de identificação dos dados, obtenção das informações para a importação e o mapeamento dos dados para os esquemas desejados, e também, a atividade dos processos de ETL para por fim dar carga dos dados em DWs, repositórios analíticos ou em arquivos XML.

Outro modelo importante para o entendimento do fluxo de processos da ferramenta ETL4NoSQL é o diagrama de atividades, que de acordo com a UML 2.0 tem como objetivo mostrar o fluxo de atividades em um único processo. O diagrama mostra como um atividade depende uma da

Quadro 3.1 Requisitos do ETL4NoSQL

Requisito	Prioridade
O sistema deve importar os dados de diversas bases relacionais e não relacionais	Importante
O sistema deve permitir a leitura e escrita dos dados importados	Importante
O sistema deve permitir mapear os dados no modelo relacional	Importante
O sistema deve permitir mapear os dados em quaisquer modelo desejado pelo usuário	Importante
O sistema deve possuir os mecanismos ETL mais conhecidos na literatura	Importante
O sistema deve possibilitar a criação de novos mecanismos ETL desejado pelo usuário	Importante
O sistema deve possuir um ambiente que possibilite a execução dos mecanismos de ETL em operações	Importante
O sistema deve permitir o reutilização dos seus mecanismos para vários cenários	Importante
O sistema deve permitir processamento distribuído	Desejável
O sistema deve permitir a importação de dados a partir de uma nuvem	Desejável

outra. Na figura 3.2 o diagrama mostra a interação dos componentes ao executar um processo de ETL, onde o estágio inicial é a importação dos dados seguido pelo mapeamento, após a obtenção dos dados necessários é possível a execução dos diversos processos de ETL em uma área de processamento para finalmente os dados serem exportados para base de destino.

3.2 Arquitetura do ETL4NoSQL

Sommerville (2007), define o projeto de arquitetura como um processo criativo em que se tenta organizar o sistema de acordo com os requisitos funcionais e não funcionais. Um estilo de arquitetura é um padrão de organização de sistema (Garlan e Shaw, 1993; Sommerville, 2007), como uma organização cliente-servidor ou uma arquitetura em camadas. Porém, a arquitetura não necessariamente utilizará apenas um estilo, a maioria dos sistemas de médio e grande porte utilizam vários estilos. Para Garlan e Shaw, há três questões a serem definidas na escolha do projeto de arquitetura, a primeira é a escolha da estrutura, cliente-servidor ou em camadas, que permita atender melhor aos requisitos. A segunda questão é a respeito da decomposição dos subsistemas em módulos ou em componentes. E por fim, deve-se tomar a decisão de sobre como a execução dos subsistemas é controlada. A descrição da arquitetura pode ser representada graficamente utilizando modelos informais e notações como a UML (Clements, et al., 2002; Sommerville, 2007).

A arquitetura do ETL4NoSQL, representada graficamente na figura 3.3, é baseada no requisito de reutilização. A possibilidade do reuso, reduz o trabalho repetitivo na implementação

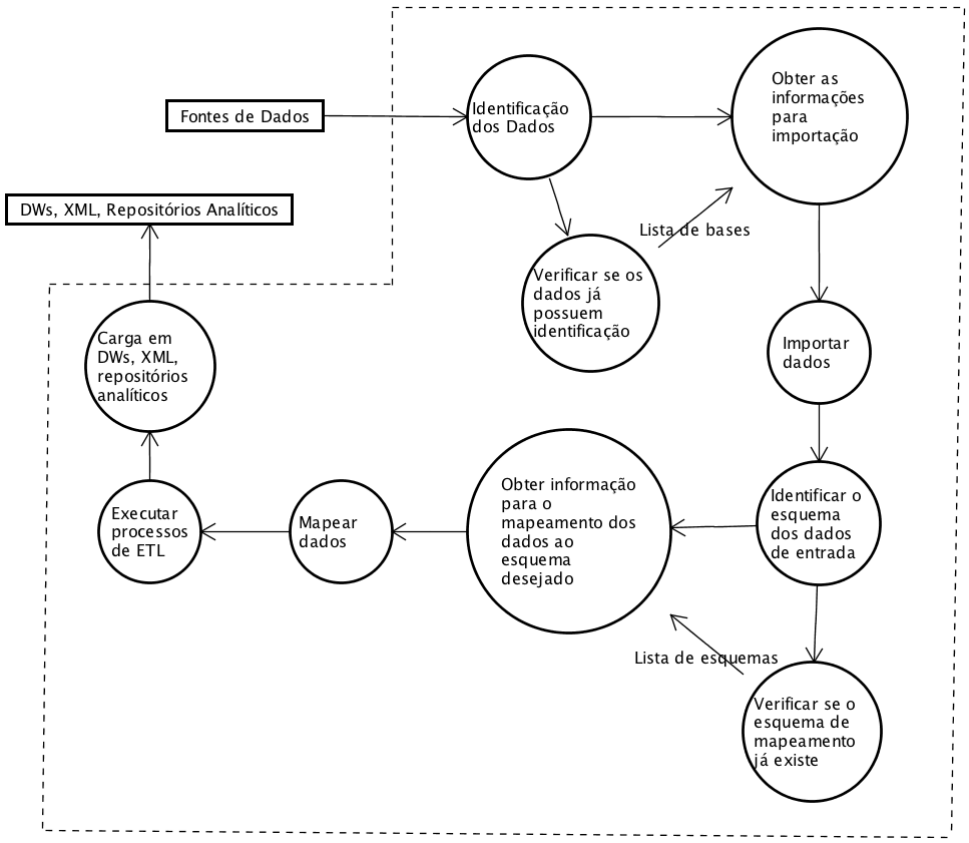

Figura 3.1 Modelo de Processos do ETL4NoSQL

de componentes e o custo de manutenção (Szyperski, et al.2002), e sua estrutura é em camadas, onde há a camada de sistema e a camada de interface. A camada de sistema lida com todas as operações internas e a camada de interface faz toda a interligação do sistema com o ambiente externo. A decomposição dos subsistemas do ETL4NoSQL é em componentes, pois componentes podem ser subsistemas ou simples objetos que podem ser reusados (Sommerville, 2007). Os componentes que integram o framework e representados na figura 3.3 são os componentes de importação, mapeamento, mecanismos ETL e Operações. Estes componentes serão melhor detalhados na seção seguinte.

3.3 Componentes do ETL4NoSQL

A engenharia de software baseada em componentes é uma abordagem fundamentada em reuso para desenvolvimento de sistemas de software, ela envolve o processo de definição, implementação e integração ou composição de componentes independentes não firmemente acoplados ao sistema. Os componentes são independentes, ou seja, não interferem na operação uns dos outros e se comunicam por meio de interfaces bem definidas, os detalhes de implementação são ocultados, de forma que as alterações de implementação não afetam o restante

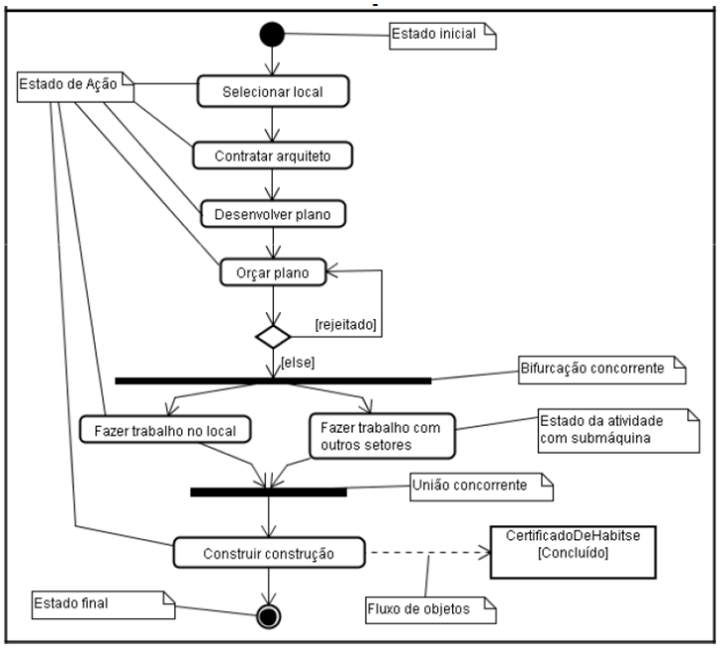

Figura 3.2 Diagrama de Atividades do ETL4NoSQL

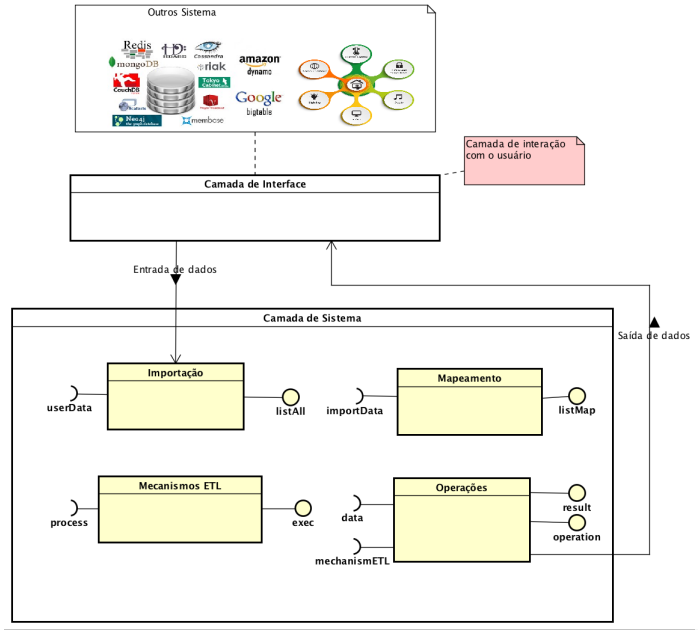

Figura 3.3 Arquitetura do Framework ETL4NoSQL

do sistema (Sommerville, 2007). Segundo [15], componentes são uma parte do sistema de software que podem ser identificados e reutilizados, onde descrevem ou executam funções específicas e possuem interfaces claras, documentação apropriada e a possibilidade de reuso bem definida. Ainda de acordo com o autor, um componente deve ser autocontido, identificável,

funcional, possuir uma interface, ser documentado e ter uma condição de reuso.

De acordo com os requisitos do ETL4NoSQL, foi possível identificar quatro importantes funcionalidades que podem ser definidas como componentes do sistema, a funcionalidade de importação, mapeamento de dados, mecanismos de ETL e o controlador de operações. Os componentes do ETL4NoSQL e suas características são apresentados nas seções seguintes, seguindo as características de componentes adotadas por [9].

3.3.1 Componente de Importação

Um dos objetivos do framework ETL4NoSQL é possibilitar a integração de várias estruturas de dados, relacionais ou não relacionais, presentes nos sistemas modernos. Para isso, a ferramenta deve permitir a leitura e escrita dos diversos SGBDs existentes que aplicam essas estruturas. A solução encontrada para isso foi desenvolver um componente programável que possibilite a importação dos dados por meio de inserção de parâmetros em linha de comando. Este componente, por ser criado utilizando o paradigma de orientação a objetos, permite também sua extensão, por meio de especialização, para que atenda a especificidade de cada cenário. As características do componente são apresentadas a seguir.

a) Interface: Componente responsável pela importação dos dados da base origem.

b) Nomeação: Import.

c) Metadados: Este componente contém as informações da base origem como a linguagem de manipulação de dados e meios para estabelecer a conexão com a base, requer uma interação com a interface para o usuário disponibilizar as informações e fornece os dados importados para outros componentes.

d) Interoperabilidade: Oferece comunicação com outros componentes por meio dos métodos listAll e userData.

e) Customização: Este componente permite customizar as formas de apresentar os dados importados, de acordo com a necessidade de cada sistema.

f) Suporte a evolução: Possibilita o suporte aos métodos de acordo com as mudanças de conexões e manipulações de bases de dados futuras.

g) Empacotamento e utilização: Os métodos são encapsulados e podem ser utilizados pela importação de sua classe e a interface com o usuário é por meio de linha de comando.

3.3.2 Componente de Mapeamento

Para viabilizar a organização dos dados em vários tipos de esquemas desejáveis pelo usuário o ETL4NoSQL oferece o componente de mapeamento. Este componente permite definir o esquema dos dados de acordo com a necessidade da aplicação almejada pelo usuário. Por meio de parâmetros de inserção em linha de comando é possível utilizar os esquemas de dados pré-definidos pelo componente, mas também, por utilizar o paradigma de orientação a objetos e

as características de reusabilidade dos componentes, é possível especializar e customizar os esquemas conforme a conveniência do usuário.

a) Interface: Componente responsável por gerar o mapeamento dos dados oferecidos pelo componente de importação para um esquema relacional.

b) Nomeação: Map.

c) Metadados: Este componente requer os dados de uma base de dados para efetuar o mapeamento.

d) Interoperabilidade: Oferece comunicação com outros componentes por meio dos métodos importData e listMap.

e) Customização: É possível customizar as regras de mapeamento para outros esquemas de dados.

f) Suporte a evolução: Possibilita o suporte aos métodos de acordo com a necessidade de alterar os esquemas dos dados.

g) Empacotamento e utilização: Os métodos são encapsulados e podem ser utilizados pela importação de sua classe e a interface com o usuário é por meio de linha de comando.

3.3.3 Componente de Mecanismos de ETL

O ETL4NoSQL é um framework de ETL que possibilita a integração de várias estruturas de dados, por isso ele deve apresentar mecanismos que viabilizem as principais operações de ETL conhecidas pela literatura. Dessa forma, para disponibilizar as operações de ETL, o ETL4NoSQL possui um componente de mecanismos de ETL que permite executar processos de ETL como extração, limpeza/transformação e carga de dados. Além das operações básicas de ETL, o componente permite a especialização e criação de mecanismos permitindo a customização das operações de ETL conforme a necessidade do usuário.

a) Interface: Componente que contém métodos que realizam as principais operações de ETL presentes na literatura.

b) Nomeação: MechanismETL.

c) Metadados: Este componente requer dados de controle para realizar as operações por meio de seus métodos.

d) Interoperabilidade: Oferece comunicação com outros componentes por meio dos métodos exec e process.

e) Customização: É possível customizar e criar mecanismos de acordo com a necessidade de cada processo de ETL.

f) Suporte a evolução: Deve possibilitar o suporte aos métodos de acordo com a necessidade de alterar os esquemas dos dados.

g) Empacotamento e utilização: Os métodos deverão ser encapsulados e poderão ser utilizados pela importação de sua classe e a interface com o usuário será por meio de linha de comando.

3.3.4 Componente de Operações

Para proporcionar o controle dos processos de ETL executados pelo framework, o ETL4NoSQL possui o componente de operações. Este componente é responsável pelo controle das operações dos processos de ETL, ele assegura a execução dos mecanismos de ETL de acordo com a necessidade do usuário. É possível também, customizar e especializar as operações deste componente.

a) Interface: Componente responsável por criar e executar processos de ETL.

b) Nomeação: Componente de Operação.

c) Metadados: Este componente deverá possibilitar a comunicação com o componente de mecanismos de ETL e deverá criar e executar processos de ETL.

d) Interoperabilidade: Deve possibilitar a comunicação entre outros componentes.

e) Customização: É possível customizar os processos de ETL criados.

f) Suporte a evolução: Deve possibilitar o suporte aos métodos de acordo com a necessidade de alterar os processos.

g) Empacotamento e utilização: Os métodos deverão ser encapsulados e poderão ser utilizados pela importação de sua classe e a interface com o usuário será por meio de linha de comando.

3.4 Considerações Finais

CAPÍTULO 4

Estudo Experimental de Software

Este capítulo provê o roteiro de experimentação de software para ferramentas de ETL utilizando dados estruturados, semi estruturados e não estruturados. A Engenharia de Software Experimental tem como objetivo aprimorar métodos, técnicas e ferramentas de Engenharia de Software a partir de métodos experimentais (Isaque Elcio de Souza, TESE - Um sistema de Inf para Geren de Projetos Experimentais em ES). As etapas definidas no processo de experimentação em Engenharia de Software proposto por [Amaral (),Isaque Elcio de Souza, TESE] consiste em etapas de definição, planejamento, operação, interpretação dos dados e empacotamento que serão melhor detalhados nas seções a seguir.

4.1 Objetivos do experimento

O objetivo principal da aplicação deste experimento é definir se o framework proposto por esta pesquisa de dissertação é uma ferramenta adequada para auxiliar no desenvolvimento de processos de ETL em dados estruturados, semi estruturados e não estruturados.

4.1.1 Objetivo da Medição

Tendo como base as ferramentas de ETL existentes na literatura, caracterizar:

1. Quais as principais funcionalidades que as ferramentas de ETL oferecem:
   (a) essas funcionalidades manipulam dados estruturados, semi estruturados e não estruturados.
   (b) essas funcionalidades não manipulam dados estruturados, semi estruturados e não estruturados.
2. Quais funcionalidades podem ser consideradas fundamentais para a produtividade na criação de processos de ETL:
   (a) quais necessitam manipular dados em grande escala.
   (b) quais não manipulam grande volume de dados.
3. Quais funcionalidades poderiam aprimorar as ferramentas de ETL.

4.1.2 Objetivos do Estudo

- Analisar as ferramentas de ETL para dados estruturados, semi estruturados e não estruturados;
- Com o propósito de caracterizar;
- Com respeito à intersecção das ferramentas de ETL existente;
- Do ponto de vista da literatura;
- No contexto de comparativo entre as ferramentas mais conhecidas no mercado atual.

4.1.3 Questões

Q1. Existem funcionalidades listadas pelas ferramentas pesquisadas que não estão presentes no ETL4NoSQL?

Métrica: A lista de funcionalidades que não estão presentes no ETL4NoSQL.

Q2. Existem funcionalidades oferecidas pelo ETL4NoSQL que não estão presentes nas ferramentas apresentadas pela literatura?

Métrica: A lista de funcionalidades que não estão presentes nas ferramentas da literatura.

Q3. Existem funcionalidades que não estão presentes no ETL4NoSQL e nas ferramentas da literatura que poderiam ser implementadas?

Métrica: A lista de funcionalidades que não estão presentes em nenhuma das ferramentas.

4.2 Planejamento

Na etapa de planejamento são definidas as hipóteses do estudo, a descrição da instrumentação, as métricas, seleção do contexto e dos indivíduos, as variáveis, a análise qualitativa e a validade do experimento. Todas elas serão descritas nas seções seguintes.

4.2.1 Definição das Hipóteses

Hipótese nula (H0): As funcionalidades oferecidas pelo ETL4NoSQL são similares às funcionalidades oferecidas pelas ferramentas presentes na literatura.

Fp - Funcionalidades do ETL4NoSQL

Fl - Funcionalidades das ferramentas da literatura

H0: Fl - (Fp $\cap$ Fl) = $\emptyset$

Hipótese alternativa (H1): A lista de funcionalidades oferecidas pelo ETL4NoSQL é diferente da lista de funcionalidades oferecidas pelas ferramentas presentes na literatura.

Fp - Funcionalidades do ETL4NoSQL

Fl - Funcionalidades das ferramentas da literatura

H1: Fl - (Fp $\cap$ Fl) $\neq \emptyset$

Hipótese alternativa (H2): A lista de funcionalidades que poderiam ser implementadas é diferente da lista de funcionalidades oferecidas pelas ferramentas na literatura e pelo ETL4NoSQL.

Fp - Funcionalidades do ETL4NoSQL

Fl - Funcionalidades das ferramentas da literatura

Fi - Funcionalidades que poderiam ser implementadas

H2: Fi - (Fp $\cap$ Fl $\cap$ Fi) $\neq \emptyset$

4.2.2 Descrição da instrumentação

Para cada funcionalidade presente nas ferramentas apresentada na literatura que são consideradas fundamentais para o funcionamento dos processos de ETL pode ser encontrada no quadro 4.1:

Para cada funcionalidade aplicar teste estatístico Chi-2 para definir:

se pode considerar que essa funcionalidade é fornecida;

se pode considerar que essa funcionalidade é útil;

se pode considerar que essa funcionalidade necessita de melhoria.

Resultado: N funcionalidades com valores (P; M; U) onde P - presença 0 - não presente; 1 - presente; U - utilidade 0 - não é útil; 1 - é útil; melhoria 0 - não necessita melhorar; 1 - necessita melhorar.

Quadro 4.1 Descrição da Instrumentação

Presença da Funcionalidade (P)	Melhoria da Funcionalidade (M)	Utilidade da Funcionalidade (U)
1. Não está presente 2. Está presente parcialmente 3. Está presente	1. Necessita melhorar 2. Não há necessidade de melhoria 3. Pode melhorar, mas não necessidade	1. É útil 2. Não é útil 3. É parcialmente útil

4.2.3 Métricas

Na tabela 4.2 são apresentadas as métricas utilizadas neste experimento.

Quadro 4.2 Métricas

No	P	M	U	Descrição da Funcionalidade	Questões
1	0	0	0	Não está presente, não necessita melhorar, não é útil	N/A
2	0	0	1	Não está presente, não necessita melhorar, é útil	Q3
3	0	1	0	Não está presente, necessita melhorar, não é útil	N/A
4	0	1	1	Não está presente, necessita melhorar, é útil	Q3
5	1	0	0	Está presente, não necessita melhorar, não é útil	Q1, Q2
6	1	0	1	Está presente, não necessita melhorar, é útil	Q1, Q2
7	1	1	0	Está presente, necessita melhorar, não é útil	Q1, Q2
8	1	1	1	Está presente, necessita melhorar, é útil	Q1, Q2

4.2.4 Seleção do contexto

De acordo com Travassos (2002), o contexto pode ser caracterizado conforme quatro dimensões:

- o processo: on-line / off-line;
- os participantes: ferramentas de ETL;
- realidade: o problema real / modelado;
- generalidade: específico / geral.

Nosso estudo supõe o processo off-line porque as ferramentas não estão sendo testadas durante todo o tempo da utilização, mas em certo instante. Os participantes são as ferramentas de

ETL encontradas na literatura. O estudo é modelado porque as funcionalidades das ferramentas não são caracterizadas durante a resolução do problema real, mas utilizando parâmetros subjetivos (ex. presença, utilidade e necessidade). As funcionalidades do ETL4NoSQL são comparadas com as ferramentas presentes na literatura, então, o contexto possui o caráter específico.

4.2.5 Seleção dos indivíduos

Como participantes para o estudo propõe-se utilizar as ferramentas encontradas na literatura. Assume-se que esses indivíduos estão presente em diversos estudos realizados e avaliados no meio acadêmico.

Para a escolha das ferramentas utilizadas neste estudo foi levado em consideração a semelhança da finalidade do uso com a ferramenta proposta. Seria conveniente utilizar para o estudo ferramentas que tem o objetivo de auxiliar processos de ETL em diversas estruturas de dados. Dessa forma, a seleção baseou-se nas características das ferramentas.

4.2.6 Variáveis

Variável independente: A lista de funcionalidades das ferramentas encontradas na literatura.
Variáveis dependentes:

1. A similaridade entre as funcionalidades oferecidas pela ferramenta proposta e as funcionalidades encontradas nas ferramentas da literatura.

   Pode receber os valores: Igual, quando todas as funcionalidades tem o valor PMU = { 1, X, X } (métricas 5-8); Diferente, quando todas as funcionalidades tem o valor PMU = { 0, X, X } (métricas 1-4) Similar, quando não se cumprem as condições de "Igual"e "Diferente". O grau de similaridade pode ser avaliado como: { 1, X, X } / { 0, X, X } + { 1, X, X } * 100%

2. A utilidade das funcionalidades similares. Mostra a parte útil das funcionalidades oferecidas pela ferramenta proposta: Parte útil: { 1, X, 1 } / { 1, X, X } * 100% Parte inútil: { 1, X, 0 } / { 1, X, X } * 100%

3. A melhoria das funcionalidades similares. Mostra a necessidade de melhoria nas funcionalidades oferecidas pela ferramenta proposta: Não necessita melhorar: { 1, 0, X } / { 1, X, X } * 100% Necessita melhorar: { 1, 0, X } / { 1, X, X } * 100%

4.2.7 Análise Qualitativa

Para analisar a informação referente às funcionalidades não oferecidas no ETL4NoSQL, mas que poderiam ser implementadas, propõe-se aplicar a análise qualitativa. Essa análise deve apresentar a lista de funcionalidades presentes nas ferramentas da literatura, que não estão presentes na ferramenta proposta, mas que são consideradas necessárias para facilitar a manipulação de dados estruturados, semi estruturados e não estruturados. Assim, essa análise deve

considerar funcionalidades com valor PMU = 0, X, X (métricas 1-4) e a opção ”É útil” para ”utilidade da funcionalidade”.

4.2.8 Validade

Validade interna: como mencionado na parte "Seleção dos indivíduos"para o estudo se propõe a utilizar ferramentas presentes na literatura, que são validadas pelo meio acadêmico. Assim, assume-se que elas são representativas para a população de ferramentas de ETL.

Além disso, para redução da influência dos fatores que não são interesse do nosso estudo e, portanto, para aumento da validade interna do estudo supõe-se utilizar dados das ferramentas mais populares da literatura, cuja a validação já tenha passado por diversas avaliações.

Validade de conclusão: para receber os valores da presença, utilidade e melhorias o teste binomial será utilizado. A verificação de hipótese será feita por meio de simples demonstração de presença ou não de funcionalidades nas listas que representam as variáveis independentes.

Validade de construção: esse estudo está caracterizado pela conformidade das funcionalidades listadas na ferramenta proposta com as funcionalidades reais necessárias para a utilização de ferramentas de ETL. As características das ferramentas de ETL presentes na literatura representa a lista de funcionalidades que uma ferramenta de ETL deve apresentar para mostrar o desempenho adequado do ponto de vista da literatura. As funcionalidades, que tem o maior relacionamento com as ferramentas de ETL do ponto de vista dos pesquisadores, foram escolhidas do conjunto total de funcionalidades das ferramentas de ETL presentes na literatura.

Validade externa: como foi mencionado nas partes "Seleção dos indivíduos"e "Validade interna"os participantes do estudo em geral podem ser considerados representativos para a população da literatura apresentada pela academia. Para avaliação do nível de importância das funcionalidades analisadas foi levada em consideração a frequência que a funcionalidade apareceu nas ferramentas da literatura.

Os materiais utilizados no estudo podem ser considerados representativos e "em tempo"para o problema sob análise, porque se compõem das funcionalidades de ferramentas de ETL presentes na literatura atual.

4.3 Operação

A etapa de operação ocorre após a etapa de planejamento do estudo experimental. Nela é exercido o monitoramento do experimento para garantir que ele esteja ocorrendo conforme foi planejado (Souza Isaque, 2015). Nesta seção serão apresentados os questionários do perfil da ferramenta de ETL e o de Funcionalidades.

4.3.1 Questionário do Perfil da Ferramenta de ETL

O quadro 4.3 mostra as questões usadas para definir o perfil das ferramentas utilizadas como indivíduos deste experimento.

Quadro 4.3 Questionário do Perfil da Ferramenta de ETL

Nome da ferramenta de ETL:	
Possui código aberto?	Sim ◯ Não ◯
Possui uma marca reconhecida no mercado?	Sim ◯ Não ◯
Tem como finalidade utilizar bancos de dados NoSQL?	Sim ◯ Não ◯
Possui interface gráfica?	Sim ◯ Não ◯
É programável?	Sim ◯ Não ◯
É integrada?	Sim ◯ Não ◯
Qual o tipo de processamento que a ferramenta executa?	Distribuído ◯ Centralizada ◯ Híbrido ◯
É extensível?	Sim ◯ Não ◯
Para qual finalidade a ferramenta procura auxiliar melhor os processos de ETL?	Modelagem ◯ Desempenho ◯

4.3.2 Questionário de Funcionalidades

Sob o ponto de vista das características das ferramentas e considerando a finalidade da ferramenta indicada acima, avalie as colunas correspondentes segundo as escalas abaixo, a presença, utilidade e melhorias quanto às funcionalidades das ferramentas apresentadas nos seus respectivos trabalhos de pesquisa, das funcionalidades listadas no questionário:

Quadro 4.4 Instrumentação para aplicar o questionário

Presença da Funcionalidade (P)	Melhoria da Funcionalidade (M)	Utilidade da Funcionalidade (U)
1. Não está presente 2. Está presente parcialmente 3. Está presente	1. Necessita melhorar 2. Não há necessidade de melhoria 3. Pode melhorar, mas não necessidade	1. É útil 2. Não é útil 3. É parcialmente útil

4.4 Resultado do Estudo

4.5 Descrição

4.6 Considerações Finais

N	Funcionalidade	Descrição	Presença			Utilidade			Melhoria		
			1	2	3	1	2	3	1	2	3
	Processo										
1	Leitura e escrita NoSQL										
2	Saída no formato relacional e não relacional										
3	Processamento em nuvem										

Figura 4.1 Questionário de Funcionalidades

CAPÍTULO 5

Considerações Finais

5.1 Principais Contribuições

5.2 Discussão

5.3 Resultados

5.4 Trabalhos Futuros

Referências Bibliográficas

[1] S. Chaudhuri and U. Dayal. An overview of data warehousing and olap technology. *SIG- MODRecord,v.26,n.1,p.65â74*, 1997.

[2] Max Chevalier, SMohammed El Malki, Arlind Kopliku, Olivier Teste, and Ronan Tournier. Implementing multidimensional data warehouses into nosql. *Proceedings of the 17th International Conference on Enterprise Information Systems, p.172-183, April 27-30*, 2015.

[3] Jerome Darmont, Omar Boussaid, AJean christian Ralaivao, and Kamel Aouiche. An architecture framework for complex data warehouses. *7th International Conference on Enterprise Information Systems (ICEISâ05), Miami, USA, pages 370â373*, 2005.

[4] Lucas de Carvalho Scabora. AvaliaÃ§Ã£o do star schema benchmark aplicado a bancos de dados nosql distribuÃdos e orientados a colunas. Master's thesis, USP - SÃ£o Carlos, April 2016.

[5] Mário Sergio da Silva. Um framework para desenvolvimento de sistemas etl. Master's thesis, Universidade Federal de Pernambuco, September 2012.

[6] FAYAD M. E. and SCHMIDT D. C. Object-oriented application frameworks. *Communica- tions of the ACM 40 (10): 32â38*, 1997.

[7] FAYAD M. E., SCHMIDT D. C., and JOHNSON R. E. Building application frameworks: object-oriented foundations of framework design. *John Wiley and Sons, New York, USA, pp. 3â29*, 1999.

[8] Martin Fowler and Pramod J. Sadalage. *NoSQL Essencial Um guia conciso para o mundo emergente de PersistÃªncia Poliglota*. Novatec, 2013. First Edition.

[9] George T. Heineman. *Component-Based Software Engineering: putting pieces together*. Addison-Wesley, 2001.

[10] Ralph Kimball and J. Caserta. *The Data Warehouse ETL ToolKit*. Robert Ipsen, 2004.

[11] Ralph Kimball and Margy Ross. *The Data Warehouse Toolkit*. Robert Ipsen, 2002. Second Edition.

[12] Xiufeng Liu, Christian Thomsen, and Torben Bach Pedersen. Cloudetl: Scalable dimensional etl for hive. *DB Tech Reports*, July 2013.

[13] Petter Nasholm. *Extracting Data From NoSQL Databases*. PhD thesis, Chalmers University of Technology, SE-412 96 Goteborg Sweeden, January 2012.

[14] VASSILIADIS P., SIMITSIS A., and GEORGANTAS P. TERROVITISB M. SKIADOPOU-LOS S. A generic and customizable frameworkfor the design of etl scenarios. *Information Systems - Special issue: The 15th international conference on advanced information systems engineering 30 (7): 492â525*, 2005.

[15] Johannes Sametinger. *Software Engineering with Reusable Componets*. Springer, 1997.

[16] BRAGA R. T. V. Um processo para construÃ§Ã£o e instanciaÃ§Ã£o de frameworks baseados em uma linguagem de padrÃµ*esparaumdomnioespecfico*.*Master*′*sthesis*,*ICMC*/*USP*,*S£oPaulo*,2003.