Um ambiente de Data Warehousing para apoiar a tomada de decisão quanto a evasão escolar na UFPE

Gabriel Mac'Hamilton Renaux Alves[1], **Robson do Nascimento Fidalgo**[1]

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Recife, PE – Brasil

{gmhra,rdnf}@cin.ufpe.br

Abstract. *The main objectives of a federal higher education institution (HEI) can be summarized to provide quality public education, academic and professional training for its students for the labor market. Although these goals coincide with those of many Brazilians, the many challenges faced by students cause many to give up on their studies and end up in the situation of dropping out, generating losses for themselves, for public higher education institutions and consequently for the country. An efficient way to manage dropouts and students is to know the data about them. To this end, this work aims to provide a proposal for a Data Warehousing environment that can be used so that the managers of the Universidade Federal de Pernambuco (UFPE) can obtain information about the students and, based on this information, make efficient decisions to reduce school dropout rates. In the development of the present work, anonymized data from the SIG@ UFPE system were used, together with necessary artificial data referring to university students.*

Keywords*: Business intelligence. Data warehouse. Higher Education. School Dropout.*

Resumo. *Os principais objetivos de uma instituição de ensino superior (IES) federal podem ser resumidos no fornecimento de ensino público de qualidade, formação e capacitação de seus estudantes para o mercado de trabalho. Embora tais objetivos coincidam com o de muitos brasileiros, os diversos desafios enfrentados pelos estudantes fazem com que muitos desistam dos estudos e acabem entrando na situação de evasão escolar, gerando prejuízos para si mesmos, para as instituições públicas de formação superior e consequentemente ao país. Um meio eficiente de gerenciar a evasão escolar e os alunos é conhecendo os dados sobre eles. Para tal, este trabalho tem como objetivo fornecer uma proposta de um ambiente de Data Warehousing que poderá ser utilizado para que os gestores da Universidade Federal de Pernambuco (UFPE) possam obter informações sobre os alunos e, baseando-se nessas informações, tomem decisões eficientes para a redução das taxas de evasão escolar. No desenvolvimento do presente trabalho foram utilizados os dados anonimizados do sistema SIG@ UFPE, aliados aos dados artificiais necessários, referentes aos alunos da universidade.*

Palavras-chave*: Business intelligence. Data warehouse. Ensino superior. Evasão escolar.*

1. Introdução

O presente trabalho tem como principal motivação a necessidade da redução da evasão escolar na Universidade Federal de Pernambuco. Tal necessidade é evidenciada por ser destacada nos objetivos do Plano Nacional de Educação (PNE) do Governo Federal e da Política de Assistência Estudantil (PAE) da UFPE, diretrizes que norteiam a universidade. No PNE, elaborado pelo Governo Federal, e aprovado pela da lei nº 13.005 de 25 de junho de 2014 [BRASIL 2014], foi elencado o ponto 12.3 em seu anexo, que busca aumentar a taxa de conclusão média dos cursos de graduação presencial em universidades públicas para 90%, consequentemente visando diminuir a taxa de evasão, que é inversamente proporcional à taxa de conclusão média. Já na UFPE a resolução Nº 15/2019 [UFPE 2019], que regulamenta a PAE, elaborada pelo conselho de ensino pesquisa e extensão da UFPE, traz no artigo 4º seus objetivos, sendo o primeiro deles garantir que os estudantes permaneçam o tempo necessário e concluam os cursos de graduação presencial com qualidade, nos quesitos de formação ampliada, produção de conhecimento, melhoria do desempenho acadêmico e da qualidade de vida, visando reduzir os índices de evasão.

A análise e entendimento dos indicadores da evasão escolar é uma das principais medidas para iniciar a solução do problema, pois com o conhecimento destas informações é possível elencar medidas viáveis para se tentar prevenir casos futuros. Porém, a análise dos dados dos alunos muitas vezes não é feita em tempo hábil, pois envolve diferentes setores da organização, gerando lentidão no processo de tomada de decisão e tornando difícil identificar soluções para o problema da evasão escolar. Diante desta realidade, o presente trabalho propõe o desenvolvimento de um ambiente de Data Warehousing utilizando os dados dos alunos da UFPE, pois com isso poderão ser fornecidos relatórios com informações relativas à evasão escolar para os gestores da universidade, de maneira mais ágil em relação à situação atual, desta forma permitindo a celeridade na tomada de decisão quanto à evasão escolar, consequentemente apoiando na solução do problema.

O presente trabalho tem como objetivo geral demonstrar o desenvolvimento de um ambiente de Data Warehousing para apoiar a gestão dos alunos e da evasão escolar, no contexto da UFPE. Para alcançar o objetivo geral foram elencados os seguintes objetivos específicos:

- Elencar os dados necessários para o desenvolvimento do projeto;
- Descrever o processo de construção do ambiente de Data Warehousing proposto no presente trabalho;
- Desenvolver um esquema estrela para demonstrar como os dados devem ficar armazenados na base de dados do projeto;
- Desenvolver duas ferramentas para demonstrar como pode ser feito o acesso aos dados da área de apresentação de dados do ambiente de Data Warehousing para a geração de relatórios, sendo estas um cubo de dados multidimensional, e um conjunto de painéis de Business Intelligence (BI).

2. Evasão escolar

O processo de evasão escolar tem uma difícil definição, pois é ocasionada e influenciada por diversos fatores relativos ao estudante. O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) define evasão escolar como a saída do aluno, antes

de se ter concluído o ano, série ou ciclo, da instituição de ensino, por questão de desistência do próprio, sem que seja considerado o motivo da desistência, caracterizando insucesso no desenvolvimento das capacidades esperadas ao final da jornada de aprendizado [INEP 2017]. De acordo com a classificação da Universidade Federal de Pernambuco, entra em situação de evasão escolar, todo aluno que está em situação de desligamento da universidade, desvinculado com a universidade, em situação de transferência interna ou em situação de transferência externa [PROPLAN-UFPE 2020].

Os principais meios elencados pela literatura para se calcular a taxa de evasão escolar de uma instituição, levam em consideração um ano em questão, a quantidade de alunos ingressantes, a quantidade de alunos retidos, e a quantidade de alunos evadidos. Dentre os indicadores elencados para apoiar no cumprimento do ponto 12.3 do anexo da lei nº 13.005 de 25 de junho de 2014 [BRASIL 2014] está a taxa de evasão, sendo calculada dividindo a quantidade de matrículas finalizadas evadidas pelas matrículas atendidas, e multiplicando este valor por 100. Já na UFPE, de acordo com um documento da [PROPLAN-UFPE 2020], o cálculo é feito dividindo-se a quantidade de alunos evadidos, pela quantidade de alunos ingressantes, dado um ano (a), e multiplicando esse valor por 100, conforme a Figura 1.

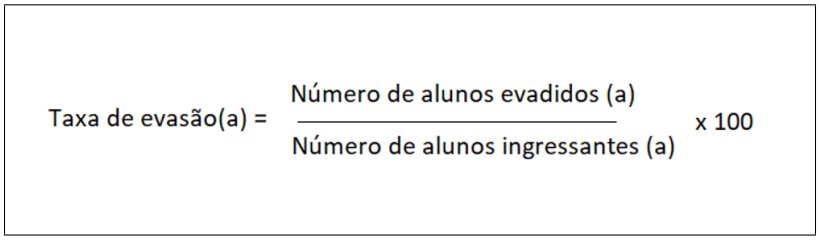

Figura 1. Cálculo da taxa de evasão escolar na UFPE

Existem diversas maneiras de se combater a evasão escolar, desde que se conheçam os dados referentes aos alunos. O Instituto Lobo de Pesquisa e Gestão Educacional indica que a própria instituição de ensino superior pode fazer o cálculo chamado de "Acompanhamento da Coorte", onde se trata dos dados de cada aluno de forma individual, que permite a análise da evasão chegando até o nível de granularidade do aluno. Também é indicado pelo Instituto Lobo que independentemente do método que seja utilizado para calcular a taxa de evasão, o fator essencial é a capacidade de medir a evolução da evasão escolar, para que assim, possam ser definidas as medidas de prevenção e combate. Foram estabelecidos pelo Instituto Lobo sete pontos gerais que contribuem para a redução da taxa de evasão escolar [LOBO 2012]:

1. Estabelecimento de um grupo de trabalho responsável por reduzir a evasão dentro da IES;
2. Avaliação das estatísticas da evasão escolar;
3. Determinação das causas da evasão escolar;
4. Estímulo da visão da IES centrada no aluno para seus colaboradores;
5. Criação de condições que atendam as expectativas que os alunos tinham quando entraram no curso da IES;
6. Tornar e manter o ambiente físico e os meios de trânsito na IES agradáveis para os alunos;
7. Criação de um programa de aconselhamento e orientação para os alunos.

3. Data Warehousing

A seguir serão apresentados os principais conceitos ligados à descrição e ao desenvolvimento de um Data Mart (DM), Data Warehouse (DW) e de um ambiente de Data Warehousing. Um Data Mart consiste na organização, estruturamento e armazenamento dos dados operacionais de um processo de negócio de uma organização, com o objetivo de fornecer uma estrutura onde é possível realizar consultas aos dados de forma facilitada e com alta performance. Já um Data Warehouse, é o conjunto de todos os Data Marts de uma organização, resultando em um conglomerado de dados e informações para o apoio à tomada de decisão. De acordo com [KIMBALL e ROSS 2013] um ambiente de Data Warehousing é composto de quatro partes:

1. Sistemas operacionais fonte (Operational Source Systems):
   Sistemas que fornecem a fonte dos dados que vão compor os Data Marts do ambiente, geralmente são os sistemas transacionais das organizações.
2. Área de preparação dos dados (Data Staging Area):
   É a área de armazenamento prévio dos dados, antes de serem modelados para a modelagem dimensional. Esta área é o espaço intermediário entre os dados brutos da fonte e os dados tratados e modelados para o consumo final, e é onde ocorre o processo de ETL (extração, transformação e carga dos dados).
3. Área de apresentação dos dados (Data Presentation Area):
   Esta é a área onde os dados estão armazenados de forma organizada para que os usuários e outros sistemas possam consultar. É nesta área onde estão todos os Data Marts do ambiente, com seus dados armazenados na forma da modelagem dimensional.
4. Ferramentas de acesso aos dados (Data access tools):
   Ferramentas que permitem acesso aos dados armazenados na área de apresentação.

Segundo [KIMBALL e ROSS 2013], a indústria chegou à conclusão de que a modelagem dimensional é a forma mais adequada de entregar dados para os usuários de sistemas de Data Warehouse devido à sua simplicidade, que garante a facilidade para que os usuários entendam os dados e para que os softwares possam navegar e trazer resultados de maneira performática. A definição de modelagem dimensional consiste em uma técnica de modelagem de bancos de dados que visa torná-lo mais simples e apresentável, para que desta forma se tenha uma maior aceitação dos usuários da área de negócio.

A modelagem dimensional é composta principalmente de medidas, tabelas de fato e dimensões. Nas tabelas de fato são armazenadas as medidas de performance numérica do processo de negócio, além de todas as referências para as tabelas de dimensão, no formato de chaves artificiais, e é através destas chaves que as tabelas de dimensão se conectam com a tabela de fato e umas com as outras. Já nas tabelas de dimensão são armazenadas descrições textuais do processo de negócio e são definidas de acordo com um assunto, por exemplo, uma tabela de dimensão de produto contém informações textuais referentes a um produto, como nome do produto, descrição e marca do produto. Caso a modelagem dimensional seja armazenada em uma base de dados relacional, o artefato gerado ao final da modelagem é chamado de esquema estrela, devido à sua estrutura semelhante a uma estrela, onde a tabela de fato fica no centro, ligada às tabelas de dimensão através dos relacionamentos de chaves estrangeiras e primárias [KIMBALL e ROSS 2013]. Pode-se encontrar na Figura 2 uma representação da estrutura do esquema estrela.

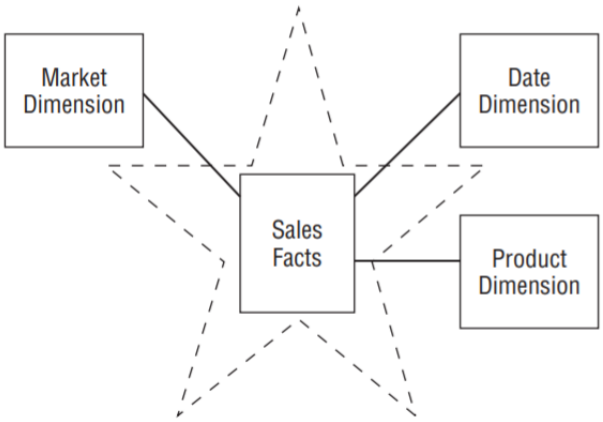

Figura 2. Representação do esquema estrela de acordo com [KIMBALL e ROSS 2013]

Para o desenvolvimento da modelagem dimensional [KIMBALL e ROSS 2013] definem quatro etapas a serem seguidas. A primeira etapa é a definição e compreensão do processo de negócio a ser modelado, visto que um Data Mart é referente a apenas um processo de negócio específico e há apenas uma modelagem dimensional por Data Mart. A segunda etapa é a declaração do grão do Data Mart, ou seja, qual o menor nível que se pode chegar nos dados, servindo assim como um norte para o desenvolvimento da modelagem. Com o grão definido, é iniciada a terceira etapa, que é a definição das dimensões do processo de negócio em questão. Para a definição das dimensões busca-se principalmente identificar os fatores ligados ao fato que está sendo analisado, como: o que aconteceu, como aconteceu, quando aconteceu, quem está envolvido, entre outros fatores. Por fim a quarta e última etapa da modelagem dimensional é a identificação das medidas de performance do processo que se desejam ser analisadas e que vão compor a tabela de fato, baseando-se principalmente no grão definido e buscando o que se quer calcular nas análises. A Figura 3 traz uma representação visual do processo de modelagem dimensional descrito.

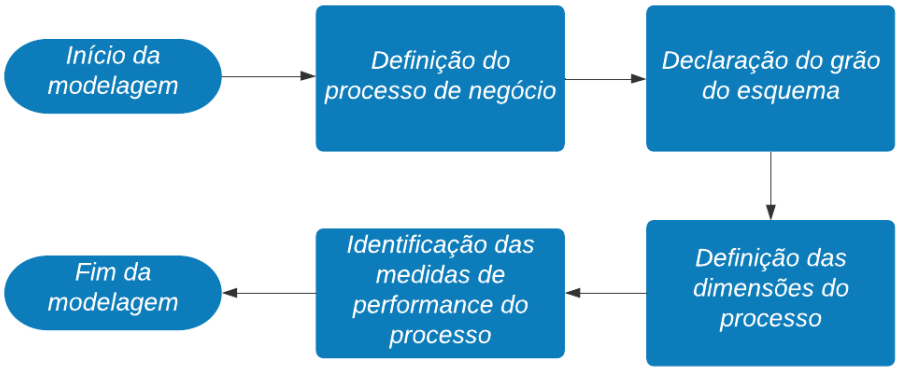

Figura 3. Etapas da modelagem dimensional

4. Materiais e Métodos

Na presente seção será descrito o processo realizado para o desenvolvimento do projeto de Data Warehousing, que possui como objetivo apresentar informações que auxiliem na análise dos dados da evasão escolar. Então, visando cumprir duas das sete práticas recomendadas por [LOBO 2012] para redução da evasão escolar, ou seja, a análise das estatísticas da evasão escolar e determinação das suas causas, foi desenvolvido um ambiente de Data Warehousing contendo um Data Mart sobre a evasão escolar. Para o projeto foram utilizados dados mascarados do SIG@ UFPE, aliados a dados simulados que descrevem os alunos do Centro de Informática da UFPE, porém a ferramenta pode ser utilizada para analisar a evasão escolar de qualquer curso da UFPE, desde que sejam disponibilizados os dados. A metodologia adotada no desenvolvimento foi a aplicação dos conceitos de [KIMBALL e ROSS 2013] para o desenvolvimento dos componentes de um ambiente de Data Warehousing e para o desenvolvimento da modelagem dimensional, conforme a Figura 4.

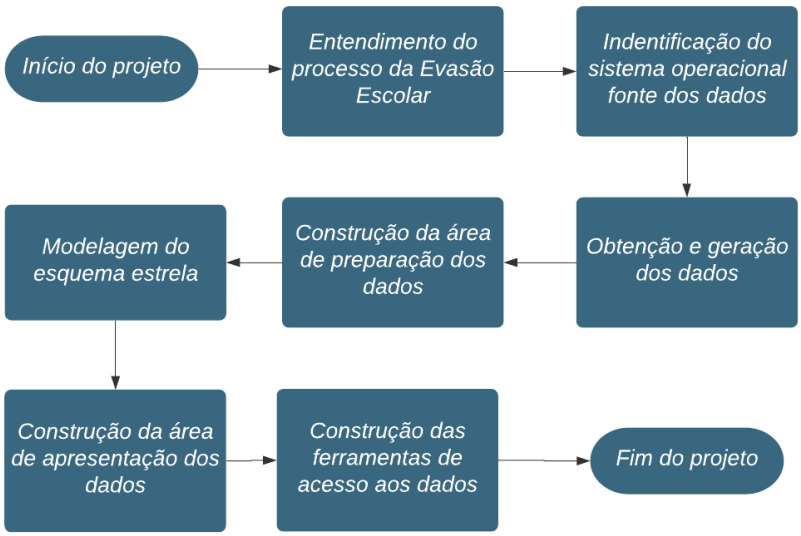

Figura 4. Fluxograma do desenvolvimento do projeto

4.1. Entendimento do processo de evasão escolar

Uma das principais fases no desenvolvimento de um projeto de Data Warehousing, é o entendimento do processo que se busca analisar. Considerando o primeiro passo para a modelagem dimensional descrito por [KIMBALL e ROSS 2013] como sendo a compreensão do processo de negócio em questão, buscou-se primariamente a conclusão deste ponto antes de seguir com o desenvolvimento do ambiente. Nesta etapa, teve-se como objetivo buscar informações sobre evasão escolar, sua definição e principais causas que levam um aluno a entrar em situação de evasão, para entender principalmente como o processo ocorre em geral. Foram utilizados como fonte de obtenção de conhecimento trabalhos realizados na área, reuniões com o público-alvo e materiais contendo os conceitos

necessários para alcançar o objetivo desta etapa. Para a compreensão de como a análise é feita na UFPE, foi realizado um estudo no relatório disponibilizado, desenvolvido pela Pró-Reitoria de Planejamento Orçamentário e Finanças da UFPE (PROPLAN - UFPE), que descreve os dados da situação do vínculo dos alunos por semestre, e serviu como base para o desenvolvimento do projeto. O relatório em questão está representado na Figura 5.

SISTEMAS DE INFORMAÇÃO

ANO DE INGRESSO	INGRESSANTES			CONCLUINTES			EVADIDOS			VINCULADOS	
	VEST/SISU	OUTRA	TOTAL	ALUNOS	TAXA DE SUCESSO(%)	TEMPO MÉDIO	EVASÃO	% EVASÃO	TEMPO MÉDIO	VINC	%
2008	0	1	1	1	100,0	20,0	-	-	-	-	-
2010	25	3	28	11	39,3	11,5	17	60,7	9,4	-	-
2011	50	-	50	27	54,0	12,0	22	44,0	6,5	1	2,0
2012	70	3	73	42	57,5	10,4	26	35,6	7,1	5	6,8
2013	71	-	71	37	52,1	10,6	19	26,8	6,5	15	21,1
2014	71	6	77	24	31,2	8,9	25	32,5	4,4	28	36,4
2015	71	-	71	9	12,7	9,1	29	40,8	4,1	33	46,5
2016	70	-	70	1	1,4	6,0	19	27,1	2,3	50	71,4
2017	70	1	71	-	-	-	17	23,9	2,0	54	76,1
2018	67	2	69	-	-	-	7	10,1	1,4	62	89,9
2019	70	-	70	-	-	-	-	-	-	70	100,0
TOTAL DE SISTEMAS DE INFORMAÇÃO	635	16	651	152	23,3	11,1	181	27,8	4,9	318	48,8

Situação em Dezembro/2019

Ingressantes = Vestibular/Sisu + Outras formas de Ingresso
Evadidos = Desligamento + Desvinculado + Tranferências Interna e externa
Vinculados = Matriculado + Matrícula Vínculo + Trancamento + Mobilidade estudantil
Concluintes = Integralizado ou Formado
Tempo Médio = Nº de semestres médio de permanência

Figura 5. Relatório da situação acadêmica dos alunos da UFPE

4.2. Ambiente de Data Warehousing

A seguir serão descritos os componentes do ambiente de Data Warehousing desenvolvido, bem como seu processo de construção e modelagem. Para a construção do ambiente, o foco principal se deu em modelar os componentes descritos por [KIMBALL e ROSS 2013] sendo eles, a identificação do sistema operacional de fonte dos dados, a construção da área de preparação dos dados, da área de apresentação dos dados e por fim das ferramentas de acesso aos dados. Na Figura 6 pode-se visualizar como o projeto ficou estruturado.

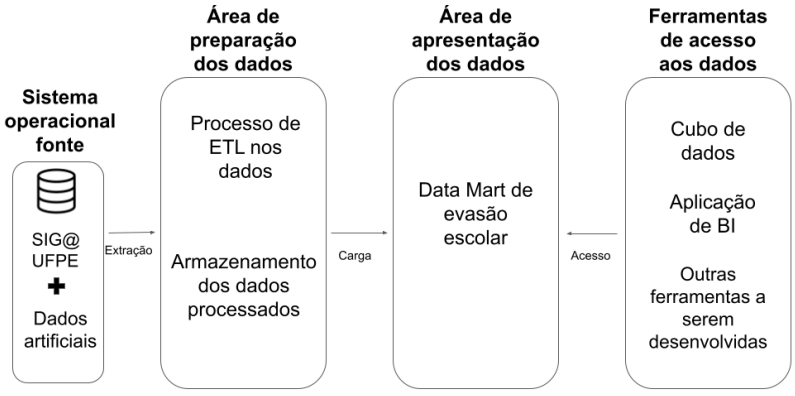

Figura 6. Estrutura do ambiente de Data Warehousing

O sistema operacional que serviu de fonte para os dados foi principalmente o SIG@ da UFPE, sistema de gestão acadêmica da universidade, que contém os dados de

todos os seus alunos. Com base em reuniões com o coordenador do curso de Sistemas de Informação e na análise do relatório da PROPLAN [PROPLAN-UFPE 2020] foi possível identificar os dados do SIG@ que seriam necessários para o desenvolvimento do projeto. A partir disso foram requisitadas, através de uma solicitação à equipe de administração do SIG@ UFPE quatro tabelas, que foram complementadas com colunas geradas artificialmente, para que fossem trazidos os dados necessários para as análises que precisariam ser realizadas no Data Mart para o cumprimento de seu objetivo. Por questões de disponibilidade e privacidade dos dados, seu conteúdo não condiz com a realidade, porém é importante que sejam utilizados dados reais da UFPE no desenvolvimento do projeto. A relação das tabelas extraídas do SIG@ e suas descrições estão dispostas na Tabela 1.

Tabela 1. Tabelas extraídas do SIG@

NOME DA TABELA	DESCRIÇÃO
alunoscin	Tabela contendo os dados dos alunos do Centro de Informática
disciplinascursadas	Tabela contendo a relação entre os alunos do Centro de Informática e as disciplinas cursadas
docentesemdisciplinas	Tabela contendo a relação de professores e disciplinas
turmasdocentes	Tabela contendo a relação das turmas de disciplinas que cada professor leciona

Levando em conta os dados trazidos do sistema operacional fonte e a compreensão do processo de negócio, puderam ser selecionados os dados que seriam utilizados para o desenvolvimento do Data Mart do ambiente. Após selecionados os dados, estes foram trazidos para o fluxo de dados da ferramenta da suíte Pentaho Community Edition, Pentaho Data Integration (PDI), uma ferramenta própria para integração, extração, tratamento e carga de dados, entre diversas outras aplicações. Dentro do PDI, os dados passaram por uma rotina de ETL, sendo extraídos da fonte, tratados para manter a consistência e qualidade dos dados, e por fim carregados em uma tabela do sistema gerenciador de banco de dados PostgreSQL, estando agora próprios para serem modelados para a área de apresentação dos dados.

Para o desenvolvimento da área de apresentação dos dados, onde ficam armazenados os dados de forma organizada para o consumo de forma performática, foi desenvolvido um Data Mart no modelo de esquema estrela a partir da modelagem dimensional dos dados contidos na área de preparação dos dados. Na modelagem do esquema estrela foram seguidos os quatro passos para a modelagem dimensional definidos por [KIMBALL e ROSS 2013], que devem ser seguidos de maneira sequencial.

Dado que o primeiro passo, a definição e compreensão do processo de negócio, já havia sido feito previamente, esta etapa foi dada como concluída e deu-se prosseguimento à próxima, sendo a declaração do grão. A decisão da declaração do grão foi feita baseando-se primariamente no conhecimento do processo da evasão escolar, e na característica transacional da fonte de dados, onde cada linha representa a situação de cada aluno em um semestre em específico. Então, como conclusão desta etapa, o grão da modelagem dimensional, o dado em nível mais atômico que é possível ser alcançado nas análises, foi definido como cada aluno por semestre.

A etapa seguinte, a definição as dimensões, foi feita primariamente baseando-se novamente no processo de negócio, e na granularidade do modelo, identificados nas etapas anteriores. O foco foi em determinar as dimensões do processo, trazendo os fatores envolvidos, e que podem ser analisados em relação à evasão escolar. Com a aplicação

deste método foi possível elencar as dimensões: tempo, aluno, situação do aluno e curso. Por questões de performance e práticas definidas por [KIMBALL e ROSS 2013] para a definição de dimensões, a dimensão de aluno foi dividida em duas tabelas, de acordo com a cardinalidade dos dados. Uma tabela foi destinada para dados com mais de cinco possibilidades de registros presentes na coluna, onde ficaram armazenados os dados como nome do aluno, cpf, faixa etária, endereço, faixa de renda e semestre de ingresso. Já a outra dimensão referente aos alunos foi criada contendo elementos com cinco possibilidades de registros presentes na coluna ou menos, como o sexo do aluno, a cota que utilizou para ingressar na universidade, se ele recebe uma bolsa ou auxílio, e qual tipo de bolsa ou auxílio. A dimensão curso possui os dados referentes ao curso do aluno, contendo o nome do curso, centro, turno, localização e campus. Na dimensão de situação do aluno ficaram armazenadas a situação geral e situação detalhada do aluno, em relação ao seu vínculo com a universidade. Por fim a dimensão de tempo possui os dados de semestre e ano, e faz referência à data da situação de vínculo do aluno. É possível encontrar um dicionário de dados do Data Mart, contendo o nome dos campos, sua descrição, a coluna que originou o campo do esquema estrela, a origem do dado, e a tabela do Data Mart onde o dado está armazenado, na Tabela 2.

Tabela 2. Dicionário de dados da área do Data Mart desenvolvido

NOME DO CAMPO	DESCRIÇÃO	COLUNA ORIGEM	ORIGEM	TABELA DM
cpf	número de CPF do aluno	nm_cpf_pess	SIG@ - UFPE	Dimensão aluno
nome	Nome do aluno	nome_aluno	SIG@ - UFPE	Dimensão aluno
fx_etária	Faixa etária do aluno	data_nascimento	SIG@ - UFPE	Dimensão aluno
tipo_logr	Tipo do logradouro do endereço do aluno	tipo_logr	SIG@ - UFPE	Dimensão aluno
logradouro	Logradouro do endereço do aluno	logradouro	SIG@ - UFPE	Dimensão aluno
bairro	Bairro do endereço do aluno	bairro	SIG@ - UFPE	Dimensão aluno
semestre_ingresso	Semestre de ingresso do aluno	ingresso	SIG@ - UFPE	Dimensão aluno
cidade	Cidade do endereço do aluno	cidade	SIG@ - UFPE	Dimensão aluno
fx_renda	Faixa de renda declarada da família do aluno	fx_renda	Gerado artificialmente	Dimensão aluno
sexo_aluno	Sexo do aluno	sexo	SIG@ - UFPE	Dimensão combinada aluno
nome_cota	Nome da cota usada pelo aluno para ingressar na universidade	nome_cota	SIG@ - UFPE	Dimensão combinada aluno
recebe_bolsa_auxilio	Se o aluno recebe bolsa ou auxílio	bolsa_auxilio	Gerado artificialmente	Dimensão combinada aluno
tipo_bolsa_auxilio	Qual auxilio/bolsa o aluno recebe	tipo_bolsa_auxilio	Gerado artificialmente	Dimensão combinada aluno
situacao_geral_aluno	Situação de vínculo do aluno: Evadido, Vinculado, Ingressante, Concluinte	situacao_geral_aluno	Gerado artificialmente	Dimensão combinada situação
situacao_detalhada_aluno	Situação detalhada do vínculo do aluno	situacao_detalhada_aluno	Gerado artificialmente	Dimensão combinada situação
curso	Curso do aluno	curso	SIG@ - UFPE	Dimensão curso
centro	Centro do curso que o aluno está matriculado	centro_curso	Gerado artificialmente	Dimensão curso
turno	Turno do curso que o aluno está matriculado	turno_curso	Gerado artificialmente	Dimensão curso
localizacao	Endereço do curso que o aluno está matriculado	local_curso	Gerado artificialmente	Dimensão curso
campus	Campus do curso que o aluno está matriculado	campus_curso	Gerado artificialmente	Dimensão curso
qtd_alunos	Coluna utilizada para contar a quantidade de alunos. O conteúdo da coluna é o número de matrícula do aluno.	cod_matricula	SIG@ - UFPE	Tabela de fato
tempo_de_permanência	Quantidade de semestres que o aluno permaneceu na universidade	tempo_de_permanência	Gerado artificialmente	Tabela de fato
semestre	Semestre.	semestre	Gerado artificialmente	Dimensão tempo
ano	Ano.	ano	Gerado artificialmente	Dimensão tempo

Para a conclusão da modelagem dimensional do esquema estrela do Data Mart, a última etapa foi a identificação das medidas de performance do processo de negócio, que ficarão armazenadas na tabela de fato. Visando as futuras análises, e baseando-se nos resultados das etapas anteriores (definição do processo de negócio, grão e dimensões) as medidas identificadas foram a de quantidade de alunos e o tempo de permanência do aluno na universidade. As medidas serão agregadas nas análises através das operações de contagem das matrículas distintas dos alunos para a medida de quantidade de alunos, e média para os tempos de permanência dos alunos na universidade. Com a finalização

desta última etapa, o esquema estrela está completo, mas é importante ressaltar que novas informações podem ser incluídas no modelo, desde que seja respeitada a granularidade de cada dimensão. O esquema estrela desenvolvido está representado na Figura 7.

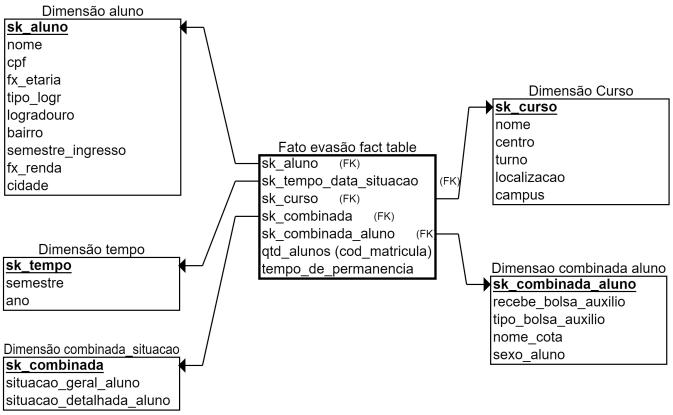

Figura 7. Esquema estrela do projeto

O último componente para a finalização do ambiente de Data Warehousing são as ferramentas de acesso aos dados, onde as ferramentas construídas devem acessar os dados disponibilizados na área de apresentação de dados. Para exemplificar algumas ferramentas de acesso aos dados que podem ser desenvolvidas, foram criadas duas neste projeto, sendo elas um cubo de dados multidimensional, uma ferramenta que permite análise em todas as dimensões do negócio, utilizando como fonte uma base de dados na modelagem dimensional, e uma aplicação de BI baseada em parâmetros, que podem ser configurados pelo usuário a fim de obter as informações necessárias. As ferramentas devem ser desenvolvidas de modo que qualquer usuário da área de negócio da organização possa utilizá-las sem dificuldade, para a obtenção de informações. Alguns exemplos de informações que podem ser obtidas com o Data Mart desenvolvido, utilizando as ferramentas de acesso aos dados, podem ser encontrados na Tabela 3.

Tabela 3. Lista parcial de informações possíveis

Informação
Quantitativo de alunos evadidos por: faixa etária, bairro, semestre de ingresso, faixa de renda, semestre, ano, tipo de evasão, curso, centro, campus, tipo de bolsa ou auxílio, cota, sexo, município.
Lista dos alunos em situação de evasão escolar
Tempo de permanência médio dos alunos em situação de evasão escolar
Taxa de conclusão por semestre
Taxa de evasão por semestre
Taxa de evasão por ano
Taxa de conclusão por ano

5. Resultados

Nesta seção, serão descritas as ferramentas de acesso aos dados desenvolvidas para exemplificar as análises que podem ser feitas a partir do Data Mart do ambiente desenvolvido. Também serão apresentadas as tecnologias usadas no desenvolvimento de cada ferramenta e a sua maneira de uso.

O cubo de dados foi desenvolvido no Pentaho Schema Workbench Community Edition (PSWCE), uma ferramenta para criação e edição de cubos de dados, e foi publicado no Saiku Analytics, uma ferramenta para visualização de cubos de dados e realização de análises multidimensionais hospedada no servidor de BI da suíte Pentaho, o Pentaho BI Server Community Edition. Com os dados disponíveis no cubo são possíveis análises com operações de drill down, drill up, slice e dice, ou seja, permitindo a navegação nos dados desde sua forma mais agregada possível até o grão dos dados, com os filtros e disposição necessários. A seguir serão demonstrados nas Figuras 8, 9 e 10 alguns exemplos de análises possíveis com os dados disponibilizados:

Ano da situação do aluno	Nome Curso	Situação Geral	Situação Detalhada	Quantidade de alunos	Tempo médio de permanência
2020	CIÊNCIAS DA COMPUTAÇÃO	CONCLUINTE	FORMADO	1	8
		EVADIDO	DESVINCULADO	1	8
			TRANSFERÊNCIA EXTERNA	2	8
		INGRESSANTE	OUTRAS FORMAS DE INGRESSO	3	10
			VESTIBULAR/SISU	7	9,857
		VINCULADO	MATRICULADO	8	9,75
	ENGENHARIA DA COMPUTAÇÃO	CONCLUINTE	FORMADO	6	9
		EVADIDO	DESVINCULADO	2	11
			TRANSFERÊNCIA EXTERNA	1	11
		INGRESSANTE	OUTRAS FORMAS DE INGRESSO	3	11
			VESTIBULAR/SISU	15	10,6
		VINCULADO	MOBILIDADE ESTUDANTIL	1	11
			TRANCAMENTO	3	11
	SISTEMAS DE INFORMAÇÃO	CONCLUINTE	FORMADO	4	8,25
		EVADIDO	DESVINCULADO	5	8
			TRANSFERÊNCIA EXTERNA	4	8
		INGRESSANTE	OUTRAS FORMAS DE INGRESSO	5	8,4
			VESTIBULAR/SISU	16	8,438
		VINCULADO	MATRÍCULA VÍNCULO	4	8
			MOBILIDADE ESTUDANTIL	4	9
			TRANCAMENTO	5	9

Figura 8. Exemplo de análise verificando as situações dos alunos por curso e por ano em formato de tabela

Ano da situação do aluno	Situação Geral	Quantidade de alunos
2020	CONCLUINTE	11
	EVADIDO	15
	INGRESSANTE	49
	VINCULADO	25

Figura 9. Exemplo de análise verificando as situações dos alunos por ano em formato de tabela

Nome Curso	Sexo	Situação Geral	Cidade	Faixa de renda	Quantidade de alunos	Tempo médio de permanência
SISTEMAS DE INFORMAÇÃO	Feminino	EVADIDO	JABOATAO DOS GUARARAPES	R$ 2.005,00 - R$ 8.640,00	1	8
			OLINDA	R$ 2.005,00 - R$ 8.640,00	2	8
			RECIFE	R$ 2.005,00 - R$ 8.640,00	5	8
	Masculino	EVADIDO	OLINDA	R$ 1.255,00 - R$ 2.004,00	1	8

Figura 10. Exemplo de análise visualizando os dados dos alunos em situação evasão escolar do curso de sistemas de informação em formato de tabela

A segunda ferramenta de acesso aos dados desenvolvida para o ambiente foi um conjunto de painéis de BI dinâmicos, que visam fornecer consultas prontas para que o

usuário visualize informações estatísticas a partir da configuração de parâmetros, resultando em uma aplicação de BI composta por três telas, uma página inicial, uma tela de relatórios semestrais e uma tela de relatórios por município. A tecnologia utilizada para o desenvolvimento dos painéis foi o plugin Sparkl do Pentaho BI Server Community Edition, um plugin próprio para o desenvolvimento e implantação de aplicações de BI. Os painéis desenvolvidos, assim como o restante do ambiente de Data Warehousing, foram construídos baseando-se principalmente no relatório da PROPLAN-UFPE disponibilizado. A relação dos indicadores usados, filtros que podem ser feitos e o painel onde estão presentes está disposta na Tabela 4.

Tabela 4. Lista informações disponíveis na aplicação de BI

Informação	Filtros	Painel
Tempo de permanência médio dos alunos em situação de evasão escolar, conclusão de curso, ingresso no curso e vinculados	Semestre, ano e curso	Semestral
Taxa de conclusão (sucesso)	Semestre, ano e curso	Semestral
Taxa de evasão	Semestre, ano e curso	Semestral
Quantitativo de alunos evadidos por tipo de conclusão (integralizado ou formado)	Semestre, ano e curso	Semestral
Quantitativo de alunos evadidos por tipo de ingresso (Vestibular/Sisu, outras formas de ingresso)	Semestre, ano e curso	Semestral
Quantitativo de alunos evadidos por tipo de vínculo (matriculado, matrícula vínculo, trancamento, mobilidade estudantil)	Semestre, ano e curso	Semestral
Quantitativo de alunos evadidos por tipo de evasão (desligamento, desvinculado, transferência externa ou interna)	Semestre, ano e curso	Semestral
Lista dos alunos em situação de evasão escolar	Semestre, ano e curso	Semestral
Lista dos alunos em situação de conclusão de curso	Semestre, ano e curso	Semestral
Lista dos alunos vinculados	Semestre, ano e curso	Semestral
Lista dos alunos ingressantes	Semestre, ano e curso	Semestral
Quantitativo de alunos evadidos por município	Nenhum	Alunos por município
Quantitativo de alunos concluintes por município	Nenhum	Alunos por município
Quantitativo de alunos por município	Nenhum	Alunos por município

Os relatórios semestrais só são apresentados após a seleção dos parâmetros disponíveis, e trazem os indicadores da taxa de evasão, tempo médio de evasão, taxa de sucesso e tempo médio de conclusão. Além dos indicadores, também são apresentadas informações sobre os alunos ingressantes, concluintes, evadidos e vinculados, respectivamente, conforme as Figuras 11, 12, 13 e 14.

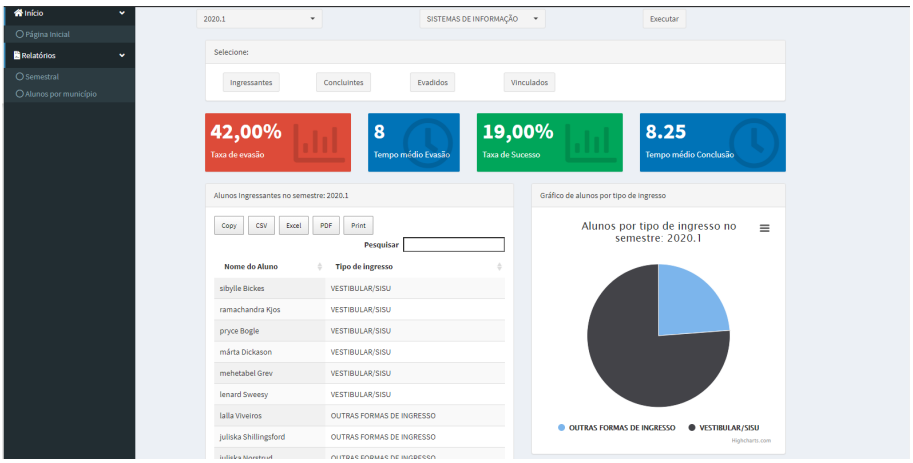

Figura 11. Tela de relatórios semestrais, após filtrados o curso e semestre, apresentando a aba de alunos ingressantes

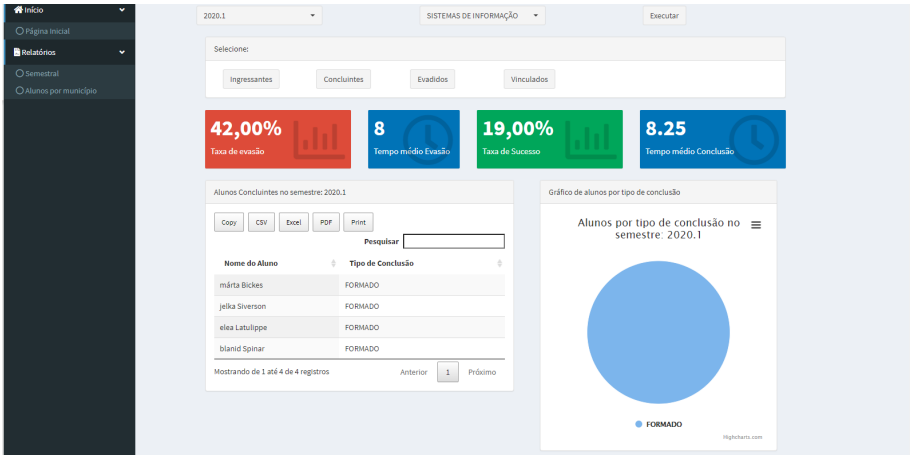

Figura 12. Tela de relatórios semestrais, após filtrados o curso e semestre, apresentando a aba de alunos concluintes

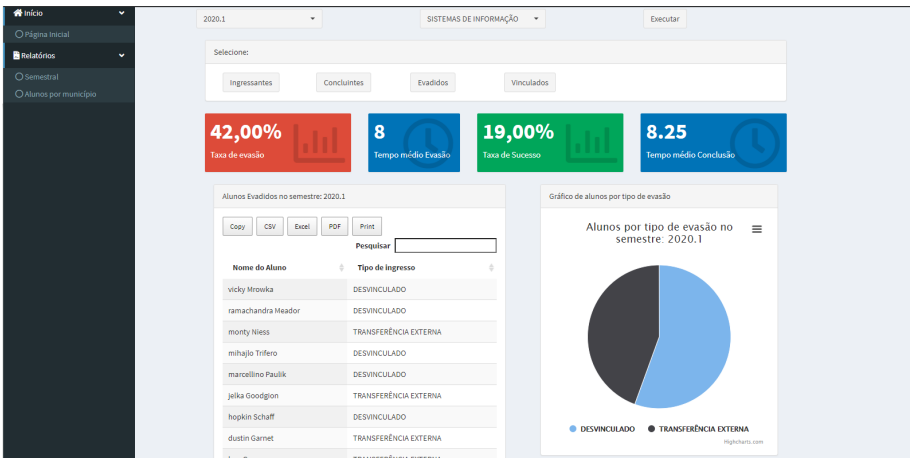

Figura 13. Tela de relatórios semestrais, após filtrados o curso e semestre, apresentando a aba de alunos evadidos

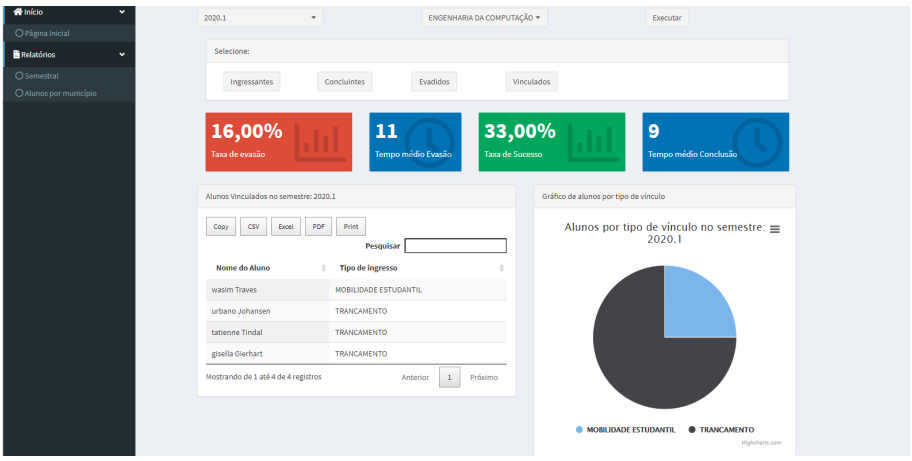

Figura 14. Tela de relatórios semestrais, após filtrados o curso e semestre, apresentando a aba de alunos vinculados

Além dos relatórios semestrais, também foi desenvolvido um relatório municipal, trazendo informações sobre os alunos em relação ao seu município de residência, conforme a Figura 15:

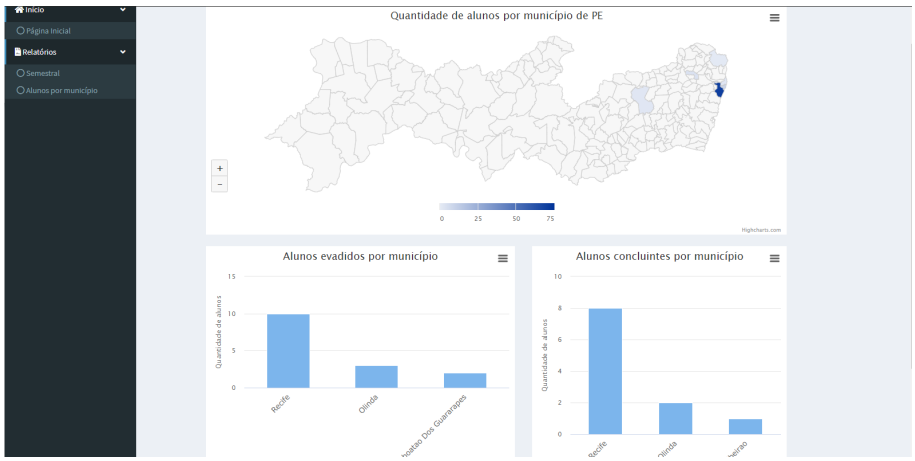

Figura 15. Relatórios por município do painel de BI

6. Trabalhos relacionados

Na atual seção serão elencados trabalhos com temática semelhante, que serviram como base e inspiração para a escrita do presente trabalho. Os trabalhos abordam o uso de técnicas de BI na área de educação, em particular, na educação superior, e foram analisados verificando pontos como a origem dos dados, técnicas utilizadas e abrangência quanto a evasão escolar. Apesar da importância de tais obras, nenhuma leva em questão o cenário e estrutura de dados da UFPE, o que torna o presente trabalho relevante para a universidade. Dentre as obras analisadas, foi identificado um padrão referente ao desenvolvimento de um Data Warehouse para o fornecimento de informações gerenciais na área de educação superior, nele consiste a criação de modelos de bases de dados dimensionais, e utilizaram como fontes: dados abertos, dados provenientes de pesquisas, e dados de sistemas SIG, sistemas provenientes de instituições federais de educação, que por sua vez fazem parte do termo de cooperação com a Universidade Federal do Rio Grande do Norte, semelhantes ao SIG@ da UFPE. Um comparativo entre os trabalhos analisados pode ser encontrado na Tabela 5.

Tabela 5. Comparativo entre os trabalhos relacionados

Autor	Fonte dos dados	Técnicas de BI utilizadas	Abrangência quanto à evasão escolar
[SANTOS 2017]	SIG-IFPR	Data Warehouse, CRISP-DM, Painéis de BI	Fornece informações que apoiem na gestão dos alunos e da evasão escolar
[GOIVEIA e FREITAS 2018]	Dados Abertos (E-MEC e UAB)	Data Warehouse, OLAP, Painéis de BI	Fornece informações que apoiem na gestão dos alunos e da evasão escolar
[JUNIOR et al. 2015]	Table Functions + Visões do SGBD Oracle da UTFPR	Data Warehouse, Painéis de BI	Busca servir como base para trabalhos futuros sobre a evasão escolar

[SANTOS 2017] desenvolveu o principal trabalho que trouxe inspiração ao estudo e desenvolvimento da atual obra. O trabalho traz uma proposta metodológica para solucionar o problema da falta de ferramentas para análise de dados educacionais no sentido da evasão escolar no IFPR. Utilizando uma metodologia adaptada de CRISP-DM para projetos de BI, o autor desenvolve um Data Warehouse, e cria painéis que apresentam informações sobre a evasão escolar no ensino superior, alimentados por dados do SIG-IFPR.

[GOUVEIA e FREITAS 2018] trazem como contribuição, a implementação de um Data Warehouse utilizando dados educacionais provenientes das fontes de dados abertos do E-MEC e da Universidade aberta do Brasil (UAB). A partir do DW criado, com os dados armazenados no SGBD MySQL, os autores utilizaram das ferramentas de BI da suíte Pentaho, como o Pentaho Schema Workbench para o desenvolvimento de cubos de dados e o plugin do Pentaho BI Server, Saiku Analytics View para a visualização dos dados em uma forma multidimensional a partir dos cubos desenvolvidos. Com isso, o trabalho tem como um de seus objetivos servir de apoio para trabalhos futuros, que almejem utilizar técnicas de machine learning para a descoberta de padrões de evasão escolar.

[JÚNIOR et al. 2015] propõem um Data Warehouse educacional para apoiar os gestores da educação superior na tomada de decisão. Os autores utilizaram os dados do SGBD Oracle da Universidade Tecnológica Federal do Paraná (UTFPR) e construíram o DW a partir de visões materializadas e table functions dentro do SGBD, pois alegam que desta forma, possuem mais facilidade na agregação das regras de negócio, e na construção dos relatórios para a camada de visualização, gerando assim um Virtual Warehouse. A partir dos dados contidos no DW construído foram criados painéis com tabelas, gráficos e filtros, permitindo que os usuários visualizem os dados atuais sobre os alunos e dados em série histórica. O trabalho tem foco principal a análise da evasão e retenção dos alunos em cursos de graduação, além de servir como base para aplicação de técnicas de machine learning para identificação de padrões de evasão escolar em trabalhos futuros, assim como o trabalho de [GOUVEIA e FREITAS 2018], citado anteriormente.

7. Considerações Finais

Nesta seção serão expostas as considerações finais do presente trabalho. Dentre as considerações, estão inclusas a contribuição esperada, tanto da proposta do trabalho quanto da sua implementação, e recomendações para trabalhos futuros, que visem complementar ou basear-se no presente trabalho.

7.1. Contribuição esperada

O presente trabalho tem como contribuição esperada fornecer uma proposta que sirva como base para a construção de um ambiente de Data Warehousing, com o objetivo de fornecer informações que apoiem a gestão da evasão escolar na UFPE. As principais vantagens e pontos de contribuição da implementação do projeto baseiam-se no cumprimento dos deveres de um DW propostos por [KIMBALL e ROSS 2013]. Com isso, destacam-se:

- A acessibilidade dos dados, ficando agora integrados em um repositório que permite consultas performáticas e geração de relatórios, além da disponibilidade para os gestores, por meio das ferramentas de acesso aos dados;

- A apresentação consistente dos dados, visto que os dados devem ser condizentes com a realidade, e com a utilização de dados de sistemas da organização e o processo de ETL, os dados ganham consistência e veracidade;
- Adaptabilidade, escalabilidade e resiliência a mudanças, dado que o projeto pode ser incrementado com novos dados, dimensões ou medidas que forem identificados e possuam compatibilidade com a granularidade do esquema estrela proposto, além da possibilidade da construção de outras ferramentas de acesso aos dados e até mesmo outros Data Marts. O modelo também pode ser utilizado para qualquer curso, ou centro da UFPE, visto que a estrutura de dados do SIG@ é a mesma para toda a universidade;
- Proteção aos dados, que pode ser adicionada através da implementação de um sistema de autenticação para as ferramentas de acesso aos dados e aos SGBD onde ficam armazenadas as áreas de apresentação e preparação dos dados do DM.

7.2. Trabalhos futuros

Dentre as sugestões para trabalhos futuros, elencam-se as seguintes para complementação do presente trabalho ou desenvolvimento de trabalhos semelhantes:

- Testar a proposta do presente trabalho com dados reais e avaliar sua eficácia em apoiar a equipe de gestão da UFPE no controle da evasão escolar;
- Testar o modelo dimensional proposto, avaliando seu desempenho na realização de consultas aos dados;
- Adicionar mais dados ao Data Mart, permitindo uma maior quantidade de cenários a serem analisados, complementando a modelagem dimensional conforme necessário;
- Utilização do modelo dimensional desenvolvido neste projeto para montar uma base de dados, e com isso realizar o treinamento de algoritmos de inteligência artificial para identificação de padrões relacionados à evasão escolar;
- Testar e adaptar, caso necessário, o modelo proposto para outras universidades públicas ou privadas;
- Desenvolvimento de Data Marts complementares de diferentes processos da UFPE, como por exemplo, processos da área financeira, ou relativos ao corpo docente, ampliando a gama de consultas e relatórios que podem ser gerados.

Referências

BRASIL (2014). Lei nº 13.005, de 25 de junho de 2014. aprova o plano nacional de educação - pne e dá outras providências. *Diário Oficial [da] República Federativa do Brasil*.

GOUVEIA, R. M. M. e FREITAS, C. N. C. (2018). ImplementaÇÃo de um data warehouse para anÁlise de dados abertos governamentais da educaÇÃo a distÂncia. *TEAR Revista de educação, ciência e tecnologia*, 7(2).

INEP (2017). Metodologia de cálculo dos indicadores de fluxo da educação superior. Technical report, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), Brasília.

JÚNIOR, J. G. D. O., BASTOS, L. C., e KAESTNER, C. A. A. (2015). Uma abordagem de data warehouse educacional para apoio à tomada de decisão. In *Anais dos Workshops do IV Congresso Brasileiro de Informática na Educação (CBIE 2015)*.

KIMBALL, R. e ROSS, M. (2013). *The Data Warehouse Toolkit*. John Wiley & Sons Inc., Indianapolis, Indiana.

LOBO, M. B. D. C. M. (2012). Panorama da evasÃo no ensino superior brasileiro: Aspectos gerais das causas e soluÇÕes. Technical report, Instituto Lobo para o Desenvolvimento da Educação, Ciência e Tecnologia, Brasília.

PROPLAN-UFPE (2020). RelatÓrio da situaÇÃo acadÊmica dos alunos ingressantes por curso segundo o anos de ingresso na ufpe, 2001-2019. Technical report, Pró-Reitoria de Planejamento, Orçamento e Finanças, Recife.

SANTOS, J. S. D. (2017). Business intelligence: Uma proposta metodolÓgica para anÁlise da evasÃo escolar em instituiÇÕes federais de ensino. Master's thesis, UNIVERSIDADE FEDERAL DO PARANÁ, CURITIBA.

UFPE (2019). ResoluÇÃo nº 15/2019. regulamenta a política de assistência estudantil da universidade federal de pernambuco. *5ª (QUINTA) SESSÃO ORDINÁRIA DO CONSELHO DE ENSINO, PESQUISA E EXTENSÃO-CEPE*.