Uma arquitetura de suporte à coleta e análise de dados educacionais em ambientes de ensino institucionais

Pedrina Célia Brasil[1], Isabel Dillmann Nunes[1]

[1] Programa de Pós-Graduação em Inovação e Tecnologias Educacionais – PPgITE
Universidade Federal do Rio Grande do Norte (UFRN) - Natal- RN - Brasil

pedrina.brasil@gmail.ufrn.br, bel@imd.ufrn.br

Abstract. *Although there is a large amount of educational data nowadays, accessing and processing these data is still a challenging task for researchers and developers of learning analysis application. Many developers use different approaches to implementing their systems. In a fragmented way, each developer uses a different approach to implement their application, even when they share an essential set of data and tools with other developers. Hence, this article proposes as a challenge the promotion of a consensual model that describes how to efficiently integrate common components and tools into the automation of a Learning Analysis process.*

Resumo. *Embora atualmente exista uma grande quantidade de dados educacionais, o acesso e o processamento a esses dados ainda é uma tarefa desafiadora aos pesquisadores e desenvolvedores de aplicações baseadas em análise de aprendizagem. De forma fragmentada, cada desenvolvedor utiliza uma abordagem diferente para implementar sua aplicação, mesmo quando ela compartilha um conjunto essencial de dados e ferramentas em comum com outros sistemas. Nesse sentido, este artigo propõe como desafio a promoção de um modelo de referência que descreva como integrar eficientemente os componentes e ferramentas comuns a automação de um processo de Análise de Aprendizagem.*

1. Introdução

O avanço e uso das Tecnologias de Informação e Comunicação (TIC) no âmbito educacional possibilitou o desenvolvimento de novas soluções personalizadas ao contexto do aluno (FERGUSON *et al.*, 2016). Em consonância com esse desenvolvimento, a adoção em larga escala de Ambientes Virtuais de Aprendizagem (AVA) — também conhecidos como Sistemas de Gerenciamento de Aprendizagem (SGA) — em contextos escolares significa que instituições de ensino podem tirar vantagem dos dados brutos desses sistemas e extrair informações úteis ao processo educacional (ALMAZROUI, 2013).

Nesse contexto, surge a área de *Learning Analytics* (LA) — em português, Análise de Aprendizagem. LA é definida como a medida, coleta, análise e relato de dados sobre os alunos e seus contextos de aprendizagem com o objetivo de entender e otimizar o aprendizado e o ambiente que este ocorre (FERGUSON *et al.*, 2016). Por meio da análise da aprendizagem é possível, por exemplo, descobrir as dificuldades de aprendizagem de um aluno e realizar intervenções pedagógicas que beneficiem seu ambiente de ensino.

Os dados utilizados em LA podem ser coletados de forma implícita ou explícita a partir das interações do aluno com o SGA. Quando o aluno conclui ou realiza uma tarefa, isso configura-se uma interação explícita do usuário com seu progresso de aprendizagem.

Por outro lado, interações implícitas são ações como a realização de atividades extracurriculares, postagem em fóruns de discussão e outras que não são diretamente consideradas como parte do progresso acadêmico do aluno (JOHNSON *et al.*, 2010).

Embora exista uma grande quantidade de dados educacionais, o acesso e o processamento a eles ainda são tarefas desafiadoras, o que inibe a implementação de aplicações de LA. Estas aplicações, geralmente, abordam a integração de múltiplos sistemas (Sistema de Biblioteca, AVA, Sistema de Registro de Alunos, *etc*), análise de uma variedade de dados sob diferentes visões de usuário (alunos, professores, educadores, *etc*), o desenvolvimento de ferramentas com diferentes propósitos (avaliar a taxa de evasão dos alunos, acompanhar o progresso dos estudantes, notificar o usuário, *etc*) e níveis de granularidade (a nível de instituição, curso, disciplina, *etc*) (SCLATER, 2017).

A maioria dos projetos que exploram o uso de LA em contextos educacionais o faz de forma específica, não considerando as várias fontes de dados e as diferentes perspectivas e níveis de granularidade da instituição (BORGES *et al.*, 2016). Mesmo quando dois projetos compartilham um mesmo conjunto de dados, devido a regras de negócio específicas, cada um utiliza-se de uma abordagem única de acesso e análise (SCLATER, 2017). Ou seja, cada abordagem e ferramenta tem seu próprio modelo de dados e metadados, que, geralmente, lida com uma fonte única de dados (em geral, um único SGA) e atende a uma necessidade específica de um tipo de usuário.

O fato dessas aplicações não explorarem diferentes perspectivas de usuário, níveis de granularidade, fontes de dados e, além disso, serem desenvolvidas de forma fragmentada, restringe a descoberta de novos conhecimentos sobre toda a instituição e seus diferentes contextos. Além de dificultar a replicação das abordagens de análise e coleta de dados já utilizadas institucionalmente.

A partir desse contexto, este artigo propõe como desafio a promoção de um modelo de referência que descreva como integrar, de forma eficiente (sem comprometer as funcionalidades dos sistemas), componentes que atendam diferentes aplicações de LA, sob múltiplas perspectivas e níveis de granularidade institucionais. Este modelo poderia ser trabalhado, por exemplo, através de uma arquitetura de software que permita a colaboração de diferentes *stakeholders* e aplicações.

O artigo está organizado da seguinte forma: a Seção 2 apresenta os conceitos chaves deste trabalho; a Seção 3 descreve os trabalhos relacionados a proposta deste trabalho; a Seção 4 articula o desafio proposto e sugere uma metodologia para atacá-lo; e na Seção 5 são traçadas as considerações finais deste trabalho.

2. Fundamentação Teórica

Nesta seção são apresentados os conceitos chaves deste trabalho, são eles: *Learning Analytics*, *Online Analytical Processing* e *Data Warehouse*.

2.1 Learning Analytics

Learning Analytics (LA) é um campo de pesquisa emergente, tendo ocorrido sua primeira conferência internacional no ano de 2011, na cidade de Banff, Alberta, Canadá (SIEMENS, 2011).

Embora não exista um consenso, LA é comumente definida como a medida, coleta, análise e relato dos dados sobre os alunos e seus contextos de aprendizagem com o objetivo de entender e otimizar o aprendizado e o ambiente que este ocorre (FERGUSON *et al.*, 2016).

Desde o seu surgimento, LA tem sido aplicada, por exemplo, para analisar cursos *on-line* e dar suporte ao desenvolvimento de sistemas de *e-learning* (ensino eletrônico) mais efetivos e, além disso, investigar como os alunos "trapaceiam"[1] o sistema (BAKER e YACEF, 2009; KOTSIANTIS, 2012; BAKER e SIEMENS, 2014).

O processo de LA é baseado em dois campos de pesquisa: *Business Intelligence* (BI) — em português, Inteligência de Negócio — e *Data Mining* (DM) — em português, Mineração de Dados (BORGES *et al.*, 2016). Enquanto o campo de BI se concentra na busca de ferramentas computacionais para auxiliar na tomada de decisão através da fusão eficiente de dados recolhidos dos diferentes sistemas de uma organização (BORGES *et al.*, 2016); DM é um campo da computação que aplica uma variedade de técnicas (*i. e.* construção de árvores de decisão, regras de indução, redes neurais, aprendizagem baseada em instâncias, aprendizagem bayesiana, lógica de programação e algoritmos estatísticos) em um conjunto de dados visando a descoberta e a apresentação de conhecimentos prévios e, potencialmente, padrões úteis de dados (FERGUSON, 2012). Quando aplicada sob um conjunto de dados educacionais, o processo de mineração é tido como *Educational Data Mining* (EDM) — em português, Mineração de Dados Educacionais. Este subcampo de pesquisa visa o desenvolvimento de métodos para explorar dados oriundos de AVA para melhoria do entendimento sobre os estudantes e o contexto que eles aprendem (FERGUSON, 2012).

No contexto deste trabalho, o termo *Learning Analytics* é usado abrangendo tanto análise de dados acadêmicos, quanto mineração de dados educacionais.

2.2 Online Analytical Processing e *Data Warehouse*

Os dados gerados por diferentes usuários são armazenados por um ou mais SGA, através de logs de acesso, bancos de dados relacionais, etc. Em geral, um SGA processa transações de dados em tempo real, eles são tidos como sistemas de OLTP (*Online Transaction Processing*). Esses sistemas processam transações de dados gerados diariamente a partir das interações dos usuários. Nesse tipo de sistema a representação dos dados é feita a nível operacional, por exemplo, a partir de bases de dados que permitem transações e consultas aos dados armazenados (BORGES *et al.*, 2016).

Aplicações de LA, no entanto, podem ser classificadas como aplicações OLAP (*Online Analytical Processing*). Tratam-se de aplicações com capacidade de analisar

[1]Usar as regras destinadas a proteger o sistema para manipulá-lo e direcioná-lo a um resultado desejado.

grandes volumes de informações, nas mais diversas perspectivas de um *Data Warehouse* (DW) (BORGES *et al.*, 2016). O OLAP faz referência às ferramentas analíticas utilizadas em BI para a visualização de informações gerenciais e para suporte às funções de análise de negócio (COREY, 2001).

O *Data Warehouse* é um "depósito de dados digitais" que serve para armazenar informações detalhadas de uma empresa, criando e organizando relatórios através de históricos que são usados para ajudar a tomada de decisões importantes. Essa tecnologia consiste em organizar em uma base os dados dos diversos sistemas de uma empresa, visando subsidiar a geração de informações úteis aos gerentes e diretores da organização, dando assim suporte a tomada de decisão (COREY, 2001).

As tecnologias OLAP e DW são destinadas a trabalharem juntas, enquanto o DW armazena as informações de forma eficiente, o OLAP deve recuperá-las com a mesma eficiência e maior rapidez.

3. Trabalhos Relacionados

Ao longo dos anos, fornecedores de sistemas educacionais têm desenvolvido métodos e ferramentas para lidar com LA (SCLATER, 2017). Como trabalhos relacionados, pode-se observar algumas iniciativas no sentido de contribuir com a organização dos dados de diferentes bases. A Fundação Apereo[2], por exemplo, é uma fornecedora de *software* educacional que oferece suporte a milhares instituições de ensino em todo o mundo. Entre os sistemas desenvolvidos e mantidos pela fundação está o sistema Sakai[3]. Em 2015, a Apereo iniciou o desenvolvimento de uma visão de plataforma aberta de análise de aprendizagem (APEREO, 2018).

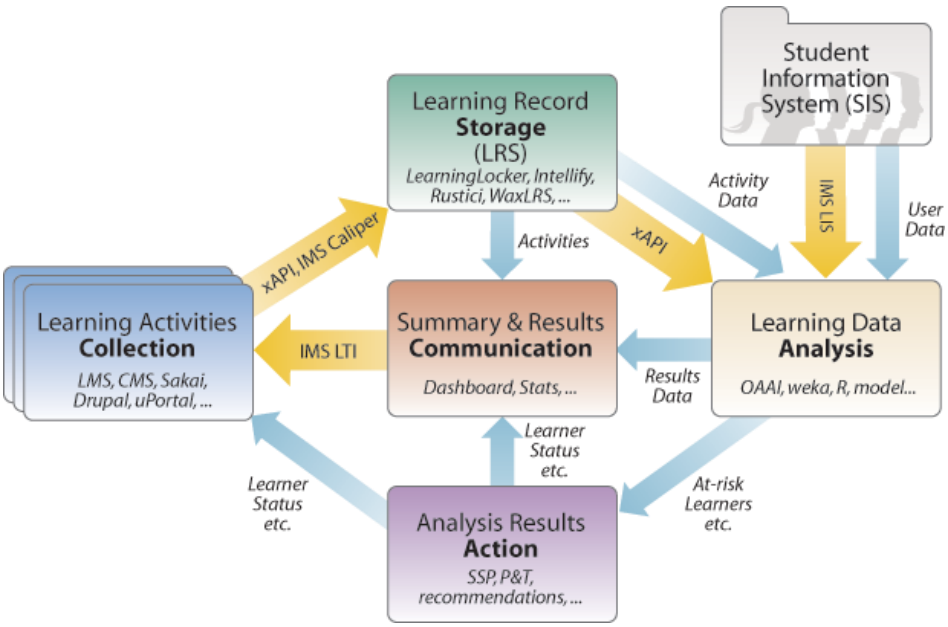

Figura 1. Componentes da Apereo *open learning analytics platform* (APEREO FOUNDATION, 2018)

[2] https://www.apereo.org
[3] https://sakaiproject.org/

Nessa visão são previstos componentes livres de *software* para cada uma das cinco atividades principais de um processo de análise de aprendizagem (APEREO FOUNDATION, 2018):

- ***Collection*** (coleção): componentes de coleções de dados baseados em padrões e tecnologias *open source*, tais como: Experience API[4] (xAPI) e IMS Caliper/Sensor API[5].
- ***Storage*** (armazenamento): um repositório padrão de Armazenamento de Registros de Aprendizagem (LRS - *Learning Record Storage*), baseado no OpenLRS[6].
- ***Analysis*** (análise): um processador de análise de aprendizado que pode manipular mineração de dados, processamento de dados, modelagem preditiva e geração de relatórios.
- ***Action*** (ação): componentes que alimentam a saída das análises e que acionam alertas ou outras ações/intervenções na plataforma.
- ***Communication*** (comunicação): componente de painel (*dashboard*) que exiba a saída do processador de análise de aprendizagem.

Essa visão prevê uma arquitetura baseada em componentes reutilizáveis e *open source* que possibilitem o desenvolvimento escalável e flexível de uma plataforma aberta de LA. As tecnologias selecionadas nessa arquitetura são de código aberto, entre elas: xAPI, IMS Caliper/Senson API, OpenLRS e outros (APEREO, 2018).

Em 2017, a visão da Apereo foi estendida pela empresa JISC[7]. A JISC é uma empresa sem fins lucrativos do Reino Unido, cujo papel é fornecer soluções tecnológicas que apõem o ensino médio e superior (JISC, 2018).

[4] https://xapi.com/

[5] https://www.imsglobal.org/activity/caliper

[6] https://github.com/Apereo-Learning-Analytics-Initiative/OpenLRS

[7] https://www.jisc.ac.uk/

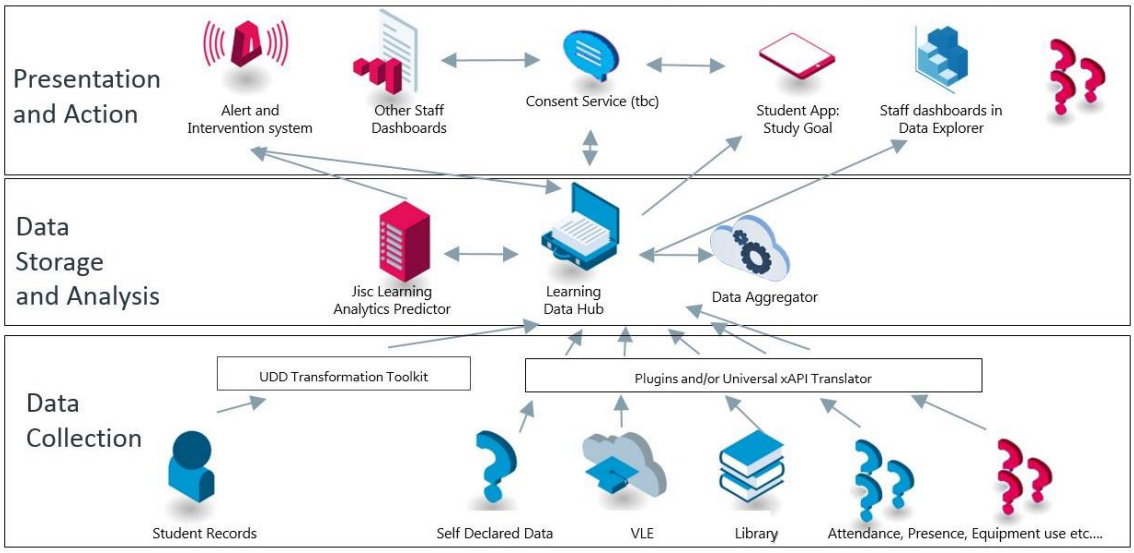

Figura 2. Arquitetura JISC para *Learning Analytics* (JISC, 2018)

Essa empresa propôs uma arquitetura baseada em serviços de software (SaaS — *Software as a Service*) dividida em 3 (três) camadas, que prevê a utilização de componentes (obrigatórios e opcionais) conforme a necessidade da instituição (JISC, 2018):

- ***Data collection*** (Coleta de dados): nesta camada estão dispostas as fontes de dados dos sistemas. A arquitetura divide esses dados em duas categorias: 1. **Dados sobre o estudante**: registram quem é o aluno, que cursos ele está cursando e quais são as suas notas. Esses dados normalmente são mantidos em um sistema de registro de alunos e são coletados em um formato padrão chamado UDD (*Universal Data Definition*); 2. **Dados de atividade**: descrevem o que um aluno faz, esteja ele utilizando um AVA, participando de uma palestra ou usando outro sistema. Isso é coletado em um padrão conhecido como xAPI. A ideia principal é que o mesmo evento seja descrito da mesma maneira, independentemente do sistema utilizado.
- ***Data Storage and Analysis*** (Armazenamento e análise de dados): nesta camada há uma central de armazenamento e processamento de dados educacionais nas nuvens. Cada instituição tem sua própria central de armazenamento e só pode ver seus próprios dados. O processador de dados de aprendizagem é um componente opcional da arquitetura. Ele pode, por exemplo, coletar os dados da central de dados e executar modelos preditivos para determinar se o aluno corre o risco de falhar.
- ***Presentation and Action*** (Apresentação e ação): nesta camada há sistemas que são alimentados pelo processador de dados. Um sistema chave a esta camada é o sistema de *dashboards*. Ele provê uma grande variação de modelos de visualização dos dados analisados, levando em consideração os diferentes *stakeholders* do sistema (alunos, professores, *etc*). Além desse, outro sistema chave é o Sistema de Consentimento, neste os alunos podem decidir que dados gostariam de compartilhar com a instituição e que tipos de intervenção

gostariam de receber. Outros sistemas opcionais à esta camada são: um sistema de intervenção e alerta; um aplicativo mobile para o estudante que provê análises diretamente ao aluno; entre outros.

Essa arquitetura também prevê o uso de um conjunto de componentes *open source* reutilizáveis que possibilitem o desenvolvimento escalável e flexível de uma plataforma aberta de LA. Entre o conjunto de tecnologias utilizadas nessa proposta estão: a xAPI, o UDD e outros. Diferente da proposta da Apereo, a modelo desenvolvido pela JISC prevê uma arquitetura SaaS baseada em 3 (três) camadas de componentes obrigatórios e opcionais, conforme a necessidade da instituição.

4. Contribuição

Ambos os modelos apresentados nos trabalhos relacionados, Apereo (Figura 1) e JISC (Figura 2), preveem o uso de componentes de código aberto e especificações detalhadas para a indexação e tratamento de dados. Embora isso promova a colaboração entre os desenvolvedores e o desenvolvimento de aplicações de LA que utilizam várias fontes de usuário, componentes em diversos níveis de granularidade e voltados para diversos tipos de usuário, dificulta a implementação dessas abordagens em instituições cujos os sistemas em produção não atendem os critérios especificados.

Dessa forma, faz-se necessário um modelo arquitetural mais abstrato, que possibilite as instituições de ensino, implementar aplicações de LA que se utilizem dos modelos de dados e metadados já utilizados institucionalmente.

A proposta deste trabalho é um modelo organizado por meio de uma arquitetura SaaS dividida em 4 (quatro) camadas de aplicação que permita a diferentes *stakeholders* e aplicações LA colaborarem entre si, como mostra a Figura 3.

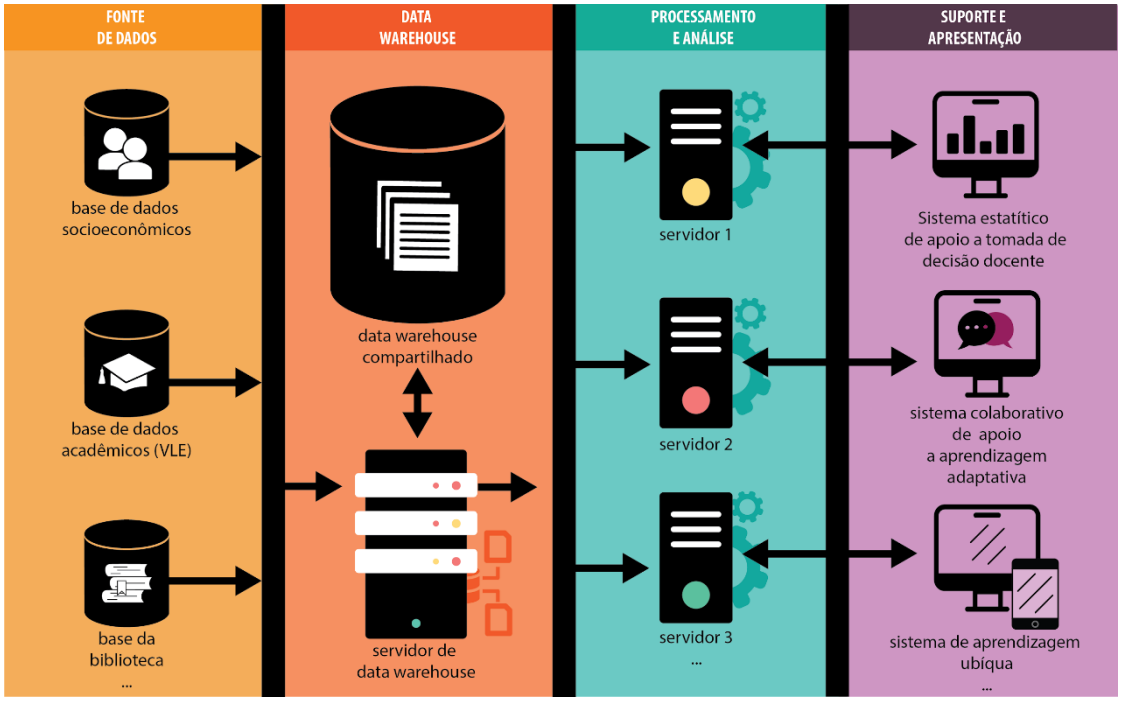

Figura 3. Arquitetura conceitual proposta

Considerando que LA tem sua origem em BI e DM, e que, normalmente, BI usa diferentes fontes de dados para suportar a tomada de decisão e definir estratégias de negócio, o modelo arquitetural apresentado tem como objetivo padronizar a coletar os dados dos diferentes sistemas da instituição, além disso, centralizar, pré-processar os dados coletados e facilitar a colaboração de aplicações OLAP da instituição considerando os modelos de dados e metadados já utilizados institucionalmente, conforme descrito a seguir:

- **Fonte de dados**: camada onde são armazenadas as diversas fontes de dados da instituição, tais como: base do sistema socioeconômico de suporte ao aluno; base do sistema de gerenciamento de aprendizagem; base do sistema da biblioteca, entre outras.
- **Data warehouse**: camada que mantém uma central *online* de DW, gerenciada por um servidor de acesso dedicado à seleção e tratamento primário de dados. Esse DW é um dos componentes centrais deste modelo arquitetural. Periodicamente, o servidor de acesso pode executar, por exemplo, uma rotina de coleta e tratamento primário de dados. Através desta, os dados dos diferentes sistemas de produção da organização são extraídos e, em seguida, normalizados e tratados conforme as políticas de privacidade e segurança da instituição. Ou seja, os dados que comprometem a privacidade e segurança dos alunos são mascarados, impossibilitando a identificação do usuário. Em seguida, esses dados são carregados em um DW que possibilite o acesso de aplicações de terceiros autorizadas pela organização, tais como: ferramentas de análises (OLAP) ou de mineração de dados. Espera-se que um DW elimine a competição dos acessos aos dados de apoio à decisão e os dados dos sistemas em produção da organização.
- **Processamento e análise**: nesta camada estão os servidores de processamento dos sistemas de análise de aprendizagem institucionais. Todo servidor de processamento terá acesso aos dados da central de armazenamento (*Data Warehouse)* de forma padrão e poderá, de maneira autônoma, processar essas informações conforme desejado. Ou seja, cada desenvolvedor compartilha um conjunto padrão de dados e ferramentas e poderá, de forma independente ou colaborativa, implementar suas aplicações de acordo com o objetivo da sua aplicação de LA, nível de granularidade, ou usuário alvo.
- **Suporte e apresentação**: camada de apresentação e suporte ao usuário. Nesta estão interfaces para auxiliar o usuário, por exemplo, na tomada de decisão docente; notificações e intervenções ao processo de aprendizagem; a partir de *dashboards* representam informações úteis sobre os índices de evasão dos alunos da instituição; sistemas *mobile* de acesso estudantil; etc.

De forma a enfrentar esse desafio em um período de dez anos, descreve-se a seguir quatro ações a serem realizadas em conjunto com a implementação e implantação do modelo apresentado (Figura 3):

1. Criar estruturas organizacionais e computacionais que suportem os diferentes processos e aplicações de análise de aprendizagem, tais como: infra estrutura

mínima de servidores de dados (*Data Warehouse*) e aplicações OLAP; e Criação do comitê estratégico de governança de dados da instituição responsável pela criação das políticas institucionais de acesso aos dados educacionais da instituição e responsável pela manutenção do *Data Warehouse*, tendo em vista os objetivos e escopo geral da organização.

2. Desenvolver ferramentas inteligentes para coleta, tratamento e agrupamento dos dados dos diferentes sistemas de produção da instituição. Possibilitando, por exemplo, o agrupamento dos dados em um *Data Warehouse online* que disponibilize os dados educacionais da instituição de maneira organizada e centralizada.
3. Engajar diferentes *stakeholders* a colaborar e desenvolver novas funcionalidades de *learning analytics* que integrem os diversos sistemas institucionais, criando um grupo de pesquisa institucional que reunia os pesquisadores e desenvolvedores (alunos, professores, técnicos administrativos, *etc*) de aplicações de LA e, a partir deste grupo, promover as aplicações desenvolvidas/utilizadas na instituição. Além disso, criar um portfólio de aplicações; realizar eventos institucionais que promovam a capacitação e utilização das aplicações de LA desenvolvidas/implantadas institucionalmente.
4. Desenvolver uma política institucional que garanta os direitos dos alunos em relação a sua privacidade e segurança.

5. Considerações Finais

Neste artigo são apresentadas as dificuldades enfrentadas no desenvolvimento de ferramentas de análise de aprendizagem. É proposto como desafio um modelo de referência que descreva como integrar, de forma eficiente (sem comprometer as funcionalidades dos sistemas), os componentes e ferramentas comuns a automação de um processo de *Learning Analytics*. Por fim, é apresentada uma proposta conceitual e um conjunto de ações que podem possibilitar a diferentes *stakeholders* e ferramentas compartilhar configurações padrão de coleta e acesso aos dados educacionais de uma instituição.

Referências

ALMAZROUI, Yousef A. A survey of Data mining in the context of E-learning. International Journal of Information Technology & Computer Science (IJITCS), v. 7, n. 3, p. 8-18, 2013.

APEREO FOUNDATION. Learning Analytics Initiative. Disponível em: <https://www.apereo.org/communities/learning-analytics-initiative> Acessado em: 23 de março de 2018.

BAKER, R. S. J. D.; YACEF, K. The state of educational data mining in 2009: a review and future visions. Journal of Educational Data Mining, v. 1, n. 1, p. 3-17, 2009.

BAKER, R.; SIEMENS, G. Educational data mining and learning analytics. In: SAWYER, K. (Ed.). The Cambridge handbook of the learning sciences. 2. ed. Cambridge: Cambridge University Press, 2014. p. 253-274.

BLIKSTEIN, Paulo. Using learning analytics to assess students' behavior in open-ended programming tasks. In: Proceedings of the 1st international conference on learning analytics and knowledge. ACM, 2011. p. 110-116.

BORGES, Vanessa Araujo; NOGUEIRA, Bruno Magalhaes; BARBOSA, Ellen Francine. A multidimensional data model for the analysis of learning management systems under different perspectives. In: Frontiers in Education Conference (FIE), 2016 IEEE. IEEE, 2016. p. 1-8.

COREY, Michael. Oracle 8i Data Warehouse: PLANEJE E CONSTRUA UMA SOLUÇAO DE ANALISE RESISTEN. Elsevier Brasil, 2001.

FERGUSON, Rebecca et al. Research evidence on the use of learning analytics: Implications for education policy. 2016.

FERGUSON, Rebecca. Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning, v. 4, n. 5-6, p. 304-317, 2012.

JISC. Learning Analytics: General Overview. Disponível em <https://docs.analytics.alpha.jisc.ac.uk/docs/learning-analytics/General-Overview> Acessado em: 23 de Março de 2018.

JOHNSON, Larry et al. The 2010 Horizon Report. New Media Consortium. 6101 West Courtyard Drive Building One Suite 100, Austin, TX 78730, 2010.

SCLATER, Niall. Learning analytics explained. Taylor & Francis, 2017.

SIEMENS, G., LONG, P.. Penetrating the Fog: Analytics in Learning and Education. EDUCAUSE review, vol. 46, n. 5, p. 31-40, 2011.