UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CÂMPUS CORNÉLIO PROCÓPIO
DEPARTAMENTO ACADÊMICO DE COMPUTAÇÃO
TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

ALEX JUNIOR NUNES DA SILVA

ANÁLISE DE BIG DATA PARA VISUALIZAÇÃO DE MÉTRICAS DOS DOCENTES DO ENSINO SUPERIOR

TRABALHO DE CONCLUSÃO DE CURSO

CORNÉLIO PROCÓPIO
2015

ALEX JUNIOR NUNES DA SILVA

ANÁLISE DE BIG DATA PARA VISUALIZAÇÃO DE MÉTRICAS DOS DOCENTES DO ENSINO SUPERIOR

Trabalho de Conclusão de Curso de graduação, do curso de Tecnologia em Análise e Desenvolvimento de Sistemas da Universidade Tecnológica Federal do Paraná – UTFPR, como requisito parcial para a obtenção do título de Tecnólogo.

Orientador: Prof. Francisco Pereira Júnior
Co-orientador: Rosangela de Fátima Pereira

CORNÉLIO PROCÓPIO
2015

FOLHA DE APROVAÇÃO

"A Folha de Aprovação assinada encontra-se na Coordenação do Curso"

Dedico este trabalho à minha famîlia, que sempre me ensinou que o caráter reto e íntegro é a virtude mais valorosa que se pode ter e o conhecimento é o bem mais importante que se pode adquirir.

AGRADECIMENTOS

Primeiramente agradeço a Deus, pois sem Ele eu nada seria. Dele por Ele e para Ele são todas as coisas.

Agradeço ao meu orientador Prof. Francisco Pereira Júnior, pela sapiência, responsabilidade e paciência com que me guiou nesta trajetória. Também a amiga Rosangela Pereira por me auxiliar, compartilhando do seu vasto conhecimento.

Aos meus colegas de sala de aula e aos meus amigos de viagem que sempre me acompanharam dando suporte nesse trajeto diário.

Aos meus amigos de trabalho que todos os dias tiveram paciência em me ensinar e acompanhar meu progresso de aquisição de conhecimentos.

A Secretaria do Curso e os demais setores da Instituição pela cooperação.

Gostaria de deixar registrado também meu reconhecimento à minha famîlia, principalmente o meu pai pelos conselhos de vida, seu exemplo e por me incentivar a estudar a área de informática. E a minha mãe pela compreensão, e também aos meus dois irmãos, pois acredito que sem o apoio deles seria muito difícil vencer esse desafio.

Enfim, a todos os que por algum motivo contribuíram para a realização desta pesquisa.

“Em todas as coisas o sucesso depende de uma preparação prévia, e sem tal preparação o falhanço é certo.”

Confúcio

RESUMO

SILVA, Alex Junior Nunes. **Análise De Big Data Para Visualização De Métricas Dos Docentes Do Ensino Superior**. 2015. 102 f. Trabalho de Conclusão de Curso (Graduação) – Tecnologia em Análise e Desenvolvimento de Sistemas. Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2015.

Os dispositivos eletrônicos têm se tornado cada vez mais inteligentes. Dados de diversas fontes são gerados a cada segundo, e com o crescimento dos dados digitais em todo o mundo, têm-se cada vez mais a necessidade de gerenciá-los utilizando soluções não-convencionais, e que atendam esta demanda de forma financeiramente viável. Tais soluções são denominadas Tecnologias para Análise de Big Data e são arquiteturas criadas com o objetivo de processar e gerenciar computacionalmente rápida, uma ampla variedade de dados agrupados em grandes quantidades.

Diversas instituições públicas e privadas têm utilizado análise de Big Data como estratégia de gestão. Essas soluções auxiliam na tomada de decisões, que quando orientada a dados, tem demonstrada grande eficiência. Todavia, esses dados precisam ser visualizados de forma eficaz, sendo que esse é um grande desafio em análise de Big Data. Uma correta visualização de dados faz com que essas soluções tenham maior importância no processo de tomada de decisões.

As instituições de ensino superior (IES) possuem problemas no gerenciamento de dados que podem ser resolvidos com o auxílio de uma solução para análise de Big Data. Dessa forma, esse trabalho tem como objetivo resolver problemas nesse contexto. Com o resultado desse trabalho pretende-se criar um banco de dados que fornecerá informações de forma simples e eficaz e elas poderão auxiliar os gestores das IES a obter informações de grande relevância podendo gerar uma melhoria no processo de gestão.

Utilizando as tecnologias para análise de Big Data, foi desenvolvida nesse trabalho uma ferramenta Web para que os gestores das IES possam visualizar informações e indicadores das atividades dos docentes de maneira adequada.

Palavras-chave: Big Data. Hadoop. Plataforma Lattes. sistemas de informações acadêmicas. Instituições de Ensino Superior. dados acadêmicos. Visualização de dados. Pentaho. Kettle. Computação distribuída. Alto desempenho.

ABSTRACT

SILVA, Alex Junior Nunes. **Analysis Of Big Data For Viewing Of Metrics Of Teachers In Higher Education**. 2015. 102 f. Work Completion of course (Graduation) - Technology Analysis and Systems Development. Federal Technological University of Paraná. Cornélio Procópio, 2015.

Electronic devices have become increasingly intelligent. Data from several sources are generated every second and with the growth of digital data worldwide, increasingly have the need to manage them using unconventional solutions, and meet this demand in a financially viable way. Such solutions are known as technologies for Big Data analysis and architectures are created in order to process and manage computationally fast, a wide variety of grouped data in large quantities.

Various public and private institutions have used analysis of Big Data as a management strategy. These solutions assist in decision making, when oriented data has shown greater efficiency. However, this data needs to be viewed effectively, and this is a major challenge in analyzing Big Data. Correct data visualization makes these solutions have greater importance in the decision-making process.

Higher education institutions (HEIs) have problems in data management that can be solved with the aid of a solution for analyzing Big Data. Thus, this paper aims to solve problems in this context. As a result of this work aims to create a database that will provide information simply and effectively, and they can assist managers of IES to obtain very relevant information and may generate an improvement in the management process.

Using the technologies for analyzing Big Data, it was developed in this study a web tool for managers of IES can view information and indicators of the activities of teachers properly.

Keywords: Big Data, Hadoop, Lattes Platform, academic information systems, higher education institutions, academic data, data visualization, Pentaho Kettle, distributed computing, High Performance.

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE FIGURAS

LISTA DE SIGLAS

API	*Application Programming Interface*
BPMN	*Business Process Modeling Notation*
CE	*Community Edition*
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CPF	Cadastro de Pessoa Física
CSS	*Cascading Style Sheets*
CSV	*Comma Separated Value*
DIRGTI	Diretoria de Gestão de Tecnologia da Informação
EAP	Estrutura Analítica do Processo
EE	*Enterprise Edition*
ETL	*Extract Transform Load*
GB	Gigabyte
GFS	*Google File System*
GPS	*Global Positioning System*
HD	*Hard Disk*
HDFS	*Hadoop Distributed File System*
HTML	*HyperText Markup Language*
IBM	*International Business Machines*
IDC	*International Data Corporation*
IES	Instituição de Ensino Superior
JSF	*Java Server Faces*
JSP	*Java Server Pages*
KDD	*Knowledge-Discovery in Databases*
KPI	*Key Performance Indicator*
LHC	*Large Hadron Collider*
MB	Megabyte
MVC	*Model-View-Controller*
ORM	*Object-Relational Map*
PDI	*Pentaho Data Integration*
RDBMS	*Relational Database Management System*
SGBD	Sistema de gerenciamento de banco de dados
SQL	*Structured Query Language*
SVG	*Scalable Vector Graphics*

TB	Terabyte
UTFPR	Universidade Tecnológica Federal do Paraná
XML	*eXtensible Markup Language*

LISTA DE ACRÔNIMOS

RFID	*Computer Aided Engineering*
CERN	*Conseil Européen pour la Recherche Nucléaire*
BI	*Business Intelligence*
YARN	*Yet Another Resource Negotiator*
DOM	*Document Object Model*
JSON	*JavaScript Object Notation*
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
DER	Diagrama de Entidade e Relacionamento
RA	Registro Acadêmico
RAM	*Randomic Acess Memory*

SUMÁRIO

1 INTRODUÇÃO

A quantidade de dados digitais vem crescendo a cada dia. Os dispositivos eletrônicos têm se tornado cada vez mais inteligentes, gerando novos dados a cada instante. Esses novos dados são criados por meio de várias fontes: interações de usuários com redes sociais, sensores, diversos registros de eventos (logs), sistemas transacionais, sistemas de GPS, leitores RFIDs entre outros. A perspectiva é que haja um aumento ainda maior desses dados com a consolidação da "internet das coisas", que é uma previsão do futuro da tecnologia onde vários dispositivos estarão conectados e comunicando-se entre si, mudando o paradigma como a conhecemos e deixando-a cada vez mais importante, útil, indispensável e ubíqua em nosso dia-a-dia.

Reforçando essas afirmações, White (WHITE, 2012) diz que estamos vivenciando a "era dos dados". O Grande Colisor de Hádrons (LHC) do CERN gera para o centro de dados cerca de 82,20 Terabytes por dia, – ou – 30 Petabytes por ano, para apoiar os pesquisadores em seus estudos ("Computing | CERN", 2015). De acordo com um estudo realizado pelo IDC (GANTZ; REINSEL, 2011), no ano de 2010 foram gerados mais de 1 Zettabyte de informações no mundo todo. Os autores em (BRYNJOLFSSON; MCAFEE, 2012) afirmaram que no ano de 2012 foram gerados 2.5 Exabytes de dados a cada dia. Essa massiva quantidade de dados tem se tornado um grande problema computacional, pois não é possível gerenciá-las com eficiência empregando as ferramentas e modelos tradicionais.

Diante desse contexto, pode-se afirmar que, em um cenário global a informação é cada vez mais valorizada (TURKINGTON, 2013). Atualmente, pelo menos duas das dez maiores empresas mundiais de tecnologia, a Google e o Facebook, faturam bilhões de dólares todos os anos vendendo o produto "Informação" ("Facebook Investors", 2014) ("Google Inverstors", 2014) (WINKLER; BARR, 2014). Todavia, para isso, tem-se a necessidade de minerar e aproveitar cada dado criado pelas diferentes fontes e interações com usuários de diversos tipos e com perfis heterogêneos.

Ao encontro dos cenários exibidos, surgem as tecnologias de análise de Big Data, que são soluções desenvolvidas para poder lidar com essas quantidades massivas de dados. A Gartner ("What Is Big Data? - Gartner IT Glossary", 2015) define Big Data como dados com: grande volume, grande velocidade de acesso e leitura e

grande variedade. Além disso, esses dados necessitam de soluções inovadoras e com baixos custos para serem processados e servirem como base para geração de conhecimento e tomada de decisões gerenciais. O termo grande volume de dados, característico de Big Data, é algo relativo, pois o que era considerado um grande volume de dados a alguns anos atrás, hoje pode ser considerado um volume normal. Portanto, a tendência é que sempre haja uma mudança de paradigma nesse conceito. A Figura 1 demonstra alguns marcos importantes na história do Big Data, por meio dela pode-se visualizar o crescimento da quantidade de dados digitais ao longo de quatro décadas.

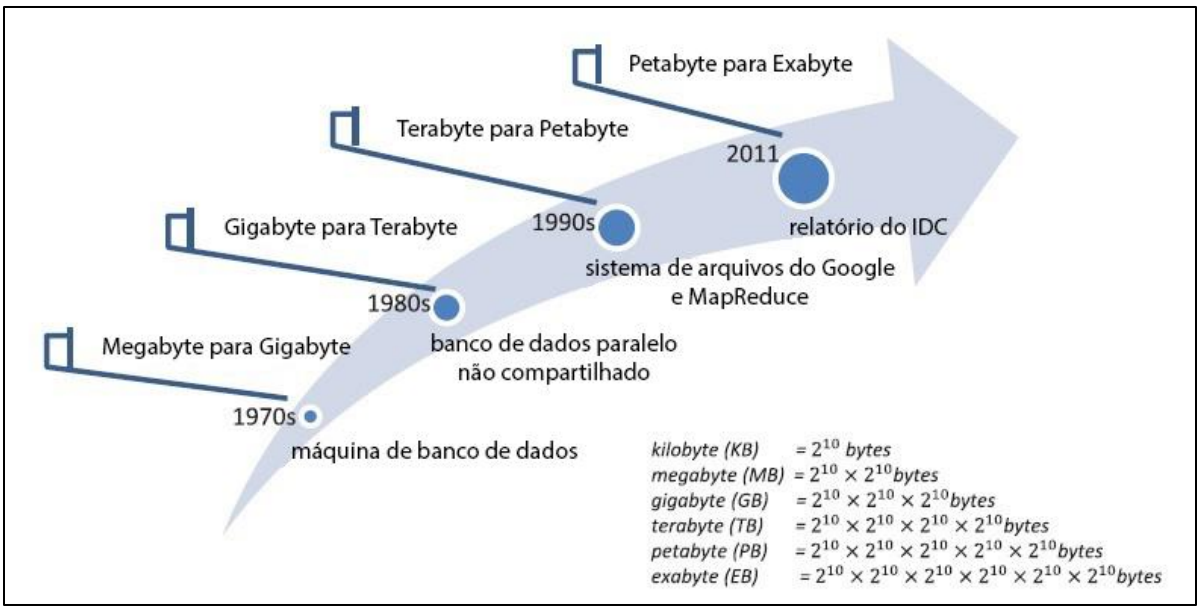


Figura 1 – Uma breve história dos maiores marcos do Big Data
Fonte: Adaptado de (HU et al., 2014).

As soluções de análise de Big Data têm sido largamente utilizadas para auxiliar o processo de tomada de decisões e geração de conhecimento (SRIVASTAVA; DONG, 2013) (TEKINER; KEANE, 2013). A tomada de decisões orientada a dados oferece informações relevantes que na maioria das vezes não são perceptíveis pela análise manual, dada a complexidade e a quantidade dos dados. Visando obter melhores resultados analíticos, em algumas empresas essas soluções são combinadas com soluções de *Business Intelligence* (BI) - soluções computacionais inteligentes que analisam dados históricos e geram novas informações para serem usadas como base para a tomada de decisões gerenciais com maior eficiência -, BI também pode servir como ferramenta para vantagem

competitiva; estratégia para o aumento de vendas e produtividade; antecipação de mudanças do mercado e ações de concorrentes; e a diminuição de prejuízos (MARINHEIRO; BERNARDINO, 2013) (DELSOTO, 2013). Embora os conceitos de BI e Big Data sejam diferentes, ambos podem ser integrados a uma única solução (DELSOTO, 2013).

Outro fator que tem sido um desafio é como visualizar essa quantidade de informação, pois, na maioria das vezes, elas são de tipos diferentes. Em alguns casos, quando possível, torna-se apropriado converter tipos diferentes em um tipo único, para facilitar uma posterior interpretação. Com uma adequada visualização das informações pode-se interpretar resultados de forma eficiente, valorizando a importância dos dados. Uma estratégia para a visualização dos dados é a utilização de *dashboards*, que são telas compostas por camadas de indicadores e mostradores do tipo tabelas, mapas e gráficos, em vários modelos e formatos.

Diversos problemas, em diferentes contextos, podem ser resolvidos com soluções de análise de Big Data, todavia, o foco desta proposta está em problemas vinculados as instituições de ensino superior (IES).

No gerenciamento das instituições de ensino superior (IES) utiliza-se sistemas de informações internos e externos que, alimentados por seus usuários, geram diariamente grandes quantidades de dados. Esses dados, muitas vezes, não são aproveitados de forma eficaz em sua totalidade. Alguns são gerados, armazenados, e posteriormente não são utilizados ou têm uma utilização mínima. Esses mesmos dados quando bem aplicados podem servir de apoio para a geração de uma base de informações consistentes, que os gestores poderão utilizar para perceber situações que antes eram desconhecidas, passando a ter maior domínio sobre vários acontecimentos na instituição. Isso pode melhorar significativamente a eficiência das atividades realizadas.

1.1 PROBLEMATIZAÇÃO

A obtenção de informações e métricas sobre desempenho e produtividade dos docentes não é uma tarefa trivial. As instituições de ensino superior (IES) utilizam vários sistemas de informação diferentes que são alimentados com dados sobre

rotinas de atividades dos profissionais acadêmicos. Além disso, o cruzamento dos dados de diferentes fontes, muitas vezes, é realizado de forma não informatizada ou sem uma ferramenta específica para tal, tornando o processo moroso e com grandes possibilidades de erros na sua obtenção. Os dados para obtenção dos indicadores de desempenho dos docentes até existem, todavia, precisam ser agregados, analisados e apresentados de uma forma eficaz, como por exemplo por meio de gráficos e *dashboards*.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

O objetivo desse trabalho é implantar uma solução de análise de Big Data para apresentar informações e métricas de desempenho do corpo docente de instituições de ensino superior, extraindo e cruzando dados de diferentes fontes. As informações servirão de conhecimento e poderão fornecer subsídios (*insights*) às chefias acadêmicas.

1.2.2 Objetivos Específicos

Para que fosse possível a conclusão desse trabalho as seguintes fases foram realizadas:

- identificar métricas de desempenho condizentes com o cenário acadêmico e de instituições de ensino superior;
- extrair dados de sistemas de informações acadêmicas;
- extrair dados da plataforma Lattes, mais especificamente do Currículo Lattes;
- armazenar e processar os dados obtidos utilizando tecnologias de análise de Big Data;

- apresentar métricas de desempenho de forma clara e objetiva, em formato interativo e dinâmico.

1.3 JUSTIFICATIVA

As informações geradas como resultado desse trabalho são de grande importância para a gestão das instituições de ensino superior. Com essas informações, as IES passam a ter uma visão consolidada do andamento das atividades do corpo docente. Além disso formarão uma base com informações consistentes que servem de conhecimento e poderão fornecer subsídios que podem auxiliar na tomada de decisões baseadas no conhecimento adquirido por meio dos dados.

As informações poderão auxiliar também na adequação às legislações vigentes, pois, como exemplo, as quantidades mínimas e máximas de aulas ou outras atividades acadêmicas serão mais transparentes e poderão ser facilmente observadas.

1.4 ORGANIZAÇÃO

A estrutura desse texto está organizada da seguinte forma: O capítulo 1 descreve a introdução, a problematização, o objetivo geral, os objetivos específicos e a justificativa desse trabalho. O capítulo 2 é composto de pesquisas relacionadas ao tema abordado nesse trabalho. No capítulo 3 estão descritos os procedimentos metodológicos (ferramentas e métodos) que foram aplicados para alcançar o objetivo geral do mesmo. No capítulo 4, algumas discussões obtidas na finalização do trabalho são apresentadas, juntamente com as considerações que agregam informações gerais às atividades descritas nesse contexto.

2 FUNDAMENTAÇÃO TEÓRICA

Nessa seção serão apresentados os principais conceitos utilizados nesse trabalho, para subsidiar o entendimento do contexto ao qual ele é inserido.

2.1 BIG DATA

Nos últimos anos muito tem se falado sobre o termo Big Data (BAO; CHEN, 2014) (TEKINER; KEANE, 2013) (SRIVASTAVA; DONG, 2013) (BRYNJOLFSSON; MCAFEE, 2012) (GOLDMAN et al., 2012) (DELSOTO, 2013) (TURKINGTON, 2013). A definição dos autores em (MANYIKA et al., 2011) diz que Big Data são conjuntos de dados que as tecnologias convencionais não conseguiriam gerenciar, dada a complexidade e tamanho desse conglomerado de dados. Outro fator importante a se considerar em Big Data é a velocidade na obtenção de informações sobre uma massiva quantidade de dados. Em algumas situações uma informação tem que estar disponível em alguns segundos, ou até mesmo em tempo real, e isso reforça ainda mais a necessidade de uma solução não convencional para essa categoria de problemas (GOLDMAN et al., 2012) (NANDIMATH et al., 2013).

Ainda não há uma clara definição quando o assunto é Big Data, pois cada autor cita uma característica diferente e o define de acordo com sua perspectiva (HU et al., 2014). Embora não existam regras e padrões definidos, uma característica comumente citada pelos autores é que para ser Big Data o conjunto de dados deve atender três atributos, denominados como “3Vs”: Volume, Variedade e Velocidade (BRYNJOLFSSON; MCAFEE, 2012). Essa definição é a mais tradicional, e também a mais aceita encontrada na literatura, e é ela que a IBM, a Gartner e alguns pesquisadores da Microsoft utilizam (HU et al., 2014). Todavia, existem outras classificações, que levam em consideração além desses, outros atributos, como por exemplo: a dos “4Vs” que acrescenta o atributo Valor (HU et al., 2014); ou ainda, a classificação dos “5Vs”, que adiciona também a Veracidade a essa lista de atributos (SRIVASTAVA; DONG, 2013).

Grande parte dos dados gerados pelos usuários e por sistemas de informações ou autonômicos são semiestruturados ou não estruturados, como por exemplo, FAX, *email*, fotos, vídeos, *logs* de sistemas (NANDIMATH et al., 2013). Por esses dados não obedecerem uma estrutura estática, definida e previsível, torna-se difícil o armazenamento, o processamento, o cruzamento, a interpretação e a geração de conhecimento sobre eles. Embora os dados tenham uma grande promessa de valor agregado, é necessário usar estratégias adequadas para extração, mineração e visualização das informações que porventura ali existam.

Fortes candidatos a solucionar essas demandas são *frameworks* como: Apache Hadoop, Apache Spark, Apache Giraph, Apache Hama, Apache Storm, GraphLab, entre outros. Essas soluções são consideradas tecnologias de análise de Big Data, pois desde o início foram projetadas para o processamento de um grande volume de dados estruturados, semiestruturados e não estruturados. Dentre os vários existentes, e alguns já citados, destaca-se o Apache Hadoop, que por algum tempo foi considerado o melhor *framework* genérico para essa finalidade, pois: tem o código aberto; pode ser constituído por estruturas de redes e computadores convencionais; possui um modelo de programação simples (*MapReduce*); implementa tolerância a falhas, o que garante continuidade na execução mesmo em condições adversas; e, principalmente, deixa transparente ao usuário as complexidades do ambiente paralelo e distribuído (GOLDMAN et al., 2012). Com o passar do tempo surgiram novas tecnologias e novos *frameworks* para análise de Big Data, mas ainda assim, o Hadoop tem sido largamente utilizado.

2.2 APACHE HADOOP

O Apache Hadoop é um *framework* que permite o armazenamento e processamento de grandes conjuntos de dados. Gerenciado pela empresa Apache Software Foundation, o Hadoop é desenvolvido sobre a linguagem de programação Java e foi inspirado nos projetos: GFS (*Google File System*) e MapReduce, ambos do Google (WHITE, 2012) (GOLDMAN et al., 2012). Ele utiliza um modelo de computação distribuída e de alto desempenho, onde vários nós (*clusters* ou *grids*) podem processar diferentes informações paralelamente, aumentando e potencializando o

poder de processamento com um custo financeiro não tão elevado se comparado com outras soluções para o mesmo fim. Hadoop é um único *framework*, composto por diversos subprojetos, que durante sua evolução, foram desenvolvidos para permitir de forma eficiente e simplificada, resolver os problemas da segurança e integridade dos dados, bem como da tolerância a falhas e da escalabilidade de cada parte (nós) do seu conjunto (GOLDMAN et al., 2012).

O *framework* é mantido com o código aberto (*open-source*), portanto, conta com uma equipe mundial de desenvolvedores que de forma mútua colaboram para a otimização dessa plataforma, melhorando as ferramentas existentes ou desenvolvendo vários novos componentes específicos (*plug-ins*). Há também o apoio de grandes corporações como a Yahoo, o Facebook, o Twitter, o LinkedIn, o The New York Times, a Adobe, o e-Bay, dentre várias outras que utilizam suas soluções ("PoweredBy - Hadoop Wiki", 2015). O Apache Hadoop é composto por vários subprojetos, porém, atualmente, em seu núcleo principal estão quatro: o Hadoop Common, o Hadoop Distributed File System (HDFS), o Hadoop MapReduce e o Hadoop YARN. A Figura 2 demonstra a estrutura básica de uma arquitetura Hadoop onde, na parte inferior está o HDFS que é o responsável pelo armazenamento dos dados de maneira distribuída entre os nós do *cluster*; na parte do meio está o YARN que gerenciará os recursos do *cluster*; na parte superior está o MapReduce, que irá processar os dados em lotes (*batch*) e ao lado dele estão as outras formas de processamento, gerenciamento e acesso a dados, como por exemplo: o Spark, Storm, Hive, Pig, HBase, entre outros.

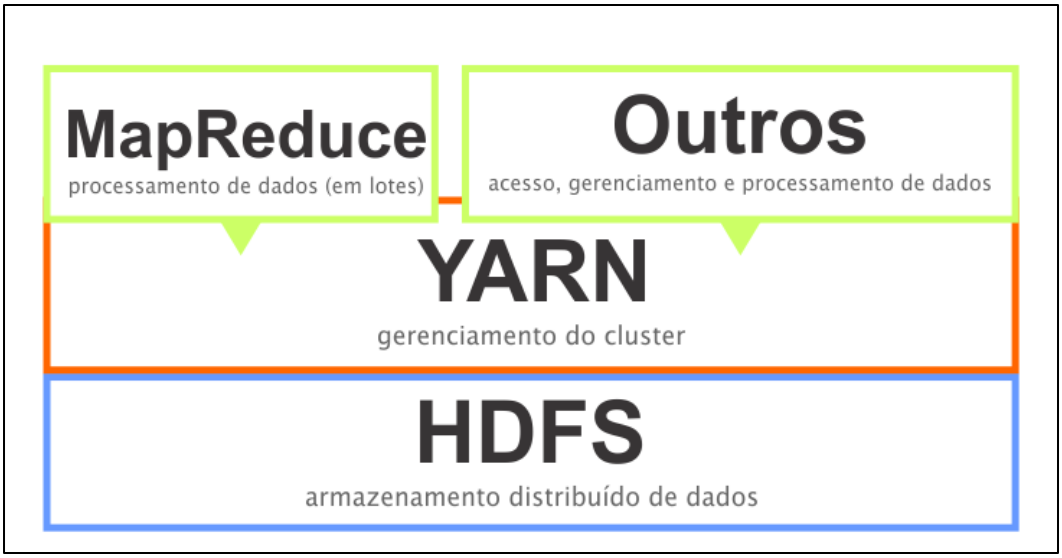


Figura 2 – Estrutura básica de uma arquitetura Hadoop

Fonte: Adaptado de "YARN - The Architectural Center of Enterprise Hadoop" (2015)

O Hadoop, como já mencionado, dispõe de vários pontos positivos: código aberto, programação simplificada (*MapReduce*), tolerância a falhas, transparência em relação as configurações de *clusters* ou *grids*. Esses pontos o fazem ser amplamente utilizado, todavia, para problemas que não podem ser paralelizáveis, ou seja, problemas que não podem ser divididos em partes menores e executados simultaneamente, não é a ferramenta mais indicada, assim como para os problemas com grande dependência entre seus dados. Outros problemas que costumam ser inviáveis para o processamento em Hadoop são aqueles que possuem uma quantidade de dados relativamente pequena pois, os custos para dividir, processar e juntar os dados prejudica o processo, deixando-o moroso (GOLDMAN et al., 2012).

Outro ponto a se destacar é que analisando a arquitetura do Hadoop tanto em White (WHITE, 2012) quanto em Goldman (GOLDMAN et al., 2012), constata-se que há uma grande dependência da localidade dos dados, ou seja, por mais que o HDFS tenha uma visão de todo o dado, o MapReduce processa somente os dados que estão armazenados localmente em cada nó.

2.2.1 Hadoop Common

O Hadoop Common é um módulo do *framework* Hadoop que serve como base para todos os outros subprojetos, ele possui os utilitários e bibliotecas básicas e necessárias para que todo o ecossistema Hadoop funcione corretamente.

2.2.2 Hadoop Distributed File System (HDFS)

Grande parte dos sistemas de informações utilizam banco de dados relacionais (RDBMS - *relational database management system*) para operações com dados transacionais. Porém, em algumas situações, estruturas comuns não conseguem processar um grande volume de informações em um tempo esperado.

Nesse caso, uma solução seria utilizar a computação distribuída, mais especificamente, *clusters* ou *grids*, para dividir o processamento e o armazenamento entre as máquinas (nós). Embora a utilização da computação distribuída seja uma opção, esses sistemas de gerenciamento de banco de dados (SGBD) podem não se comportar de maneira esperada em um ambiente de computação distribuída, sendo que sua implementação na maioria das vezes não é transparente, e é necessário realizar todo o gerenciamento de falhas e a escalabilidade desses SGBDs, o que pode ser muito complexo e trabalhoso. Para esses casos, pode-se utilizar SGBDs ou os sistemas de arquivos distribuídos, como HDFS, que foram criados para esses ambientes.

O HDFS é um sistema de arquivos distribuído. Ele abstrai algumas complexidades de uma "clusterização" e é capaz de deixar invisível ao programador o armazenamento de dados entre múltiplos componentes (nós).

O HDFS se encarrega de fazer a divisão dos dados em partes menores (por padrão, tamanhos fixos de 64 MB), chamadas de blocos, pedaços de entrada, ou somente pedaços (WHITE, 2012). O próprio HDFS realiza a distribuição dos arquivos e o gerenciamento entre os nós de armazenamento (GOLDMAN et al., 2012), esses nós podem ser criados utilizando *hardware*/computadores comuns (GOLDMAN et al., 2012).

2.2.3 Hadoop MapReduce

O Hadoop MapReduce pode ser considerado como o centro do *framework*, é a parte onde os dados são processados. Ele simplifica o desenvolvimento de aplicações paralelas e distribuídas, fazendo com que não sejam necessários conhecimentos avançados nesse tema, uma vez que o próprio Hadoop se encarrega de resolver as suas complexas configurações. Dessa forma, o usuário do Hadoop pode focar seus esforços nas regras de negócios, deixando que o MapReduce se encarregue do processamento (GOLDMAN et al., 2012).

De acordo com uma função definida pelo usuário, os fragmentos de dados (pedaços de entrada) presentes no HDFS são mapeados (*Map*) em cada um dos nós e é atribuída uma chave, ou seja, um valor único que identifica e diferencia esse

fragmento do dado principal. Esses pedaços são então ordenados, armazenados no disco local do nó e depois da execução do mapeamento, uma cópia do resultado de cada um deles é enviada para o HDFS do nó que irá fazer a redução (*Reduce*). Ao receber os pedaços de dados mapeados e ordenados, eles são fundidos e a função de redução é aplicada sobre eles (WHITE, 2012).

A Figura 3 representa o fluxo de processamento do MapReduce com múltiplas tarefas de redução. As caixas maiores, com as linhas tracejadas, simulam a existência de mais de um nó. As setas com as linhas tracejadas, exibem o fluxo de dados dentro de cada nó e as setas com as linhas contínuas exibem o fluxo de dados entre os nós.

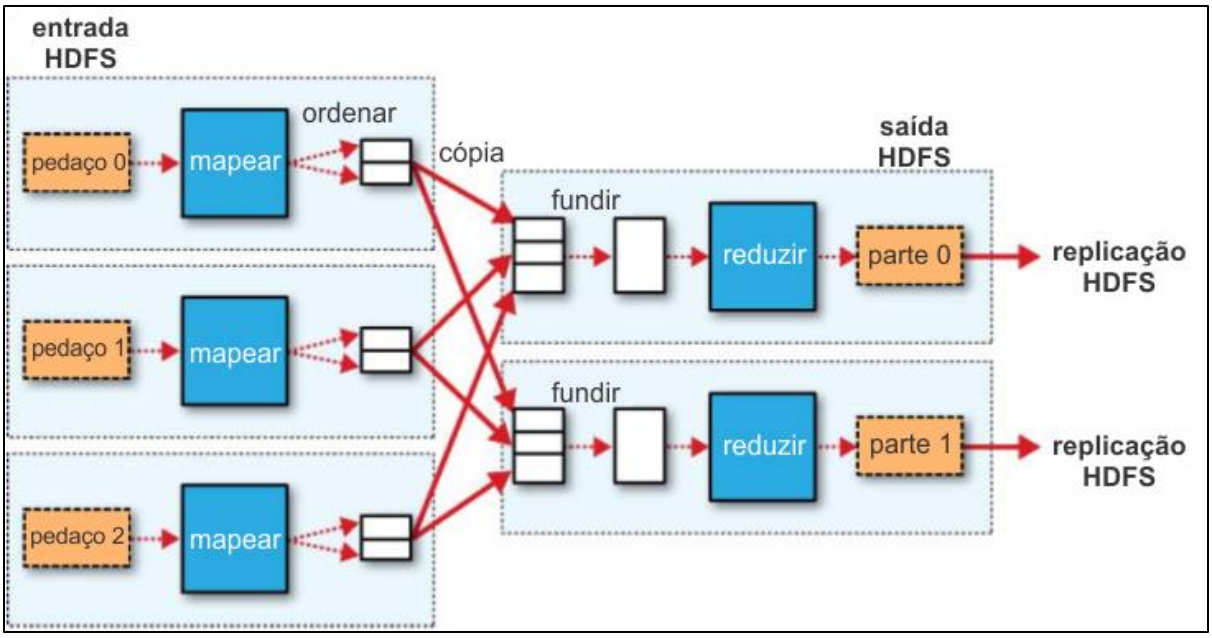


Figura 3 – Fluxo de processamento do MapReduce com múltiplas tarefas
Fonte: Traduzido de WHITE (2012, p. 123).

2.2.4 Hadoop YARN

Na primeira geração do Hadoop, o MapReduce era utilizado como uma aplicação e como um motor (*engine*) de execução, com o *JobTracker* sendo o mestre (*master*) e o *TaskTracker* sendo o escravo (*slave*) (GOLDMAN et al., 2012). O problema é que com esse motor de execução, o Hadoop permitia somente a execução de aplicações MapReduce, ou seja, somente aplicações em lote. Isso foi resolvido

com o YARN (*Yet Another Resource Negotiator*), também chamado de MapReduce 2.0 ou ainda MRv2, que é um módulo presente a partir da segunda geração do Hadoop. O YARN é uma solução para o gerenciamento de *clusters* Hadoop, separando a parte do gerenciamento/monitoramento das tarefas do *cluster* – que antes era feito pelo *JobTracker*; da parte do processamento - que antes era feito apenas pelo MapReduce (VAVILAPALLI et al., 2013).

O YARN é uma camada intermediaria entre o MapReduce e o HDFS, e oferece suporte a múltiplas aplicações e serviços, portanto, com ele, é possível estender o Hadoop não apenas para a utilização com dados em lotes, processados em MapReduce, mas também com dados em tempo real (*on-line*), dados em memória, dados de grafos, e de transmissões (*streaming*). Após o YARN o MapReduce passou a ser apenas uma das aplicações executadas no Hadoop.

2.3 SUÍTE PENTAHO

O Pentaho é um pacote de sistemas de informações utilizado em inteligência de negócios (*Business Intelligence* ou BI). É uma solução que conta com 2 versões: uma gratuita e *open-source* (versão *Comunnity Edition* ou CE) e outra versão paga (versão *Enterprise Edition* ou EE). Desenvolvida com a linguagem de programação Java, a ferramenta é uma das mais usadas e com maior reputação dentre as soluções de BI existentes (MARINHEIRO; BERNARDINO, 2013).

O *Pentaho Data Integration* (PDI), também conhecido como *Kettle*, é o módulo de ETL (*Extraction, Transformation and Load*) da suíte Pentaho. Portanto ele é o responsável por toda a manipulação e estruturação dos dados que as outras aplicações do Pentaho irão utilizar.

Embora o PDI seja parte da suíte Pentaho, ele pode trabalhar sem a necessidade das outras aplicações (*stand-alone*). Dessa maneira, ele pode ser utilizado para realizar os processos de ETL de maneira descomplicada (MARINHEIRO; BERNARDINO, 2013). Por possuir uma integração com soluções Hadoop, o resultado do PDI pode alimentar diretamente soluções de análise de Big Data.

O PDI trabalha com dois tipos de atividades, as transformações (*transformations*), que são uma série linear de operações aplicadas sobre os dados; e os trabalhos (*jobs*), que são os gerenciamentos das transformações. Os trabalhos controlam o fluxo de dados entre as transformações e os possíveis erros que possam acontecer nessas transformações, conforme demonstrado na Figura 4. Tanto nas transformações quanto nos trabalhos, o PDI utiliza componentes visuais para separar e identificar cada estágio de uma atividade, dessa forma, cada componente tem sua característica e é responsável por uma etapa. Por exemplo, analisando a Figura 4, no estágio (SetarVariaveis) as variáveis que serão repassadas para as transformações são configuradas, já no estágio (EmailSucesso) é configurado para que o PDI envie um e-mail para os endereços informados com a mensagem que o *job* foi executado com sucesso. Dessa forma, pode-se observar que etapas completamente diferentes, são separadas em estágios diferentes, o que facilita a manutenção posterior.

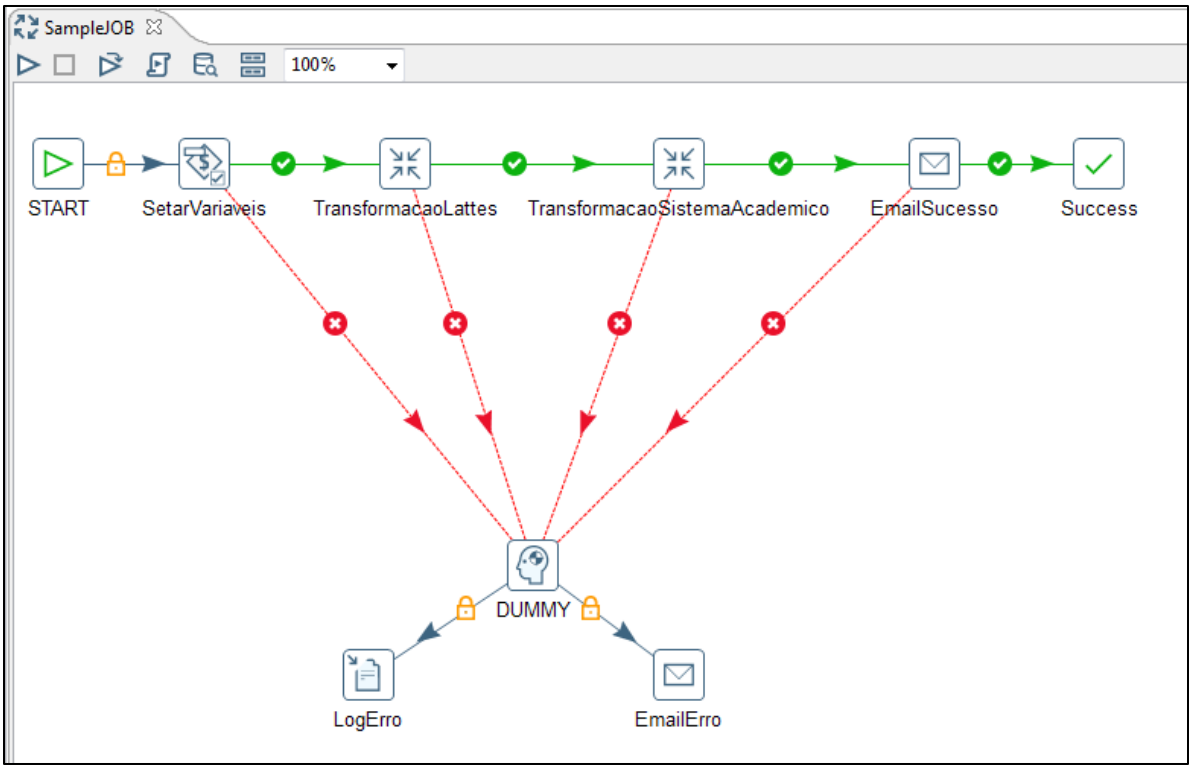


Figura 4 – Trabalho (*job*) gerenciando transformações no PDI
Fonte: Autoria própria

O PDI é um software criado para dados estruturados (CASTERS; BOUMAN; DONGEN, 2010), portanto, não é o ideal utilizá-lo em dados não estruturados. Este trabalho irá utilizar apenas dados estruturados, todavia, em

trabalhos futuros, dados não estruturados poderão ser incluídos, e dessa maneira, o PDI deverá ser substituído por outra ferramenta ou até mesmo utilizar somente o Hadoop.

Embora o Pentaho possua algumas ferramentas para visualização de dados, tabelas e *dashboards*, elas não serão utilizadas nesse trabalho, pois não funcionam fora da arquitetura do Pentaho (*stand-alone*).

2.4 VISUALIZAÇÃO DE DADOS

Muito tem se falado sobre as tecnologias para o processamento de Big Data. Todavia, uma categoria de extrema importância em análise de Big Data é a visualização de dados. Por mais que os dados tenham sido capturados, limpos e processados, ainda sim é necessário visualizá-los de maneira adequada. Uma correta visão pode dar utilidade a dados que se vistos de maneira equivocada podem até atrapalhar ou gerar interpretações incoerentes (FRY, 2008). Muito embora existam linhas de pesquisas específicas sobre visualização de dados, o foco desse trabalho é utilizar estruturas/formas já existentes nas ferramentas.

2.4.1 D3.JS

O D3.js (D3 para *Data-Driven Documents*) é uma biblioteca desenvolvida sobre a linguagem de programação JavaScript, especializada na geração de elementos visuais, muito utilizada para manipular documentos que têm dados como base (documentos orientados a dados). O D3.js combina 3 diferentes padrões *Web*, que são HTML, SVG e CSS para manipular o DOM (*Document Object Model*) - DOM é uma convenção para a representação de objetos em páginas HTML-, isso faz com que os navegadores de internet modernos tenham suporte nativo a execução de soluções criadas com D3.js (MURRAY, 2013). Como fonte de dados o D3.js utiliza arquivos de texto (JSON, XML, CSV) (LEE; JO; KIM, 2014).

(BRYNJOLFSSON; MCAFEE, 2012) afirmam que é necessário medir algo para poder gerenciá-lo, ou seja, para gerenciar uma corporação é necessário medir alguns de seus indicadores. Soluções de Big Data muitas vezes são usadas para medir e comparar indicadores (métricas e KPIs) de empresas. Como o objetivo do D3.js é a visualização de dados (ZHU, 2013), ele pode ser integrado nessas soluções, criando *dashboards* interativos compostos por gráficos e mapas com elevada riqueza visual, o que pode tornar as informações fáceis de serem conhecidas, medidas, compreendidas e interpretadas.

2.5 DADOS ACADÊMICOS

O objetivo dessa seção e abordar os dados presentes no ambiente de instituições de ensino, levando em consideração sistemas de informações internos e externos.

2.5.1 Plataforma Lattes

A Plataforma Lattes é uma plataforma composta por sistemas de informação criada pelo físico Césare Mansueto Giulio Lattes e gerenciada pelo Instituto Stela, em parceria com o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), que é responsável por agrupar bases de dados com informações sobre perfis acadêmicos de pesquisadores brasileiros, grupos de pesquisas e sobre as instituições de ensino superior do país (GUEDES, 2001) (PAPER; CATARINA, 2012). Os dados disponibilizados na plataforma são públicos e ela tem como principal elemento o Currículo Lattes (ALVES; YANASSE; SOMA, 2012).

Os dados são acessados através de páginas e portais Web, portanto, a interação desses dados com os de outros sistemas de informação pode não ser uma tarefa trivial.

O Currículo Lattes é atualmente a principal base de dados de currículos acadêmicos do país. Tem o objetivo de centralizar e padronizar informações e registros de vida pregressa e atual da comunidade científica brasileira, onde arquiva, gerencia e disponibiliza dados de mais de 2.000.000 de perfis de pesquisadores, discentes e docentes, que em sua maioria atuam no Brasil (ALVES; YANASSE; SOMA, 2011). Uma motivação para que essa base esteja sempre atualizada é que os editais das agências de fomento utilizam esses dados para disponibilizar recursos para pesquisadores e grupos de pesquisadores. Ainda, essa base também é utilizada para colher dados para o reconhecimento de cursos de graduação e a avaliação de programas de pós-graduação (FERRAZ; QUONIAM; ALVARES, 2014) (ALVES et al., 2015).

2.5.2 ScriptLattes

O scriptLattes é uma ferramenta de código aberto desenvolvida sobre a linguagem de programação Python utilizada para a extração de informações dos Currículos Lattes. Ele procura e extrai dados de um grupo desejado de pesquisadores por meio das páginas HTML, onde as informações de Currículos Lattes são disponibilizadas de maneira livre, e as armazenam em novos arquivos HTML ou arquivos textos (formato txt e/ou csv) (MENA-CHALCO; JUNIOR, 2009) (FERRAZ; QUONIAM; ALVARES, 2014). Esses novos arquivos gerados pelo scriptLattes apresentam os dados pré-minerados, tratados e agrupados, podendo exibir as produções bibliográficas, artísticas e técnicas, os projetos de pesquisa e orientações, os títulos, prêmios, citações e coautorias, mapas de geolocalizações e redes de colaborações entre os pesquisadores cadastrados (MENA-CHALCO; JUNIOR, 2009) (CHALCO, 2014).

O sistema scriptLattes é utilizado para gerar as extrações de dados e as visualizações das informações nele, todavia, suas informações extraídas podem servir como base para outros sistemas.

2.5.3 Sistema Acadêmico

O Sistema Acadêmico[1] da Universidade Tecnológica Federal do Paraná é um sistema de informação gerenciado pela Diretoria de Gestão de Tecnologia da Informação (DIRGTI) da UTFPR. Ele funciona sobre uma plataforma Web e concentra dados do ecossistema da instituição de ensino. Nesse sistema os docentes realizam lançamentos de frequência dos alunos, cadastram os conteúdos que estão sendo aplicados em aula e lançam as notas finais individuais de cada aluno. As matrículas dos alunos, as composições de turmas e informações sobre o curso também são cadastradas nesse sistema, portanto, com os dados arquivados nele, pode-se extrair diversas informações do ambiente acadêmico. O sistema acadêmico utiliza um RDBMS como estrutura de armazenamento de dados.

2.6 MÉTRICAS

Muitas vezes é necessário que gerentes e gestores de corporações estabelecerem limites e/ou metas, ou mesmo medir e/ou determinar as variações no desempenho de algumas atividades ou processos. Para que esses objetivos sejam alcançados, uma estratégia é definir e utilizar métricas. Elas podem auxiliar na exibição do estado do processo; anteceder de maneira precisa tendências ou problemas que possam vir a acontecer; estimar tempo; energia de trabalho; e, custos dos projetos. Portanto, as métricas são utilizadas para manter os interessados informados e inteirados sobre o estado e o andamento das tarefas relacionadas aos seus projetos, sendo que elas podem ser fundamentais para o seu sucesso (KERZNER, 2013).

Kerzner (KERZNER, 2013) diz que uma métrica deverá possuir algumas características básicas como ter um propósito/alvo, disponibilizar informações úteis e refletir o verdadeiro estado do projeto. Ele ainda afirma que embora existam vários tipos de métricas, pode-se subdividi-las em apenas duas:

[1] Pode ser acessado através do link https://sistemas.utfpr.edu.br/

- os indicadores de resultados (*Result Indicators* ou RIs), que indicam o que foi realizado; e,
- os indicadores de desempenho (*Key Performance Indicators* ou KPIs), que indicam a capacidade de performance de uma empresa.

Resumidamente, métrica pode ser utilizada como um termo genérico, já o KPI geralmente é utilizado para algo mais específico. Uma grande diferença é que na primeira, o foco da informação está no tempo presente; já no segundo, KPI, o foco da informação está no futuro, geralmente fornecendo predições sobre um determinado indicador (KERZNER, 2013).

Tanto os KPIs quanto as métricas podem ser exibidas em *dashboards*, *scorecards* e relatórios, para que suas visualizações e interpretações sejam fáceis (KERZNER, 2013).

2.6.1 Métricas Acadêmicas

Pode-se entender como métricas acadêmicas os indicadores referentes ao ensino, a pesquisa e a extensão, essas métricas nem sempre são fáceis de serem obtidas, todavia, são de grande importância para a gestão acadêmica. Pode-se mensurar dados como: a quantidade de aulas ministradas por cada docente; quantidade de orientações (trabalhos de graduação, estágios curriculares obrigatórios, especializações, mestrados, doutorados); publicações; quantidade de projetos externos às IES; consultorias em empresas; bancas de apresentações e defesas de trabalhos.

Os dados acima citados podem ainda ser refinados, como por exemplo a quantidade de aulas ministradas por cada docente. Essas aulas são diferentes umas das outras de acordo com o assunto, o período do curso que elas estão sendo aplicadas ou ainda a quantidade de alunos da turma. Portanto, pode-se aplicar pesos diferentes em cada uma dessas aulas, qualificando os dados e, tornando mais apurada a análise dessas métricas.

No quesito pesquisa, já existem alguns indicadores mais formais pois, para classificar as revistas científicas existentes no Brasil, a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), desenvolveu o Qualis, que

é um índice que varia do nível mais baixo (C) até o mais elevado (A1). Esses níveis categorizam as revistas científicas de acordo com critérios específicos da área de pesquisa na qual ela atua (OLIVEIRA et al., 2015). O Qualis pode ser utilizado como uma métrica acadêmica, podendo de acordo com a classificação (estrato) Qualis, diferenciar as publicações de um pesquisador ou de um grupo de pesquisadores.

Tanto o Qualis quanto as demais métricas podem compor bases de dados, que por sua vez podem servir de base para a geração de gráficos e *dashboards* quem auxiliam os gestores na extração de conhecimento (*knowledge-discovery in databases* ou KDD) sobre as atividades acadêmicas praticadas pelos docentes de uma IES.

2.7 TRABALHOS RELACIONADOS

Nessa seção são apresentados os trabalhos e pesquisas com os assuntos relacionados a esse trabalho.

2.7.1 Stela Experta

A Stela Experta é uma plataforma comercializada pela empresa TEKIS Tecnologias Avançadas Ltda e criada pelo instituto Stela, que é o mesmo desenvolvedor da Plataforma Lattes. Ela tem por objetivo apoiar a tomada de decisões nas IES, fornecendo dados estratégicos sobre os currículos dos pesquisadores cadastrados na Plataforma Lattes (PAPER; CATARINA, 2012).

A Stela Experta é uma plataforma para conhecimento de alguns KPIs, todavia, ela utiliza apenas os dados cadastrados nas bases do Currículo Lattes. Dessa maneira, caso seja necessário realizar o cruzamento com dados de outros sistemas, o processo não é simples ou trivial, e na medida em que a quantidade desses sistemas ou os dados presentes neles vai aumentando, o processo pode ficar bem mais complexo ou até impossível de se fazer sem uma ferramenta específica. A Figura 5 apresenta a tela principal do Stela Experta onde é possível verificar as informações que podem ser obtidas através desse sistema, na Figura 6 é exibido o menu de acesso

a todas as funcionalidades do Stela Experta, pode-se ver que não existe informações que venham de outros sistemas, apenas dados do Currículo Lattes.

Figura 5 – Tela principal do Stela Experta
Fonte: STELA EXPERTA (2015).

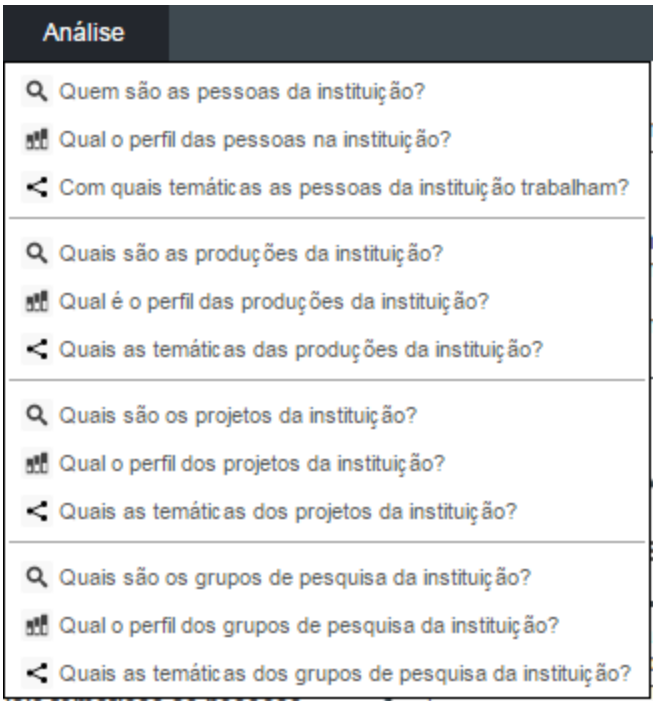


Figura 6 – Menu de acesso as funcionalidades do sistema Stela Experta
Fonte: STELA EXPERTA (2015).

A Figura 7 demonstra na parte superior um gráfico do tipo barras e na posterior outro do tipo colunas, com informações acerca do perfil das pessoas que

compõem o ambiente de uma IES. Uma limitação do sistema Stela Experta é que seus gráficos são gerados no formato (*flash*) com a extensão *Shockwave Flash File (SWF)*, portanto, não é possível faze-los funcionar em sistemas operacionais móveis modernos (Android, IOS, Windows Phone) presentes em *tablets* e *smartphones*, pois esses, não dão suporte para o *Adobe Flash Player* que é o *plugin* responsável por poder reproduzir esses documentos.

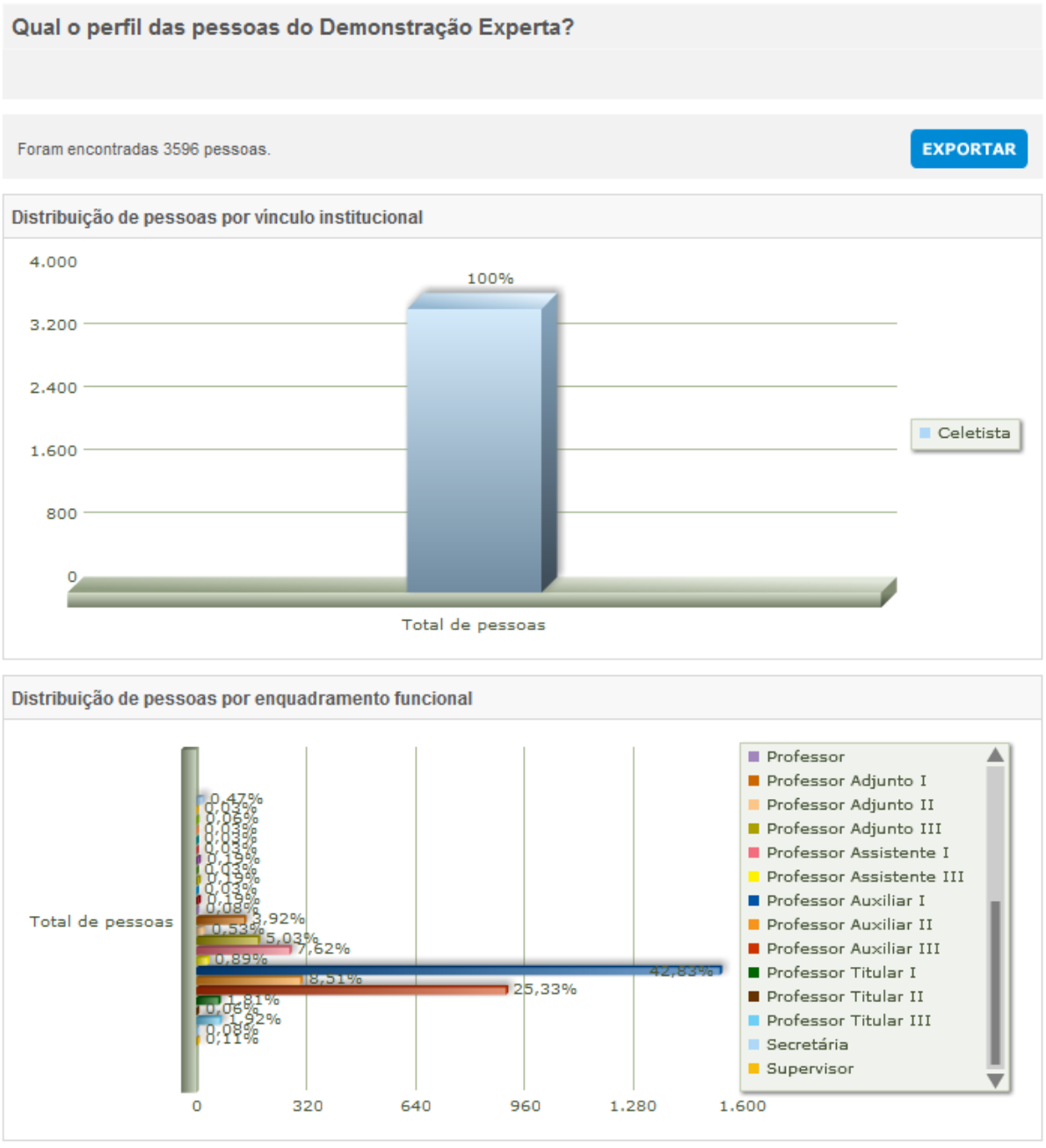


Figura 7 – Gráficos sobre perfil de pessoas do sistema Stela Experta

Fonte: STELA EXPERTA (2015).

2.7.2 Outros Estudos Relacionados

Em (ALVES; YANASSE; SOMA, 2011) os autores apresentam o Sucupira, que é um sistema de informação criado para extrair informações do Currículo Lattes e apresentá-las aos usuários finais. O principal objetivo desse sistema é criar uma rede social de pesquisadores relacionando-os de acordo com áreas de pesquisas, regiões e publicações. Através desse estudo pode-se constatar a eficiência do sistema para a integração de dados do Currículo Lattes. Entretanto, uma limitação encontrada é que ele só extrai informações dessa plataforma, ou seja, como seu foco é muito específico para a plataforma, ele não consegue responder algumas perguntas em relação à produtividade de docentes que estão externas ao Currículo Lattes. Dessa forma, fica faltando uma opção para cruzar esses dados de Currículo Lattes com os de outros sistemas de informações internos ou externos das instituições de ensino superior. A Figura 8 demonstra uma consulta acerca dos pesquisadores da IES UTFPR, nela pode-se verificar que é possível filtrar os docentes de acordo com o ano de atuação, a IES e programas, ou ainda selecionar um docente específico. Os dados exibidos são o nome completo do docente e quais as categorias que ele atua na IES, e esse relatório pode ser utilizado quando deseja-se saber quais os docentes que compõem um determinado programa.

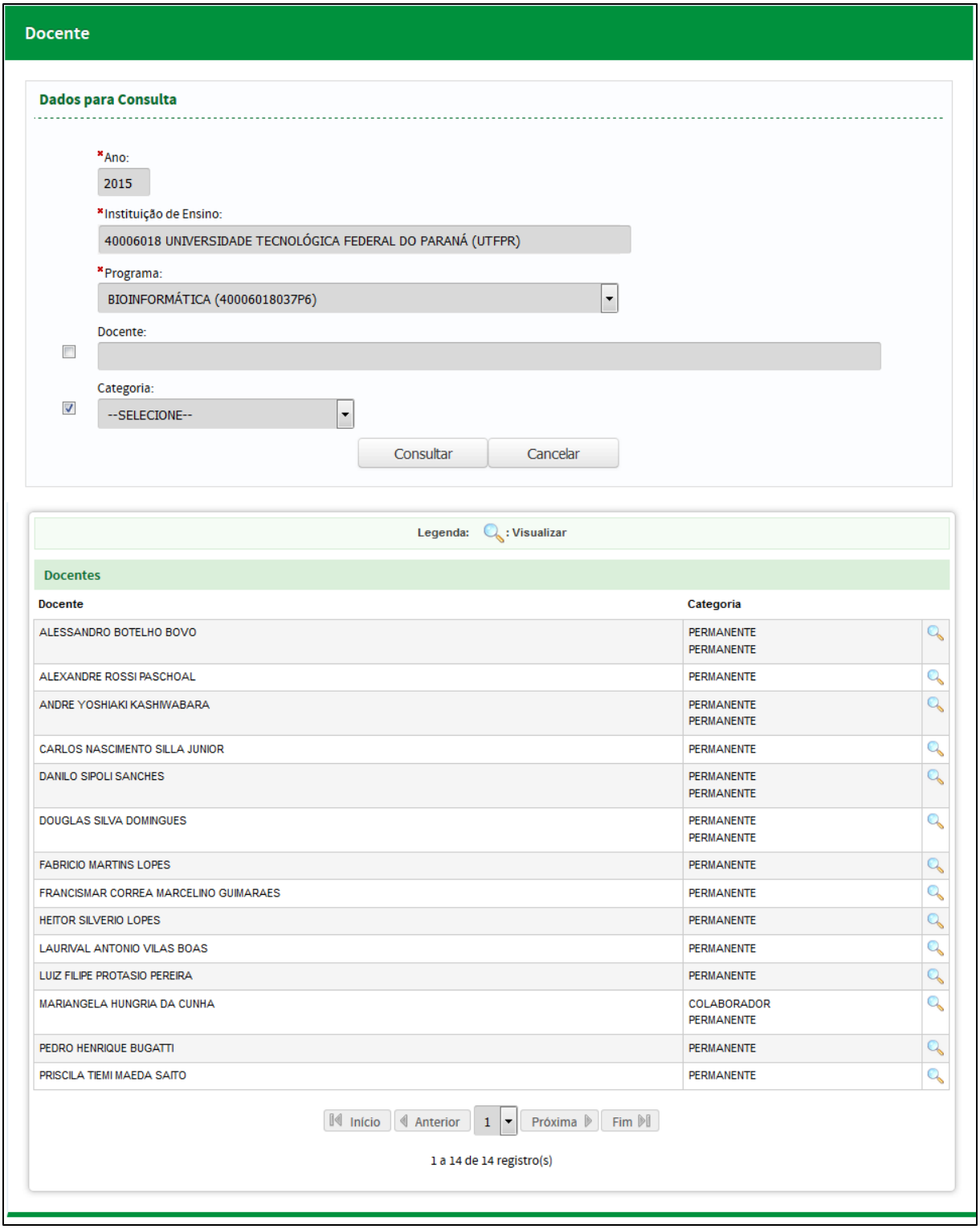
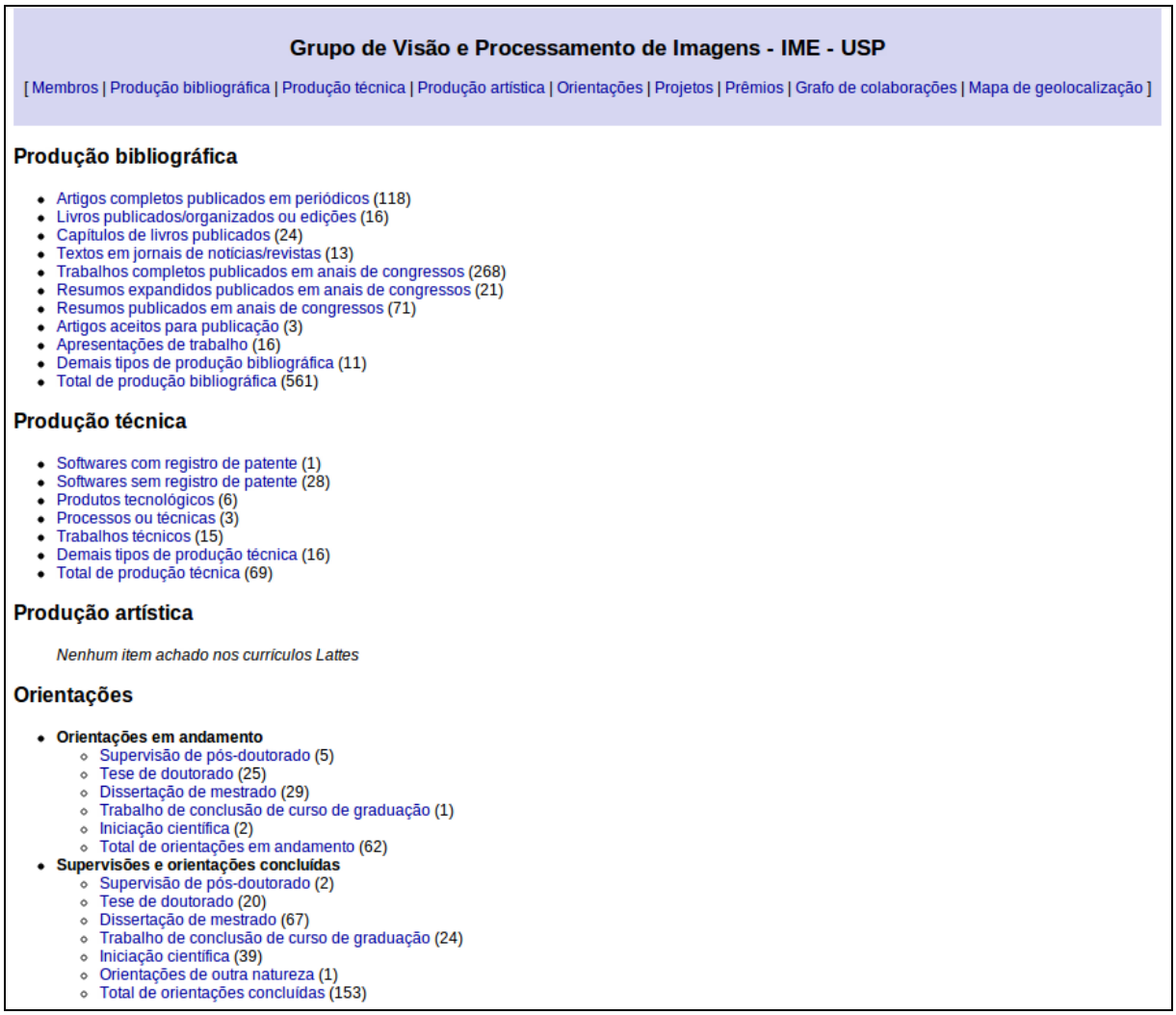


Figura 8 – Consulta de pesquisadores da UTFPR na plataforma Sucupira
Fonte: CAPES (2015)

Em (MENA-CHALCO; JUNIOR, 2009) os autores apresentam a ferramenta scriptLattes, explicando o objetivo de criação da mesma. Nesse artigo os autores demonstram a estrutura do scriptLattes, explanando como ele faz a seleção e o

processamento de dados do Currículo Lattes, o tratamento das redundâncias, a geração dos grafos de colaborações, geração de mapas geográficos de pesquisadores e a geração de relatórios no sistema. O sistema scriptLattes é utilizado para extrações de dados e as visualizações das informações, todavia, suas informações extraídas podem servir como base para outros sistemas. A Figura 9 apresenta um exemplo de resultado gerado pelo scriptLattes, as produções acadêmicas do grupo de pesquisadores sobre o tema de Visão e Processamento de Imagens do Instituto de Matemática e Estatística (IME) da Universidade de São Paulo (USP). Com base na Figura 9, pode-se analisar e destacar que o sistema é bastante intuitivo e as informações são exibidas sumarizadas por categorias e subcategorias que torna o sistema bastante prático e útil para a exibição dos dados do Currículo Lattes.



Grupo de Visão e Processamento de Imagens - IME - USP

[Membros | Produção bibliográfica | Produção técnica | Produção artística | Orientações | Projetos | Prêmios | Grafo de colaborações | Mapa de geolocalização]

Produção bibliográfica

- Artigos completos publicados em periódicos (118)
- Livros publicados/organizados ou edições (16)
- Capítulos de livros publicados (24)
- Textos em jornais de notícias/revistas (13)
- Trabalhos completos publicados em anais de congressos (268)
- Resumos expandidos publicados em anais de congressos (21)
- Resumos publicados em anais de congressos (71)
- Artigos aceitos para publicação (3)
- Apresentações de trabalho (16)
- Demais tipos de produção bibliográfica (11)
- Total de produção bibliográfica (561)

Produção técnica

- Softwares com registro de patente (1)
- Softwares sem registro de patente (28)
- Produtos tecnológicos (6)
- Processos ou técnicas (3)
- Trabalhos técnicos (15)
- Demais tipos de produção técnica (16)
- Total de produção técnica (69)

Produção artística

Nenhum item achado nos currículos Lattes

Orientações

- **Orientações em andamento**
  - Supervisão de pós-doutorado (5)
  - Tese de doutorado (25)
  - Dissertação de mestrado (29)
  - Trabalho de conclusão de curso de graduação (1)
  - Iniciação científica (2)
  - Total de orientações em andamento (62)
- **Supervisões e orientações concluídas**
  - Supervisão de pós-doutorado (2)
  - Tese de doutorado (20)
  - Dissertação de mestrado (67)
  - Trabalho de conclusão de curso de graduação (24)
  - Iniciação científica (39)
  - Orientações de outra natureza (1)
  - Total de orientações concluídas (153)

Figura 9 – Resultado do scriptLattes. Publicações de um grupo de pesquisadores do IME-USP
Fonte: (CHALCO, 2014)

3 PROCEDIMENTOS METODOLÓGICOS

A metodologia a seguir foi criada com base nos “7 estágios de visualização de dados” de Ben Fry (FRY, 2008), adaptada a um projeto para análise de Big Data.

O estudo de Fry descreve 7 etapas para um projeto de visualização de dados:

- aquisição (obtenção)
- estruturação
- filtragem
- mineração
- representação
- refinamento
- interação. A etapa de refinamento foi abstraída nesse projeto de Big Data pois as etapas de apresentação e de interação contemplam de maneira intrínseca o assunto.

Além das 6 etapas extraídas do estudo de Fry, foi incluída a etapa de armazenamento, que é de suma importância para problemas de Big Data.

3.1 DESCRIÇÃO DO AMBIENTE

Para a execução desse trabalho os computadores utilizados foram: um computador desktop, com processador Intel™ Core™ I5, 6 GB de memória RAM e HD de 1 TB e com o sistema operacional Linux CentOS7; e um notebook com processador Intel™ Core™ I3, 4 GB de memória RAM e HD de 500 GB, com o sistema operacional Microsoft™ Windows™ 10.

O *hardware* utilizado para o processamento dos dados nesse trabalho, foi suficiente, todavia, como o trabalho aborda análise de Big Data e os dados podem crescer exponencialmente, outras estruturas de hardware devem ser utilizadas.

No computador desktop o sistema operacional utilizado foi o Linux Centos7 no notebook, Microsoft Windows 10. Tanto para o gerenciamento do processo, quanto para o desenvolvimento, os softwares usados foram:

- XMind 6 versão 3.5.3, para criação das EAPs;
- Kanbanize Web, para criação do quadro de Kanban;
- o PDI na versão 6.0, para manipulação dos dados;
- o scriptLattes na versão 8.10, para extração de dados do Currículo lattes;
- Hadoop na versão 2.6, para tratamento e processamento dos dados;
- D3.js versão 3.5.6, para a criação dos gráficos;
- Bootstrap 3.3.5 para a criação de páginas com os *dashboards*.

Todos os softwares utilizados foram adquiridos em suas versões oficiais e o ambiente de trabalho foi configurado exclusivamente para esse trabalho. O PDI é executado em um computador com o sistema operacional Linux com Máquina Virtual Java (Java *Virtual Machine* ou JVM) versão 8 *update* 66, com o Kit de Desenvolvimento Java (JDK) na versão 8u65.

Também, neste mesmo computador, foi instalado e configurado o Apache Hadoop. A configuração se deu no modo pseudo-distribuído, onde todas as configurações para simular o ambiente de um *cluster* são empregadas, porém, todo o processamento é realizado localmente por apenas um nó.

A Figura 10 exibe um diagrama contendo as ferramentas utilizadas em cada etapa do processo de desenvolvimento, onde, da esquerda para a direita destacam-se: a direção do fluxo que o processo tem, qual a etapa da metodologia é abordada e quais as ferramentas utilizadas naquela etapa específica. A etapa de gerenciamento está presente em todas as outras etapas, por esse motivo, ela é destacada no diagrama com uma cor diferente das demais. Embora ela não faça parte explicitamente das 7 etapas do projeto, a etapa de gerenciamento também utiliza softwares como ferramentas para sua administração.

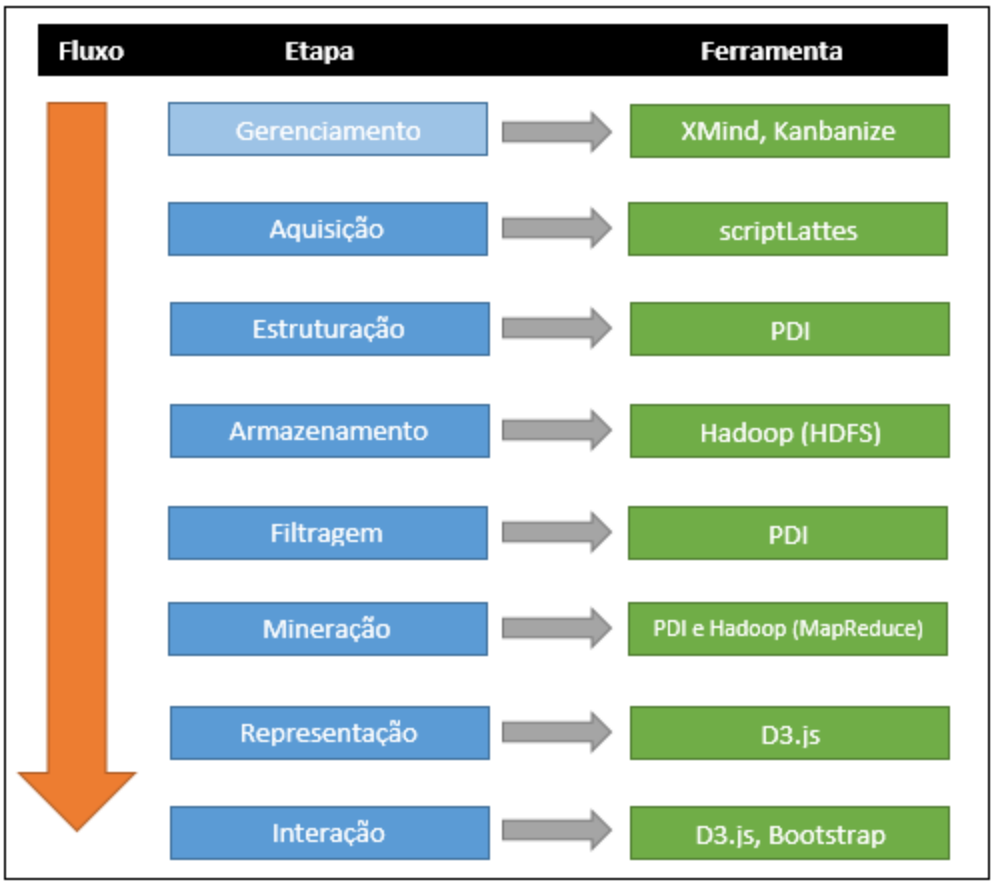


Figura 10 – Ferramentas utilizadas em cada etapa do processo de desenvolvimento

3.2 GERENCIAMENTO DO PROJETO

Para o gerenciamento do processo de análise de Big Data desse trabalho, foi criada uma Estrutura Analítica do Projeto (EAP) indicando de maneira linear, cada uma das fases a serem realizadas até a conclusão do projeto, como pode ser visto na Figura 11. As atividades mais específicas dessa EAP tornaram-se atividades que compuseram um quadro eletrônico de Kanban, utilizando a ferramenta Kanbanize[2]. Kanban são quadros compostos com cartões que representam uma tarefa (*task*) do projeto a ser realizada e acompanhada. A Figura 12 apresenta o quadro de Kanban com todas as tarefas que compõem o processo no estado de reservadas (*backlog*).

Com o Kanban, cada etapa pôde ser facilmente visualizada e o processo acompanhado nas 4 colunas: a coluna de reserva (*backlog*), onde todos os processos iniciaram; a coluna de "a fazer", com os processos prontos para serem realizados; a

[2] Pode ser acessada através do site http://kanbanize.com

coluna “fazendo”, com os processos em execução; e por fim, a coluna “feita”, com os processos finalizados. Cada etapa do projeto, descritas abaixo, conterá a EAP do processo e o quadro de Kanban composto por cada uma das tarefas.

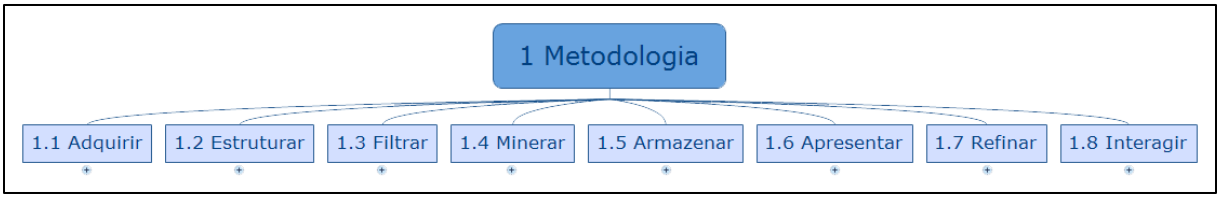


Figura 11 – EAP geral do processo

Figura 12 – Quadro de Kanban exibindo as atividades do processo

3.3 PROCESSO

A seguir é apresentado o processo utilizado, dividido em 7 etapas e exibindo o objetivo e atividades existentes em cada uma delas.

3.3.1 Etapa 1 - Adquirir

O primeiro passo para um projeto de análise de Big Data é a obtenção dos dados. Nessa etapa foi identificada: quais as fontes de dados seriam utilizadas; e, quais dados seriam coletados dessas fontes. Para melhor organização dos dados coletados, definiu-se a nomenclatura padrão dos arquivos de entrada e uma estrutura de diretórios que os organizaram de acordo com alguma regra predeterminada (data, fonte, formato, propósito, etc).

A EAP com a visão linear referente a primeira etapa (adquirir) é demonstrada na Figura 13, composta por 3 tarefas de menor granularidade que formam 3 tarefas em quadro de Kanban, conforme Figura 14.

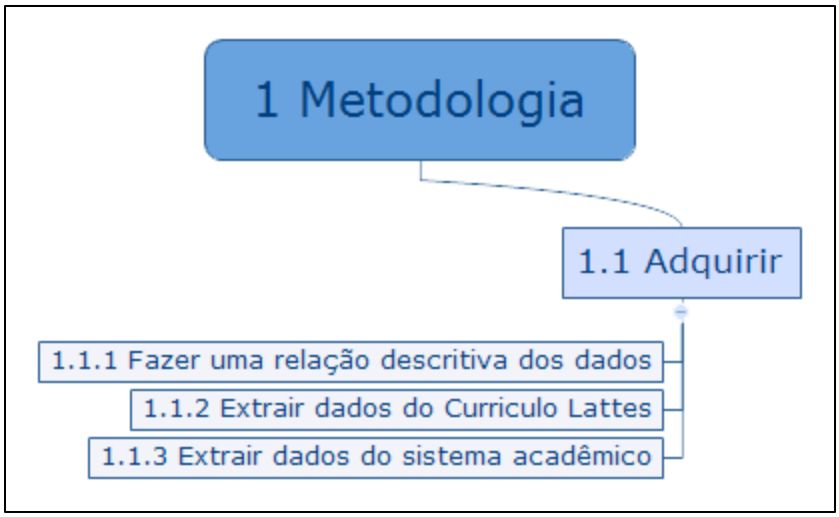


Figura 13 – EAP da etapa de aquisição de dados

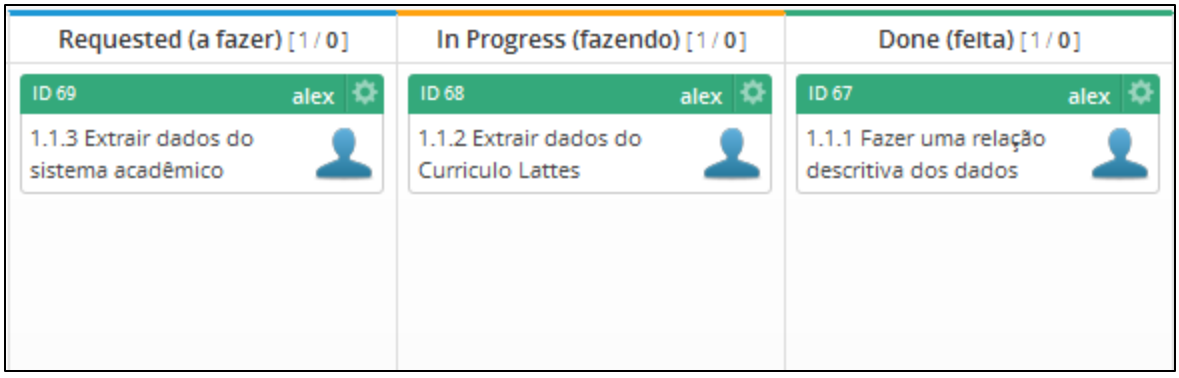


Figura 14 – Quadro de Kanban da etapa de aquisição

Para essa etapa os seguintes resultados foram entregues:

- Relação descritiva dos dados
- Arquivos de dados adquiridos para a análise

O Quadro 1 exibe os resultados da etapa de aquisição, mostrando a relação descritiva dos dados.

Nome do arquivo	Fonte	Tamanho	Formato	Localização	Informações coletadas
publicacoesPorMembro	scriptLattes	6,97 KB	CSV	/dados/lattes	Publicações em periódicos e congressos
sis_academico	Sistema acadêmico	Desconhecido	(SGBD)	127.0.0.1:5432	Aulas ministradas e frequência de alunos

Quadro 1 – Relação descritiva dos dados utilizados no processo

No processo de aquisição dos dados do Currículo Lattes foi utilizado o software PDI, esse processo é demonstrado na Figura 15. No primeiro estágio (csvPublicacoesPorMembro) é apontado a localização do arquivo físico CSV que será utilizado. Caso acontença um erro em algum nesse estágio o fluxo de dados é interrompido, o estágio (juntaErros) recebe uma chamada e envia para o estágio (logDeErro) que gera um arquivo texto com os logs informando detalhes da falha ocorrida no processo. Esse log gerado pode servir como base para que o administrador do sistema possa realizar, posteriormente as devidas manutenções ou correções no processo.

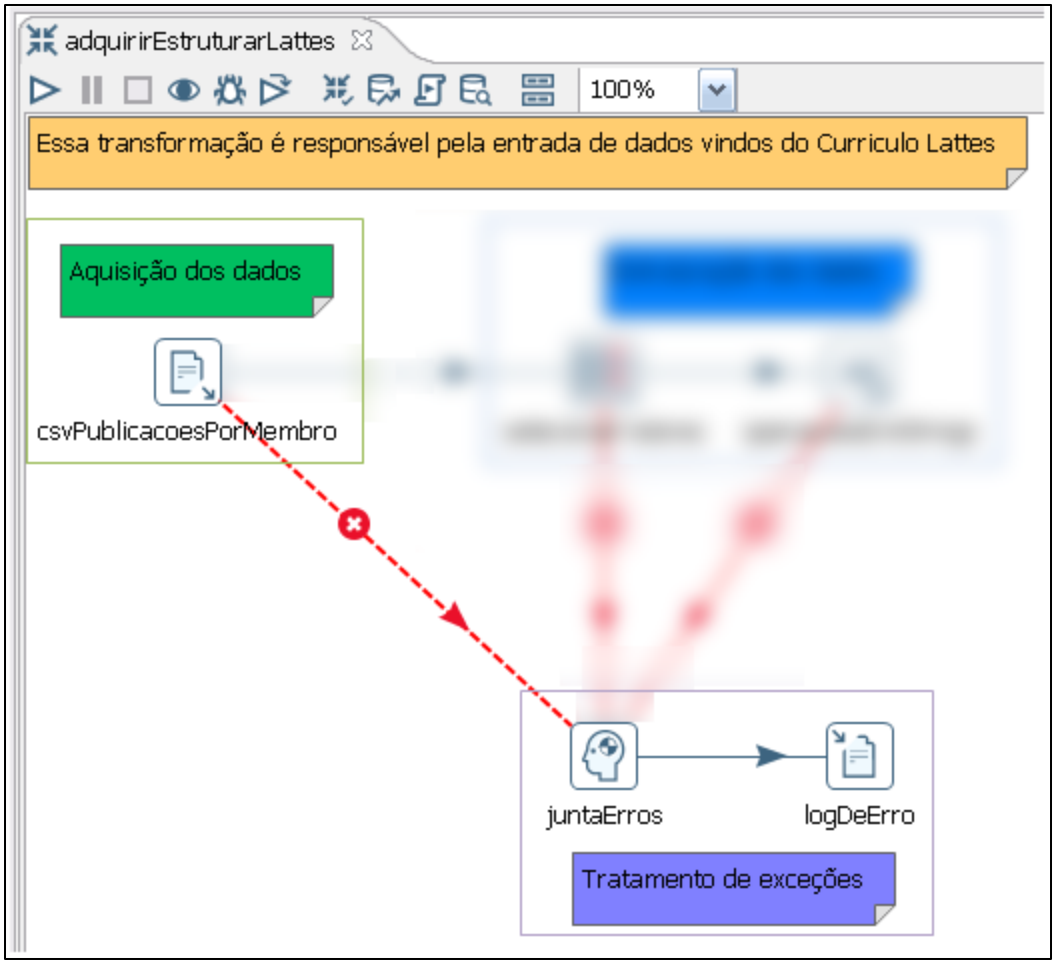


Figura 15 – Transformação PDI de Aquisição/Estruturação do scriptLattes com destaque no estágio de aquisição

No processo de aquisição dos dados do sistema acadêmico da UTFPR também foi utilizado o software PDI para extrair os dados do SGBD. Embora tenha-se o objetivo de utilizar os dados reais presentes no sistema acadêmico, por questões burocráticas, de privacidade e de segurança, a UTFPR não permitiu o acesso aos dados. Assim, nesse trabalho, o banco de dados foi simulado em uma estrutura utilizando o SGBD PostgreSQL. O Diagrama de Entidade e Relacionamento (DER) desta base de dados, pode ser visualizado no Apêndice A desse trabalho.

O processo de extração é mostrado na Figura 16, onde no estágio (SGBDSistemaAcademico) é apontado a localização do banco de dados e a tabela que deseja-se utilizar. Para cada perspectiva de tabela do SGBD (tabela principal + tabelas com junções) que deseja-se utilizar os dados é necessária uma transformação do PDI.

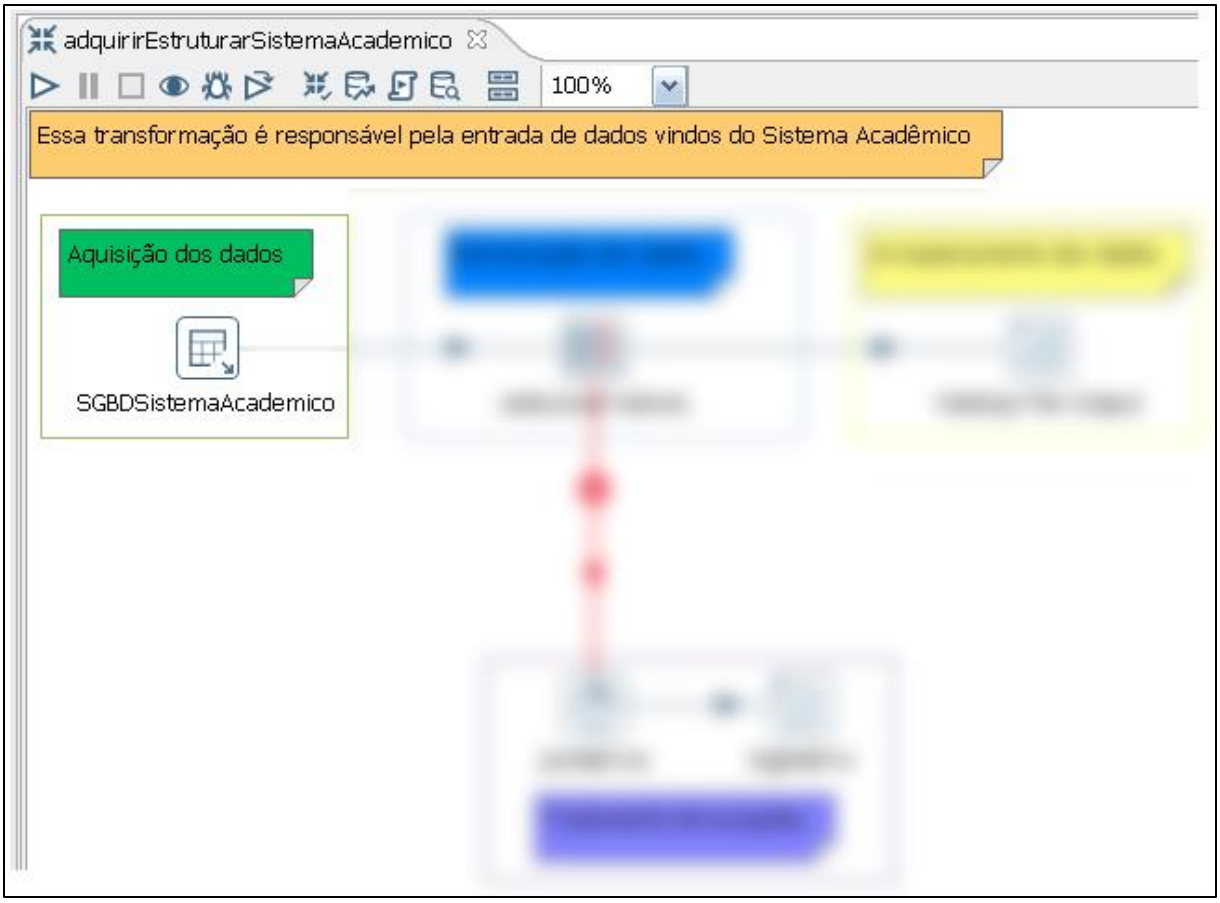


Figura 16 – Transformação PDI de Aquisição/Estruturação do Sistema Acadêmico com destaque no estágio de aquisição

Na Figura 17, o estágio (SGBDSistemaAcademico) é detalhado exibindo a consulta (query) SQL responsável por obter os dados referentes aos docentes e as aulas ministradas por eles, armazenadas no SGBD do sistema acadêmico. Essa consulta irá retornar os dados das aulas ministradas, tais como: o conteúdo ministrado e a data da aula, quanto aos docentes que ministram essas aulas, a consulta também irá trazer suas informações, tais como: nome do docente, o código único que o identifica na instituição (registro acadêmico ou RA), o CPF e sua data de nascimento.

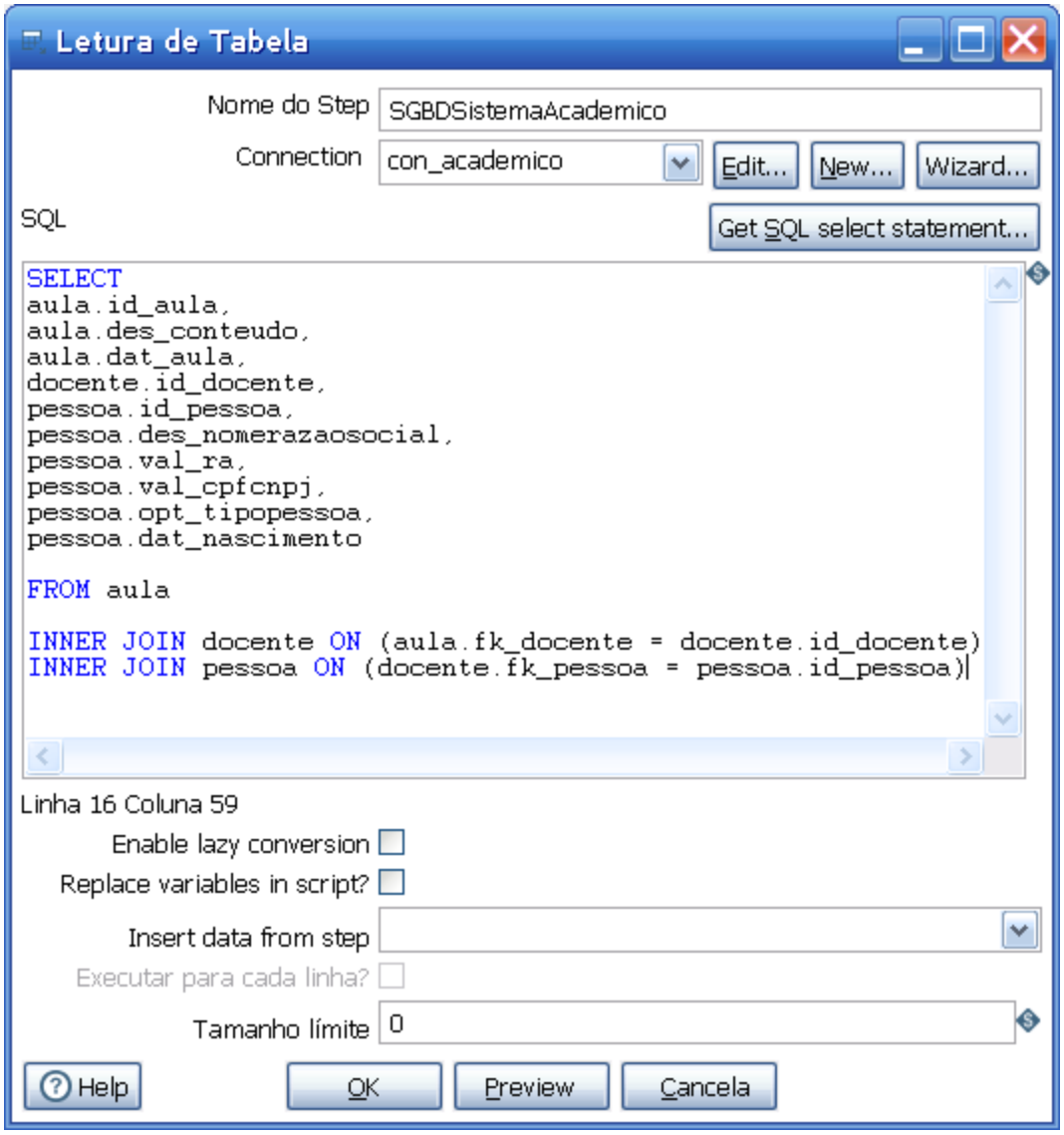


Figura 17 – Consulta dos dados do SGBD do sistema acadêmico

3.3.2 Etapa 2 - Estruturar

Uma vez coletados os dados, eles estarão em memória temporária (RAM) e o próximo passo é estruturá-los, preparando-os para armazená-los em disco e disponibilizá-los para as próximas fases. Esta fase pode ser considerada esta como a primeira fase de pré-processamento, pois a estruturação dos dados permitiu separar os campos do arquivo, definir um campo delimitador para os dados e ordená-los, de acordo com a necessidade. Nessa etapa também que foram definidos os tipos para cada campo de dados no novo arquivo. No final, o dado foi etiquetado e

consequentemente tornou-se mais útil para um programa manipulá-lo e representá-lo de alguma maneira.

Para essa etapa os seguintes resultados foram entregues:

- relação descritiva dos campos
- dados estruturados em um formato adequado

Nome do arquivo	Delimitador	Ordem e tipo dos campos
qualis_{codigododocente}.csv	\,	CLASSIFICATION {string}, QUANTITY {numeric}, PERIODICOS {string}
quantidadepublicacoes_{co-digododocente}.csv	\,	CLASSIFICATION {string}, QUANTITY {numeric}
quantidadeorientacoes_{co-digododocente}.csv	\,	CLASSIFICATION {string}, QUANTITY {numeric}
quantidadedeaulasmes_{codig ododocente}.csv	\,	CLASSIFICATION {string}, QUANTITY {numeric}
divisaotempo_{codigododo-cente}.csv	\,	CLASSIFICATION {string}, QUANTITY {numeric}

Quadro 2 - Relação descritiva dos dados resultantes da etapa de estruturação

Para essa fase podem ser utilizadas ferramentas de apoio como a ferramenta de ETL Pentaho Data Integration (PDI) e Hadoop. Todavia, optou-se pelo PDI, pois ele é mais simples, e visualmente, mais fácil de realizar manutenções. Foi analisada a estrutura dos dados para que os mesmos atendam ao modelo proposto, definido no protocolo para inserção de dados.

Na Figura 18 é apresentada a EAP do processo na etapa de estruturação.

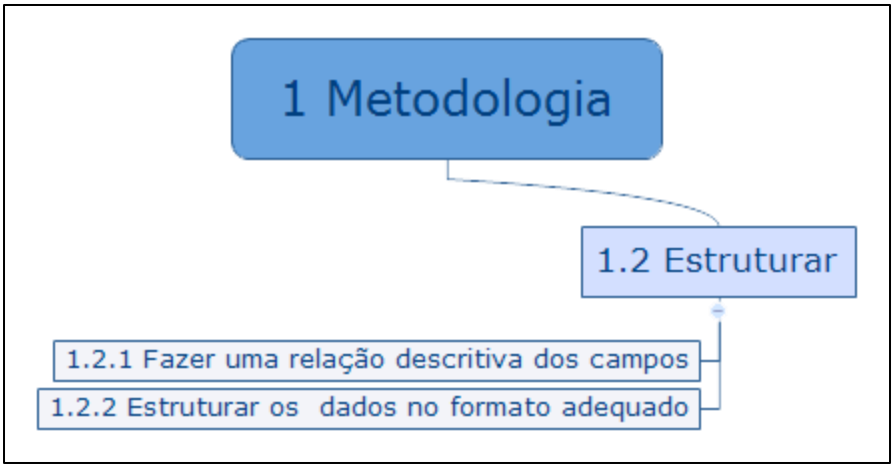


Figura 18 – EAP da etapa de estruturação

A Figura 19 apresenta o processo de estruturação dentro da estrutura do Kanban.

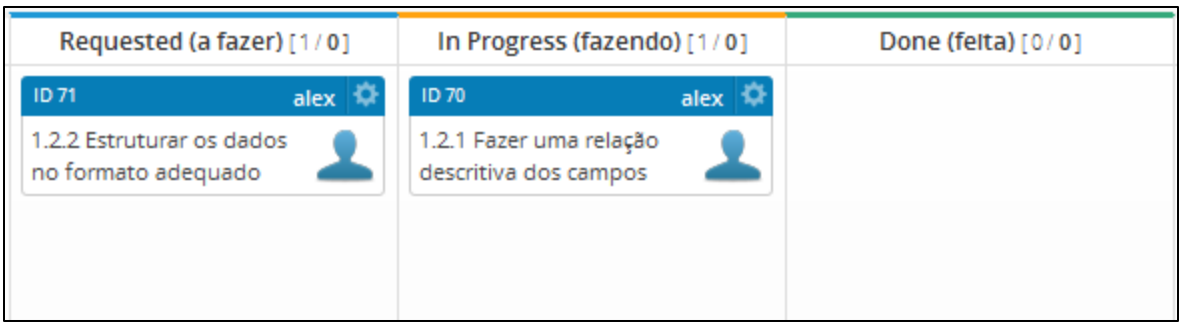


Figura 19 – Quadro de Kanban da etapa de estruturação

A Figura 20 ilustra o processo de estruturação com a ferramenta PDI. No primeiro estágio (selecionarValores) os dados vindos da fase anterior (Aquisição) são preparados, portanto, é informado o tipo e se ele possui algum prefixo.

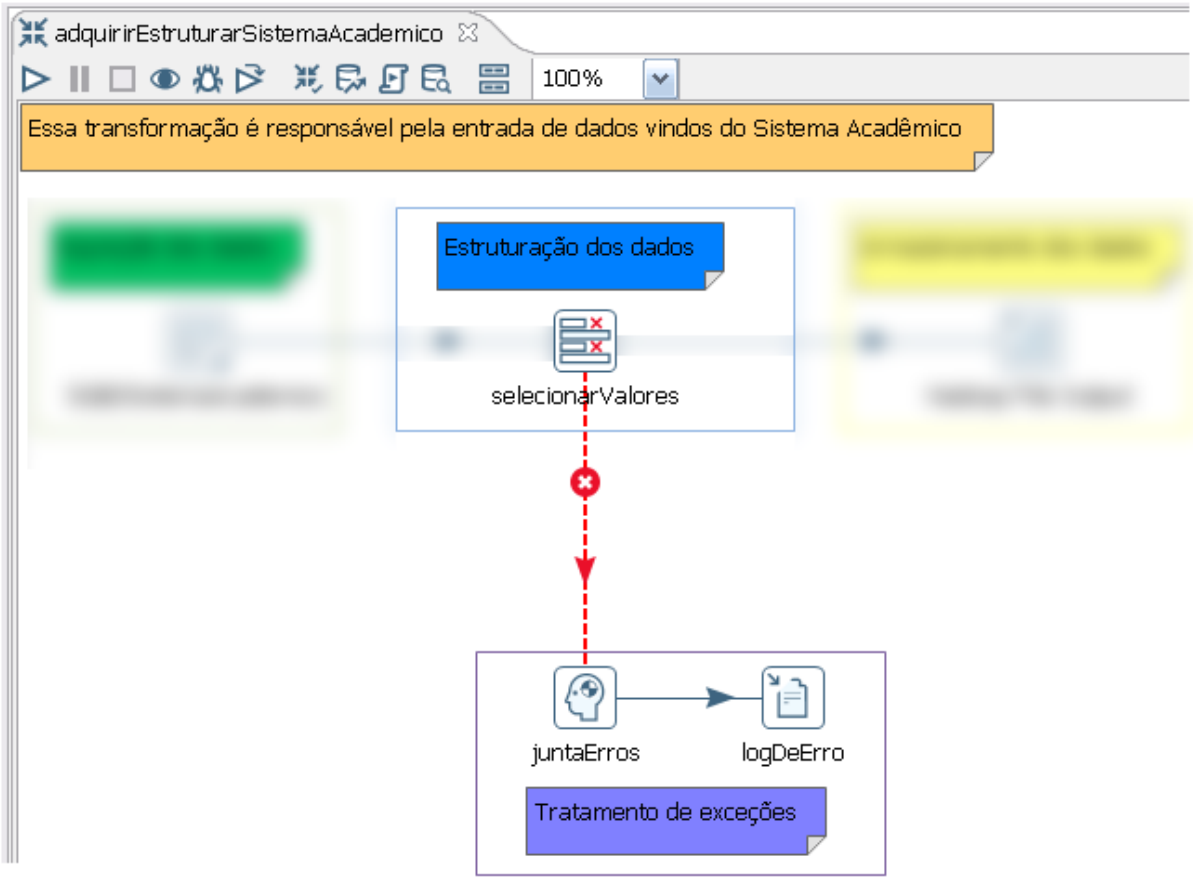


Figura 20 – Transformação PDI de Aquisição/Estruturação com destaque no estágio de estruturação

3.3.3 Etapa 3 - Armazenar

A etapa de armazenamento é uma das mais importantes do processo, principalmente quando fala-se de Big Data, pois desde o início, antes de qualquer processamento ou pré-processamento os dados precisam ser adequadamente armazenados.

Essa etapa também é responsável por arquivar os dados preparados, em uma estrutura apropriada para que eles possam facilmente ser consultados e utilizados posteriormente. Essa etapa é necessária para que as informações que servirão de base para as próximas fases estejam disponíveis em um tempo de resposta curto. O armazenamento também é necessário para que os processos anteriores não sejam reexecutados desnecessariamente.

O armazenamento temporário dos dados também deverá ser estruturado, pois para que o processamento ocorra, os dados têm que ser lidos constantemente e, e no contexto de Big Data, como o volume de dados é grande, é interessante que os dados que estejam em uma estrutura adequada (formatada e/ou tabulada) para que o processamento não sofra atrasos (gargalos) no tempo de leitura desses dados, favorecendo para que a aplicação tenha um bom desempenho.

A Figura 21 apresenta a EAP na etapa de armazenamento, e a Figura 22 apresenta o quadro de Kanban da mesma etapa.

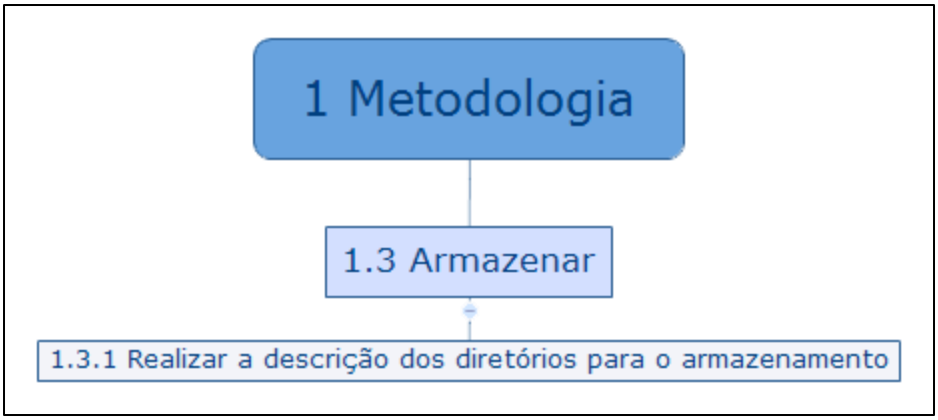


Figura 21 – EAP da etapa de armazenamento

Figura 22 – Quadro de Kanban da etapa de armazenamento

Embora o processo tenha sido descrito de maneira linear, essa etapa é uma exceção, pois no final de cada uma das etapas o armazenamento de dados deverá ser necessário.

Para essa etapa os seguintes resultados deverão ser entregues:

- Descrição da organização dos diretórios;
- Arquivo de dados com a persistência.

Nome do arquivo	Tipo	Localização no HDFS
qualis_{codigododocente}	csv	/dados/lattes/estruturados
quantidadepublicacoes_{codigododocente}	csv	/dados/lattes/estruturados
quantidadeorientacoes_{codigododocente}	csv	/dados/sistemaacademico/estruturados
quantidadedeaulasmes_{codigododocente}	csv	/dados/sistemaacademico/estruturados
divisaotemp_{codigododocente}	csv	/dados/sistemaacademico/estruturados

Quadro 3 - Descrição da organização dos diretórios no HDFS

Depois de coletados e estruturados, os dados foram enviados para o HDFS diretamente pelo PDI, utilizando o estágio (saidaHDFS), conforme ilustrado na Figura 23. Os dados foram arquivados em uma pasta específica do projeto, todavia o PDI não cria a estrutura de diretórios, sendo que essa foi feita manualmente, utilizando o

terminal do Linux e executando nele o comando *mkdir* do HDFS, conforme demonstra a Figura 24.

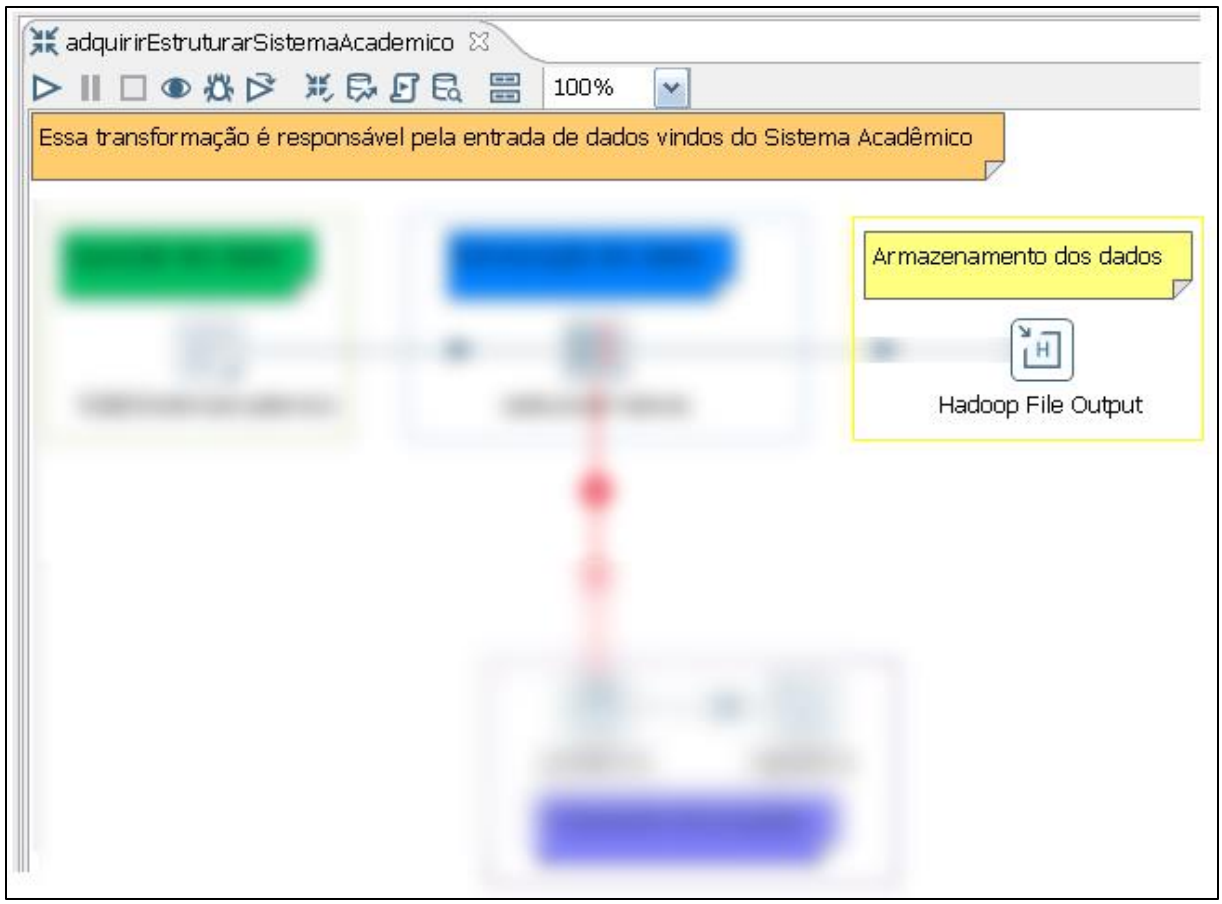


Figura 23 – Transformação PDI de Aquisição/Estruturação com destaque no estágio de armazenamento

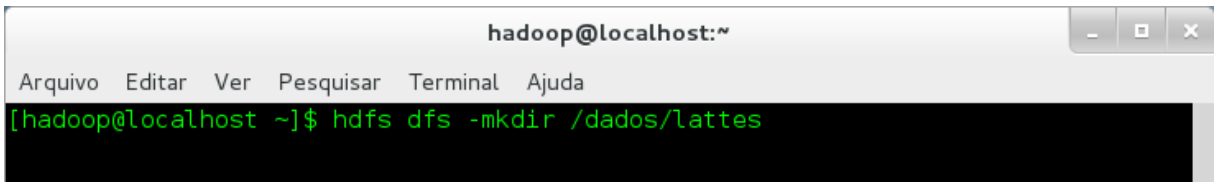
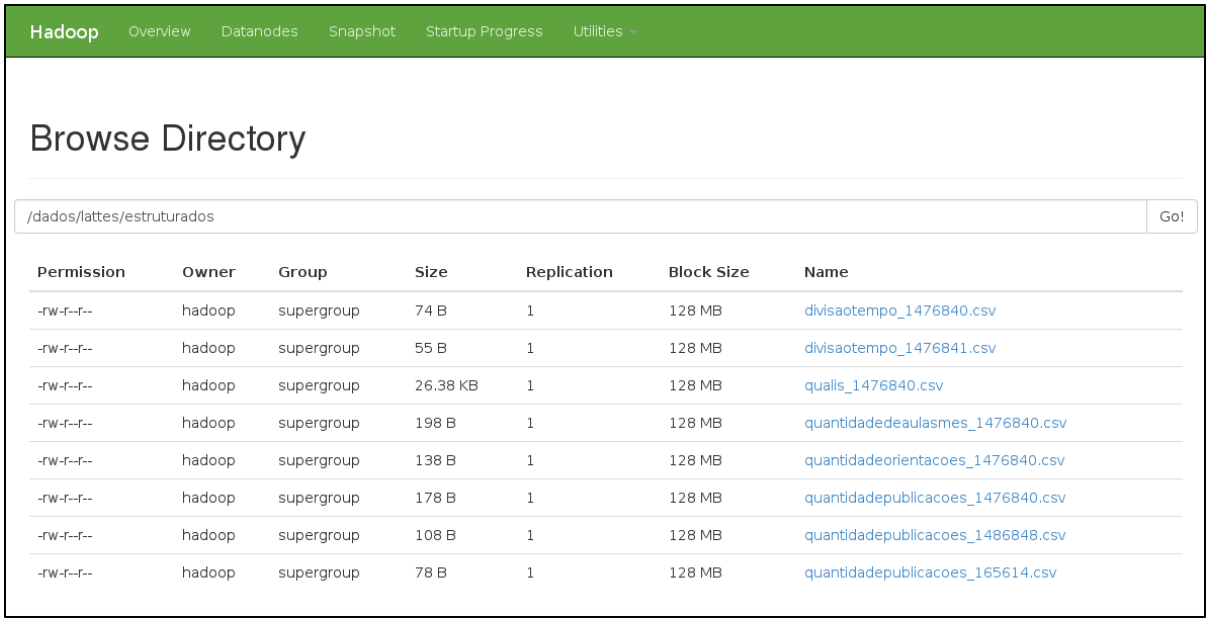


Figura 24 – Comando utilizado para criação de diretórios no HDFS

A Figura 25 expõe a interface Web com os dados do Currículo Lates estruturados e armazenados no HDFS.



Permission	Owner	Group	Size	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	74 B	1	128 MB	divisaotempo_1476840.csv
-rw-r--r--	hadoop	supergroup	55 B	1	128 MB	divisaotempo_1476841.csv
-rw-r--r--	hadoop	supergroup	26.38 KB	1	128 MB	qualis_1476840.csv
-rw-r--r--	hadoop	supergroup	198 B	1	128 MB	quantidadedeaulasmes_1476840.csv
-rw-r--r--	hadoop	supergroup	138 B	1	128 MB	quantidadeorientacoes_1476840.csv
-rw-r--r--	hadoop	supergroup	178 B	1	128 MB	quantidadepublicacoes_1476840.csv
-rw-r--r--	hadoop	supergroup	108 B	1	128 MB	quantidadepublicacoes_1486848.csv
-rw-r--r--	hadoop	supergroup	78 B	1	128 MB	quantidadepublicacoes_165614.csv

Figura 25 – Página Web do HDFS, com os dados estruturados do scriptLattes

A Figura 26 demonstra os mesmos dados armazenados no HDFS, porém, listados diretamente pelo terminal do Linux utilizando o comando "-ls + espaço + diretório desejado":

 # hdfs dfs -ls /dados/lattes/estruturados

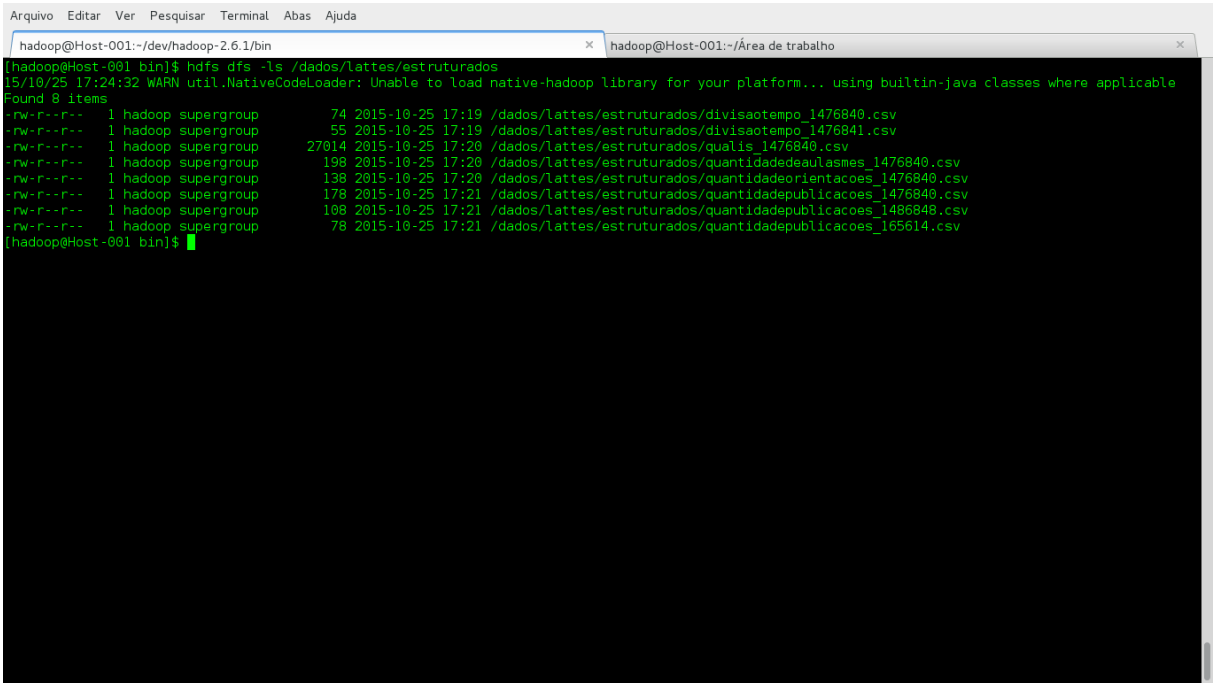


Figura 26 – Terminal do Linux exibindo os dados armazenados no HDFS

3.3.4 Etapa 4 - Filtrar

Durante a etapa de aquisição os dados foram coletados sem passar por nenhum processo de inspeção. Por esse motivo é provável que informações desnecessárias tenham sido coletadas juntamente com os dados úteis para a análise, assim, esta etapa envolve filtrar os dados para remover partes não relevantes ao uso. Além disso, mesmo os campos que serão utilizados na etapa seguinte podem necessitar um novo pré-processamento, como a remoção de espaços, remoção de caracteres especiais, conversão para minúsculo/maiúsculo e padronização/normalização de valores.

A Figura 27 expõe a EAP da etapa de filtragem dos dados provenientes das etapas anteriores, a Figura 28 demonstra o quadro de Kanban da mesma etapa.

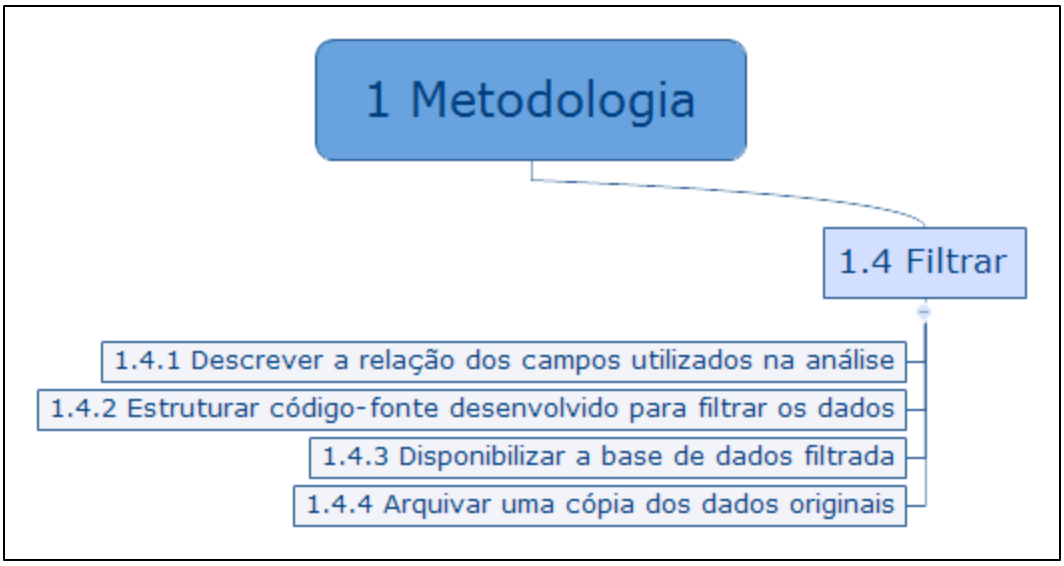


Figura 27 – EAP da etapa de Filtragem dos dados

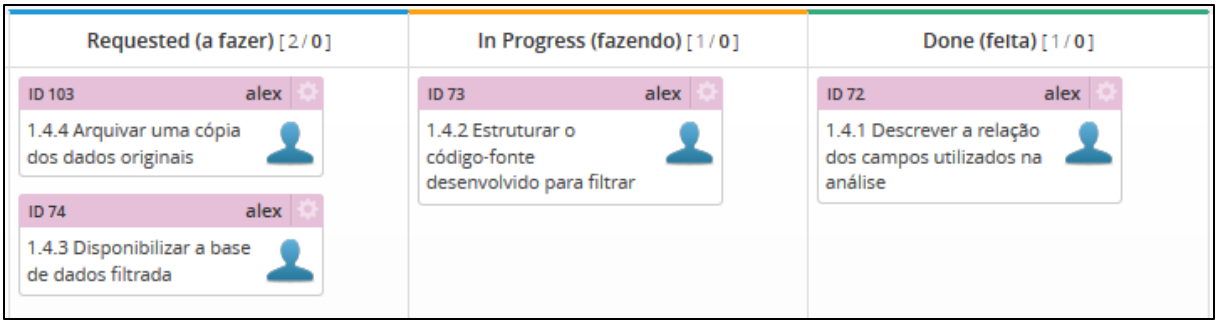


Figura 28 – Quadro de Kanban da etapa de filtragem dos dados

Para essa etapa os seguintes resultados foram entregues:

- Relação dos campos utilizados na análise
- Código-fonte desenvolvido para filtrar os dados
- Base de dados filtrada
- Cópia dos dados originais

Nome do arquivo	Campos utilizados
qualis_{codigododocente}	Classificação do congresso/periódico, quantidade de publicações, nome do congresso/periódico
quantidadepublicacoes_{codigododocente}	Tipo da publicação, quantidade de publicações do tipo
quantidadeorientacoes_{codigododocente}	Mês de referência, quantidade de orientações
quantidadedeaulasmes_{codigododocente}	Mês de referência, quantidade de aulas
divisaotempo_{codigododocente}	Tipo (prática ou teórica), quantidade(percentual)

Quadro 4 - Relação dos campos utilizados na análise

A Figura 29 demonstra o processo de filtragem dos dados utilizando o PDI No primeiro estágio (entradaHDFS) os dados são obtidos diretamente de uma estrutura HDFS. No segundo estágio (operacoesEmStrings) são executadas operações para remover espaços vazios no começo e/ou no final das *Strings*, remover caracteres especiais e transformar em maiúscula ou minúscula. No terceiro estágio (ordenarDados) os dados são então ordenados de acordo com uma regra definida. No quarto estágio (saidaHDFS) os dados filtrados são persistidos no HDFS, invocando a etapa de armazenamento.

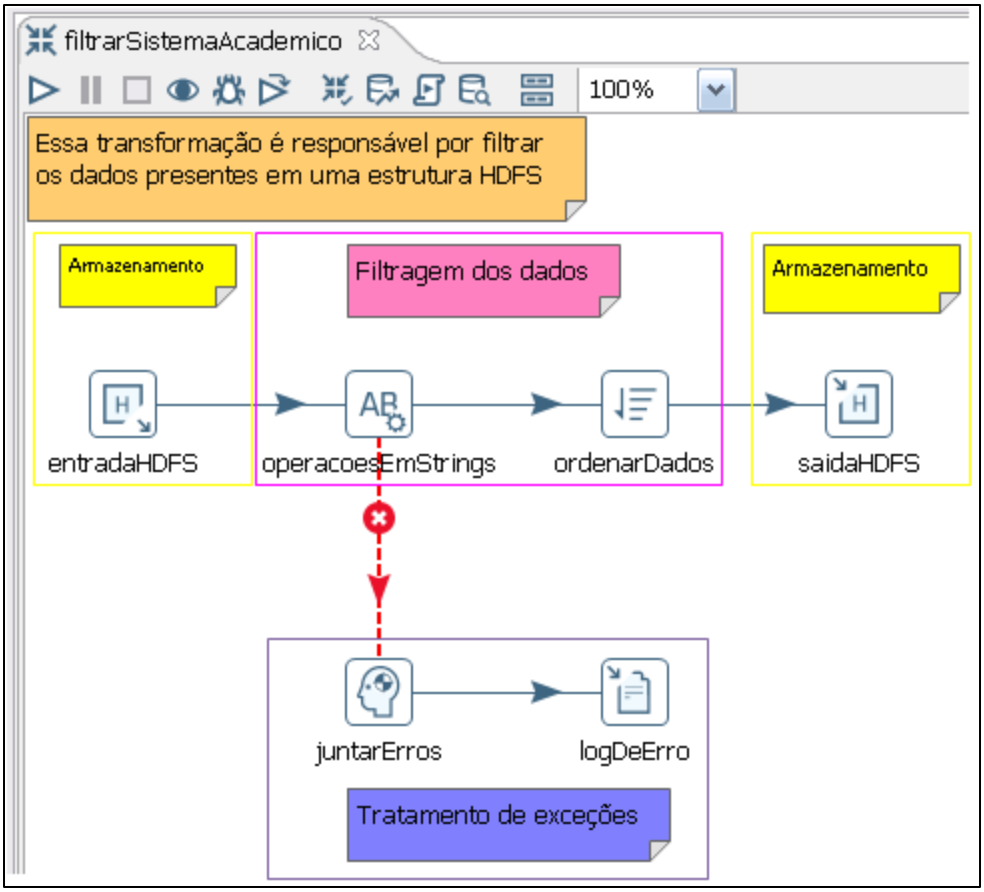


Figura 29 – Processo de filtragem dos dados no PDI

A Figura 30 demonstra o processo de filtragem dos dados utilizando o PDI. Essa transformação tem como objetivo padronizar os dados de publicações em eventos/periódicos dos docentes. No primeiro estágio da filtragem (isNotResumoExpandidoEmCongresso) verifica se o tipo da publicação não é um Artigo Expandido em Congresso. Caso não seja, a informação é apenas salva no HDFS (saidaHDFS) da mesma maneira. Caso seja, o estágio (limpaPeriodico) apaga todas as informações de periódicos que não são relevantes para as análises, o estágio (setPeriodicoNulo) deixa o valor na coluna onde continham as informações sobre os periódicos nulos, o estágio (seNulo) verifica na coluna dos periódicos, caso seja nulo, ele apaga toda a coluna, deixando os dados no padrão das demais publicações que não possuem essa coluna.

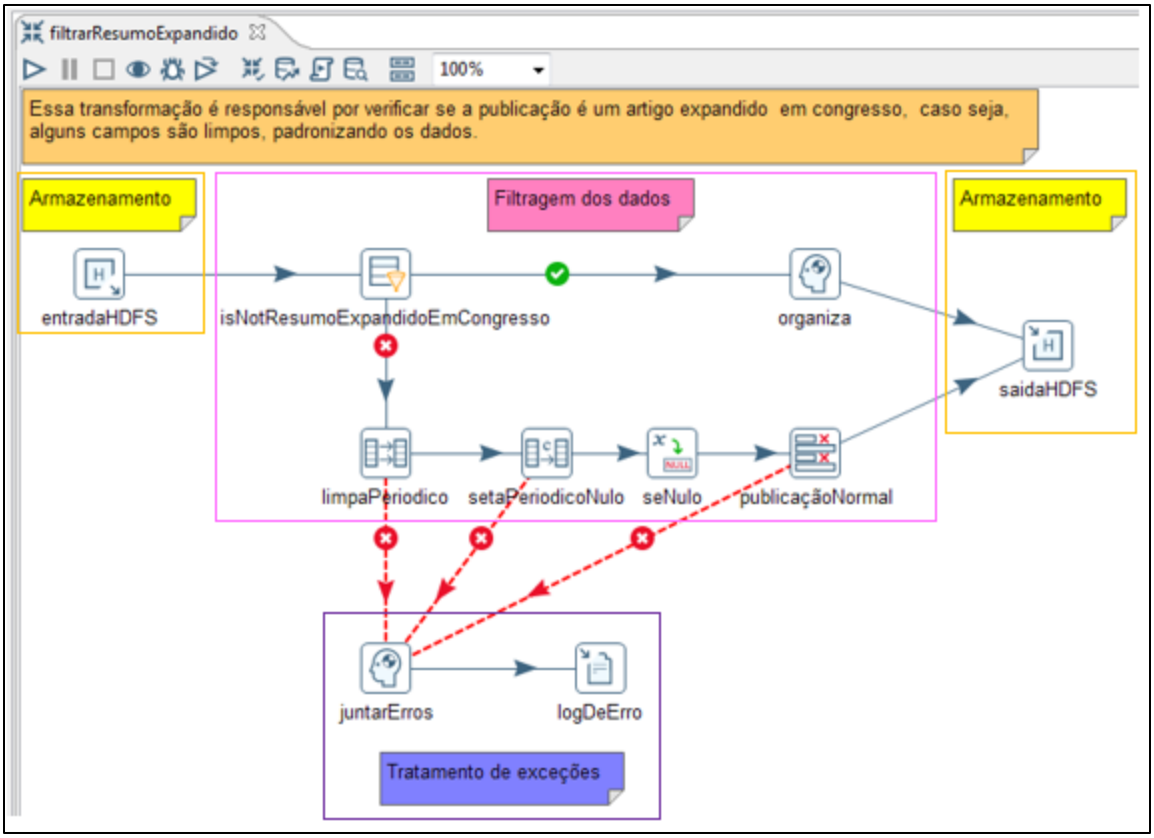


Figura 30 – Processo de filtragem dos dados de resumos expandidos em congressos no PDI

3.3.5 Etapa 5 - Minerar

A etapa de mineração (*Data Mining*) é a responsável por gerar os resultados que serão utilizados como base de entrada na visualização dos dados. Para a execução dessa etapa é primordial que se saiba previamente qual técnica de mineração de dados e/ou estatística é a mais adequada para gerar os resultados desejados. Uma vez respondida essa pergunta, torna-se necessário implementar uma solução que utilize os dados da etapa anterior como base de entrada, aplique uma técnica específica sobre essa base, e gere os resultados em um novo arquivo de saída.

Aplicar mineração de dados e/ou métodos estatísticos para obter alguns padrões sobre os dados pode, muitas vezes, não ser uma tarefa muito simples, todavia, é essencial para ofertar uma boa base de dados para ser visualizada.

A Figura 31 apresenta a EAP da etapa de mineração dos dados e a Figura 32 o quadro de Kanban dessa etapa.

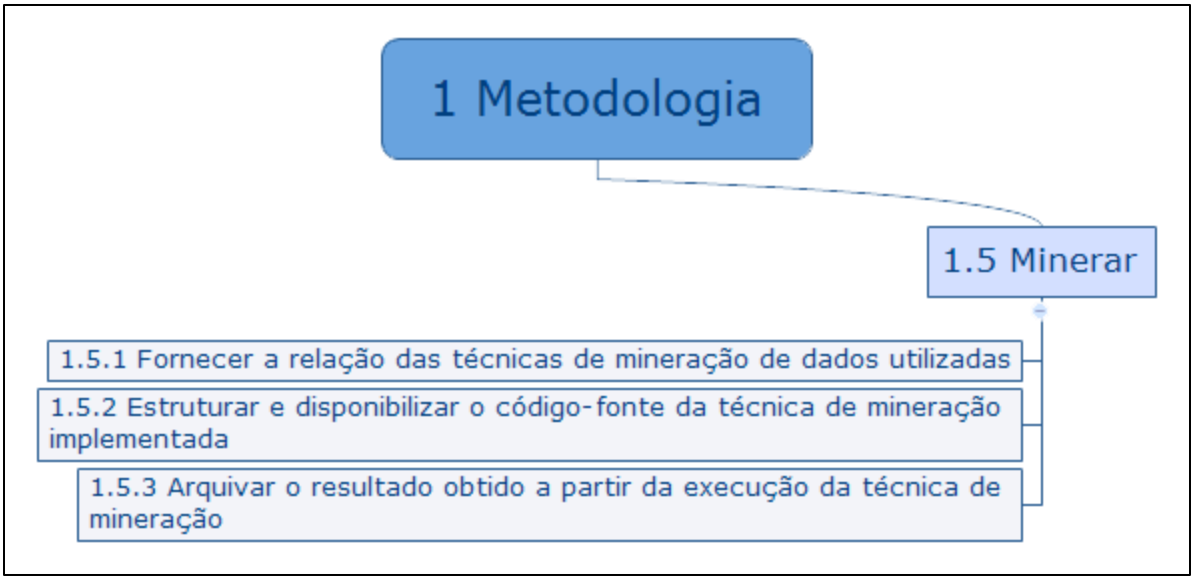


Figura 31 – EAP da etapa de mineração

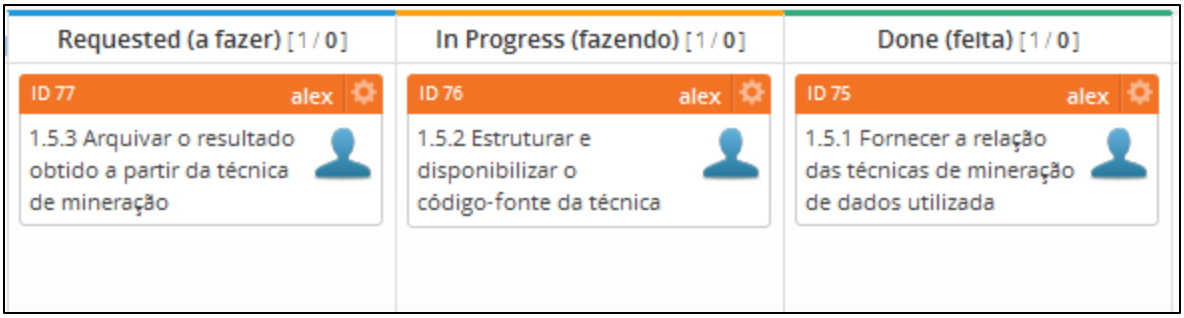


Figura 32 – Quadro de Kanban da etapa de mineração

Para essa etapa os seguintes resultados foram entregues:

- Relação das técnicas de mineração de dados utilizadas
- Código-fonte da técnica de mineração implementada
- Resultado obtido a partir da execução da técnica de mineração

Técnica utilizada	Objetivo
Contagem	Contar qual o total de publicação de cada tipo

Quadro 5 – Relação das técnicas de mineração de dados utilizadas

Para essa fase foram utilizadas ferramentas de apoio como o arcabouço Hadoop, por questões de desempenho e escalabilidade. Também foi executada a técnica de mineração de dados no PDI para fins de comparação.

Para obter-se a quantidade de publicações de cada tipo, a estratégia utilizada foi contar as informações textuais que diz qual tipo de publicação aquele docente possui, portanto, se um docente possui 5 artigos em congresso, 5 vezes irá aparecer no arquivo de dados dele o termo "*artigoEmCongresso*", dessa forma, utilizar um contador de palavras naquela coluna de dados que identificam o tipo da publicação é eficiente e foi a técnica utilizada.

Para realizar o processo de mineração utilizando o MapReduce, um algoritmo para contar uma quantidade de palavras específicas em um texto foi utilizado. Esse algoritmo é muito comum de ser encontrado em exemplos de utilização do MapReduce. O comando na Figura 33, compila o código-fonte contendo esse algoritmo. Como resultado da compilação é gerado o arquivo *bytecode* com a extensão "*.class*".

```
[hadoop@Host-001 bin]$ hadoop com.sun.tools.javac.Main WordCount.java
```

Figura 33 – Geração de um *bytecode* Java

Os arquivos *bytecode* Java gerados pelo processo anterior são empacotados e compactados em um Arquivo Java (Java *ARchive* ou JAR), conforme mostra a Figura 34.

```
[hadoop@Host-001 bin]$ jar cf wc.jar WordCount*.class
```

Figura 34 – Empacotamento dos *bytecodes*

Após o algoritmo ser preparado, ele é executado sobre os dados presentes no HDFS, nos quais deseja-se obter a quantidade de palavras. As figuras 35A, 35B e 35C exibem o algoritmo de contar palavras sendo executado no Hadoop.

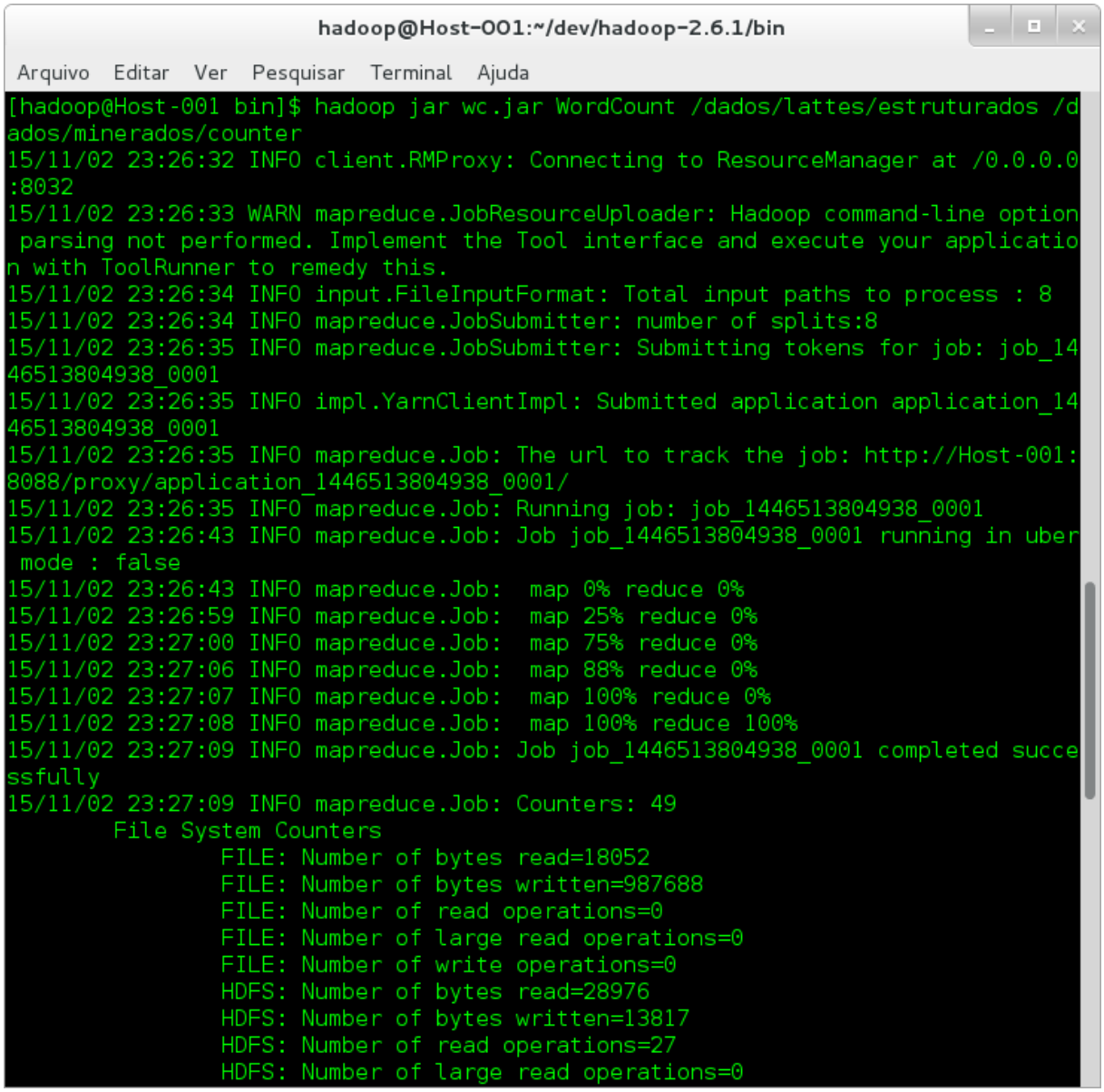


Figura 35A – Execução do algoritmo para contagem de palavras no MapReduce, parte 1

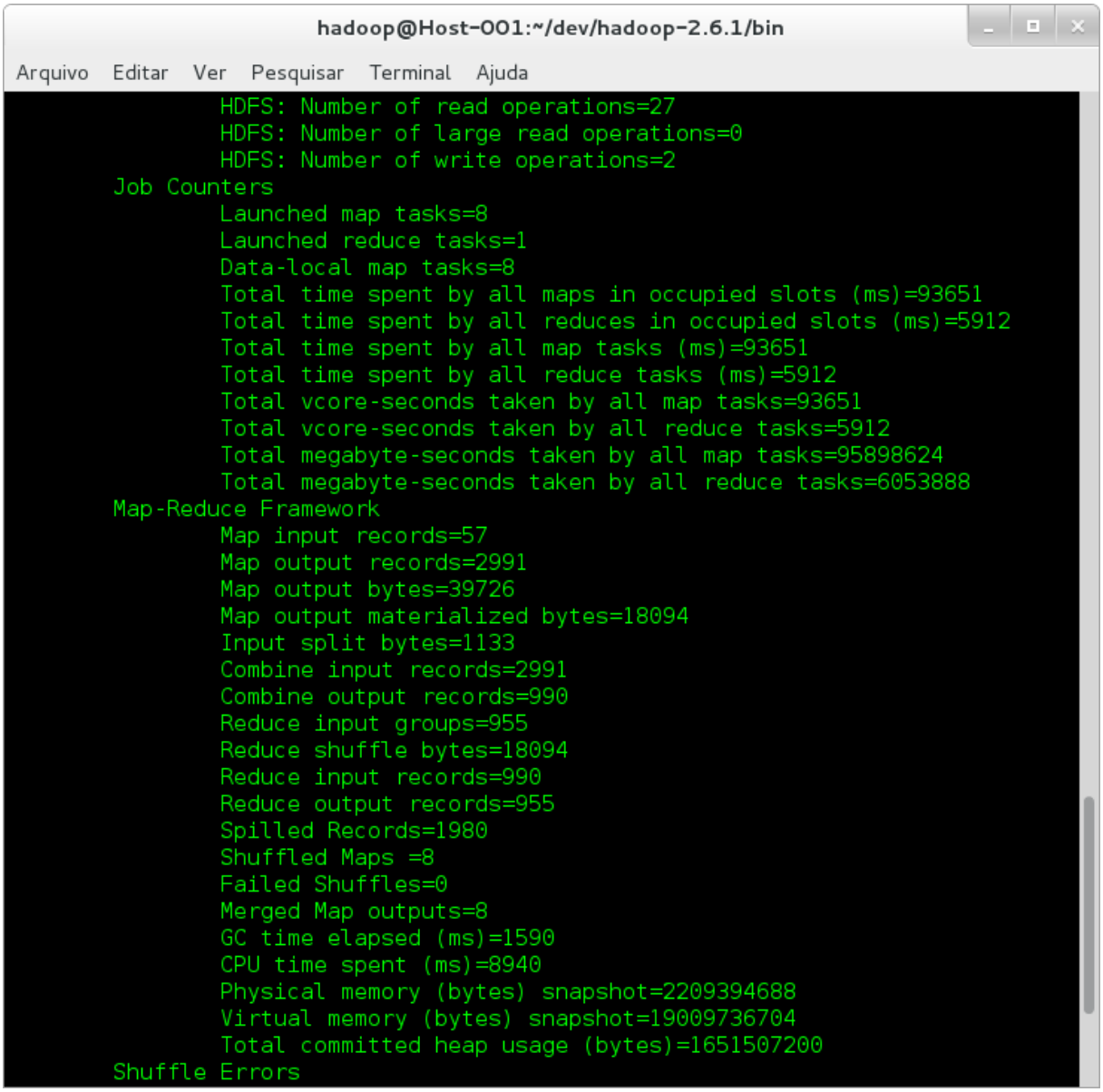


Figura 35B – Execução do algoritmo para contagem de palavras no MapReduce, parte 2

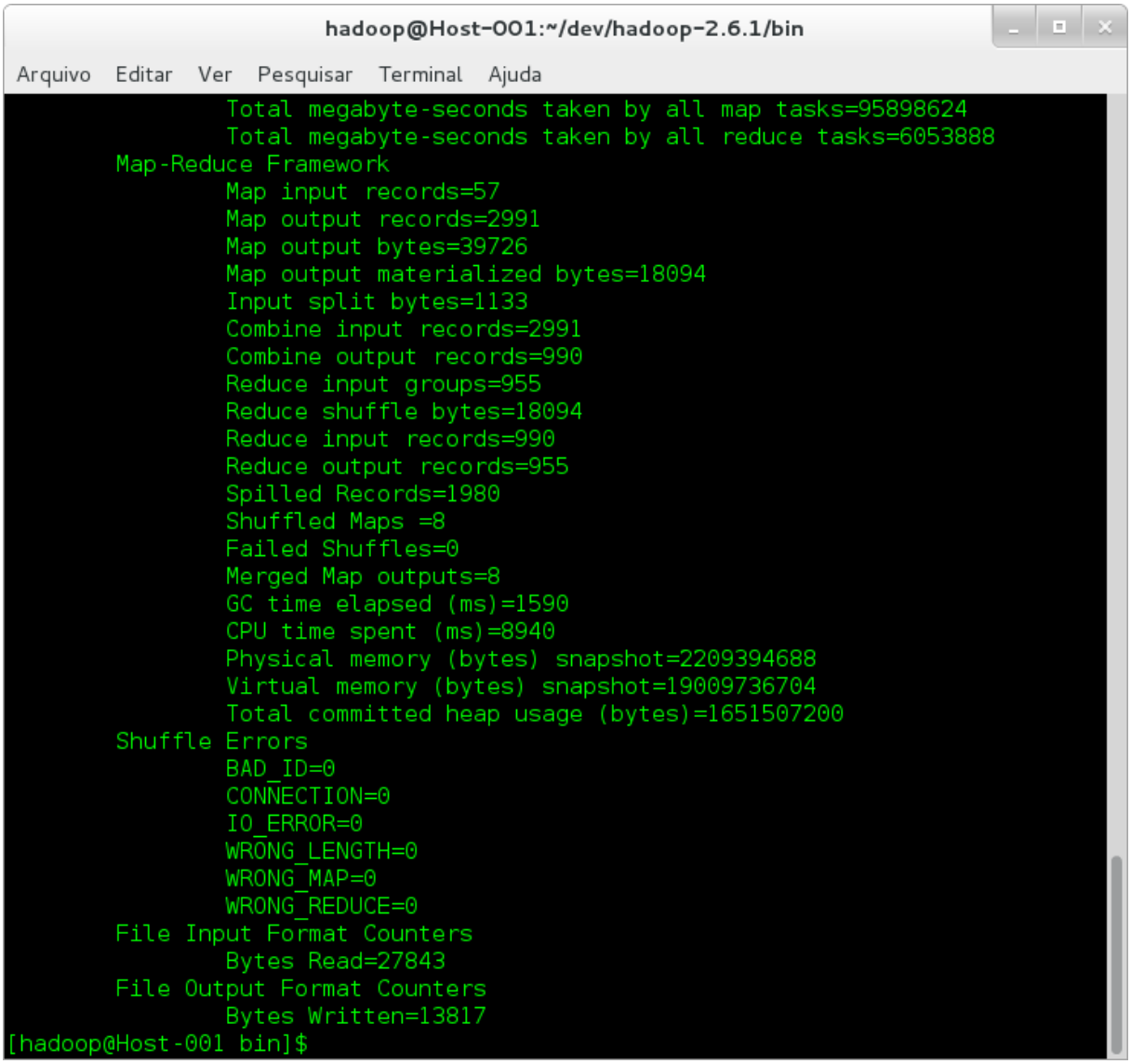


Figura 35C – Execução do algoritmo para contagem de palavras no MapReduce, parte 3

A execução do algoritmo pode ser visualizada na tela de gerenciamento do YARN, conforme mostra a Figura 36.

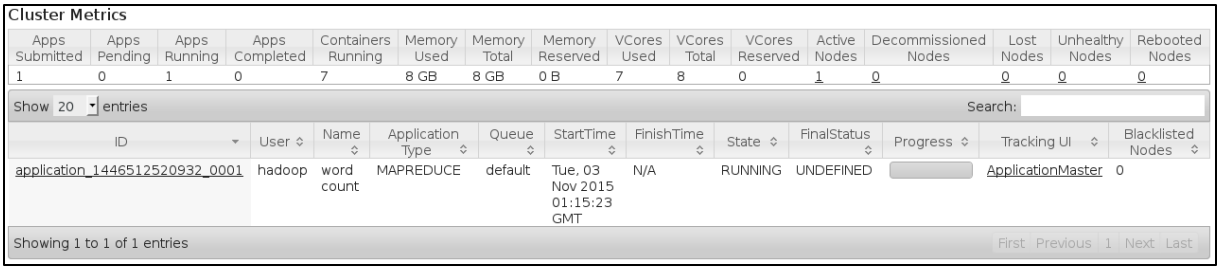


Figura 36 – Gerenciamento de execução das aplicações no YARN

O processo de mineração por contagem também é realizado no PDI para fins de comparação. A Figura 37 exibe o processo realizado, sendo que no primeiro estágio da mineração de dados (obtemVariavel) um valor do tipo *String* é passado para a transformação através do *job* que a chama. Nesse valor deverá haver o nome do docente desejado para que o próximo estágio verifique se o autor da linha lida naquele momento é o mesmo armazenado na variável, caso não seja o dado é encaminhado para o estágio (nothing) e ignorado. Caso seja, ele passa para o próximo estágio (ordenaLinhas) que irá ordenar os dados de modo ascendente de acordo com os dados da coluna (tipo) que informa o tipo da publicação (artigo em periódico, trabalho completo em congresso, dentre outros). Após ordenados, os dados são enviados para o estágio (agrupaPorTipoEConta). Esse estágio é a chave para o resultado do processo de mineração, pois ele agrupa os dados de acordo com a coluna (tipo) e realiza um somatório da quantidade de registros de cada tipo. Após a contagem os dados são enviados para o estágio (selecionaValores) que deixa somente os 2 campos necessários: type, com o tipo da publicação e quantity, com a quantidade de publicações daquele tipo. Esses dados são então enviados para o estágio (saidaHDFS) que persiste esses resultados na estrutura HDFS.

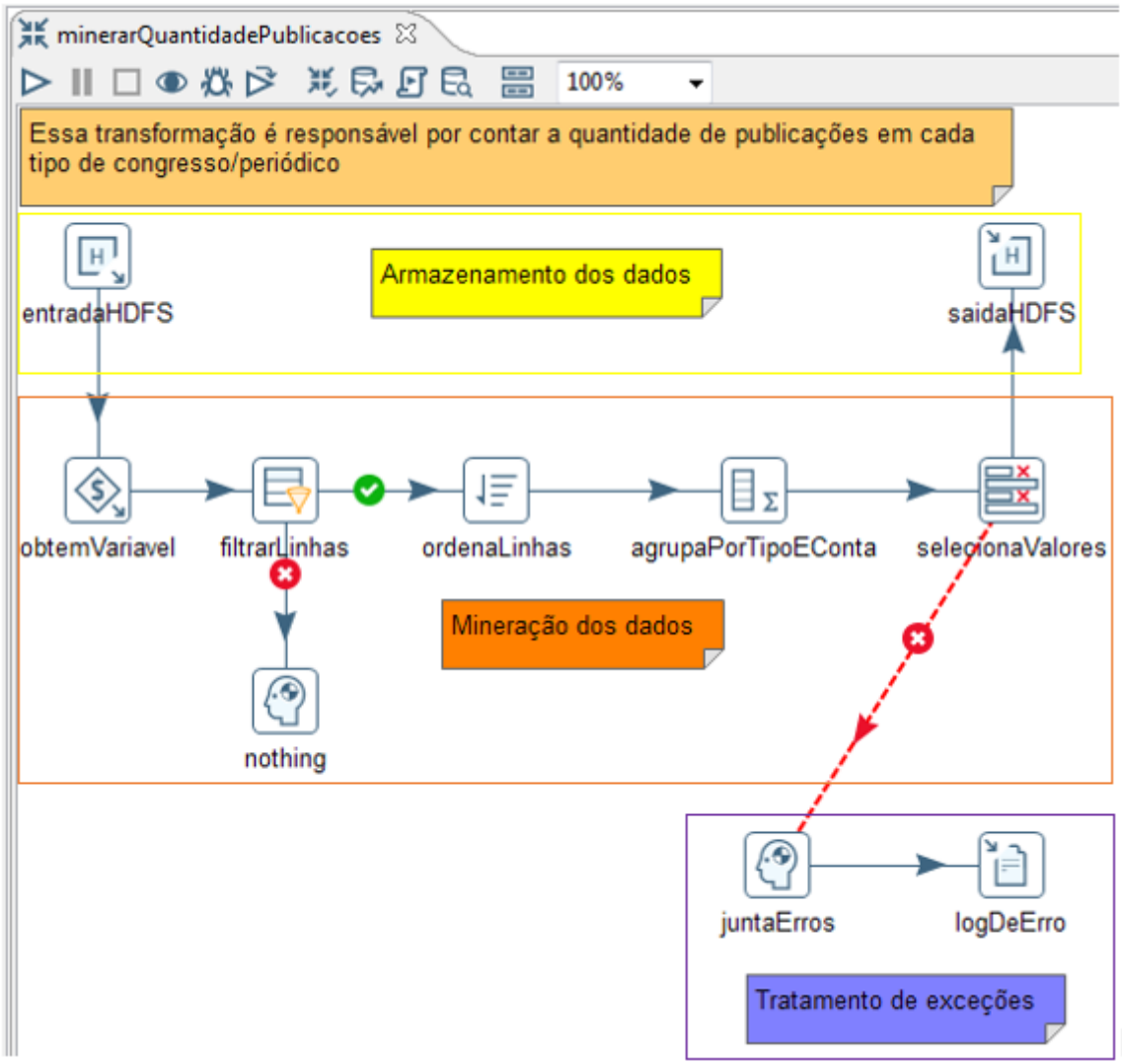


Figura 37 – Mineração de dados no PDI, contagem de publicações

3.3.6 Etapa 6 - Apresentar

As cinco primeiras etapas da metodologia proposta tiveram como objetivo a preparação dos dados que deverão ser visualizados graficamente nesta fase. Somente nessa etapa (6) é que os primeiros gráficos serão gerados. O objetivo principal dessa etapa é gerar uma primeira representação gráfica básica dos dados, que permita verificar se a escolha do tipo do gráfico, bem como dos dados em si foram adequadas. Dependendo do resultado, esta etapa pode fazer repensar as etapas anteriores.

Nessa etapa, os gráficos foram desenvolvidos sobre a linguagem de programação Javascript utilizando a biblioteca D3.js para que os indicadores fossem apresentados de maneira simples. A Figura 38 mostra a EAP da etapa de apresentação e a Figura 39 exibe o quadro de Kanban da etapa.

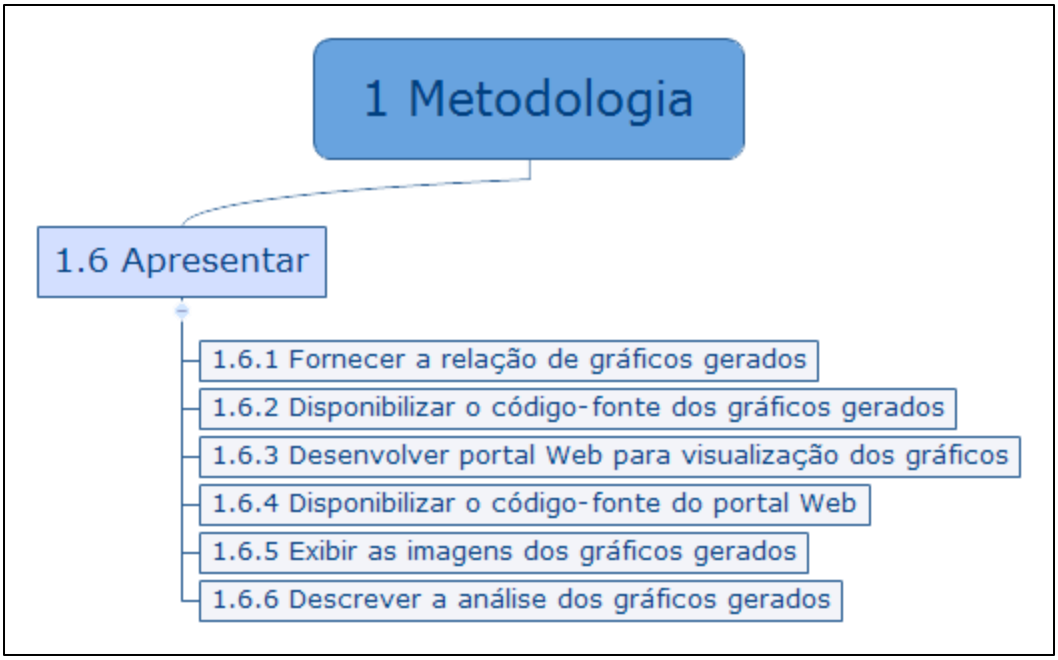


Figura 38 – EAP da etapa de apresentação

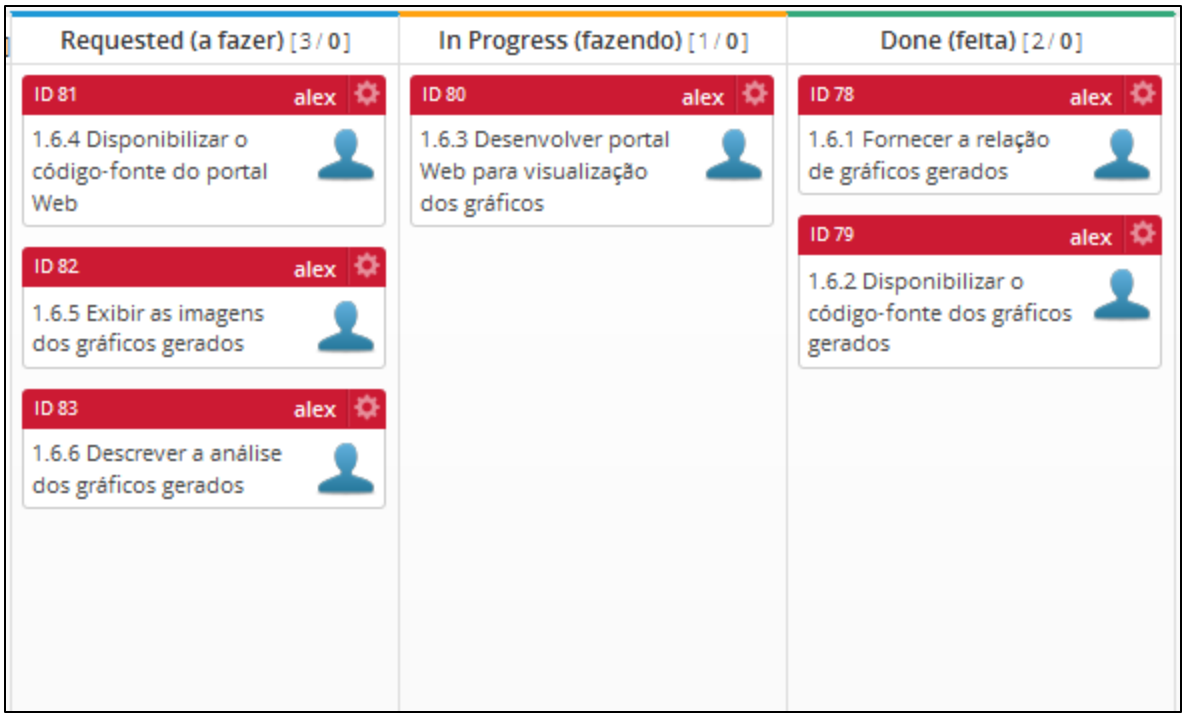


Figura 39 – Kanban da etapa de apresentação

Para essa etapa os seguintes resultados foram entregues:

- relação de gráficos gerados
- código-fonte dos gráficos gerados
- código-fonte do portal Web
- portal Web para visualização dos gráficos
- imagens dos gráficos gerados
- análise dos gráficos gerados

Gráfico	Objetivo
Setores	Exibir a divisão do tempo entre as atividades práticas e teóricas dos docentes.
Setores e barras	Exibir a quantidade de aulas de um grupo de docentes e comparar entre si.
Linha	Exibir a quantidade de aulas de cada docente em cada mês.
Área	Exibir a quantidade de projetos de orientações em trabalhos de diplomação e/ou estágios.
Rosca (*donut*)	Exibir a quantidade de publicações agrupando por Qualis.
Barras	Exibir a quantidade de publicações agrupando por tipo de evento.

Quadro 6 – Relação dos gráficos gerados

O Gráfico 1 é um gráfico do tipo setores com as informações referentes a quantidade de aulas práticas/teóricas ministradas pelo docente, nele é possível fazer a comparação entre as duas modalidades de aula sendo que sempre a modalidade com a maior quantidade terá uma maior área no gráfico.

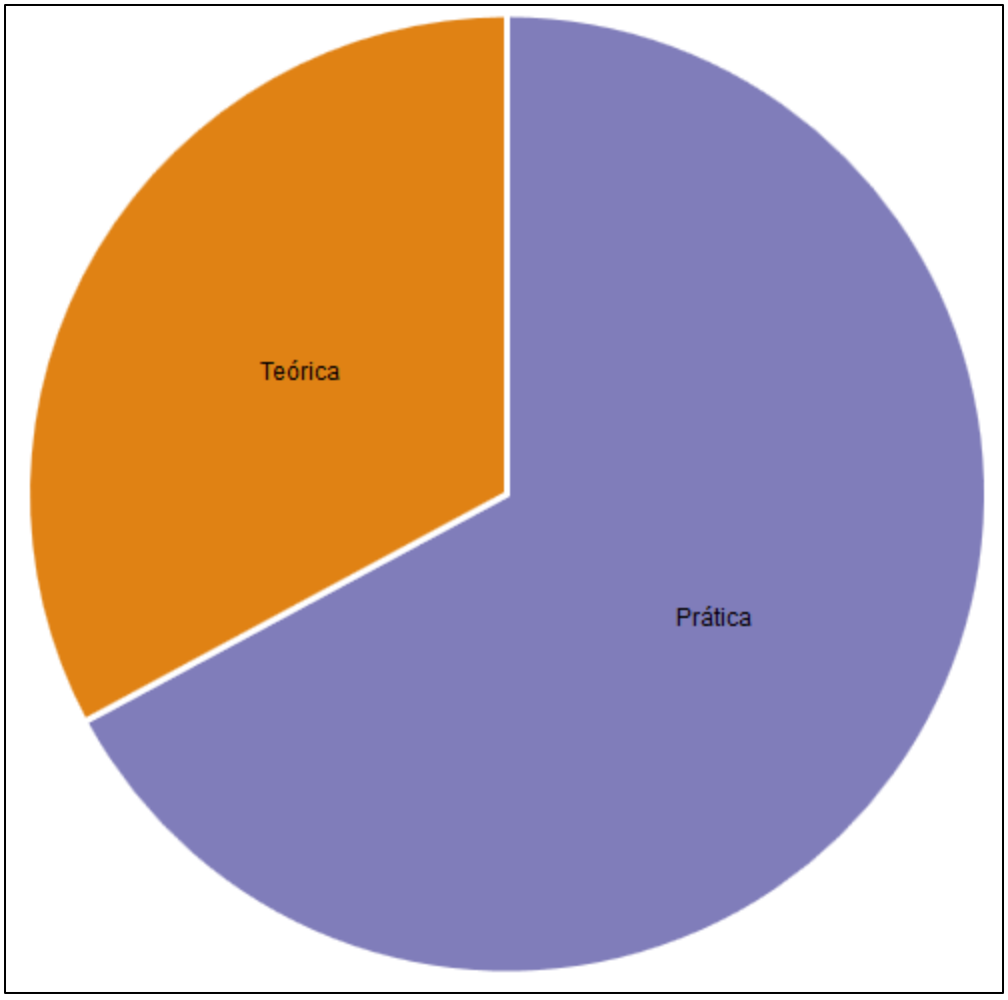


Gráfico 1 - Quantidade de aulas práticas e teóricas do docente

O Gráfico 2 é um gráfico de linhas com as informações referentes a quantidade de aulas ministradas em relação a cada mês de um ano. No eixo X (horizontal) do gráfico são informados os meses do ano, e no eixo Y (vertical) são exibidos os números com as quantidades de aulas ministradas naquela linha. Os picos do gráfico indicam uma maior quantidade de aulas. No Gráfico 2, pode-se observar que o mês de Julho teve uma baixa quantidade de aulas, pois nesse mês os discentes estavam em férias, dessa forma a quantidade de aulas foram 0. Já nos meses de Junho e Dezembro a quantidade de aulas ministradas por esse docente foi elevada, pois nesses meses o gráfico obteve picos.

A linha exibida nesse gráfico representa a ligação entre os distintos pontos, com ela pode-se observar que há uma grande queda do mês de Junho (pico) para o mês de Julho (vale), pode-se também observar que do mês de Setembro a Novembro, o crescimento da quantidade de aulas foi bastante grande, o que não foi tão expressivo de Novembro a Dezembro.

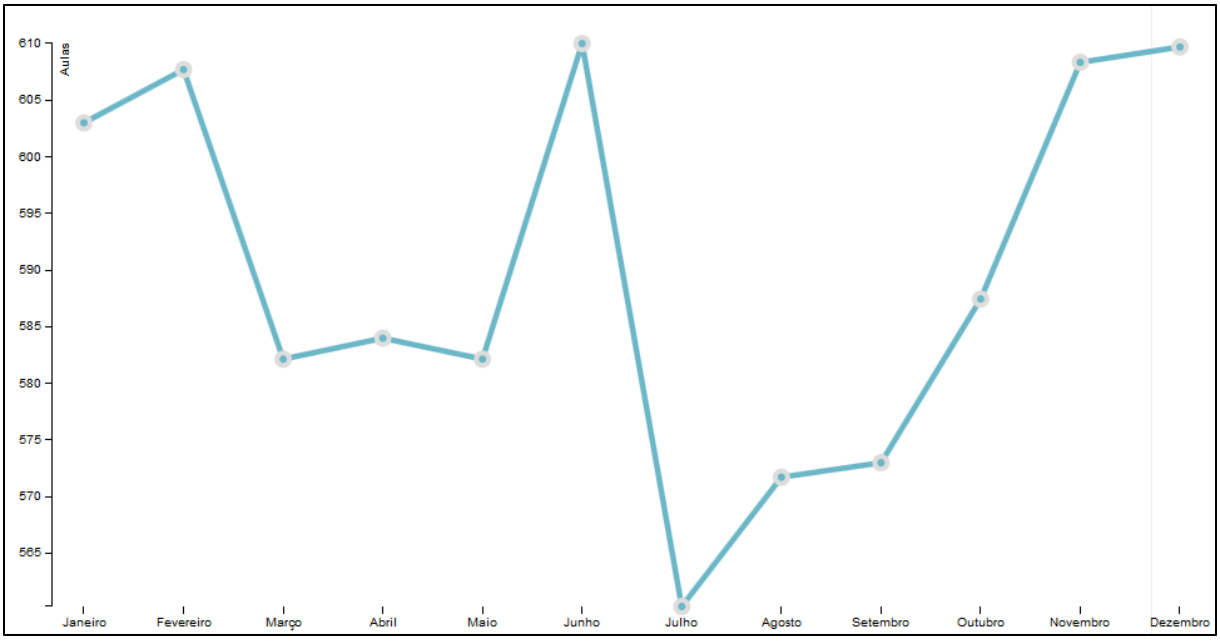


Gráfico 2 - Quantidade de aulas ministradas em cada mês (gráfico de linhas)

Para fins de comparação o Gráfico 3 exibe as mesmas informações do Gráfico 2, todavia, o formato de gráfico utilizado é o gráfico de barras, com isso é possível verificar que no gráfico de linhas a tendência pode ser visualizada mais facilmente.

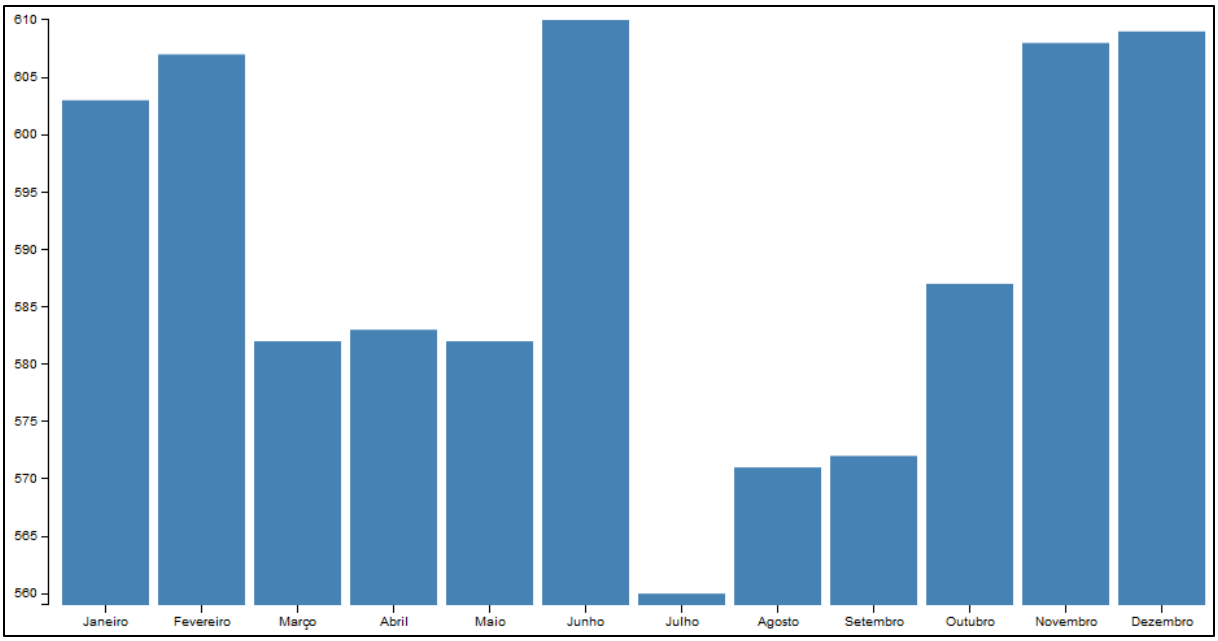


Gráfico 3 - Quantidade de aulas ministradas em cada mês (gráfico de barras)

O Gráfico 4 é um gráfico de área com as informações referentes a quantidade de orientações em trabalhos de conclusões de cursos e em orientações para estágios curriculares obrigatórios que o docente está atuando. No eixo X (horizontal) do gráfico são informados os meses do ano, e no eixo Y (vertical) são exibidas as quantidades de orientações naquela linha. Os picos do gráfico indicam uma maior quantidade de orientações, quanto mais clara for o preenchimento da área do gráfico, maior será a quantidade de orientações daquele docente, os vales do gráfico e as cores mais escuras representam uma menor quantidade de orientações daquele docente, como é o exemplo do mês de Julho, onde há a troca de semestre e o docente não possui nenhum discente sobre orientação.

Os círculos presentes no gráfico representam o ponto em que cada valor está. No mês de Junho há um pico no gráfico que indica que foi o mês que o docente mais teve projetos em orientação, 10 no total. A linha exibida no gráfico exibe a ligação entre cada um dos pontos, com ela pode-se facilmente observar que do mês de Janeiro até o mês de Fevereiro, o crescimento foi expressivo, pois até o mês de Janeiro o docente só tinha 1 projeto em orientação dos anos anteriores, no mês de Fevereiro onde as aulas daquele ano letivo iniciaram, o docente teve um aumento de 9 projetos em orientações.

Entre os meses de Junho e Julho foi grande a queda na quantidade de projetos de orientações ativos, pois nesse período 5 alunos de estágio curricular obrigatório defenderam seus projetos e foram aprovados. No mês de Agosto além dos 5 alunos que o docente já tinha em orientação foram acrescentados mais 3 alunos, no mês de Outubro os 3 alunos de trabalho de conclusão de curso defenderam seus projetos de trabalho (pré-banca final) e foram aprovados, já no mês de Novembro os 5 alunos restantes defenderam seus trabalhos finais e foram aprovados, deixando o docente com nenhum projeto em orientação no mês de dezembro.

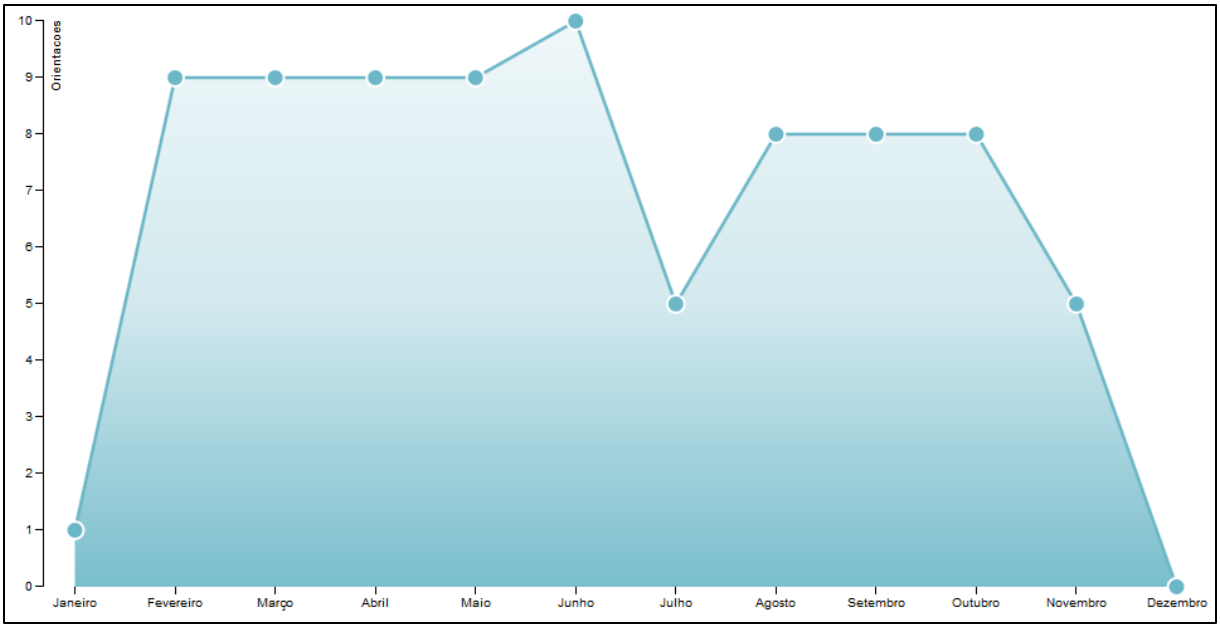


Gráfico 4 - Quantidade de trabalhos de graduação e estágios orientados em cada mês

O Gráfico 5 é um gráfico de rosca (*donut*) com as informações referentes a qual a classificação (Qualis) do evento/periódico que contém as publicações das quais o docente possui. O gráfico é dividido em pedaços, chamados de fatias. Cada fatia possui uma cor diferente e equivale a uma classificação do Qualis. O tamanho dessa fatia está relacionado a quantidade de publicações daquela categoria, sendo que, quanto maior o tamanho da fatia, maior a quantidade de publicações naquela categoria. No Gráfico 5 pode-se observar que o docente possui uma grande quantidade de publicações em periódicos com o Qualis não identificado, sendo que essas publicações são responsáveis por quase a metade de todas as suas publicações.

Deve-se observar que para realizar a comparação entre um gráfico de rosca com as publicações de um docente e outro gráfico do mesmo formato com dados sobre outro docente, os tamanhos de cada fatia podem ser diferentes, pois o tamanho da fatia é definido pela série de dados individual de cada gráfico, portanto, um mesmo valor em diferentes gráficos pode ter tamanhos de fatias diferentes de acordo com os dados no qual o gráfico está inserido.

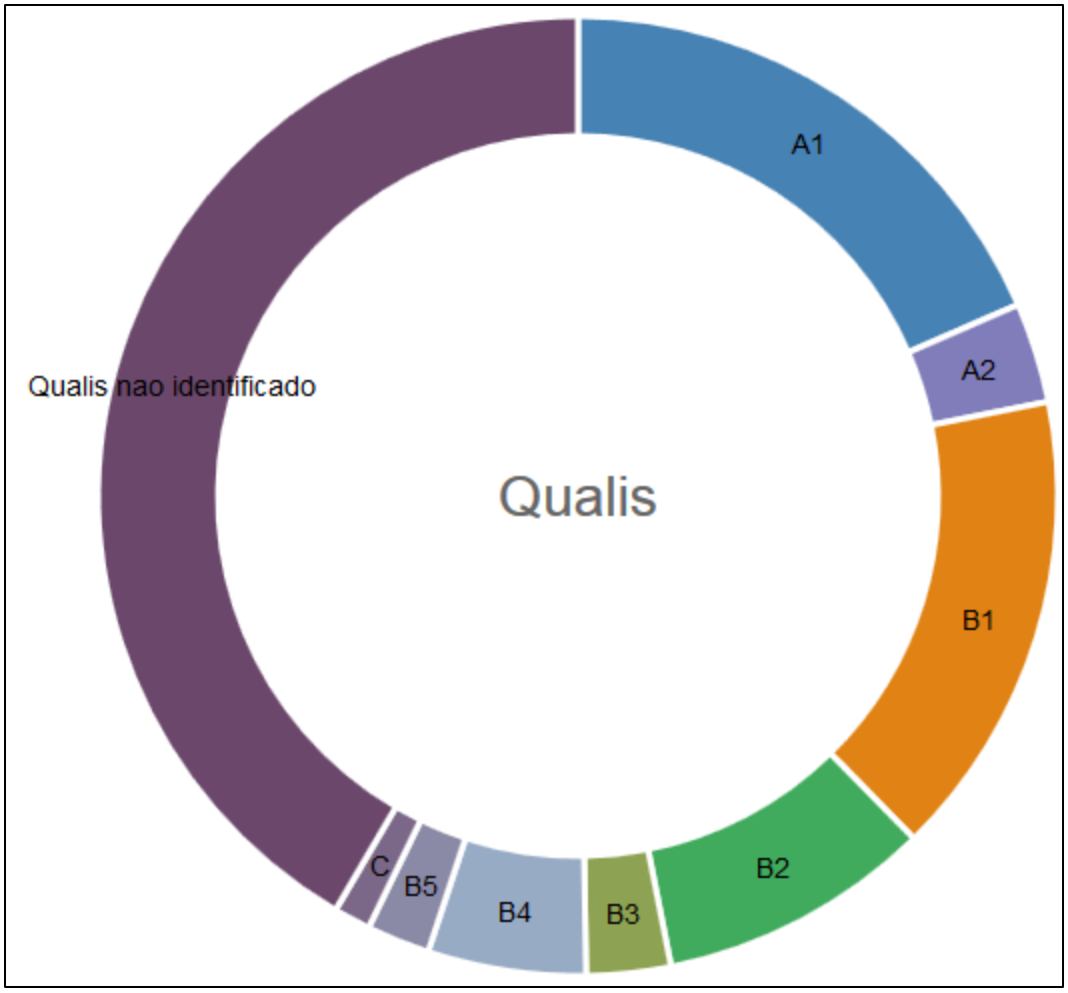


Gráfico 5 - Quantidade de publicações classificadas por Qualis

O Gráfico 6 é um gráfico de barras com informações referentes a quantidade de trabalhos externos à IES de acordo com cada tipo, tais como: trabalho em congresso, artigo em periódico e etc. No eixo X (horizontal) do gráfico são informados os tipos dos trabalhos, no eixo Y (vertical) são exibidas as quantidades das publicações naquela linha. Os picos do gráfico indicam uma maior quantidade, já os vales indicam uma menor quantidade de trabalho. No Gráfico 6 pode-se observar que o docente selecionado, possui uma maior quantidade de trabalho (11) do tipo trabalhos completos em congresso, e uma menor quantidade de trabalho (1) do tipo apresentação em congresso.

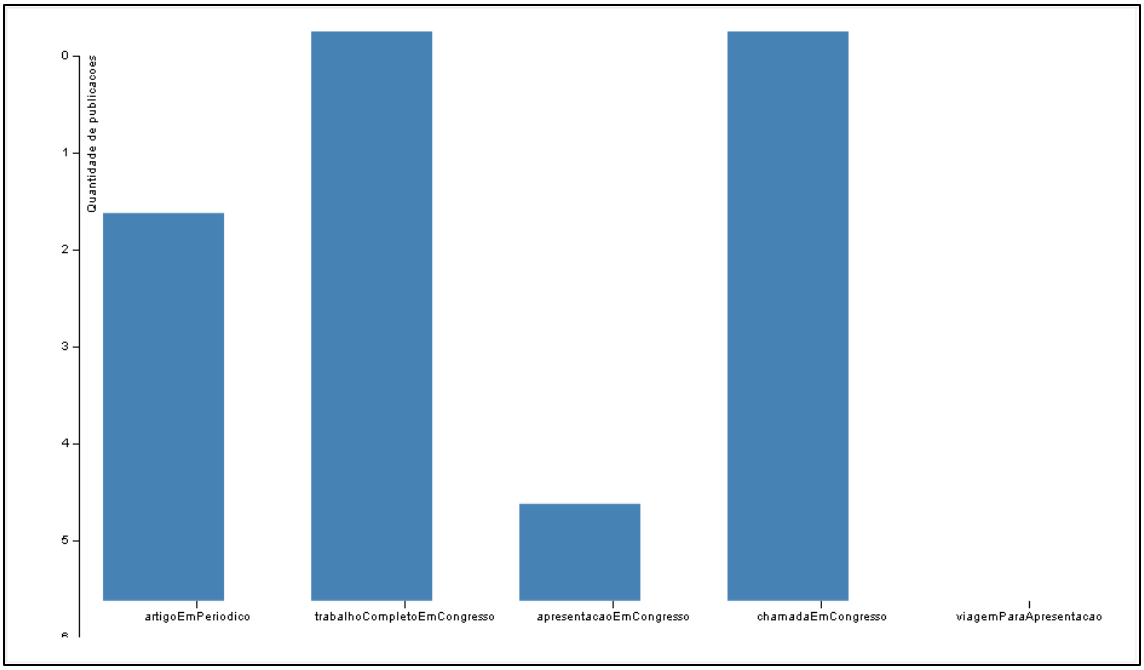


Gráfico 6 - Quantidade de publicações classificadas por tipos

O Gráfico 7 é composto por dois gráficos diferentes sobrepostos em uma mesma tela. O primeiro é do tipo barras e exibe a quantidade de aulas de um docente no período de 1 ano, o eixo Y dele está localizado no canto esquerdo da imagem. Esse gráfico possui as mesmas características e os mesmos dados do Gráfico 3.

O segundo gráfico que sobrepõe ao primeiro, é do tipo linha e exibe a quantidade de orientações executadas por um docente, embora o tipo de gráfico seja diferente, ele possui as mesmas características do Gráfico 4. Seu eixo Y fica localizado no lado direito da imagem.

O Gráfico 7 torna fácil o processo de realizar comparações entre a quantidade de aulas e as orientações, com isso, pode-se observar que no mês de Junho o docente teve a maior quantidade de aulas, porém não teve tantas orientações como em outros meses. Diferente do mês de Janeiro onde, a quantidade de aulas, bem como a quantidade de orientações foram acima da média anual.

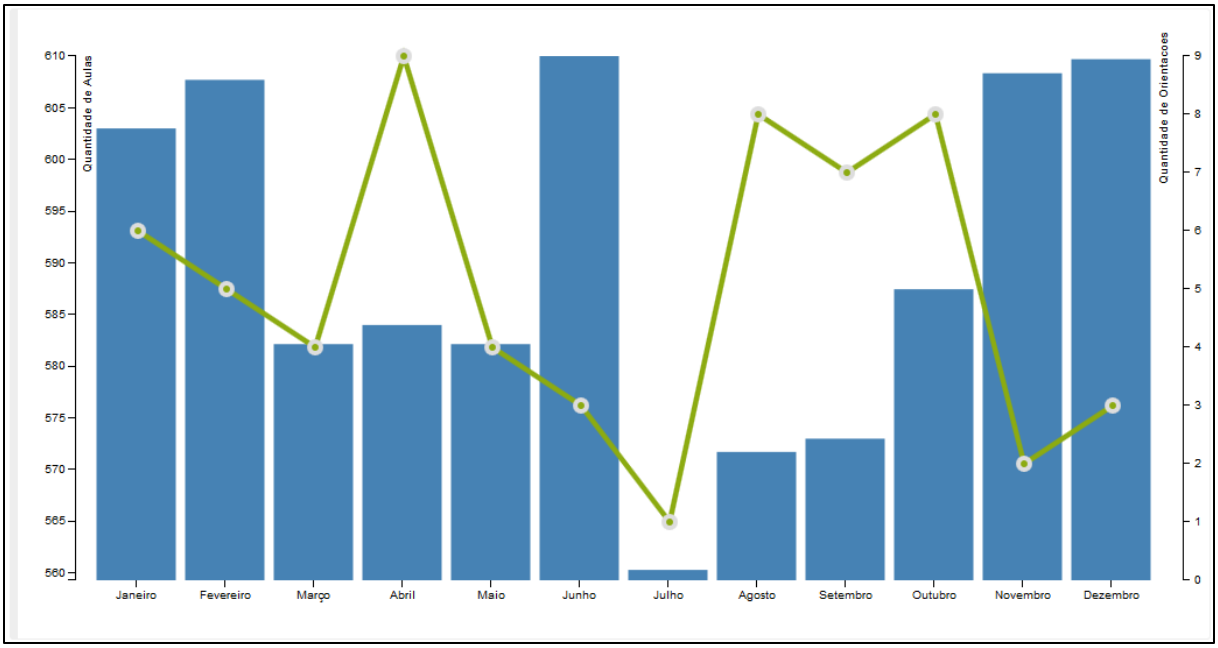


Gráfico 7 – Gráfico com 2 séries de dados. Quantidade de aulas e orientações

No Gráfico 8, os dados foram manipulados (pois nesse trabalho a base de dados do sistema acadêmico é simulada) para exibir uma situação que pode vir a acontecer. Nela, o docente teve um baixo número de aulas no primeiro semestre do ano, porém, um grande número de alunos e/ou projetos em orientações, o que poderia justificar o declínio na quantidade de aulas. No segundo semestre do ano o docente teve um aumento na quantidade de aulas ministradas, porém uma queda nas orientações, se comparado ao semestre anterior. Dessa forma, pode-se inferir, que no ano representado, o docente possui uma divisão aparentemente justa entre dois indicadores (quantidade de aulas e quantidade de orientações).

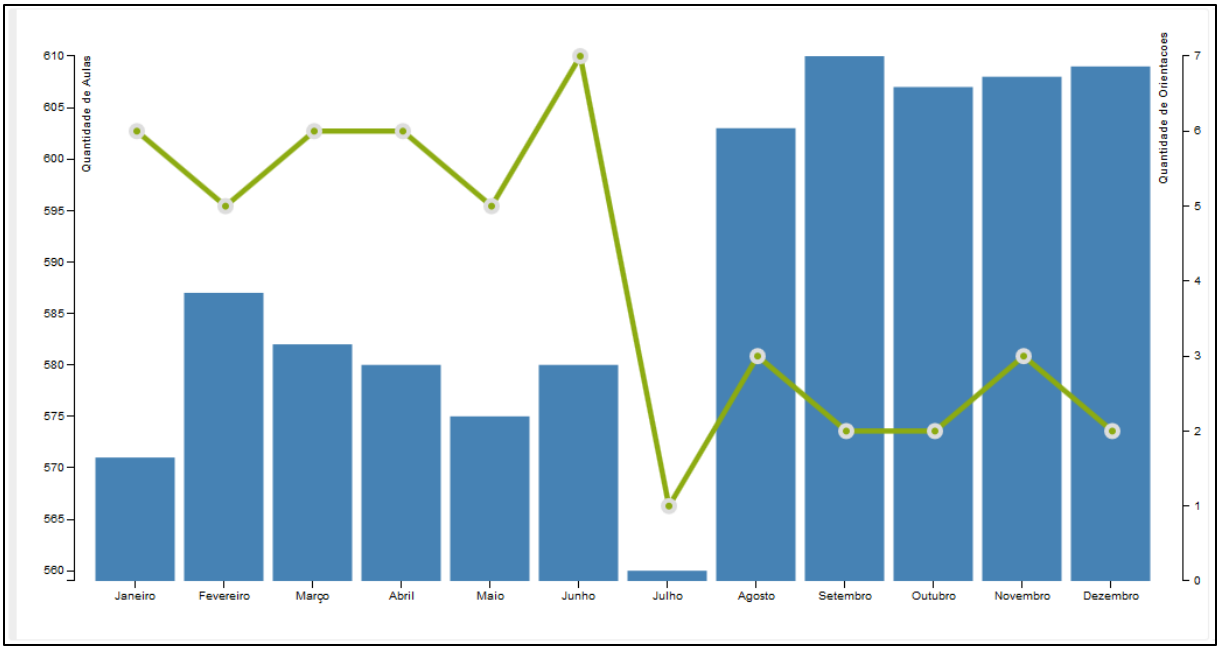


Gráfico 8 – Gráfico com 2 séries de dados. Quantidade de aulas e orientações, exibindo equilíbrio entre os indicadores

O Gráfico 9 exibe um *dashboard* composto por gráficos que informam: a quantidade de aulas; quantidade de orientações e a comparação entre aulas práticas e teóricas. Com esse *dashboard* os três gráficos podem ser visualizados de uma única vez. Os filtros de dados utilizados para esses gráficos são o docente e o período (ano) em que as informações pertencem, esses filtros são compartilhados entre os 3 gráficos.

Gráfico 9 – *Dashboard* que exibe a quantidade de aulas, orientações e divisão do tempo

O Gráfico 10 exibe outro *dashboard* que mostra: a quantidade de aulas de um grupo de docentes no período de um ano. No gráfico de barras, primeiramente são exibidas as quantidade de aulas em cada mês no período de um ano; no gráfico de setores é exibida uma comparação entre a quantidade de aulas no período de um ano de cada docente do grupo; na tabela, a esquerda do gráfico de setores, tem-se um resumo contendo: o nome do docente, um somatório com a quantidade de suas aulas no ano, e o percentual do resultado da comparação entre a quantidade de aulas daquele docente com os demais docentes do grupo que compõe o *dashboard*.

A interação que existe nesse gráfico é apresentada na etapa “7 - Interagir” dessa metodologia. Ela altera completamente a aparência dos gráficos que compõem o relatório.

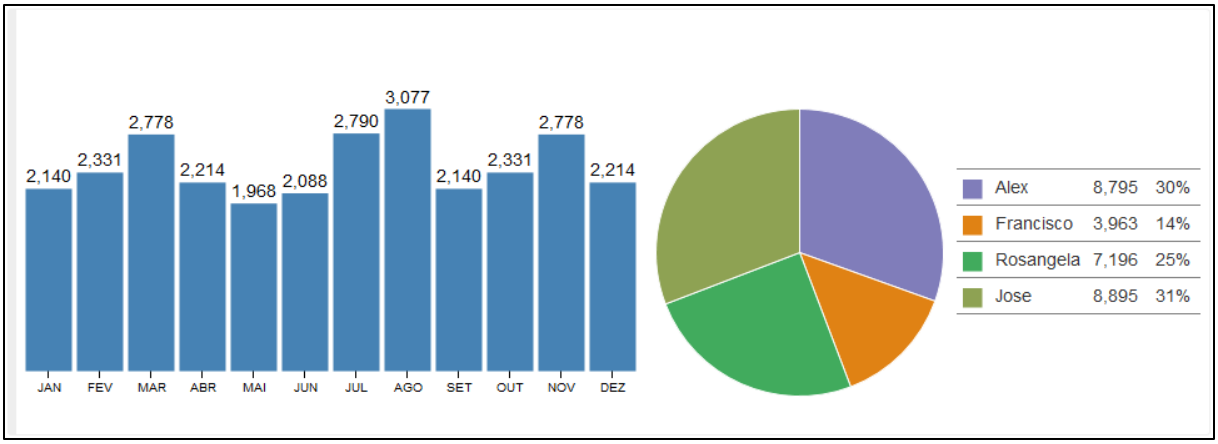


Gráfico 10 – *Dashboard* que exibe a quantidade de aulas de um grupo de docentes

3.3.7 Etapa 7 - Interagir

Até a etapa 6, o objetivo foi aperfeiçoar a qualidade da visualização de dados, entretanto, atualmente, é essencial que um portal de visualização de dados permita que o usuário interaja com o gráfico, permitindo a visualização em diferentes perspectivas e formatos, permitindo que o usuário controle ou explore visualmente os dados. Ações como zoom, navegação, seleção, alteração do modelo gráfico, podem melhorar a experiência do usuário no entendimento da análise dos gráficos representados.

A Figura 40 exibe a EAP com as atividades da etapa de interação, a Figura 41 demonstra o quadro de Kanban dessa etapa.

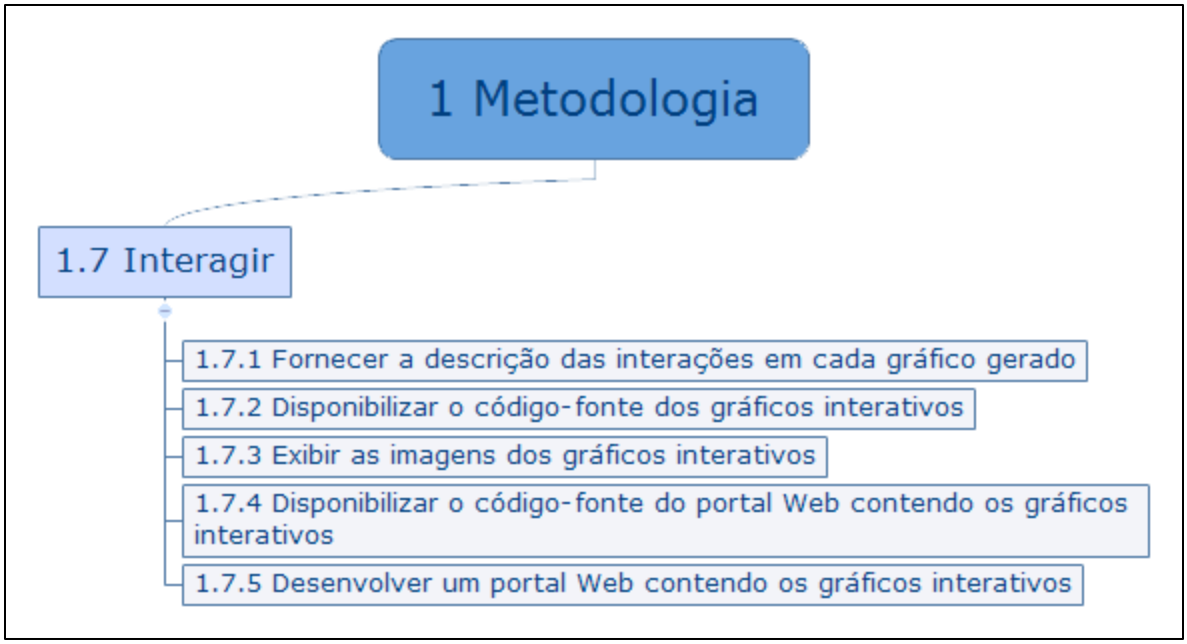


Figura 40 – EAP da etapa de interação

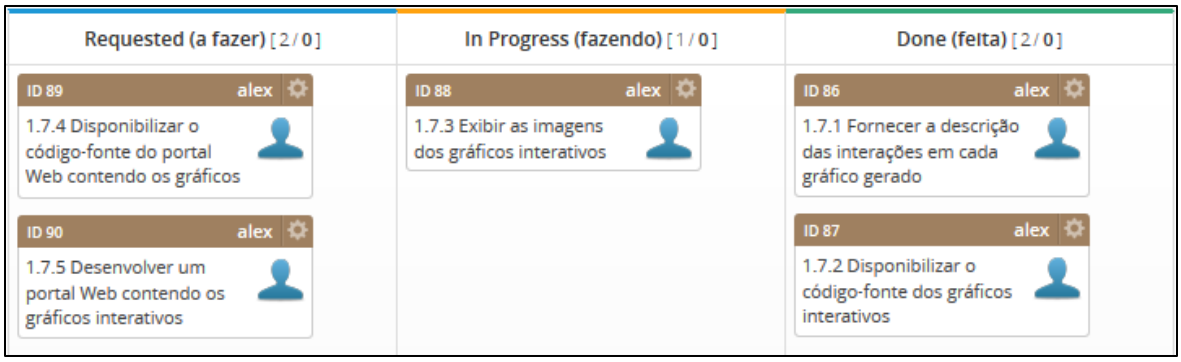


Figura 41 – Quadro de Kanban da etapa de interação

Para essa etapa os seguintes resultados foram entregues:

- Descrição das interações em cada gráfico gerado
- Código-fonte dos gráficos interativos
- Imagens dos gráficos interativos
- Código-fonte do portal Web contendo os gráficos interativos
- Portal Web contendo os gráficos interativos

Gráfico	Interações
Todos	Os gráficos possuem efeitos que fazem que seu carregamento inicial seja feito de maneira suave.

Todos	Quando o usuário passar o mouse em cima de cada parte do gráfico, uma pequena janela (*popup*) irá aparecer com os valores detalhados daquela seção do gráfico.
Rosca (*donut*)	Ao clicar em uma fatia do gráfico, irá abrir uma lista detalhando os dados, ou seja, os periódicos daquela classificação nos quais o docente possui publicações.
Barras (*dashboard* do grupo de docentes)	Ao passar o mouse sobre uma das barras do gráfico presente no *dashboard*, os dados dos outros 2 itens do *dashboard* (gráfico de setores e a tabela) são filtrados pelo mês em que a barra focada pertence. Por exemplo, se o usuário passar o mouse sobre a barra que representa o mês de Agosto, o gráfico de setores e a tabela irá exibir as informações provenientes do mês de agosto.
Setores (*dashboard* do grupo de docentes)	Ao passar o mouse sobre um dos setores do gráfico presente no *dashboard*, o gráfico de barras irá exibir os dados referentes aquele docente do qual o setor representa no *dashboard*.

Quadro 7 – Descrição das interações realizadas em cada gráfico

A Figura 42 demonstra a interação que exibe os valores da série, ao passar o mouse sobre o gráfico, a Figura 43 exibe também os valores, porém no gráfico de setores.

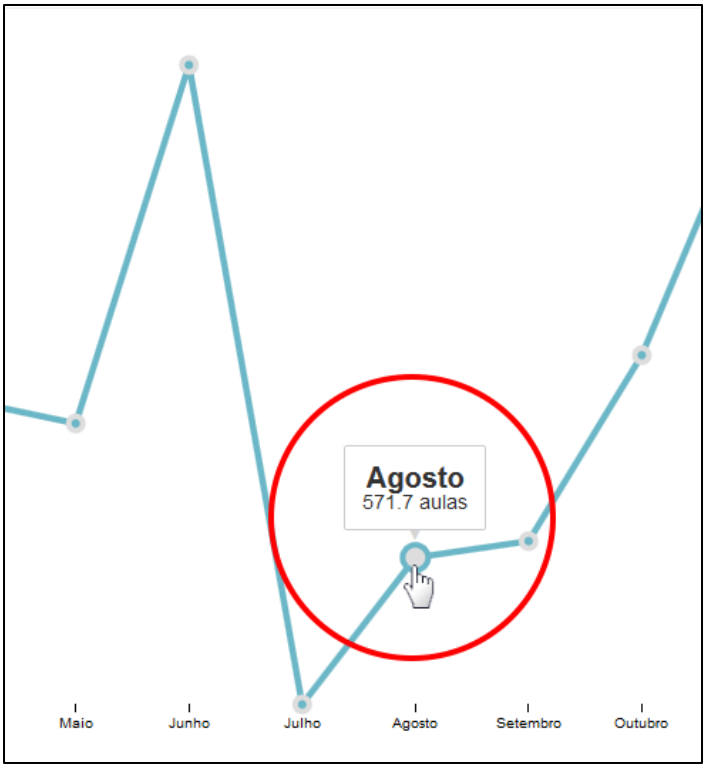


Figura 42 – Interação ao passar o mouse sobre uma série do gráfico de linhas

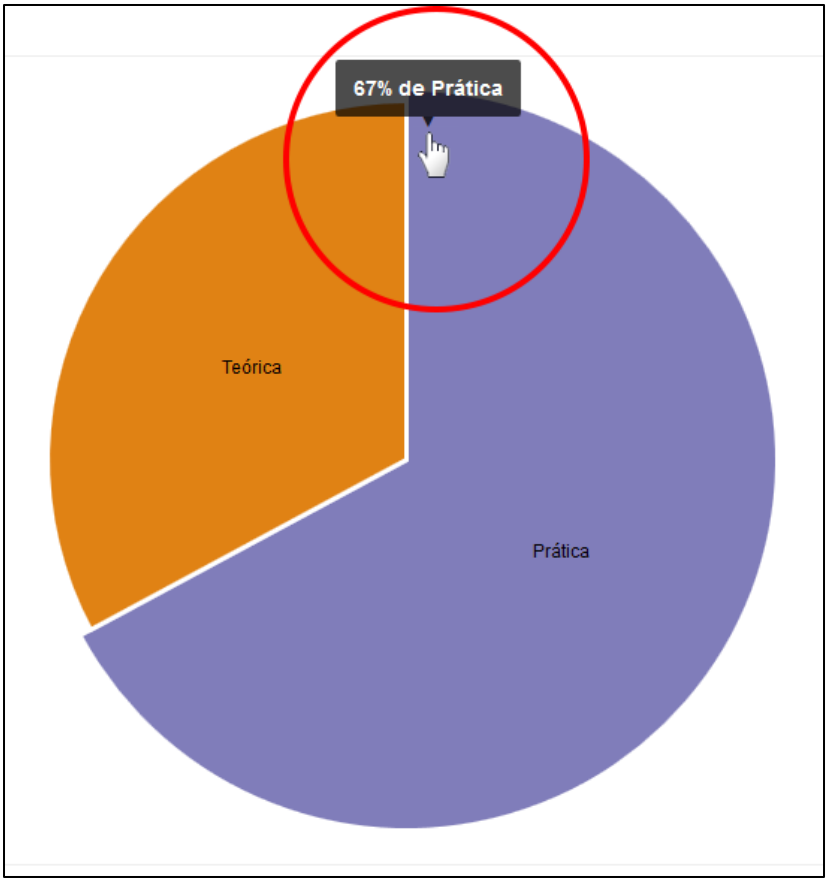


Figura 43 – Interação ao passar o mouse sobre uma série do gráfico de setores

A Figura 44 exibe a interação ao clicar sobre uma fatia do gráfico de rosca com as informações acerca da classificação (Qualis) das publicações de um docente.

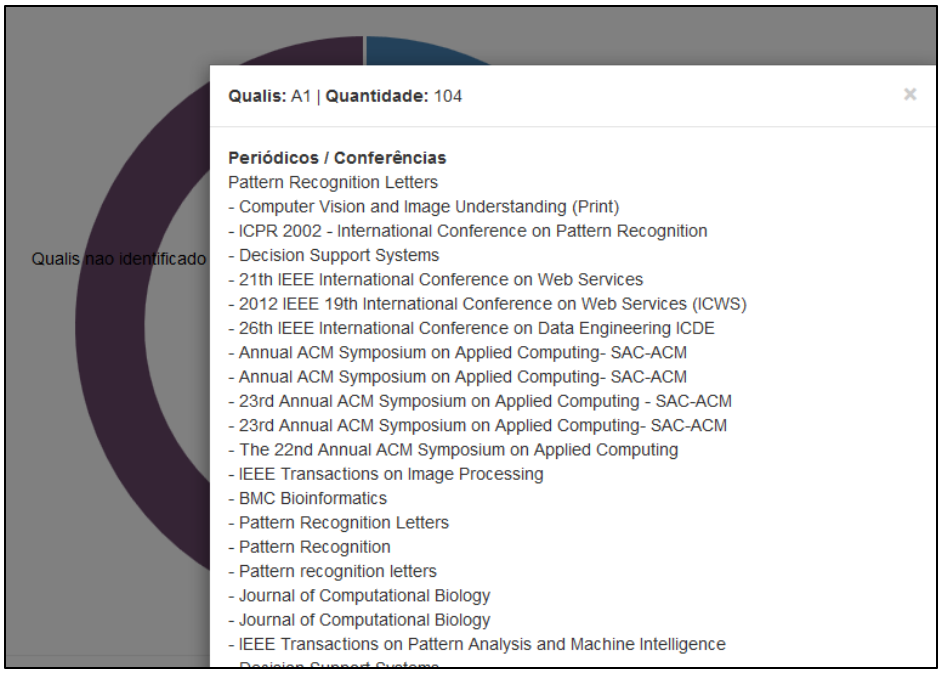


Figura 44 – Interação ao clicar em uma série do gráfico de setores

A Figura 45 representa as interações permitidas no gráfico. Ao passar o ponteiro do *mouse* sobre uma das barras, do gráfico de barras presente no *dashboard* de quantidade de aulas de um grupo de docentes, os dados da área destacada na cor vermelha, são filtrados de acordo com o mês que representa a barra escolhida (foco do *mouse*).

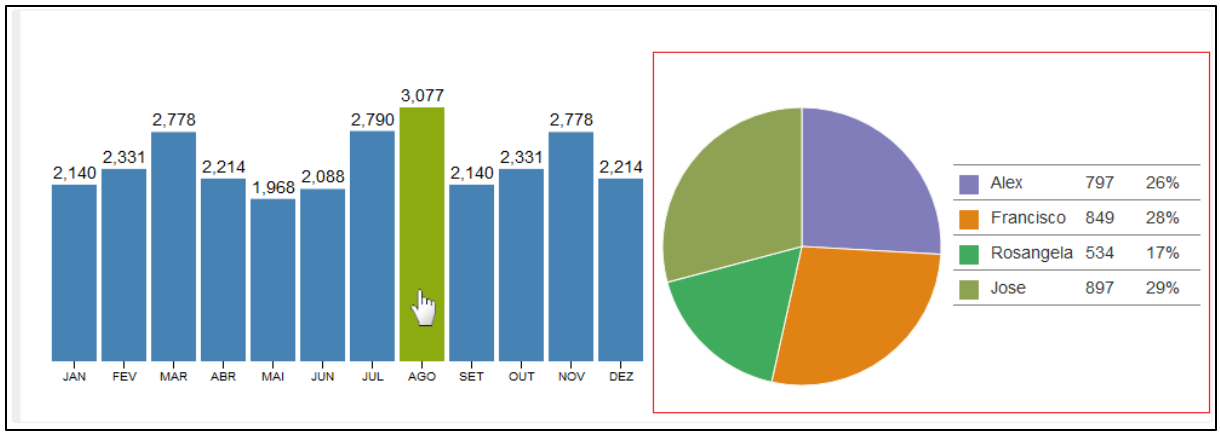


Figura 45 – Interação ao passar o mouse sobre uma barra no *dashboard* de informações sobre um grupo de docentes

Já a Figura 46 demonstra a interação que acontece quando o usuário passa com o ponteiro do *mouse* sobre um setor do gráfico de setores presente no *dashboard* de quantidade de aulas de um grupo de docentes. Ao passar o ponteiro do *mouse* as cores e os valores do gráfico de barras (destacado na cor vermelha) são filtrados de acordo com o docente representado pelo setor escolhido (foco do *mouse*).

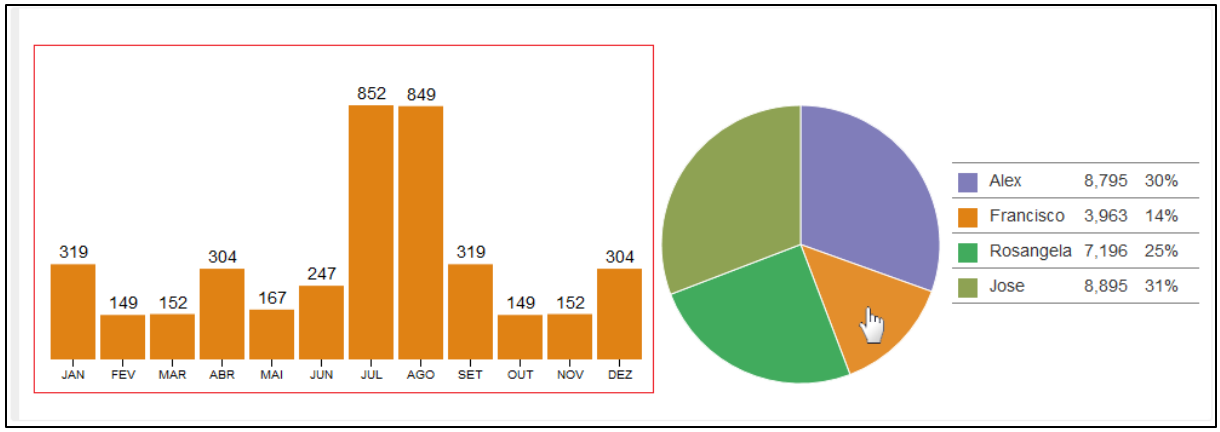


Figura 46 – Interação ao passar o mouse sobre um setor no *dashboard* de informações sobre um grupo de docentes

3.4 SMARTIES

Para exibição dos *dashboards* e gráficos, foi desenvolvido um sistema de informações denominado SmartIES. Analisando as 7 etapas anteriormente descritas, pode-se afirmar que esse sistema engloba as etapas 6 - Visualizar e 7 - Interagir.

A Figura 47 exibe o fluxo de dados percorrido desde a aquisição até a etapa final que está destacada em cor vermelha nesse diagrama. As setas, exibem qual a direção em que os dados estão tramitando. A parte inferior da Figura 47 mostra o local onde os dados que serão extraídos estão armazenados, o Currículo Lattes e o Sistema Acadêmico. Em uma camada acima são exibidas as ferramentas utilizadas para fazerem a extração dos dados armazenados, que no caso do Currículo Lattes é a ferramenta scriptLattes, já no caso do sistema acadêmico a base de dados é acessada diretamente pelo PDI e/ou Hadoop não tendo nenhuma ferramenta intermediária. Na etapa de processamento dos dados são utilizados o Hadoop e o PDI, tanto para os dados do Sistema Acadêmico, quanto para os dados do Currículo Lattes. Após o processamento dos dados, eles são exibidos na camada de Apresentação e interação. Nessa camada está o sistema SmartIES.

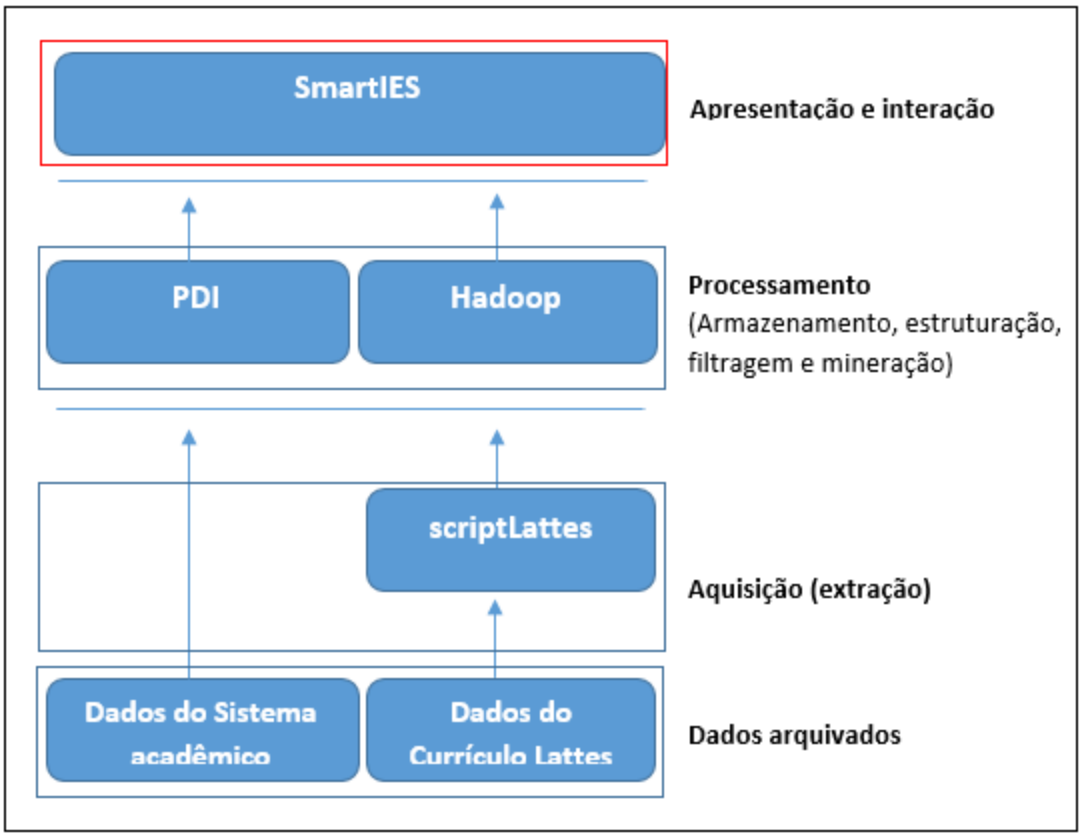


Figura 47 – Interação ao clicar em uma série do gráfico de setores

O SmartIES é um sistema Web que foi desenvolvido utilizando o padrão de arquitetura de *software* *Model-View-Controller* (MVC), que separa em camadas a representação do código da interação com o usuário.

Na camada de modelagem (*Model)*, foi utilizada a linguagem de programação Java. Na persistência de dados no SGBD foi utilizado a API Java *Persistence* API (JPA) juntamente com o *framework* de Mapeamento de Objeto-Relacional (ORM), Hibernate, dessa forma, toda a estruturação de SGBD foi realizada nas classes da camada *Model* do projeto. Ainda na camada *Model*, todas as classes responsáveis por acessar o SGBD diretamente, estão utilizando o padrão de projetos chamado Objeto de Acesso aos Dados (*Data Access Object* ou DAO), separando as regras de negócios, do acesso aos dados. O SGBD utilizado nesse sistema é o Posgresql por possui grande robustez e possuir o código aberto.

Na camada de visualização (*View*) foi utilizada a tecnologia Java *Server Pages* (JSP) que torna possível a geração de páginas HTML ou XML dinâmicas, sendo possível inserir códigos Java diretamente nessas páginas. Embora existam outras tecnologias e *frameworks* que simplificam o desenvolvimento de páginas Web

com Java, como por exemplo o Java *Server Faces* (JSF), o JSP foi utilizado devido a sua boa performance em ambiente Web. Foi utilizado também o HTML na versão 5 e o CSS na versão 3, bem como a linguagem de programação Javascript que é largamente utilizada nesse sistema, tanto na geração dos gráficos (D3.js) quanto na exibição de páginas e interação com os usuários.

Na camada de *Controller* onde as mediações entre o *Model* e as *Views* são realizadas, foi utilizado o conceito de *Servlets*, sendo que todas as requisições HTTP passam pelos *Controllers* do sistema, sendo que nenhuma página da camada de visualização é acessada diretamente. A estrutura MVC do projeto é exibida na Figura 48, nesse diagrama é possível verificar que o Navegador Web, nunca se realiza nenhuma requisição HTTP diretamente para a camada *View*, sendo que sempre irá passar por um *Servlet* da camada *Controller*, com isso, nenhum Usuário que não esteja logado, consegue acessar o sistema, pois todos os *Servlets* verificam os acessos.

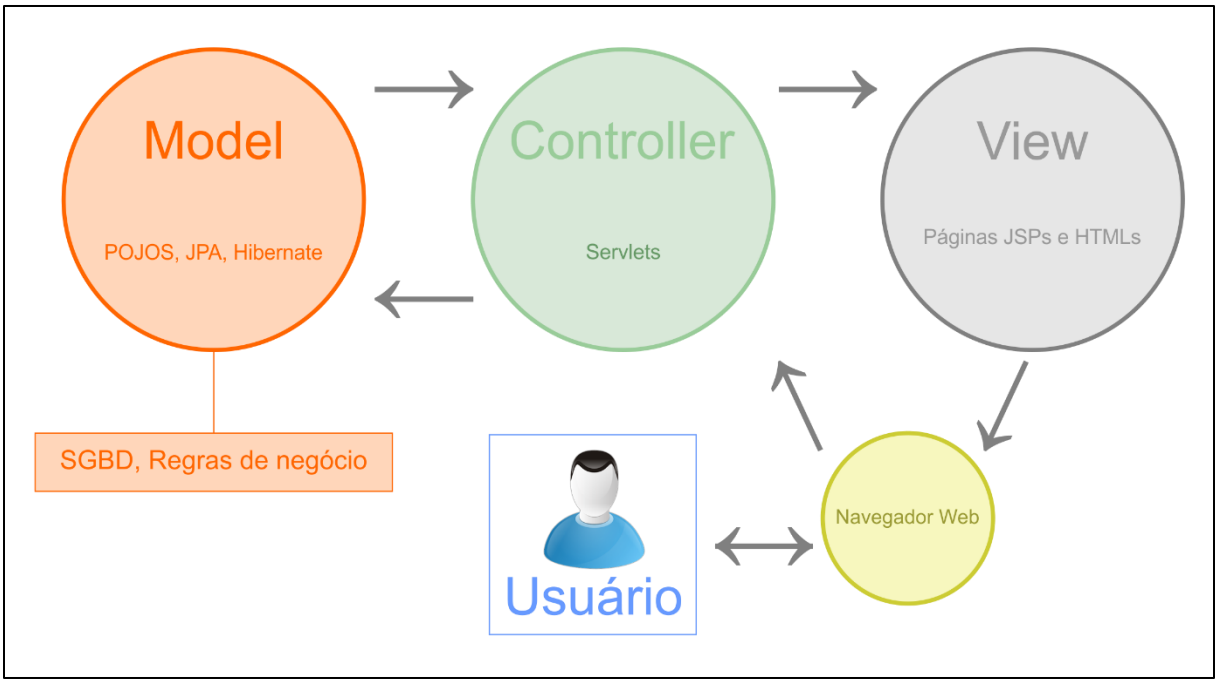


Figura 48 – Estrutura MVC do sistema SmartIES

O sistema SmartIES é um sistema responsivo, portanto ele pode ser acessado com diferentes resoluções e diferentes tamanhos de telas o que torna bastante útil para ser acessado tanto de computadores de mesa (*desktops*) e *notebooks*, quanto para dispositivos móveis: *tablets* e *smartphones*. Para que esse

recurso fosse possível no sistema, foi utilizada a biblioteca Bootstrap, que é um *framework* sobre a linguagem Javascript e CSS que simplifica a geração de conteúdo Web que se adapta a diferentes tamanhos de telas e resoluções.

A Figura 49 exibe, em uma resolução padrão de computadores *desktops* e *notebooks* (1024 por 768 pixels), a página Web do SmartIES onde está o *dashboard* que exibe os dados de quantidade de aulas, orientações e a divisão do tempo de um docente. Já a Figura 50 exibe a mesma página, com o mesmo *dashboard*, utilizando os mesmos dados, porém em uma resolução menor (400 por 768 pixels) comumente encontrada em *smartphones*. Pode-se observar que ao diminuir a resolução, os gráficos permanecem com o mesmo formato, porém eles irão se adequar ao tamanho da tela sem que seja necessária nenhuma alteração do usuário.

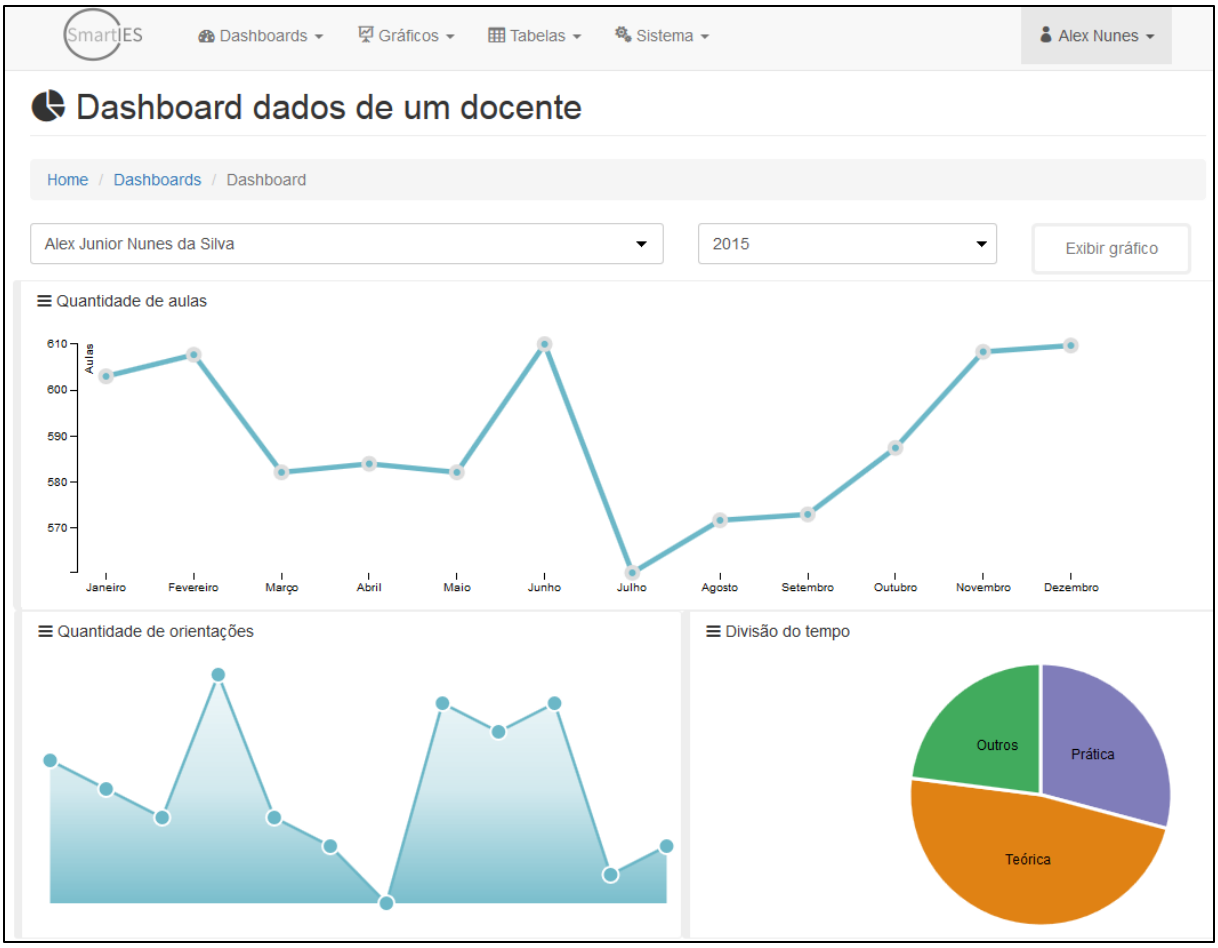


Figura 49 – *Dashboard* exibido em uma tela com grande resolução

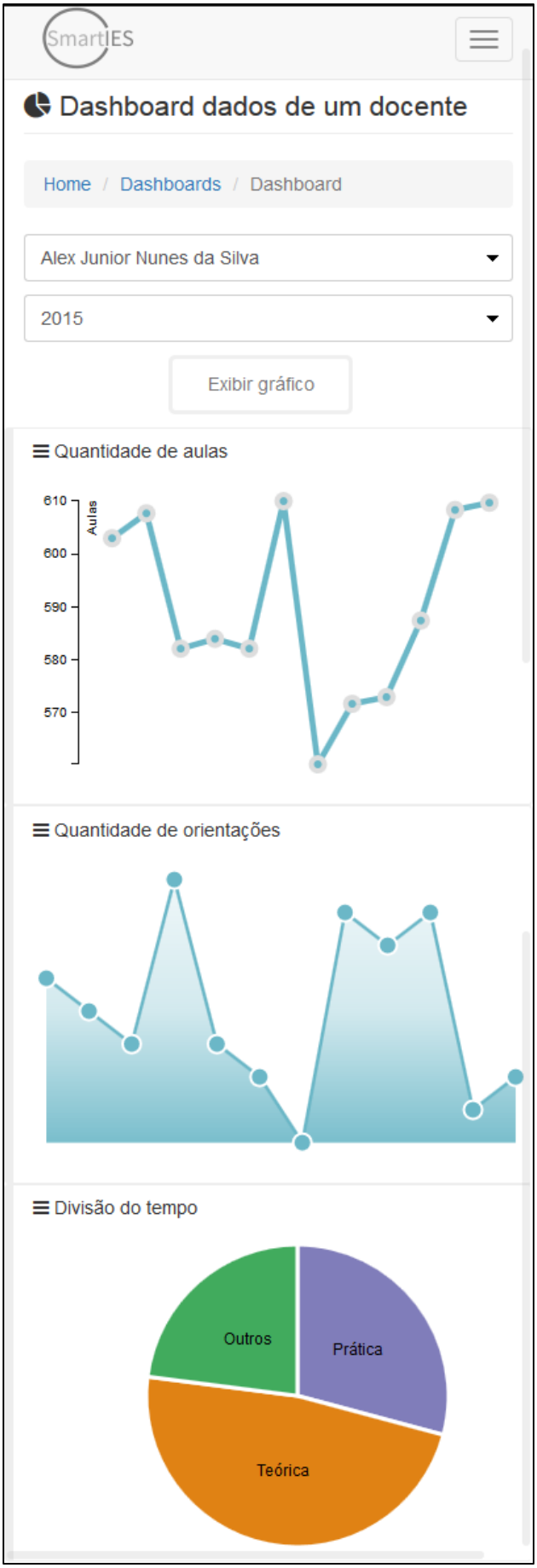


Figura 50 – *Dashboard* exibido em uma tela de baixa resolução

O menu principal do sistema também é alterado quando a tela é redimensionada, os itens do menu que antes eram dispostos na direção horizontal, passam a ficar na vertical. A Figura 51 exibe o comportamento do menu após a interação do usuário (clique) no botão que expande o menu (destacado na cor vermelha).

Figura 51 – Menu expandido do sistema SmartIES em telas com baixa resolução

3.5 RESUMO DAS ATIVIDADES

As figuras 52A e 52B exibe um diagrama de Notação de Modelo de Processo de Negócios (BPMN) com a arquitetura geral do processo de desenvolvimento desse trabalho. Nela é possível identificar cada etapa, que corresponde a uma cor no diagrama e cada atividade é representada por um retângulo que deve entregar no mínimo um artefato, os artefatos que necessitam de armazenamento são destacados com a cor amarela. Embora o processo seja linear, a etapa de armazenamento é uma exceção, pois ao final de várias etapas é necessário o armazenamento dos resultados.

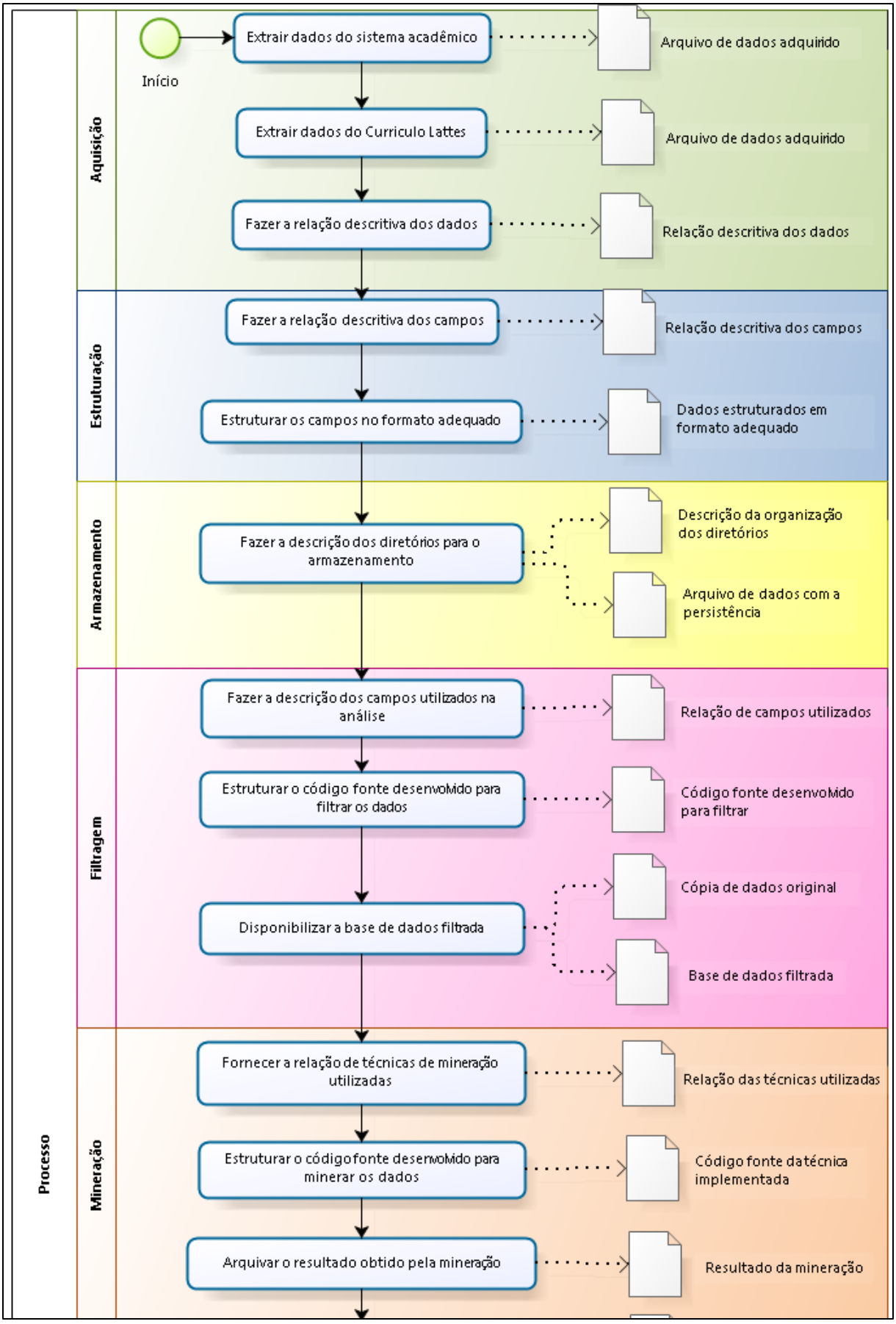


Figura 52A – Diagrama BPMN com a arquitetura geral do processo, parte 1

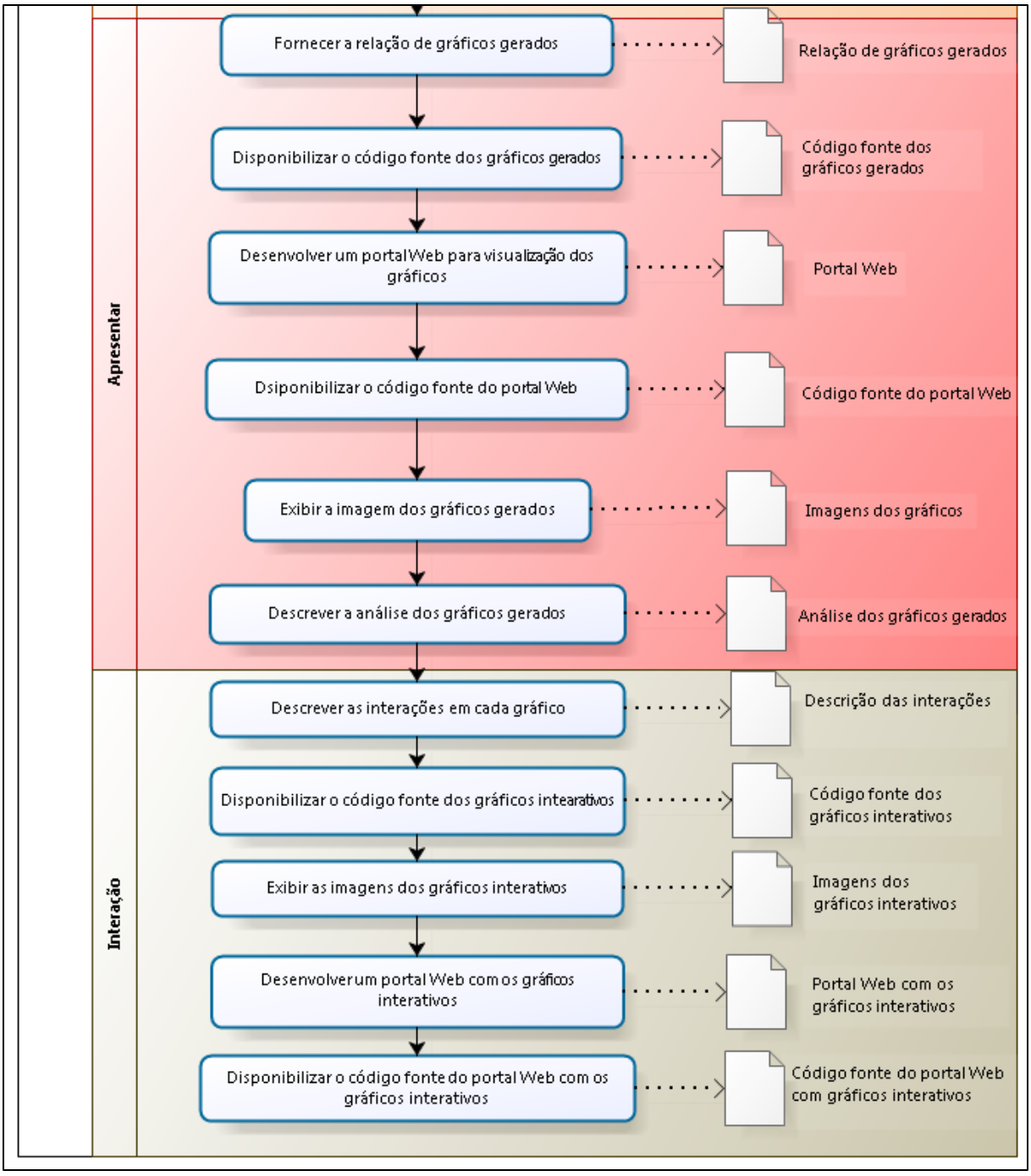


Figura 52B – Diagrama BPMN com a arquitetura geral do processo, parte 2

4 RESULTADOS E DISCUSSÕES

Nessa seção são apresentados e discutidos os resultados obtidos no desenvolvimento do trabalho.

4.1 DISCUSSÕES

Na etapa 5 do projeto (mineração), houve uma comparação entre o PDI e o MapReduce do Hadoop para processar os mesmos dados. Com essa comparação pode-se constatar que o PDI é mais simples de ser utilizado e seus componentes (estágios) auxiliam bastante em tarefas comuns, pois eles dividem as responsabilidades do processo e as mesmas podem ser visualmente gerenciadas de maneira individual.

Usando o PDI também não são necessários conhecimentos avançados em linguagens de programação para tarefas mais simples e comuns, pois os seus componentes já as fazem eliminando a necessidade de serem programadas. Como por exemplo a contagem de uma quantidade de valores, que para o PDI, basta informar o campo para agrupar os valores e o campo que contém os valores que o componente "*Group By*" a fará, conforme ilustra a Figura 53. Caso fosse realizada no MapReduce, um algoritmo que contasse os valores deveria ser implementado. Em uma etapa do projeto, houve uma comparação entre o PDI e o Hadoop para processar alguns dados, com isso pode-se constatar que o PDI é muito mais simples de ser utilizado e seus componentes (estágios) auxiliam bastante em tarefas comuns, todavia, o PDI não trabalha de maneira paralela ou em *cluster*, por isso, para um volume grande de dados o Hadoop é o mais indicado.

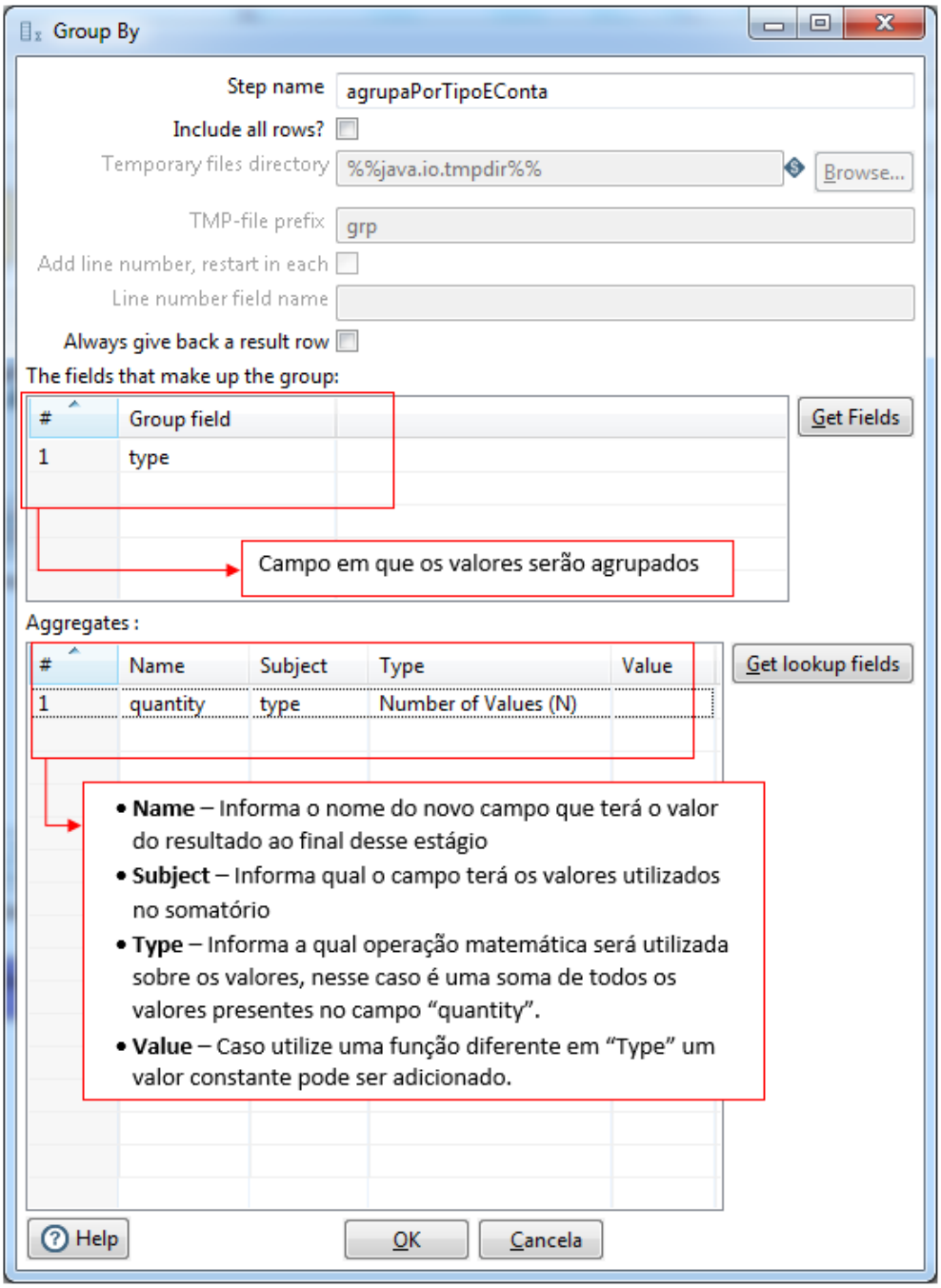


Figura 53 – Detalhamento do componente *Group By* do PDI

Na comparação realizada houve uma perda de performance ao utilizar o MapReduce, pois o volume de dados manipulado era relativamente pequeno. O processo completo da etapa de mineração no PDI gastou cerca de 4,2 segundos, já no Hadoop, o tempo gasto para o processamento da mesma quantidade de informações no mesmo ambiente (*hardware*), foi de cerca de 16,7 segundos. Todavia, o PDI não trabalha de maneira paralela ou em cluster, por isso, para um volume grande de dados o Hadoop é o mais indicado.

O desenvolvimento desse trabalho apresenta uma solução para fornecer auxílio na melhoria das atividades acadêmicas realizadas na UTFPR campus Cornélio Procópio, todavia, esse trabalho pode ser utilizado para a implantação em outras IES, desde que o processo de desenvolvimento atenda os 7 passos dos tópicos descritos na metodologia desse trabalho.

4.2 PROBLEMAS ENCONTRADOS

Diversas atividades foram realizadas até a conclusão desse trabalho, um ponto importante a se destacar é que conhecimentos em sistemas operacionais Linux, tais como: instalação e configuração de programas, permissões de diretórios e *firewalls*, manipulação de arquivos, dentre outros, foram indispensáveis para se trabalhar com as ferramentas utilizadas nesse trabalho.

Uma dificuldade encontrada foi a documentação e o gerenciamento do projeto, pois muitas das técnicas utilizadas em Engenharia de *Software* possuem especificidades que não se adaptam as soluções de análise de Big Data, por esse motivo, foi adaptado uma metodologia de visualização de dados ao processo desse trabalho.

Outro ponto a se destacar, foi a configuração do Apache Hadoop, que embora possua uma documentação eficiente, os problemas que aparecem exigem conhecimento da estrutura do Hadoop e a compreensão de como os componentes interagem entre si.

Para utilizar o D3.js, foi necessário adquirir um grande conhecimento na linguagem de programação Javascript, pois embora ele seja uma biblioteca criada para simplificar o processo de geração de gráficos, os algoritmos criados para a plotagem de cada gráfico tiveram que ser desenvolvidos ou adaptados.

4.3 CONCLUSÃO

O grande volume de geração de dados está mudando a cada dia a perspectiva da Tecnologia da Informação, dessa forma, é necessário cada vez mais soluções que consigam extrair valores desses grandes aglomerados de dados.

As soluções de análise de Big Data, devem ser utilizadas não apenas em instituições privadas, mas também em públicas, pois ambas possuem problemas com dados que podem ser resolvidos e processos podem ser otimizados com uma tomada de decisões com maior assertividade. Ainda um outro fator positivo é que muitas soluções robustas são de código aberto e não possuem licenças com valores elevados, o que pode ser um excelente atrativo paras as IES e seus pesquisadores.

Para uma solução de análise de Big Data é necessário a aquisição dos dados, o processamento para a preparação e estruturação deles e a exibição desses dados de maneira apropriada. Para o desenvolvimento desse trabalho diferentes tecnologias foram compreendidas e utilizadas.

Com a ferramenta desenvolvida nesse trabalho, os dados de desempenho dos docentes podem ser visualizados de maneira mais adequada para um grande conjunto de dados, por conseguinte, processos internos da instituição podem ser melhorados para otimizar a utilização no tempo de trabalho dos docentes.

4.4 LIMITAÇÕES

A solução descrita nesse trabalho é alimentada apenas por dados estruturados ou semiestruturados. Para utilizar dados não estruturados, o PDI deverá ser substituído por outra ferramenta que consiga atender a esses dados.

4.5 TRABALHOS FUTUROS

Embora as fontes de dados utilizadas nesse trabalho tenham sido somente o Currículo Lattes e o sistema acadêmico, em trabalhos posteriores outras fontes de dados poderão ser utilizadas e acrescentadas à solução desenvolvida, pois como trata-se de uma solução de análise de Big Data, onde o grande volume de dados já é

esperado. Quanto mais informações diferentes forem acrescidas ao banco de dados do sistema criado, uma maior base as análises terão, portanto mais refinadas as informações finais serão.

Esse trabalho contou com uma etapa que utiliza a mineração de dados (*data mining*), todavia, a etapa foi extremamente simples pois a mineração não foi o foco principal do trabalho. Porém em trabalhos futuros os dados gerados pela solução desenvolvida por esse, poderão servir de base para a previsão de eventos através de algoritmos de inteligência artificial, como por exemplo, com base na tendência da quantidade de aulas no ano, a quantidade de turmas que o docente atende e os projetos externos, pode-se prever que ele terá uma sobrecarga de atividades enquanto outro docente poderá ter seu tempo subutilizado, assim a divisão de responsabilidades poderia ser melhor realizada.

REFERÊNCIAS

ALVES, A. D. et al. **Mapeamento de Competências Tecnológicas: Technological Competencies Mapping: A Petrobras' Information System for aiding the decision process based on the Lattes Platform**Information Systems and Technologies (CISTI), 2015 10th Iberian Conference on. **Anais**...2015

ALVES, A. D.; YANASSE, H. H.; SOMA, N. Y. SUCUPIRA: A system for Information extraction of the Lattes Platform to identify academic social networks. **6th Iberian Conference on Information Systems and Technologies (CISTI 2011)**, p. 1–6, 2011.

ALVES, A. D.; YANASSE, H. H.; SOMA, N. Y. LattesMiner: uma linguagem de domínio específico para extração automática de informações da Plataforma Lattes. **XII Workshop de Computação Aplicada -WORCAP 2012**, 2012.

BAO, F.; CHEN, J. Visual framework for big data in d3.js. **Proceedings - 2014 IEEE Workshop on Electronics, Computer and Applications, IWECA 2014**, p. 47–50, 2014.

BRYNJOLFSSON, E.; MCAFEE, A. Big data: the management revolution. **Harvard Business Review**, v. 90, n. 10, p. 60–68, 128, 2012.

CAPES. **Plataforma Sucupira**. Disponível em: <https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/docente/listaDocente.jsf>. Acesso em: 3 nov. 2015.

CASTERS, M.; BOUMAN, R.; DONGEN, J. VAN. **Pentaho Kettle Solutions - Building Open Source ETL Solutions with Pentaho Data Integration**. 1 Edition ed. Indianapolis, IN: Wiley Publishing, Inc., 2010.

CHALCO, J. M. **scriptLattes**. Disponível em: <http://scriptlattes.sourceforge.net/description.html>. Acesso em: 4 nov. 2015.

Computing | CERN. Disponível em: <http://home.web.cern.ch/about/computing>. Acesso em: 27 mar. 2015.

DELSOTO, D. A Influência Do Big Data No Business Intelligence. v. d, 2013.

Facebook Investors. Disponível em: <http://investor.fb.com/releasedetail.cfm?ReleaseID=861599>. Acesso em: 24 abr. 2015.

FERRAZ, R. R. N.; QUONIAM, L.; ALVARES, L. M. A. D. R. Avaliação de redes multidisciplinares com a ferramenta scriptlattes: os casos da nanotecnologia, da dengue e de um programa de pós-graduação Stricto Sensu em Administração. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 19, n. 40, p. 67, 2014.

FRY, B. **Visualizing data**. First Edit ed. Sebastopol, CA: O'Reilly Media, Inc., 2008.

GANTZ, J.; REINSEL, D. Extracting Value from Chaos. **IDC iView**, n. June, p. 1–12, 2011.

GOLDMAN, A. et al. Apache Hadoop - Conceitos Teóricos e Práticos, Evolução e Novas Possibilidades. **Csbc**, 2012.

Google Inverstors. Disponível em: <http://investor.google.com/earnings/2014/Q2_google_earnings.html>. Acesso em: 24 abr. 2015.

GUEDES, C. A. **Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Currículo Lattes: Perguntas e Respostas.** Disponível em: <http://www.pucrs.campus2.br/manuais/dicas_lattes.pdf>. Acesso em: 20 abr. 2015.

HU, H. et al. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. **IEEE Access**, v. 2, p. 652–687, 2014.

KERZNER, H. **Project Management Metrics, KPIs, and Dashboards: A Guide to Measuring and Monitoring Project Performance**. First Edit ed. New Jersey: John Wiley & Sons, 2013.

LEE, S.; JO, J.; KIM, Y. **Performance Testing of Web-Based Data Visualization**IEEE Int'l Conference on Systems, Man, and Cybernetics. **Anais**...Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on, 2014

MANYIKA, J. et al. Big data: The next frontier for innovation, competition, and productivity. **McKinsey Global Institute**, n. June, p. 156, 2011.

MARINHEIRO, A.; BERNARDINO, J. Analysis of Open Source Business Intelligence Suites Análise de Suites Open Source Business Intelligence. 2013.

MENA-CHALCO, J. P.; JUNIOR, R. M. C. scriptLattes: an open-source knowledge extraction system from the Lattes platform. **Journal of the Brazilian Computer Society**, v. 15, n. 4, p. 31–39, 2009.

MURRAY, S. **Interactive Data Visualization for the Web**. First Edit ed. Sebastopol: O'Reilly Media, 2013.

NANDIMATH, J. et al. Big data analysis using Apache Hadoop. **2013 IEEE 14th International Conference on Information Reuse & Integration (IRI)**, p. 700–703, 2013.

OLIVEIRA, A. B. et al. Comparação entre o Qualis/Capes e os índices H e G: o caso do portal de periódicos UFSC. **Informação & Informação**, v. 20, n. 1, p. 70, 2015.

PAPER, C.; CATARINA, S. Gestão Estratégica de Informações Curriculares em ICTIs. n. September 2015, p. 0–15, 2012.

PoweredBy - Hadoop Wiki. Disponível em: <http://wiki.apache.org/hadoop/PoweredBy>. Acesso em: 14 abr. 2015.

SRIVASTAVA, D.; DONG, X. Big data integration. **… 2013 IEEE International Conference on Data …**, p. 1245–1248, 2013.

STELA EXPERTA. **Stela Experta**. Disponível em: <http://www.stelaexperta.com.br/experta/index.html>. Acesso em: 4 set. 2015.

TEKINER, F.; KEANE, J. A. Big data framework. **Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013**, p. 1494–1499, 2013.

TURKINGTON, G. **Hadoop Beginner's Guide**. 1ª Edition ed. BIRMINGHAM - MUMBAI: Packt Publishing Ltd., 2013.

VAVILAPALLI, V. K. et al. Apache Hadoop YARN : Yet Another Resource Negotiator. **ACM Symposium on Cloud Computing**, p. 16, 2013.

What Is Big Data? - Gartner IT Glossary. Disponível em: <http://www.gartner.com/it-glossary/big-data>. Acesso em: 14 abr. 2015.

WHITE, T. **Hadoop: The Definitive Guide, 3rd edition**. Third Edit ed. Sebastopol: O'Reilly Media, 2012. v. 54

WINKLER, R.; BARR, A. **Google Reports Better-Than-Seen Revenue Growth - WSJ**. Disponível em: <http://www.wsj.com/articles/google-reports-22-revenue-growth-1405628219>. Acesso em: 24 abr. 2015.

YARN - The Architectural Center of Enterprise Hadoop. Disponível em: <http://br.hortonworks.com/hadoop/yarn/>. Acesso em: 22 out. 2015.

ZHU, N. Q. **Data Visualization with D3.js Cookbook**. Birmingham: Packt Publishing Ltd., 2013.

APÊNDICE A – Diagrama de Entidade e Relacionamento (DER) Simulado do Sistema Acadêmico

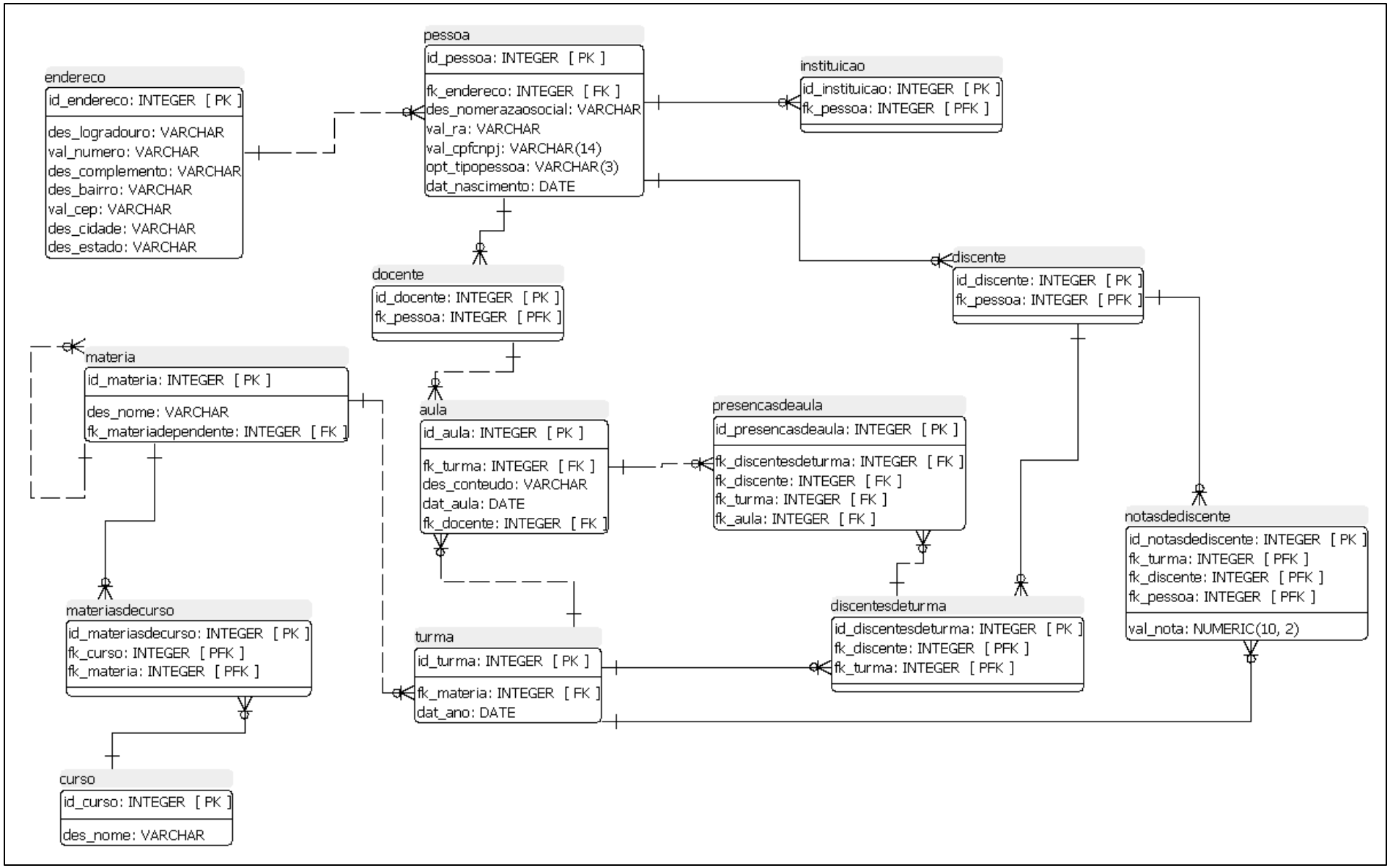


Diagrama de Entidade e Relacionamento (DER) Simulado do Sistema Acadêmico.