UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

CÂMPUS CORNÉLIO PROCÓPIO

CURSO SUPERIOR EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

ALEX JUNIOR NUNES DA SILVA

**BIG DATA COM HADOOP**

PROPOSTA DE TRABALHO DE CONCLUSÃO DE CURSO

CORNÉLIO PROCÓPIO

2015

ALEX JUNIOR NUNES DA SILVA

**BIG DATA COM HADOOP**

Proposta de Trabalho de Conclusão de Curso de graduação, apresentado ao 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 em Análise e Desenvolvimento de Sistemas.

Orientador: Prof. Francisco Pereira Júnior

CORNÉLIO PROCÓPIO

2015

**RESUMO**

SILVA, Alex Junior Nunes. IMPLANTAÇÃO DE SOLUÇÃO DE BUSINESS INTELLIGENCE PARA GERENCIAMENTO ACADÊMICO DE DISCENTES. 2014. 14 f. Proposta de Trabalho de Conclusão de Curso (Graduação) – Tecnologia em Análise e Desenvolvimento de Sistemas. Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2014.

Atualmente, os discentes da Universidade Tecnológica Federal do Paraná (UTFPR), não possuem um sistema de acompanhamento acadêmico no qual possam obter informações analíticas detalhadas, gráficos e tabelas sobre seu rendimento no decorrer do curso, sendo assim, esse trabalho visa implantar uma solução de *Business Intelligence* (BI), que são soluções muito utilizadas em empresas para analisar dados de históricos e apoiá-las nas tomadas de decisões, de uma forma semelhante ao *Student Relationship Management* (SRM), que são soluções implantadas em empresas para otimizar o relacionamento com clientes (CRM) adaptadas ao meio acadêmico. Com a implementação dessa solução, tem-se como objetivo a diminuição de uma das causas do insucesso escolar e consequentemente a diminuição do nível de evasão escolar em cursos superiores de universidades públicas que tem sido bastante elevado.

**Palavras-chave:** CRM, SRM, Student Relationship Management, Inteligência de Negócios, Business Intelligence, BI, Sistemas de Suporte a Decisões.

**LISTA DE SIGLAS**

|  |  |
| --- | --- |
| HDFS | *Hadoop Distributed File System* |
| GPS | *Global Positioning System* |
| SGBD | Sistema de Gerenciamento de banco de dados |
| IDC |  |
| LHC | *Large Hadron Collider* |

**LISTA DE ACRÔNIMOS**

|  |  |
| --- | --- |
| CERN | *Conseil Européen pour la Recherche Nucléaire* |
|  |  |

**SUMÁRIO**

|  |  |
| --- | --- |
| **1 INTRODUÇÃO.................................................................................................** | **6** |
| 1.1 PROBLEMATIZAÇÃO................................................................................... | 7 |
| 1.2 OBJETIVOS.................................................................................................. | 8 |
| 1.2.1 Objetivo geral............................................................................................. | 9 |
| 1.2.2 Objetivos específicos................................................................................. | 9 |
| 1.3 JUSTIFICATIVA............................................................................................ | 9 |
| 1.5 ORGANIZAÇÃO............................................................................................ | 10 |
| **2 FUNDAMENTAÇÃO TEÓRICA.......................................................................** | **11** |
| 2.1 *Customer Relationship Management................*............................................ | 11 |
| 2.2 *Student Relationship Management*............................................................... | 11 |
| 2.3 *Business Intelligence.*................................................................................... | 11 |
| 2.4 TRABALHOS RELACIONADOS................................................................... | 12 |
| **3 PROCEDIMENTOS METODOLÓGICOS........................................................** | **13** |
| 3.1 FERRAMENTAS........................................................................................... | 13 |
| 3.2 MÉTODOS.................................................................................................... | 13 |
| **REFERÊNCIAS...................................................................................................** | **14** |
|  |  |
|  |  |

# 1 INTRODUÇÃO

A quantidade de dados computacionais vem crescendo a cada dia. Os dispositivos eletrônicos têm se tornado cada vez mais inteligentes, gerando novos dados a cada instante, esses dados são gerados de várias fontes: interações de usuários com redes sociais, sensores, diversos *logs* de atividades, sistemas transacionais, sistemas de GPS. 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 imperceptível 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 de informações por ano para apoiar os pesquisadores em seus estudos (“Computing | CERN”, 2015). De acordo com um estudo (GANTZ; REINSEL, 2011) realizado pelo IDC, no ano de 2010 foram gerados mais de 1 Zettabyte de informações no mundo todo. (MCAFEE; BRYNJOLFSSON et al., 2012) afirmam que no ano de 2012 foram gerados 2.5 Exabytes de dados a cada dia, essa massiva quantidade de informações tem se tornado um grande problema computacional, pois não é possível gerenciá-las com eficiência empregando as ferramentas e modelos comuns de banco de dados (SGBD).

Com essa grande quantidade de dados, estamos inseridos em um mundo onde a informação é cada vez mais valorizada. 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”. Por isso, tem-se a necessidade de minerar e aproveitar cada dado gerado pelas diferentes fontes e interações com usuários de diversos tipos e com perfis heterogêneos.

Partindo então dessas motivações, temos o conceito de *Big Data,* que são soluções desenvolvidas para poder lidar com essas quantidades massivas de dados, esse amplo volume é uma característica muito encontrada nessas soluções, porém, em qual tamanho pode-se considerar um grande volume de dados?

Outro fator que tem sido um grande problema é como podemos visualizar essas informações, pois na maioria das vezes os dados são de diferentes tipos e em alguns casos é mais interessante até converter em um tipo único para que facilite a posterior interpretação, portanto o desafio é transformar essa miríade de dados em informação.

Nesse artigo abordaremos e utilizaremos o arcabouço *Apache Hadoop*, porém existem diversos outros com diferentes características, todavia, atualmente o *Apache Hadoop* é considerado o melhor framework para esse fim (GOLDMAN et al., 2012)

## 1.1 PROBLEMATIZAÇÃO

Aqui estará minha problematização.

1.2 OBJETIVOS

1.2.1 OBJETIVO GERAL

O objetivo desse trabalho é.

1.2.2 OBJETIVOS ESPECÍFICOS

* Objetivo específico 1;
* Objetivo específico 2;
* Objetivo específico 3.

1.3 JUSTIFICATIVA

As conclusões obtidas ao final desse trabalho refletem importantes resultados para a UTFPR – Cornélio Procópio, uma vez que, os métodos atuais de gestão de informações da carreira acadêmica, tais como o portal do aluno[[1]](#footnote-1) podem ser otimizados para entregar dados que possam colaborar de forma direta na tomada de decisões dos discentes. Com as conclusões desse trabalho têm-se a proposta de utilizar os procedimentos de BI para contribuir diretamente com diminuição de uma das causas da evasão de estudantes de cursos superiores em universidades públicas, que é a falta do planejamento do percurso acadêmico.

1.5 ORGANIZAÇÃO

A estrutura desse texto está organizada da seguinte forma: capítulo 1 descreve a introdução, a problematização, o objetivo geral desse trabalho, os objetivos específicos, a justificativa, os procedimentos metodológicos.

O capítulo 2 é composto da fundamentação teórica e os estudos relacionados ao tema abordado nesse trabalho.

No capítulo 3 estão descritos os procedimentos metodológicos (ferramentas e métodos) que deverão ser realizados para a conclusão do objetivo geral.

# 2 FUNDAMENTAÇÃO TEÓRICA

2.1 *BIG DATA*

O termo Big Data não é novo, porém nos últimos anos muito tem se falado sobre ele (Tiago Cordeiro), entretanto, ainda não há uma clara definição, pois cada autor cita uma característica diferente e o define de acordo com sua perspectiva, todavia, embora não existam regras e padrões definidos, uma característica comumente citada pelos autores é que para ser uma solução de *Big Data* ela deve atender cinco atributos denominados como “5Vs”, que são volume, variedade, velocidade, veracidade e valor. Grande parte desses dados são semiestruturados ou não estruturados (NANDIMATH et al., 2013), o que torna difícil à interpretação e a geração de conhecimento sobre eles, embora os dados possuam muito valor agregado, é necessária uma correta extração, mineração e visualização das informações que eles podem disponibilizar.

A definição de (MANYIKA et al., 2011) diz que soluções de Big Data são um grande conjunto de dados que as tecnologias comuns de bancos de dados (SGBDs) não conseguem gerenciar dada a complexidade e tamanho desses aglomerados. Outro fator importante é que um dos atributos para considerar uma solução como de *Big Data* é a velocidade, portanto isso reforça a ideia de que para obter dados em meio a uma quantidade massiva de informações, é necessária uma solução não convencional (GOLDMAN et al., 2012; NANDIMATH et al., 2013). Dessa forma podemos usar arcabouços de processamento e armazenamento paralelo.

2.2 APACHE HADOOP

O Apache Hadoop é um robusto *framework,* gerenciado pela empresa *Apache Software Foundation*, é inspirado no projeto GFS e no MapReduce do Google, desenvolvido sobre a linguagem Java e que tem como objetivo o armazenamento e o processamento de grandes conjuntos de dados, ele utiliza um modelo de computação distribuída e de alto desempenho, onde vários nós (*clusters*) 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. Por ser um único arcabouço composto por diversos subprojetos, ele pretende 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.

O arcabouço 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 que utilizam suas soluções. O Apache Hadoop é composto por vários subprojetos, porém, em seu núcleo temos três principais: o *Hadoop* *Common*, o *Hadoop MapReduce* e o *Hadoop Distributed File System* (HDFS).

2.2.1 HADOOP COMMON

O Hadoop Common é um módulo do arcabouço Hadoop que serve como base para outros subprojetos, ele possui os utilitários necessários para que eles funcionem.

2.2.2 HADOOP MAPREDUCE

O Hadoop MapReduce pode ser considerado como o centro do arcabouço, é a parte onde são processados os dados. Ele abstrai as complexidades de um processamento distribuído, simplificando o processo, como o próprio nome já diz, em apenas duas etapas o “*Map*” e o “*Reduce*”, onde na primeira, os dados serão mapeados sendo que um aglomerado de dados será convertido em outro aglomerado de dados porém agora os elementos dividem-se em tuplas compostas por chaves e valores. O *Reduce* é alimentado pelas saídas do *Map* sendo que ele coleta as tuplas já mapeadas, as combinam e as reduz a um conjunto menor de tuplas, portanto sempre deve ser seguido essa ordem primeiro o mapeamento depois a redução dos dados.

2.2.3 HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

O *Hadoop Distributed File System* é um sistema de arquivos distribuído, onde cada nó é uma parte desse sistema, ele abstrai algumas complexidades de uma clusterização, ele é capaz de deixar invisível ao programador esses múltiplos nós, o próprio HDFS se encarrega de fazer a divisão dos dados em partes menores (blocos) e realiza a distribuição entre os nós. Com blocos de dados menores o MapReduce pode se encarregar de fazer todo o processamento necessário, tornando-o possível mesmo com uma quantidade inicial de dados grande.

2.3 D3.JS

Definição do D3.js

2.4 TRABALHOS RELACIONADOS

3 PROCEDIMENTOS METODOLÓGICOS

3.1 FERRAMENTAS

Será implementado uma versão do *framework* de *Business* *Intelligence* Pentaho, na versão *Comunity* *Edition* (CE) 5.2.0 que está sob a licença *open source* (PPL).

3.2 MÉTODOS

Para que seja alcançado o objetivo geral será necessário:

* Estudar a estrutura do banco de dados do portal do aluno
* Criar um processo de ETL, que consiste e extrair dados de uma fonte, transformar esses dados, tratando para que as informações fiquem como desejadas e carregar as mesmas em uma nova base de dados
* Criar e popular um *Data* *Warehouse*[[2]](#footnote-2)
* Desenvolver relatórios
* Desenvolver painéis indicadores de dados (*dashboards*)
* Desenvolver cubos, que cruzam diferentes informações exibindo em uma só tabela ou gráfico.

**REFERÊNCIAS**

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

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.

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

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

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

1. Estrutura que exibe o histórico escolar do aluno com informações referentes a notas, quantidade de curso concluídos e pendentes. [↑](#footnote-ref-1)
2. Armazém de dados que possibilitam que sejam depositados e sumarizados grandes volumes de informações não voláteis referentes às atividades de uma organização e que facilite posteriormente a análise desses dados por ferramentas como as de BI. [↑](#footnote-ref-2)