ATLANTE
Cuadernos de Educación y Desarrollo (2ª Época)

Marzo 2019 - ISSN: 1989-4155

ANÁLISE DE BASE DE DADOS HISTÓRICA DO PROUNI, USANDO MINERAÇÃO DE DADOS

Dionatas Pereira Silva;

Graduando Sistemas de Informação (Centro Universitário Padre Anchieta - UNIANCHIETA), Jundiaí/SP. dionatas3@gmail.com

Juliano Schimiguel;

Universidade Cruzeiro do Sul (São Paulo/SP) e Centro Universitário Padre Anchieta – UNIANCHIETA (Jundiaí/SP). schimiguel@gmail.com

Para citar este artículo puede utilizar el siguiente formato:

Dionatas Pereira Silva y Juliano Schimiguel (2019): "Análise de base de dados histórica do Prouni, usando mineração de dados", Revista Atlante: Cuadernos de Educación y Desarrollo (marzo 2019). En línea:

https://www.eumed.net/rev/atlante/2019/03/base-datos-historica.html

RESUMO

Utilizar *big data* para criar soluções inteligentes é um grande desafio e tem exigido mais interatividade entre as diversas áreas do conhecimento, processar grandes volumes de dados é uma parte desse desafio, com a diversidade de dados gerados de diferentes fontes, é fundamental o uso de tecnologias big data. O objetivo do artigo é descrever *big data*, as tecnologias que processam e armazenam essa gigantesca massa de dados, descrever *data mining*, ou mineração de dados e tratar dados históricos utilizando as bibliotecas da linguagem de programação *Python: Pandas e matplotlib.* A metodologia utilizada para elaboração do artigo foi um estudo bibliográfico sobre *big data* e *data mining* e também um caso de uso que comparou os dados históricos das bolsas concedidas na região do CIMBAJU **(**Consórcio Intermunicipal dos Municípios da Bacia do Juqueri) e do Brasil. A conclusão deste trabalho é a apresentação do universo *big data* e suas tecnologias, mostrando teoricamente o que é *big data* e também *data mining* além de uma descrição do ProUni e a comparação dos dados históricos que foram analisados.

Palavras-Chave: Big data, mineração de dados, Python, Pandas, matplotlib, ProUni, data mining, dados.

ABSTRACT

Using big data to create smart solutions is a major challenge and has demanded more interactivity among the various areas of knowledge, processing large volumes of data is a part of this challenge,

with the diversity of data generated from different sources, is Fundamental use of big data technologies. The purpose of the article is to describe big data, the technologies that process and store this gigantic mass of data, describe data mining, and handle historical data using the libraries of the Python programming language: Pandas and Matplotlib. The methodology used to elaborate the article was a study on big data and data mining and also a case of use that compared the historical data of the fellows contemed in the region of Cimbaju (Intermunicipal consortium of Municipalities of the Juqueri basin) and of Brazil. The conclusion of this work is the presentation of the Big Data universe and its technologies, theoretically showing what is big data and also data mining in addition to a description of the ProUni and the comparison of the historical data that were analyzed.
Keywords: Big data, data mining, Python, Pandas, Matplotlib, ProUni, data

1. INTRODUÇÃO

A grande volumetria de dados gerados diariamente reflete a situação desafiadora que é extrair conhecimento de bases históricas que podem ser potencialmente geradoras de *insights* significativos para tomadas de decisões bem-sucedidas.

Hoje no universo *Big data* já é normal se falar em tabelas de milhões de registros e processamentos *near-real-time* onde processos *ETL* disponibilizam esses dados para serem consumidos.

A era *big data* como é conhecido esse novo momento da tecnologia da informação, onde o crescimento exponencial dos dados inviabiliza uma simples análise colunar acaba gerando novas situações e problemáticas que se relacionam com questões que vão desde como armazenar essa massa gigantesca de dados a como utilizar esses dados de maneira eficiente.

Esse número gigantesco constantemente passa a ser maior por conta da cultura que tem sido enraizada onde quase nada é descartado e quase tudo é mantido, sejam fotos, comentários em redes sociais ou e-mails que já não são mais deletados pelo fato de os provedores não terem mais um limite de armazenamento tão baixo como há alguns anos atrás. Para se ter uma ideia de como a produção massiva de dados está sendo gerada.

"O volume de informações geradas pela sociedade é assustador. Por exemplo, o Twitter sozinho gera diariamente doze terabytes de tuites. Os medidores inteligentes de energia, que começam a substituir os seus ancestrais analógicos, geram 350 bilhões de medições por ano." (TAURION, 2013, p.4)

A facilidade de conexão à internet, o baixo custo de armazenamento de dados em plataformas *cloud computing*, assim como dados de internet das coisas que relaciona dados a aparelhos domésticos, carros e outros acessórios, tudo isso contribui para explicar a geração desta massiva fonte de dados a qual praticamente todas as pessoas que se conectam à internet fazem parte. Muitas são as fontes de dados hoje em dia, a variedade desses dados é muito ampla e consequentemente a complexidade de utilizar esses dados se torna um desafio que muitos profissionais da área da tecnologia da informação tentam tratar diariamente.

Há aproximadamente vinte ou trinta anos atrás a maior parte dos dados gerados eram armazenados em formato analógico, como fitas de vídeo cassete, disquetes, entretanto recursos como esses que ainda existem passaram a ser substituídos por tecnologias digitais. Não precisar de

um acessório físico que limita a quantidade de dados a ser armazenados e até mesmo a possibilidade de que esses dados possam ser perdidos são boas razões para que a migração de dados tenha caminhado para o formato digital, para armazenamentos virtuais, onde o custo por espaço é muito menor e a possibilidade de perda é bastante reduzida.

Hoje a internet também é fator determinante para a expansão desta grande massa de dados "*O poder de armazenamento, os recursos computacionais e o acesso à internet oferecidos por esses dispositivos ampliaram não somente a quantidade de dados únicos gerados, mas também a quantidade de vezes que eles eram compartilhados*. " (MARQUESONE, 2016, p.23)

Obter informação com qualidade e no tempo certo, assim como extrair conhecimento de base dados nos dias atuais tornou-se a linha tênue entre a vida e a morte de empresas que querer permanecer competitivas, de governos ou órgãos públicos que precisam ser eficientes, organizações que necessitam de informações em tempo real.

Devido a toda essa massa de dados que é gerada diariamente, existe algum tipo de ferramenta que trata esses dados? Existem maneiras ou técnicas de tornar toda essa informação palpável a ponto de alguém apenas olhar para uma tela e conseguir entender o que tudo aquilo significa? Existe algum tipo de padrão nessa massa de dados? Todas essas perguntas podem ser resumidas em uma frase: Como visualizar, entender padrões, extrair conhecimento e utilizar dados que crescem exponencialmente em uma velocidade tão rápida?

Assim como as ferramentas de *big data* que podem contribuir para extração de valor de grande massa de dados, a mineração de dados, através de técnicas e visualizações gráficas pode ajudar na criação de conhecimento, ao se fazer analises olhando apenas para tabelas, dependendo da quantidade de dados disponíveis mesmo que depois de um certo tratamento e classificação não é tão eficiente e razoável, algo que já justifica o tratamento dos dados de forma visual, os relatórios visuais, *dashboards* contribuem para uma melhor tomada de decisão tornando-a mais bem amparada.

O estudo de *big data* e as tecnologias para o processamento de dados se tornaram extremamente importantes, inclusive para outras áreas do conhecimento, pois o mundo atual que é pautado na informação e no mundo digital necessariamente converge para o conhecimento da tecnologia da informação. Diversos artigos vêm sendo publicadas demonstrando casos de uso de *big data,* como os artigos da consultoria *Gardner*, Hekima, *SAS*, *Intel*, muitos gigantes da tecnologia como *Google*, *Facebook* fazem uso de ferramentas *big data,* assim como a gigante *Netflix*. "*Análise de big data é uma estratégia baseada em tecnologia que permite a coleta de insights mais profundos e relevantes dos clientes, parceiros e sobre o negócio — ganhando assim uma vantagem competitiva*. " (INTEL, 2017, p.4)

O objetivo geral deste trabalho é fazer uma análise de dados utilizando bases históricas do Programa universidade para todos (ProUni) onde será feita uma comparação quantitativa para criar *insights* referentes a região do consórcio intermunicipal dos municípios da bacia do Juqueri(CIMBAJU) em relação ao Estado Brasileiro, explicar o que é o ProUni e trazer uma breve introdução do porquê foi importante o programa ter sido criado, além de tratar dos assuntos *big data* e *data mining*, assim como suas técnicas e tecnologias.

Foi utilizada como metodologia no trabalho um estudo teórico sobre *big data*, *data mining* utilizando a bibliografia de referência e também foi feito um caso de uso com bases históricas do ProUni, que compreendem os anos no intervalo de 2005 até 2016 e que estão disponíveis em formato *.CSV*, usando as bibliotecas *Pandas e Matplotlib* do *Python* 3 para gerar gráficos e *data frames* em uma granularidade federativa e regional, assim como gerar *insights* dessas bases com relação ao CIMBAJU, Consórcio Intermunicipal dos Municípios da Bacia do Juqueri , que tem como integrantes: Franco da Rocha, Caieiras, Cajamar, Francisco Morato e Mairiporã. Para trabalhar com essas ferramentas foi utilizada a última versão do Júpiter que é uma aplicação *IDE* para programação em *Python*.

O artigo tem sua estrutura formada pelo resumo geral, seguido da introdução dos temas e a contextualização, partindo para uma explicação mais específica de *big data* e suas tecnologias, *data mining* e suas técnicas e finaliza a parte teórica tratando do programa universidade para todos, o ProUni, com uma breve análise da educação no ensino superior no Brasil. O artigo tem uma continuação no caso de uso onde são feitas as análises dos dados históricos do ProUni e onde compara os dados entre a região do CIMBAJU e o Estado Brasileiro. O artigo é finalizado nas considerações finais.

2. BIG DATA

O valor dos dados tem muito a ver com o seu tratamento e com a velocidade a qual se cria uma informação e um conhecimento a partir do mesmo, da mesma maneira que produtos de setores primários, ingredientes para criação de alimentos e peças na indústria para criação de maquinário, os dados tem seu valor e necessitam de um tratamento para se tornar informação útil, sendo muito importante o tempo gasto para gerar tão informação.

Big data não é dar um chute e tentar a sorte para saber o que irá acontecer, pois muitos dados são gerados massivamente a cada dia e analisar padrões dessa massa de dados pode gerar informações e posteriormente conhecimento para tomada de decisões mais acertadas em diversos setores, tanto público quanto privado. As empresas vêm analisando como podem usufruir de tecnologias *big data*, tentando compreender os conceitos e as tecnologias, aos poucos as organizações tanto privadas quanto públicas passam a compreender o ganho que se tem ao tratar os dados de uma maneira integrada com conceitos e tecnologias *big data*.

O conceito de *big data* contribui com a integração dos dados, permitindo empresas e até mesmo o setor público tomar decisões acertadas e não apenas intuitivas. Tratar dados e cruzá-los pode trazer muita transparência para o setor público, onde este muitas vezes tem seus departamentos trabalhando de maneira isolada, ao tratar esses dados com *big data*, muitas soluções podem surgir, assim como já tem acontecido em cidades como *New York* que aderiu ao uso de *big data* para trazer soluções inteligentes para o departamento de segurança e Singapura para gerar solução inteligentes para o seu departamento de trânsito.

Para usufruir de todo esse conceito de *big data*, é necessária uma grande capacidade computacional, uma vez que *big data* tende a fazer análise de uma grande quantidade de dados, o

ambiente na nuvem é extremamente recomendado, pois pode-se escalar processamento e memória, o uso de tecnologias como *Haddop* e *Spark* são realidades no universo de *big data*.

Big data pode trazer valor para empresas tratando dados históricos alocados em *data warehouses*, assim como fazer predições através de algoritmos e analisar padrões nos dados, analise de tuites, comentários de redes sociais como *Facebook* ou até mesmo vídeos do *YouTube*.

Os dados são provenientes das mais diversas fontes, tanto de sistemas próprios usados internamente nas empresas, como dados gerados pela internet das coisas, sensores, mídias sociais, *big data* de fato é baseado nisso, no volume de dados gerados, nesta variedade, pois as fontes desses dados são as mais variadas, a velocidade a qual esses dados são gerados também é importante salientar, pois dados como os de mídias sociais são muitas vezes gerados em tempo real, assim como imagens de vídeos de lugares públicos, portando *big data* está arquitetado na variedade, velocidade e volume desses dados.

Para se utilizar *big data*, não necessariamente precisa-se estar trabalhando com dados nas casas do *petabytes*, pois apesar de o volume ser um dos atributos que compõem a ideia de *big data*, existe também a questão da relação da variedade desses dados, pois empresas que ainda estão no patamar dos *gigabytes* ainda assim podem necessitar de soluções *big data*, uma vez que esses dados podem não ser estruturados, onde os tratamentos através de tecnologias *big data* são muito eficazes.

Figura 2.1 Estrutura big data

Fonte: (MARQUESONE, 2016)

Para muitos autores a definição de *big data* se dá pelo que é chamado “3v” e em alguns casos “5v”, tal definição é uma abreviação para as palavras que se iniciam com a letra “v”: velocidade, variedade, volume e também os dois últimos vês: veracidade e valor.

> *É comum, ao ouvir pela primeira vez o termo Big Data, pensarmos que ele está unicamente relacionado a um grande volume de dados (o que é normal, já que o nome diz exatamente isso). Entretanto, o volume de dados não é sua única característica. Além dessa, pelo menos outras duas propriedades devem ser consideradas: a variedade e a velocidade dos dados. Tais propriedades são popularmente denominadas os 3 Vs de Big Data.* (MARQUESONE, 2016, p.27)

Compreender a definição de *big data* é importante para se entender o porquê é necessário utilizar tecnologias de *big data* e porque não apenas continuar com as tecnologias que podem tratar as demandas de hoje. A medida que se vai entendendo o que significa *big data*, passa-se a aceitar

com mais credibilidade que sim, os dados estão exponencialmente sendo gerados, estes também estão represados sem tratamento dentro das companhias e até mesmo dentro dos órgãos públicos.

Tratar esses dados e saber filtrá-los pode levar a um outro ponto da definição de *big data* que é a inclusão da veracidade e valor dos dados, onde os dados necessariamente têm que ser validados para serem utilizados e ter valor. *Big data* tem tudo a ver com o volume de dados gerados e como esses dados podem ser tratados através de suas tecnologias de enriquecimento e criação de informação, quando se fala de volume de dados, algo que mudou completamente ao passar dos anos, pode-se destacar a velocidade que esse volume de dados está sendo gerado e por fim a variedade desse volume de dados que é criado tão rapidamente.

Tratando se de velocidade de geração dos dados. "*A velocidade diz respeito não somente a da produção do dado em si, mas a velocidade do processamento e produção de informação e conhecimento*". (AMARAL, 2016, p.10)

Além da velocidade, destaca-se o volume, a criação de dados que é tida como exponencial pelos autores e realmente é, pois, comparando-se a geração de dados historicamente, nota-se que algumas condicionais foram importantes para tal criação deste volume. "*O atributo volume é a característica mais significativa no conceito de Big Data. Ele faz referência à dimensão sem precedentes do volume de dados.* " (MARQUESONE, 2016, p.30). Para se ter uma ideia 90% dos dados existentes foram criados nos últimos dois anos segundo pesquisas de consultorias de TI.

Analisando uma estatística da representatividade desses dados, que em sua grande maioria são dados não estruturados, neste caso, tal ponto será tratado quando for abordada a variedade de dados que é um outro componente da definição de *big data*, tratando-se da origem desses dados que são provenientes das mais diversas fontes, percebe-se que as redes sociais são importantes geratrizes dessa massa de dados.

"A rede social Facebook contabilizou em junho de 2016 uma média de 1.13 bilhão de usuários, 2.5 bilhões de compartilhamentos e 2.7 bilhões de "curtidas" diariamente. A rede social de compartilhamento de fotos Instagram recebe atualmente cerca de 80 milhões de fotos por dia." (MARQUESONE, 2016, p.30)

A variedade dos dados pode ser verificada analisando tal estatística, pois o que se vê é que dada a dimensão tempo em segundos, dias, o volume de dados gerados pelas redes sociais, principalmente, pode ser dos mais diversos tipos como fotos, vídeos, textos.

Os dados podem ser estruturados ou não, hoje um grande desafio em *big data* é justamente tratar os dados não estruturados. Os dados semiestruturados são os que possuem uma estrutura prévia, como um arquivo *Json*, já os dados não estruturados são os vídeos, imagens, áudio e alguns formatos de texto.

Esses dados não estruturados ou até mesmo os semiestruturados se tornam complexos para se processar com ferramentas tradicionais de gerenciamento de banco de dados, pois o banco de dados relacional que é o modelo mais utilizado necessita de uma estrutura definida previamente. Para melhor entendimento, para que um dado vá para um banco de dados relacional, necessariamente precisa-se definir o *schema*, objeto e o tipo de dado que fará a composição deste objeto.

Dados semiestruturados ou não estruturados não são compatíveis com o modelo relacional, trazendo assim desafios para que estes sejam tratados através de tecnologias *big data*, tais dados necessitam de um modelo flexível quando se trata de sua estrutura, tecnologias que não sejam rígidas como a relacional que exige um prévio tratamento e que se adequem para a volumetria que é gerada diariamente, impossibilitando o uso de um banco de dados tradicional.

Outro ponto focal de *big data* que se pode destacar é a velocidade, hoje os dados são gerados pelas mais diversas fontes e também em uma escala gigantesca. A velocidade se relaciona com sua criação e também o seu tratamento. "*Estatísticas mencionam que, em apenas 1 minuto, mais de 2 milhões de pesquisas são realizadas no buscador Google, 6 milhões de páginas são visitadas no Facebook e 1.3 milhão de vídeos são vistos no YouTube*" (MARQUESONE, 2016, p.43)

O tratamento que se faz em um dado de uma maneira rápida e eficaz, pode trazer um outro ponto a ser analisado na definição de *big data*: o valor. O dado tratado pode ser de grande valia para uma organização se estiver disponível no momento certo. Uma tomada de decisão embasada em uma informação concreta é muito mais valiosa que uma decisão tomada apenas por intuição. "*Dados são os recursos naturais da sociedade da informação, como o petróleo para a sociedade industrial. Tem valor apenas se tratados, analisados e usados para tomada de decisões.* " (TAURION, 2013, p.54).

Devida a diversidade das fontes geradoras de dados, um ponto a ser também considerado e que é colocado em evidência por alguns autores é a veracidade. Dados vindo de todas as partes e a todo momento, uma vez consumidos e utilizados, caso não sejam adequadamente tratados e validados, podem ser um grande problema para quem os utiliza.

"*O atributo veracidade está relacionado à confiabilidade dos dados. Pelo fato de que Big Data está inserido em um contexto de dados em grande volume e variedade, é comum a existência de dados inconsistentes. "* (MARQUESONE, 2016, p.46)

Os dados que são estruturados, semiestruturados e não estruturados são criados tanto pela interferência humana ou não, pois podem ser criados de forma automatizada. Pessoas geram dados massivamente interagindo em redes, dados gerados por processos não necessariamente precisam de intervenção humana, são dados produzidos por processos de computadores, internet das coisas, são processos automáticos de coleta de dados.

Big data não significa apenas volume como foi visto, além disso outro ponto que pode trazer alguma confusão é que ao se falar de *big data*, não se fala apenas de dados não estruturados, apesar de estes serem a maioria dos dados gerados.

Gerar valor a partir dos dados é a principal tarefa de *big data*, por este motivo, projetos que necessitem de soluções e ainda assim não utilizam um conjunto de dados na casa dos *petabytes* podem ser beneficiados com soluções de *big data*.

3. TECNOLOGIA POR TRÁS DO BIG DATA

Tratando-se de projetos *big data* que normalmente tratam ou trabalham com ingestão de uma grande volumetria de dados, novamente, não necessariamente *petabytes*, pois *gigabytes* de texto já é muito dado.

> *Projetos de Big Data carregam grandes volumes de dados em um sistema de arquivo como o HDFS, em seu formato nativo, mesmo que em princípio não se enxergue valor. Estes dados formam os conhecidos data lakes, ou lagos de dados. Posteriormente, parte destes dados pode ser transformados e carregados em um data mart tradicional. Outra forma que podemos olhar uma solução de Big Data é sob sua arquitetura básica. Neste contexto, temos quatro elementos: fontes de dados, carga, armazenamento, análise e visualização (ou apresentação).* (AMARAL, 2016, p.15)

Trabalhar com projeto de *big data* é explorar a imensa massa de dados que está dentro ou fora das empresas, órgãos governamentais, quando se trabalha com *big data* a primeira etapa é a de coleta dos dados, sejam eles dados internos ou externos. Esses dados podem estar disponíveis internamente em banco de dados transacionais que as empresas usam, como pode ser também dados externos de câmeras de segurança, mídias sociais, sensores, radares.

Uma vez que os dados são coletados, estes necessitam receber os tratamentos devidos e serem integrados à necessidade do projeto. Uma vez que o dado é tratado e armazenado e estando à disposição, uma fase analítica pode ser implementada, onde se fará a tradução dos resultados obtidos ao se utilizar esses dados que foram validados, visualização dos dados por meio de *dashboards* fazem parte desta parte analítica e também de apresentação dos resultados. A visualização de dados também teve uma significativa evolução, pois até um tempo atrás, tomadores de decisões se baseavam muito mais em relatórios simples, como total de vendas, quantidade de clientes, gastos, lucros, dados mais brutos e menos granulares. "*Os relatórios evoluíram para se tornarem sofisticados sistemas de informações gerenciais (SIGs), incluindo a geração de diferentes tipos de gráficos e painéis com diferentes informações (dashboards).* " (LOH, 2014, p.7)

Apesar de muitas empresas e até mesmo os órgãos públicos terem muitos dados internos disponíveis para colaborar com inteligentes tomadas de decisões, muitas vezes o sucesso de se utilizar tais dados esbarra justamente no fato desses dados não estarem integrados, pois estão em diferentes bases, demora para analise, imagine que uma informação seria importante antes de ocorrer um problema, tratando um dado e tendo esta informação, possivelmente este problema seria evitado, entretanto havendo demora na análise e no tratamento deste dado não faz mais sentido tê-lo e este perde valor.

Os dados não serem apresentados com clareza pode também afetar a utilização e um tomador de decisão deixar de utilizá-los, pois não foi claro ou completo o suficiente, desta maneira acabam não utilizando esses dados para obter valor.

Tratando-se do setor público que parece funcionar de maneira arcaica em muitas das cidades do Brasil, onde os sistemas não são interligados e há pouca ação do ponto de vista do uso da informação, alguns serviços prestados deixam a desejar, em 2012 a lei de acesso à informação, lei nº12.527/2011, a qual regulamenta o direito de acesso às informações públicas entrou em vigor, trazendo a possibilidade da sociedade de modo geral, seja pessoa física ou jurídica ter acesso aos dados públicos.

A análise dos dados públicos poderia ser feita e interligada entre os diversos órgãos trazendo mais garantia e acerto nas decisões dos administradores da máquina pública, evitando o desperdício

desnecessário de recursos, agilidade na disponibilização das informações a serem passadas ao público.

Um grande acerto do setor jurídico é a maior parte das comarcas estarem migrando para o mundo digital, onde os advogados podem despachar e ver processos pela internet, o público ter acesso aos processos que não são segredo de justiça, uma vez que anteriormente seria necessário ir ao fórum buscar tal informação e o controle das operações ser automatizado, trazendo segurança e produtividade.

"O potencial do Big Data na administração pública só será alcançado com mudanças organizacionais, culturais e de processos. Os governos passam a ter a possibilidade de tomar decisões baseados em fatos e operar com muito mais eficiência. " (TAURION, 2013, p.103)

Big data demanda uma alta disponibilidade, ou seja, uma vez que uma tecnologia está em uso, esta não pode estar indisponível ou pelo menos ter menos riscos desta indisponibilidade. A virtualização, utilização de máquinas virtuais, *clusters* e recursos de computação na nuvem diminui as chances de um serviço ter indisponibilidade por falta de recursos.

Utilizar computação na nuvem possibilita que projetos diferentemente de infraestruturas *on-primese*, tenham a condição de escalar os recursos horizontalmente, uma vez que se tem um *cluster* que é a composição de diversos nós de computação e necessariamente ser preciso mais memória e processamento, basta escalar, ou seja, requisitar ao fornecedor mais memória e mais processamento e pagar por isso apenas enquanto utiliza.

A disponibilidade de um serviço acarreta diretamente nos custos de um projeto, pois se os serviços se tornam constantemente indisponíveis, outros recursos passam a ser negativamente afetados, como por exemplo os recursos: pessoal, tempo, fornecedores.

" Uma disponibilidade alta, por exemplo, de 90%, significando que o sistema poderá estar indisponível cerca de 36 dias por ano é intolerável até para a aplicação menos crítica. Já uma disponibilidade de 99,999% representa uma indisponibilidade de 5,26 minutos por ano. " (AMARAL, 2016, p.46)

Das tecnologias envolvidas em *big data* o projeto *Hadoop* é extremamente importante, tendo como inspiração os trabalhos do Google, idealizado por Doug Cutting e Mike Cafarella, o *Hadoop* foi lançado pelo *Yahoo* em 2008 como projeto *open source*. Hoje o ecossistema *Hadoop* que é a composição de outros projetos juntamente com o *Hadoop* é mantido pela *Apache*, organização sem fins lucrativos.

O projeto *Hadoop* foi inspirado nos projetos do Google, *Google File System*, *GFS* e também no paradigma de programação *MapReduce*. O *Hadoop* é composto por dois projetos, *Hadoop MapReduce*(*HMR*) que foi inspirado no *MapReduce*, *software* utilizado para acelerar as pesquisas no Google. Além do *HMR* o *Hadoop* é composto pelo *Hadoop Distributed File System* (*HDFS*) que é o sistema de arquivos distribuídos e otimizados para trabalhar com dados não estruturados no *Hadoop*, este no caso teve sua origem inspirada no projeto *GFS*.

O *Hadoop* tem uma grande importância pela sua capacidade de armazenar e processar grandes massas de dados, sejam dados estruturados ou não, uma vez que os dados são gerados exponencialmente, o ambiente *Hadoop* é ator principal neste cenário. O modelo do *Hadoop* computacional que possibilita escalabilidade faz com que o processamento de grandes volumes de

dados seja feito com grande êxito. *Hadoop* é extremamente eficiente pois é tolerante a falhas, quando um dos nós cai, outros nós passam a executar o trabalho do nó indisponível, fazendo com que o processo continue.

No *Hadoop* os dados são armazenados no *HDFS*, para processar dados no *Hadoop* necessariamente os dados precisam estar lá, os dados são divididos em pequenos blocos e pesquisados paralelamente, ou seja, ao mesmo tempo, tornando o processo extremamente rápido.

> *O outro componente do Hadoop é o MapReduce. É o coração do Hadoop. É o paradigma de programação que possibilita escalabilidade massivamente paralela em centenas ou milhares de servidores. O próprio termo MapReduce representa duas tarefas distintas que os programas Hadoop executam. A primeira tarefa é mapear os dados, ou seja, acessar um conjunto de dados e convertê-los em outro conjunto onde os elementos individuais são quebrados em tuplas (pares chave/valor).* (TAURION, 2013, p.183)

Diversas empresas que trabalham com *cloud computing* disponibilizam o sistema *Hadoop* e seu ecossistema para utilização, essas empresas disponibilizam o sistema através de suas estruturas *clouding* e cobram o uso por hora, desta forma quem está utilizando paga uma ferramenta, recurso apenas o tempo que se utiliza não necessitando criar toda uma infraestrutura para tratar um problema pontual.

O *Hadoop* tem um ecossistema que são outros projetos integrados e relacionados, onde ao se utilizar o sistema *Hadoop* pode-se utilizar também estes projetos. Dentre os projetos do ecossistema *Hadoop* pode ser destacado o *Hbase*, banco de dados *NOSQL*, que utiliza o *HDFS*. O *Hive* também faz parte do ecossistema, onde pode-se criar tabelas e processar queries como se fosse uma ferramenta *SQL* tradicional. Outro projeto que faz parte é o *PIG*, uma linguagem de alto nível. Esses são alguns dos exemplos que fazem parte do Ecossistema *Hadoop*, veja na figura uma tabela de modo geral:

Camada funcional do Hadoop	Subprojetos
Modelagem e desenvolvimento	MapReduce, Pig, Mahout
Armazenamento e gestão de dados	HDFS, Hbase, Cassandra
Data Warehousing e queries	Hive, Sqoop
Coleta, agregação e análise de dados	Chukwa, Flume
Metadados, tabela e esquemas	HCatalog
Cluster management, job scheduling e workflow	Zookeeper, Oozie, Ambari
Serialização de dados	Avro

Figura 3.1 Ecossistema Hadoop

Fonte: (MARQUESONE, 2016)

Tratando-se ainda de tecnologias por trás do *big data*, não se pode excluir o *Apache Spark*. O *Spark* é uma ferramenta de código aberto para o tratamento de dados em projetos *big data*, utiliza o *cache* na memória para armazenamento e execução para ter uma melhor performance. Tratando-se de velocidade de processamento o *Spark* pode ser muito mais rápido que o *Hadoop*, além de mais veloz que o *Hadoop*, o *Spark* vem com diversas bibliotecas como suporte para consultas *SQL*, fluxo

de dados, *Machine Learning* e processamento de gráficos. Muitas empresas gigantes utilizam *Spark* para processamento de dados e hoje este é o maior projeto de código aberto para processar dados.

" Spark, irmão do Hadoop, é uma excelente plataforma de processamento em memória para projetos de tempo real ou próximo ao tempo real. Hoje, um servidor com 6 TB de memória pode suportar grandes volumes de dados. " (AMARAL, 2016, p.43)

O *Spark* deve ser tratado como uma alternativa em relação ao *Hadoop* e não como um simples substituto, diferentemente do *Hadoop* que grava os processamentos em disco, o *Spark* os detém na memória *ram*, entretanto pode-se utilizar o disco uma vez que o *Spark* executa processamentos em disco caso não haja mais disponibilidade de memória.

Assim como o *Hadoop* o *Spark* utiliza o *HDFS* para armazenamento dos dados processados. Outro pronto positivo ao se utilizar *Spark* é que este *framework* suporta linguagens de programação como: *Python*, *Scala*, *Java*, *Clojure*, R. Essas linguagens de programação são utilizadas para tratamento de dados e facilita muito o processamento em *Clusters* que utilizam *Spark* para tratar grandes massas de dados.

Tanto o *Hadoop* quanto o *Spark*, se não tivessem seus ecossistemas, muito provavelmente não seriam utilizados em um curto intervalo de tempo, uma vez que ambos os *frameworks* foram criados a pouco tempo, no caso do *Spark* o código foi aberto em 2010, ou seja, recentemente. Um grande acerto tanto do *Haddop* quanto do *Spark* é aproximar os desenvolvedores que já conhecem linguagens de tratamento e consulta de dados como *SQL* que através do *Hive* por exemplo no *Hadoop* conseguem fazer consultas como se estivessem usando o *SQL* através do *HQL*, falando de *Spark* que também tem seu ecossistema, mas totalmente integrado, performático e que possibilita o uso direto de linguagens de programação convencionais, onde além de por exemplo estar usando *Python*, pode usar *data frames* importando bibliotecas do *Spark SQL* e utilizar *queries* tradicionais sem a necessidade de uma migração de conhecimento muito grande.

Podem ser destacados no ecossistema do *Spark*, o *Spark Streaming* que possibilita o processamento de dados em tempo real, *Spark SQL* que possibilita tratar conjuntos de dados no mesmo esquema de ferramentas tradicionais e também trabalhar com a extração de dados em diferentes formatos para consultas.

O *Spark* tem *RDD* como conceito que são conjuntos de dados distribuídos, funciona como se fosse uma tabela em um banco de dados. *Data frames* também podem ser criados e podem receber as diversas funções de agregação do *SQL* para processar e tratar dados.

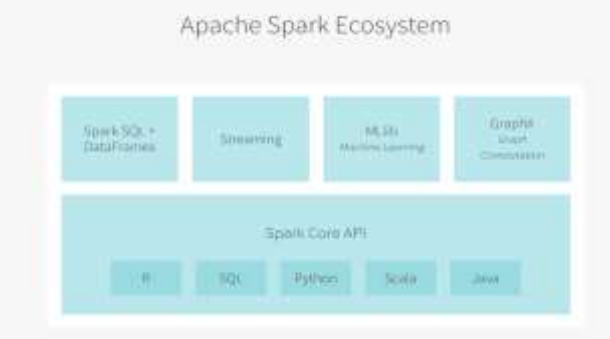


Figura 3.2 Ecossistema do Spark

Fonte: (DATABRICKS, 2017)

Além das bibliotecas destacadas no ecossistema do *Spark* e as linguagens de programação que compõem a o *Spark Core*, pode-se falar da biblioteca de *Machine Learning*, *MLib*, que trabalha com algoritmos de aprendizagem, classificação, regressão dentre outros e também a biblioteca *GraphX* que possibilita a interação com dados através da criação de grafos.

4. DATA MINING

Big data proporciona uma volumetria de dados gigantesca, apesar de todos esses dados estarem disponíveis pode ser muito estressante para quem precisa analisa-los pois são muitas variáveis para ver ou pensar. A procura por padrões pode ser uma vantagem para quem busca o sucesso, estudar durante a semana e buscar por um padrão de estudo pode ajudar a refletir em melhores resultados, um time de futebol quando busca um padrão de jogo a longo prazo durante um campeonato pode ter um melhor rendimento, o mercado quando identifica um padrão de venda pode focar seus recursos para esse padrão e ter um reflexo positivo em vendas.

Buscar padrões é uma maneira inteligente de capturar informações no oceano de dados existente, o processamento de técnica de mineração de dados ou *data mining* é composto por três frentes que se complementam: a estatística, inteligência artificial e *machine learning*.

Até pouco tempo atrás empresas queriam ter uma informação diária da quantidade de vendas e quais foram os produtos mais vendidos por departamento, hoje o mercado é extremamente concorrente e só esse tipo de informação não basta, muitas empresas hoje precisam saber em tempo real o que está afetando suas vendas ou a queda dessas vendas, assim como o perfil do cliente e se tem um momento exato que esse cliente consome um determinado produto, se existe relação de venda de um produto com outro, ou seja, muitas variáveis e condições a serem analisadas e ter o conhecimento de relações temporais, aproximações pode ser um passo à frente muito importante.

Para buscar padrões utilizando *data mining* existem diversas técnicas e algoritmos que são utilizados. Os modelos são divididos em: descritivo, prescritivo e preditivo. Com os métodos *data mining* pode-se agrupar registros semelhantes, detectar anomalias, como tentativas de compras *online* usando um cartão de crédito em uma região diferente da normal, detectar relação entre dados.

Dentre as técnicas de *data mining* podem ser destacadas nos modelos descritivos: regras de associação, *clustering*, grupos de afinidade, nos modelos preditivos podem ser destacadas: regressão, árvore de decisão, redes neurais e para os modelos prescritivos tem-se: análises preditivas e suas regras, otimização de *marqueting*.

Data mining é isso, um campo multidisciplinar, que trabalha maneiras de automatizar a busca por padrões, relações, tendências utilizando técnicas e algoritmos em bases de dados para prever resultados futuros.

Para utilização de técnicas de *data mining* as bases de dados precisam passar por fases de tratamento, para se tirar valor de um dado este precisar ser confiável, desta maneira surge um processo chamado *RDD* (*Knowledge Discovery* in *Databases*), ou extração de conhecimento.

> *A entrada do processo é um banco de dados e a saída um conjunto de conhecimentos. A etapa principal é a de Mineração ou Análise dos Dados*

> *(Data Mining). A análise nunca é feita sobre todos os dados e sim sobre amostras. Para tanto, é necessário ter antes uma etapa de preparação dos dados, a partir da base de entrada. Nesta etapa, os dados serão tratados (limpeza, integração, deduplicidade) e amostras diferentes serão geradas. (LOH, 2014, p.19)*

Gerar conhecimento através de técnicas de mineração de dados é objetivo para ser competitivo e evitar futuras perdas e projetar maiores ganhos, como também elaborar prognósticos futuras em diversas áreas.

O processo de descoberta de conhecimento passa por algumas fases para que o dado possa ser minerado, coletar dados objetivos que tenham influencia no projeto, para esta fase, usa-se ferramentas *ETL* (Extração, Transformação e Carga), como *Apache Nifi* e *Talend*, aí é onde se faz toda limpeza do dado para dar carga em outras bases para utilização posterior. Feita a coleta de dados, ter os dados com o formato correto é muito importante, no processo de descoberta também é essencial a qualidade, *garbage in garbage out*, se o dado é ruim o resultado, a informação será ruim.

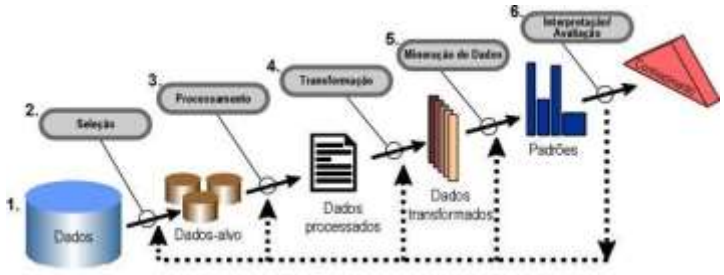


Figura 4.1 Processo de descoberta de conhecimento

Fonte: (researchgate, 2015)

5. TÉCNICAS USADAS PARA DATA MINING

Técnicas de mineração de dados são métodos para encontrar padrões e poder extrair valor de dados que foram previamente tratados para receber algum tipo de mineração. A técnica de associação é uma das técnicas mais conhecidas, tem como objetivo avaliar valores que podem aparecer juntos em transações, identificar relação entre esses valores dentro de uma mesma classe. "*Esta técnica é a mais famosa e ficou conhecida depois que uma rede de supermercados, ao utilizar uma ferramenta de Data Mining com esta técnica, descobriu que, nas 6as-feiras, quem comprava fraldas também comprava cerveja*. " (LOH, 2014, p.128)

Uma outra técnica bastante conhecida é a análise de regressão e modelos de predição, técnica estatística que elabora um modelo para explicar a relação entre as variáveis, podendo ser duas ou mais. "*A principal vantagem de poder determinar a relação entre duas variáveis é poder realizar previsões sobre o comportamento futuro das variáveis, calculando um valor quantitativo futuro ou até mesmo podendo prever acontecimentos (eventos) que ainda não ocorreram.* " (LOH, 2014, p.132)

Outliers é um método onde se analisa os casos atípicos em uma determinada conjuntura, um comportamento estranho e incomum, situações as quais fogem do padrão de ocorrência, como uma

compra usando um cartão de crédito, onde corriqueiramente o usuário faz compras muito diferentes da que foi feita.

Está técnica também é conhecida como detecção de desvios. "*A técnica de detecção de desvios utiliza funções ou intervalos médios (padrões), mas seu objetivo é estar atento ao que se desviou dos valores médios, os outliers. Em alguns casos, eles são mais importantes que os casos normais.* " (LOH, 2014, p.134)

Séries temporais baseiam-se em comportamentos que se repetem, ou uma parte ao longo do tempo, com isso pode-se tentar tratar uma previsão em um curto período de tempo, um bom exemplo seria o comportamento do valor de mercado de uma ação, onde seria possível predizer o valor desta ação em um futuro intervalo, dentro de um período curto de tempo.

A técnica de agrupamento, seleciona os elementos mais similares e menos similares e agrupa em diferentes grupos, ou seja, para cada elemento parecido ou muito similar agrupa-se em uma classe, podendo haver uma ou mais classes, grupos.

6. PROUNI

A educação pública historicamente é o calcanhar de Aquiles brasileiro, apesar de todo esforço de grupos ligados à educação como professores, cursinhos populares, pesquisadores. Existe uma dificuldade que vai além do esforço da comunidade escolar, o pouco investimento na educação que constitucionalmente tem seu ciclo fundamental garantido não se mostra suficiente para atender a demanda e proporcionar uma educação básica de qualidade.

Um grande problema de não ter educação básica de qualidade para a população de baixa renda é que as universidades públicas que apesar de ter boa qualidade ofertam poucas vagas e a competitividade candidato por vaga é muito alta, como a universidade é pública uma pequena parte da população que pode ter uma educação básica de qualidade por meio do setor privado acaba ficando com essas vagas.

"Os principais obstáculos que impedem essa camada da população de obter êxito nos processos seletivos são o número ainda reduzido de Instituições de Ensino Superior (IES) públicas, a elevada relação candidato/vaga, bem como a formação deficiente na educação básica da rede pública. " (CUNHA e col., 2016, p.239)

As estatísticas da FUVEST de 2017 mostram que 63.1% dos aprovados na primeira fase para ingressar na Universidade de São Paulo cursaram todo ensino médio em escola particular contra 21.6% que cursaram todo ensino médio em escola pública. O ProUni possibilita que alunos de baixa renda acessem o ensino superior através das bolsas de estudos ofertadas pelas instituições de ensino, o que compreende a mensalidade que o bolsista fica isento de pagar ou paga apenas uma parte, quando a bolsa não é integral.

Para conseguir uma bolsa no ProUni é necessário fazer o Exame Nacional do Ensino Médio, o ENEM, para que na abertura das inscrições do programa cada estudante faça a inscrição e concorra à vaga na universidade de sua escolha, desde que tal instituição esteja participando do programa, neste processo existe a concorrência por meio da nota obtida no ENEM, além disso existe uma nota mínima para poder pleitear uma das vagas. O MEC se faz claro quanto as exigências de

ingresso no ProUni "*Para concorrer a uma bolsa, o estudante deve participar do Exame Nacional do Ensino Médio (Enem), na edição imediatamente anterior ao processo seletivo do ProUni, e obter a nota mínima nesse exame, estabelecida pelo MEC. (MEC, 2005, p.1)* "

Existem outras exigências de extrema importância, o candidato precisa ter renda familiar de até três salários mínimos por pessoa, outras condições que o candidato precisar garantir, não necessariamente todas é: ter cursado o ensino médio completo ou parcialmente em escola pública ou utilizando-se de bolsa de estudos integral em escola particular, ser pessoa com deficiência, ser professor da rede pública de ensino básico em exercício efetivamente, neste caso o candidato precisa concorrer às vagas de cursos de licenciatura, normal superior ou pedagogia. Quando o candidato é professor da rede pública de ensino básico a renda familiar não é considerada.

As instituições de ensino que fazem parte do programa são responsáveis por aferir a veracidade das informações do candidato. "*Cabe à instituição de ensino, na figura do coordenador do ProUni, a aferição dos documentos apresentados pelo candidato para a comprovação das informações prestadas em sua inscrição no programa.* " (MEC, 2005, p.1)

O programa tem uma contrapartida para as instituições de ensino que participam que é conceder isenção em tributos como: Imposto de renda das pessoas Jurídicas, Contribuição Social sobre o Lucro Líquido, Contribuição Social para o financiamento da Seguridade Social e Contribuição Social para o Financiamento da Seguridade Social. Para que passe a valer a isenção dos tributos de IRPJ, CSLL, Cofins e PIS, a instituição precisa assinar o termo de adesão de 10 anos de vigência, a instituição necessita estar com os cursos regularmente cadastrados no INEP e também deve ter regularidade fiscal, analisada pelo MEC, mediante a consulta ao Cadastro Informativo de Créditos não quitados do Setor Público Federal (Cadin),

O ProUni foi um acerto para a o ensino superior brasileiro, assim como fez bem para muitos alunos que acessaram o ensino superior através do programa contribuiu com instituições de ensino através dos incentivos fiscais. Os alunos das instituições que fazem parte do ProUni precisam apresentar bom rendimento para que fiquem no programa e não tenham a bolsa de estudos cancelada. *"Durante o curso, o bolsista do ProUni deverá apresentar aproveitamento acadêmico de, no mínimo, 75% (setenta e cinco por cento) nas disciplinas cursadas em cada período letivo, sob pena de encerramento da bolsa.* " (MEC, 2005, p.1)

Um outro ponto importante é que mesmo se o candidato não atingir a nota para concorrer a uma bolsa integral, pode usar o Fies, programa de financiamento estudantil, para custear parte da mensalidade quando conseguir uma bolsa parcial de 50%. O ProUni também usa o sistema de cotas para inscrição no programa, utilizando uma porcentagem das bolsas de estudos disponibilizados para os cotistas, os candidatos cotistas podem escolher entre concorrer pela lista de cotas ou pela ampla concorrência. "*O ProUni reserva, em processo seletivo, bolsas às pessoas com deficiência e aos autodeclarados pretos, pardos ou índios*. " (MEC, 2005, p.1)

O programa também oferece uma bolsa-permanência para os bolsistas que estudam em período integral em cursos acima de seis semestres e presencial, a bolsa-permanência é concedida automaticamente no início de cada semestre. "*O ProUni instituiu, em 2006, a Bolsa-Permanência,*

destinada a ajudar no custeio das despesas educacionais dos estudantes. É um benefício, de até R$ 300,00 mensais, concedido a estudantes com bolsa integral". (MEC, 2005, p.1)

O programa universidade para todos foi tão bem aceito que o governo federal criou o SISU, sistema de seleção unificada que também utiliza o ENEM para que os candidatos possam pleitear vagas em universidades públicas.

7 ESTUDO DE CASO

7.1 MATERIAIS E MÉTODOS

Para a elaboração do artigo foi utilizada a pesquisa bibliográfica onde foram destacadas citações pontuais dos autores, a pesquisa bibliográfica consiste na pesquisa de material existente sobre o assunto que se está sendo estudado, como: livros, revistas, artigos de internet.

"*A pesquisa bibliográfica é elaborada a partir de material já publicado, constituído principalmente de: livros, revistas, publicações em periódicos e artigos científicos, jornais, boletins, monografias, dissertações, teses, material cartográfico, internet.* " (PRODANOV; FREITAS, 2013, p.54)

Além da pesquisa bibliográfica o artigo é composto por um estudo de caso, onde faz uma análise dos dados históricos do ProUni, o estudo de caso trata de coletar e analisar informações, podendo ser uma análise qualitativa ou quantitativa. "*Um Estudo de Caso, independentemente de qualquer tipologia, orientará a busca de explicações e interpretações convincentes para situações que envolvam fenômenos sociais complexos*" (PRODANOV; FREITAS, 2013, p.61)

Para fazer tal análise foi utilizada a linguagem de programação *Python* e as bibliotecas: *Matplotlib* e *Pandas* que fazem parte do pacote *Anaconda Navigator version 5.3*, onde foram feitos os processamentos de todos os arquivos .csv disponíveis no site oficial do ProUni, além dos processamentos, foram gerados diversos gráficos para a visualização dos dados. Durante o desenvolvimento dos *scripts* o ambiente *Jupyter notebook* foi utilizado para fazer o uso da linguagem de programação e das bibliotecas sendo executado no navegador *Google Chrome* versão *69.0.3497.100* em ambiente local, a escolha por todas essas ferramentas se deu pelo fato das mesmas fazerem parte do universo *big data***.**

Para leitura que antecedeu o processamento dos arquivos .csv no *Python* o *Microsoft Excel 2016* foi utilizado para verificar se os arquivos não estavam corrompidos, a ferramenta *online Draw.io* foi utilizada para criar o diagrama de caso de fluxo o qual mostra o processo *ETL* que é implementado para extração, leitura e transformação dos dados.

7.2 DESENVOLVIMENTO

O estudo de caso segue um fluxo de extração, leitura e transformação de dados, onde os arquivos estão disponíveis na Internet, utilizando um navegador comum para acessá-los e fazer download, após isso a leitura é feita utilizando a linguagem de programação e as bibliotecas, assim como na transformação e na visualização dos dados processados.

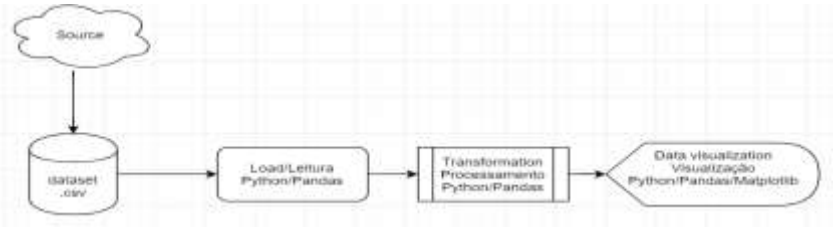


Figura 7.2.1 Fluxo dos dados

Os dados Históricos do ProUni foram processados utilizando Python 3, as bibliotecas *Pandas* e *Matplotlib* foram importadas e os *data frames* foram gerados consumindo os arquivos .csv que foram disponibilizados na página do ProUni.

```
In [ ]: import matplotlib.pyplot as plt
        %matplotlib inline
        import pandas as pd
        import numpy as np
        #Lendo os arquivos csv e criando dataframes para serem tratados

        df_PROUNI_2016= pd.read_csv('/Users/Blueshift014/Documents/Base Prouni/PDA_PROUNI_2016_CSV.CSV', sep=';', encoding='latin-1')
        df_PROUNI_2015= pd.read_csv('/Users/Blueshift014/Documents/Base Prouni/PDA_PROUNI_2015_CSV.CSV', sep=';', encoding='latin-1')
        df_PROUNI_2014= pd.read_csv('/Users/Blueshift014/Documents/Base Prouni/PDA_PROUNI_2014_CSV.CSV', sep=';', encoding='latin-1')
        df_PROUNI_2013= pd.read_csv('/Users/Blueshift014/Documents/Base Prouni/PDA_PROUNI_2013_CSV.CSV', sep=';', encoding='latin-1')
        df_PROUNI_2012= pd.read_csv('/Users/Blueshift014/Documents/Base Prouni/PDA_PROUNI_2012_CSV.CSV', sep=';', encoding='latin-1')
        df_PROUNI_2011= pd.read_csv('/Users/Blueshift014/Documents/Base Prouni/PDA_PROUNI_2011_CSV.CSV', sep=';', encoding='latin-1')
        df_PROUNI_2010= pd.read_csv('/Users/Blueshift014/Documents/Base Prouni/PDA_PROUNI_2010_CSV.CSV', sep=';', encoding='latin-1')
        df_PROUNI_2009= pd.read_csv('/Users/Blueshift014/Documents/Base Prouni/PDA_PROUNI_2009_CSV.CSV', sep=';', encoding='latin-1')
        df_PROUNI_2008= pd.read_csv('/Users/Blueshift014/Documents/Base Prouni/PDA_PROUNI_2008_CSV.CSV', sep=';', encoding='latin-1')
        df_PROUNI_2007= pd.read_csv('/Users/Blueshift014/Documents/Base Prouni/PDA_PROUNI_2007_CSV.CSV', sep=';', encoding='latin-1')
        df_PROUNI_2006= pd.read_csv('/Users/Blueshift014/Documents/Base Prouni/PDA_PROUNI_2006_CSV.CSV', sep=';', encoding='latin-1')
        df_PROUNI_2005= pd.read_csv('/Users/Blueshift014/Documents/Base Prouni/PDA_PROUNI_2005_CSV.CSV', sep=';', encoding='latin-1')
        #Concatenando todos os arquivos para criação de uma única base
        df_PROUNI_Historico = pd.concat([df_PROUNI_2016,df_PROUNI_2015,df_PROUNI_2014, df_PROUNI_2013, df_PROUNI_2012, df_PROUNI_2011,
                                         df_PROUNI_2010, df_PROUNI_2009, df_PROUNI_2008, df_PROUNI_2007, df_PROUNI_2006, df_PROUNI_2005])
        #Gerando os dataframes por cidades para depois gerar o dataframe do CIMBAJU

        df_PROUNI_MORATO = df_PROUNI_Historico[(df_PROUNI_Historico['MUNICIPIO_BENEFICIARIO_BOLSA']=='FRANCISCO MORATO')]
        df_PROUNI_CAIEIRAS = df_PROUNI_Historico[(df_PROUNI_Historico['MUNICIPIO_BENEFICIARIO_BOLSA']=='CAIEIRAS')]
        df_PROUNI_FRANCO_DA_ROCHA = df_PROUNI_Historico[(df_PROUNI_Historico['MUNICIPIO_BENEFICIARIO_BOLSA']=='FRANCO DA ROCHA')]
        df_PROUNI_CAJAMAR = df_PROUNI_Historico[(df_PROUNI_Historico['MUNICIPIO_BENEFICIARIO_BOLSA']=='CAJAMAR')]
        df_PROUNI_MAIRIPORA = df_PROUNI_Historico[(df_PROUNI_Historico['MUNICIPIO_BENEFICIARIO_BOLSA']=='MAIRIPORÃ')]
        #Criação do dataframe CIMBAJU com a composição dos dados integrais das cidades de: Francisco Morato, Franco da Rocha, Caieiras,
        #Cajamar e Mairiporã
        df_PROUNI_CIMBAJU = pd.concat([df_PROUNI_MORATO, df_PROUNI_CAIEIRAS, df_PROUNI_FRANCO_DA_ROCHA,
```

Figura 7.2.2 Leitura dos arquivos .csv no Python

Uma vez criado o *data frame* histórico que contém todos os registros e o *data frame* do CIMBAJU que contém os registros das cidades da região da bacia do Juqueri, foi conferida a quantidade de registro que o *data frame* histórico detém, assim como o do CIMBAJU.

```
In [38]: df_PROUNI_Historico['ANO_CONCESSAO_BOLSA'].count()

Out[38]: 1989317

In [41]: df_PROUNI_CIMBAJU['ANO_CONCESSAO_BOLSA'].count()
Out[41]: 8146
```

Figura 7.2.3 Contagem dos registros nos dataframes

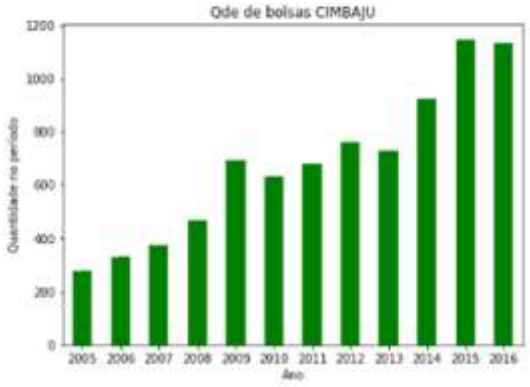
Com a contagem feita o total de bolsas concedidas durante o intervalo de anos de 2005 até 2016 foi de um milhão novecentos e oitenta e nove mil trezentos e dezessete. Portanto quase 2 milhões de bolsas foram ofertadas no Brasil, enquanto que na região do CIMBAJU foi de oito mil cento e quarenta e seis. Analisando o gráfico, também pode ser analisado que os anos de 2015 e 2016 foram os anos que houve mais concessões de bolsas tanto no Brasil quanto no CIMBAJU, sendo o ano de 2015 o ano que mais bolsas foram concedidas.

```
In [46]: #Linha de codigo que plota grafico do tipo de bolsa em formato de barra e mostra a quantidade
df_PROUNI_Historico.groupby('ANO_CONCESSAO_BOLSA').count().plot(kind ='bar',
figsize =(7,5), grid=False, rot =0, color = 'green', legend=False)
plt.title('Qde de bolsas Brasil')
plt.xlabel('Ano')
plt.ylabel('Quantidade no período')
plt.show()
df_PROUNI_Historico['ANO_CONCESSAO_BOLSA'].value_counts()
```


```
Out[46]: 2015    252650
         2016    239262
         2014    223598
         2013    177326
         2012    176764
         2011    170766
         2009    161369
         2010    152733
         2008    124621
         2006    109025
         2007    105574
         2005     95629
```

Figura 7.2.4 Contagem de bolsas concedidas por ano no Brasil

```
In [45]: #Linha de codigo que plota grafico do tipo de bolsa em formato de barra e mostra a quantidade
df_PROUNI_CIMBAJU.groupby('ANO_CONCESSAO_BOLSA').count().plot(kind ='bar',
figsize =(7,5), grid=False, rot =0, color = 'green', legend=False)
plt.title('Qde de bolsas CIMBAJU')
plt.xlabel('Ano')
plt.ylabel('Quantidade no período')
plt.show()
df_PROUNI_CIMBAJU['ANO_CONCESSAO_BOLSA'].value_counts()
```




```
Out[45]: 2015    1145
         2016    1132
         2014     923
         2012     759
         2013     728
         2009     694
         2011     682
         2010     631
         2008     470
         2007     374
         2006     329
         2005     279
```

Figura 7.2.5 Contagem de bolsas concedidas por ano no CIMBAJU

A próxima consulta nos *data frames* mostra que das cinco universidades que mais concederam bolsas de estudo no Brasil e no CIMBAJU durante os anos de 2005 até 2016 foi a Universidade Paulista.

```
In [51]: df_PROUNI_Historico['NOME_IES_BOLSA'].value_counts().sort_values(ascending=False).head(5)
Out[51]: UNIVERSIDADE PAULISTA                86947
         UNIVERSIDADE PITÁGORAS UNOPAR       86254
         UNIVERSIDADE ANHANGUERA - UNIDERP   49300
         UNIVERSIDADE ESTÁCIO DE SÁ          43006
         UNIVERSIDADE SÃO JUDAS TADEU        38771
```

```
In [53]: df_PROUNI_CIMBAJU['NOME_IES_BOLSA'].value_counts().sort_values(ascending=False).head(5)
Out[53]: UNIVERSIDADE PAULISTA                1269
         UNIVERSIDADE NOVE DE JULHO           1140
         UNIVERSIDADE SÃO JUDAS TADEU          714
         CENTRO UNIVERSITÁRIO PADRE ANCHIETA   325
         FACULDADE FLAMINGO                    312
```

Figura 7.2.6 Top 5 das instituições de ensino no Brasil e CIMBAJU

Olhando apenas para as instituições que concederam bolsas integrais, há alteração no *TOP* 5, sendo a Universidade Pitágoras UNOPAR como a primeira colocada no Brasil e a Universidade Paulista ainda como a primeira na região do CIMBAJU, por conta desta nova condição, algumas

instituições deixaram de integrar o *TOP* 5, como a São Judas Tadeu no TOP 5 do Brasil e a Faculdade Flamingo e a própria São Judas no CIMBAJU.

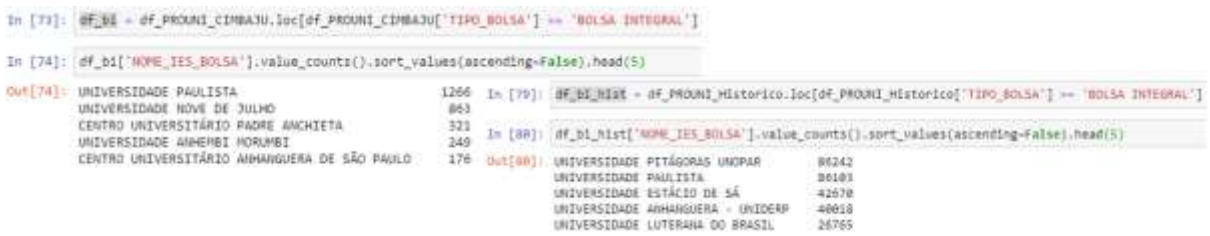


Figura 7.2.7 Top 5 das instituições de ensino com bolsa integral no Brasil e CIMBAJU

A região Sudeste do Brasil é onde se concentra a maior parte da população, seguida da região Nordeste, entretanto seguido do Sudeste, a região Sul que é uma região pequena e com um menor número populacional tem sido a mais beneficiada.

A região Sudeste foi contemplada com novecentos e setenta e seis mil cento e oito bolsas de estudos, neste caso o CIMBAJU está incluído, da região do CIMBAJU, Francisco Morato foi contemplada com duas mil seiscentos e oito bolsas de estudos, números que compreendem o intervalo de 2005 até 2016.

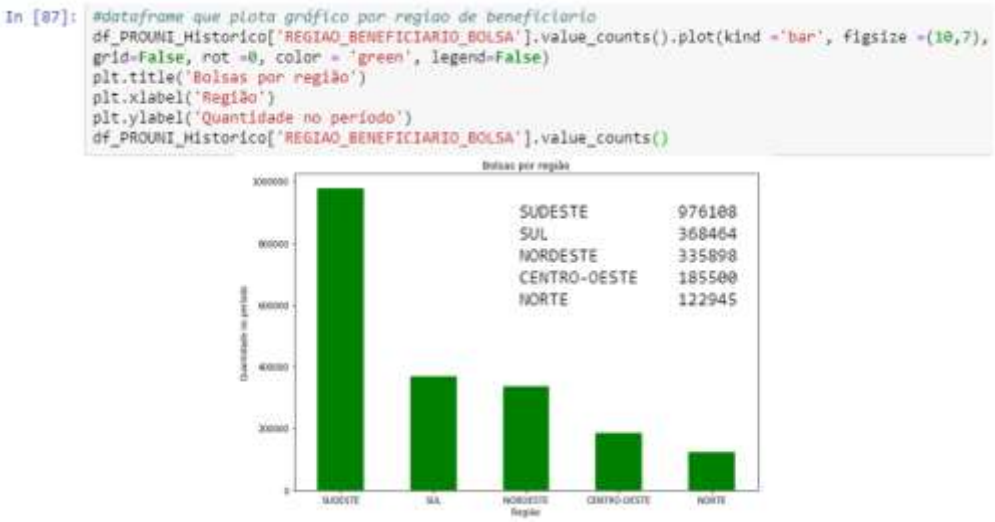


Figura 7.2.8 Concessões por região no Brasil

Em relação a região do CIMBAJU, a cidade de Francisco Morato foi a cidade com maior quantidade de bolsas concedidas, dentre as cinco cidades da região, Francisco Morato é a cidade com maior quantidade populacional e a única que faz parte da grande São Paulo, indicativo que pode apontar o porquê ser a cidade que teve mais bolsas concedidas assim como a região Sudeste

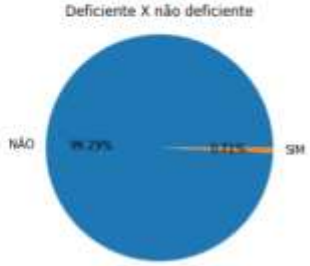
```
In [90]: #dataframe que plota grafico por regiao de beneficiario
df_PROUNI_CIMBAJU['MUNICIPIO_BENEFICIARIO_BOLSA'].value_counts().plot(kind ='bar', figsize =(10,7),
grid=False, rot =0, color = 'green', legend=False)
plt.title('Bolsas por município')
plt.xlabel('Município')
plt.ylabel('Quantidade no período')
df_PROUNI_CIMBAJU['MUNICIPIO_BENEFICIARIO_BOLSA'].value_counts()
```


```
FRANCISCO MORATO    2688
FRANCO DA ROCHA     2154
CAIEIRAS            1426
MAIRIPORÃ           1001
CAJAMAR              957
```

Figura 7.2.9 Concessões por município no CIMBAJU

Fazendo uma análise em todas as regiões do Brasil, a quantidade de bolsas que foram concedidas para pessoas que se declararam deficientes foi extremamente baixa, menos de 1% das bolsas foram destinadas para pessoas com deficiência. As mulheres formam a maioria dos contemplados durante os anos, sendo 53.40% do total.

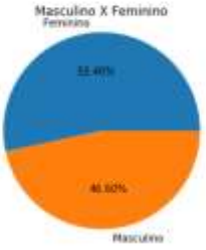
```
df_PROUNI_Historico.BENEFICIARIO_DEFICIENTE_FISICO.value_counts().plot(kind='pie', autopct='%.2f%%')
plt.title('Deficiente X não deficiente')
plt.axis('equal')  # Para deixar o grafico redondo
df_PROUNI_Historico['BENEFICIARIO_DEFICIENTE_FISICO'].value_counts()
```




```
NÃO     1975119
SIM       14198
```

Figura 7.2.10 Deficiente x não deficiente no Brasil

```
df_PROUNI_Historico.SEXO_BENEFICIARIO_BOLSA.value_counts().plot(kind='pie', autopct='%.2f%%')
plt.title('Masculino X Feminino')
plt.axis('equal')  # Para deixar o grafico redondo
df_PROUNI_Historico['SEXO_BENEFICIARIO_BOLSA'].value_counts()
```



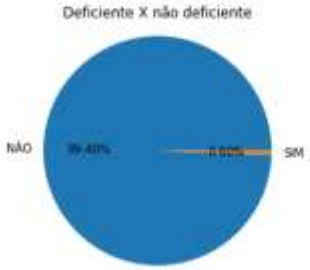

```
Feminino     1062394
Masculino     926923
```

Figura 7.2.11 masculino x feminino no Brasil

Fazendo a mesma análise olhando para a região do CIMBAJU, o número de bolsas concedidas para pessoas com deficiência é um pouco menor sendo 0,60%, menos que 1% do total,

vale ressaltar que durante todos os anos apenas 49 bolsas foram concedidas nesta região para pessoas que se declararam deficientes.

```
df_PROUNI_CIMBAJU.BENEFICIARIO_DEFICIENTE_FISICO.value_counts().plot(kind='pie', autopct='%.2f%%')
plt.title('Deficiente X não deficiente')
plt.axis('equal')  # Para deixar o gráfico redondo
df_PROUNI_CIMBAJU['BENEFICIARIO_DEFICIENTE_FISICO'].value_counts()
```

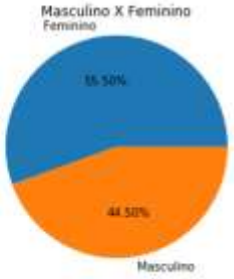



```
NÃO    8097
SIM      49
```

Figura 7.2.12 Deficiente x não deficiente no CIMBAJU

As mulheres também são a maioria na região, 55.50% do total, uma diferença de pouco mais de 2% em relação a base geral onde o total de mulheres é 53.40%.

```
df_PROUNI_CIMBAJU.SEXO_BENEFICIARIO_BOLSA.value_counts().plot(kind='pie', autopct='%.2f%%')
plt.title('Masculino X Feminino')
plt.axis('equal')  # Para deixar o gráfico redondo
df_PROUNI_CIMBAJU['SEXO_BENEFICIARIO_BOLSA'].value_counts()
```




```
Feminino     4521
Masculino    3625
```

Figura 7.2.13 Masculino x feminino no CIMBAJU

Durante os onze anos analisados na base de dados o curso mais escolhido em todas as regiões do Brasil foi Administração. O gráfico mostra exatamente os *top* 5 dos cursos.

```
df_cursos_x_regiao=pd.crosstab(df_PROUNI_Historico["NOME_CURSO_BOLSA"],
df_PROUNI_Historico["REGIAO_BENEFICIARIO_BOLSA"],margins=True).sort_values(by='All', ascending = False).head(6)
df_cursos_x_regiao.rename(columns={"All": "Total"}, inplace=True)
df_cursos_x_regiao
```

REGIAO_BENEFICIARIO_BOLSA	CENTRO-OESTE	NORDESTE	NORTE	SUDESTE	SUL	Total
NOME_CURSO_BOLSA						
All	185088	334249	122514	973909	367506	1983266
Administração	27124	46537	16320	129341	55277	274599
Direito	21608	34573	9227	77140	32946	175494
Pedagogia	15634	22351	10746	72019	26358	147108
Ciências Contábeis	12440	21084	8411	43490	20547	105972
Enfermagem	8407	19725	4998	30657	8419	72206

Figura 7.2.14 Top 5 de cursos por região do Brasil

Tratando-se da região do CIMBAJU foi o curso de Pedagogia o mais escolhido. O gráfico mostra exatamente os *top* 5 dos cursos.

```
df_cursos_x_cidades=pd.crosstab(df_PROUNI_CIMBAJU["NOME_CURSO_BOLSA"],
df_PROUNI_CIMBAJU["MUNICIPIO_BENEFICIARIO_BOLSA"],margins=True).sort_values(by='All', ascending = False).head(6)
df_cursos_x_cidades.rename(columns={"All": "Total"}, inplace=True)
df_cursos_x_cidades
```

MUNICIPIO_BENEFICIARIO_BOLSA NOME_CURSO_BOLSA	CAIEIRAS	CAJAMAR	FRANCISCO MORATO	FRANCO DA ROCHA	MAIRIPORÃ	Total
All	1423	957	2800	2147	1001	8128
Pedagogia	128	94	338	243	56	859
Administração	132	82	254	237	118	823
Direito	111	72	150	134	79	546
Gestão De Recursos Humanos	55	50	121	81	35	342
Logística	47	52	137	84	22	342

Figura 7.2.14 Top 5 de cursos por cidade no CIMBAJU

Os dados do ProUni apontaram que as concessões no intervalo histórico foram em maior parte para bolsistas de raça branca, tanto no Brasil quanto na região do CIMBAJU, seguidos dos pardos e depois dos pretos. Analisando o gráfico as concessões para raça indígena e preta são muito menores comparando com pardos e brancos, neste caso menos da metade de ambas as raças obtiveram a concessão da bolsa de estudo.

```
df_PROUNI_CIMBAJU['RACA_BENEFICIARIO_BOLSA'].value_counts().plot(kind ='bar', figsize =(10,7),
grid=False, rot =0, color = 'green',legend=False)
plt.title('Quantidade de concessões por raça na região do CIMBAJU 2005-2016')
plt.xlabel('Raça')
plt.ylabel('Quantidade no período')
plt.show()
df_PROUNI_CIMBAJU['RACA_BENEFICIARIO_BOLSA'].value_counts()
```

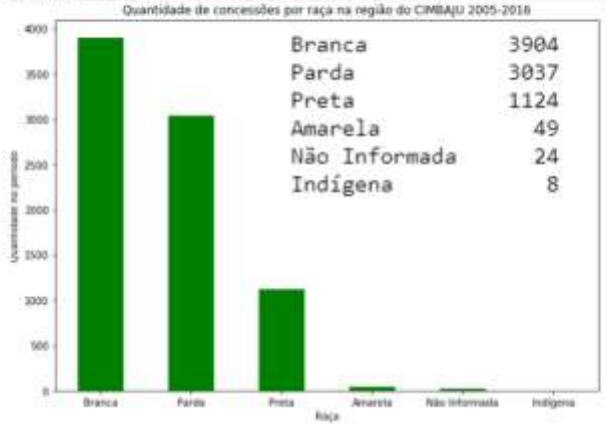


Figura 7.2.15 Quantidade de concessões por raça no CIMBAJU

```
df_PROUNI_Historico['RACA_BENEFICIARIO_BOLSA'].value_counts().plot(kind ='bar', figsize =(10,7),
grid=False, rot =0, color = 'green',legend=False)
plt.title('Quantidade de Concessões por raça no Brasil 2005-2016')
plt.xlabel('Raça')
plt.ylabel('Quantidade no período')
plt.show()
df_PROUNI_Historico['RACA_BENEFICIARIO_BOLSA'].value_counts()
```

Figura 7.2.16 Quantidade de concessões por raça no Brasil

7.3 RESULTADOS

Analisando a base de dados do ProUni as mulheres formaram a maior parte dos contemplados, resultado que se alinha com o censo da educação superior de 2016, onde aponta que as mulheres são maioria no ensino superior. "*Dados do Censo da Educação Superior de 2016, última edição do levantamento, revelam que as mulheres representam 57,2% dos estudantes matriculados em cursos de graduação". (INEP, 2016, p.1).* Um resultado muito esperado na análise dos dados foi a região Sudeste ter a maior quantidade das bolsas concedidas, pois além de ser uma região economicamente forte com a maior parte das instituições de ensino é onde também se concentra a maior parte da população brasileira. "*A Região Sudeste está em primeiro lugar em número de alunos matriculados em cursos presenciais, com mais de 3 milhões ou 47% do total, apresentando, em 2014. " (SEMESP, 2016, p.33)*

Com pouco mais de 8 mil bolsas de estudo concedidas na região do CIMBAJU durante os 11 anos analisados através da base de dados pode-se concluir que o total de bolsas para a região do CIMBAJU foi extremamente baixo, mesmo a região não tendo um número expressivo de instituições de ensino superior e sendo uma região economicamente fragilizada, ainda assim São Paulo é o Estado que está na dianteira da região Sudeste em número de matriculas no ensino superior. Só no ano de 2014 São Paulo registrou mais de 1,5 milhão de matriculas no ensino superior, apenas em cursos presenciais. "*O Estado de São Paulo foi o grande responsável por esse índice, apresentando 1,7 milhão de matrículas em cursos presenciais (56% do total na região) " (SEMESP, 2016, p.33)*

Um dado que chamou muito a atenção foi a quantidade de pessoas com deficiência que conseguiram uma bolsa de estudos durante esses 11 anos analisados, menos de 1% do total. Na região do CIMBAJU apenas 49 bolsas concedidas para pessoas com deficiência, o que seria menos de 5 bolsas por ano. Números que coincidem com a realidade do país. "*Apenas 0,45% do total de 8 milhões de matrículas no ensino superior são de alunos com deficiência. Na rede privada, o percentual é ainda menor, 0,35%, enquanto na rede pública ele chega a 0,73%" (*ENSINO SUPERIOR, 2017, p.8)

O Brasil é o segundo país com mais negros (pretos e pardos) no mundo, ou seja, a maior parte da população é preta ou parda e os dados do ProUni refletiram esta realidade, somando-se pretos e pardos, estes compõem a maioria nas concessões das bolsas de estudo, o ProUni utiliza o sistema de cotas raciais e isso também pode explicar o bom número de concessões para os negros. É importante ressaltar que segundo pesquisas do IBGE de modo geral quando se fala do ensino superior os negros são extrema minoria em relação aos brancos, mesmo sendo a maior parte da população. Os indígenas são minoria tanto na região do CIMBAJU quanto no Brasil.

O expressivo número do ProUni de quase 2 milhões de concessões de bolsa de estudos até 2016 se consolidou pelo sucesso do programa onde pode-se notar que ao passar dos anos o número de concessões anual teve uma tendência de alta, onde a partir de 2014 as concessões ultrapassaram as 200 mil bolsas concedidas, mostrando que o programa tem tudo para continuar e que mesmo em anos de recessão econômica o programa continuou forte e sólido.

8. CONSIDERAÇÕES FINAIS

O mundo está em constante transformação, agora com a expansão da tecnologia e a geração massiva de dados, trabalhar com novas soluções como as de Big *data* para filtrar dados e criar informação de valor é mais do que necessário. O conhecimento é a chave para implementar soluções inovadoras. Uma sociedade que em muitos lugares está vivendo uma experiência migratória fazendo com que os centros urbanos se tornem extremamente ineficientes, enquanto que em outros lugares o envelhecimento e a baixa taxa de natalidade está se tornando um problema, poderia ser muito mais beneficiada pela tecnologia da informação e suas novas soluções.

Compreender os problemas sociais e aplicar as soluções inovadoras que podem ser implementadas utilizando *big data,* com conhecimento produzido *near-real time* pode ser uma das chaves para se viver em um mundo melhor, globalizado e com os problemas enfrentados atualmente reduzidos consideravelmente. *Big data* se tornou uma realidade e as tecnologias que são utilizadas para dar sustentação são muito próximas tecnicamente de ferramentas já conhecida não sendo necessária um aprendizado técnico tão penoso e tão demorado.

Os resultados das análises feitas na base dados do ProUni também mostraram que o Brasil precisa buscar soluções para problemas como a falta de inclusão no ensino superior, pois o índice de pessoas com deficiência é extremamente baixo, outro ponto é continuar investindo em regiões como Nordeste e Norte, pois são regiões historicamente carentes e necessitam de incentivos para que o ensino superior ganhe uma projeção positiva para os próximos anos, além dos investimentos e incentivos nas universidades públicas vale propor medidas para que mais instituições particulares possam se fazer presentes e acessíveis.

Outro fator conclusivo e negativo foi o baixo número de bolsas concedidas na região do CIMBAJU durante os 11 anos analisados, cidades como Francisco Morato, Franco da Rocha e Caieiras que tem fácil acesso a cidade de São Paulo e um número consideravelmente grande de habitantes, onde se concentra o maior número de instituições no Estado, poderiam ter um número maior de concessões, vale lembrar que Francisco Morato e Franco da Rocha sempre estão figurando como uma das cidades mais carentes da grande São Paulo.

Utilizando pandas foi muito prático e eficiente fazer buscas e plotar gráficos em uma base de quase 2 milhões de registros em uma máquina local, utilizar ferramentas e tecnologias *big data on cloud* para tratamento de dados em escalas gigantescas seria da mesma forma uma maneira prática e eficiente. Ao se utilizar *Python* e *Pandas* foi muito simples conseguir trabalhar com funções já existentes em *SQL* para fazer consultas e cruzar os dados da base o que pode refletir como uma resposta positiva para projetos que forem implementados utilizando *Pyspark* que é a utilização de *Python* e *data frames* em ambiente *big data*.

Para trabalhos futuros será muito importante continuar focando nas estratégias, técnicas e novas tecnologias para tratamento de dados, tratar de forma mais analítica algoritmos e técnicas de *data mining e machine learning* e também eventos *near-real time* como *IoT*(internet das coisas), trabalhar com casos de uso de ferramentas *streaming*, como o Streaming Analytics, *Spark Streaming*, *Storm*, *Kafka* e *databases* não relacionais como o *Hbase*, fazer um estudo focado em arquitetura *lambda* onde pode-se tratar de todos esses assuntos em uma única arquitetura.

9. REFERÊNCIAS

AMARAL, Fernando. **Big Data: Uma Visão Gerencial: Para Executivos, Consultores e Gerentes de Projetos**. Casa do Código. Edição do Kindle, 2016

CUNHA, M. S; LESTON, Sena; ALMEIDA, Vivianne. **O Acesso à Educação Superior por Meio do PROUNI: a Perspectiva de Egressos do Curso de Direito.** Disponível em: *https://seer.utp.br/index.php/a/article/view/288/289*. Acesso em < 20/09/2018>

Databricks. **What is Apache Spark?** Disponível em*: https://databricks.com/spark/about*. Acesso em < 22/09/2018>

Ensino Superior. **Matrículas de alunos com deficiência representam menos de 0,5% do total**, 2018. Disponível em: *http://www.revistaensinosuperior.com.br/matriculas-de-alunos-com-deficiencia-representam-menos-de-05-do-total/*. Acesso em < 03/11/2018>

Fasete. **POLÍTICAS PÚBLICAS DE ACESSO À EDUCAÇÃO SUPERIOR: um estudo sobre o P rouni em uma IES privada do munícipio de Paulo Afonso-BA**, 2014. Disponível em *https://www.fasete.edu.br/revistarios/media/revistas/2014/8/politicas_publicas_de_acesso_a_educacao_superior.pdf.* Acesso em < 18/09/2018>

Fuvest. **Questionário de avaliação socioeconômica,** 2017. Disponível em: *ttp://acervo.fuvest.br/fuvest/2017/FUVEST_2017_qase_conv_car_fuvest_2017.pdf*. Acesso em < 23/09/2018>

Inep. **Mulheres são maioria na Educação Superior brasileira**, 2018. Disponível em: *http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/mulheres-sao-maioria-na-educacao-superior-brasileira/21206*. Acesso em < 04/11/2018>

Inep. **Censo da educação superior – principais resultados**, 2016. Disponível em: *http://download.inep.gov.br/educacao_superior/censo_superior/documentos/2016/censo_superior_tabelas.pdf*. Acesso em < 04/11/2018>

Intel. **Saiba mais sobre big data?** Disponível em*: https//www.intel.com.br/contente/ dam/ www/public/lar/br/pt/documents/articles/90318386-1-por.pdf*. Acesso em < 06/08/2018>

LOH, Stanley. **BI na era do big data para cientistas de dados: indo além de cubos e dashboards na busca pelos porquês, explicações e padrões**. Casa do Código. Edição do Kindle, 2014

MARQUESONE, Rosangela. **Big Data: Técnicas e tecnologias para extração de valor dos dados**. Casa do Código. Edição do Kindle, 2016.

Planalto. **LEI No 11.096, DE 13 DE JANEIRO DE 2005.** Disponível em: *http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2005/lei/l11096.htm*. Acesso em < 20/09/2018>

PRODANOV, Cleber; FREITAS, Ernani. **Metodologia do trabalho científico: Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico**. Universidade Feevale. 2ª edição, 2013

Researchgate. **Big data e data Science: Admirável mundo novo,** 2015. Disponível em: *https://www.researchgate.net/publication/289253933_Big_Data_e_Data_Science_Admiravel_Mundo_Novo*. Acesso em < 06/08/2018>

SAS. **Hadoop O que é e qual sua importância?,** 2017. Disponível em: *https://www.sas.com/pt_br/insights/big-data/hadoop.html*. Acesso em < 20/09/2018>

SAS. **Mineração de Dados O que é e qual sua importância?,** 2017. Disponível em: *https://www.sas.com/pt_br/insights/analytics/mineracao-de-dados.html*. Acesso em < 08/08/2018>

Semesp. **Mapa do ensino superior no Brasil**, 2016. Disponível em: *http://convergenciacom.net/pdf/mapa_ensino_superior_2016.pdf*. Acesso em < 02/11/2018>

TAURION, Cezar. **Big Data**. Brasport. Edição do Kindle, 2013

Udacity. **O que é data mining? Entenda esse campo que envolve inteligência artificial e data Science,** 2018. Disponível em: *https://br.udacity.com/blog/post/data-mining*. Acesso em < 18/09/2018>