ANÁLISE COMPORTAMENTAL POR INTERMÉDIO DAS TECNOLOGIAS DE BIG DATA

Daniel dos Santos Bernardo[1]
Danilo Romualdo Borlone[2]
Artigo recebido em junho de 2015

RESUMO

O mundo está cada vez mais interligado. Novas tecnologias vêm sendo criadas para conectar e diminuir a distância entre as pessoas. Com a ascensão das tecnologias há um crescimento exponencial na geração de dados variados, estruturados ou não. Analisar estes dados em busca de informações úteis para o negócio, visando ainda, a obtenção do comportamento de um grupo de indivíduos sobre uma questão em exame, é o que este artigo aborda. Isso a partir de demonstrações e descrições das tecnologias de Big Data, explorando ainda, os conceitos fundamentais aplicados ao contexto atual e, ao final, apresenta de forma básica o emprego de algumas das principais técnicas da plataforma IBM® InfoSphere® BigInsights™, que engloba diversos recursos para o auxílio na busca por *insights* relevantes. Explorando as potencialidades do Big Data é possível obter em tempo hábil resultados de processamentos de diferentes tipos de dados, traçando perfis comportamentais e classificações de sentimentos sobre uma determinada marca, produto ou serviço, além de permitir a aproximação com o público alvo.

Palavras-chave: Big Data. Análise social. BigInsights. BigSheets.

ABSTRACT

The world is increasingly interconnected. New technologies are being created to connect and bridge the gap between people. With the rise of technology there is an exponential growth in the generation of variable data, structured or unstructured. Analyze this data for useful information for business, yet in order to obtain the behavior of a group of individuals on a matter in question is what this article addresses. This from demonstrations and descriptions of Big Data technologies, exploring further the fundamental concepts applied to the current context and in the end provides a basic way the use of some of the main techniques of IBM® InfoSphere BigInsights ™ platform, which encompasses many resources to aid in the search for relevant insights. Exploring the Big Data capabilities can be obtained in a timely manner results of processing different types of data, tracing behavioral profiles and ratings feelings about a particular brand, product or service, and allows the approach to the target clients.

Keywords: Big Data. Social Analysis. BigInsights. BigSheets.

[1] Egresso da Faculdade de Tecnologia da Zona Sul. E-mail: danielsbernardo@gmail.com.
[2] Egresso da Faculdade de Tecnologia da Zona Sul. E-mail: daniloborlone @gmail.com.

1 INTRODUÇÃO

Geram-se milhares de dados por segundo, sejam através de publicações na internet e mídias em geral, ou por transações, dispositivos, conversas etc. Muitas dessas informações são geradas publicamente e de forma voluntária, sendo significativa parte a expressão de opiniões, experiências, localizações e infinidades de outras ramificações, isso, principalmente nas redes sociais. Esses procedimentos deixam "rastros" que quando bem analisados se tornam poderosa ferramenta de análise comportamental e social. Em caso de utilização em tempo real, se configuram úteis inclusive para prevenção de tendências e suporte para tomada de decisões críticas, tanto governamentais quanto privadas. O processamento desta enorme quantidade de dados em velocidades aceitáveis exige soluções específicas, que já vêm sendo discutidas amplamente. Uma delas é o Big Data.

As empresas estão tendendo a constituir ou melhorar suas estratégias de negócios com base nessas informações coletadas. Nota-se que entender e utilizar os dados já coletados em conjunto com novas coletas seria essencial para se conhecer melhor as preferências do mercado. Contudo as soluções Big Data não se limitam apenas a isso, também se apresentam extremamente importantes para áreas críticas da sociedade, como por exemplo: a educação, a saúde, a segurança e os avanços científicos.

Implementações de Big Data para análise comportamental já vêm sendo aperfeiçoadas, haja vista o projeto da IBM de análise de "sentimentos" em *twitters*, utilizado na copa das confederações de 2013 ou as sugestões inteligentes de produtos por lojas virtuais. Porém, mesmo assim, enfrentam-se desafios complexos, ainda a solucionar.

O objetivo deste artigo é descrever e apresentar as tecnologias e técnicas utilizadas para a análise comportamental por meio de tecnologias Big Data e também, verificar as informações apresentadas como resultado dos processamentos desses sistemas.

2 REFERENCIAL TEÓRICO

A cada dia, em todo o mundo, se geram enormes volumes de dados, oriundos de fontes variadas. De acordo com estudo elaborado pela IBM em 2012, estima-se que se publiquem 2,5 quintilhões de dados diariamente na Web. Esse volume de dados não estruturado é gerado

por meio de postagens nas redes sociais, fóruns, portais de notícias, vídeos, coordenadas geográficas, fornecidas por tecnologias envolvendo o GPS, leituras de etiquetas RFID e geolocalização a partir do IP (CAMBOIM, 2014).

De acordo com Qmee (2014), essa massa de dados que é gerada na internet pode ser visualizada por meio de infográfico (pode ser acessado no sítio), apresentando em 60 segundos o demonstrativo das informações geradas a partir dos principais serviços da internet, representados pelo Google, Facebook, Twitter, dentre outras empresas. O trabalho revela que a cada 60 segundos são realizadas dois milhões de pesquisas no Google, no Facebook se geram 350 GB de dados, sendo distribuídos entre postagens, *likes*, fotos e vídeos. No Instagram, rede social para compartilhamento de fotografias, publica-se por minuto cerca de 3.600 fotos; no LinkedIn, rede social para profissionais, realizam-se onze mil consultas por minuto em perfis profissionais; e, na loja virtual da Amazon é obtido aproximadamente U$83.000 em vendas no mesmo tempo.

Mas, não é apenas as redes sociais ou a própria internet em si que são capazes de gerar informações plausíveis de processamento e análise, outros dispositivos físicos, como automóveis, aeronaves, rodovias, robôs, etiquetas de rádio frequência (RFID), dentre outros, também podem ser encarados como fontes de dados.

Estima-se que do início da civilização até o ano de 2013 a humanidade criou em torno de cinco hexabytes de dados, valor no presente se alcança em questão de dias (VILLELA, 2013).

2.1 O que fazer com os dados

Estima-se que o brasileiro gaste em média três horas e 41 minutos do seu dia navegando na internet, sendo que 26% das pessoas a acessam diariamente. Ainda de acordo com a Secretaria de Comunicação Social da Presidência da República (SECOM-PR, 2014), responsável pela pesquisa, 78% dos usuários que realizam o acesso pelo menos uma vez por semana à internet são jovens com idade entre 16 e 25 anos.

Alguns estudos publicados no The Econominst, sítio especializado em economia e política, sugerem que com o uso de programas *analytics* (inteligência de negócio), dotados de capacidade de processar grandes volumes de dados, se torna possível a aplicação de técnicas

para a previsão de comportamento humano em nível de população. Hoje se contam com dados envolvendo personalidade, gostos e afinidades dos usuários postados de maneira pública na internet, perfis estes que caso utilizados adequadamente auxiliam qualquer empresa a entender melhor o seu consumidor. Com esse entendimento do comportamento, as organizações podem eficientemente gerenciar as suas estratégias de negócios, melhorando assim sua competitividade junto ao mercado.

Encontram-se disponíveis ferramentas que visam cumprir a proposta de análise e previsão comportamental e de tendências, como exemplo o Windsom, que faz essa análise de dados a partir do Facebook.

2.2 Privacidade

Entre novembro e dezembro de 2014 o jornal britânico The Guardian e o americano The Washigton Post denunciaram uma rede de vigilância do governo dos EUA nas comunicações internacionais, fato que acabou aumentando a exposição e o debate sobre o assunto. Na Web é estipulado que a análise de massas de dados envolvendo mídias, pesquisas e redes sociais são suficientes para que se identifique um indivíduo.

Com o programa americano de espionagem chamado PRISM, que utiliza Métodos Sustentáveis de Integração de Projetos, agentes da agência de segurança nacional americana NSA (National Security Agence) teriam o acesso direto aos servidores de grandes empresas que atuam na Web, como: Google, Facebook, Microsoft, Apple e Skype, além de acesso aos dados da Verizon, uma das maiores empresas de Telecomunicações dos EUA. Nesse programa se coletam os dados de e-mails, históricos de navegação, mensagens de *chats* e arquivos que foram transferidos entre os usuários. Segundo o seu funcionamento, analisam-se esses dados em busca de suspeitos de crimes e de redes de terrorismo. Com isto ficou evidente para o mundo como todos estão expostos ao usar os serviços básicos, principalmente envolvendo a rede mundial de computadores, reforçando ainda mais as questões de privacidade dos usuários interligados.

A questão da privacidade é um dos problemas que empresas terão que lidar na exploração dos potenciais do Big Data, conforme Oliveira e Marcos (2014).

2.3 Veracidade das postagens

Outro paradigma ligado ao Big Data é quanto à veracidade das informações que são postadas nas redes sociais e na internet em geral. Estão em aberto as discussões sobre se os usuários das redes sociais realmente postam informações integralmente honestas sobre si mesmo ou o que desejam ou anseiam.

2.4 Definição de Big Data

Muitas vezes mal compreendido em sua essência, o termo Big Data está cada vez mais popular. Definem-se Big Data, em termos gerais, como imensos conjuntos de dados variados que, por aproveitamento melhor do processamento dessa massa de dados em busca da geração da informação, auxiliam as tomadas de decisões críticas. Para isso, exigem-se ferramentas e ambientes especialmente preparados para lidar com esses grandes volumes. Considerando que toda e qualquer informação nestes meios deverão ser encontradas, analisadas e aproveitadas em tempo hábil e a custo aceitável.

No mundo altamente competitivo e globalizado, toda informação se torna um diferencial expressivo. Na área dos negócios, geram-se milhares de dados a cada segundo, seja por transações financeiras, publicações em redes sociais ou por uma infinidade de outras ações. Se as instituições souberem como utilizar os dados que possuem ou que estão disponíveis, capacitam-se a melhorarem os seus produtos, a otimizarem as suas estratégias de marketing, a se tornarem mais produtivas, a evitarem desperdícios de recursos, a disponibilizarem serviços de maneira mais satisfatória para o cliente e com isto ultrapassar os seus concorrentes. Para isto, a instituição necessitaria de ferramentas capazes de traçar perfis comportamentais e padrões de sentimentos de seus clientes. A utilização dos conceitos de Big Data tornaria mais viável estas soluções.

A proposta de Big Data não se refere apenas a volume de dados, oferece abordagem ampla no tratamento desses dados que se originam nos variados meios. Nesses meios os dados se dividem em duas classes: os dados estruturados e os dados não estruturados, estes conhecidos como No SQL (Not Only SQL).

Visando esclarecer sua definição, especialistas resumem os aspectos em 3 Vs:

a) Volume;
b) Velocidade;
c) Variedade.

Acresce-se aos aspectos anteriores, dois fatores: veracidade e valor.

O atributo volume se refere à quantidade de dados relativamente grande, sejam não estruturados e/ou estruturados, possuindo ou não ligações entre si. É presenciado, neste momento, um forte aumento na quantidade de dados gerados e fornecidos. Aplicativos empresariais tiveram um crescimento anual de 60% com relação à entrada de informações. É estimado que em uma empresa com mil funcionários se gerem anualmente 1.000 terabytesde dados. A tendência para os próximos anos é de aumento. Calcula-se que até 2020 o volume de informação cresça cerca de cinquenta vezes. Dentre as diversas fontes de dados responsáveis por estes altos índices se encontram as redes sociais, textos em geral, leitores de RFID, imagens, vídeos, entre outras.

O quesito velocidade abrange a necessidade da obtenção e processamento dos dados em grandes volumes efetuados em tempo hábil, isso para seja realmente implementada solução viável. É consenso, dentre os pesquisadores de Big Data, que a realização destas ações seja feita em tempo real ou muito próximo disto. Para realizar analogia do quanto a velocidade do processamento influencia o mercado, é possível analisar o quanto é desconfortável qualquer estorvo na hora de passar o cartão de crédito em estabelecimento comercial. Cada minuto a mais no tempo aguardando o resultado da operação passa a ser um dos fatores determinantes para a volta ao local ou a utilização do serviço da operadora novamente. E a situação poderia piorar ainda mais se ao invés de minutos as operações transacionais com cartão demorassem horas para a aprovação.

O outro aspecto considerado é a variedade, que como o próprio nome diz, refere-se à diversidade dos dados, e também, à capacidade de tratar e relacionar os variados tipos de dados existentes. Para análises cada vez mais coerentes é essencial se obter volumes de dados bem diversificados quanto à informação. A variedade de dados deve ser interpretada como

pertinente a sua forma de coleta e tipo, sendo: estruturado, não estruturado, imagens, áudios, vídeos etc.

Considerando os fatores acrescentados por parte dos pesquisadores da área a veracidade dos dados é importante, pois não seria útil ter uma estrutura implementada para suportar grande volume de dados variados, obtidos e processados em tempo real, caso os dados ou as informações geradas como resultados não fossem confiáveis. Por isto são estudados processos e técnicas que reflitam e garantam ao máximo a possível consistência destes dados.

O último fator é o resultado da união de todos os demais: o valor. O Big Data precisa gerar valor para quem o adota, se isto não está sendo alcançado, o investimento está sendo comprometido.

O Big Data une diversas ferramentas e técnicas para extração de informações e conhecimentos úteis para as instituições que o implementarem seguindo os seus fundamentos. Por isso o assunto vem cada vez mais ganhando espaço por todo mundo.

2.5 Introdução às ferramentas e aos paradigmas utilizados em Big Data

O planejamento de quais ferramentas e modelos seguir ou adaptar para uma solução especifica de Big Data depende do tipo da necessidade e do contexto (MYSORE; KHUPAT; JAIN, 2013).

Para realizar uma compreensão e estudo mais aprimorado para a adoção de um "tipo" de Big Data é recomendado levar em conta algumas características que auxiliam no processo de compreensão do ambiente (veja Figura 1), destacando-se, como principais:

a) Forma como os dados são adquiridos (máquinas, transações, imagens, vídeos, áudios, sensores etc.);
b) Como os dados são processados;
c) Frequência de disponibilização de novos dados. Destaca-se que cada informação se diferencia a partir de como é adquirida, ou seja, conforme sua fonte. Em geral os dados costumam ser classificados quanto ao formato do conteúdo, tipo

(exemplo: transações; *logs*; históricos etc.), frequência de disponibilização, objetivo e necessidade de velocidade de processamento.

Basicamente a implementação "simplista" do Big Data se resume em partes de infraestrutura, para armazenamento e processamento de uma enorme gama de dados, e em analítica, apoiada em tecnologias como Hadoop e MapReduce, percussores de novas e mais fáceis técnicas de programação destinadas ao trabalho de armazenamento de dados estruturados ou não e seu processamento distribuído (CHEDE, 2012).

Nesses contextos se ressaltam os bancos de dados No SQL (Not Only SQL) para o tratamento adequado de dados não estruturados (não seguindo padrões e conceitos de base de dados Relacionais).

Há ainda os projetos complexos envolvendo tecnologias de Stream Computing, implementado onde processamentos em lote não são suficientes, exigindo execuções real time.

Na Figura 1, mostram-se a classificação e a estrutura do Big Data em todos os níveis.

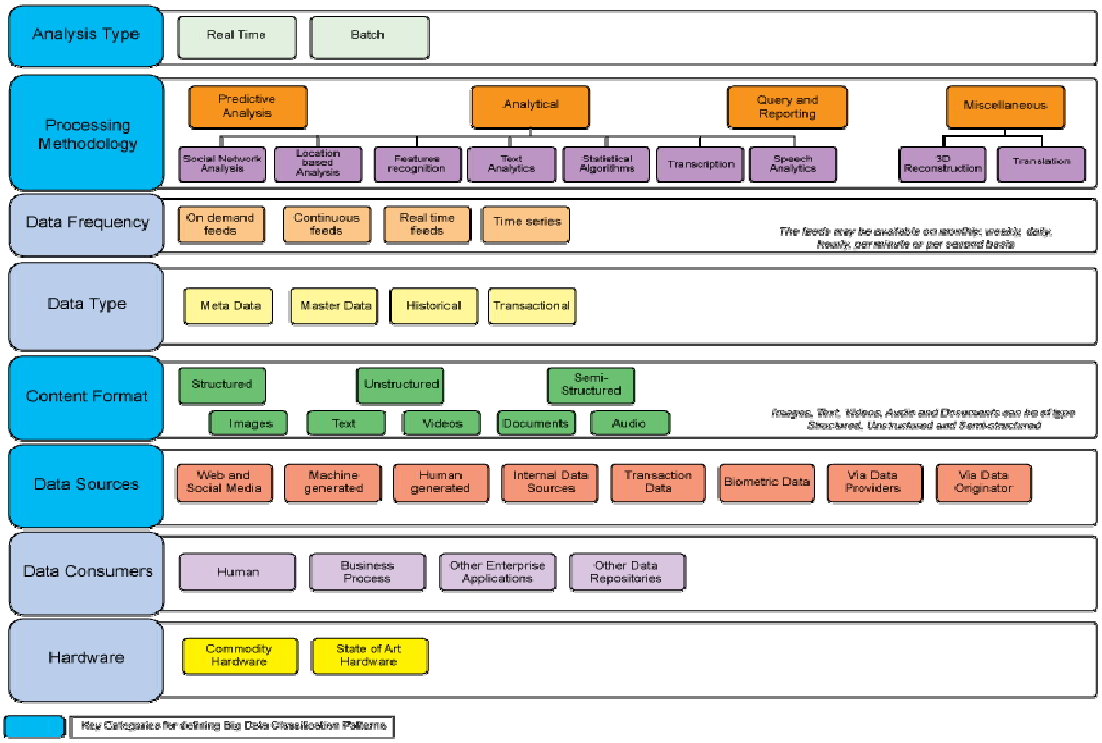

Figura 1 - Classificação de Big Data
Fonte: IBM

Na sequência, as definições sobre alguns dos principais componentes que atualmente constituem a solução Big Data, de acordo com Saracco, 2011.

2.5.1 Hadoop

Uma das principais estruturas utilizadas em implementações de Big Data, em versões modificadas ou não, o Hadoop é o sistema que permite que os dados sejam processados em alto desempenho de forma distribuída, ou seja, é multitarefa e possibilita o processamento de imensas quantidades de dados ao mesmo tempo, para finalidades e usuários diferentes, sendo ainda tolerante a falhas.

O Hadoop inclui o módulo ligado a MapReduce (modelo de programação apresentado pelo Google), onde os dados de entrada são divididos em conjuntos distintos, sendo tratados de forma independente um do outro, a partir de tarefas de mapeamento (*map tasks*). As saídas dos *map tasks* são enviadas para serem utilizadas como entradas das tarefas de redução (*reduce tasks*), sendo que, normalmente, tanto as entradas quanto as saídas das tarefas são armazenadas, considerando que a própria ferramenta se encarrega de monitorar e de se autorecuperar caso ocorram falhas (Figura 2).

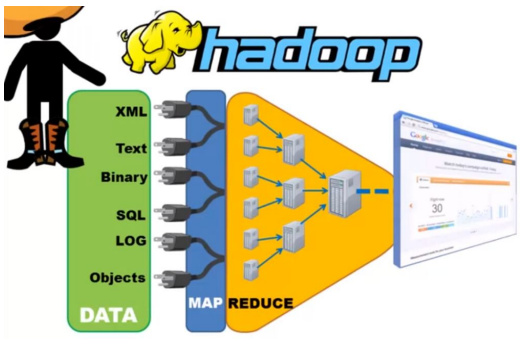

Figura 2 - MapReduce
Fonte: INTRICITY

Além disso, é importante ressaltar que o Hadoop trabalha sob sistema de arquivos próprio, baseado no Google File System, chamado de HDFS (Hadoop Distributed File System), tendo o seu foco na confiabilidade, usabilidade, desempenho e alto gerenciamento. Simplificando: possui sistema de arquivo distribuído, onde um dado seria distribuído e replicado entre os nós da estrutura montada.

Sua definição bruta disponibilizada na própria página do projeto, em tradução livre:

> Framework que permite processamento distribuído de imensa quantidade de dados através de clusters de computadores utilizando um único modelo de programação. Foi arquitetado para de um único servidor replicar a tarefa para milhares deles, que por sua vez disponibilizam recursos de processamento e armazenamento. Ao invés do hardware comandar a alta disponibilidade, a biblioteca por si só é desenhada para detectar e lidar com as falhas na camada da aplicação, e assim gerenciando o serviço de alta disponibilidade no top de um conjunto de computadores, todos poderão estar propensos a falha. (HADOOP, 2014).

HBase

Muitas vezes denominado como "Hadoop Database", pois até 2010 era um subprojeto do Hadoop, o HBase é uma solução de armazenamento de dados No SQL, ou seja, um banco de dados para dados não estruturados e parcialmente estruturados. É usado como fonte de trabalho MapReduce, assim como o Bigtable do Google. Além disso, possui alta integração com Hadoop e o seu sistema de arquivos, para armazenamento. Executa-se o HBase em três modos de trabalho: autônomo, pseudo-distribuído e completamente distribuído.

Hive

Ferramenta de *data warehousing* desenvolvida para o ambiente Hadoop, sua finalidade é extrair, transformar, carregar e consultar dados em arquivos no Hadoop. Trabalha com conceitos semelhantes aos de banco de dados relacional, considerando tabelas, colunas e partições. Também possui uma espécie de "SQL" para dados não estruturados, o HiveQL. Compilam-se as consultas do HiveQL para processamento como tarefas MapReduce do Hadoop.

Oozie

O Oozie é um gerenciador de tarefas do Hadoop, efetuando o agendamento e controla o fluxo de trabalho da execução das tarefas.

ZooKeeper

Funciona como fornecedor de serviço centralizado, que mantém informações de configuração e nomenclatura, além de prover sincronização distribuída e serviços de grupo. No geral seu objetivo é gerenciar a complexidade dos diversos serviços em execução.

Splunk

O Splunk é um software corporativo gratuito que objetiva o uso simplificado, referente ao trabalho e tarefas sobre Big Data. Está sendo amplamente utilizado para obter a inteligência operacional que aprofunda a compreensão dos negócios, melhora o serviço e o tempo de atividade, reduz os custos e os riscos contra a segurança digital.

Cassandra

Software de gerenciamento de banco de dados de código aberto que trata os dados estruturados e os não estruturados. Sua característica marcante é a escalabilidade e a alta disponibilidade fornecida, sem comprometer o desempenho, além da tolerância a falhas.

Flume

Serviço confiável, distribuído e disponível, objetiva mover de forma eficiente enormes quantidades de dados, conforme são produzidos. É recomendado em operações que visam o deslocamento de *logs* de diversos sistemas para o HDFS (Hadoop Distributed File System).

Avro

Tecnologia que realiza serialização dos dados utilizando JSON (subconjunto JavaScript) para definição dos tipos e protocolos. O Avro é conhecido pela compactação binária que efetua nos resultados de sua serialização.

Pig

Plataforma que fornece linguagem de alto nível para a realização de programas que efetuam análise de dados. Sua forma de trabalho é converter, por meio do seu compilador, os programas expressos pelo desenvolvedor em sequências de tarefas MapReduce, permitindo execução otimizada no Hadoop.

JAQL

Desenvolvida pela IBM, a JAQL é linguagem de consulta, semelhante ao SQL. Apesar de ser bem útil para dados estruturados, sua interface foi construída para trabalhar bem com o HBase.

Lucene

O projeto Lucene é uma API Java para indexar e procurar dados de texto, e é bem visto pelo alto desempenho e pela escalabilidade que possui para executar as tarefas a que se propõe, por intermédio de algoritmos de busca eficientes e poderosos.

BigSheets

Com interface parecida a planilhas, o BigSheets realiza descobertas e explorações de dados e conhecimento de forma fácil, tendo como ideia permitir que usuários da área de negócios façam as tarefas de análise e controlem o Hadoop.

Analítica de Texto

Biblioteca do BigInsights da IBM que facilita o processo de análise de textos, fornecendo conjunto de ferramentas para auxiliar no desenvolvimento de aplicações relacionadas a proposta.

Stream computing

Considerado um novo paradigma para o Big Data quanto à variedade dos dados, aborda soluções que permitem que os volumes de dados sejam processados em fluxo, objetivando a execução da aplicação em tempo real (SAKR, 2013).

Para isso, se executam consultas contínuas, contrariando desta forma o padrão amplamente aceitável e trabalhado atualmente: execução de consulta sobre um amontoado de dados para coletar resultados. É realizada a consulta antes de se ter necessariamente uma determinada quantidade de dados, sendo assim, permite-se que o fluxo contínuo de dados influencie a formação do resultado apresentado (Figura 3).

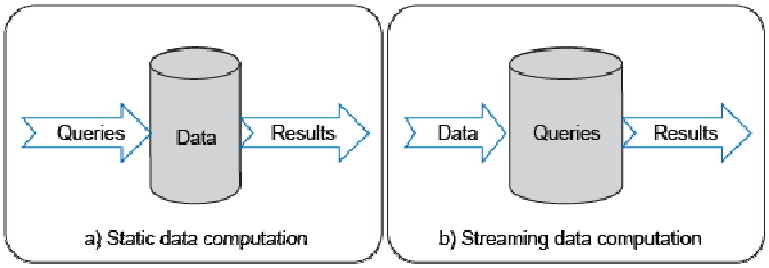

Figura 3 - Computação estática X Computação de fluxo
Fonte: IBM

2.5.2 Análise de sentimentos

Objetivando *insights* sobre visões, opiniões e sentimentos sobre determinado tema, marca, produto, serviço, evento ou pessoa por meio de ferramentas computacionais, a análise de sentimentos é o processo que examina os conteúdos de forma que possa dar uma resposta plausível quanto aos itens examinados, realizando a leitura dos mesmos.

O grande desafio computacional a análise de sentimentos está em torno da precisão de seus resultados, pois o exame se torna complexo quando se leva em conta fatores de extrema importância para o pleno entendimento, como é observado ao se deparar com frases e/ou palavras com significados ambíguos ou irônicos (tendo significados diferentes de acordo com o contexto em que foi dita), além das gírias ou dialetos (MANFROI, 2014).

Para exemplificar a complexidade, a Figura 4 apresenta uma mensagem do Twitter.

Figura 4 - Mensagem do Twitter

Na mensagem o autor expressa por intermédio de ironia seu descontentamento com a sua conexão de internet, escrevendo sobre a desconexão da rede no momento em que o arquivo em processo de download já está quase totalmente transferido, mas que por conta da falha no serviço de internet haverá a necessidade de baixar o arquivo novamente, desde início. Uma análise computacional que interpretasse unicamente a frase e seus adjetivos poderia entender isto como um *insight* positivo para o provedor de serviço de internet, guiando os relatórios para falsos positivos, não sendo, portanto, solução adequada para o contexto.

Modelos de análise não bem planejados também acabam apresentando "efeitos colaterais" indesejáveis, conforme discussão iniciada no caso relatado por Mirvish (2011), onde os trabalhos, prêmios conquistados e ações realizadas pela atriz Anne Hathaway, quando abordadas nas mídias, supostamente influenciam os "robôs" que auxiliam na valorização da empresa cujo nome é Berkshire Hathaway, sendo que a empresa e a artista não se relacionam.

A automação das tarefas de extrações de análises requer que o *software* executor desta funcionalidade tenha limites bem definidos e saiba tratar as ambiguidades, tentando eliminá-las (ZADROZNY e KODALI, 2013).

2.5.2.1 Parte técnica da análise de sentimentos

Existem técnicas que abordam fórmulas probabilísticas e documentos explicando técnicas de marketing em relação às análises de sentimentos, que resumidamente, seria a "categorização" do texto em polaridades bem definidas ou a amplitude delas.

Basicamente a polaridade é dividida em:

a) Positiva;
b) Neutra;
c) Negativa.

Já a amplitude de polaridade é a escala em que seu valor define a proximidade de uma das polaridades, localizada em uma das extremidades.

A classificação da polaridade é feita em diversas partes e camadas, como por exemplo, realizada por sentença, palavra ou por documento. Algoritmos especializados nessas classificações geralmente utilizam a polarização das camadas mais baixas como entrada para definição das camadas mais acima, ou seja, utilizam a polarização de palavras para classificar frases, a partir da classificação das frases classificam as sentenças e parágrafos, com a classificação das sentenças é feito a polarização do documento. Como notado, o processo possui certa complexidade envolvida.

A extração de sentimentos é obtida por meio de diversos algoritmos combinados, sendo os principais os baseados em:

a) Frequência: Nesta técnica as palavras que mais aparecem no documento refletem na classificação da polaridade.
b) Presença do termo: Polariza o sentimento do recurso analisado de acordo com as palavras únicas.
c) N-grams: Realiza análise tentando entender o contexto da palavra dentro do documento.
d) Sintaxe: incorporam análise sintáxica dentro das configurações de recursos.
e) Partes de discurso: Quebra uma sentença em diferentes partes, de acordo com a sua classificação gramatical: pronome; adjetivo; verbo; substantivo e etc. Após a classificação a polaridade é feita com base nos adjetivos extraídos.

Ademais, ainda se faz necessário a polarização que considere domínios e contextos para apresentação de resultados mais eficazes e coerentes. Para isto se utilizam algumas técnicas para obtenção da classificação a este nível, como os algoritmos de análise:

a) Léxica: Fornece a polaridade para cada palavra.
b) Aprendizagem: A polaridade da palavra é definida conforme uma série de fatores estipulados pelo algoritmo, sendo que, o sistema aprende o que fazer conforme classificações anteriores, ou seja, aprendizagem de máquina.

2.6 Big Data com IBM® InfoSphere® BigInsights™

O InfoSphere BigInsights é a plataforma desenvolvida pela IBM englobando o conjunto de *softwares* que possibilitam a realização de análises de grandes volumes de dados. Esta solução vem sendo amplamente utilizada no meio corporativo para alcançar implementação do Big Data. O BigInsights inclui uma versão modificada do Hadoop além de outras ferramentas de Big Data, como o Hive e o Pig. Seu diferencial está em facilitar a utilização do Hadoop e o desenvolvimento de aplicações Big Data. Pode-se dizer que esta plataforma da IBM é melhoramento das tecnologias *open source* da área, provendo interfaces mais amigáveis para soluções complexas (ver Figura 5).

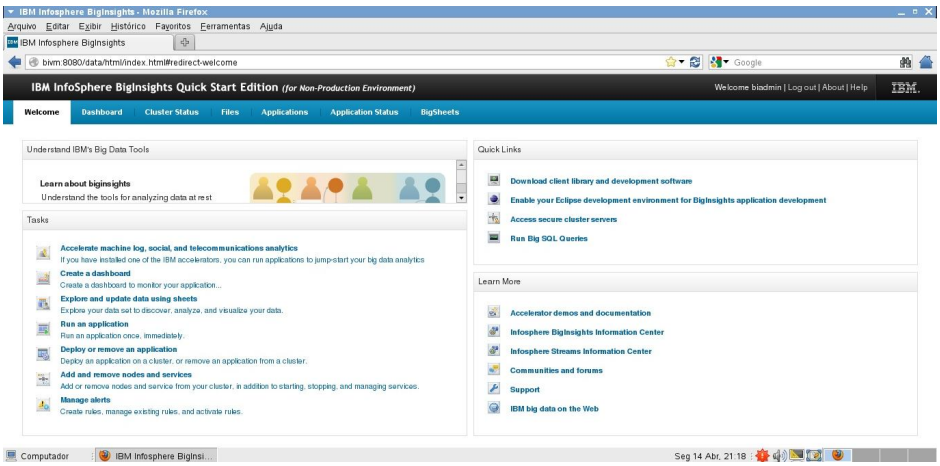

Figura 5 - Página principal do Web Console do BigInsights

Um dos *softwares* incluídos na plataforma da IBM que vem ganhando grande destaque é o BigSheets (ver Figura 6), como ressaltam Saracco e Dawra (2012), permitindo que não apenas os programadores manipulem, explorem ou visualizem dados armazenados no sistema, além disto a solução conta com poderosas funções e macros para extrair *insights*, e também é capaz de criar gráficos e exportar informações.

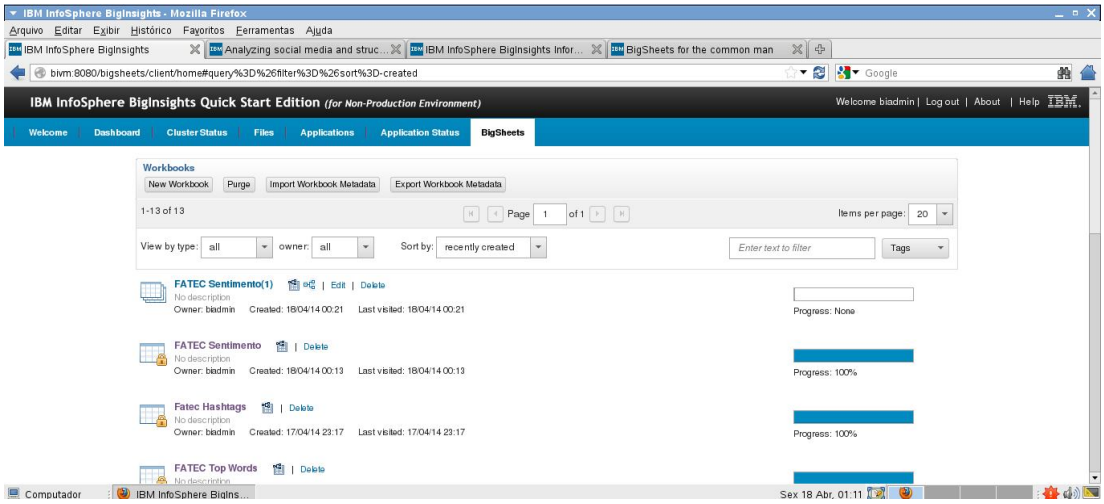

Figura 6 - Tela principal do BigSheets no BigInsights

2.6.1 Aplicando técnicas de análise comportamental com o BigSheets

A utilização da interface Web da BigSheets possibilita as extrações comportamentais e sentimentais a partir de funções e macros com algoritmos capazes de obter dos dados em análises de menções a nomes, endereços, organizações, e-mails, localizações, número de

telefones, etc. Além disso possui grande suporte a gráficos, para representação dos dados, permitindo melhor análise dos dados.

É comum que os arquivos analisados pelo BigSheets passem por camadas de extrações, como as feitas por aplicações produzidas em AQL (Annotation Query Language). As aplicações normalmente combinam visualizações para a obtenção dos resultados, sendo que, estas visualizações utilizam expressões regulares, apoiam-se nas APIs já existentes, como a LanguageWare para análise de textos, dicionários de dados contendo palavras de um determinado domínio ou contexto, dentre outros, que auxiliam na realização de consultas mais eficazes, conhecidas por sua capacidade de realizar análise textual em busca de padrões e informações relevantes para o projeto em que está trabalhando o ambiente de Big Data.

Exemplificando, a partir de arquivo contendo mensagens do Twitter, redes sociais e blogs, onde consta a palavra FATEC, (Figura 7). Se considerado o potencial Big Data, esta quantidade não seria digna de catalogada como uma solução de grande volume, uma vez que o arquivo contém poucos kilobytes, enquanto a análise de grande volume geralmente processa petabytes. O intuito, no entanto, foca-se na apresentação da aplicação da técnica em si, podendo ser replicada em ambientes variados, independente do tamanho da massa de dados trabalhada.

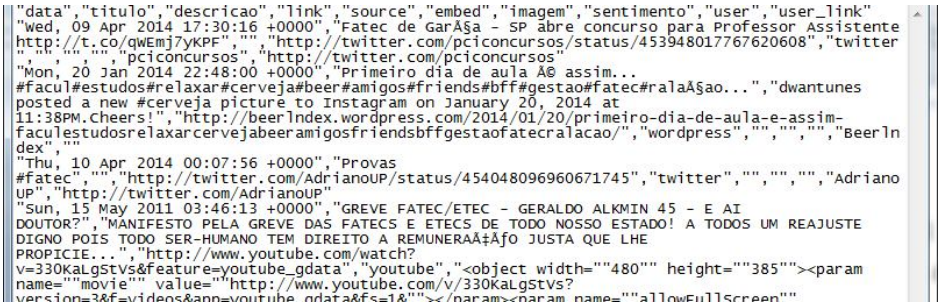

Figura 7 - Parte do arquivo de amostragem

Com o Workbook "pai" (master) criado, se aplicável, pode ser realizado algumas operações quanto aos dados criando um Workbook "filho" (child): a partir deste é permitido a realização da junção/combinação de mais de um Workbook, agrupamento dos dados, criação de colunas que têm seus preenchimentos considerando determinadas regras, como por exemplo, a função básica:

IF(SEARCH('SEBRAE*', #user) >0, 'SEBRAE', #usuario1)

A fórmula procura na coluna *user* (campo indicado com o prefixo '#') todos os valores iniciados com a palavra SEBRAE e unificam os mesmos na coluna usuario1. O asterisco (*) foi colocado após a palavra SEBRAE, mas ainda dentro das aspas simples, para garantir que todas as variações de SEBRAE sejam interpretadas como um único usuário, sendo assim, as postagens dos usuários SEBRAE-SP, SEBRAE-RJ ou SEBRAE-MG seriam consideradas como unicamente sendo do SEBRAE.

O BigSheets oferece grande variedade de formas para explorar as Workbooks, existem diversas funções e macros já desenvolvidos para suportar determinadas tarefas de extrações, tendo como categorias principais: *entities* (entidades); xml; math (matemática); html; *datetime* (data e hora); *text* (texto); e url.

Uma vez que os dados estejam adequados para sua finalidade é comum converter a análise processada em gráficos, simplificando desta forma a visualização para obter *insights*. O BigSheets conta com diversos tipos de gráficos, sendo que um é o mais indicado para atender a uma determinada apresentação de informação.

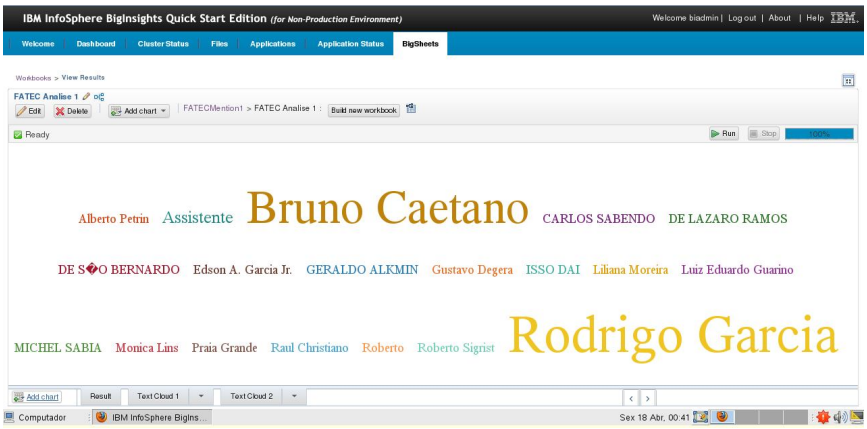

Figura 8 - Nomes citados em twitters contendo a palavra FATEC, extraídos pelo macro de entidade "Person"

Além dos gráficos tradicionais (barras horizontais e verticais, áreas, linhas etc.), as soluções disponibilizam gráficos melhorados, como o Text Cloud, que gera lista de *tags* (etiquetas contendo *links*) Em casos onde se necessitam demonstrações de valores baseados

em áreas geográficas, existe a possibilidade da geração de gráficos em formato mapa (Figura 9).

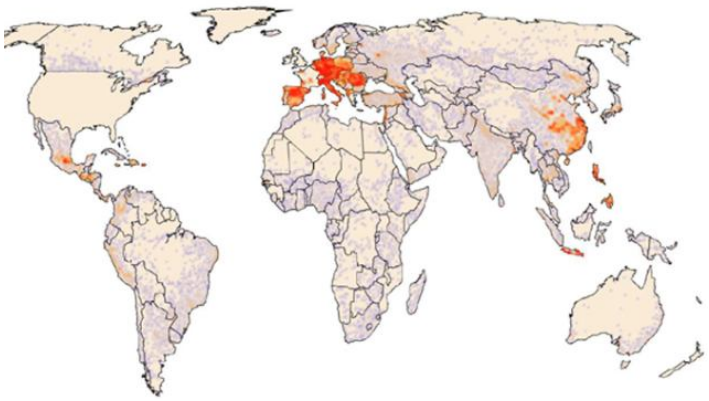

Figura 9 - Gráfico de análise geográfica

3 METODOLOLGIA

A metodologia adotada neste artigo se baseia na revisão bibliográfica, descrição de componentes do Big Data, coleta de amostras, análises de casos e análises qualitativas.

4 DISCUSSÃO E RESULTADOS

Para análise comportamental no contexto social se dá enfoque à plataforma InfoSphere BigInsights da IBM, mais especificadamente à solução BigSheets.

Em termos gerais o BigSheets atende como tecnologia de implementação da linha de pesquisa de grandes volumes de dados, permitindo de forma eficiente e intuitiva que se realizem análises comportamentais da amostragem de mensagens/postagens do Twitter, redes sociais e blogs contendo a palavra FATEC.

Alguns dos *insights* obtidos a partir da execução das funcionalidades existentes no ambiente, os nomes de pessoas (processamento feito por meio da chamada Function/Entities/Person) mais citados nas mensagens/postagens foram: Bruno Caetano e Rodrigo Garcia, sendo que, a cidade ou o estado mais mencionado foi São Paulo.

Realizando pesquisa mais profunda para procurar entender a causa dos nomes Bruno Caetano e Rodrigo Garcia terem sidos mais citados nas mensagens da amostragem, verificou-se que Bruno Caetano (Diretor Superintendente do SEBRAE) apresentou palestra na comemoração de 7 anos da FATEC Zona Sul no dia 4 de abril de 2014. O nome Rodrigo Garcia (secretário do governo nas áreas de Desenvolvimento Econômico, Ciência, Tecnologia e Inovação) anunciou a construção de nova instituição da FATEC na cidade de Ferraz de Vasconcelos, fato amplamente repercutido na mídia em geral.

Figura 10 - *Hashtags* mais usadas em mensagens contendo a palavra FATEC

Outro conhecimento obtido pela análise da amostragem foi a indicação de que as 3 *hashtags* (palavras contidas em mensagens do Twitter, com o prefixo #) mais utilizadas pelos usuários são #Cerveja, #FATEC e #Amigos. A Figura 10 mostra a referência gráfica (Burbble Cloud) da visualização desses resultados.

Utilizando o BigSheets seria ainda possível a descoberta de muitos mais conhecimentos sobre o comportamento dos usuários que mencionam ou tenham alguma relação com a instituição, além de realizar a classificação da polarização sentimental (das menções) em positiva, neutra ou negativa.

A solução de Big Data quando configurada em ambiente de produção explorando os grandes volumes de dados provindos de variadas fontes em tempo aceitável se torna ferramenta essencial para as tomadas de decisões e o alcance de relevantes *feedbacks*.

Aproveitando a ascensão da computação em nuvem é importante ressaltar que existem sítios especializados na análise de grandes massas de dados em mídias sociais, aplicando-se Big Data, como: Social Mention (www.socialmention.com) e Sentiment140 (www.sentiment140.com).

CONCLUSÕES

Este artigo objetivou descrever e apresentar as tecnologias Big Data utilizadas para análise comportamental, identificando contexto social, técnicas chave e paradigmas enfrentados no processo de extração de *insights*.

Entende-se que as tecnologias de Big Data representam umas das mais importantes soluções para a obtenção e a análise das informações indispensáveis para as áreas de negócios, considerando que atualmente se geram cada vez mais dados de diversos tipos, formas e dispositivos.

Até a pouco tempo atrás, muitas empresas quando necessitavam de *feedback* ou *input* para suas tomadas de decisões recorriam a seus dados recolhidos por intermédio de pesquisas, opiniões publicadas por especialistas do segmento em revistas e sítios, fóruns em geral, visões de familiares e conhecidos, além de alguns dados internos da companhia. No presente, as análises comportamentais vêm sendo exploradas mais ativamente pelas empresas, dando apoio às tomadas de decisões, tendo ainda, mais outro fator a ser considerado: a opinião dos clientes fornecidas na internet e meios diversos, obtidas e acessadas velozmente.

Assim, as organizações ganham vantagens competitivas, permitindo além de observar as opiniões dos seus próprios produtos e serviços, mas também verificar avaliações sobre os seus concorrentes.

Com o BigInsights, as tarefas de administração, desenvolvimento, extração e visualização de informações se tornam tarefas intuitivas e fáceis, por conta do amplo suporte e complementos existentes no pacote. Com o *software* BigSheets, incluso na plataforma do InfoSphere BigInsights, não é exigido vasta experiência em Hadoop ou tecnologias relacionadas ao Big Data para que se consiga explorar o potencial do ambiente.

Ressalta-se, finalmente, que as técnicas e conceitos tratados introduzem aos principais usos da tecnologia de Big Data, e podem ser replicados em diversos ambientes e plataformas diferentes, sejam de licença proprietária ou livre.

A parte prática se limitou a uma única ferramenta, escolhida por ser uma das mais intuitivas, mas que atendeu suficientemente ao proposto.

REFERÊNCIAS

CAMBOIM, S. **Das ilhas de informação geográfica ao Big Data**. Disponível em: http://mundogeo.com/blog/2013/06/05/das-%E2%80%9Cilhas%E2%80%9D-de-informacao-geografica-ao-big-data/. Acessado em: junho de 2014.

CHEDE, C.. **Você realmente sabe o que é Big Data?** Disponível em: https://www.ibm.com/developerworks/community/blogs/ctaurion/entry/voce_realmente_sabe_o_que_e_big_data?lang=en. Acessado em: março de 2014.

JONES, T.. **Planejamento no Hadoop**. Disponível em: http://www.ibm.com/developerworks/br/library/os-hadoop-scheduling/. Acessado em março de 2014.

MANFROI, L. Psicologia **e Comportamento nas Redes Sociais: da Web 1.0 à Big Data**. Disponível em: http://pt.slideshare.net/lucmanfroi/psicologia-e-comportamento-nas-redes-sociais-da-web-10-big-data. Acesso em: março de 2014.

MYSORE, D; KHUPAT, S; JAIN, S. **Arquitetura e Padrões de Big Data, Parte 1: Introdução à Classificação e à Arquitetura de Big Data**. Disponível em: http://www.ibm.com/developerworks/br/library/bd-archpatterns1/. Acessado em: novembro de 2014.

OLIVEIRA D; MARCOS S. **Dados, dados e mais dados: o fenômeno Big Data**. Disponível em: http://ulbra-to.br/encena/2013/10/17/Dados-dados-e-mais-dados-o-fenomeno-Big-Data. Acessado em: outubro de 2014.

QMEE Service and Technology. **What happens online in 60 seconds?** Disponivel em: http://blog.qmee.com/qmee-online-in-60-seconds/. Acessado em março de 2014.

SAKR, S. **Uma Introdução ao InfoSphere Streams**. Disponível em: http://www.ibm.com/developerworks/br/library/bd-streamsintro/. Acessado em março de 2014.

SARACCO, C. **Understanding InfoSphere BigInsights: An introduction for software architects and technical leaders**. Disponível em: http://www.ibm.com/developerworks/data/library/techarticle/dm-1110biginsightsintro/index.html. Acessado em: outubro de 2014.

SARACCO, C; DAWRA, A. **Analisando a Mídia Social e os Dados Estruturados com o InfoSphere BigInsights**. Disponível em: http://www.ibm.com/developerworks/br/data/library/techarticle/dm-1206socialmedia/. Acesso em junho de 2014.

WEBINSIDER.. **Big Data: a informação e o seu alto valor estratégico**Disponível em: http://webinsider.com.br/2013/04/02/big-data-a-informacao-e-o-seu-alto-valor-estrategico/. Acessado em: março de 2014.

ZADROZNY, P; KODALI, R. **Big Data Analytics Using Splunk: Deriving Operational Intelligence from Social Media, Machine Data, Existing Data Warehouses, and Other Real-Time Streaming Sources.** Nova Iorque : Apress, 2013.