UNIVERSIDADE PAULISTA

FLÁVIO ROCHA VALENÇA
RAFAEL DE SOUZA BATISTA
WESLLEY DO ESPIRITO SANTO

BIG DATA

SANTOS
2015

FLÁVIO ROCHA VALENÇA
RAFAEL DE SOUZA BATISTA
WESLLEY DO ESPIRITO SANTO

BIG DATA

Trabalho para apresentação da matéria de Sistemas Distribuídos apresentado a Universidade Paulista – UNIP.

Orientador: Especialista Vinícius Eduardo Ferreira dos Santos Silva

SANTOS
2015

FLÁVIO ROCHA VALENÇA
RAFAEL DE SOUZA BATISTA
WESLLEY DO ESPIRITO SANTO

BIG DATA

Trabalho para apresentação em Sistemas Distribuídos apresentado a Universidade Paulista – UNIP.

Aprovado em:

BANCA EXAMINADORA

_____ __/__/___

Especialista Vinícius Eduardo Ferreira dos Santos Silva

Universidade Paulista – UNIP

RESUMO

O grupo apresentará nesse trabalho, conceitos, aplicações e a importância do Big Data nos dias de hoje para área de TI e para o mundo. Para realizar esse trabalho o grupo fez diversas pesquisas sobre o Big Data e tudo o que gera ao seu redor. Junto com o Big Data, mostraremos o conceito de 5V que são: volume, velocidade, valor, variedade e veracidade. Traremos curiosidades sobre o zettabytes, que é uma unidade de medida computacional capaz de armazenar grande volume de dados. No trabalho, será mostrado o Hadoop, uma plataforma desenvolvida para análise de grande volume de dados, sejam eles estruturados ou não. Informações sobre mercado de trabalho, para quem é adequado usar esse recurso, como usá-lo da melhor maneira. Esse é o nosso resumo e objetivo do grupo, mostrar um pouquinho do que é, do que foi, e o que será do Big Data para os próximos anos, com certeza muito promissora e melhor aperfeiçoada.

ABSTRACT

The group will present this work, concepts, applications and the importance of big data these days for IT and for the world. To accomplish this work the group has done several studies on Big Data and everything that generates around. Along with the Big Data, show the concept of 5V which are: volume, speed, value, variety and veracity. We will bring fun facts about zettabytes, which is a unit of measure computer capable of storing large amounts of data. At work, Hadoop, a platform developed for analysis of large volumes of data, whether structured or not will be displayed. Information on the labor market, for whom it is appropriate to use this feature, how to use it the best way. This is our summary and purpose of the group, show a little of what is, what was and what will be the Big Data for years to come, with very promising and best improved certainty.

LISTA DE ILUSTRAÇÃO

LISTA DE ABREVIATURAS E SIGLAS

SIM Subscriber Identity Module (Módulo de identificação do assinante)
IBM International Business Machines (Máquinas de Negócio Internacionais)
GPS Global Positioning System (Sistema de posicionamento global)
GFS Google File System (Sistema de arquivos Google)
NDFS Nutch Distributed File System (Sistema de arquivos distribuído Nutch)
HDFS Hadoop Distributed File System (Sistema de arquivos distribuído Hadoop)
HD Hard Disk (Disco Rígido)
NoSQL Not Only Structured Query Language (Não somente Linguagem de Consulta Estruturada)
MPP Massively Parallel Processing (Processo Paralelo Massivo)

SUMÁRIO

1 INTRODUÇÃO

O presente trabalho tem o objetivo apresentar o quanto importante é o conceito de big data, do qual hoje estamos todos envolvidos e nem nos damos conta. Trazendo algumas curiosidades e como é o seu funcionamento.

Esperamos poder contribuir para o aprendizado dos leitores e com isso quem sabe despertar um novo profissional.

2 DESENVOLVIMENTO

2.1 O que é big data?

Big data é um termo utilizado para descrever grandes volumes de dados e que ganha cada vez mais relevância à medida que a sociedade se depara com um aumento sem precedentes no número de informações geradas a cada dia. As dificuldades em armazenar, analisar e utilizar grandes conjuntos de dados tem sido um considerável gargalo para as companhias. Os volumes de informação digital vêm aumentando consideravelmente, em 2011 (1,7 *zettabytes*[1]), 2012 (2,7 *zettabytes*) e em 2015(8 *zettabytes*).

2.2 Curiosidade sobre o zettabytes

Atualmente, cerca de 15 *petabytes* de dados estruturados e não estruturados são gerados todos os dias. Entre eles destacam-se vídeos, comentários em redes sociais, conteúdos, de blogs e dispositivos móveis. 1ZB(*zettabyte*) equivalem a 1 bilhão de HDS[2] iguais a de um *desktop*, 75 milhões de *iPads* de 16 GB, 0,5ZB equivalia a toda a internet de 2009, 42ZB seria o tamanho aproximado total, se todas as palavras ditas pela humanidade em toda a sua historia, fosse digitadas.

2.3 Conceito de 5V

2.3.1 Volume

Estamos falando de quantidades de dados realmente grandes, que crescem exponencialmente e que, não raramente, são subutilizados justamente por estarem nestas condições. Estima-se que esse volume dobre a cada 18 meses.

2.3.2 Velocidade

Para dar conta de determinados problemas, o tratamento dos dados deve ser feito em tempo hábil, muitas vezes em tempo real. Se o tamanho do banco de dados for um fator limitante, o negócio pode ser prejudicado: imaginamos, por exemplo, o transtorno que uma operadora de cartão de crédito teria e causaria, se demorasse horas para aprovar uma transação de um cliente pelo fato de o seu sistema de segurança não conseguir analisar rapidamente todos os dados que podem indicar

[1] Medida de armazenamento

[2] É uma sigla inglesa cujo significado Hard Disk ou traduzido para Disco Rígido

uma fraude.

2.3.3 Variedade

Os volumes de dados que temos hoje são consequência também da diversidade de informações. Temos dados em formato estruturados, isto é, armazenados em bancos como *PostgreSQL* e *Oracle*, e dados não estruturados oriundos de inúmeras fontes, como documentos, imagens, áudios, vídeos e assim por diante. É necessário saber tratar a variedade como parte de um todo, por exemplo, um tipo de dado pode ser inútil se não for associado a outros.

2.3.4 Veracidade

Também temos que considerar, pois não adianta muita coisa lidar com a combinação de volume, velocidade e variedade se houver dados não confiáveis. É necessária que haja processos que garantam o máximo possível a consistência dos dados. Voltando ao exemplo da operadora de cartão de crédito, imagine o problema que a empresa teria se o seu sistema bloqueasse uma transação genuína por analisar dados não condizentes com a realidade.

2.3.5 Valor

Informação não é só poder, informação também é patrimônio. A combinação de volume, velocidade, variedade, veracidade, além de todo e qualquer outro aspecto que caracteriza uma solução de *Big Data*, se mostrará inviável se o resultado não trouxer benefícios significativos e que compensem o investimento.

2.4 Aplicações de Big Data

Um startup, em São Paulo, usa o *Big Data* para oferecer ao criador de gado uma ferramenta inédita. Enquanto a realidade da maioria das fazendas do Brasil ainda é o papel e caneta, aqui eles criaram um aplicativo de coleta de dados para dispositivos móveis. O trabalhador coloca tudo ali: quantidade de bois e vacas, quais são as melhores produtoras de leite, peso, tudo. E é nas etapas seguintes a essa coleta de dados mais precisa que entra a solução de *Big Data*. Com o cruzamento desses dados, a solução usa mecanismos e algoritmos para interpretar esse grande conjunto de informações para que o pecuarista possa tomar suas decisões. Ainda hoje, para quem não aderiu à tecnologia, é difícil identificar a lucratividade real da atividade pecuária. Com o problema identificado, a startup viu a necessidade de

oferecer uma solução para que o criador de gado pudesse ter total controle sobre as informações coletadas da sua criação.

Outra experiência interessante com *Big Data* foi testada durante a Copa das Confederações de 2014. O exemplo misturou futebol, torcida e toda informação não estruturada proveniente das redes sociais. A solução era capaz de ler publicações do *Twitter* e *Facebook* incluido a capacidade para compreender gírias. Com o resultado, Felipe Scolari, técnico da Seleção, poderia ter o sentimento da torcida em tempo real ali no banco de reservas. Hoje a solução existe para qualquer empresa ou marca que queira medir sua reputação através das mídias sociais.

Ainda no esporte, outra ferramenta de *Big Data* aplicada no mundo do tênis analisou oito anos de informações dos Grand Slams, um total de 41 milhões de pontos marcados, para identificar padrões e estilos dos jogadores. Com isso, alguns dos melhores players do mundo já estenderam o treino para além das quadras. A norte-americana Serena Williams usa o *Big Data* para avaliar e comparar suas adversárias.

O *Flu Trends* do *Google* é outro exemplo. Baseado nos dados do seu buscador, a empresa desenvolveu um projeto no qual conseguiu identificar tendências de propagação de gripe antes de números oficiais refletirem a situação. A mesma técnica pode ser aplicada para analisar inflação, desemprego e muitas outras coisas.

Outra novidade é o uso em computação cognitiva, na qual a máquina passa a reproduzir o pensamento humano. Um dos primeiros testes foi em um “jogo do milhão” nos Estados Unidos, o computador com sistema cognitivo desafiou os maiores vencedores do programa e venceu.

A companhia *Skybox* tira fotos de satélites e vende a seus clientes informações com tempo real sobre a disponibilidade de vagas de estacionamento livres numa cidade em determinada hora ou quantos navios estão ancorados no mundo neste momento.

O projeto *Global Pulse*, das nações unidas, vai utilizar um programa que decifre a linguagem humana na analise de mensagens de texto e posts em redes sociais para prever o aumento do desemprego, o esfriamento econômico e epidemias de doenças.

Uma varejista americana, *Dollar General*, controla as combinações de produtos que seus clientes põem no carrinho. Ganhou eficácia e ainda descobriu

várias curiosidades como quem compra *Gatorade* tem mais chances de comprar também laxante.

A *spring Nextel* saltou da ultima para a primeira posição do ranking de satisfação dos usuários de celular no EUA ao integrar dados de todos os seus canais de relacionamento. De quebra cortou pela metade os gastos com *Call Center*.

No terremoto do Haiti, pesquisadores americanos fizeram uso da geolocalização de dois milhões de chips SIM, para auxiliar nas missões humanitárias.

Um hospital Canadense utilizou de uma tecnologia proposta pela *IBM* e da universidade de Ontario, para o monitoramento dos quadros de bebes prematuros, permitindo aos médicos antecipar as ameaças às vidas das crianças.

Em busca dos melhores lugares para instalar turbinas eólicas, uma empresa dinamarquesa analisou *petabytes* de dados climáticos do nível das marés, mapas de desmatamentos, entre outros. No fim o que costumava demorar semanas durou apenas algumas horas.

E diversos tipos de setores, como por exemplo, uma rede de vestuário que controla em tempo real seu fluxo de mercadoria e cruza os dados dos *GPS* dos caminhões dos seus fornecedores.

Descobrimento do pré-sal, devido a sua velocidade, que agilizava os processamentos de dados sísmicos captados pelas sondas que procuram petróleo no fundo do mar. Como são milhões as variáveis, o trabalho exige intermináveis simulações de imagens.

2.5 Hadoop

2.5.1 O que é Hadoop?

O *Hadoop* é uma plataforma *open source* desenvolvida especialmente para processamento e análise de grandes volumes de dados, sejam eles estruturados ou não estruturados. O projeto é mantido pela *Apache Foundation*, mas conta com a colaboração de várias empresas, como *Yahoo!*, *Facebook*, *Google* e *IBM*.

Pode-se dizer que o projeto teve início em meados de 2003, quando o *Google* criou um modelo de programação que distribui o processamento a ser realizado entre vários computadores para ajudar o seu mecanismo de busca a ficar mais rápido e livre das necessidades de servidores poderosos e caros. Esta tecnologia recebeu o nome de *MapReduce*.

Alguns meses depois, o *Google* apresentou o *Google File System* (GFS), um sistema de arquivos especialmente preparado para lidar com processamento distribuído e, como não poderia deixar de serem no caso de uma empresa como esta, grandes volumes de dados[3].

Em 2004, uma implementação *open source* do *Google File System* (GFS) foi incorporada ao *Nutch*[4]. O Nutch enfrentava problemas de escala, não conseguia lidar com um volume grande de páginas e a variação do *Google File System* (GFS), que recebeu o nome *Nutch Distributed Filesystem* (NDFS), se mostrou como uma solução. No ano seguinte, o *Nutch* já contava também com uma implementação do *MapReduce*.

Na verdade, o *Nutch* fazia parte de um projeto maior, uma biblioteca para indexação de páginas chamada *Lucene*. Os responsáveis por estes trabalhos logo viram que o que tinham em mãos também poderia ser usado em aplicações diferentes das buscas na Web. Esta percepção motivou a criação de outro projeto que engloba características do *Nutch* e do *Lucene*: o *Hadoop*, cuja implementação do sistema de arquivos recebeu o nome de *Hadoop Distributed File System* (HDFS).

2.5.2 Por que o Hadoop?

É um projeto *open source*, o fato é que permite a sua modificação para fins de customização e o torna suscetível a melhorias constantes graças à sua rede de colaboração. Por causa desta característica, vários projetos derivados ou complementares foram e ainda são criados.

Proporciona economia, já que não exige o pagamento de licenças e suporta *hardware* convencional, permitindo a criação de projetos com máquinas consideravelmente mais baratas.

Conta, por padrão, com recursos de tolerância a falhas, como replicação de dados.

É escalável: havendo necessidade de processamento para suportar maior quantidade de dados, é possível acrescentar computadores sem necessidade de realizar reconfigurações complexas no sistema.

Pode ser usado em conjunto com bancos de dados *NoSQL*. A própria *Apache Foundation* mantém uma solução do tipo que é uma espécie de subprojeto do

[3] Em grandezas de terabytes ou mesmo petabytes

[4] Um projeto de motor de busca para a Web

Hadoop.

Imagem 1: Logo Hadoop

Fonte: Google imagens, 2015

A denominação Hadoop tem uma origem inusitada, este é o nome que o filho de Doug Cutting, principal nome por trás do projeto, deu ao seu elefante de pelúcia amarelo.

É a opção de maior destaque, mas não é a única. É possível encontrar outras soluções compatíveis com *NoSQL* ou que são baseadas em *Massively Parallel Processing* (MPP).

2.6 Todas as empresas estão preparadas para utilizar?

Segundo a opinião do renomado físico alemão Andreas Weigend, uma das maiores autoridades mundiais em *Big data,* é preciso ser criativo, Isso significa que precisamos de competências diferentes para lidar com essa nova realidade. E quando falo em ensinar as pessoas a trabalhar com dados, não estou falando de matemática ou estatística, mas sobre como interpretá-los para que façam as perguntas pertinentes.

2.7 O que as pessoas devem saber?

Com a tecnologia batendo na porta de todo o mundo, onde grande parte da população e até mesmo a mais pobre possui celulares, desktops, e com isso pode obter qualquer tipo de informação, o segredo hoje em dia para o sucesso está na maneira de fazer a pergunta, é uma mudança já que em relação ao passado, você aprendia que era importante dar a melhor resposta.

2.8 Quais empresas estão fazendo as perguntas certas?

O *Google* e *Amazon*, por exemplo, vem fazendo excelente trabalho. Ambas possui uma incrível estratégia sobre o que fazer com os dados. Rapidamente, a partir de dados obtidos, elas já sabem se aquele projeto ou estratégia dará êxito em curto espaço de tempo, essa é a melhor maneira e o segredo do sucesso.

2.9 Empresas nacionais estão atentas ao Big Data

Uma pesquisa divulgada pela *IBM* mostra que mais da metade das empresas nacional já está se movimentando para estruturar estratégias de *Big Data*, para controlar melhor o grande fluxo de dados em suas operações.

De acordo com o estudo, que entrevistou 65 empresas no Brasil, 51% delas está definindo estratégias e atividades que serão colocadas em prática e 24% já implantando seus projetos pilotos. Os 25% restantes ainda não começaram. Para Sergio Loza, Diretor de Smarter Analytics Latam da IBM, o mercado está em fase acelerada de amadurecimento, com as empresas estudando o conceito de *Big Data*.

"As corporações estão enxergando os benefícios de investir em análise de dados, mapeando o planejamento e iniciando a implementação de grandes projetos", destaca.

De acordo com os resultados, quando pensam em Big Data 22% dos entrevistados pensam em novos tipos de dados e análises, enquanto 19% pensam em informações em tempo real.

Conforme aponta a IBM, mais da metade dos entrevistados (63%) notam que o uso da análise de dados tem criado uma vantagem competitiva para suas organizações. Se comparado com estudo semelhante realizado em 2010, o aumento foi de 70% (na ocasião, 37% enxergavam este cenário).

No entanto, nem tudo no Big Data está na pauta das empresas. Dados advindos de mídias sociais foram lembrados por apenas 2% dos entrevistados brasileiros. Para a IBM, isto mostra uma falta de conhecimento por parte das organizações de como as redes sociais podem impactar o negócio. Menos da metade das empresas brasileiras disseram já coletar e analisar perfis sociais. A maioria prefere concentrar seus esforços em fontes internas já existentes.

"As organizações estão comprometidas com a melhora da experiência do consumidor e em compreender suas preferências e comportamento, mas para que isso ocorra de fato é preciso também interagir e analisá-lo no ambiente online", explica Loza.

3 CONCLUSÃO

Conforme estudamos sobre o *Big Data*, hoje no mercado é essencial a utilização desse recurso para grandes empresas e com poder financeiro, pois a manutenibilidade requer grandes recursos financeiros.

Por outro lado, quem utiliza o *Big Data*, obtém informações preciosas, para aperfeiçoar o ramo de onde trabalha atender melhor o usuário, saber da concorrência, informações que pode avaliar o que poderá acontecer no futuro, resolver problemas desconhecidos ou até mesmo criar grandes estratégias de marketing.

A tendência, é que se use cada vez mais essa ferramenta, pois cada dia que passa, usamos e estudamos cada vez mais coisas sobre tudo o que acontece, sejam pra natureza, remédios, estudo humano, estudo didático, portanto, precisamos cada vez mais guardar essas informações sendo o *Big Data* então cada vez mais útil no nosso cotidiano.

REFERÊNCIAS

ADAMI, A. Big Data - Informática - InfoEscola. **Info Escola**. Disponivel em: <www.infoescola.com/informatica/big-data/>. Acesso em: 03 fev. 2015.

ALECRIM, E. O que é Big Data? **Info Wester**, 2013. Disponivel em: <http://www.infowester.com/big-data.php>. Acesso em: 10 fev. 2015.

DIGITAL, O. Olhar Digital: Conheça as aplicações reais do Big Data. **Olhar Digital**, 2013. Disponivel em: <http://olhardigital.uol.com.br/pro/video/conheca-as-aplicacoes-reais-do-big-data/39376>. Acesso em: 20 fev. 2015.

EXAME. “Big data tomará decisões pelas empresas”, diz cientista. **Exame**, 2014. Disponivel em: <http://exame.abril.com.br/tecnologia/noticias/falta-inteligencia-as-empresas-no-big-data-diz-weigend>. Acesso em: 02 fev. 2015.

GLOBO, O. Como funciona o Big Data. **O Globo**. Disponivel em: <http://www.oglobo.globo.com/infograficos/bigdata/>. Acesso em: 04 fev. 2015.

IBM. IBM - Infográfico: O que é o Big Data? - Brasil. **IBM Corporation**. Disponivel em: <http://www.ibm.com/midmarket/br/pt/infografico_bigdata.html>. Acesso em: 02/02/2015 fev. 2015.