FACULDADE DE ADMINISTRAÇÃO E NEGÓCIOS DE SERGIPE – FANESE

MBA EM ADMINISTRAÇÃO DE BANCO DE DADOS

MARCOS VINICIUS DOS SANTOS

INTRODUÇÃO AO CONCEITO DE BIG DATA

ARACAJU

2016

MARCOS VINICIUS DOS SANTOS

INTRODUÇÃO AO CONCEITO DE BIG DATA

ARACAJU

2016

SUMÁRIO

RESUMO

A quantidade de informação gerada atualmente é muito grande e precisa ser tratada pois as empresas precisam de dados para auxiliar na tomada de decisão para suas estratégias de vendas, marketing. Porém a maior parte dessas informações é classificada como não estruturada, como por exemplo, dados de redes sócias, e-mails, imagens. Os bancos de dados mais antigos foram projetados para tratar de dados estruturados e armazenar os registros em linhas e colunas, um exemplo desses dados seriam dados internos de um sistema de controle de estoque. Diante da necessidade de tratar esses dados não estruturados surgiu o conceito de Big data, com uma série de ferramentas que trata justamente do armazenamento, processamento e análise de uma quantidade enorme de dados originados de várias fontes e de vários tipos. O objetivo desse trabalho é apresentar o conceito de Big data, fases do processo, algumas ferramentas e riscos da utilização.

PALAVRAS CHAVE: INFORMAÇÃO; TOMADA DE DECISÃO; DADOS ESTRUTURADOS/NÃO ESTURURADOS; BIG DATA.

1. INTRODUÇÃO

O Big data está cada vez mais em evidência devido à capacidade de tratar a enorme quantidade de dados que são geradas atualmente, onde a maior parte desses dados não são estruturados, ou seja, não são dados controlados que estão armazenados em bancos de dados modelados para sistemas, são postagens em redes sociais, vídeos, e-mails, navegação em sites. Um exemplo da utilização de tratamento de dados não estruturados é que uma marca hoje consegue monitorar a navegação dos seus clientes dentro do seu site de vendas e a partir disso passar a sugerir produtos de acordo com o interesse dos clientes, consegue também medir o nível de satisfação com seus produtos buscando informações em redes sociais e e-mails. Para conseguir fazer essas medições e sugestões é necessário processar, fazer cruzamentos de dados de uma quantidade grande quantidade de informação. Mas como analisar esse tipo de informação se ela não estaria armazenada nos bancos de dados das empresas? Esse tipo de informação valiosa precisa ser armazenada, processada e analisada, diante dessa necessidade o Big data surgiu, ainda está em processo de amadurecimento, mas várias empresas já estão montando suas estratégias de negócios a partir de suas próprias implementações de Big data.

2. REFERENCIAL TEÓRICO

Segundo o ISACA(2016), Big Data representa uma tendência em tecnologia que está abrindo caminho para um novo modelo de compreensão do mundo e do processo decisório de negócios. A quantidade de informação gerada atualmente é muito grande, em alta velocidade, e originada de diversas fontes. Os sistemas atuais não são capazes de analisar toda essa informação, pois nem sempre são dados simples de serem tratados, como por exemplo, dados de um sistema ERP, que controla todo o funcionamento de uma empresa. A maior parte dos dados não são estruturados como: vídeos, imagens, mensagens de redes sociais, dados de sensores, monitoramento de cliques em sites, smartphones e vários outros dispositivos que forneça algum tipo de informação. Além da complexidade do tipo dos dados outro fator importante é que não

estão armazenados nos bancos de dados das empresas, estima-se que 90% dos dados digitais disponíveis não estão sendo aproveitados de maneira correta. Diante deste cenário surge o conceito de big data, que segundo TAURION(2013) é um conjunto de tecnologias, processos e praticas que permitem as empresas analisarem dados que antes não tinham acesso e tomar decisões ou mesmo gerenciar atividades de forma muito mais eficiente. Dados são recursos naturais da sociedade da informação, porém só possuem valor de forem tratados, analisados e usados para tomada decisão.

A INFOWESTER diz que, informação é poder, logo se uma empresa souber como utilizar os dados que em mãos, poderá entender como melhorar um produto, como criar uma estratégia de venda mais eficiente, como cortar gastos, enfim otimizar a utilização de recursos e oferecendo um produto ou serviço de melhor qualidade para seus clientes.

2.1 Os ‘Vs’ do Big Data

O termo big data não é tão recente, então à medida que o tempo passa algumas características são adicionadas e/ou melhoradas, inicialmente a justificativa se dava em cima de três variáveis:

1. **Volume:**
   Que corresponde a enorme quantidade de informação que é manipulada e que cresce cada vez mais.
2. **Velocidade**
   O tratamento dos dados deve ser feito de forma bastante rápida, em alguns casos em tempo real.
3. **Variedade**
   Corresponde a diversificação de tipos de dados, dados estruturados, semiestruturados e não estruturados.

As variáveis mais recentes são:

4. **Veracidade**
   Os dados devem ser confiáveis. Não adianta fazer um processamento de uma quantidade enorme de dados de forma rápida se os dados não são confiáveis.

5. **Valor**

   A questão do valor é a justificativa da execução da analise, não faz sentido executar todo o procedimento se o resultado final da operação não for favorável, não trouxer nenhuma melhoria, seria apenas desperdício de recursos.

2.2 Fases do Big Data

Existem empresas que tentam utilizar o Big Data de forma descoordenada, sem processos definidos. Certamente essas empresas falharão em algum momento, segundo TAURION(2013), Big Data não é apenas comprar pacotes de tecnologia, mas uma nova maneira de explorar o imenso volume de dados que circula dentro e fora das empresas, isso embute transformações nos processos de negócios, fontes de dados, infraestrutura de tecnologia, capacitações e mesmo mudanças organizacionais na empresa e em TI. Big Data cria valor para as empresas descobrindo padrões e relacionamentos entre dados.

As fases do processo são as seguintes:

1. **Coleta de dados**

   Nessa fase, as variáveis predominantes são volume e variedade. Volume porque todos os dados oriundos de diversas fontes são depositados no repositório, e a variedade se refere a vários tipos de dados que serão adicionados. Um exemplo prático dessa fase seria uma empresa de varejo obter os dados de seus produtos, sua marca, em comentários de redes sócias.

2. **Limpeza e formatação**

   É importante validar os dados extraídos, remover os dados incompletos ou com inconsistência, e formatar os dados de alguma forma que possam ser manipulados de forma mais fácil.

3. **Integração e agregação**

   Como os dados já foram tratados, nesse momento ocorre os cruzamentos das informações de acordo com a questão a ser resolvida.

4. **Análise e interpretação**

Nessa fase é feita a analise do cruzamento das informações, verificando se foram encontradas respostas, indícios de como agregar mais valor ao negócio, se as questões informadas foram resolvidas.

2.3 Recursos Tecnológicos

Do ponto de vista tecnológico existem dois grupos de ferramentas para montar a solução de Big data, o primeiro se refere à parte de análise dos dados também conhecido como “analytics”, e o segundo é a parte de infraestrutura de fato, onde a grande massa de dados é armazenada e processada. É necessário atenção aos componentes de analytics pois é quem transforma os dados em algo de valor para o negócio. Big Data Analytics não significa eliminar os tradicionais sistemas de BI que existem hoje, mas, pelo contrário, devem coexistir.

Sobre a analise dos dados uma tecnologia que se destaca é o Hadoop, que é um projeto da comunidade Apache(//hadoop.apache.org) criado em 2005 que foi inspirado no trabalho do Google em seu projeto GFS(Google File System) junto com o paradigma de programação MapReduce. Segundo seu próprio site (hadoop.apache.org) o Hadoop é definido como uma estrutura que permite o processamento distribuído de grandes conjuntos de dados em clusters de computadores que utilizam modelos de programação simples. Ele é projetado para ampliar a partir de um único servidor para milhares de máquinas, cada um oferecendo computação e de armazenamento local. O Haddop inclui alguns módulos:

- **Hadoop Common**: que é um conjunto bibliotecas e utilitários que são suporte a outros módulos do Hadoop.
- **Hadoop Dsitributed File System(HDFS™):** É um sistema de arquivos otimizado para atuar em dados não estruturados e com grande volume de dados. Os dados são armazenados em blocos divididos em pedaços menores e distribuídos por diversos servidores, com isso a pesquisa se torna muito mais rápida pois o processamento é feito de forma simultânea e em paralelo.
- **Hadoop YARN:** que é um framework para agendamento de atividades e recursos, e um gerenciador de recursos para serviços em cluster.

- **Hadoop MapReduce:** é o coração do Hadoop, um framework para processamento de grandes volumes de dados de forma paralela. O termo mapReduce representa duas tarefas executadas pelo Hadoop, a primeira é o map onde um conjunto de dados é processado e mapeado. A segunda é o reduce onde o resultado do mapeamento é processado e reduzido, gerando um resultado.

Sobre o armazenamento, desde quando foi projetado o modelo relacional tem sido bastante utilizado, hoje ainda é o mais utilizado, porém ele foi projetado para armazenar dados estruturados, em tabelas e linhas. Acontece que no big data a maior parte dos dados não são estruturados, diante dessa necessidade surgiram os bancos de dados com modelos não estruturados NoSql(Não somente SQL), que foram projetados para manipular dados estruturados e não estruturados,

Segundo JUDITH (2013) os bancos de dados não relacionais possuem as seguintes características em comum:

- **Escalabilidade:** se refere à capacidade de gravar dados de várias fontes de dados de forma simultânea.
- **Modelo de dados e consultas:** Em vez de linhas, colunas, estruturas de chave, os banco não relacionais utilizam frameworks para armazenar e pesquisar os dados de forma inteligente.
- **Desenho de persistência:** Persistência é um elemento crítico em banco de dados. Devido a alta velocidade, variedade, e volume dos dados, esse tipo de banco de dados usam diferentes mecanismos para persistir o dado.
- **Interface diversificada:** Oferece uma grande variedade de mecanismos de conexão para programadores e administradores de banco dados, incluindo ferramentas de análise e relatórios.
- **Consistência eventual:** Enquanto os bancos de dados relacionais usam ACID (Atomicidade, consistência, isolamento e durabilidade) como mecanismos para garantir a consistência dos dados, os nãos relacionais utilizam BASE (“Basically Avaliable” basicamente disponível, “soft state” tolerante a falhas, “eventual consistence” consistência eventual). A consistência eventual é a mais

importante, pois é responsável pela resolução de conflitos quando dados estão em movimento entre nós em uma implementação distribuída.

Existem várias modelos de banco de dados não relacionais, os principais modelos segundo JUDITH(2013) estão listados abaixo:

- **Banco de dados Chave-Valor:** O mais simples dos bancos NoSQL, baseado em chave-valor, não exige um esquema de dados(similar ao relacional) e oferece grande flexibilidade e escalabilidade. Os dados são armazenados como “Strings” e mapeados por uma chave. Exemplos de banco de dados nesse modelo são o RIAK, Redis, TokioCabinet, CouchBase.
- **Banco de dados orientado a documento:** Esse tipo de banco de dados é muito útil quando é necessário produzir um conjunto de relatórios e eles precisam ser montados de forma dinâmica a partir de elementos que mudam constantemente. Exemplos de banco de dados nesse modelo MongoDB, couchdb.
- **Banco de dados orientado a colunas:** Ao invés de cada registro da tabela ficar armazenado em uma linha, o registro passa a ser armazenado em colunas separadas. Essa forma de armazenamento tem algumas vantagens, como exemplo a capacidade de compressão dos dados, se formos analisar a compressão de um banco onde os registros são armazenados em linha, encontraremos em uma mesma linha diferentes tipos (domínios) o que torna o processo mais complicado, já no banco orientado a colunas, cada coluna irá conter o mesmo tipo (domínio) de dado. De acordo com algumas pesquisas o nível de compressão alcançado em bancos orientados a colunas chega a ser de 60% a 70% mais eficiente que nos bancos orientados a linhas. Exemplo de banco de dados nesse modelo, HBase.
- **Banco de dados orientado a grafos:** A ideia desse modelo é representar os dados como grafos dirigidos. Possui três componentes básicos: os nós(vértices do grafo), os relacionamentos(as arestas) e as propriedade(ou atributos) dos nós e relacionamentos. Exemplo de banco de dado nesse modelo, Neo4j.

- **Banco de dados espaciais:** é um banco de dados utilizado para armazenar informações geográficas. Exemplo de banco de dados nesse modelo, PostGis

O cloud computing(computação nas nuvens) também deve ser levando em consideração, pois como a quantidade de dados é muito grande e pode variar, fica difícil estipular uma capacidade de armazenamento local, nesse caso poderia utilizar nuvens públicas onde seria permitido utilizar servidores virtuais sob demanda no momento do tratamento dos dados

2.4 Recursos Humanos

É natural que com tantas tecnologias surgindo, sejam criadas novas oportunidade de trabalho, novas funções, se tratando de big data, um novo cargo que surgiu e merece destaque é o de “data scientist”, cientista de dados, que diante do trabalho que será executado demanda da pessoa conhecimento das áreas de ciência da computação, matemática e estatística. O cientista de dados pode ser definido como um profissional de alto nível de formação, com curiosidade de fazer descobertas no mundo big data.

Um dos grandes desafios do big data será ter profissionais capacitados para exercer a função de cientista de dados, como já trabalham na área de dados por exemplo na área de business intelligence(BI), alguns profissionais se auto intitulam cientistas, porém deve ser levado em consideração que, BI, trabalha com dados históricos, dados estruturados, então é possível fazer comparações sobre algum momento antigo junto ao atual e obter algumas respostas de desempenho por exemplo. No big data muitas vezes os dados serão analisados em tempo real e com e com a informação obtida será possível fazer previsões sobre determinado assunto, auxiliando na tomada de decisão. Uma pequena comparação mostra a diferença, um profissional de BI geralmente está habituado com ferramentas de data warehouse, usa SQL junto a bancos de dados relacionais, como Oracle, Sql Server. Já o cientista de dados deve ter conhecimento de estatística, matemática, entender do negocio, conhecer as tecnologias como hadoop, e bancos de dados não relacionais.

Abaixo segue uma imagem com algumas das tecnologias utilizadas para cada cargo:

Analista de BI	Cientista de dados
Cognos, modelo relacional, banco de dados SQLServr, Oracle, DB2	Hadoop, modelos relacionais e NoSQL, bancos de dados não relacionais e in-memory
Modelagem relacional/estruturada	Inclui também modelagem não estruturada. Modelagem analítica é essencial.
Desenvolve queries estruturados sobre dados passados.	Cria perguntas e busca relacionamentos entre fatos aparentemente desconexos.

Fonte: TAURION(2013)

2.5 Riscos

Para a implantação de um projeto de big data é necessário um investimento consideravelmente alto, porém devem ser levados em consideração os riscos envolvidos nesse projeto, como por exemplo, produção dos dados de forma correta, local do armazenamento, proteção dos dados, legalidade. È necessário uma integração de todos os setores da empresa, alinhando ao negócio e focando no resultado. Segundo o ISACA(2016) dados imprecisos, incompletos ou manipulados de modo fraudulento representam um risco crescente à medida que as empresas se tornam mais dependentes dos dados para direcionar a tomada de decisões e avaliar resultados.

É muito importante manter uma política de proteção aos dados, pois o vazamento de alguns desses dados podem gerar grandes danos a empresa. Além da proteção, outra questão é a utilização de certas fontes de dados, como por exemplo, dados pessoais, mesmo que estejam em redes sociais, muitas empresas utilizam esse

tipo de informação para acompanhar o nível de satisfação do cliente com determinado produto, suas preferências, necessidades, sugestão de produtos, é preciso verificar se não existe alguma lei que regulamenta a utilização desses dados, pois cada país pode determinar suas próprias leis.

3. CONCLUSÃO

Como a quantidade de informação tende a crescer cada vez mais, o trabalho com Big data vai continuar sendo feito, sempre buscando otimizar o processo, auxiliando as empresas a conseguir indicadores para auxiliar na tomada de decisões, provavelmente as empresas que ainda não aderiram, em algum momento devem aderir, visto que oferece uma série de vantagens. A maior parte das tecnologias utilizadas ainda hoje é de código aberto, ou seja, não é necessário comprar licenças, o que não significa que será barato, claro que vão existir módulos proprietários, mas que deve ser avaliada a questão do custo x beneficio.

Devem ser levados em consideração os riscos de um projeto desses, é necessária uma integração das áreas envolvidas, todo um cuidado para não gerar informação invalida, vazamento de informação e problemas com legalidade com a utilização dos dados.

Abstract

The amount of information generated is currently too large and needs to be addressed, because companies need data to assist in decision making for their sales strategies, marketing. But most of the information is classified as unstructured, such as data members networking, e-mails, images. Older databases are designed to handle structured data and store the records in rows and columns, an example of such data would, internal data of a stock control system. Faced with the need to deal with these unstructured data the concept of Big data came with a series of tools that just deals with the storage, processing and analysis of a huge amount of data coming from various sources and of various kinds. The aim of this paper is to present the concept of Big data, process steps, some tools, and risks of using.

KEY WORDS: INFORMATION; DECISION TAKING; STRUCTURED DATA / NO ESTURURADOS; BIG DATA.

REFERÊNCIAS

HURWITZ, JUDITH ; NUGENT, ALAN; HALPER, FER, KAUFMAN MARCIA; **BIG DATA FOR DUMMIES** JOHN WILEY & SONS, INC. 2013

TAURION, CEZAR **; BIG DATA** ; BRASPORT, 2013

http://www.infowester.com/bigdata.php ACESSO EM 13 DE MAIO 2016

https://www.fiap.com.br/2015/07/27/fiapx/governancadebigdataanalytics/ ACESSO EM 13 DE MAIO 2016

https://www.fiap.com.br/fiapx/cursos/big-data-desafios-oportunidades-e-tendencias/apresentacao ACESSO EM 15 DE MAIO 2016

https://pt.wikipedia.org/wiki/Big_data espaciais ACESSO EM 16 DE MAIO 2016

ISACA. DISPONÍVEL EM: http://www.isaca.org/Knowledge-Center/Research/Documents/Big-Data_whp_Por_0413.pdf . ACESSO EM 19 DE MAIO 2016.

https://pt.wikipedia.org/wiki/Banco_de_dados_espaciais ACESSO EM 19 DE MAIO 2016

http://www.devmedia.com.br/introducaoaosbancosdedadosnosql/26044 ACESSO EM 19 DE MAIO

https://isaiasbarroso.wordpress.com/2012/06/20/bancodedadosorientadoacolunas/ ACESSO EM 19 DE MAIO