# O que é Big Data – Da Teoria à Implementação

Publicado por: Byron Kiourtzoglou em Desenvolvimento de Software 25 de abril de 2013

https://www.javacodegeeks.com/2013/04/what-is-big-data-theory-to-implementation.html

Seria o Big Data a última tendência em quase todos os domínios negócio? É apenas um modismo ou veio para ficar?

De fato, o termo "Big Data" tem um significado bastante simples, é exatamente o que diz: um banco de dados muito grande. Mas quão grande deve ser um BigData? A resposta mais exata é "tão grande quanto você possa imaginar"!

Mas como pode este banco de dados er tão grande?

Se considerarmos a IoT, podemos afirmar que os dados podem vir de todos os lugares e em velocidades enormes: sensores RFID que reúnem dados de tráfego, sensores usado para reunir informações sobre o tempo, pacotes GPRS de telefonia celulares, mensagens de sites de mídia social, imagens e vídeos digitais, registros de compra de transações on-line, e muito mais! Big Data se refere a um enorme banco de dados que pode conter informações de todas as fontes possíveis e que produz dados do nosso interesse.

No entanto Big Data é mais do que simplesmente uma questão de tamanho, é também uma oportunidade para discernir e identificar tipos de dados novos e emergentes, com conteúdo relevante e interrelacionado, que podem tornar as empresas mais ágeis quando são utilizados para responder às perguntas que eram consideradas previamente fora do nosso alcance. Nessa abordagem, o Big Data é caracterizado por quatro aspectos principais: Volume, Variedade, Velocity, e Veracidade (valor), conhecidos como "os quatro V's do Big Data". Vamos examinar brevemente o que cada um deles representa e quais os desafios que ela apresenta:

## Volume

Volume se refere à quantidade de conteúdo que uma empresa deve ser capaz de capturar, armazenar e acessar. 90% dos dados existentes no mundo hoje foram gerada somente nos últimos dois anos. Organizações de hoje estão sobrecarregados com volumes de dados, acumulando facilmente terabytes – ou mesmo petabytes – de dados de todos os tipos, muitos dos quais precisam ser organizados, analizados e armazenados com segurança.

## Variedade

80% dos dados disponíveis no mundo são semiestruturados. Sensores e sistemas diversos, dispositivos inteligentes e mídias sociais estão gerando esses dados através de páginas da Web, arquivos em blogs, fóruns de mídia social, arquivos de áudio e vídeo, sequências de cliques (www), e-mails, documentos e assim por diante. As soluções tradicionais existentes (aplicativos) para análise de dados funcionam muito bem com informação estruturada, como os dados em um banco de dados relacional bem construído. No Big Data, a variedade nos tipos de dados representa uma mudança fundamental na forma como os dados são armazenados, e a análise precisa ser feito para apoiar os processo de discernimento e de tomada de decisão nos dias de hoje e. Portanto, Variedade representa os vários tipos de dados que não podem ser facilmente capturados e gerenciados em um banco de dados relacional tradicional, mas que podem ser facilmente armazenados e analisados com as tecnologias de Big Data.

## Velocidade

Velocidade exige metdologia de análise de dados praticamente em tempo real, também conhecido como "2 minutos pode ser tarde demais!". Ganhar uma vantagem competitiva significa identificar uma tendência ou uma oportunidade em minutos ou mesmo segundos antes do seu concorrente. Outro exemplo se relaciona com os processos sensíveis ao tempo, como a captura de fraudes onde a informação deve ser analisada como ela flui em sua empresa, a fim de maximizar o seu valor. Dados sensíveis ao tempo têm um prazo de validade muito curto, levando as organizações à sua análise em tempo quase real.

## Veracidade (Valor)

Lidar com dados (organizar) é uma forma de criar oportunidades e agregar valor. Dados têm tudo a ver com apoio à decisão, então quando você está diante de decisões que podem ter um grande impacto sobre o seu negócio, você vai querer ter acesso ao máximo de informações possíveis para tomar a decisão mais correta. No entanto, o volume de dados por si só não fornece a confiança suficiente para os tomadores de decisão para agir sobre a informação. A veracidade e qualidade dos dados são as fronteiras mais importantes para abastecer novos discernimentos e idéias. Assim, estabelecer a confiança em soluções de Big Data, provavelmente, apresenta o maior desafio deve superar a introduzir uma base sólida para a tomada de decisões bem sucedido.

Considerando- se que a base de soluções existente para nteligência de negócios (business intelligence) e data warehouse não foram projetados para suportar os quatro de V's, as soluções para Big Data estão sendo desenvolvidas para enfrentar esses desafios.

As principais ferramentas baseadas em código aberto Java que estão disponíveis hoje como suporte a Big Data são brevemente apresentadas a seguir:


**Hadoop HDFS** HDFS (Hadoop Distributed File System) é o armazenamento distribuído primário utilizado para aplicações Hadoop. Um cluster HDFS consiste principalmente de um *NameNode* que gerencia os metadados do sistema de arquivos e *DataNodes* que armazenam os dados reais. HDFS é projetado especificamente para armazenar grande quantidade de dados, por isso é otimizado para armazenar/acessar um número relativamente pequeno de arquivos muito grandes, em comparação com os sistemas de arquivos tradicionais que são otimizados para lidar com um grande número de arquivos relativamente pequenos.


**Hadoop MapReduce** MapReduce é uma estrutura de software para facilitar a criação de aplicações que processam grandes quantidades de dados paralelados (conjuntos de dados multi-terabyte) em grandes clusters (milhares de nós) de hardware convencional, de forma confiável e tolerante a falhas.


**Apache HBase** HBase é o banco de dados Hadoop para armazenamento de grandes volumes de dados, distribuído escalável. Fornece acesso de leitura/gravação aleatória em tempo real em BigData e é otimizado para hospedagem de tabelas muito grandes - bilhões de linhas X milhões de colunas - em clusters de hardware convencional. O núcleo do Apache HBase é uma versão de armazenamento distribuído orientado a colunas, modelado como o Google Bigtable: Um sistema de armazenamento distribuído para dados estruturados (Chang et al). Assim como Bigtable aproveita o armazenamento de dados distribuídos fornecido pelo Google File System, HBase fornece capacidades Bigtable semelhantes, com o Hadoop e HDFS.


**Apache Cassandra** O Apache Cassandra é uma base de dados linear de alto desempenho, escalável e de alta disponíbilidade, que pode ser gerenciada em hardware convencional ou em infraestrutura de nuvem, tornando-se a plataforma perfeita para dados de origem crítica. O Cassandra dá o melhor suporte para replicação em vários datacenters, proporcionando menor latência para os usuários e a tranquilidade de saber que você pode sobreviver às interrupções regionais. O modelo de dados do Cassandra oferece a conveniência dos índices de coluna com o desempenho de atualizações estruturado em log, além de suporte robusto para a desnormalização e visões materializadas, com armazenamento em cache próprio e poderoso.


**Apache Hive** Apache Hive é um sistema de armazenamento de dados para Hadoop que facilita bastante a compactação de dados, consultas ad-hoc, e a análise de grandes conjuntos de dados armazenados em sistemas de arquivos compatíveis com Hadoop. Hive fornece um mecanismo para projetar estruturas sobre esses dados e consultar os dados usando uma linguagem baseada em SQL, a HiveQL. Ao mesmo tempo, esta linguagem também fornece plugins aos programadores tradicionais que utilizam MapReduce permitindo ligar os seus mapeadores e redutores personalizados quando é inconveniente, ou quando essas ferramentas são ineficientes em HiveQL.

**Apache Pig** Apache Pig é uma plataforma para análise de grandes conjuntos de dados. Ele consiste de uma linguagem de alto nível para expressar programas de análise de dados, juntamente com a infra-estrutura para a avaliação desses programas. A propriedade mais evidente dos programas em Apache Pig é que a sua estrutura é passível de paralelização substancial, que por sua vez lhes permite lidar com grandes conjuntos de dados. A camada de infra-estrutura do Apache Pig consiste em um compilador que produz sequências de programas para MapReduce, e a camada de linguagem é constituída de uma linguagem textual denominada Pig Latin, que foi desenvolvida com o propósito de facilitar a programação, otimização de oportunidades e extensibilidade.


**Apache Chukwa** Apache Chukwa é um sistema de coleta de dados de código aberto para monitorar grandes sistemas distribuídos. Ele é construído sobre o HDFS e no framework MapReduce e herda a escalabilidade e robustez do Hadoop. Chukwa também inclui um conjunto de ferramentas flexível e poderoso para a exibição, monitoramento e análise de resultados, permitindo melhor uso dos dados coletados.

**Apache Ambari** Apache Ambari é uma ferramenta baseada na Web para o provisionamento, gerenciamento e monitoramento de clusters de Apache Hadoop, incluindo suporte para o Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig e Sqoop. Ambari também fornece um painel de controle para visualização de estado do cluster – Cluster Health – na forma de mapas de temperatura, além da capacidade de exibir visualmente os aplicativos MapReduce, Pig e Hive, com recursos para diagnosticar suas características de desempenho de uma maneira amigável (user-friendly).

**Apache Zookeeper** Apache ZooKeeper é um serviço centralizado para manter as informações de configuração, denominação, proporcionando sincronização distribuída e serviços de grupo. Todos esses tipos de serviços são utilizados de alguma forma ou de outra por aplicações distribuídas. Em suma, Apache ZooKeeper é um serviço de coordenação de alto desempenho para

aplicações distribuídas como aqueles executados em um cluster do Hadoop.

**Apache Sqoop** Apache Sqoop é uma ferramenta concebida para a transferência eficiente de dados em massa entre Apache Hadoop e bases de dados estruturadas, tais como bancos de dados relacionais.

**Apache Oozie** Apache Oozie é um sistema de fluxo de trabalho (Workflow) programavel escalável, confiável e extensível para gerenciar tarefas do Apache Hadoop. As tarefas do Oozie Workflow são representadas por grafos acíclicos dirigidos (DAGs). As terfas do Coordenador Oozie são workflows Oozie recorrentes, disparadas por tempo (frequência) e disponibilidade dos dados. Oozie é integrado com o restante da estrutura Hadoop pelo suporte aos vários tipos de tarefas independentes do Hadoop (como Java MapReduce, streaming MapReduce, Pig, Hive, Sqoop e Distcp), bem como tarefas específicas do sistema (como programas Java e Shell scripts).

**Apache Mahout** Apache Mahout é uma biblioteca escalável que opera como máquina de aprendizagem e de mineração de dados. Atualmente Mahout apoia principalmente quatro casos de uso:

Mineração para Recomendação: Acompanha o comportamento dos usuários e a partir desses dados tenta encontrar itens que os usuários podem gostar.
Clustering: Toma documentos de texto (por exemplo) e os organiza em grupos de documentos relacionados por tópico.
Classificação: aprende a partir de documentos classificados que documentos de uma categoria específica se parecem e é capaz de atribuir documentos não marcados para o (espero) categoria correta existente.
Mineração para Itens de Ocorrência Frequente: Toma um conjunto de grupos de itens (termos em uma sessão de consulta, conteúdo do carrinho de compras, etc) e identifica quais itens individuais geralmente aparecem juntos. (“Outros compradores desse item também compraram:...)

Apache HCatalog Apache HCatalog é um serviço de gerenciamento de tabelas e armazenamento de dados criados usando Apache Hadoop. Isso inclui:

Proporcionar um mecanismo de compartilhamento de esquema e tipo de dados.
Fornecendo uma uma tabela de abstração para que os usuários não precisem se preocupar com onde ou como os seus dados são armazenados.
Fornecer interoperabilidade entre ferramentas de processamento de dados, tais como Pig, MapReduce e Hive.

É isso aí; Big Data, uma breve introdução teórica e uma matriz compacta de implementação abordagens centradas na superação dos problemas de uma nova era - a era que nos obriga a fazer perguntas maiores!

Byron Kiourtzoglou
Byron é um engenheiro de software mestre trabalha nos domínios de TI e Telecom. Ele é um desenvolvedor de aplicativos em uma ampla variedade de aplicações/serviços. Atualmente, ele está agindo como o líder da equipe e arquiteto técnico para a criação e integração de serviços plataforma proprietária para ambas as indústrias de TI e Telecom, além de uma solução de análise em tempo real in-house para BigData. Ele está sempre fascinado por SOA, serviços de middleware e desenvolvimento móvel. Byron é co-fundador e editor executivo no código Java Geeks.

# O maior problema em BigData: é muito difícil inserir os dados

Por Jason Hiner, em Big Data Analytics . 10 de abril de 2016

http://www.zdnet.com/article/big-datas-biggest-problem-its-too-hard-to-get-the-data-in/

***Enquanto big data passou a ser mais um termo de marketing do que uma tecnologia, ele ainda tem um enorme potencial inexplorado. Mas, uma grande questão tem de ficar resolvido em primeiro lugar.***

Imagem: iStockphoto / kieferpix

**(texto sem revisão – tradução direta do Google Translator)**

A maioria das empresas estão nadando em mais dados do que eles sabem o que fazer com. Infelizmente, muitos deles associar esse fenômeno se afogando com o próprio big data. Tecnologicamente, big data é uma coisa muito específica - o casamento de dados estruturados (informações de propriedade da sua empresa) com dados não estruturados (fontes públicas, tais como fluxos de mídia sociais e governo alimenta).

Quando você sobrepor dados não estruturados em cima de dados estruturados e usar a análise de software para visualizá-lo, você pode obter insights que nunca foi possível antes - prever as vendas de produtos, melhores clientes-alvo, descobrir novos mercados, etc.

Big data não está sofrendo com a falta de ferramentas que assolaram que apenas alguns anos atrás, quando a fazer big data significava ter cientistas de dados sobre pessoal e mexer com ferramentas de código aberto como o R e Hadoop .

VER: O Poder da Internet das coisas e Big Data (ZDNet / TechRepublic relatório especial)

Hoje, existem toneladas de empresas que competem uns com os outros para ajudar a visualizar big data - de especialistas como Tableau, Qlik, TIBCO, e MicroStrategy para end-to-end jogadores como Microsoft, IBM, SAP e Oracle.

Mas, de acordo com os executivos de TI no Médio Porte CIO Forum / Médio Porte CMO Fórum na semana passada em Orlando, um dos maiores problemas que as empresas estão tendo com todas essas plataformas Analytics é a ingestão de dados em si.

Um CIO disse: "Nosso maior problema em TI é como vamos conseguir dados nela. É aí que estas coisas são realmente uma dor."

Apropriadamente, essa afirmação é apoiada por dados.

De acordo com um estudo realizado pelo especialista em integração de dados Xplenty, um terço dos profissionais de inteligência de negócios gastam 50% a 90% do seu tempo a limpeza de dados brutos e prepara-se para introduzi-lo em plataformas de dados da empresa. Isso provavelmente tem muito a ver com o motivo apenas 28% das empresas pensam que estão gerando valor estratégico de seus dados .

VER: 12 modelos de dados CIOs e CMOs pode começar a construir juntos (TechRepublic)

O problema de limpeza de dados também significa que alguns dos mais amplamente procurados profissionais da área de tecnologia agora estão gastando uma grande parte do seu tempo fazendo o trabalho de entorpecimento mental de triagem através de e organização de conjuntos de dados antes que eles nunca se analisou.

Isso é, obviamente, não muito escalável e limita severamente o potencial de big data. E como nós ficar melhor e melhor a recolha de mais dados - com a ajuda da Internet das coisas - o problema só piora.

Existem três soluções possíveis para o problema:

1. A grande software de análise de dados fica melhor --Since muitas dessas empresas têm investido pesadamente em big data para os últimos cinco anos, é improvável que haja avanço nas ferramentas qualquer momento em breve que irá aliviar o fardo sobre a limpeza de dados , mas devemos esperar melhorias incrementais.

2. preparadores de dados tornam-se os paralegais da ciência dados --no mesma forma que paralegais auxiliar os advogados, assumindo importantes, tarefas de nível mais baixo, preparadores de dados poderia fazer o mesmo para os cientistas de dados. Nós já estamos vendo isso em um grau. Leia o artigo TechRepublic, «rotulagem de dados» é o novo trabalho de colarinho azul da era AI?

3. AI vai ajudar a limpar os dados --A outra possibilidade é que o software e os algoritmos serão escritos para limpar, classificar e categorizar os dados. Isso é mais definitivamente vai acontecer, mas também devemos esperar que ele não vai ser uma bala de prata. Microsoft, IBM e Amazon estão investindo no uso de seres humanos para fazer a rotulagem de dados que o software não pode lidar com - e esses são três dos campeões mundiais de automação e algoritmos.

Do ZDNet Monday Morning Opener é o nosso salva de abertura para a semana em tecnologia.

Como um site global, este editorial publica na segunda-feira AEST 08:00 em Sydney, Austrália, que é 18:00 horário da costa leste no domingo em os EUA. É escrito por um membro do conselho editorial global da ZDNet, que é composta por nossos editores de chumbo em toda a Ásia, Austrália, Europa e os EUA.