4. BIG DATA, DATA WAREHOUSE e BUSINESS INTELLIGENCE

4.1. INTRODUÇÃO

Os três conceitos, **Big Data, Data Warehouse (DW) e Business Intelligence (BI),** se interligam e a perspectiva é que, cada vez mais, as maiores empresas do mundo utilizem as análises geradas por esses tipos de tecnologias para ter uma visão mais analítica do seu negócio e, assim, poder tomar as melhores decisões para o crescimento.

Três conceitos que tem, em comum, uma nova maneira de lidar com dados – sempre levando em conta um volume gigantesco de informações em diversos formatos que contribuem, de forma **estruturada ou não estruturada**, para a tomada de decisões estratégicas. Talvez isso seja o ponto intercessor de união dessas tecnologias. O objetivo final de qualquer um deles é dar um diferencial competitivo para as corporações, mas a maneira como serão utilizados é que fará toda a diferença (www.cetax.com.br).

Dados Estruturados são organizados em linhas e colunas, geralmente são encontrados em banco de dados relacionais, usado para sistemas comerciais.

Dados Não-Estruturados referem-se a dados que não podem ser organizados em linhas e colunas, como por exemplo: E-mail, imagens, vídeos, redes sociais, buscas, geolocalização, e-commerce etc.

4.2. BIG DATA

A busca por padrões comportamentais em um grande volume de dados (**massa de dados**) disponíveis na rede mundial de computadores tornou-se a grande tendência que as grandes empresas adotaram para gerar lucros e aumentar seu rol de consumidores (PETRY; VILICIC, 2013, p.75).

Big Data é a forma de manipular um imenso volume de dados, oriundos das mais variadas fontes, em alta velocidade de processamento e resposta. O manuseio de um expressivo montante de informações não estruturadas, como imagens, vídeos, SMS, fica inviável se para tal tarefa forem utilizadas ferramentas e técnicas que estão disponíveis no mercado, mas que não se norteiam nos princípios Big Data (PETRY; VILICIC, 2013, p.75).

Figura 1. Ilustração de Big Data

O **Big Data** significa um volume ou conjunto de soluções tecnológicas com um volume de dados, variedade, complexidade e velocidade grandiosos e que têm como principal característica serem **não estruturados**. Isso significa que são **não relacionais** e estão fora do ambiente corporativo. É um tipo de tecnologia que permite analisar dados em tempo real e podem ser provenientes de diferentes fontes e formas – mensagens instantâneas, redes sociais, gravações de logs, imagens, e-mails são apenas algumas delas (www.cetax.com.br).

Big Data (“Mega dados”), em tecnologia da Informação, refere-se a um grande conjunto de dados armazenados. E pode-se basear em **3V’s**: Velocidade, Volume e Variedade, que serão descritos no transcorrer desse material.

Big Data é um termo amplamente usado atualmente para nomear conjuntos de dados muito grande ou complexos, que os aplicativos de processamento de dados tradicionais não conseguem lidar. Para atuar com Big Data, deve-se compreender os desafios de se trabalhar na área, que incluem: Análise, Captura, Curadoria de Dados, Pesquisa, Compartilhamento, Armazenamento, Transferência, Visualizações e informações acerca da privacidade dos dados (www.deal.com.br).

Esta nova tecnologia veio para modificar a forma como era realizada, até então, a manipulação dos dados, que a cada instante aumenta de tamanho, substituindo a canalização para um **Data Warehouse** de um seleto conjunto de dados, pela análise na locação das informações não necessitando mais a reunião dos mesmos em um único lugar, tarefa esta que soluciona o problema de lentidão que é gerado pela movimentação de um grande fluxo de informações.

Ao adotar os preceitos Big Data de escalonamento de dados, a resposta é muito mais rápida e eficiente, já que eles possibilitam uma consulta às informações de maneira mais flexível, não limitando a quantidade de consultas, tampouco restringindo as formas de como serão realizadas as pesquisas (COSTA *et al.*, 2012).

Algoritmos cada vez mais complexos e eficientes, capazes de encontrar padrões em acessos de músicas e fazer sugestões de outras que aquele usuário ainda não conhecia. Indicar possíveis novos amigos em redes sociais mediante uma análise minuciosa em seu rol de amigos e até mesmo relacionar em frações de segundos e gerar um relatório de um mesmo produto ou oriundo de lugares distintos para que o usuário possa realizar uma comparação, facilitando assim a compra daquele serviço ou produto (PETRY; VILICIC, 2013, p.76).

De acordo com o relatório do *World Economic Forum*, o celular vem sendo considerado a grande promessa de gerador de conteúdo para a utilização dos processos de *Big Data* na criação de novas tendências, novas formas de gerar lucros, já que o aparelho de telefonia móvel é um produto que pode ser adquirido por indivíduos de todas as classes sociais. O celular se apresenta como o produto que mais contribui para a formulação de conteúdo, ele permite que as grandes empresas e organizações consigam chegar até os níveis mais baixos de uma sociedade (WEF, 2012, p.2).

Outra ferramenta que as grandes organizações utilizam, e que médias e pequenas empresas também aderiram, é a disponibilização de produtos gratuitos nas redes sociais com o intuito de gerar, a partir de sua utilização, dados de forma voluntária gerando um perfil de consumidor, contendo suas preferências, necessidades, *hobbys*. Conteúdos esses de grande valor e fonte de matéria-prima para a geração de lucros. (WEF, 2012, p.3).

Big Data não se destina apenas às grandes empresas e multinacionais, as pequenas e médias empresas também podem se beneficiar das vantagens e facilidades que o *Big Data* proporciona aos consumidores dos seus serviços (WEF, 2012, p.2).

Serviços estes que abrangem uma gama imensurável de emprego, podendo ser utilizados na área da saúde, educação, segurança pública, qualidade de vida social, não se limitando apenas ao segmento lucrativo e comercial dos negócios e na busca por mais clientes.

Tal ideia da amplitude da aplicação dos conceitos Big Data fica sintetizada no relatório do *World Economic Forum*, realizado em 2012, que elenca, entre outros, os setores da saúde, educação, serviços financeiros, agricultura, onde o emprego do Big Data agrega eficácia, economia e otimização de recursos.

O **Hadoop** é um projeto que oferece uma solução para problemas relacionados à **Big Data**, tendo em seu núcleo duas partes essenciais: o **Hadoop** Distributed Filesystem (**HDFS**), que é um sistema de arquivos distribuído e confiável,

responsável pelo armazenamento dos dados, e o próprio **Hadoop MapReduce**, responsável pela análise.

Hadoop é uma plataforma de software em Java de computação distribuída voltada para clusters e processamento de grandes massas de dados. Foi inspirada no MapReduce e no GoogleFS (GFS), da Apache.

MapReduce: é um modelo de programação desenhado para processar grandes volumes de dados em paralelo, dividindo o trabalho em um conjunto de tarefas independentes.

Google File System (GFS) é um sistema de arquivos escalável para aplicações de distribuição intensiva de dados. Ele é usado para organizar e manipular grandes arquivos e permitir que aplicações consigam usar os recursos necessários. Exemplos: YouTube, Google Earth, GMail, Google Maps, Google, dentre outras.

4.3. DATA WAREHOUSE

Pode-se considerar o **Data Warehouse** como sendo o precursor dessas Tecnologias, que armazena dados consolidados de diversas fontes, mas interligados pelo ambiente de uma corporação. São os chamados dados estruturados, que têm como principal objetivo a precisão e qualidade, que darão suporte à tomada de decisões de qualquer empresa (www.cetax.com.br).

Historicamente ele é o mais utilizado e, portanto, o mais estudado. Existem diversos autores que escreveram sobre Data Warehouse, talvez os mais conhecidos sejam Bill Inmon e Ralph Kimball, onde cada um defende uma Arquitetura.

Os **Data Warehouse** (Armazém de Dados), são grandes depósitos de dados que armazenam as informações de empresas de forma consolidada. Sua origem data dos anos 80 em instituições acadêmicas que inspiraram os sistemas de data warehouse corporativos. Tal solução evoluiu e hoje integra as funcionalidades essenciais de sistemas de Business Intelligence. Seu grande princípio é integrar dados de diferentes sistemas em atualização periódica de longo prazo, que possibilita a visualização de relatórios de períodos de duração mais prolongada (www.deal.com.br).

As informações organizadas em um Data Warehouse possibilitam a produção de relatórios e análises em séries históricas, dentre outras funcionalidades. Com base nos dados produzidos, a partir de uma base confiável, é possível tomar decisões gerenciais assertivas com embasamento em relatórios precisos e amparados por informações sistematizadas.

O DW é um conceito base para montagem de um sistema de dados utilizados em BI, onde a corporação pode unificar todos os seus sistemas para ter uma

base única para montagem de relatórios, posteriormente atividades de Data Mining (Mineração de Dados) também podem ser aplicadas a esse banco de dados.

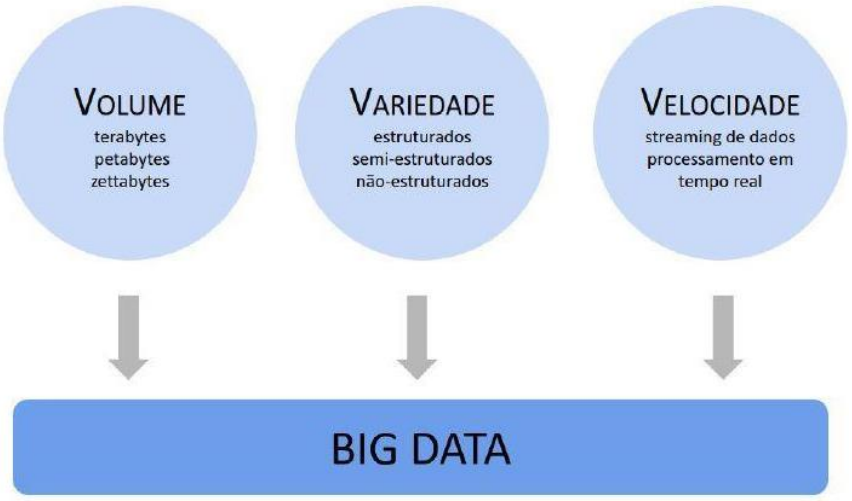
4.4. BUSINESS INTELLIGENCE

O **Business Intelligence** (BI) é a utilização de dados de uma empresa para rotinas de inteligência de negócio, sua visão mais comum é afunilamento ainda maior dos dados coletados do Data Warehouse, que chegam de forma exata e útil para a tomada de decisões. O BI transforma os dados brutos em informações úteis para analisar não só negócios como também as principais estratégias da corporação (www.cetax.com.br).

Business Intelligence (Inteligência de Negócios), é um conceito que define o processo de coleta, organização, análise, compartilhamento e monitoramento de informações que permitem oferecer suporte a gestão de negócios. É importante compreender que o BI explica dados exatos de eventos que já ocorreram (www.deal.com.br).

Em geral, o **Business Intelligence** ou simplesmente BI, é um termo Guarda-Chuva que pode abrigar muitos outros temas, como Data Mining, Reporting, Dashboards (são painéis que mostram métricas e indicadores importantes para alcançar objetivos e metas traçadas) e tudo aquilo que pode ser usado para consolidar inteligência dentro de uma organização.

4.5. CONCEITO DOS 3Vs em BIG DATA



4.5.1. Volume

O volume se refere à quantidade de informações digitais que são produzidas pelos usuários e/ou processos de aplicações. Esta é a característica que fica mais em evidência, já que o volume é a matéria-prima dessa nova tendência. Basicamente os processos *Big Data* são empregados com o objetivo de encontrar novos padrões e tendências em tais volumes de dados, agindo em um campo onde as ferramentas disponíveis no mercado não conseguem trabalhar satisfatoriamente com tamanha quantidade de informações (PETRY; VILICIC, 2013, p.74-75).

Figura 2. Ilustração do Volume

4.5.2. Variedade

Os dados que fomentam esse imenso volume de informações, em sua grande maioria são dados não estruturados, são informações oriundas de fontes mistas de dados, como, por exemplo, fotos, músicas, mensagens, informações geoprocessadas, comentários em redes sociais, histórico de páginas da *Internet*, cadeia de relacionamentos em uma rede social, ativações de leitoras de códigos de barras, entre outras. Essa variedade de informações pode ser alocada em um ***Data Warehouse*** que será alvo do emprego das aplicações de *Big Data* e essas informações poderão possuir alguma relação entre si, gerando um novo padrão ou uma nova tendência.

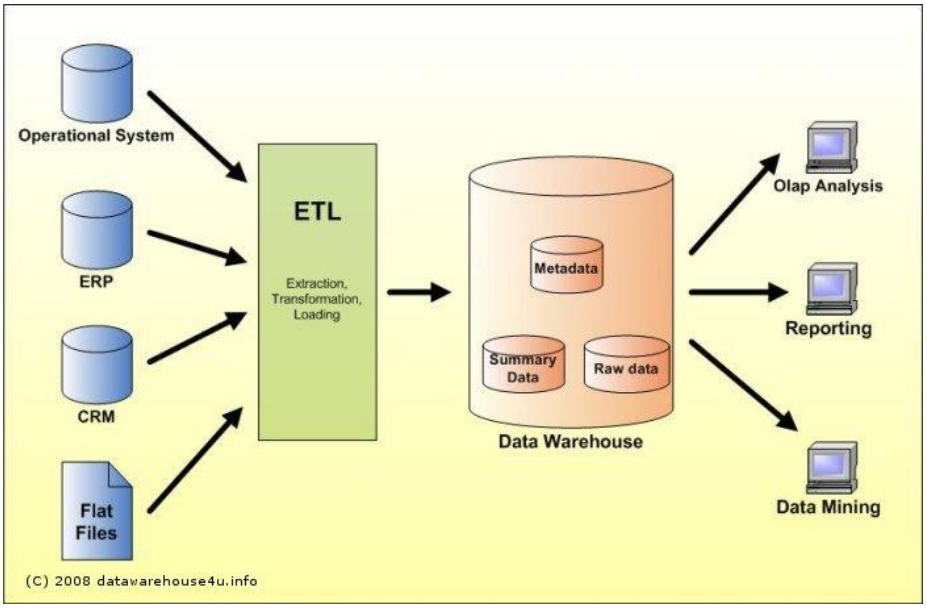


Figura 4. Ilustração da Variedade

4.5.3. Velocidade

A velocidade possui paridade muito grande com o grau de importância dos itens anteriores. Ela tipifica a rapidez com que as informações são criadas, selecionadas e alocadas. Atualmente, a resposta em tempo real se tornou a grande necessidade e exigência de todos os envolvidos em um processo digital, citando como exemplo o setor público, privado, as áreas de telecomunicações, os trabalhadores e clientes, entre outros.

Figura 3. Ilustração da Velocidade

4.6. BOAS PRÁTICAS

4.6.1. Saúde

Estudar um banco de dados alimentado com informações de saúde, onde contenha históricos de doenças em um mesmo paciente, demarcar locais aonde existam incidências de patologias, são os meios para a utilização prática do *Big Data*, com o propósito de economizar recursos e aperfeiçoar trabalhos realizados pelos agentes de saúde.

Com o relatório destes levantamentos e estudos, pode-se traçar e prever surtos de doenças, elaborando campanhas de cunho educativo, conseguir relacionar enfermidades que surjam em consonância com outra que, aparentemente, não possuía relação, além da projeção de possíveis doenças que um mesmo paciente possa vir a contrair devido ao seu padrão de vida e/ou histórico médico (WEF, 2012, p.2).

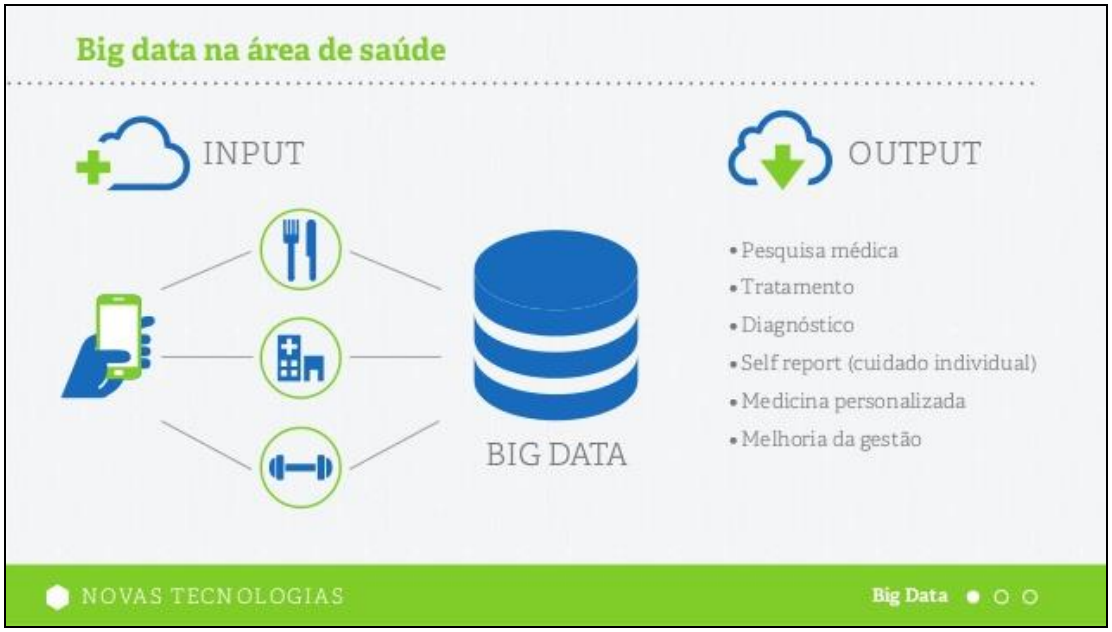


Figura 5. Ilustração de Boas Práticas na Saúde

4.6.2. Educação

Relatórios gerados a partir das técnicas de *Big Data* sobre bancos de dados com informações provenientes da educação, oriundos dos históricos de escolas públicas e privadas, universidades, históricos dos alunos, podem

apontar falhas na forma de ensino, no modo como recursos estão sendo empregados, podendo ser apontados métodos mais eficientes e específicos para uma formação educacional de qualidade.

O emprego das técnicas *Big Data* permite a descoberta de relações entre informações cujos vínculos entre si não poderiam ser visualizados se forem analisadas em pequena escala (WEF, 2012, p.2).

Figura 6. Ilustração de Boas Práticas na Educação

4.6.3. Agricultura

Todos os processos que englobam o setor agrícola de um país, tais como aquisição de insumos e subsídios, comercialização de produtos, pagamentos, condições climáticas, entre outros, quando lançados e armazenados na área digital, possibilitam a aplicação dos processos de *Big Data*.

Tal emprego pode gerar informações de grande relevância para o produtor e até mesmo para o governo, já que podem ser detectadas novas tendências na produção alimentícia e nos investimentos. Além de gerar informações mais precisas sobre capacidade, e, consequentemente, à garantia e disponibilidade de armazenamento da produção, reduzindo os níveis de desperdício e deterioração, préstimos financeiros que se enquadram nas reais necessidades do produtor (WEF, 2012, p.2).

Figura 7. Ilustração de Boas Práticas na Agricultura

4.6.4. Serviços Financeiros

A grande quantidade de informações digitais gerada de históricos de compras realizadas na *Internet* permite, após analisada, criar perfis dos consumidores. Determinar hábitos financeiros, revelando formas eficientes de abordagem para a venda de linhas de crédito financeiro.

Mapear tipos de serviços econômicos limitando-os em regiões e zonas, possibilitando a formulação de produtos financeiros mais específicos para atender as necessidades dos clientes (WEF, 2012, p.2).

Figura 8. Ilustração de Boas Práticas em Serviços Financeiros

4.6.5. Direito do Consumidor

A facilidade de aquisição de produtos e serviços digitais, somando-se com a produção quase que incontrolável e despercebida de informações via celular, *e-mail*, redes sociais, vídeo conferências, pelos usuários, torna-se necessária à criação e regulamentação de leis que defendam os direitos de privacidade e propriedade das informações particulares de cada cliente.

Tal medida se pauta na preservação e no resguardo de possíveis crimes como o roubo ou utilização indevida de informações privadas (Marco Civil da Internet). Outra medida que pode ser adotada, essa pelo poder público, é o incentivo correto e direcionado às empresas, motivando-as a fazer o uso dos dados em prol da sociedade (WEF, 2012, p.3).

Figura 9. Ilustração de Boas Práticas em Direito do Consumidor

4.6.6. Competividade Comercial

Com a imersão no mundo *Big Data*, as grandes organizações conseguirão produzir aplicações cada vez mais completas, acessíveis, eficientes e de fácil manuseio, deixando atrativa a sua compra por parte da sociedade que busca tais benefícios.

As grandes organizações que, em sua grande maioria são do setor privado, devem ser incentivadas a trabalharem também com a produção de *softwares* com código aberto (*Open Source*), visto que, os programas *Open Source* permitem a construção de novas ideias e o aperfeiçoamento das tecnologias, além de ser uma saída para a quebra do monopólio (WEF, 2012, p.3

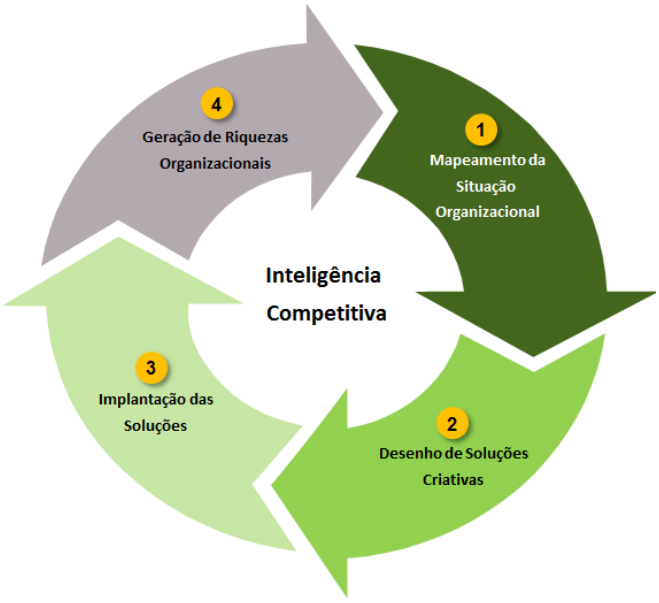


Figura 10. Ilustração de Boas Práticas em Competividade Comercial

4.6.7. Mão de Obra Especializada

Os processos que englobam o *Big Data* necessitam de profissionais extremamente capacitados a realizar as tarefas de mineração e análise de dados (Analista de Dados), na maioria das vezes competem tais atribuições aos cientistas e engenheiros da computação.

Figura 11. Ilustração de Boas Práticas em Mão de Obra Especializada

Porém, os altos custos na contratação - e até mesmo a escassez desses trabalhadores no mercado - atrapalham os avanços na área de *Big Data*. Devido a esse empecilho, até mesmo as grandes organizações encontram dificuldades para ter acesso aos conhecimentos necessários para criar novos padrões e novas técnicas de coleta, análise, mineração de dados.

4.7. TIPOS DE DADOS E PERSONAGENS

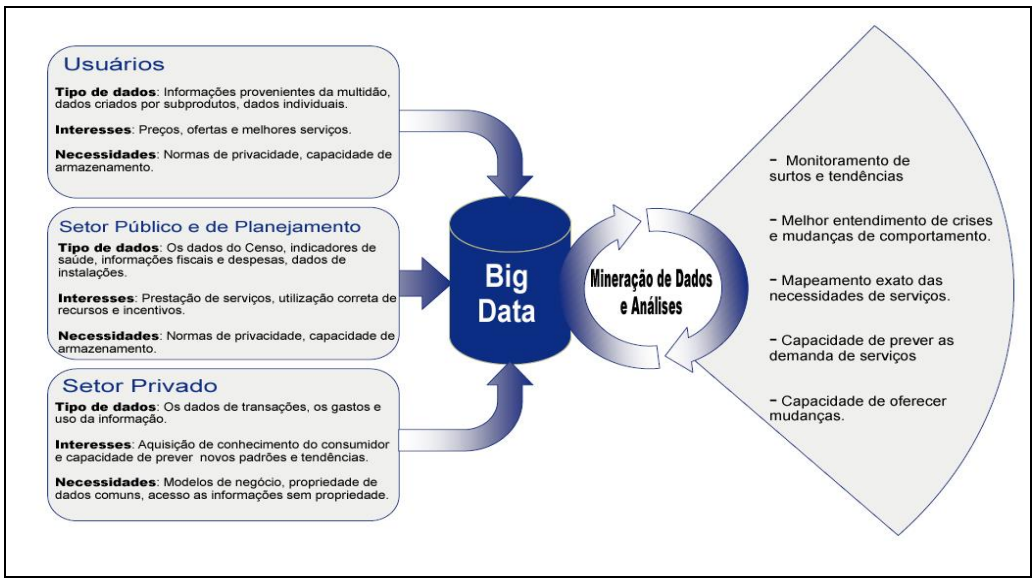


Figura 12. Tipos de Dados e Personagens

4.8. ARMAZENAMENTO DE DADOS

De acordo com White (2010, p.3), o grande desafio na implementação de *Big Data* fica atrelado à forma de armazenamento e acesso às informações.

Atualmente as unidades possuem capacidade de armazenar em média 1 *terabyte*, mas a velocidade de transferência não seguiu a mesma proporção evolutiva, sendo calculada em aproximadamente 100 Mb/s, o que levaria cerca de 2 horas e meia para executar o processo de leitura de todo o *terabyte* de dados, esse tempo aumenta ainda mais quando a tarefa se torna a escrita dos dados no disco.

Segundo Costa *et al.* (2012), arquiteturas do tipo “nuvem computacional” (*cloud computing*) são os sistemas mais adequados para trabalhar com essa grande quantidade de dados que vem sendo gerada, já que suas estruturas são de baixo valor aquisitivo, com um grau elevado em escalabilidade, flexível e ágil, além da facilidade de execução no tocante à manipulação e ao controle de enormes quantidades de informações.

O armazenamento em nuvens, até o ano de 2020, será o método mais utilizado para alocar informações digitais. As grandes potências na era tecnológica (*Amazon*, *IBM*, *Facebook*, *Microsoft*, entre outras), projetam investimentos para aumentar seus centros de dados, que operam de forma distribuída, replicando dados com o intuito de reduzir o tempo de resposta às requisições feitas pelos usuários (COSTA *et al.*, 2012).

O **Hadoop** é um projeto que oferece uma solução para problemas relacionados à Big Data, tendo em seu núcleo duas partes essenciais: o Hadoop Distributed Filesystem (HDFS), que é um sistema de arquivos distribuído e confiável, responsável pelo armazenamento dos dados, e o próprio **Hadoop MapReduce**, responsável pela análise e processamento dos dados. Ambos possuem a confiabilidade como uma marca, o que torna o sistema muito robusto para aplicações que envolvem dados massivos e importantes para as organizações que o utilizam.

Figura 13. Armazenamento de Dados

4.9. SEGURANÇA DOS DADOS

Segundo Petry e Vilicic (2013), atualmente cada usuário da grandiosa rede mundial de computadores possui uma vida digital, onde os eventos cotidianos do individuo, em sua grande maioria, são registrados por meio de fotografias e vídeos, depois são armazenados digitalmente, gerando um perfil digital de cada individuo, o que acarreta em uma grande vulnerabilidade e excesso de exposição da vida particular. O cuidado com os dados divulgados *online* tem que ser redobrado, já que muitas vezes estas informações estão em maior quantidade se comparadas com as informações que estão armazenadas dentro de casa (EMC, 2012).

Uma pesquisa realizada pela universidade de Cambridge conseguiu espantosos resultados na utilização de técnicas *Big Data*, relacionando informações disponibilizadas no *Facebook* com o intuito de descobrir informações omitidas. Segundo o estudo 95% dos testes, foi possível descobrir a etnia do usuário (cor/raça), em 88% dos resultados revelou o sexo do individuo; e a posição política e religiosa foi descoberta em 80% dos casos (PETRY; VILICIC, 2013, p.76).

A segurança em *Big Data* acaba se tornando uma vertente muito delicada, já que a ambição de conhecimento e criação de novas técnicas para manipular e até mesmo aperfeiçoar as ferramentas *Big Data* existentes atualmente no mercado causam uma vulnerabilidade

acentuada, principalmente nos casos onde os códigos dos *softwares* são abertos e de fácil acesso a todos.

Figura 14. Segurança dos Dados

4.10. ARQUITETURA GERAL