**NÍCOLAS HENRIQUE VIEIRA TOLEDO**

**SEBASTIÃO BATISTA DE ANDRADE NETO**

**ESTATÍSTICAS POLÍTICAS**

**EM TEMPO REAL COM BIG DATA**

**UNIVERSIDADE DO VALE DO SAPUCAI**

**POUSO ALEGRE**

**2015**

**NÍCOLAS HENRIQUE VIEIRA TOLEDO**

**SEBASTIÃO BATISTA DE ANDRADE NETO**

**ESTATÍSTICAS POLÍTICAS**

**EM TEMPO REAL COM BIG DATA**

Pré-projeto de pesquisa apresentado à disciplina de TCC 1 do Curso de Sistemas de Informação como requisito parcial para obtenção de créditos.

**UNIVERSIDADE DO VALE DO SAPUCAI**

**POUSO ALEGRE**

**2015**

# QUADRO TEÓRICO

## Big Data

De acordo com Emerson Alecrim (2013), a princípio, podemos definir o conceito de Big Data como sendo conjuntos de dados extremamente amplos e que, por este motivo, necessitam de ferramentas especialmente preparadas para lidar com grandes volumes, de forma que toda e qualquer informação nestes meios possa ser encontrada, analisada e aproveitada em tempo hábil.

De maneira mais simplista, a ideia também pode ser compreendida como a análise de uma massa de dados para a geração de resultados relevantes que, em volumes menores, dificilmente seriam alcançados.

## Banco de dados NoSQL

De acordo com a revista Java Magazine (2013), NoSQL é um movimento que promove soluções de armazenamento de dados não relacionais. Ele é composto por diversas ferramentas que, de forma particular e específica, resolvem problemas como tratamento de grande volume de dados, execução de consultas com baixa latência e modelos flexíveis de armazenamento de dados, como documentos XML ou JSON.

NoSQL não têm como objetivo substituir os bancos de dados relacionais, mas apenas propor algumas soluções que em determinados cenários são mais adequadas. Desta forma é possível trabalhar com tecnologias NoSQL e banco de dados relacionais dentro de uma mesma aplicação.

Os tipos de bancos de dados NoSQL que são mais usados são:

* Chave-Valor: Armazenam objetos indexados por chaves, e possibilitam sua busca a partir de sua respectiva chave.
* Documentos: Conjunto de documentos no formato JSON. Os documentos são tratados como objetos únicos, os quais possuem campos com os respectivos valores.
* Coluna: Formado por colunas que contem um conjunto de informações, semelhante a uma tabela.
* Grafos: Formado por dados distribuídos em forma de vértices e arestas, os quais possuem atributos tanto nas arestas quanto nos vértices.

Os bancos de dados NoSQL são implantados quando os bancos de dados relacionais já não estão suportando a demanda de dados, ou seja, estão apresentando lentidão em consultas e processamentos. Um NoSQL pode substituir um banco de dados relacional por completo, ou simplesmente, ser implantado como um “suporte”, realizando as consultas/processamentos em um grande volume de dados onde a velocidade é necessária. O tipo de NoSQL a ser é escolhido após uma análise crítica do gestor de TI da organização, dependendo da necessidade da organização.

## Hadoop

Segundo Avoyan Hovhannes (2013), o Hadoop é uma implementação de código aberto do paradigma de programação Map-Reduce. Map-Reduce é um paradigma de programação introduzido pelo Google para processar e analisar grandes conjuntos de dados. Todos esses programas que são desenvolvidos nesse paradigma realizam o processamento paralelo de conjuntos de dados e podem, portanto, ser executados em servidores sem muito esforço. A razão para a escalabilidade desse paradigma é a natureza intrinsecamente distribuída do funcionamento da solução. Uma grande tarefa é dividida em várias tarefas pequenas que são então executadas em paralelo em máquinas diferentes e então combinadas para chegar à solução da tarefa maior que deu início a tudo.

## Implantação das Tecnologias

Segundo Marcos Pichatelli (2013) O Big Data pode se tornar caro de processar e armazenar se implantado em bancos de dados tradicionais. Para resolver esse problema novas tecnologias usam soluções open source e plataformas de hardware de custo acessível para armazenar os dados de maneira mais eficiente, paralelizar trabalhos e entregar poder de processamento.

Ainda segundo Marcos Pichatelli (2013) alguns passos necessários para se conseguir o pleno potencial de Big Data:

* **Coletar:** O dado é coletado das fontes de informação e distribuído por meio de múltiplos nós, por exemplo em um arquitetura grid, cada um dos quais processa um subconjunto de dados em paralelo.
* **Processar:** O sistema então usa o mesmo paralelismo gerenciado para ter um desempenho computacional mais rápido em cada nó. Depois, cada nó transforma os resultados das pesquisas em informações mais consumíveis para serem usadas tanto pelos seres humanos (em caso de análise) quando pelas máquinas (em caso de interpretação de resultados em larga escala).
* **Gerenciar:** Geralmente o processamento de Big Data é heterogêneo, originado a partir de diferentes sistemas transacionais. Quase todos os dados precisam ser entendidos, definidos, anotados, limpos e auditados para fins de segurança.
* **Medir:** As análises de negócios devem determinar uma métrica e devem ser acompanhadas constantemente. Geralmente as companhias medem o quanto um dado pode ser integrado/relacionado com um comportamento de consumo ou registro histórico; e como essa integração ou correção aumenta ou diminui com o tempo.
* **Consumir:** O resultado da análise dos dados deve atender a demanda original. Por exemplo, se o resultado for de algumas centenas de terabytes de interações em redes sociais, ele pode demonstrar como seus clientes compram produtos complementares. Então, deve haver regras de como os dados de mídias sociais são acessados e atualizados. O mesmo serve para o acesso de dados máquina-a-máquina (M2M).
* **Armazenar:** Como a tendência “data-as-a-service” ainda toma forma, cada vez mais os dados permanecem em um único lugar, enquanto os programas de acesso a essas informações se movem. Mesmo que os dados sejam armazenados para o curto prazo de processamento em lote ou para o longo prazo de retenção, as soluções de armazenamento devem ser deliberadamente dirigidas.
* **Governar:** A governança de dados engloba as políticas e fiscalização de informações por meio de uma perspectiva de negócios. Como definido, a governança de dados se aplica a cada um dos seis estágios de entrega de Big Data.

Realmente, o Big Data não é uma solução barata e de fácil implantação. É necessário haver um estudo de caso, para saber se a organização realmente precisa dessa tecnologia. Além do mais, é necessário ter profissionais especializados na área para ocorrer a implantação, e obter