Aula 1

Big Data

Conversa Inicial

Prof. Luis Henrique Alves Lourenço

1

2

Era dos dados

Principais conceitos

Etapas

3

4

O avanço da tecnologia permitiu tanto que fossem gerados um volume imenso de dados quanto a análise de tais dados Transição entre o mundo analógico e o mundo digital

 "Em 1995 menos de 1% dos dados estavam armazenados em formato digital" (Marquesone, 2016)

 "Em 2007 a quantidade de dados digitais já era de 94%" (Marquesone, 2016)

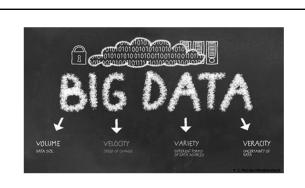
Novas tecnologias

- A influência da internet no volume de dados produzidos diariamente
 - Diversos novos serviços
- Dispositivos móveis
 - Novas formas de interação

- Internet das Coisas
 - Centenas de bilhões de dispositivos conectados
- Lei de Moore
 - "A capacidade de processamento dobra a cada 18 meses"

7 8

Os Vs em Big Data



9 10

Volume

Como extrair valor de um volume imenso de dados?

Variedade

- Diversas fontes de dados em vários formatos
- Dados estruturados
- Dados semiestruturados
- Dados não estruturados

11 12

Velocidade

- Dados são produzidos o tempo todo
- Muitos dados perdem o valor com o tempo
- Soluções podem exigir respostas em tempo real

Valor

- Maior valor potencial
- Quais dados devem ser priorizados?

13 14

Mais Vs?

- Variabilidade
- Validade
- Vulnerabilidade
- Volatilidade
- Visualização
- **-** ...

Definição

Big Data é o conjunto de práticas e técnicas que envolvem a coleta e análise de imensos volumes de dados confiáveis e variados com a velocidade necessária para a extração de informações valiosas

15 16

Obtenção e armazenamento de dados

Obtenção dos dados

- Dados internos
- Dataficação
- Dados de sensores
- Dados de fontes externas

Armazenamento

- Escalabilidade
- Disponibilidade
- Flexibilidade

NoSQL

- Dados semiestruturados ou não estruturados
- Bancos de dados não relacionais
 - Chave-valor
 - Documentos
 - Colunas
 - Grafos

19 20

Processamento de dados

Escalabilidade

- Aumentar a capacidade de processamento
- Escalabilidade vertical
- Escalabilidade horizontal

22 21

Hadoop

- Baixo custo
- Escalabilidade
- Tolerância a falhas (partição)
- Balanceamento de carga
- Comunicação entre máquinas
- Alocação de máquinas

Tecnologias distribuídas

- HDFS
- MapReduce

Processamento em tempo real

- Baixa latência
- Consistência
- Alta disponibilidade

Spark

- Spark Core
- GraphX
- Spark SQL
- MLib

25 26

Componentes Spark

- Driver program
- Cluster manager
- Workers

Análise e visualização

27 28

Análise de dados

- A evolução da tecnologia permitiu o armazenamento e processamento de dados que antes eram ignorados
- Um cientista de dados deve ser capaz de interpretar os dados

Processo de análise de dados

- Entendimento do negócio
- Compreensão dos dados
- Preparação dos dados
- Modelagem dos dados
- Avaliação do modelo
- Utilização do modelo

Visualização de dados

- Visualização exploratória
- Visualização explanatória

Processo de visualização de dados

- Aquisição
- Estruturação
- Filtragem
- Mineração
- Representação
- Refinamento
- Interação