Aula 1

Big Data

Prof. Luis Henrique Alves Lourenço

1

Conversa Inicial

2

Big Data

- **Contextualização**
- **Principais conceitos**
- **Etapas**

3

Era dos dados

4

- **O avanço da tecnologia permitiu tanto que fossem gerados um volume imenso de dados quanto a análise de tais dados**

5

- **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**

6

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

7

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

8

Os Vs em Big Data

9

10

Volume

- **Como extrair valor de um volume imenso de dados?**

11

Variedade

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

12

Velocidade

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

13

Valor

- Maior valor potencial
- Quais dados devem ser priorizados?

14

Mais Vs?

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

15

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

16

Obtenção e armazenamento de dados

17

Obtenção dos dados

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

18

Armazenamento

- Escalabilidade
- Disponibilidade
- Flexibilidade

19

NoSQL

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

20

Processamento de dados

21

Escalabilidade

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

22

Hadoop

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

23

- Tecnologias distribuídas
  - HDFS
  - MapReduce

24

Processamento em tempo real

- Baixa latência
- Consistência
- Alta disponibilidade

25

Spark

- Spark Core
- GraphX
- Spark SQL
- MLib

26

- Componentes Spark
  - *Driver program*
  - *Cluster manager*
  - *Workers*

27

Análise e visualização

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

29

Processo de análise de dados

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

30

Visualização de dados

- Visualização exploratória
- Visualização explanatória

31

Processo de visualização de dados

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

32