BIG DATA

INTRODUÇÃO

Humberto Sandmann
humberto.sandmann@gmail.com

Humberto Sandmann

humberto.sandmann@gmail.com

Possui graduação em Ciências da Computação pelo Centro Universitário da Faculdade de Engenharia Industrial (FEI), mestrado e doutorado em Engenharia Elétrica pela Universidade de São Paulo, com parte do doutorado realizado no Max Planck Institute for Dynamics and Self-organization em Göttingen (Alemanha). Tem experiência e estudos realizados na área de sistemas dinâmicos em redes neurais artificiais e processamento de sinais biológicos. Além disso, atua na área de ciência da computação, com ênfase em inteligência artificial. As principais áreas de esforços são: sistemas dinâmicos, redes neurais, processamento de sinais, reconhecimento de padrões e aprendizado de máquina. Atualmente, é sócio em uma empresa de tecnologia, a Selsantech, que atua na área de IoT (internet das coisas), inteligência computacional e big data, também, é professor no MBA da FIAP e Toledo e na gradução da FIAP e ESPM

Roteiro do curso

Aula 1: Introdução

Conceitos, mercado, tendências e arquitetura

Aula 2: Ferramentas

Hadoop, MongoDB, Neo4j e Solr

Aula 3: HBase, Hive e tratamento de dados

Aula 4: Machine learning

Perceptron - MLP, DeepLearning, SOM,

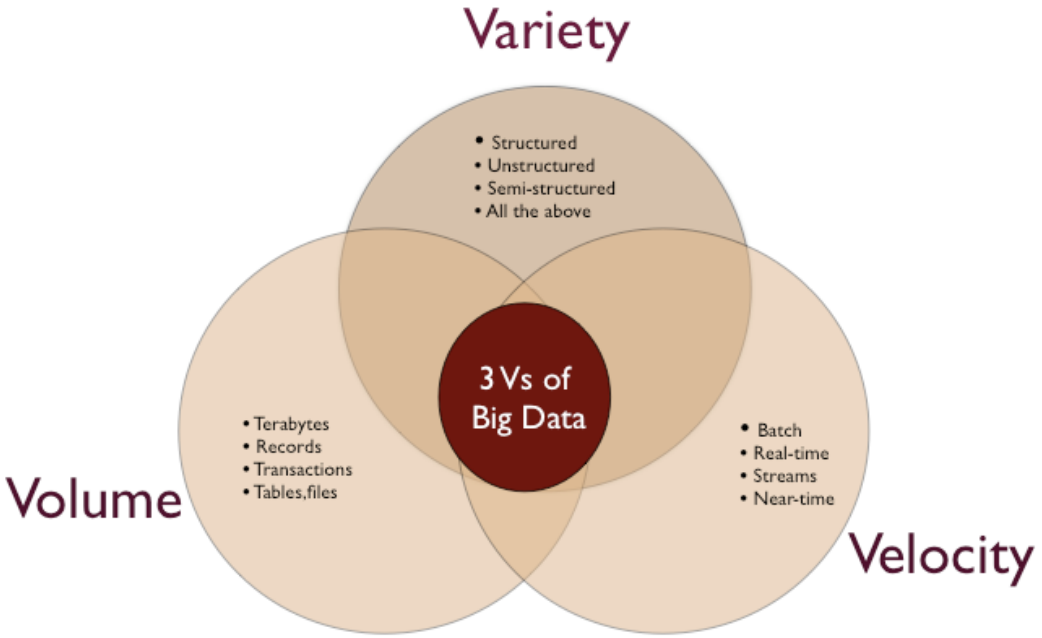
Redes probabilística (Redes Bayesianas)

Definição

Volume

Variedade

Velocidade



Conceitos dos 3Vs

Big Volume

- estruturas simples (SQL)
- estruturas complexas (non-SQL)

Big Variedade

- um grande número de fonte de dados a serem integradas

Big Velocidade

- diversas fontes provendo dados simultaneamente

Domínio de atuação

Análises são feitas em:

- centenas ou milhares de nós
- Petabytes de dados

Big Análises

Operações matemáticas complexas

- aprendizado de máquina (machine learning)
- agrupamento (clustering)
- detecção de tendências (predição)

Complexidade computacional alta (n^x)

- Multiplicidade matricial
- Decomposições matriciais
- Regressões lineares ou não-lineares

Exemplo

Estudo de relação de covariância dentre duas ações ao longo de um intervalo de tempo

- Ações diárias A e B
- 10 anos (~2000 dias)

  Resultado: 2 x 2000 dados para cálculo recursivo

Porém, quando pensamos em 4000 ações:

  4000 x 2000 dados a serem cálculados recursivamente

Exemplo de aplicações

Seguro de automóvel

- sensores em carros

  (comportamento de direção, localizações de risco, etc)

Customização de propagandas

Aplicações científicas

- genoma, imagens de satélites, astronomia, previsão do tempo, neurociência, etc.

Mudança de Domínio da Análise

Small Math  Big Math

muitos domínios

vem para ficar

Ferramentas

- Matlab (Octave)
- R
- SAS
- Microstrategy

Pontos importantes

- Gerenciamento de dados fraco ou não existente
- Sistema de arquivo de Armazenamento
- Escalável e paralelizável

Hadoop

Gerencia e centraliza o acesso a dados espalhados por diversos nós de armazenamento (+4000 por exemplo)

- Implementa o design pattern map/reduce
  - Yarn: ferramenta para gerenciamento do cluster
  - Map/reduce
- Implementa um repositório unificado de nós
  - HDFS: Hadoop File System
- Apenas viável se a performance não for problema
- OpenSource: server de referência para outras implementações

Bibliografia

Apache Hadoop
http://hadoop.apache.org/

MIT Big Data Iniciative
http://bigdata.csail.mit.edu/