Big Data – Hands On

# Objetivo

Este treinamento prático “mão na massa” ao mundo Big Data oferece uma abordagem exclusiva para ajudá-lo a agir sobre os dados (Data Mining) para obter ganhos reais nos negócios. O foco não é o que uma ferramenta pode fazer, mas o que você pode fazer com a saída da ferramenta. Obtenha as habilidades necessárias para armazenar, gerenciar, processar e analisar grandes quantidades de dados, desenvolver modelos utilizando tecnicas de machine learning (Regressão Logística, Redes Neurais, SVM, Random Forest, Gradient Boosting, XGBoost, entre outras). Tudo isso sobre todo Ecossistema Hadoop(HDFS, YARN/MapReduce, Hive, Spark, Pig, Sqoop, entre outras) e serviços disponíveis da AWS(Clusters, Lambda, Machine Learning, entre outras).

# Requisitos

Recomenda-se alguma experiência com linguagens de programação (R, Python, SAS, entre outras) e conhecimentos básicos em estatística.

# Conteúdo

1. Introdução ao Big Data
   1. O que é ?
   2. Onde é utilizado e por quem ?
   3. Como gerar valor com essa tecnologia ?
   4. Overview Aplicações de Machine Learning (Cenários Reais)
2. Arquitetura Big Data (Laboratórios 01 e 02)
   1. Visão geral do Ecossistema Big Data
      1. Hadoop
      2. Teradata + SAS
   2. Configuração de Cluster Hadoop
   3. Ferramentas de monitaramento do Cluster (Ganglia e CloudWatch)
   4. Configuração/tipo do Cluster otimizado para diferentes aplicações de Big Data
   5. Otimização de custos utilizando AWS EMR Spot
   6. Ferramentas para Mineração e Visualização dos dados
      1. Zeppelin
      2. Jupyter
3. Processamento de Grande Volume de Dados (Laboratorio 03)
   1. Tipos de armazenamento dos dados
      1. Csv
      2. Parquet
      3. Json
   2. Extração, Transformação e Carga dos Dados (ETL)
      1. Conversão de CSV para Parquet
      2. Conversão de tipo de dado (numerico, string, data, timestamp)
      3. Particionamento
      4. Armazenamento (S3 | RedShift | RDS | Teradata)
4. Mineração dos Dados (Data Mining) (Laboratório 04)
   1. Introdução ao Aprendizado de Máquina (Machine Learning)
      1. Tipos de Dados
      2. Visualização dos Dados
      3. Tratamento dos Dados
      4. Seleção de Variáveis
      5. Modelos Supervisionados
         1. Naive Bayes
         2. Regressão Linear
         3. Regressão Logística
         4. Árvore de Classificação/Regressão
         5. Random Forest | Gradient Boosting | XGBoost
         6. Redes Neurais
      6. Modelos Não Supervisionados (Clusterização)
         1. K-means
         2. Hierarquico
         3. Fuzzy
         4. DBSCAN
      7. Avaliação de Desempenho dos Modelos
      8. Implantação de Modelos
      9. Escoragem Online
      10. Escoragem Batch
5. Case de Negócio (Laboratório 05)

Neste laboratório é dado ao aluno um caso real de aplicação de problema de Big Data e utilizaremos todas ferramentas apresentadas no curso para resolver o problema. No final medimos o valor gerado ao negócio por cada solução e premiamos o vencedor com chocolate.

* 1. Entendimento do problema de negócio
  2. Desenvolvimento do Modelo
  3. Avaliação do modelo
  4. Aplicação
  5. Entrega