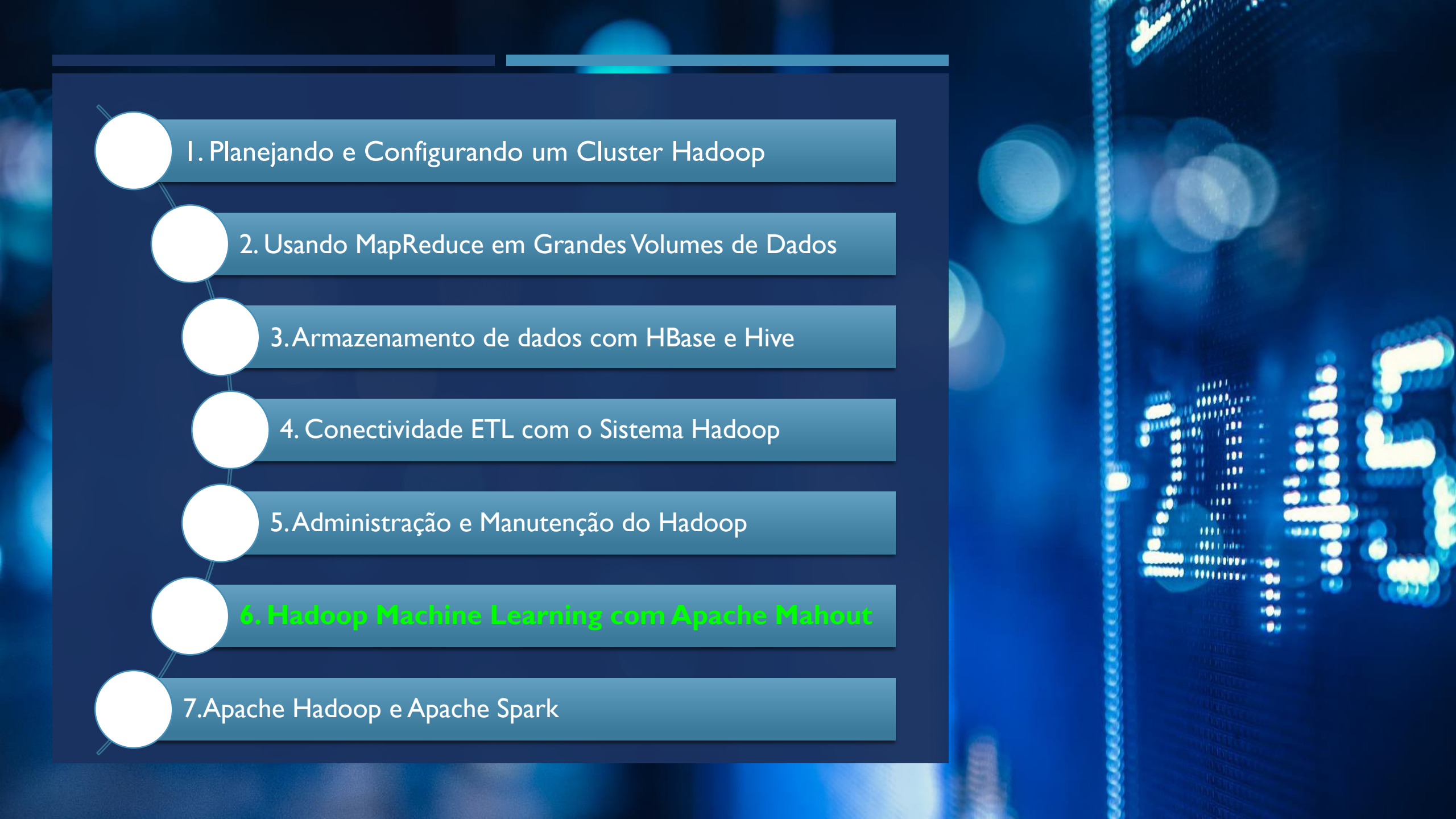
ENG DE DADOS COM HADOOP E SPARK 6


1. Planejando e Configurando um Cluster Hadoop
2. Usando MapReduce em Grandes Volumes de Dados
3. Armazenamento de dados com HBase e Hive
4. Conectividade ETL com o Sistema Hadoop
5. Administração e Manutenção do Hadoop
6. Hadoop Machine Learning com Apache Mahout
7. Apache Hadoop e Apache Spark

6.1 CONHECENDO O APACHE MAHOUT

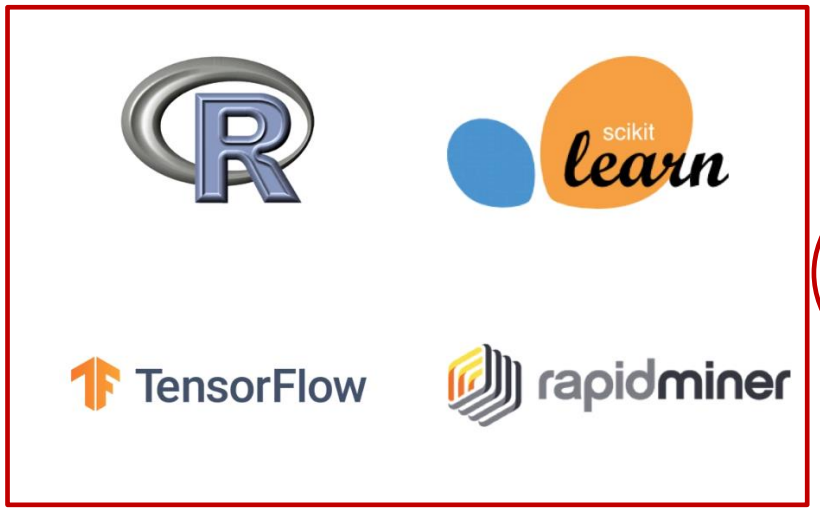
Principais Algoritmos de Machine Learning disponíveis no Apache Mahout:

- Algoritmos de Classificação
- Sistemas de Recomendação
- Clustering
- Redução de Dimensionalidade

Tipos de Algoritmos Suportados:

- Algoritmos Sequenciais
  - Regressão Logística
  - Modelos Ocultos de Markov
  - Perceptrons de Multi-camadas
- Algoritmos Paralelos
  - Random Forest
  - Naive Bayes
  - K-Means

6.2 APACHE MAHOUT X OUTROS FRAMEWORKS DE MACHINE LEARNING



- Foi criado para manipular grandes conjuntos de dados!
- Possui uma forte integração com Apache Spark!
- Os algoritmos do Mahout são escritos para funcionar sobre o Hadoop e dessa forma, eles funcionam em ambiente distribuído.
- O Framework Mahout está pronto para uso e permite realizar mineração de dados em grandes conjuntos de dados.
- Ele é eficiente na análise de grandes conjuntos de dados.
- Suporta a execução do algoritmo de classificação Naïve Bayes de forma distribuída.

- Possui diversas implementações de Clustering, Como: K-means, Fuzzy K-Means, Canopy e Mean-Shift.
- Inclui bibliotecas para manipulações de vetores e matrizes.

INSTALAÇÃO E CONFIGURAÇÃO DO APACHE MAHOUT EM ANEXO

1 – Modelo de Classificação com Naïve Bayes

2 – Modelo de Clusterização com K-Means

O Apache Mahout é uma biblioteca machine learning, open source que tem como principais objetivos operar como uma máquina de recomendação, clustering e classificação.

Quando falamos de algoritmos de Machine Learning é importante destacar que nem todos serão capazes de executar sobre um ambiente de cluster, simplesmente porque não foram criados para processamento paralelo e distribuído.

Os algoritmos do Mahout são escritos para funcionar sobre o Hadoop e dessa forma, eles funcionam em ambiente distribuído.

THANKS