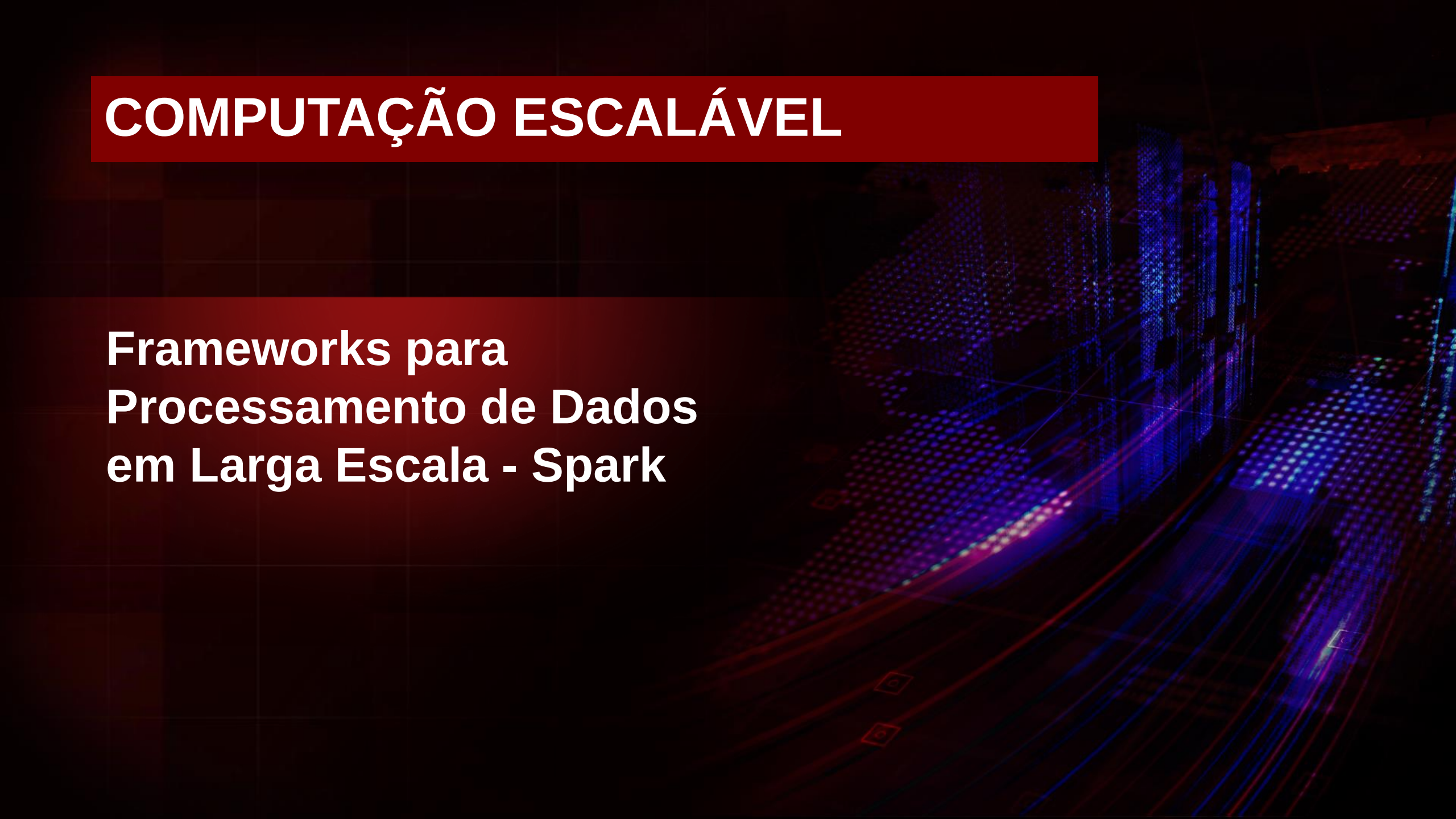
COMPUTAÇÃO ESCALÁVEL

Frameworks para Processamento de Dados em Larga Escala - Spark



ROTEIRO

- Motivação
- O que é Apache Spark?
- Como funciona?
- Ecossistema do Spark
- Benefícios

MOTIVAÇÃO

- Aumento de datasets na Web
  - Cliques
  - Requisições de servidores
  - Dados de localização de dispositivos móveis
  - Dados de grafos
    - Redes de telefonia
    - Redes sociais
    - Redes de computadores
  - Internet das Coisas
    - RFID
    - Sensores
  - Dados financeiros

MOTIVAÇÃO

- Dados, muitos dados…
  - Armazenamento de conteúdo multimídia (Youtube)
  - Bolsas de valores
  - Dados de usuários de redes sociais (Instagram, Facebook, etc)
- Grande difusão do armazenamento baseado em nuvem e dos recursos computacionais
  - Para o processamento de grandes datasets
- Já vimos as características de Cloud
  - Mais barata
  - Escala dinamicamente
  - Computação sob demanda

O QUE É O APACHE SPARK?

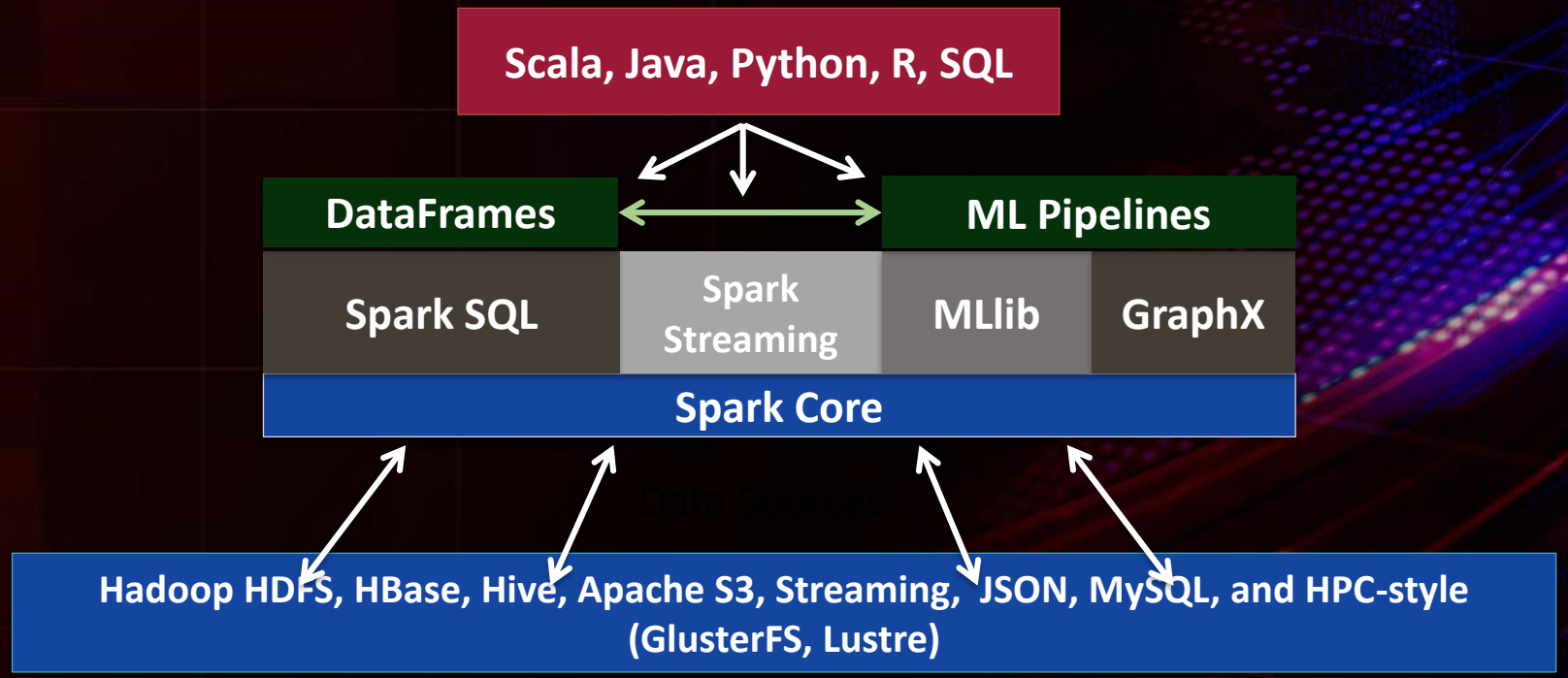
- É uma alternativa ao Map Reduce para algumas aplicações
- Um sistema de computação em cluster com baixa latência
- Usado para grandes conjuntos de dados
- Usado em conjunto com o Hadoop FS
- O MapReduce permite a análise de big data usando clusters grandes e não confiáveis
- MapReduce carece de primitivas eficientes para compartilhamento de dados

O QUE É O APACHE SPARK?

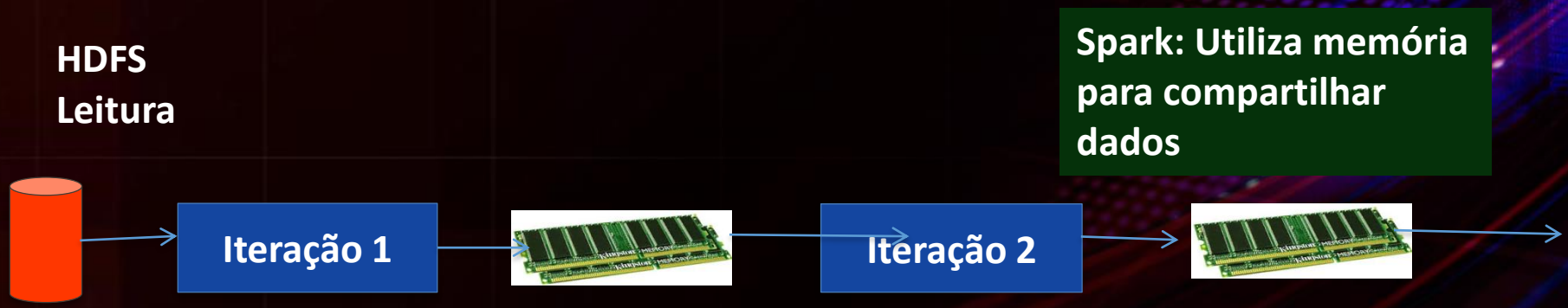
- É neste contexto que o Spark entra em cena
  - Ao invés de carregar os dados do disco para cada consulta, por que não fazer o compartilhamento de dados na memória?
  - Acesso à memória é mais rápido se comparado ao acesso ao disco
- Há APIs para linguagens como:
  - Java
  - Scala
  - Python
- Cache de dados na memória
  (para algoritmos iterativos, gráficos e de aprendizado de máquina, etc.)

O QUE É O APACHE SPARK?

- O Apache Spark suporta análise de dados, aprendizado de máquina, gráficos, dados de streaming, etc. Ele pode ler/gravar de uma variedade de tipos de dados e permite o desenvolvimento em várias linguagens



COMO FUNCIONA?



COMO FUNCIONA?

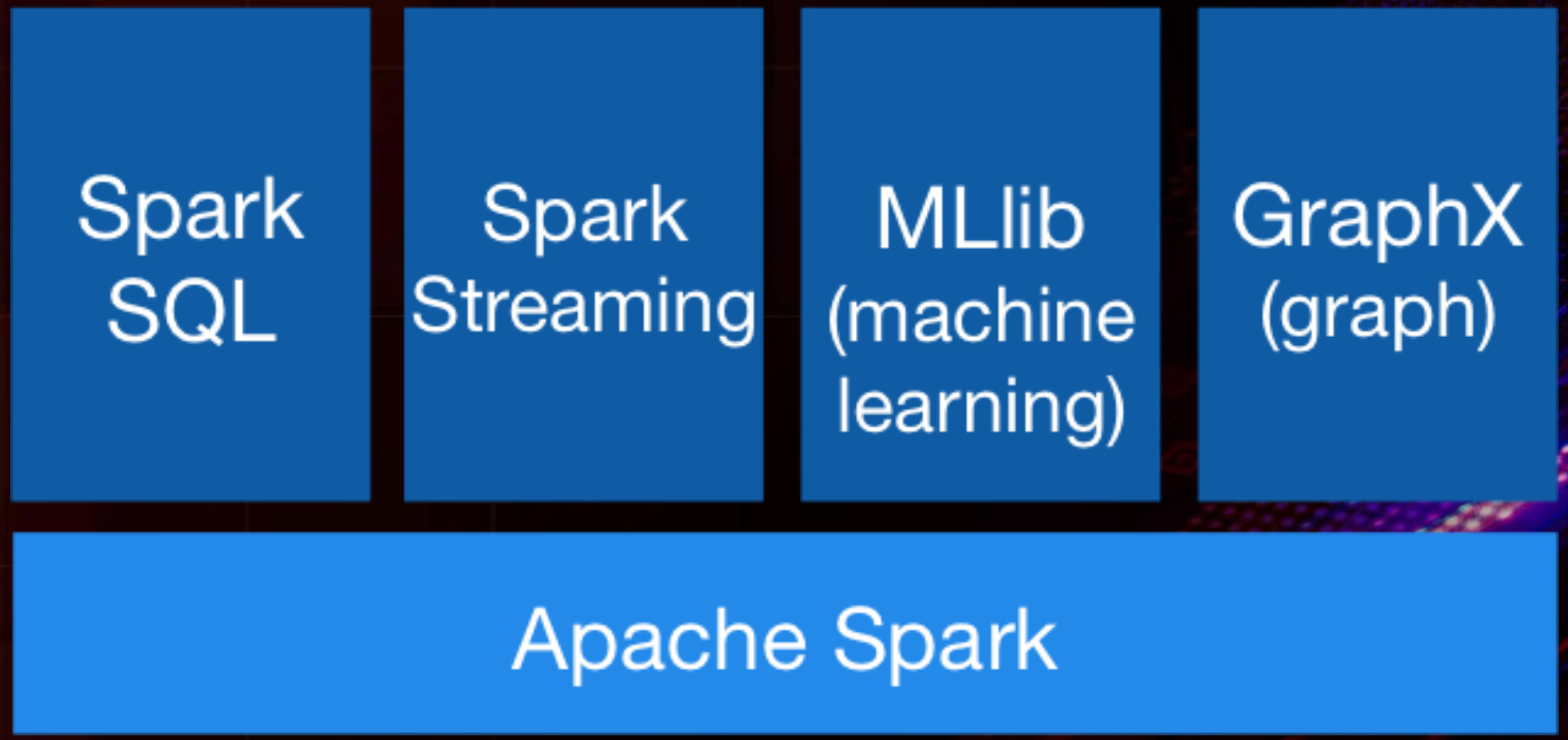
- O Spark aborda o problema do acesso compartilhado aos dados com:
  - Conjuntos de dados distribuídos resilientes (RDDs)
  - O RDD permite que as aplicações mantenham conjuntos de trabalho na memória para reutilização
- RDD se destaca como um modelo de programação
  - Tolerante a falhas
  - Computação distribuída
  - Compartilhamento em memória

- *Spark é uma implementação do RDD*

COMO FUNCIONA?

- Com o RDD
  - Somente leitura, coleções particionadas de objetos
  - Um array imutável distribuído
  - Criado por meio de transformações paralelas em dados e armazenamento estável
  - Pode ser armazenado em cache para reutilização eficiente

ECOSSISTEMA DO SPARK



ECOSSISTEMA DO SPARK

Spark Core: é o motor de execução da plataforma Spark. Ele fornece recursos de computação distribuídos na memória

Spark SQL: é um mecanismo para o Hadoop Hive que permite que consultas não modificadas do Hadoop Hive sejam executadas até 100 vezes mais rápido em implantações e dados existentes

ECOSSISTEMA DO SPARK

Spark Streaming: é um mecanismo que permite aplicativos interativos e analíticos poderosos no streaming de dados

MLLib: é uma biblioteca de aprendizado de máquina escalável

GraphX: é um mecanismo de computação gráfica construído em cima do Spark

BENEFÍCIOS

- Generaliza o MapReduce
- Modelo de programação mais rico
- Menos sistemas para dominar
- Melhor gerenciamento de memória
- Menor movimentação de dados que leva a um melhor desempenho para análises complexas

BIBLIOGRAFIA

1. http://www.prathapkudupublog.com/2018/02/modules-in-apache-spark.html
2. https://data-flair.training/blogs/apache-spark-ecosystem-components/
3. https://medium.com/expedia-group-tech/an-introduction-to-apache-spark-f0795f2d5201
4. https://www.infoq.com/br/articles/apache-spark-introduction/
5. https://docs.microsoft.com/pt-br/azure/databricks/getting-started/spark/quick-start