# BCC 362 – Sistemas Distribuídos

**BRENDA SOTERO**

**MARCOS RODRIGUES**


APACHE
Spark
TM

# O QUE É O SPARK?

# O QUE É SPARK?

## APACHE SPARK

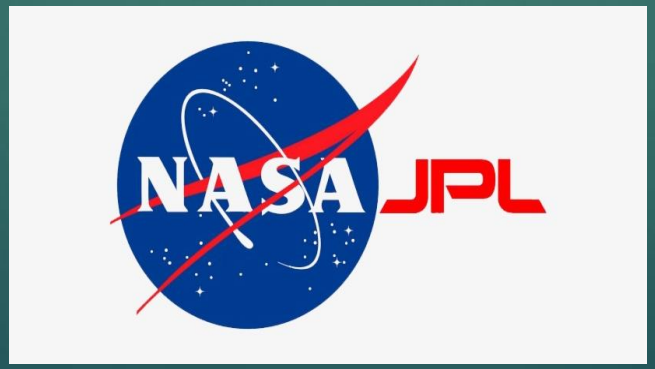
- Spark é um framework para processamento de Big Data construído com foco em velocidade, facilidade de uso e análise sofisticada. Foi desenvolvido pelo AMPLab da Universidade da Califórnia e teve seu código fonte aberto como projeto da Apache Software Foundation.

# O QUE É SPARK?

## EMPRESAS:








# O QUE É SPARK?

## CURIOSIDADES

- A maioria das empresas rodam com milhares de máquinas;
- Trabalha bem na casa dos PetaBytes;
- Já foi usado para ordenar 100TB, três vezes mais rápido que o MapReduce;
- Ganhou o Daytona GraySort Benchmark de 2014 ordenando 1PB.

# O QUE É SPARK?

## CARACTERÍSTICAS

- Spark estende o MapReduce:
  - Armazenamento de dados em memória;
  - Processamento próximo ao tempo real;
- Otimiza o uso de operadores de grafos;
- Avaliação sob demanda de consultas de Big Data contribui com a otimização do fluxo global de processamento de dados;
- Fornece APIs concisas e consistentes em Scala, Java e Python.

# O QUE É SPARK?

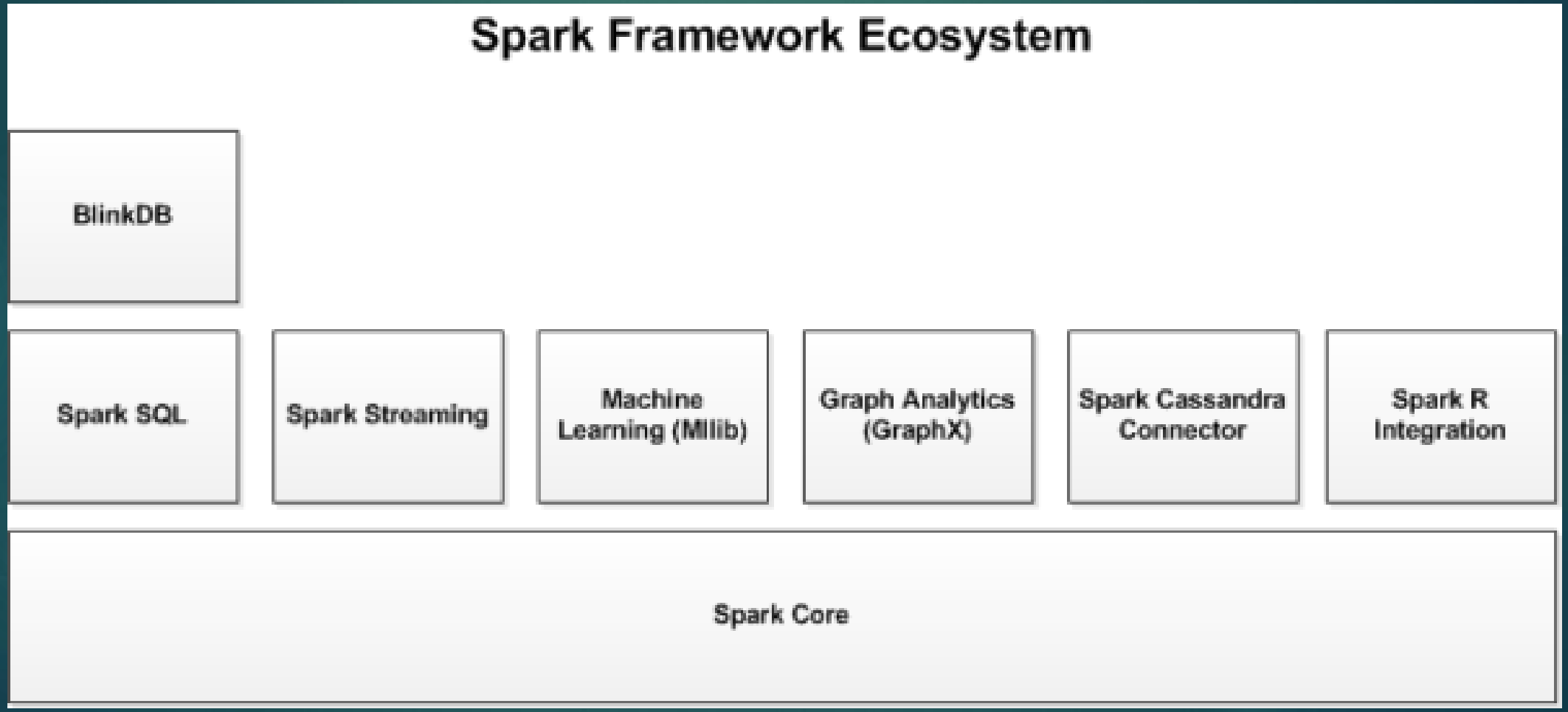
## LINGUAGENS

- Scala;
- Java;
- Python;
- Clojure;
- R.

# COMO ELE FUNCIONA?

# COMO ELE FUNCIONA?

## ECOSSISTEMA DO SPARK



# COMO ELE FUNCIONA?

## COMPONENTES

- Spark Streamming, que possibilita o processamento de fluxos em tempo real;
- O GraphX, que realiza o processamento sobre grafos;
- O SparkSQL para a realização de consultas e processamento sobre dados no Spark;
- MLlib (Machine Learning), é a biblioteca de aprendizado de máquina, com diferentes algoritmos para as mais diversas atividades, como clustering;

# COMO ELE FUNCIONA?

## COMPONENTES

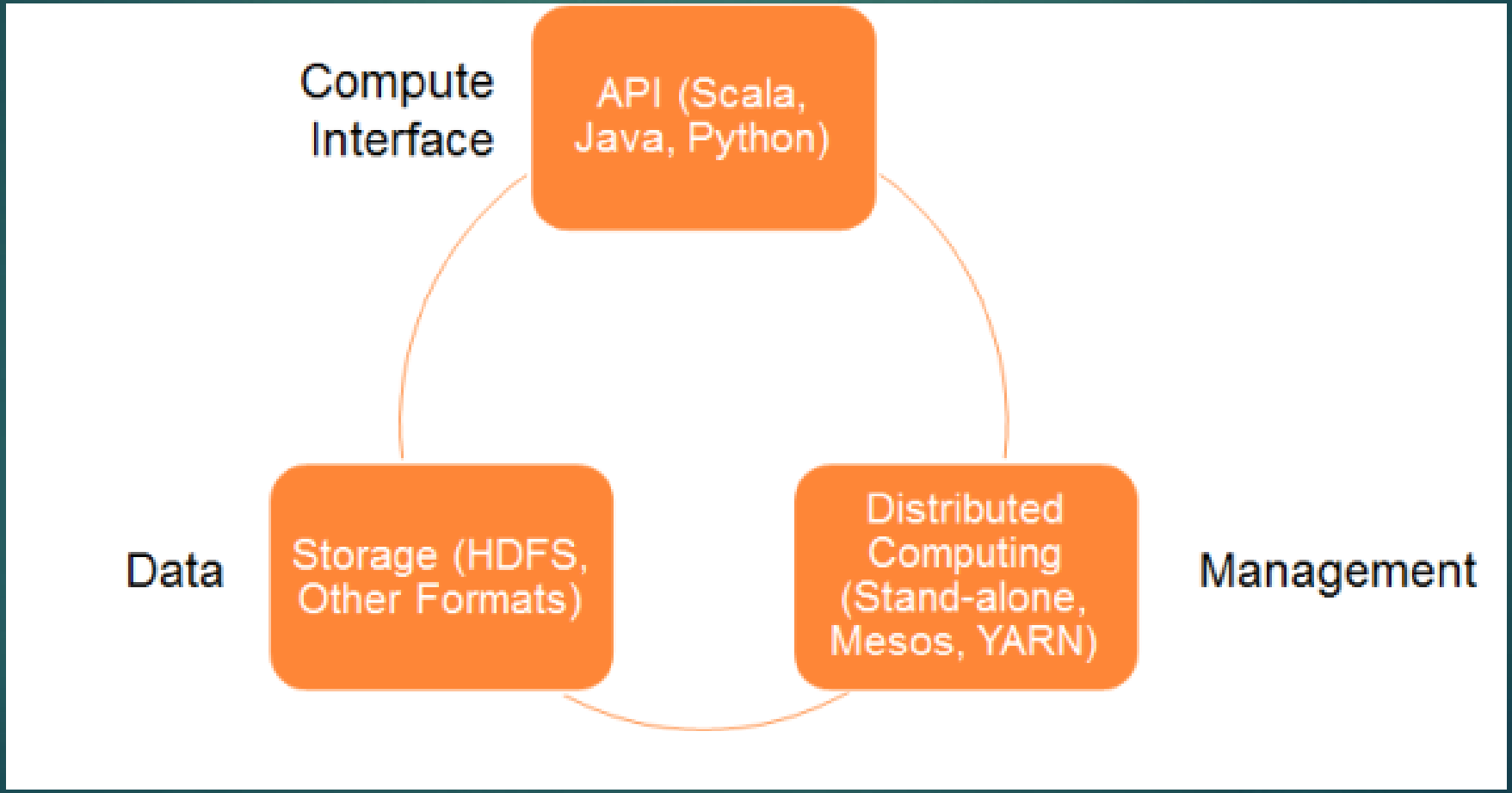
- BlinkDB é uma engine SQL para consultas por amostragem e pode ser usado para execução de consultas interativas em grandes volumes de dados;

- Tachyon é um sistema de arquivos distribuídos em memória que permite o compartilhamento de arquivos através e frameworks de cluster;

- Cassandra Spark Connector e Spark R são adaptadores de integração.

# COMO ELE FUNCIONA?

## ARQUITETURA



# COMO ELE FUNCIONA?

## ARQUITETURA

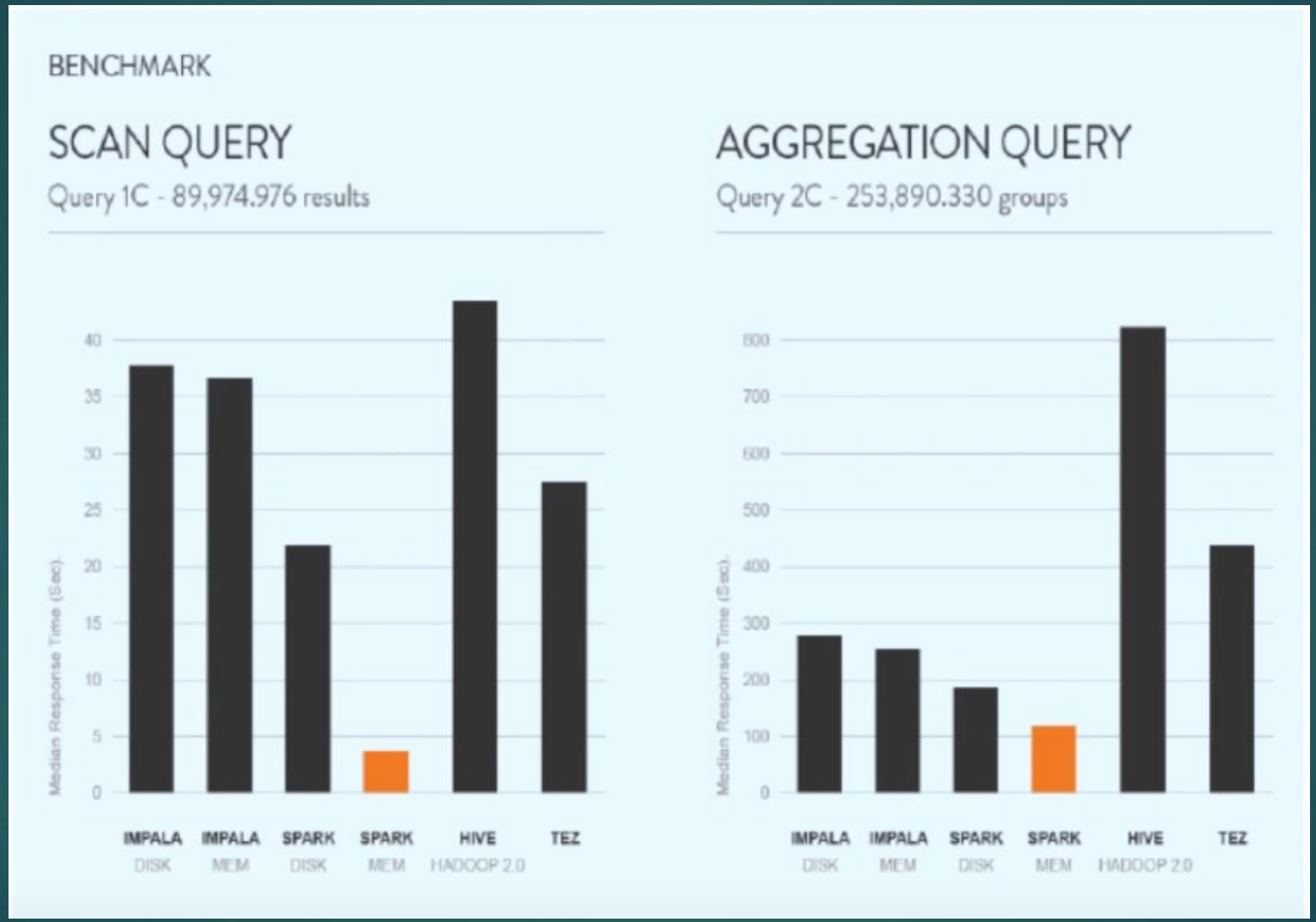
- Armazenamento de dados: usa o arquivo HDFS, compatível com Hadoop, Hbase, Cassandra etc;
- API: desenvolvimento de aplicações;
- Gerenciamento de Recursos: pode ser implantado como máquina local ou uma estrutura de computação distribuída.

# COMO ELE FUNCIONA?

## GERENCIADORES DE CLUSTER

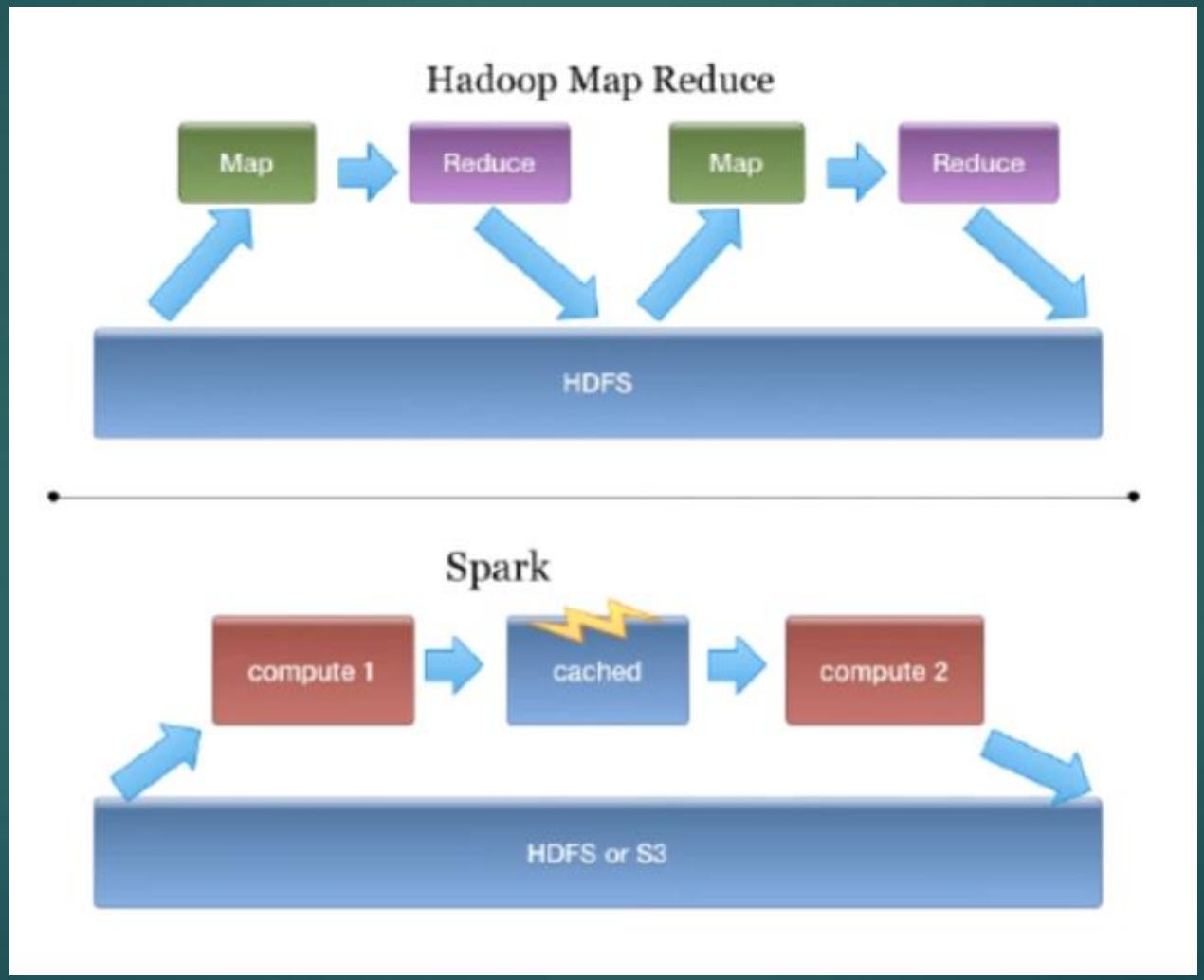
- Standalone Deploy Mode;
- Apache Mesos;
- Hadoop Yarn;
- Kubernetes.

# Eficiente?


BENCHMARK
SCAN QUERY
Query 1C - 89,974.976 results
Median Response Time (Sec)
0
5
10
15
20
25
30
35
40
IMPALA DISK
IMPALA MEM
SPARK DISK
SPARK MEM
HIVE HADOOP 2.0
TEZ
AGGREGATION QUERY
Query 2C - 253,890.330 groups
Median Response Time (Sec)
0
100
200
300
400
500
600
700
800
IMPALA DISK
IMPALA MEM
SPARK DISK
SPARK MEM
HIVE HADOOP 2.0
TEZ

# Eficiente?

- Pode ser até x100 mais rápido que o Map Reduce



# APLICAÇÃO