Matheus R Gomes e Eduardo Soares

Big Data

Grande quantidades de dados

File System

Principais sistemas de arquivos

- **GlusterFS**
- Ceph
- Hadoop
- GPFS
- NFS (filesystem)
- **Apache Spark**

Apache Spark

Apache Spark é uma engine capaz de processar em alta velocidade uma grande quantidade de dados.

É baseada no MapReduce e aprimora seu modelo para usar mais tipos de computação com alta eficiência.

Apache Spark

Características

- Cluster de computação que aumenta a velocidade da aplicação;
- Fácil de usar;
- Escrever aplicações em Java, Scala, Python, e SQL;
- Combinar diferentes bibliotecas de dados (DataFrames, MLlib, GraphX, etc.) na mesma aplicação;
- Suporte ao Hadoop;

Apache Spark

Suporte ao Hadoop

Completamente compatível com Hadoop e sistemas HDFS. Similar ao Hadoop, o Spark usa um modelo de armazenamento distribuído através de um grande cluster de servidores.

Spark pode usar o Hadoop de duas maneiras: **Armazenamento** e **Processamento**.

Apache Spark

Suporte ao Hadoop

- O Spark pode rodar utilizando seu próprio cluster ou utilizando Mesos, Hadoop YARN, ou na nuvem.
- Pode acessar dados do HDFS, Cassandra, HBase, Hive ou outras fontes do Hadoop.

Quem está usando?

Capital One

IBM

citi

amazon

Google

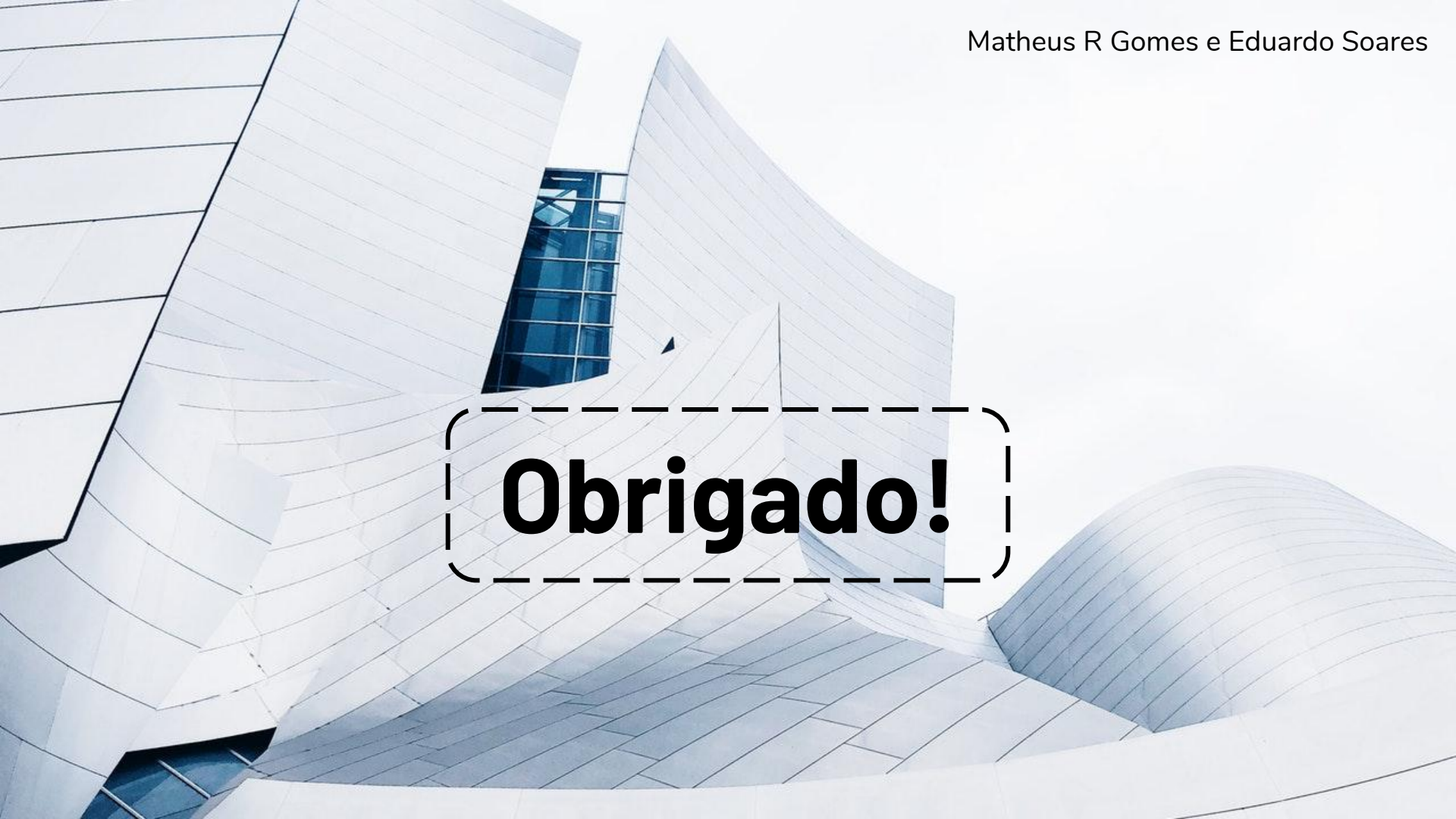
GlusterFS

É um sistema de arquivos para a construcão de clusters de armazenamento, de código fonte aberto.

Agrega múltiplas unidades de armazenamento remotas em um único volume.

As unidades de armazenamento, chamadas bricks, são distribuídas pela rede em um único sistema de arquivos paralelo, permitindo uma escalabilidade de milhares de bricks e vários petabytes de armazenamento.

Matheus R Gomes e Eduardo Soares



Obrigado!