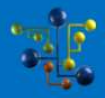
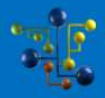


Engenharia de Dados com Hadoop e Spark



Bem-vindo(a)

Instalação e Configuração do Ambiente Hadoop

Ambiente Hadoop

Ambiente Hadoop

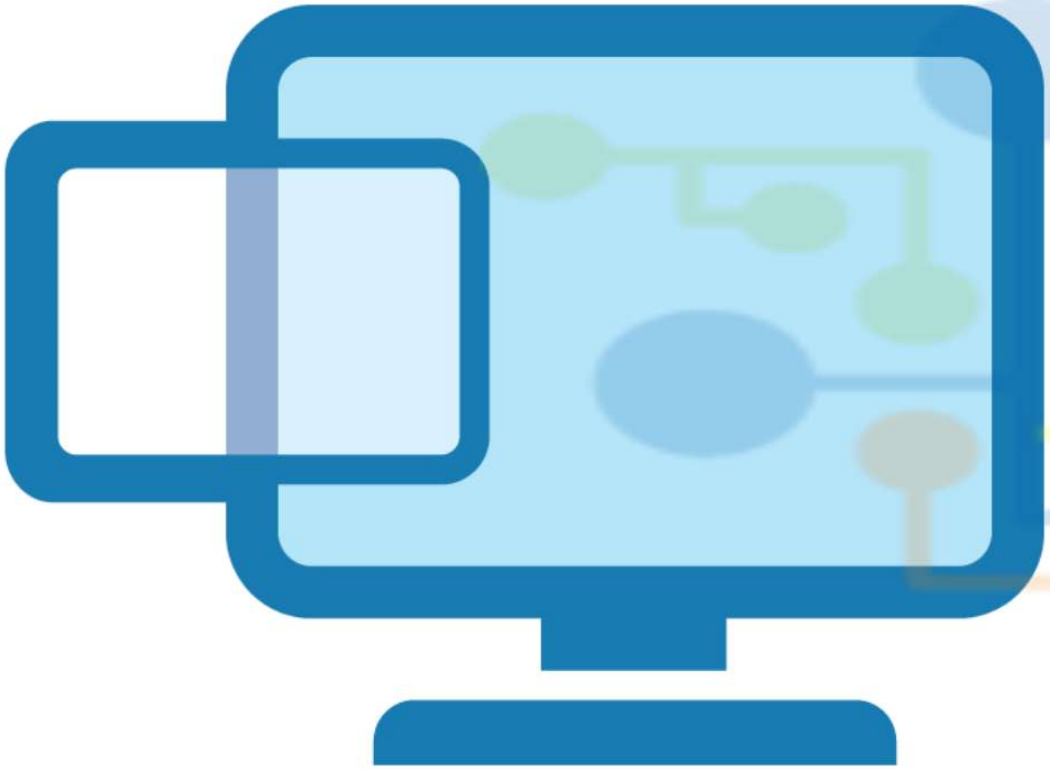
Ambiente Hadoop

cloudera

Hortonworks

- Somente para máquinas 64 bits
- Requerem computadores com no mínimo 8 GB de RAM

Ambiente Hadoop



Virtual Machine

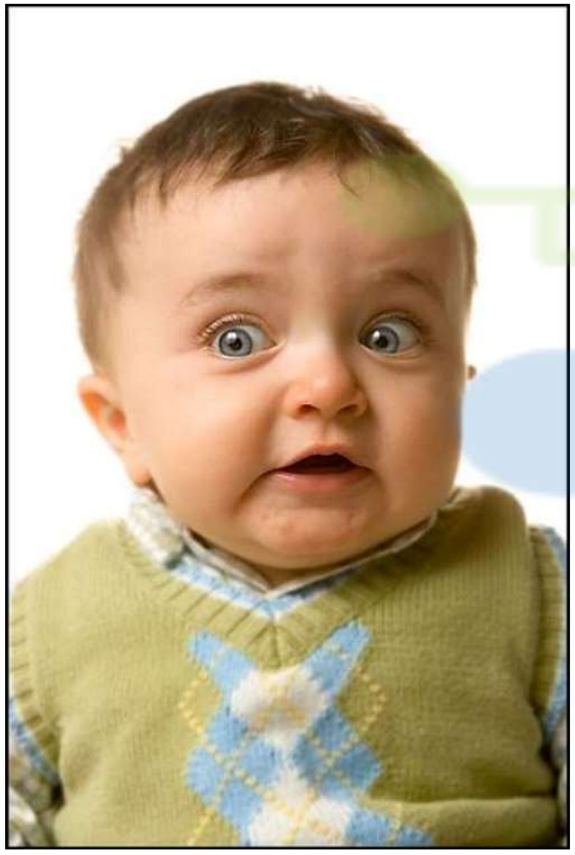
Ambiente Hadoop

E o que faremos neste e no próximo capítulo?

Ambiente Hadoop

Infraestrutura de TI para Big Data

Ambiente Hadoop



Um pequeno detalhe!

Ambiente Hadoop

Atividades

Parte 1

- Criar uma máquina virtual
- Instalar o sistema operacional Linux
- Instalar utilitários (Java, ssh, ferramentas)
- Instalar o MySQL
- Instalar o Hadoop
- Configurar o HDFS e o MapReduce
- Executar um job MapReduce no HDFS

Parte 2

- Instalar e configurar o Zookeeper
- Instalar e configurar o Hbase
- Instalar e configurar o Hive
- Instalar e configurar o Pig
- Instalar e configurar o Sqoop
- Instalar e configurar o Spark
- Instalar e configurar o Flume

Ambiente Hadoop

Ambiente Hadoop

Atividades

- Instalar e configurar a máquina virtual Cloudera
- Instalar e configurar a máquina virtual Hortonworks
- Instalar e configurar o Hadoop com Docker

Ambiente Hadoop

Mas e se eu não quiser instalar o Hadoop?

Pode ser que, por qualquer razão, você não queira atravessar este processo e não queira instalar o Hadoop. Não há problema algum. Você pode fazer o download da máquina virtual pronta que será o resultado de todo o processo que você vai acompanhar neste e no próximo capítulo (~ 10 GB).

Ambiente Hadoop

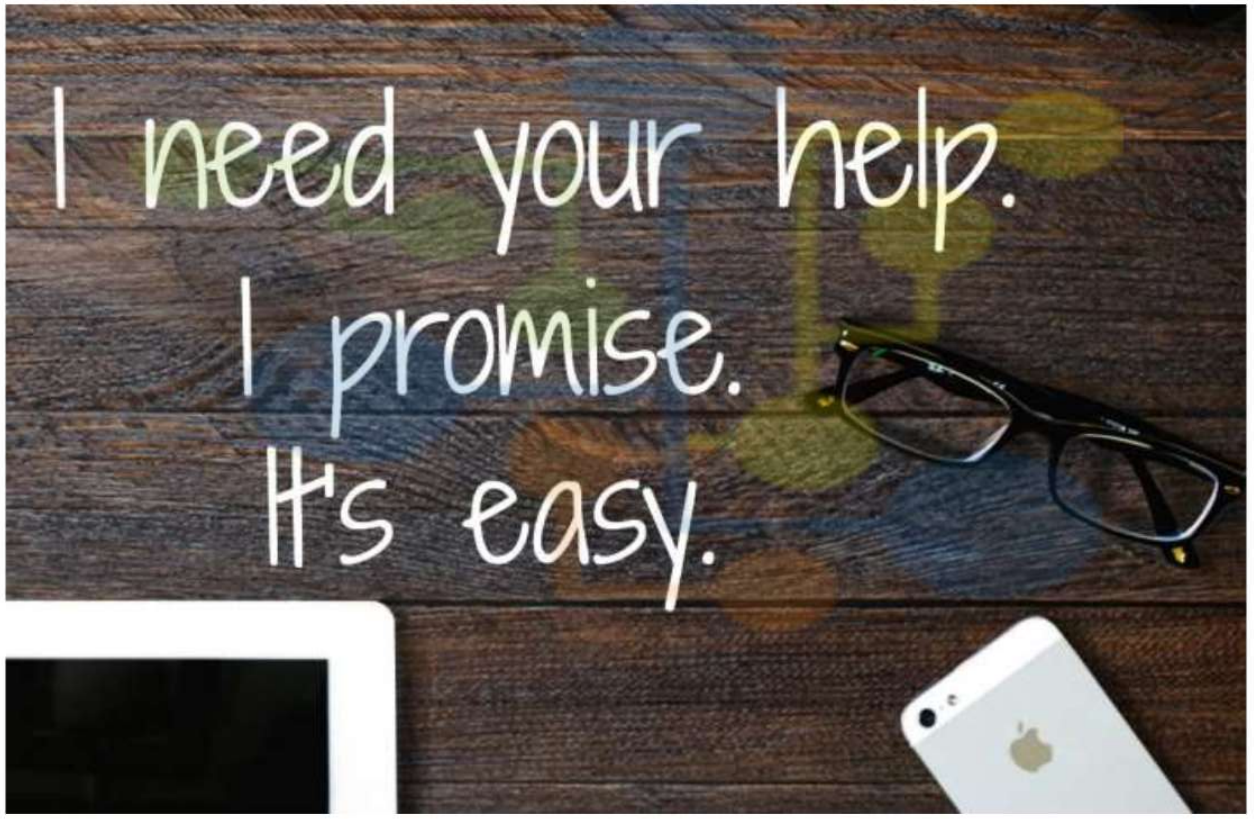
Mas eu recomendo que você atravesse o caminho.
O aprendizado será realmente um diferencial.

Ambiente Hadoop

Ambiente Hadoop

Alunos das Formações DSA tem acesso, gratuito e exclusivo, ao curso:
Introdução ao Sistema Operacional Linux.

Ambiente Hadoop



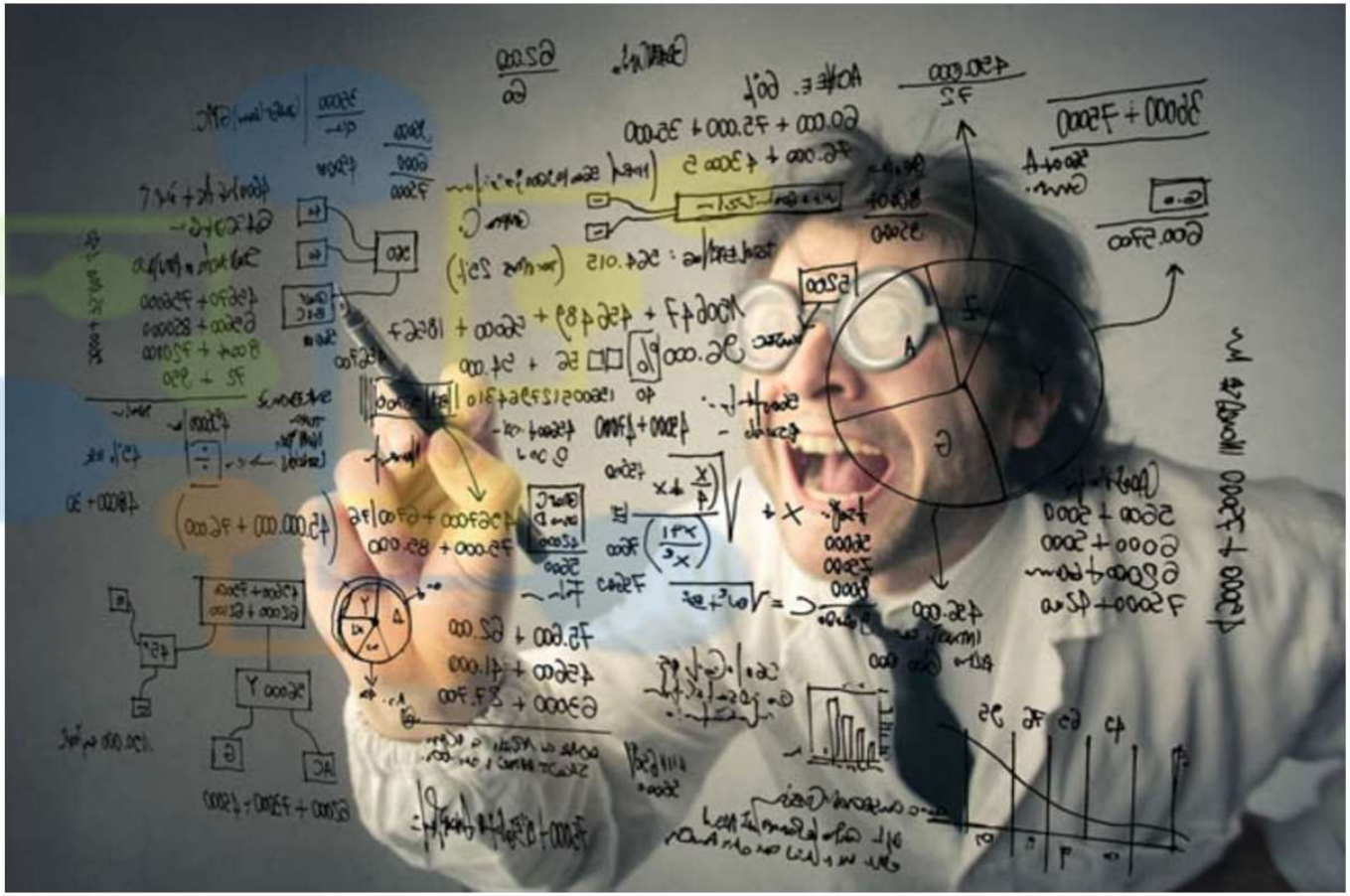
Ambiente Hadoop

Por que não usamos um ambiente na nuvem?

Por que Cientistas de Dados Precisam Conhecer o Hadoop?

Cientistas de Dados e Hadoop

Diferentes pessoas usam diferentes ferramentas para diferentes propósitos.



Cientistas de Dados e Hadoop

1

Hadoop é open source

Cientistas de Dados e Hadoop

2

Hadoop oferece o framework mais completo para armazenamento e processamento de Big Data

Cientistas de Dados e Hadoop

3

A líder mundial em bancos de dados relacionais, a Oracle, oferece soluções de Big Data Analytics com Hadoop

Cientistas de Dados e Hadoop

4

A líder mundial em sistemas operacionais, a Microsoft, oferece soluções corporativas em nuvem, com Hadoop

Cientistas de Dados e Hadoop

5

O Hadoop é mantido pela Apache Foundation, mas recebe contribuição de empresas como Google, Yahoo e Facebook

Cientistas de Dados e Hadoop

6

Um Cientista de Dados deve conhecer bem o paradigma de processamento MapReduce

Cientistas de Dados e Hadoop

7

Hadoop normalmente aparece como um dos skills mais procurados em um Cientista de Dados

Cientistas de Dados e Hadoop

8

Por se tratar de uma tecnologia avançada, faltam profissionais de Hadoop no mercado

Cientistas de Dados e Hadoop

9

Hadoop é usado por algumas das maiores empresas do mundo

Cientistas de Dados e Hadoop

10

O Big Data ainda está na sua infância. Onde vamos armazenar todos esses dados?

Cientistas de Dados e Hadoop

Convencido sobre a importância de aprender Hadoop?

Cientistas de Dados e Hadoop

Não?

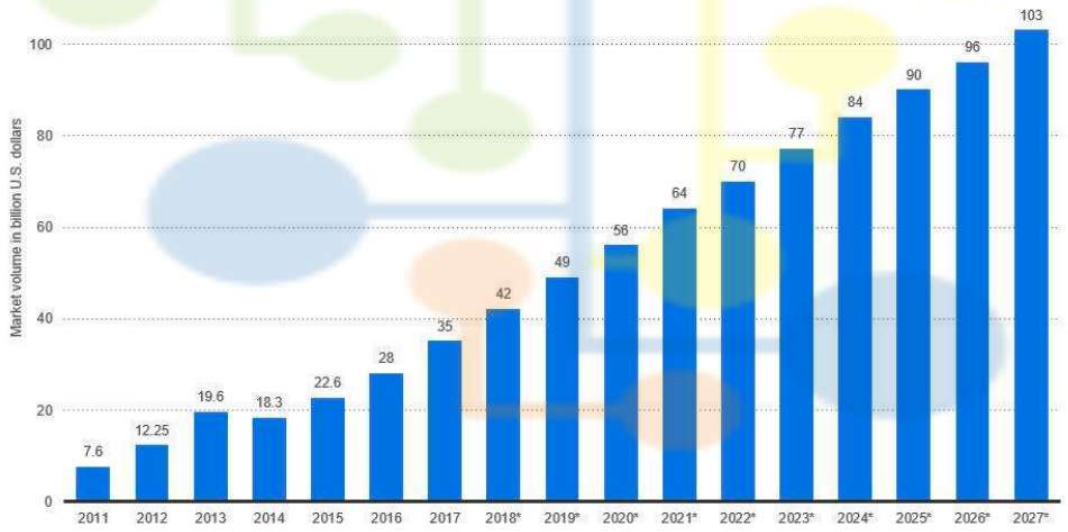
Cientistas de Dados e Hadoop

Ainda tenho mais argumentos!

Cientistas de Dados e Hadoop

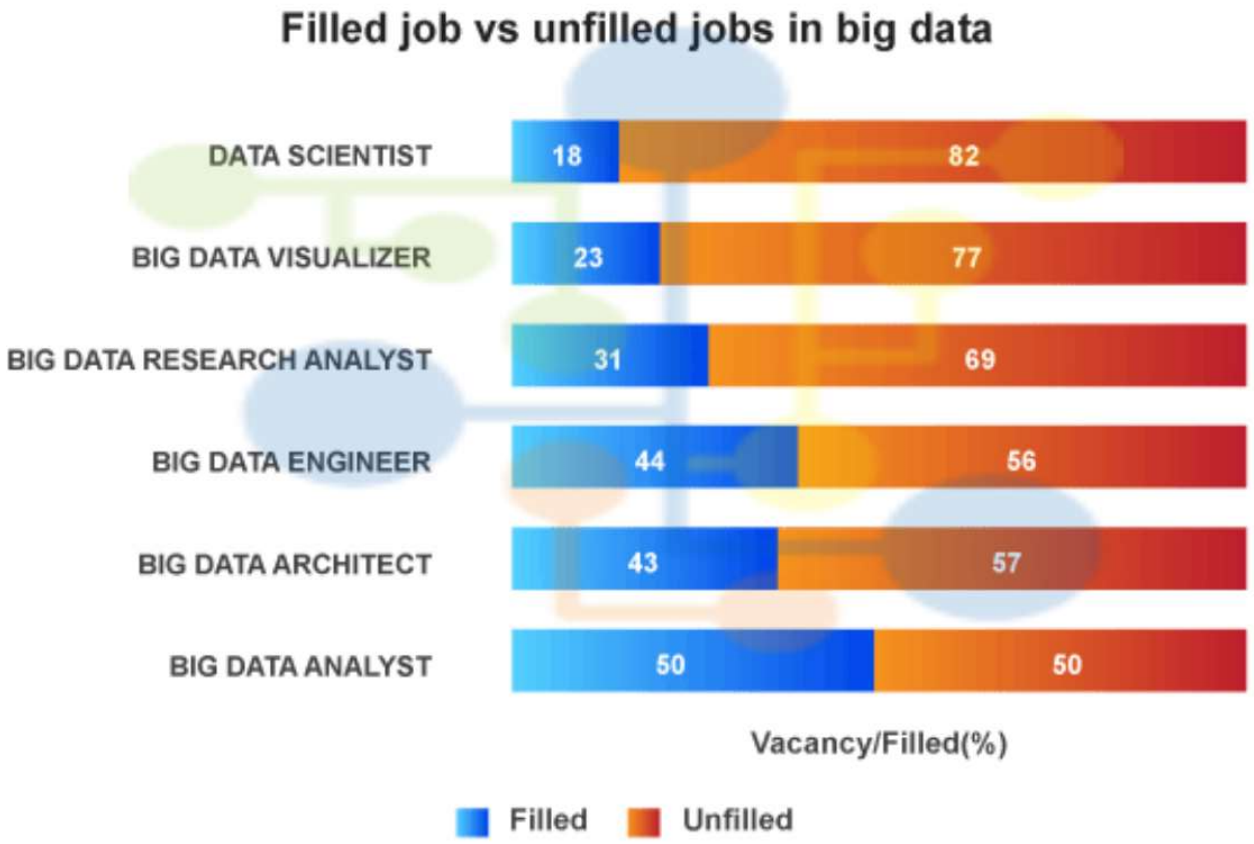
Forecast Revenue Big Data Market Worldwide 2011-2027

Big Data Market Size Revenue Forecast Worldwide From 2011 To 2027 (in billion U.S. dollars)



statista

Cientistas de Dados e Hadoop



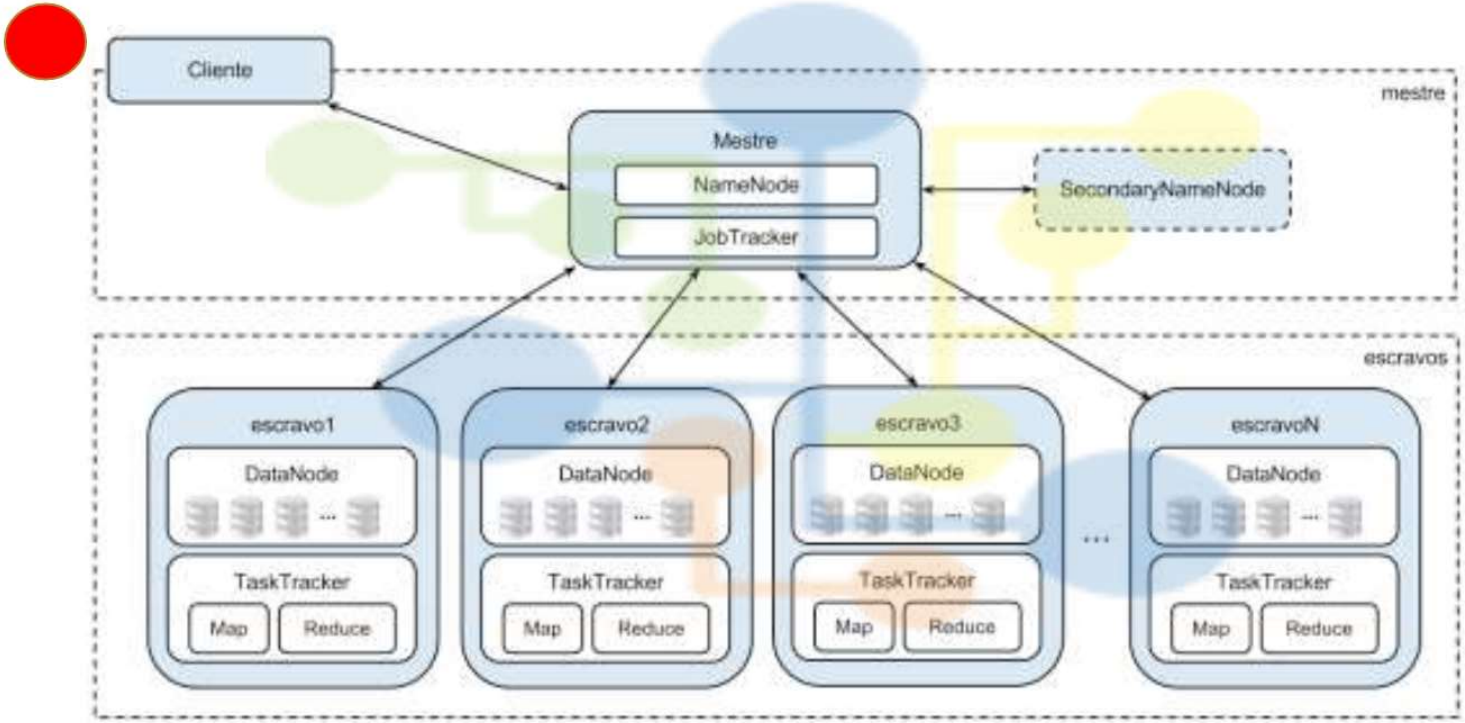
Cientistas de Dados e Hadoop

Aprender ou não Hadoop é uma escolha sua.

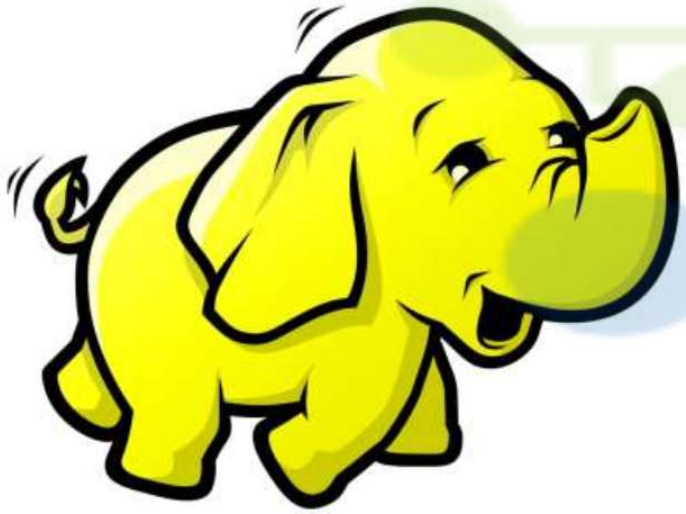
Mas com certeza este conhecimento será um grande diferencial na sua carreira e na sua compreensão sobre como armazenar e analisar Big Data.

Modos de Execução do Hadoop

Modos de Execução do Hadoop



Modos de Execução do Hadoop



- Modo Local (Standalone)
- Modo Pseudo-Distribuído (Pseudo-Distributed)
- Modo Totalmente Distribuído (Fully Distributed)

Modos de Execução do Hadoop

- core-site.xml
- hdfs-site.xml
- mapred-site.xml

Modo Local
(Standalone)

Modo Pseudo-Distribuído
(Pseudo-Distributed)

Modo Totalmente Distribuído
(Fully Distributed)

Modos de Execução do Hadoop

Modo Local
(Standalone)

Modos de Execução do Hadoop

Modo Pseudo-Distribuído
(Pseudo-Distributed)

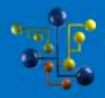
Modos de Execução do Hadoop

Modo Totalmente Distribuído
(Fully Distributed)

Modos de Execução do Hadoop

- core-site.xml
- hdfs-site.xml
- mapred-site.xml

Component	Property	Standalone	Pseudo-distributed	Fully distributed
Core	fs.default.name	file:/// (default)	hdfs://localhost/	hdfs://*namenode*/
HDFS	dfs.replication	N/A	1	3 (default)
MapReduce	mapred.job.tracker	local (default)	localhost:8021	*jobtracker*:8021



Obrigado