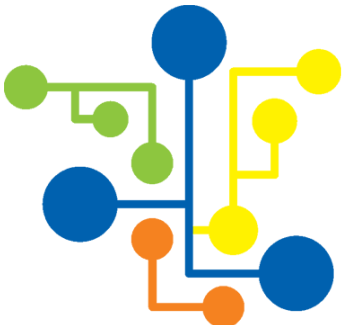


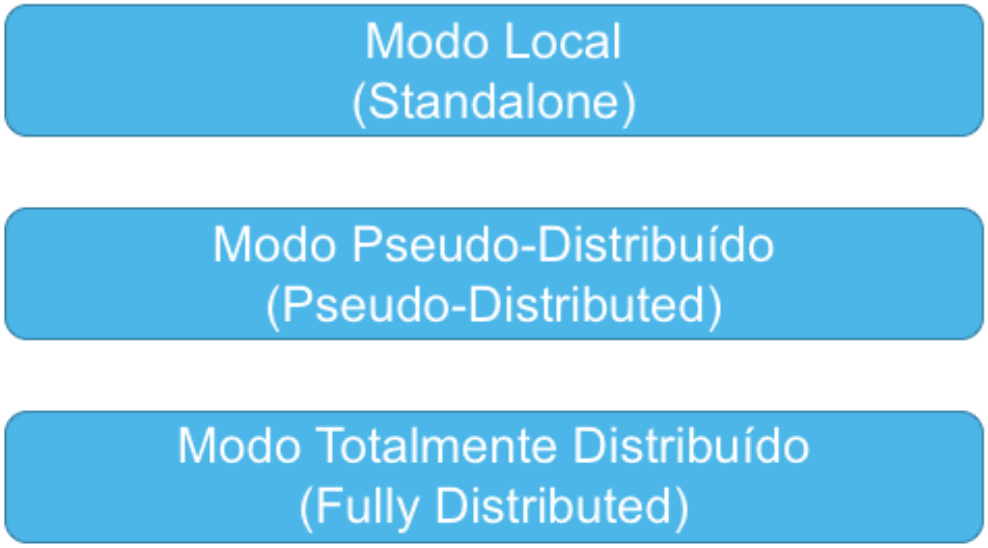
Data Science Academy

www.datascienceacademy.com.br

Data Lake – Design, Projeto e Integração

Modos de Execução do Hadoop

Embora uma aplicação Hadoop seja tipicamente executada em um conjunto de máquinas, ela pode também ser executada em um único node. Essa possibilidade permite adotar configurações simplificadas para as fases iniciais de implementação e testes, visto que depurar aplicações distribuídas não é algo trivial. Posteriormente, outras configurações mais sofisticadas podem ser utilizadas para usufruir de todas as vantagens oferecidas pelo framework Hadoop. São três modos possíveis de execução do Hadoop:



Para alternar entre essas configurações é necessária a edição de três arquivos: core-site.xml, hdfs-site.xml e mapred-site.xml. Veremos isso mais a frente!

Modo Local (Standalone)

O Hadoop é por padrão configurado para ser executado no modo local. Dessa maneira, se essa for a sua opção escolhida, os parâmetros nos arquivos de configuração não precisam de alterações. Esse modo é o mais recomendado para a fase de desenvolvimento, onde normalmente ocorre a maior incidência de erros, sendo necessária a realização de vários testes de execução. Nessa configuração, todo o processamento da aplicação é executado apenas na máquina local. Dessa forma simplificada, fica mais fácil para o usuário realizar a depuração de seu código, aumentando sua produtividade.

Modo Pseudo-distribuído

Uma segunda alternativa para executar uma aplicação Hadoop é o modo Pseudo-distribuído. Nesse modo são aplicadas todas as configurações, semelhantes às necessárias para execução em um cluster, entretanto, toda a aplicação é processada em modo local, por isso o termo Pseudo-distribuído ou também chamado “cluster” de uma máquina só. Embora não seja executado realmente em paralelo, esse modo permite a sua simulação, pois utiliza todos os processos de uma execução paralela efetiva: NameNode, DataNode, JobTracker, TaskTracker e SecondaryNameNode.

Modo Totalmente Distribuído

Por fim, o terceiro e último modo de execução é utilizado para o processamento distribuído da aplicação Hadoop em um cluster de computadores real. Nessa opção, como no modo pseudo-distribuído, também é necessário editar os três arquivos de configuração, definindo parâmetros específicos e a localização do SecondaryNameNode e dos nós escravos. Todavia, como nesse modo temos diversos computadores, devemos indicar quais máquinas irão efetivamente executar cada componente. Esse é o modo com o qual vamos construir nosso Data Lake na sequência deste capítulo.