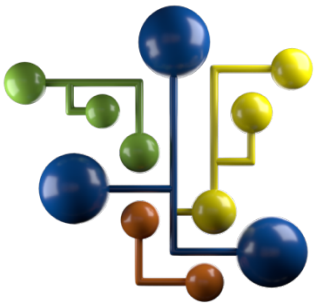


Data Science Academy

www.datascienceacademy.com.br

Engenharia de Dados com Hadoop e Spark

Melhores Práticas de Monitoramento do Cluster Hadoop

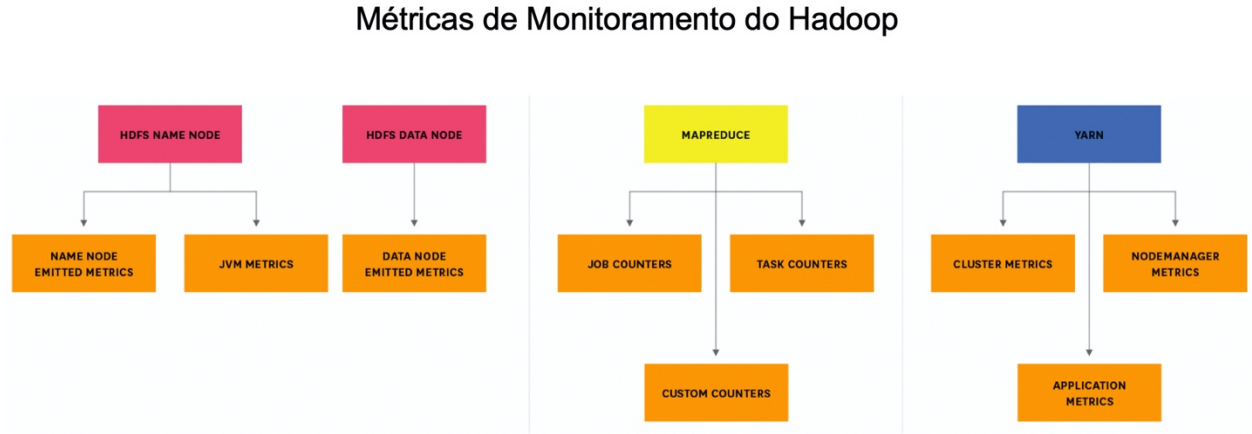
De uma perspectiva de operações, clusters Hadoop são incrivelmente resilientes em relação a falhas no sistema. O Hadoop foi projetado para ser tolerante a falhas e faz isso muito bem.

A falha de um DataNode pode não representar um problema, enquanto a falha no NameNode é algo crítico. Em outras palavras, indicadores de um único DataNode são menos importantes que os indicadores gerais de serviço de todo o cluster.

Quando configurado corretamente, um cluster Hadoop pode armazenar e processar quantidades massivas de dados, na casa de petabytes de dados. O monitoramento do cluster Hadoop normalmente é feito através do monitoramento de sub-componentes chave em 4 grandes categorias.

Métricas de Monitoramento do Hadoop

Cada um desses grupos se subdivide em grupos menores com diversas métricas que podem ser coletadas e monitoradas.



Os DataNodes se comunicam com o NameNode a cada 3 segundos, o que é configurável em um dos parâmetros nos arquivos de configuração. Essa comunicação, além de levar o estado do datanode, envia outras estatísticas que podem ser coletadas e usadas para monitorar o estado geral do cluster e do HDFS.

Diversas métricas podem ser coletadas e monitoradas (nesta tabela abaixo você encontra alguns exemplos de algumas das métricas que podem ser coletadas). Quando usamos o Apache Ambari, essas métricas são coletadas e apresentadas em um dashboards. Na seção de links úteis vc encontra a lista completa de todas as métricas de monitoramento do Hadoop.

Métrica a ser coletada
CapacityRemaining
CorruptBlocks / MissingBlocks
VolumeFailuresTotal
NumLiveDataNodes / NumDeadDataNodes
FilesTotal
TotalLoad
BlockCapacity / BlocksTotal
UnderReplicatedBlocks
NumStaleDataNodes