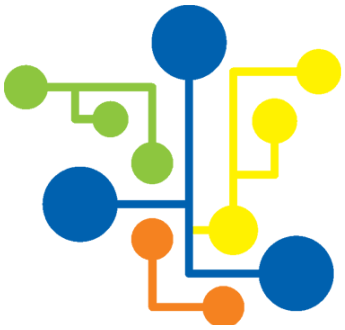


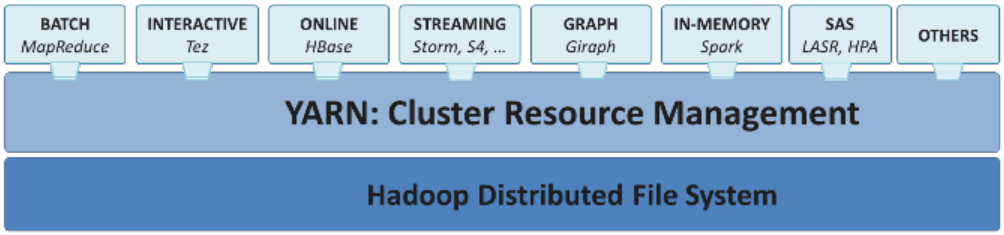
Data Science Academy

www.datascienceacademy.com.br

Data Lake – Design, Projeto e Integração

Gestão de Recursos do Cluster com YARN

O Apache YARN - " Yet Another Resource Negotiator" é a camada de gerenciamento de recursos do Hadoop. O YARN foi introduzido no Hadoop 2.x. e agora melhorado no Hadoop 3.x. O YARN permite diferentes mecanismos de processamento de dados, como processamento de gráficos, processamento interativo, processamento de fluxo e processamento em lote para executar e processar dados armazenados no HDFS (Hadoop Distributed File System). Além do gerenciamento de recursos, o YARN também é usado para agendamento de trabalhos. O YARN amplia o poder do Hadoop para outras tecnologias em evolução, para que eles possam aproveitar as vantagens do HDFS (sistema de armazenamento mais confiável e popular do planeta) e do cluster econômico.



O Apache YARN também é considerado como o sistema operacional de dados do Hadoop 2.x. A arquitetura baseada em YARN do Hadoop fornece uma plataforma de processamento de dados de uso geral que não se limita apenas ao MapReduce. Ele permite que o Hadoop processe outro sistema de processamento de dados para fins específicos, diferente do MapReduce. Ele permite executar várias estruturas diferentes no mesmo hardware em que o Hadoop é implementado.

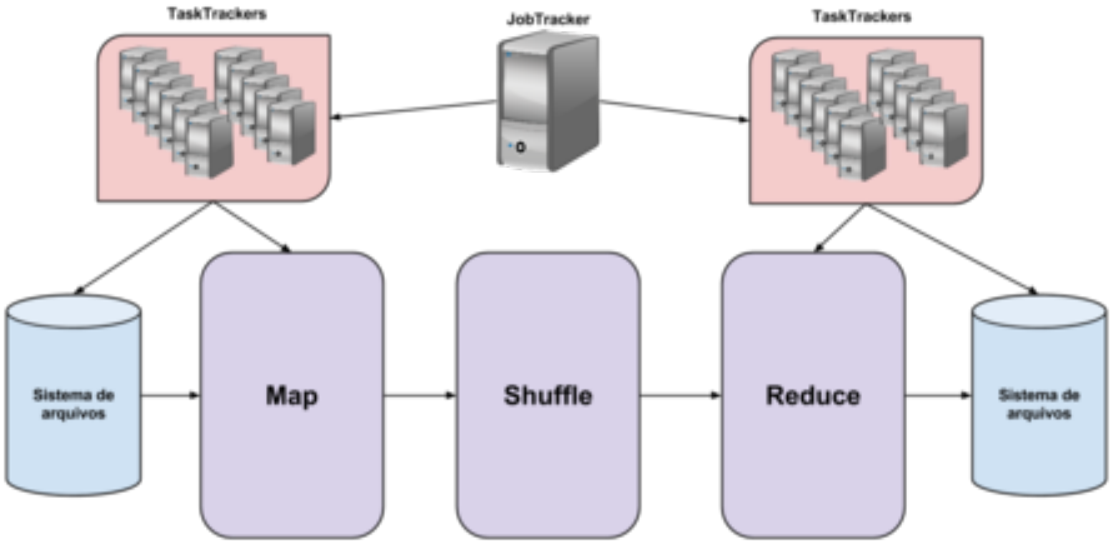
Como Surgiu o YARN?

Junto com o HDFS, o MapReduce tem sido a base de computação em larga escala na última década, para aplicações analíticas de Big Data. Com o tempo o Hadoop amadureceu se tornando um ambiente computacional estável e genérico o suficiente para o processamento de praticamente qualquer aplicação. No entanto, à medida que o Hadoop passou a ser adotado de forma mais ampla,

outras especializações se tornaram necessárias e ficou evidente que o MapReduce não era muito adequado para processamento iterativo, típico de aplicações de Machine Learning.

O MapReduce é paralelizável, fácil de entender, e compreende basicamente 2 processos: mapeamento e redução e oculta os detalhes da computação distribuída, permitindo assim a construção de aplicações relativamente fáceis para processamento em larga escala. No entanto, para conseguir coordenação e tolerância a falhas, o MapReduce utiliza um modelo de execução de extração de dados que exige escritas intermediárias de volta no HDFS.

Essa escrita frequente em disco, representa custo de tempo em qualquer sistema de computação; como resultado, embora seja extremamente seguro e resiliente, o MR também é mais lento em cada tarefa de Mapeamento e Redução.

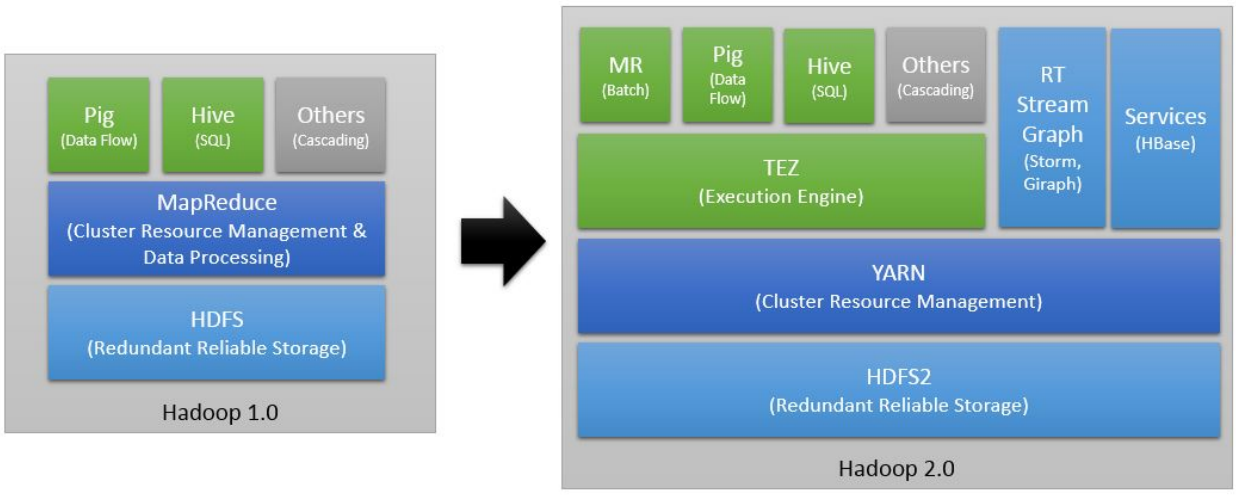


Pior ainda, quase todas as aplicações precisam encadear vários jobs de MR em muitos passos, criando um fluxo de dados em direção ao resultado final desejado. Isso resulta em quantidades enormes de dados intermediários escritos no HDFS, que não são necessários ao usuário, gerando custos adicionais no que diz respeito ao uso de disco.

Para tratar esses problemas, o Hadoop passou a usar um framework mais genérico de gerenciamento de recursos e processamento, o YARN.

Anteriormente a aplicação MR alocava os recursos (processadores, memória) aos jobs especificamente. O YARN oferece um acesso mais genérico aos recursos para aplicações Hadoop. O resultado é que ferramentas especializadas não precisam ser decompostas em uma série de jobs de MR e podem ser mais complexas. Ao generalizar o gerenciamento do cluster, o modelo de programação incialmente pensado para o MR pôde ser expandido para incluir novas abstrações e operações.

O Yarn é uma evolução do MapReduce onde as funções do JobTracker são repartidas em deamons independentes. Uma das funções principais do MapReduce é a de partilhar os dados para as funções de Map e Reduce, a outra função é gerenciar as falhas e procurar nós disponíveis para executar a função onde houve falha. Para isso o Yarn muda um pouco a nomenclatura do nó master e o apelida de Resource Manager (RM) ou Application Master (AM), onde cada função MapReduce é uma aplicação definida pelo nó mestre e o resource manager fica responsável por reordenar os nós no caso de falhas dos nós escravos. Com isso, o YARN agora faz a gestão do MapReduce, através de 2 processos: o resource manager e o application master.



Na versão 3 do Hadoop, o YARN ganhou melhorias e correções, consolidando assim sua função de gestão de recursos do cluster Hadoop.