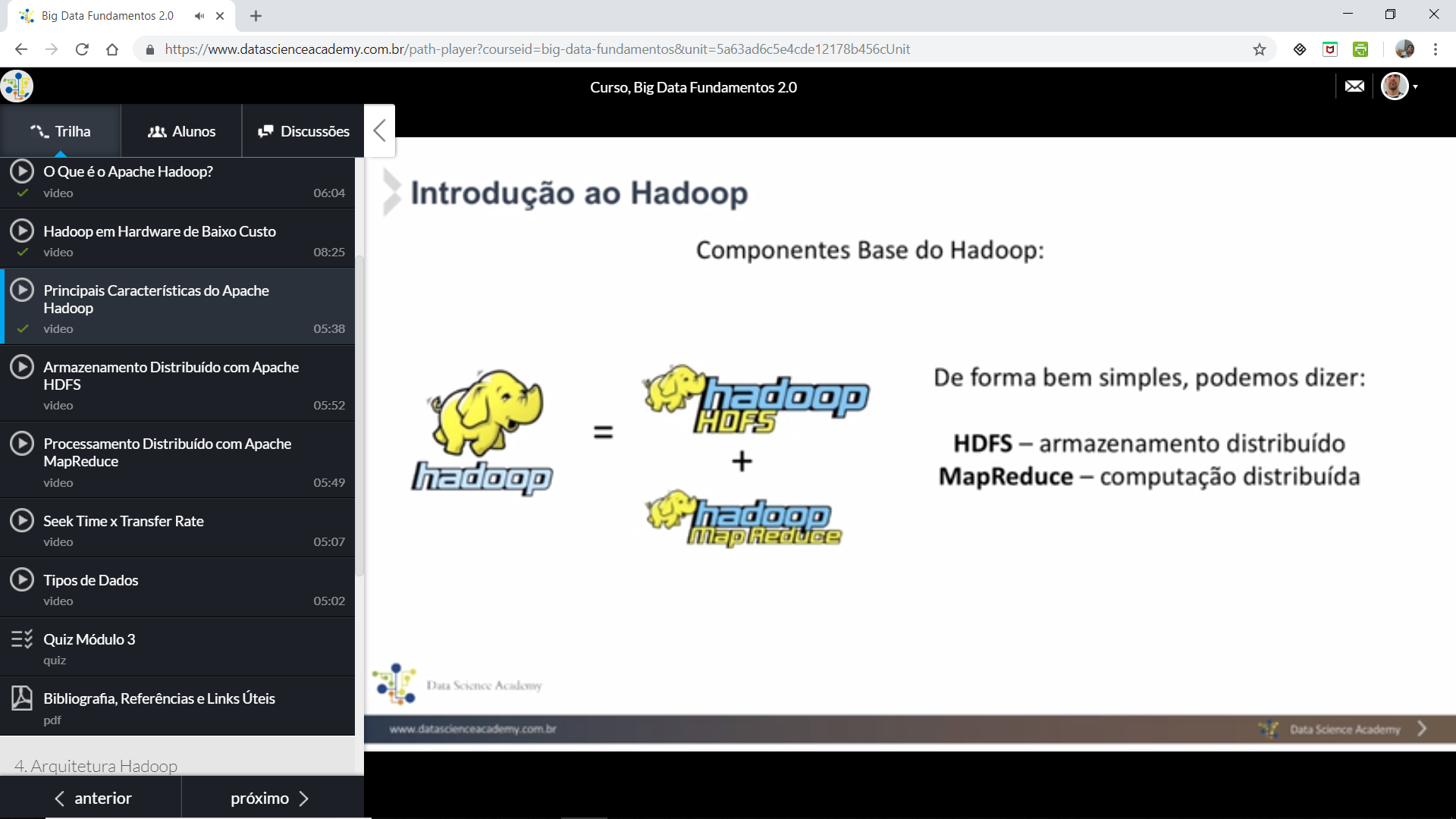
Capítulo 3 – Hadoop em hardware de baixo custo

HDFS: gerenciador de armazenamento em cluster (gravação e leitura de dados)

Haddop Yarn: auxiliar na gestão do cluster

MapReduce: processamento de dados em ambiente distribuído.



HDFS - Sistema de arquivos distribuído que armazena dados em máquinas dentro do cluster, sob demanda, permitindo uma largura de banda muito grande em todo o cluster. O HDFS (Hadoop Distributed File System) é um sistema de arquivos distribuído, projetado para armazenar arquivos muito grandes, com padrão de acesso aos dados streaming, utilizando clusters de servidores facilmente encontrados no mercado e de baixo ou médio custo.

O HDFS possui o conceito de blocos, tal como no Unix, mas seus blocos normalmente têm tamanho de 64MB. Um arquivo muito grande pode ter blocos armazenados em mais de um servidor. Com este conceito de blocos de tamanho fixo fica mais fácil calcular as necessidades de armazenamento.

O HDFS tem 2 tipos de Nós: Master (ou Namenode) e Worker (ou Datanode). O Master armazena informações da distribuição de arquivos e metadados. Já o Worker armazena os dados propriamente ditos.

O YARN é o pré-requisito para o Enterprise Hadoop, fornecendo gerenciamento de recursos e uma plataforma central para fornecer ferramentas consistentes de operações, segurança e controle de dados nos clusters do Hadoop.  
  
O YARN também amplia o poder do Hadoop para as tecnologias novas e incumbentes encontradas no data center, de modo que elas possam aproveitar o armazenamento e o processamento de escala linear e custo-benefício.  
  
YARN - Trata-se de uma plataforma de gerenciamento de recursos responsável pelo gerenciamento dos recursos computacionais em cluster, assim como pelo agendamento dos recursos.  
  
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.  
  
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.