如果你对 Hadoop 和 Hive 不熟悉且想了解更多，推荐你浏览一下相关项目的官方网站、视频和网络上的其他资源，以及一些优秀的图书。比如 Edward Capriolo 等人的 *Programming Hive*（O'Reilly 出版）就是一本很好的指南。

下面我们先讨论某些 Hadoop 和 Hive 的概念，以便为 Presto 的使用提供足够的背景知识。

Hadoop 的核心是由 HDFS 和应用软件（如 Hadoop MapReduce）组成，这些应用程序与 HDFS 中的数据进行交互。Apache YARN 用于管理 Hadoop 应用程序所需的资源。Hadoop 是一种领先的多机集群系统，它用于大规模数据集的分布式处理。它能够对系统进行伸缩，同时在计算机集群之上提供高可用服务。

最初，数据处理是通过编写 MapReduce 程序来完成的。开发者遵循一种特定的编程模型，以让数据处理能够自然地分布在集群中。这种模型运行得很好，也很稳健。然而，编写 MapReduce 程序来分析问题很麻烦。对于依赖 SQL 和数据仓库的现有基础设施、工具和用户来说，很难迁移到 MapReduce 上。

Hive 提供了 MapReduce 之外的另一种使用方式。它最初是用来在 Hadoop 之上提供一个 SQL 抽象层，从而使用类似于 SQL 的语法与 HDFS 中的数据进行交互。这样，大量了解 SQL 的用户便可以与存储在 HDFS 中的数据进行交互。

Hive 数据以文件的形式存储在 HDFS 中，通常叫作**对象**。这些文件使用各种格式，如 ORC、Parquet 等。这些文件以 Hive 所设定的特定目录和文件布局来存储，如分区表和分桶表。我们把这种布局称为 Hive 风格的表格式。

Hive **的元数据**描述了存储在 HDFS 中的数据如何映射到 schema、表和列中，并通过 SQL 进行查询。这些元数据信息保存在 MySQL 或 PostgreSQL 等数据库中，可以通过 Hive Metastore **服务**（HMS）进行访问。

Hive 运行时提供了类似 SQL 的查询语言和分布式执行层来执行查询。Hive 运行时将查询翻译成一组可以在 Hadoop 集群上运行的 MapReduce 程序。随着时间的推移，Hive 的演进使得查询可以翻译到其他的执行引擎，如 Apache Tez 和 Spark 等。

Hadoop 和 Hive 在业界得到了广泛的应用。随着它们的使用，HDFS 格式已经成为许多其他分布式存储系统所支持的格式，如 Amazon S3 和 S3 兼容的存储、Azure Data Lake Storage、Azure Blob Storage 以及 Google Cloud Storage 等。

6.4.2 Hive连接器

Presto 的 Hive 连接器允许你连接到 HDFS 对象存储集群。它利用 HMS 中的元数据，查询和处理存储在 HDFS 中的数据。