大数据处理平台 Spark 研究

温向慧　西北师范大学计算机科学与工程学院

摘要：随着大数据时代的到来，传统的单机模式已经不能满足大规模数据分析处理的需求。Spark 是专门针对海量数据设计的通用并行计算引擎。Spark 启用了弹性分布式数据集 RDD，能够在内存中进行多次迭代计算，其高端的设计理念，为大型应用程序的构建奠定了基础。

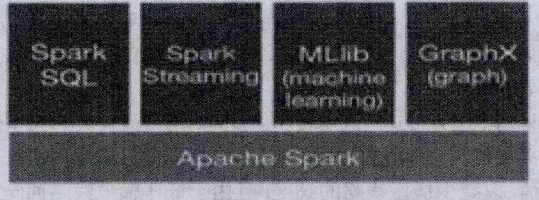
关键字：海量数据　Spark　RDD　迭代计算

1 引言

Apache Spark 是由加州伯克利分校 AMP 实验室开发的，用 scala 语言实现的一种通用计算框架，具有运行速度快、使用方便、适应性好、易于部署等特点。Spark 实现了一个集群的分布式内存抽象 (RDD)，RDD (Resilient Distributed Dataset) 是一个只读的记录分区的集合，运行于内存中。Spark 使用有向无环图 (DAG) 设计，与 Hadoop 相比，其操作简单，使用简洁的代码就能处理大规模数据问题。它可以访问不同的数据源，包括 HDFS,Cassandra,HBase 和 S3,Spark 可以使用其独立集群模式，也可以运行在 EC2，Hadoop YARN 或 Apache Mesos 上。

2 Spark 生态系统

Spark 生态系统如下图所示，包含多个组件：Spark SQL、Spark Streaming、MLlib Graph X 等。SparkSQL 用于查询 Spark 程序中的结构化数据,Spark Streaming 用于实时流处理，MLlib 用于机器学习中，Graph X 用于图计算，它们能够使用 RDD 无缝的集成，形成一站式的处理平台，使应用程序的开发变得简单。



Spark 生态系统

2.1 SparkSQL 技术

Spark SQL 是 Spark 框架的一部分，用于查询和分析结构化的海量数据。它提供了一个分布式的 SQL 查询引擎 DataFrames，是一种分布式数据集合，由“命名列”组织而成，相当于关系型数据库中的数据表。DataFrames 和 SQL 提供了访问各种数据源的常用方法，这些数据源包括 Hive，Avro，Parquet，ORC，JSON 和 JDBC。另外 SQL 接口还可以与不同数据源的数据交互。Spark SQL 在使用时先将外部数据源转化为 DataFrames，再进行查询和转换，最后将处理结果存储或展示，实用性较好。

2.2 SparkStreaming 技术

SparkStreaming 是一个高吞吐量、高容错的实时流处理系统。它不是直接的流式处理，而是将数据流切分成短小的批处理作业，例如以 1 秒为时间片切分，每个时间片数据都是一个 RDD，可以使用 RDD 的转换、行动操作来处理每个时间片数据。每个 RDD 都会产生一个 Job 处理，最后的结果也是返回多个时间片数据。SparkStreaming 支持从多种数据源获取数据，包括 Kafka、Kinesis、Twitter、TCP sockets、Flume 以及 ZeroMQ，从数据源获取数据之后，可以用 Map、Reduce、Join 和 Filter 等高级操作处理大规模复杂数据，最后将处理结果存储或展示。由于 Spark 是短小的批处理方式，所以对一些实时性要求较高的应用来说不适合比较适合实时处理与历史处理相结合的应用场景。

2.3 MLlib 技术

MLlib 是 Apache Spark 可扩展的机器学习库，其中包含许多常用的机器学习算法、实用程序和工具类，机器学习算法有分类、聚类、回归、推荐、决策树、主题建模等，实用程序包括特征转换、模型评估等，还有一些其他工具如：分布线性代数、统计。因为 Spark 的优势是迭代计算，所以对于一些多次迭代的机器学习算法，SparkMLlib 的效果远远优于 MapReduce。同时，MLlib 的出现让机器学习的门槛降低，使一些对 ML 算法不了解的用户也能方便的处理数据。

2.4 GraphX 技术

GraphX 是基于 Spark 的图计算框架，存储单位是 RDD，可以用于大规模的图计算，如社交网络关系等。GraphX 主要描述的是有向图，即包括顶点和边两种属性的图，它提供了三种视图，分别是：顶点 (Vertex)、边 (Edge) 和边三元组 (EdgeTriplet)，图计算就是在以上三种视图上进行的。GraphX 实现了一些常用的图算法模型，如相邻顶点收集算法、PageRank 算法、图中三角形统计算法、pregel 图计算框架等等。在 GraphX 上实现的一系列经典的图算法使得用户在 Spark 上编写程序更加简单。

3 结束语

在大数据环境下，传统的单机模式已不能处理海量数据。Hadoop 虽然能处理大规模数据，但它更加擅长离线的批量数据，且耗时长。Spark 既能处理流式数据又能处理批量数据，它使用 RDD 的内存抽象，使得代码的编写变得简洁，以其内存计算的优势，大大加快了数据处理速度，拥有的各个组件具有各自的优势，各组件数据也能通过 RDD 交互，构成了一站式的大数据分析处理平台。由此可看出，Spark 拥有先进的设计理念，是大数据处理平台的首选。