Presto 基本概念：Driver, Split 和 Pipeline

在使用 Presto 时，我们经常会听说 Query、Stage、Task 等概念，很多人会搞不清楚这些概念，所以会导致一些误解，本文将简单地介绍一下这些基本的概念是指

Statement

Statement语句。其实就是指我们输入的SQL语句。Presto支持需要ANSI标准的SQL语句。这种语句由子句(Clause)、表达式（Expression）和断言(Predicate)组成。

Presto为什么将语句(Statement)和查询(Query)的概念分开呢？

因为在Presto中，语句和查询本身就是不同的概念。语句指的是终端用户输入的用文字表示的SQL语句；当Presto执行输入的SQL语句时，会根据SQL语句生成查询执行计划，进而生成可以执行的查询(Query)，而查询代表的是分布到所有的Worker之间执行的实际查询操作。

Query

Query即查询执行。当Presto接收一个SQL语句并执行时，会解析该SQL语句，将其转变成一个查询执行和相关的查询执行计划。一个查询执行代表可以在Presto集群中运行的查询，是由运行在各个Worker上且各自之间相互关联的阶段（Stage）组成的。

那么SQL语句与查询执行之间有什么不同呢？
其实很简单，你可以认为SQL语句就是提交给Presto的用文字表示的SQL执行语句。而查询执行则是为了完成SQL语句所表述的查询而实例化的配置信息、组件、查询执行计划和优化信息等。一个查询执行由Stage、Task、Driver、Split、Operator和DataSource组成。这些组件之间通过内部联系共同组成了一个查询执行，从而得到SQL语句表述的查询，并得到相应的结果集。

Stage

Stage即查询执行阶段。当Presto运行Query时，Presto会将一个Query拆分成具有层级关系的多个Stage，一个Stage就代表查询执计划的一部分。例如，当我们执行一个查询，从Hive的一张具有1亿条记录的表中查询数据并进行聚合操作时，Presto会创建一个Root
Stage（后面会介绍，该Stage就是Single Stage），该Stage聚合其上游Stage的输出数据，然后将结果输出给Coordinator，并由Coordinator将结果输出给终端用户。

通常情况下Stage之间是树状的层次结构。每个Query都有一个Root Stage。该Stage用于聚集所有其他Stage的输出数据，并将最终的数据反馈给终端用户。需要注意的是，Stage并不会在集群中实际执行，它只是Coordinator用于对查询计划进行管理和建模的逻辑概念。每个Stage（除了Single Stage和Source Stage）都会有输入和输出，都会从上游Stage读取数据，然后将产生结果输出给下游Stage。需要注意的是；Source Stage没有上游，它从Connector获取数据，Single
Stage没有下游，它的结果直接输出给Coordinator，它由Coordinator输出给终端用户。

Presto中的Stage共有4种，具体介绍如下：

- Coordinator_Only
  ：这种类型的Stage用于执行DDL或者DML语句中最终的表结构创建或者更改。
- Single：这种类型的Stage用于聚合子Stage的输出，并将结果数据输出给终端用户。
- Fixed
  ：这种类型的Stage用于接受其子Stage产生的数据并在集群中对这些数据进行分布式的聚合或者分组计算。
- Source
  ：这种类型的Stage用于直接连接数据源，从数据源读取数据，在读取数据的同时，该阶段也会根据Presto对查询执行计划的优化完成相关的断言下发(Predicate PushDown)和条件过滤等。

说明：一个SQL查询可以被分解 多个前后关联的Stage,在这里我们约定：按照数据的流向，越靠近数据源的Stage越处于上游，越远离数据源的Stage越处于下游。

Exchange

Exchange的字面意思就是“交换”。Presto的Stage是通过Exchange来连接另一个Stage的。Exchange用于完成有上下游关系的Stage之间的数据交换。在Presto中有两种Exchange：Output Buffer和Exchange Client。生产数据的Stage通过名为Output Buffer的Exchange将数据传送给其下游的Stage。消费数据的Stage通过名为Exchange从上游Stage读取数据。
如果当前Stage是Surce类型的Stage，那么该Stage则是直接通过相应的Connector从数据源读取数据的。而该Stage则是通过名为Source
Operator的Operator与Connector进行交互的，例如，一个Source
Stage直接从HDFS获取数据，那么这种操作不是通过Exchange
Client来完成的，而是通过运行于Driver中的Source Operator来完成的。

Task

从前面的介绍中可以知道，Stage并不会在Presto集群中实际运行，它仅代表针对于一个SQL语句查询执行计划中的一部分查询的执行过程，只是用来对查询执行计划进行管理和建模。Stage在逻辑上又被分为一系列的Task，这些Task则是需要实际运行在Presto的各个Worker节点上的。
在Presto集群中，一个查询执行被分解成具有层次关系的一系列的Stage，一个Stage又被拆分为一系列的Task。每个Task处理一个或者多个Split。每个Task都有对应的输入和输入。一个Stage被分解为多个Task，从而可以并行地执行一个Stage。Task也采用了相应的机制：一个Task也可以被分解为一个或者多个Driver，从而并行地执行一个Task。

Driver

一个Task包含一个或者多个Driver。一个Driver其实就是作用于一个Split的一系列Operator的集合。因此一个Driver用于处理一个Split，并且生成相应的输出，这些输出由Task收集并且传送给下游Stage中的Task。一个Driver拥有一个输入和一个输出。

Operator

一个Operator代表对一个Spit的一种操作，例如过滤、加权、转换等。一个Operator依次读取一个Split中的数据，将Operator所代表的计算和操作作用于Split的数据上，并产生输出。每个Operator均会以Page为最小处理单元分别读取输入数据和产生输出数据。Operator每次只会读取一个Page对象，相应地，每次也只会产生一个Page对象。

Split

Split即分片，一个分片其实就是一个大的数据集中的一个小的子集。而Driver则是作用于一个分片上的一系列操作的集合，而每个节点上运行的Task，又包含多个Driver,从而一个Task可以处理多个Split。其中每一种操作均由一个Operator表示。分布式查询执行计划的源Stage(Source Stage)通过Connector从数据源获取多个分片。Source Stage对Split处理完毕之后，会将输出传递给其下游Stage（通常其下游Stage的类型为Fixed或者Single）。
当Presto执行一个查询的时候，首先会从Coordinator得到一个表对应的所有Split。然后Presto就会根据查询执行计划，选择合适的节点运行相应的Task处理Split。

Page

Page是Presto中处理的最小数据单元。一个Page对象包含多个Block对象，而每个Block对象是一个字节数组，存储一个字段的若干行。多个Block横切的一行是真实的一行数据。一个Page最大为1MB，最多16*1024行数据。