Tinkerpop OLAP 以及图计算

金世钰

2019.08.11

1 Tinkerpop 框架以及 gdm

关于 Tinkerpop，其官方网站对它自己的定位是：开源的图计算框架。其作用是通过以提供 API 以及相关工具的方式，简化开发者创建图应用（个人觉得包括图数据库及其处理器和其他相关应用）的难度。Tinkerpop 是一个位于不同的图数据库和图处理器上的抽象层，可以视作实现一个图数据库的公约，只要将 Tinkerpop 这个抽象层具体化，就可以开发出自定义的图应用。gdm 则是基于 Tinkerpop 图计算框架开发的一款数据库产品，

Tinkerpop 框架从大的方向来说，包含 OLTP 和 OLAP 两块，分别代表着在线事务处理和在线分析处理，如果要基于 Tinkerpop 实现一个图数据库，那么 OLTP 这一块是必不可少的，而 OLAP 是可选的。对一个数据库来说，CRUD 无可或缺并且恰好属于 OLTP 的范畴，所以对基于 Tinkerpop 开发图数据库产品的开发者而言，OLTP 部分的接口必须实现，OLAP 则提供了除 OLTP 之外的其他功能，比如基于某种算法对整个数据库里面的数据进行分析得出某些想要结论，这对一个图数据库来说不是必须提供的功能，属于可选部分。Tinkerpop 自身的架构（图 1）对外提供了 OLTP 和 OLAP 的接口（Provider API），交由图数据库的开发者去具体实现，当然 Tinkerpop 官方也提供了一个对这些 API 的实现-TinkerGraph。

目前基于 Tinkerpop 实现的图产品（图数据库）包括以下列表中几个比较典型的图数据库，而且在 DB-engines 这个全球图数据库产品排名上，下方列表中的很多产品也是名列前茅（图 2），可以看出，Tinkerpop 在图数据库界极受欢迎，同时表现也很优秀。

- Microsoft Azure Cosmos DB

- OrientDB
- ArangoDB
- JanusGraph
- Amazon Neptune
- Baidu HugeGraph
- Alibaba GDB

2 Tinkerpop OLAP

Tinkerpop 的 OLAP 部分主要用于全图分析，使用了批量同步并行计算模型（Bulk Synchronous Parallel，BSP），同时规定了一系列和 BSP 模型中相对应的编程接口，这一部分内容将在下面阐述。

2.1 BSP 与 Tinkerpop OLAP

批量同步并行计算模型（Bulk Synchronous Parallel，BSP），由哈佛大学的 Leslie Valiant 在 20 世纪 80 年代提出，是一个用于设计并行算法的桥接模型。一个 BSP 模型包括以下三个要点

1. 拥有计算能力和内存事务的组件（比如处理器）
2. 一个能在上述组件中按照路由传递信息的网络
3. 一个能同步上述组件的硬件设施

说的直白点，一个 BSP 模型就是一系列的处理器，每个处理器都具有计算能力且配置着高速的本地内存，所有处理器之间可以通过通信网络之间进行信息交流。一个基于 BSP 模型的算法，最依赖的是上述的第三个要点，一个 BSP 算法的执行由很多超步组成，每个超步都包含三部分

1. 并行计算

2. 通信

3. 栅栏同步

2.2 Tinkerpop 单节点 OLAP 的架构与实现

2.3 Tinkerpop 分布式 OLAP 架构设想