分布式元数据管理

王超　北京东方国信科技股份有限公司

2020年11月29日

提 纲

- 分布式系统架构
  - 集群架构
  - 元数据分层管理
  - 数据分布
  - 读写流程
- 分布式元数据管理机制
  - 自动创建Schema机制
  - 远程元数据管理机制
- 总结

资料整理
火龙果软件
荐

IoTDB基本概念

Storage Group（存储组/数据库）

- 独立的物理存储和资源管理
- 可管理多个设备
- root.sgcc/root.ln

Device（设备）

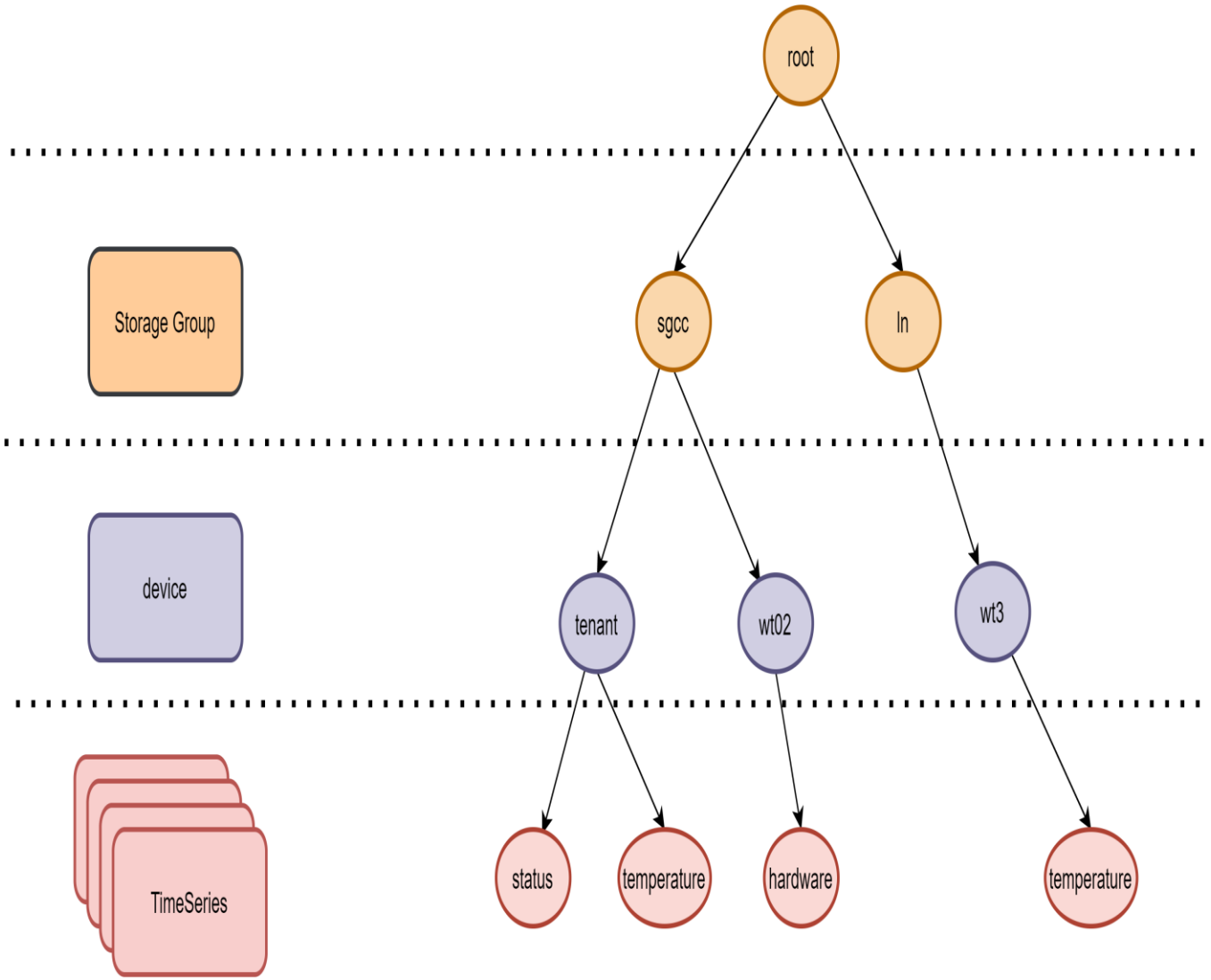
- 一个工业设备实例，可以有多个属性，隶属于某个存储组
- root.sgcc.tenant

Measurement（测点）

- 一个设备上的多个传感器，采集不同的数据
- status/temperature

Time Series（时间序列）

- Device + Measurement（设备+测点）
- root.sgcc.tenant.status



资料整理
火龙果软件
荐

IoTDB基本概念

元数据

- Storage Group（存储组），物理存储资源独立
- Measurement（测点），包括类型、编码方式、压缩方式等
- 树形结构，存储时间序列路径，可支持模糊检索

元数据

- root.ln
- status/temparature, dataType, EncodingType, Compressor
- root.sgcc.tenant.status

数据

- 数据点，即每个传感器采集过来的数据，每个时间点都会带一个时间戳

数据

- <1000, 1.4>,<1001,2.5>...
- <1000, false>,<1001,true>....

资料整理
火龙果软件
荐

分布式系统几大问题

系统架构

- 主从架构（实现比较清晰），典型系统的有Hadoop
- 对等架构（去中心化，扩展性好），典型的系统有Cassandra
- 介于两者之间，轻量级主从架构，典型的系统有Ceph，IoTDB分布式架构也是介于两者之间

数据分布

- 哈希分布，典型的系统有Cassandra
- Range分区，典型的系统有HBase
- Range分区+哈希分布，典型的有InfluxDB

元数据管理

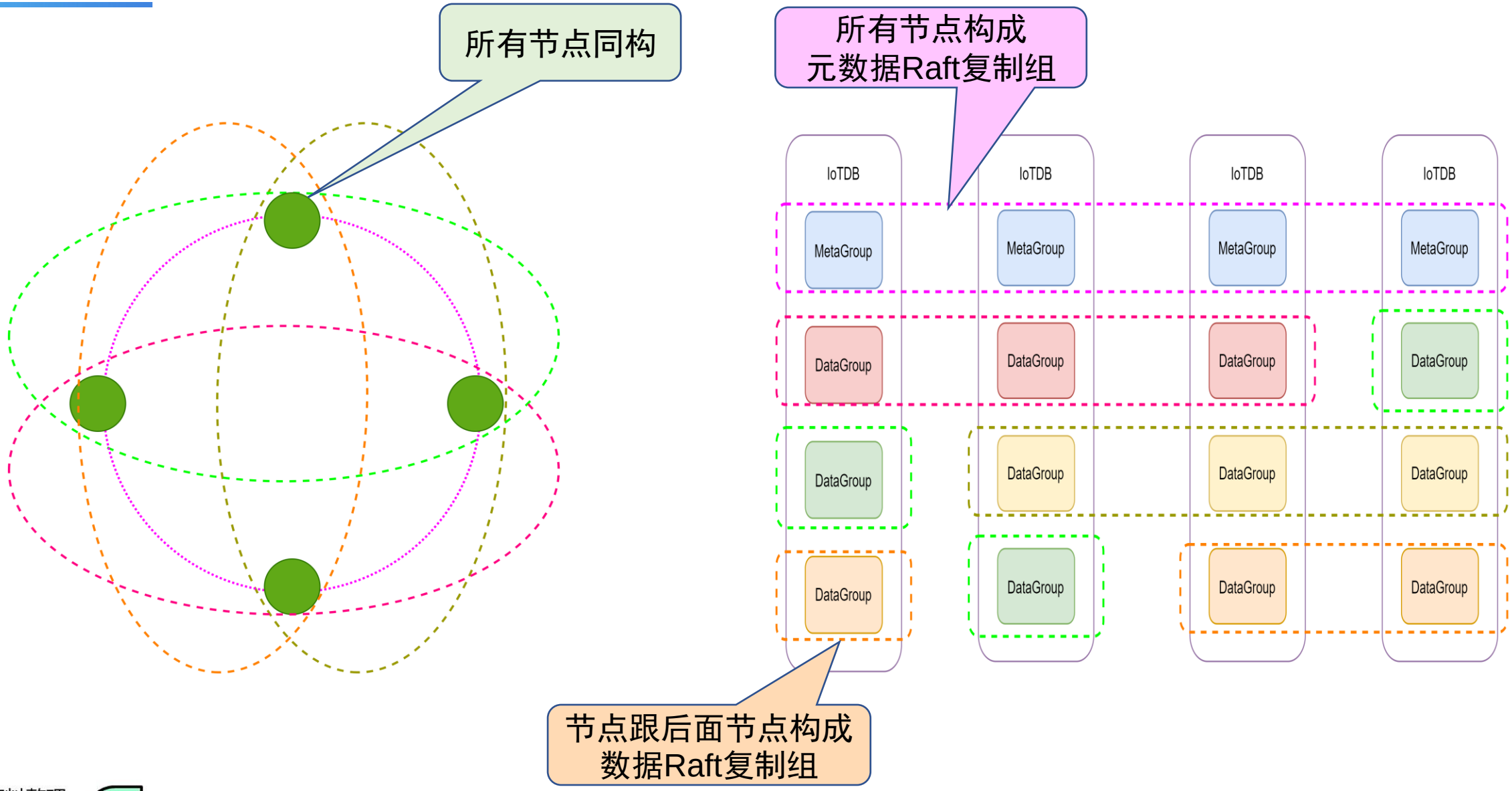
- 集中管理，典型的有HDFS，NameNode管理所有元数据
- 分散管理，典型的有Cassandra，所有节点都维护一部分元数据
- 分层管理，典型的有Ceph，对象先被映射到PG里（一层），PG再映射到OSD里（两层）

数据高可靠

- 单领导者复制
  - 树形复制，Paxos和Raft复制副本数据
  - 链式复制，典型的有HDFS写入时pipeline链式复制
  - 分发复制

资料整理
火龙果软件
荐

集群架构



资料整理
火龙果软件
荐

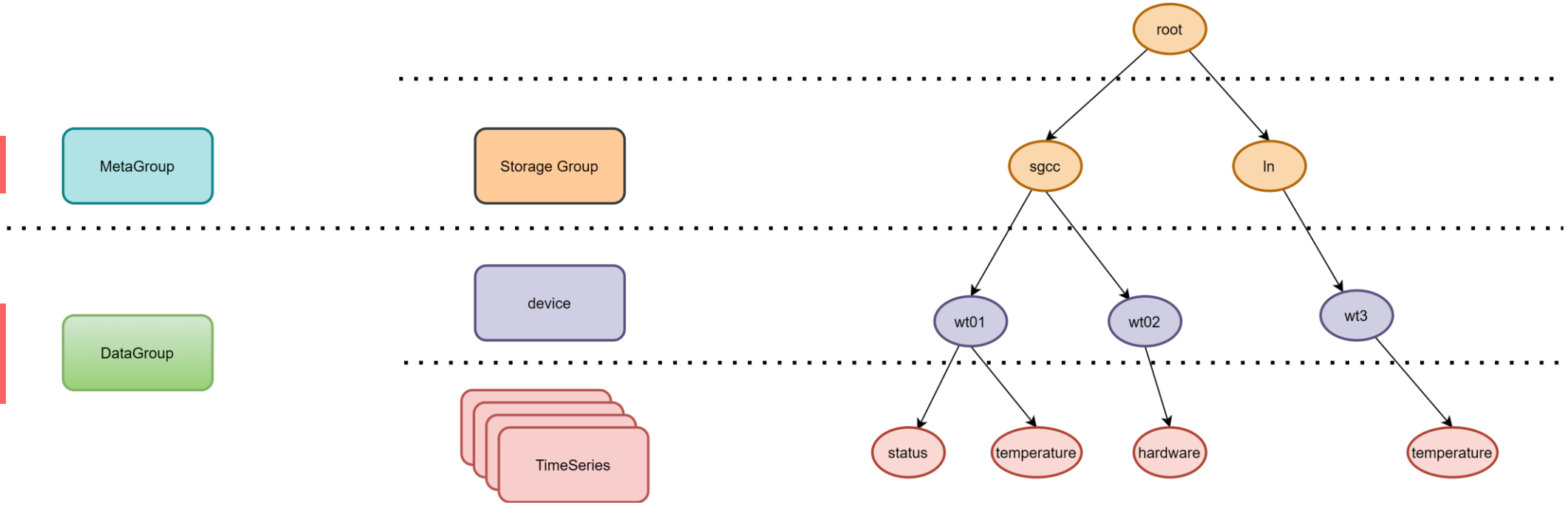
集群架构-元数据分层管理

时间序列元数据分散到各个数据复制组，避免单节点热点

元数据分散到各个节点，能尽量cache在内存中，充分利用各个节点的内存，加快访问

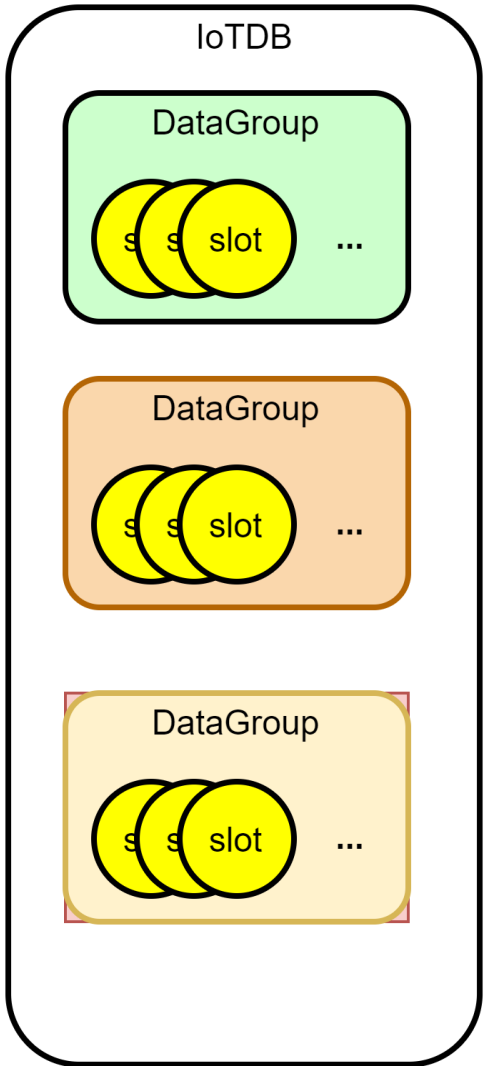
存储组元数据

时间序列元数据

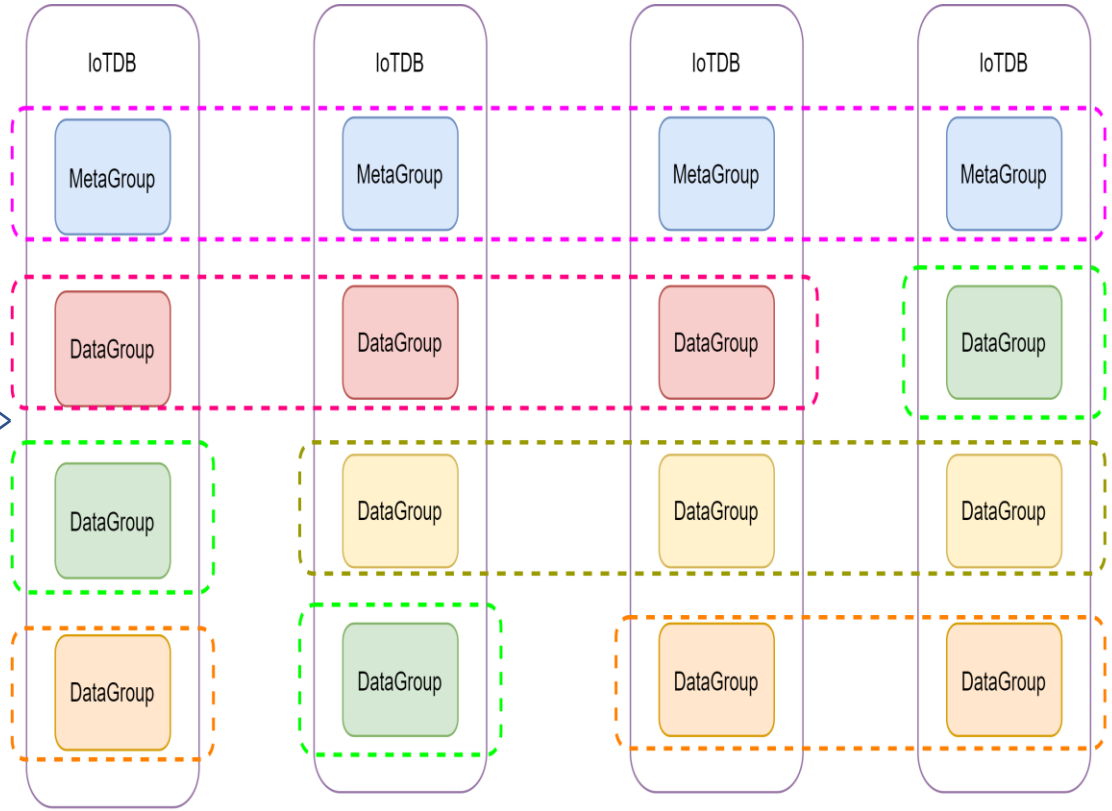


资料整理
火龙果软件
荐

集群架构-数据分布



全集群10000个slot，一致性hash均匀分布到各个DataGroup

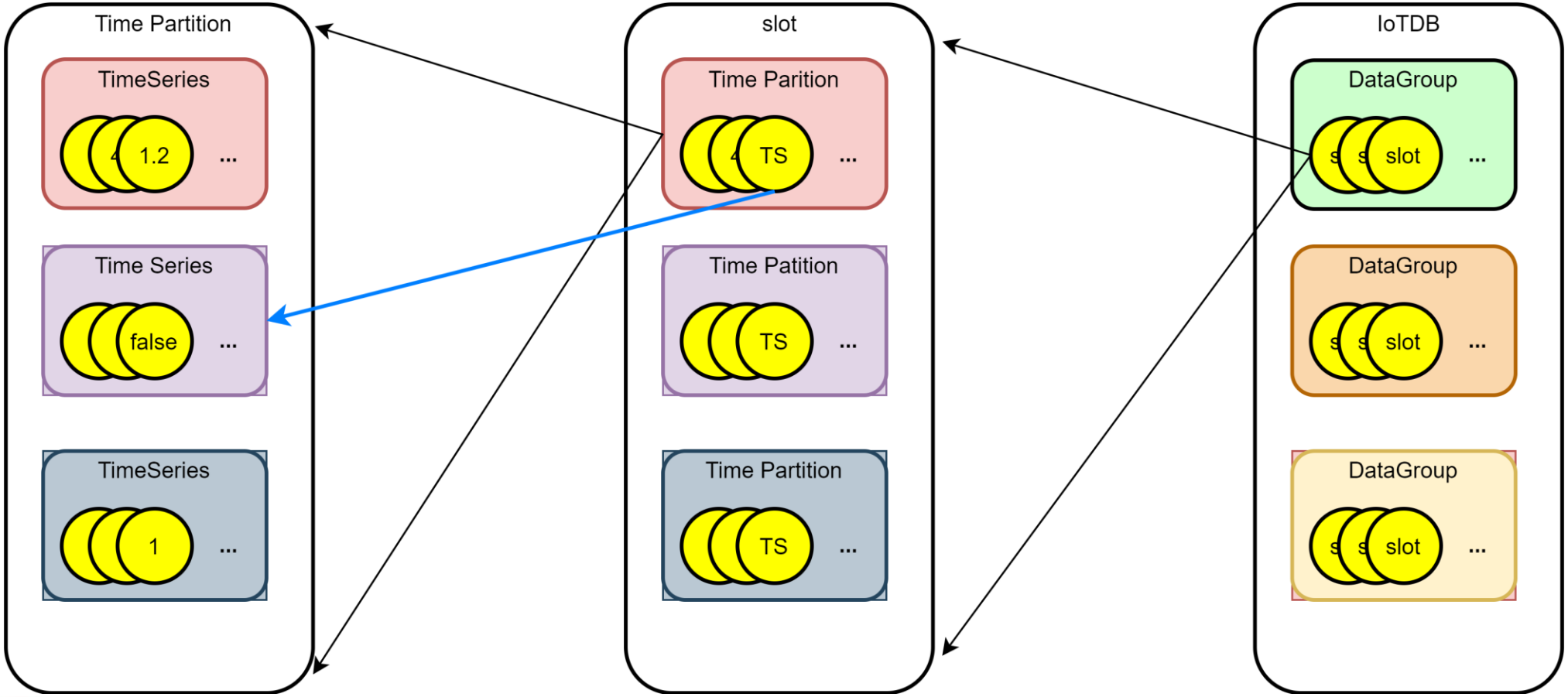


资料整理
火龙果软件
荐

集群架构-数据分布

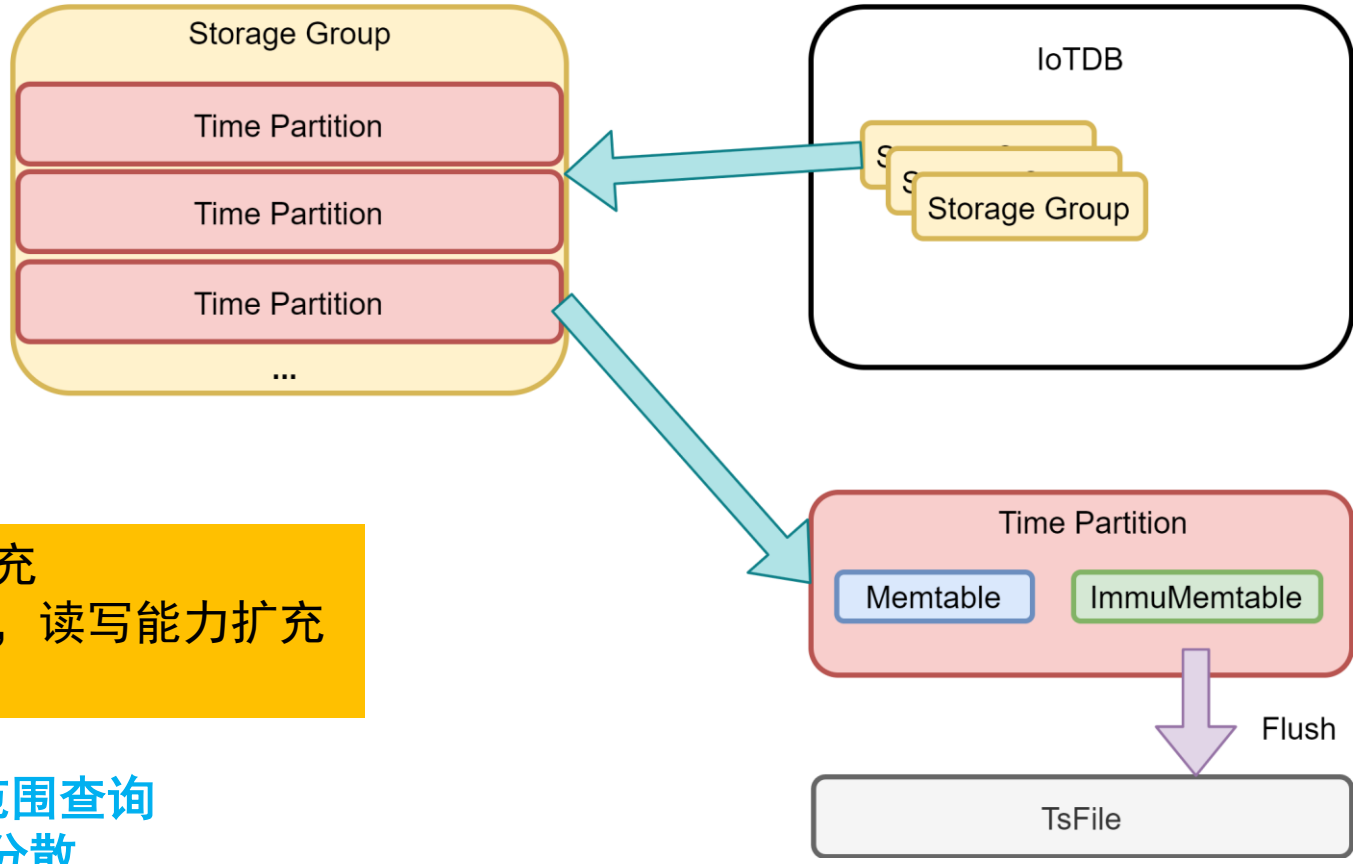
时间分区包含多个时间序列某段时间的数据

slot里包含多个时间分区



资料整理
火龙果软件
荐

集群架构-数据分布



- 数据分散到多个节点，容量扩充
- 读写请求也会分散到多个节点，读写能力扩充
- 对于过期数据，能快速删除

先按时间Range分区，利于时间范围查询
再按存储组Hash分区，利于压力分散

Slot

TimePartitionNum = TimeStamp % PartitionInterval

Hash(Storage Group +TimePartitionNum) % maxSlotNum

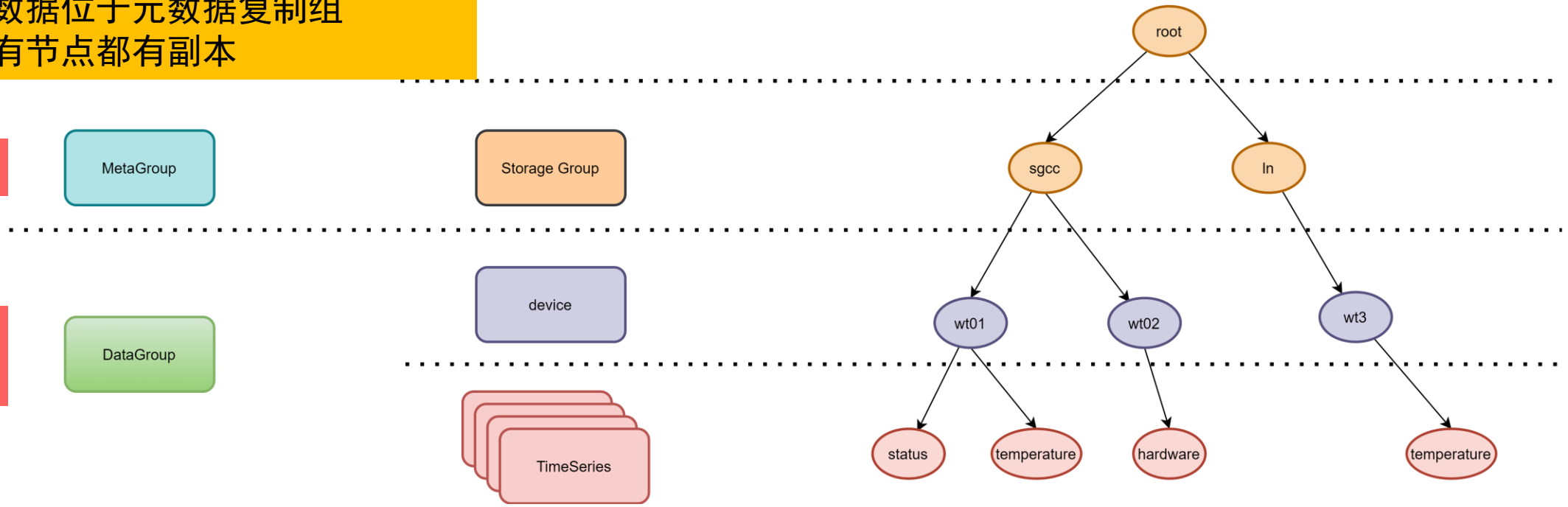
资料整理
火龙果软件
荐

集群架构-元数据分布

存储组元数据位于元数据复制组
全集群所有节点都有副本

存储组元数据

时间序列元数据



时间序列元数据跟数据分布一样的方式

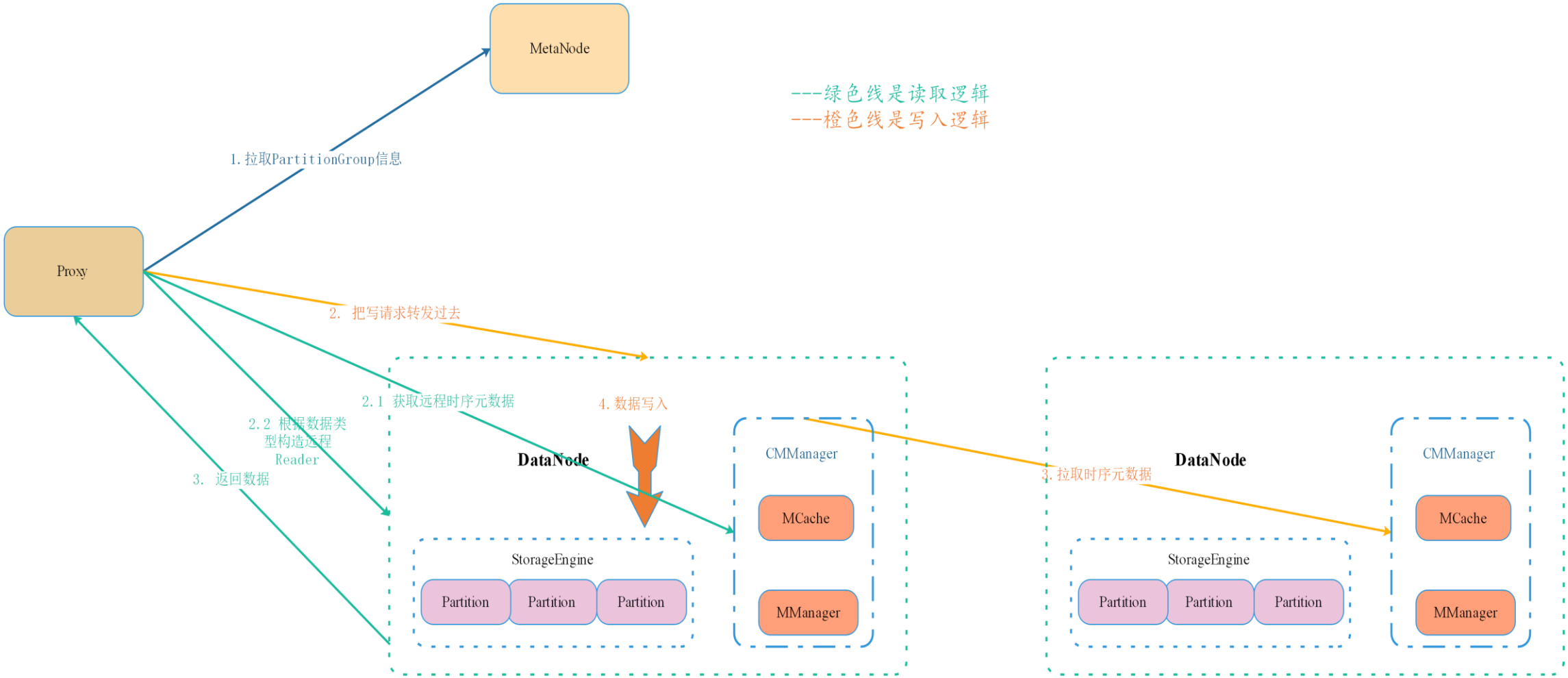
Slot

TimePartition = **0L**

Hash(Storage Group +TimePartition) % maxSlotNum

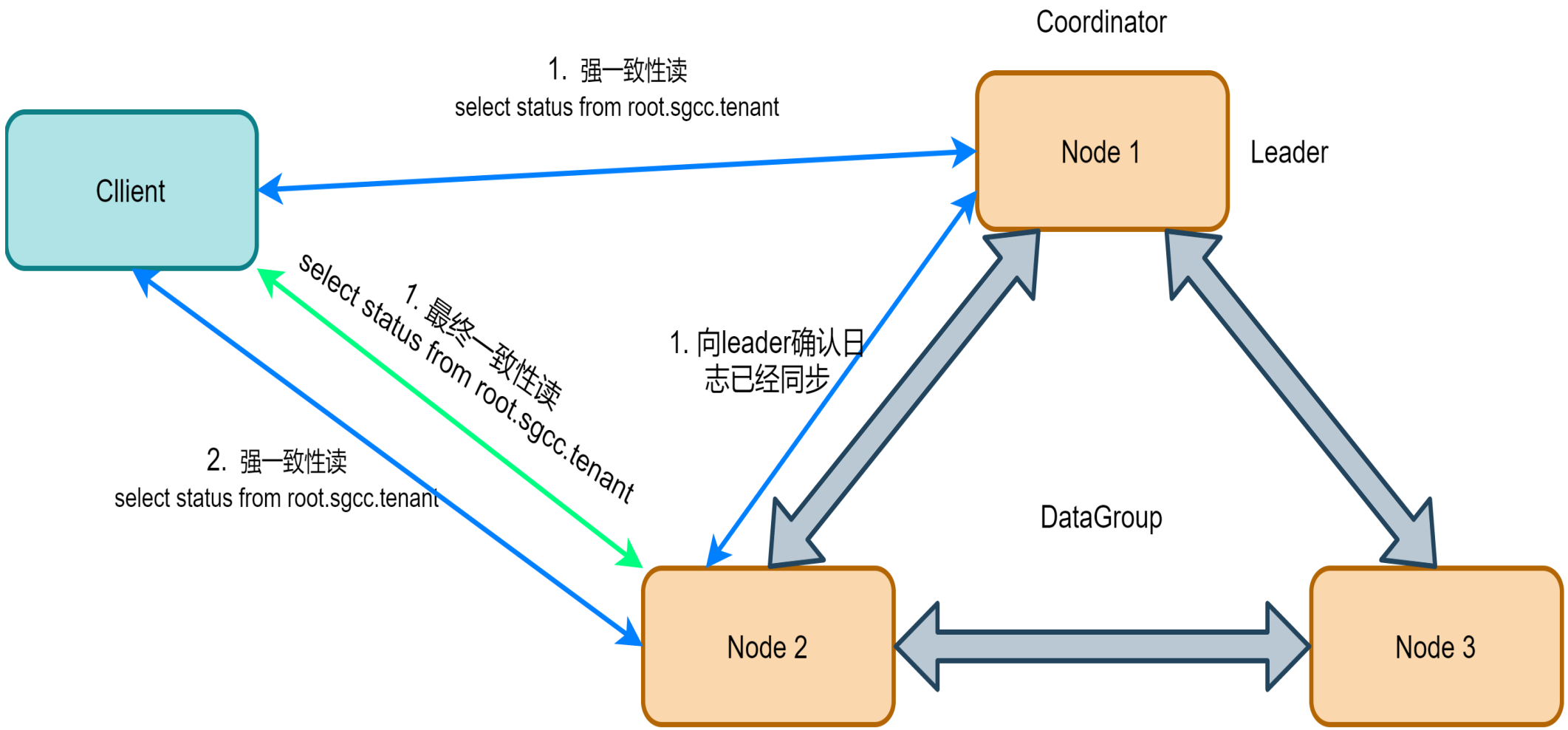
资料整理
火龙果软件
荐

集群架构-读写流程



资料整理
火龙果软件
荐

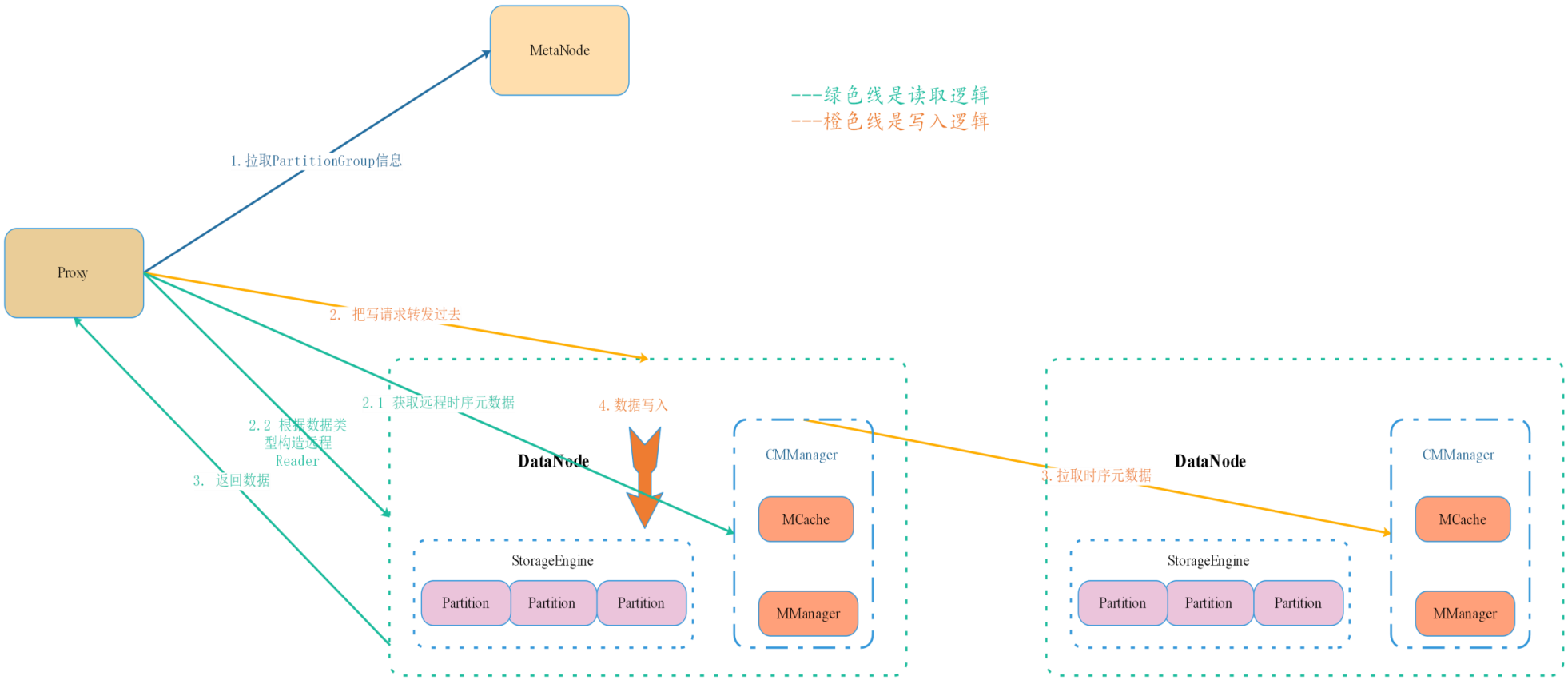
集群架构-读写流程



支持强一致性读和最终一致性读

资料整理
火龙果软件
荐

分布式元数据管理机制-两个问题



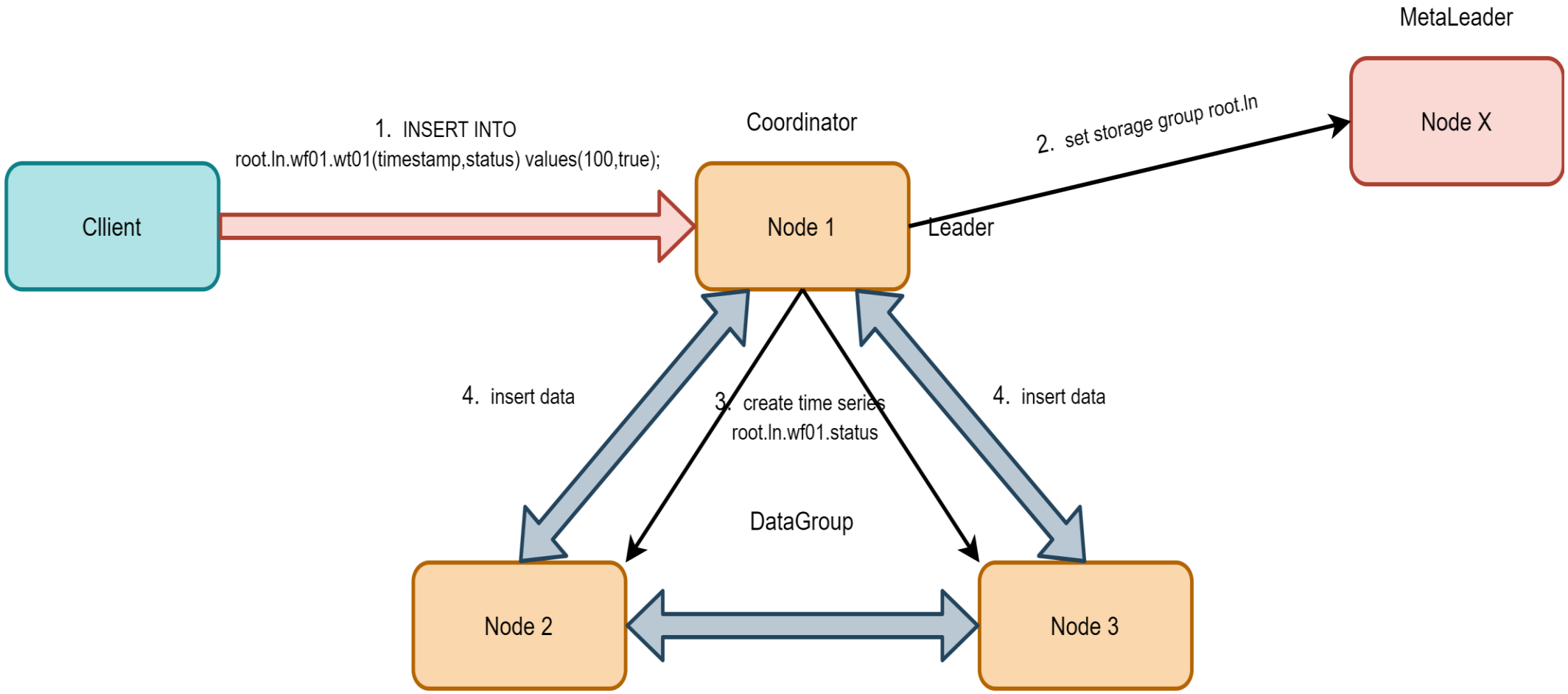
写入时时序元数据未创建怎么处理？

时序元数据不在本地，重启后怎么处理？

资料整理
火龙果软件
荐

分布式元数据管理机制-元数据自动推导创建

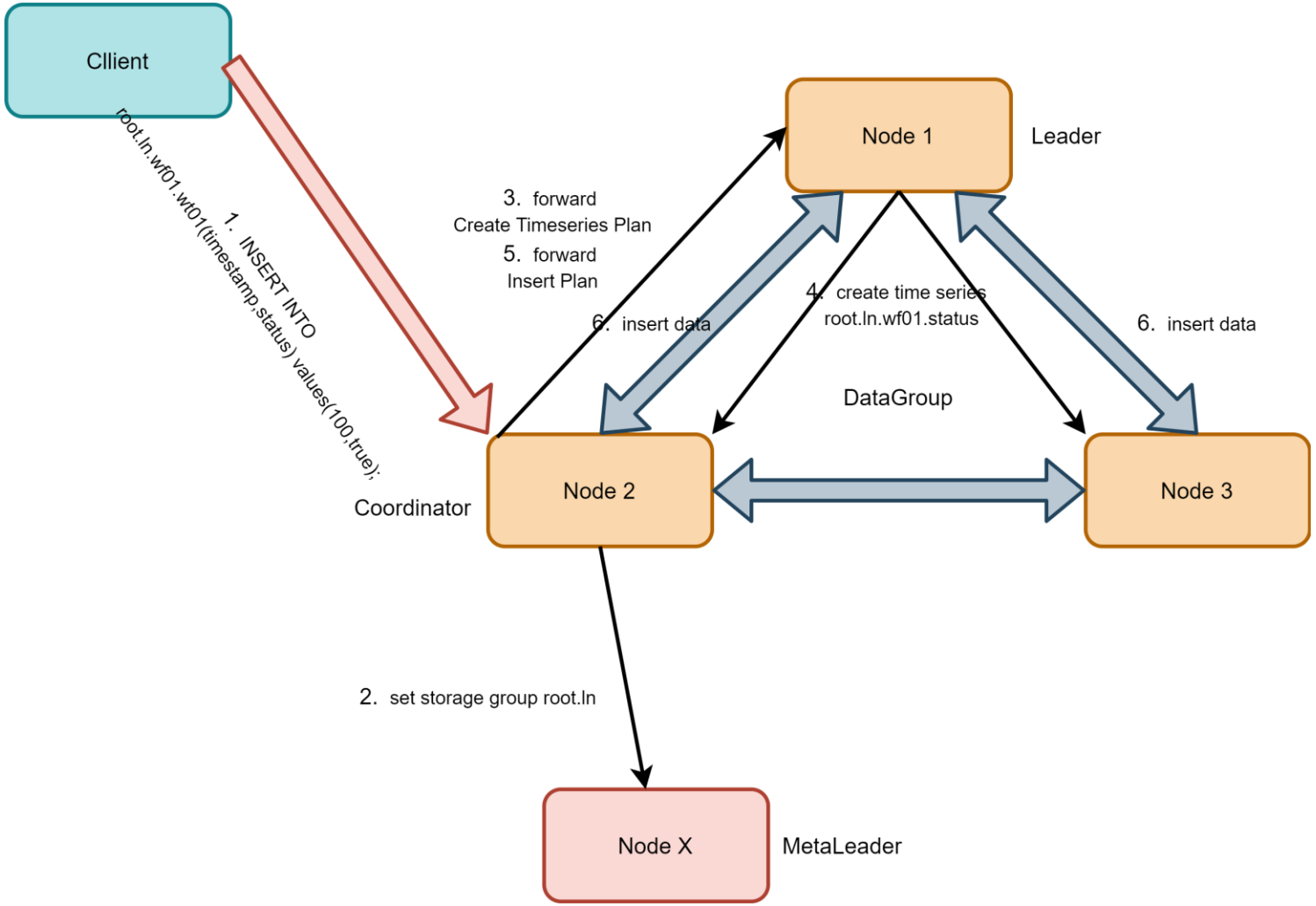
协调节点是数据所在DataGroup的leader



资料整理
火龙果软件
荐

分布式元数据管理机制-元数据自动推导创建

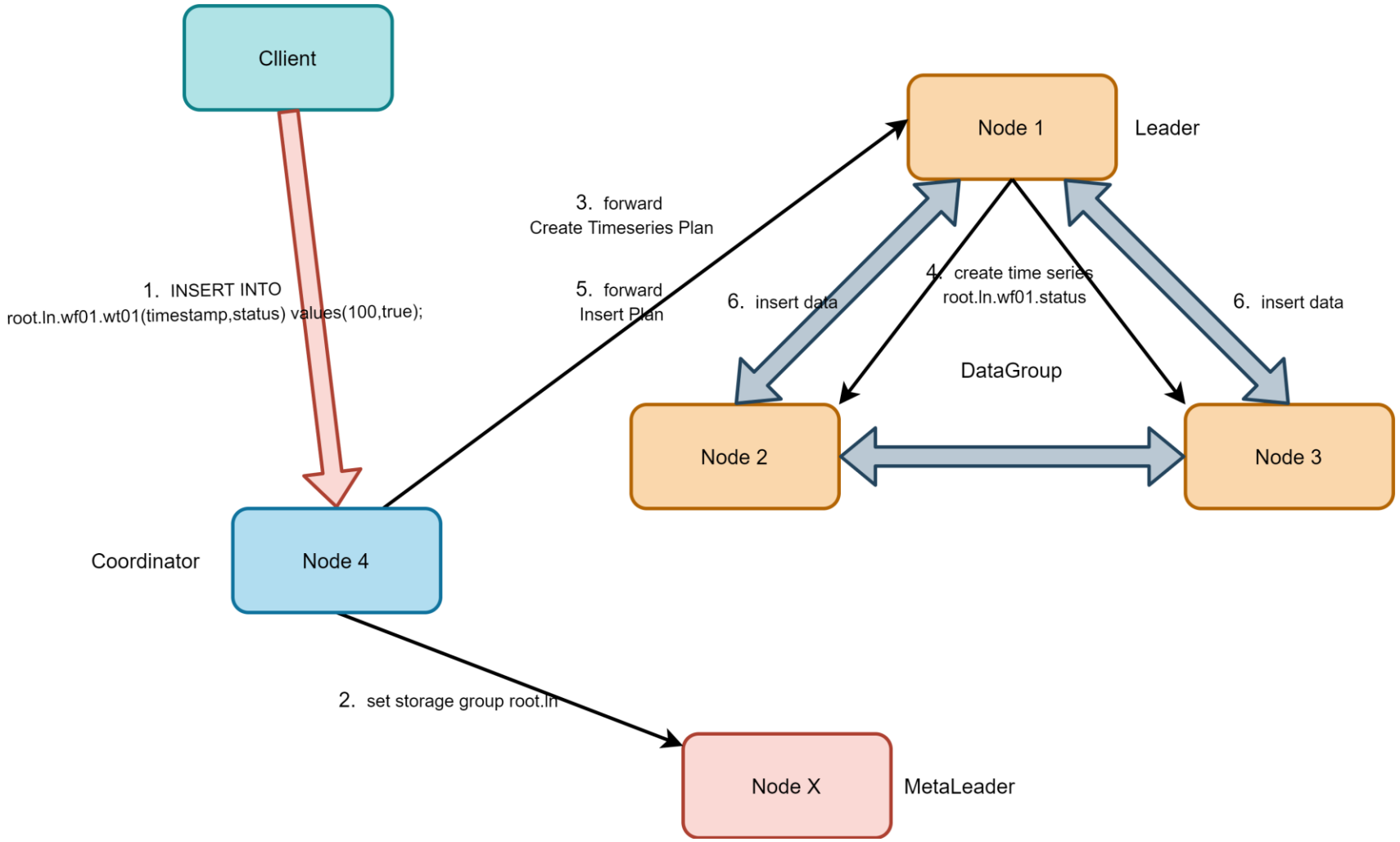
协调节点是数据所在DataGroup的Follower



资料整理
火龙果软件
荐

分布式元数据管理机制-元数据自动推导创建

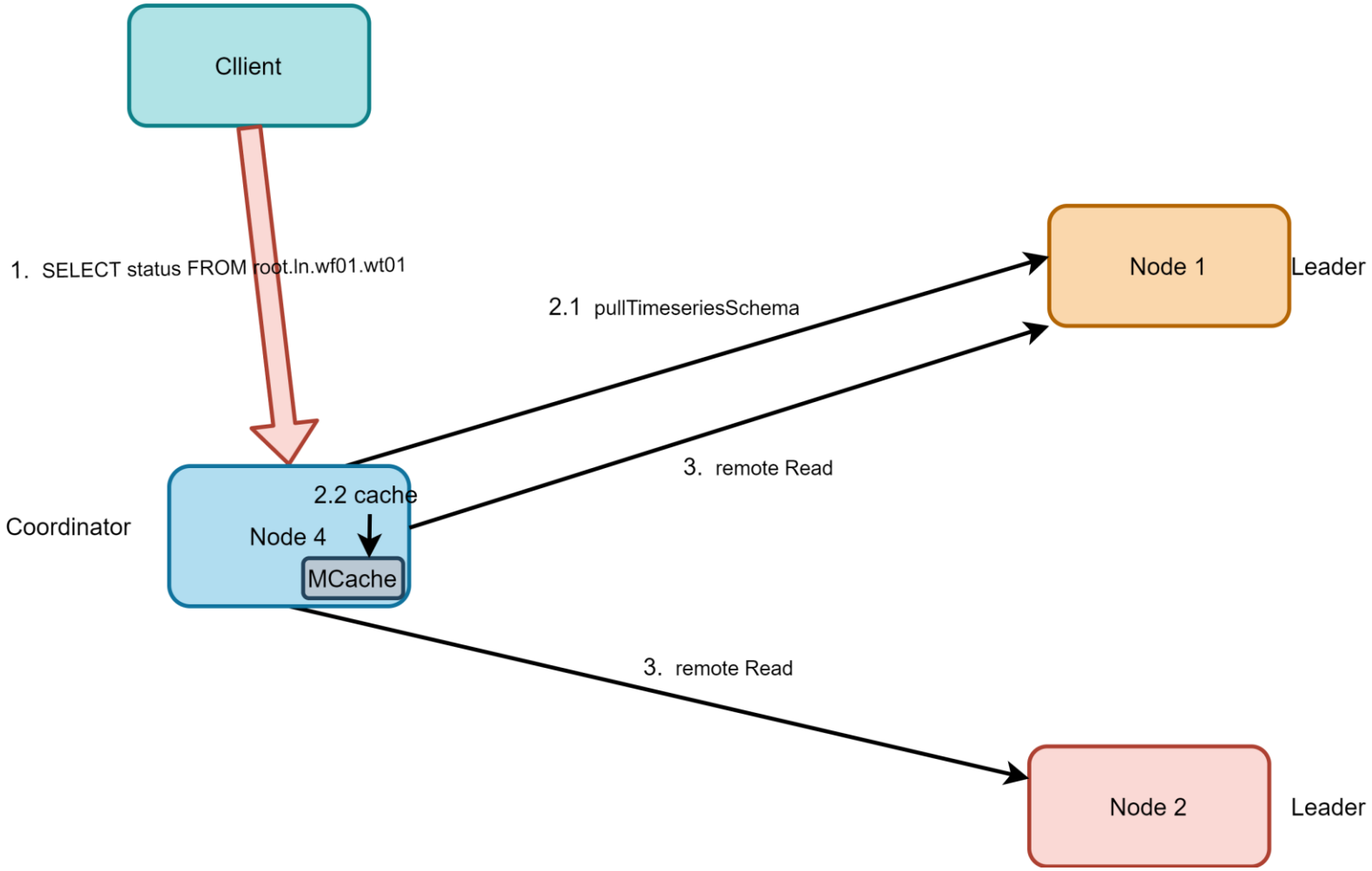
协调节点非数据所在DataGroup



资料整理
火龙果软件
荐

分布式元数据管理机制-元数据拉取和缓存

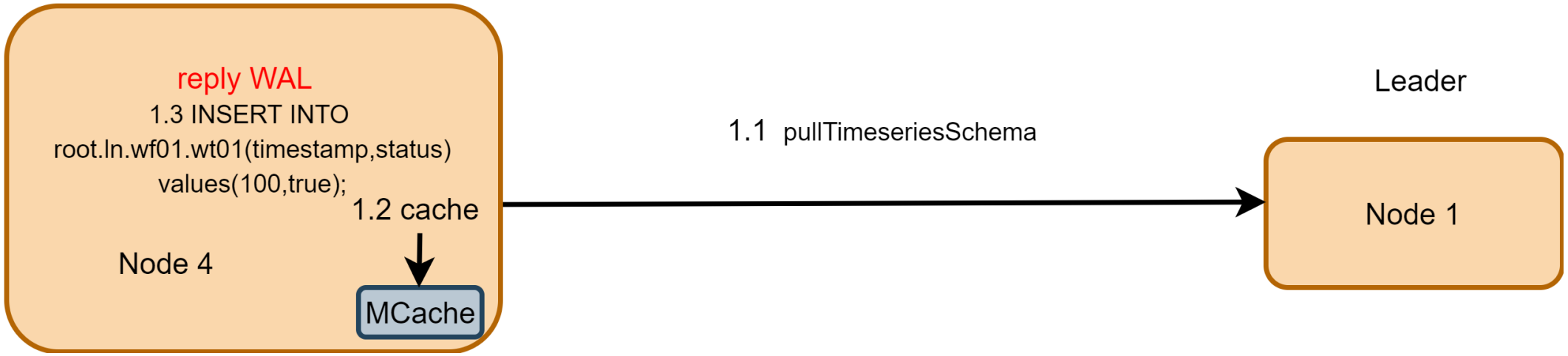
协调节点不属于元数据所在DataGroup，写入和查询时都会拉取元数据



资料整理
火龙果软件
荐

分布式元数据管理机制-元数据拉取和缓存

协调节点不属于是数据所在DataGroup，重启时重新WAL时需要拉取元数据



根源是MCache不会进行持久化，因为会带来性能损失

资料整理
火龙果软件
荐

总结

节点同构

- 无热点节点
- 安装部署简单

元数据分层管理

- 无热点访问
- 充分利用全集群RAM，提高读取效率

数据均匀分布，高可靠

- 一致性hash算法均匀分布数据
- Raft复制组保障数据高可靠

资料整理
火龙果软件
荐

谢 谢 ！