



SPG引擎层的能力和规范

创邻科技CTO 周研

CCKS 2023

浙 江 创 邻 科 技 有 限 公 司

Createlink Technology Co., Ltd



知识图谱的发展历程





● 使用 Triplestore

- 天然支持RDF和OWL
- 标准化的查询语言SPARQL
- 包含丰富的语义关系和推理机制

● 使用 Property Graph

- RDF格式的导入和导出
- 通过插件支持RDF和OWL
- 主流查询语言为Cypher，以及未来的GQL

● 自定义实现



- 具备“属性”的抽象使得数据组织更加简洁
- 免索引邻接提供了更高的查询性能
- 更直观、易于学习的查询语言，如Cypher和GQL
- 通常可以提供丰富的图算法库
- 支持分布式和横向扩展，支持大规模数据量
- 丰富的运维管理能力和成熟的商业解决方案

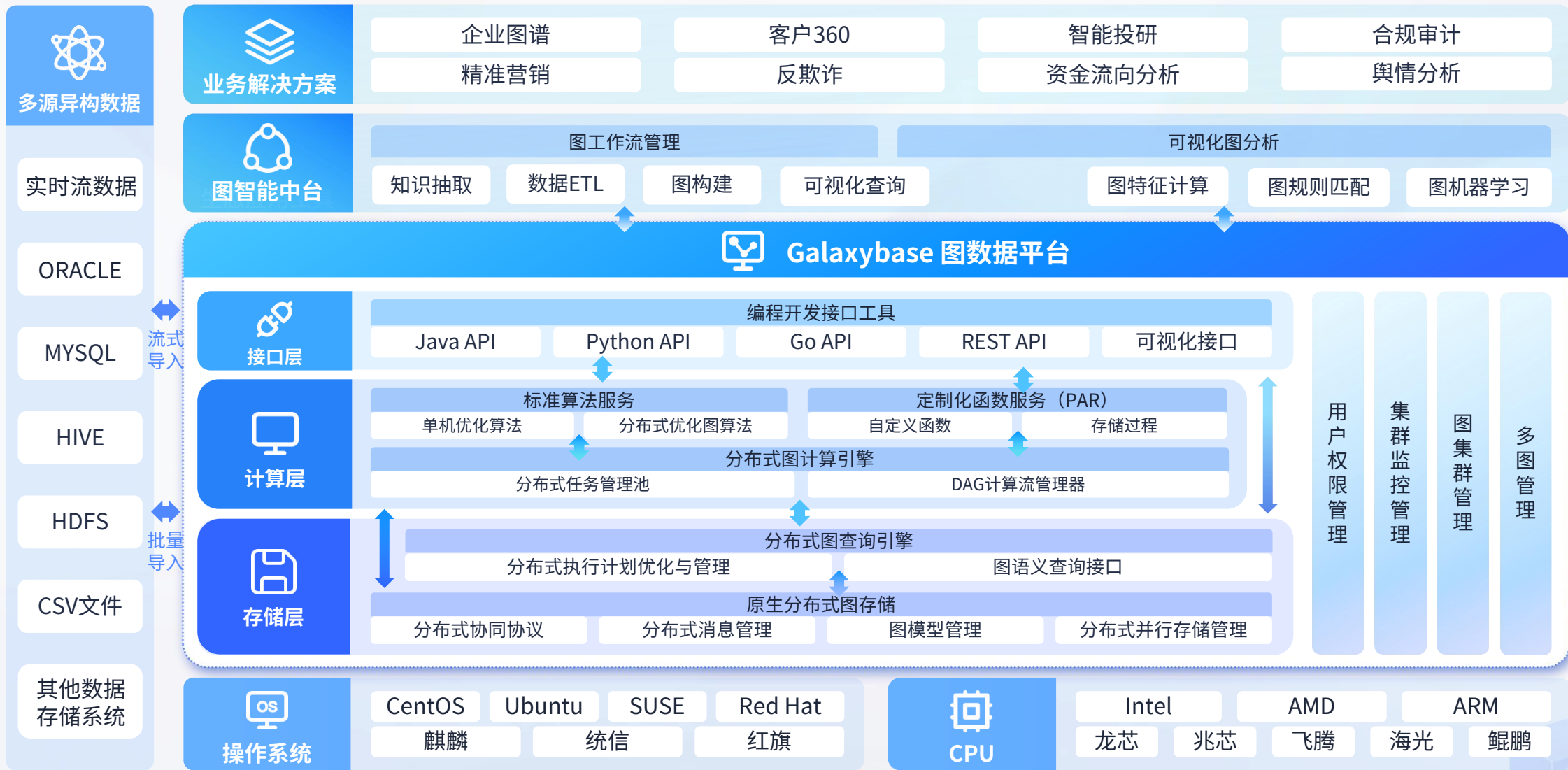




Galaxybase: 国产高性能图平台



Galaxybase是存储和计算内核100%自主知识产权的国产高性能分布式图平台。





标杆用户





- Schemaless还是Strongly-typed
- 相同基础数据、不同使用场景下的图谱复用问题
- 实体之间、关系之间存在逻辑依赖带来的一致性问题
- 业务目标的迁移变化导致schema 及数据的持续膨胀
- 子图的定义和融合问题

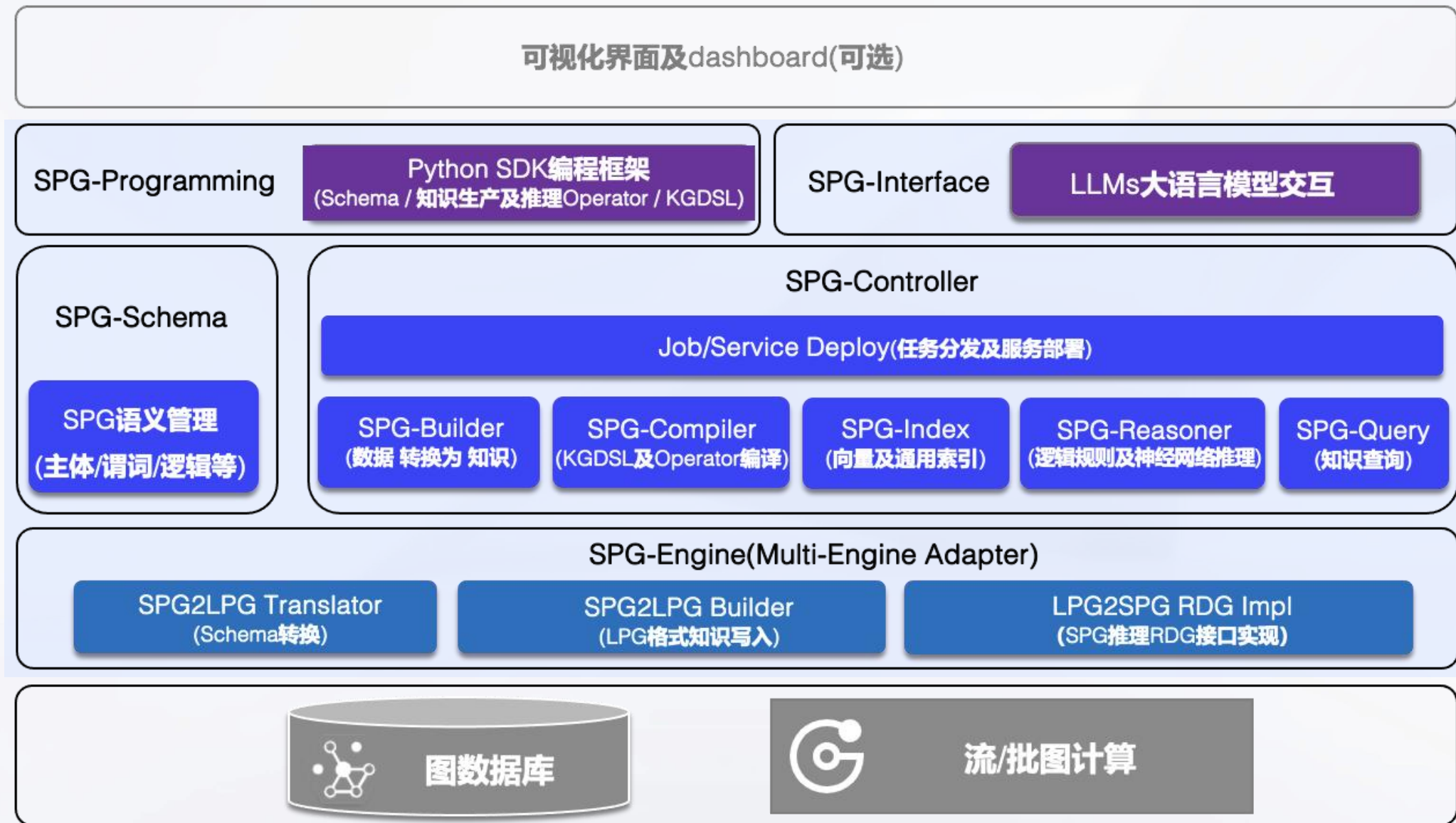
静态的数据和动态的业务需求之间的矛盾



- RDF/OWL学习成本高，Cypher/GQL缺少推理能力
- 进行业务逻辑开发时还要兼具查询性能优化的能力
- 图模式（Schema）的选择会极大的影响产品性能和易用性
- 对于“事件”这样的随时间演化的数据缺少标准处理机制
- 事件传导推理结论的可解释性不够直观

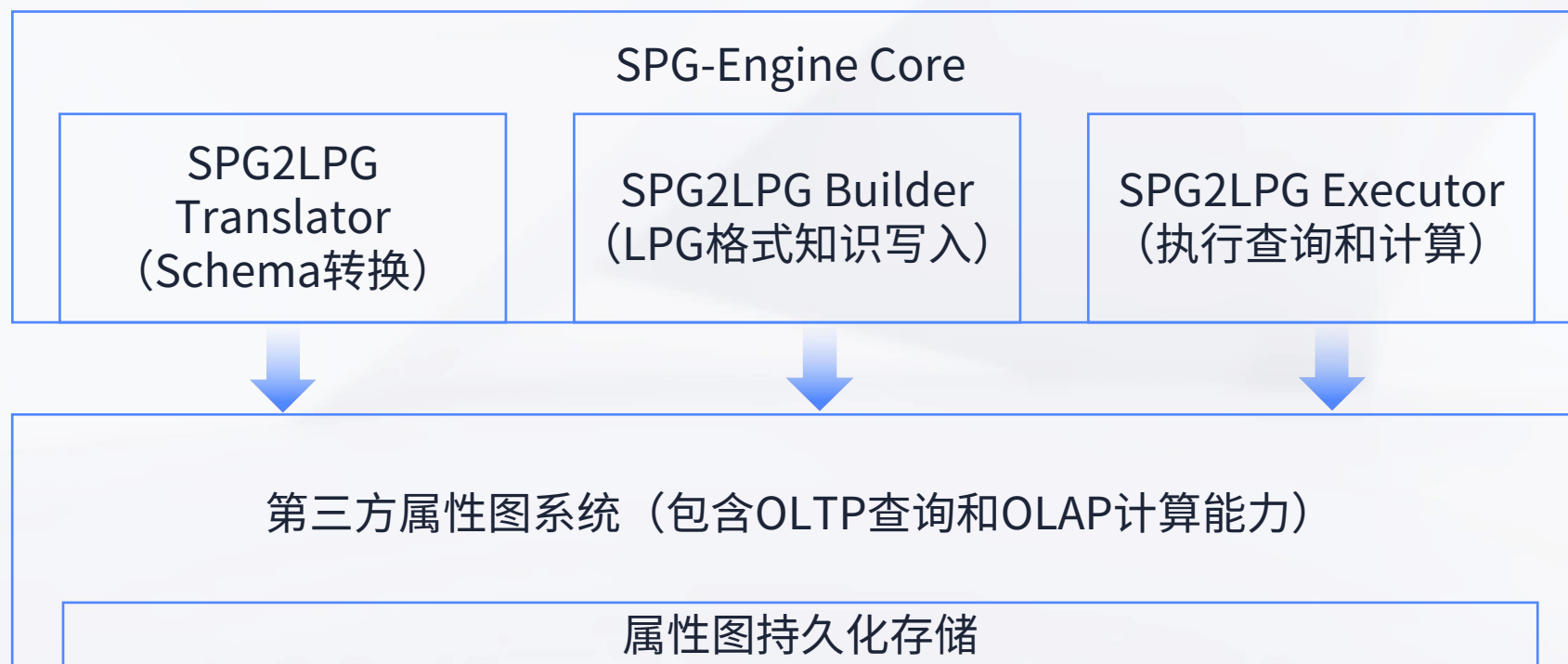
高易用性和强表达力之间的矛盾

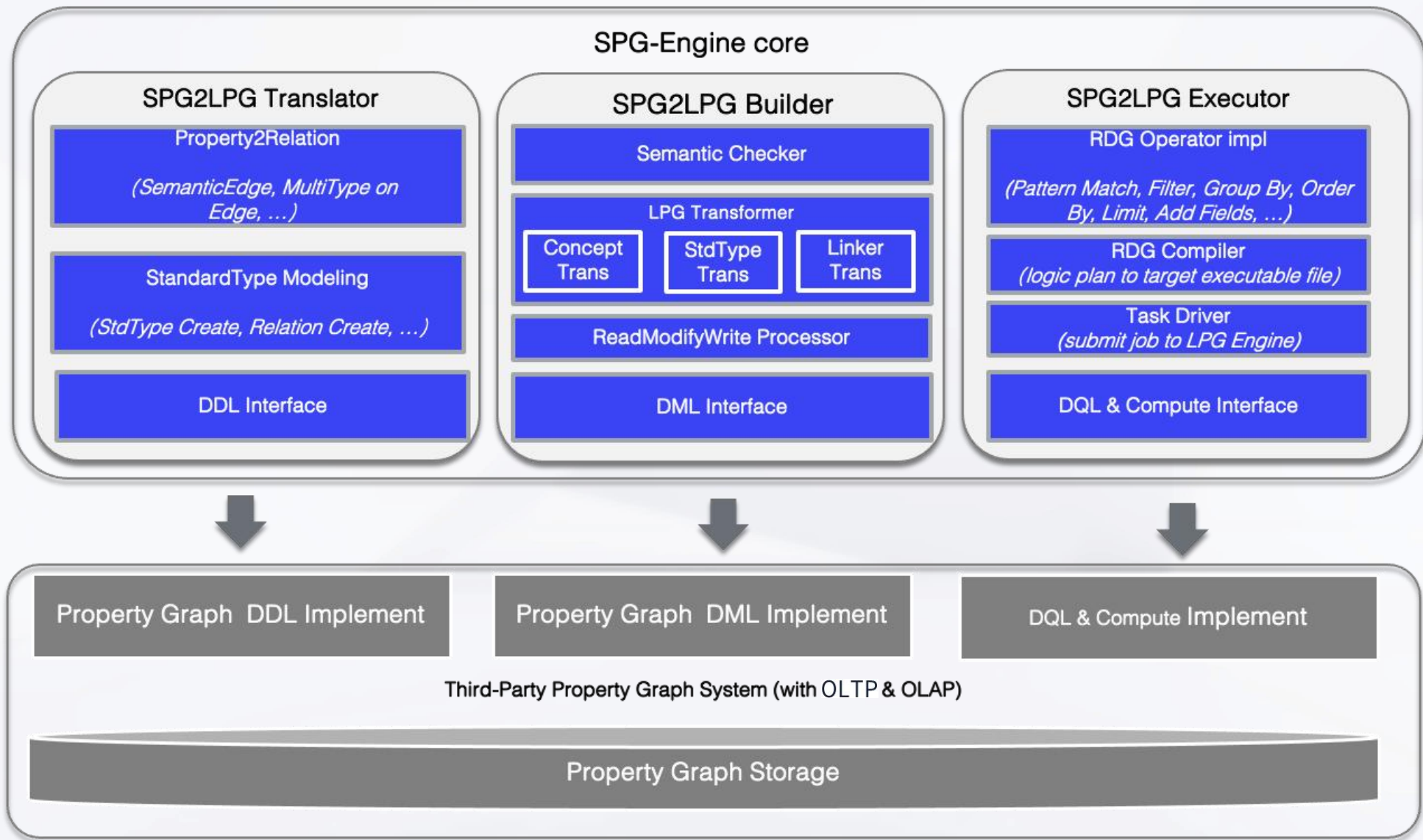






SPG-Engine层是将SPG的推理和计算转换到实际的LPG系统中执行的模块。SPG底层依赖通常包括图存储、图查询、图计算等基础能力，由LPG的图服务厂商提供。为了满足基于SPG的知识图谱推理和服务能力的要求，我们将对引擎能力的要求分为基础能力和进阶能力。





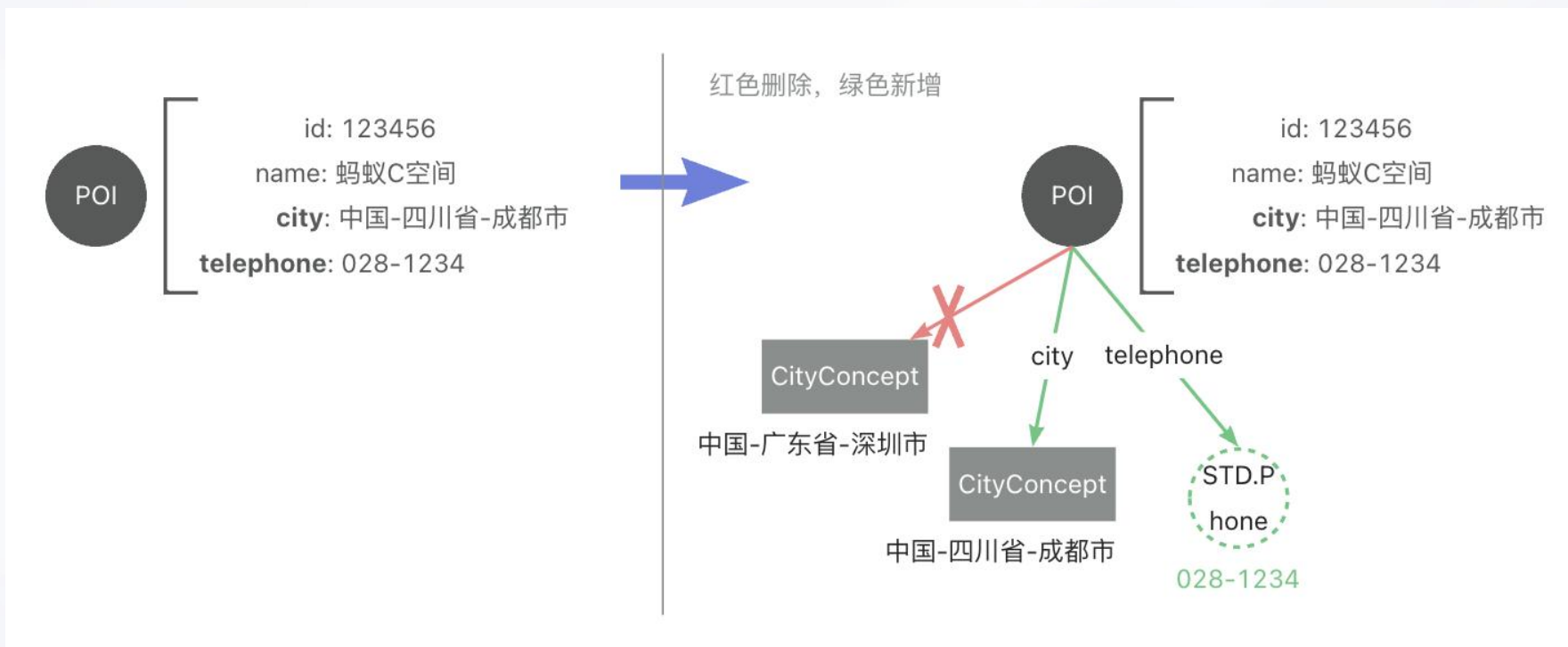


SPG2LPG Translator主要负责将SPG的Schema转换为LPG的Schema格式

SPG Meta Model	作用	LPG Meta Model
Class	实体类型	Node
Concept	概念	Node
NormalizedProp	标准属性	Node
Event	事件	Node
leadTo	概念间的因果关系谓词	Edge
hypernym	概念间的上下位谓词	Edge
object	事件的客体	Edge
subject	事件的主体	Edge
subClassOf	实体类型的层级关系	Edge
RelationShip	关系类型	Edge
subRelOf	关系类型间的层级关系	Edge
MC	关系的多元约束	Edge

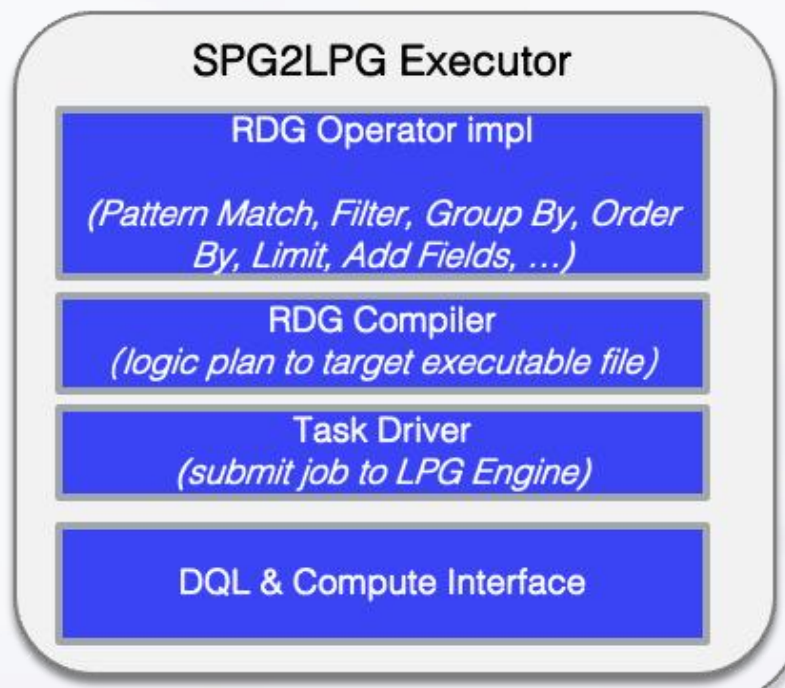


SPG2LPG Builder将SPG格式的数据转换为LPG格式的数据，包括导入实体、删除实体、导入关系、删除关系、导入概念和删除概念等操作。





SPG2LPG Executor主要执行由SPG-Reasoner下发的基于RDG算子（Resilient Distributed Graph）组成的执行计划，执行计划组织方式为树状结构，按照后序遍历方式一次执行树上算子。



- RDG Operator Impl, RDG模型算子实现层，依据底层的LPG引擎分别实现各个算子定义的功能，例如Pattern Match、Filter等
- RDG Compiler, RDG编译器，将SPGReasoner下发的执行计划转换成底层LPG可执行的二进制文件
- Task Driver, 将RDG Compiler转换成的二进制文件提交到LPG Engine执行，该模块需要和具体引擎接口对接



编号	算子名	算子作用
1	patternMatch	给定一个子图模式和起点实例，获得该起点的子图实例（RDG），当不满足时返回null
2	expandInto	从已获取的子图实例（RDG）中，挑选一个实体实例作为起点实例
3	filter	给定一个判断表达式，判断当前获取的子图实例是否满足，若不满足，则该子图实例丢弃
4	groupByAndAgg	聚合计算操作，针对已有子图实例，进行聚合统计等行为计算，例如统计某个点的邻居数目
5	orderByAndLimit	排序并截断操作
6	limit	不排序，截断操作
7	shuffleAndFilter	根据不同不同点类型视角拆分和聚合子图数据
8	addFields	Rule中计算的临时变量存储
9	dropFields	移除无用的临时变量
10	join	将不同的子图实例（RDG）合并在一起
11	linkedExpand	调用三方服务做链指
12	ddl	实现对点、边的定义
13	select	输出指定的子图结果
14	cache	缓存当前RDG，为内部计算时使用算子

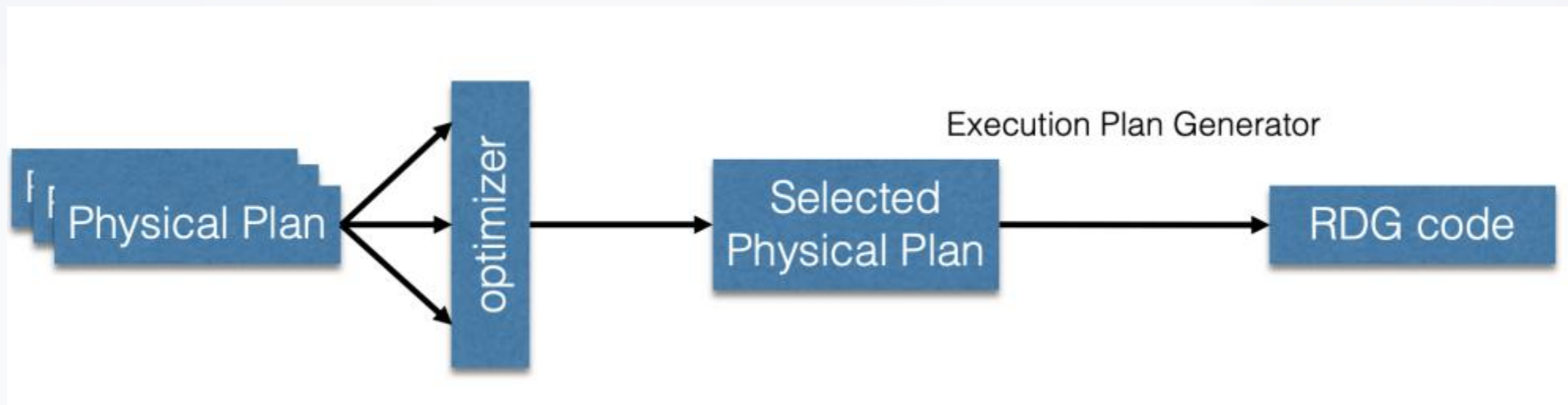


判断用户是否为一个多设备用户

```
Define (s:User)-[p:belongTo]->(o:UserClass/ManyDeviceUser) {  
  GraphStructure {  
    (s)-[t:has]->(u:Device)  
  }  
  Rule {  
    has_device_num("持有设备数目") = group(s).count(u.id)  
    R1("持有设备超过 100 个"): has_device_num > 100  
    R2("年龄大于 18 岁"): s.age > 18  
  }  
}
```

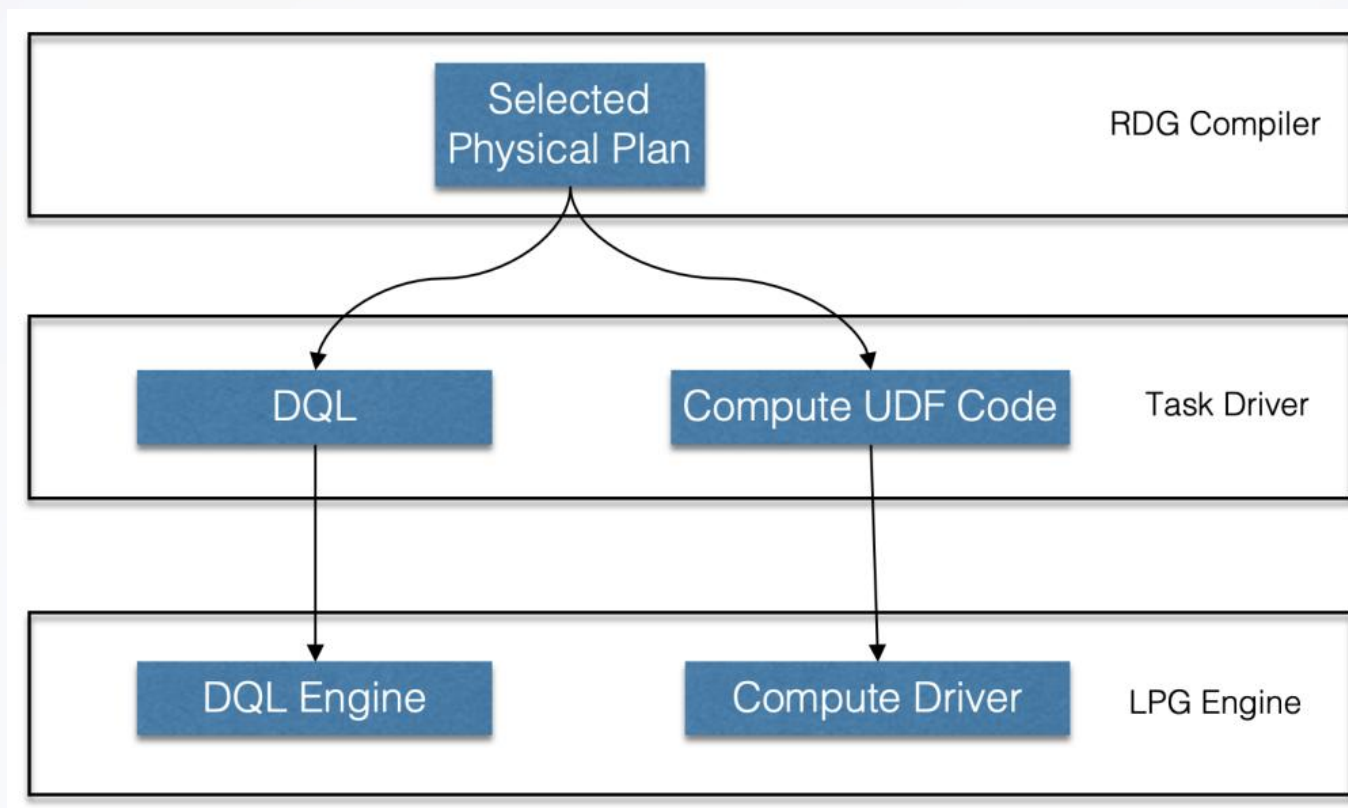


```
└─DDL(ddlOp=Set(AddPredicate(PredicateElement(belongTo,p,(s:User),EntityElement(ManyDeviceUser,UserClass)))  
  └─Filter(rule=LogicRule(R2,"年龄大于 18 岁",BinaryOpExpr(name=BGreaterThan)))  
    └─Filter(rule=LogicRule(R1,"持有设备超过 100 个",BinaryOpExpr(name=BGreaterThan)))  
      └─GroupByAndAgg(group=Set(NodeVar(s,null))  
        └─PatternMatch(pattern=PartialGraphPattern(s,Map(s -> (s:User), u -> (u:Device)),Map(s -> Set((s)->[t:has]-
```



RDG 算子表达的是对一个RDG 的原子操作，将RDG 算子树转换成底层引擎可执行代码，需要配合执行计划树经过Execution Plan Generator生成。





根据第三方属性图系统（LPG）的接口类型不同，分为如下两种不同场景。

- 场景1：LPG提供类似Cypher等DQL查询语言
- 场景2：LPG提供类似Spark等计算编程框架，可通过嵌入用户自定义算子实现计算

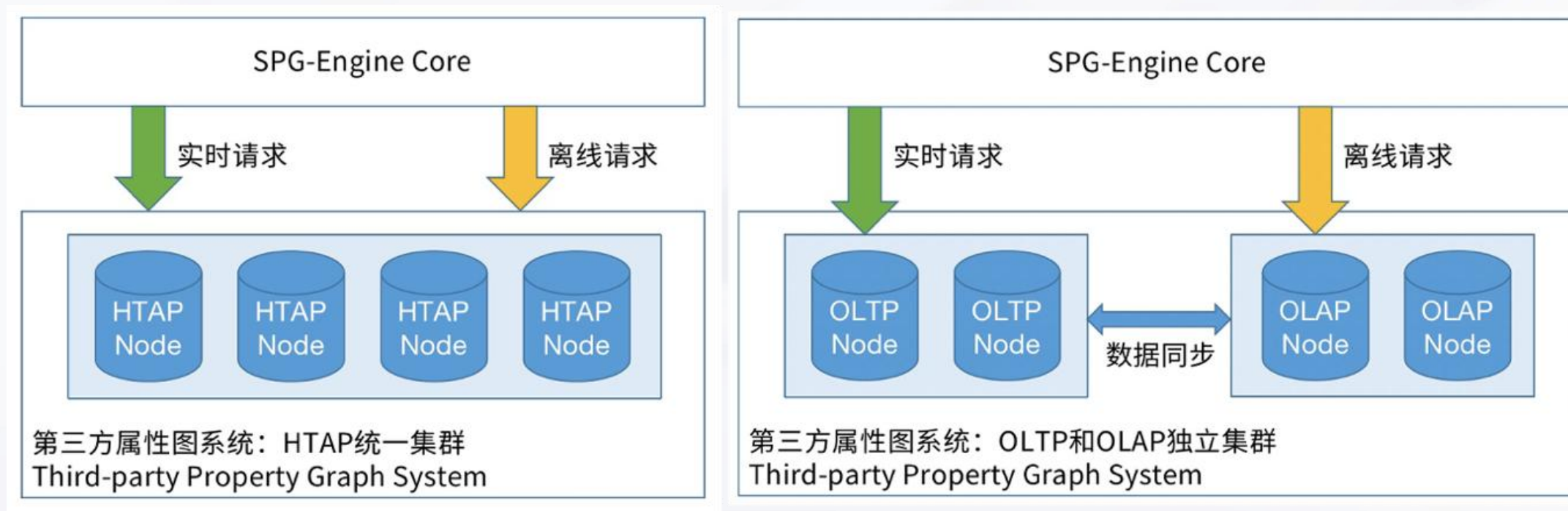




与第三方属性图系统进行对接



第三方属性图系统是一个独立的服务进程，应支持分布式部署，具备独立的集群安装、部署、管理、监控、运维方式，以提供基于web的ui界面为佳。该图系统通过一组适配接口和SPGController所在进程进行交互，这组适配接口即为SPG-Engine LPG Adapter。





SPG对第三方属性图系统的基本要求



能力模型	子能力	详细描述
图模型	支持属性图模型	提供点 (Vertex)、边 (Edge)、属性 (Property) 的存储和查询, 点、边具备类型 (Type) 或者标签 (Label)
	支持边特性	有向边和无向边, 支持两点间存在同类型的多条边
	支持属性类型	支持基本属性类型定义: 整型、浮点型、字符串等基本类型 支持复杂属性类型定义: 高精度数值 (BigDecimal)、时间戳 (DateTime)、地理坐标 (GeographicPoint) 等复杂属性类型 支持容器属性类型定义: 列表 (List)、集合 (Set)、映射 (Map)、属性组 (PropertyGroup) 等
	支持图隔离要求	在同一个图服务中支持多个图的存储、查询、计算
图数据导入	数据导入方式	支持全量数据、增量数据批量导入; 支持流式导入
	数据来源支持	支持从现有关系库中, 通过JDBC连接进行导入, 可自定义用于导入的SELECT语句; 支持常规文件数据源, 如CSV格式文件、JSON格式文件 支持常见大数据存储文件格式, 如HDFS上的Parquet格式文件
	映射方式	支持从单一原始表映射多个实体和关系的能力
	导入约束	进行批量导入时无须停止图引擎服务
属性索引	支持常见类型索引	对常见属性类型 (整型、浮点型、字符串型、时间戳型等) 进行索引
	支持组合索引	具备对多个属性字段进行组合索引
	支持模糊索引	对字符串型属性创建和使用全文索引
事务要求	支持事务的ACID特性	提供不低于读已提交 (Read Committed) 的隔离级别, 应用层通过显式加锁的方式保证执行顺序的可序列化 (Serializable) 隔离级别
图查询语言	具备DQL能力	支持标准化查询语言GQL或者支持提供DQL查询能力
部署模式	支持灵活部署模式	对于较小的数据量, 支持主备模式的部署方式; 对于较大的数据量, 支持分布式分片 (Sharding) 的部署方式



SPG对第三方属性图系统的进阶要求



创邻科技
CREATE LINK



GALAXYBASE

- 支持触发器
- 支持用户自定义函数/过程/算法
- 支持时序数据的存储和查询
- 具备多层次图谱的能力
- 支持跨图的实体/关系类型映射和转换
-





- 推进SPG在实际应用场景中落地
- 接入更多第三方属性图系统，赋能实时推理
- SPG能力评测方案
- 与LLM结合的Engine层



关注公众号了解更多



语义增强可编程图谱框架

洞察关联数据，创造无限可能