# GalaxyBase 图构建数据导入文档

# 1. 文档说明

建议选手首先了解 GalaxyBase 前端可视化建图流程，了解通过 GalaxyBase 平台创建图谱的过程，对 GalaxyBase 建图有初步认知。具体建图搭建流程详情可参见《Galaxybase 开发者版本文档》中“Galaxybase Studio 使用说明”全部章节。

相关内容文档链接：

https://www.galaxybase.com/document?file=dev&docid=6

需要注意的是，本次竞赛规定选手必须通过代码实现数据到图谱建设这一自动化过程。所以本文档主要描述了如何通过建图工具进行图构建的方法，引导选手实现自动化建图意图。

# 2. 自动化建图流程简要介绍

自动化建图旨在通过代码完成创建图模型并导入数据等操作，需要选手通过代码依次生成图模型定义文件（schema.json）和图数据映射文件（mapping.json）并程序执行图数据导入工具（galaxybase-console-buildgraph.jar）完成建模及数据导图。

以下逐一说明“schema.json 样例与讲解”、“mapping.json 样例与讲解”、“Galaxybase-console-buildgraph.jar 图构建样例与讲解”

## 2.1 schema.json 样例与讲解

如同关系型数据库需要先定义表名和表头一样，图数据库也需要先定义图的 schema。

schema.json 是 Galaxybase 用来描述图模型信息的文件。图模型信息包括图名称、点类型、边类型、点和边的属性以及属性值的数据类型。

下图是一个描述社交网络中不同人之间相互认识关系的图模型。*Person（人）点类型具有 id、name、sex、age、city、country 六个属性，其中 id 属性为数据唯一标识（一个点类型下不同点的唯一标识，不可重复）；Know（知道）边类型具有 score 一个属性。*


图模型对应的 schema.json 样例如下：

```
{
  "version": 0,//图的版本信息
  "graphName": "SocialNetwork",//图的名称，全局唯一。
  "vertexes": {
    "Person": {
      "type": "Person",//点类型名
      "primaryId": "id",//作为数据唯一标识的属性名
      "property": [//点的属性集合
        {
          "name": "name",//属性名
          "type": "STRING"//属性值的数据类型
        },
        {
          "name": "sex",
          "type": "STRING"
        },
        {
          "name": "age",
          "type": "INT"
        },
        {
          "name": "city",
          "type": "STRING"
        },
        {
          "name": "country",
          "type": "STRING"
        }
      ]
    }
  },
  "edges": {
    "Know": {
      "type": "Know",//边类型名
```

```
        "fromType": "Person",//起始点的类型名
        "toType": "Person",//终止点的类型名
        "directed": true,//是否有向，boolean 值
        "allowRepeat": true,//边的去重性
        "property": [//边的属性集合
          {
            "name": "score",
            "type": "DOUBLE"
          }
        ]
      }
    }
}
```

schema.json 分为 version、graphName、vertexes、edges 四部分，分别定义图版本、图名称、点类型、边类型。点边类型的属性（property）在点类型部分、边类型部分分别定义。

## 2.2 mapping.json 样例与讲解

mapping.json 是 Galaxybase 用来描述图模型与数据源的映射关系的文件，可以将数据源文件中的数据列和图模型中的点类型、边类型、点和边的属性建立映射。数据源文件必须放在 docker 的 data 目录的子目录下（允许 data/scoailNetwork/person.csv，不允许 data/person.csv），data 目录可以挂载至宿主机上。

以下是 socialNetwork 数据集的片段，它包括 person.csv 和 know.csv 两个文件，都具有表头，列分隔符为逗号。

- person.csv 片段。该文件记录的是点信息，如表头所示，第 0 列是数据唯一标识，第 1 列是姓名，第 2 列是性别，第 3 列是年龄，第 4 列是国籍。

| Id | 姓名 | 性别 | 年龄 | 国籍 |
|---|---|---|---|---|
| 1 | 1 | 女 | 16 | 美国 |
| 2 | 2 | 男 | 17 | 法国 |
| 3 | 3 | 女 | 18 | 英国 |
| 4 | 4 | 男 | 19 | 俄罗斯 |
| 5 | 5 | 女 | 20 | 加拿大 |
| 6 | 6 | 男 | 21 | 德国 |
| 7 | 7 | 女 | 22 | 日本 |

| | | | | |
|---|---|---|---|---|
| 8 | 8 | 男 | 23 | 意大利 |
| 9 | 9 | 女 | 24 | 中国 |

- know.csv 片段。该文件记录的是边信息，第 0 列是起始点数据唯一标识，第 1 列是终止点数据唯一标识，第 2 列是边属性值。

| from | to | 关注度 |
|---|---|---|
| 1 | 2 | 3 |
| 1 | 3 | 4 |
| 1 | 4 | 5 |
| 1 | 5 | 6 |
| 1 | 6 | 7 |
| 1 | 7 | 8 |
| 1 | 8 | 9 |
| 1 | 9 | 10 |
| 1 | 10 | 11 |

我们期望将 socialnetwork 数据集中 person.csv 文件映射到图模型中的 Person（ id、name、sex、age、city、country）点上，并为 Person 点的 id、name、sex、age、country 这几个属性赋值。同时我们希望将 know.csv 文件映射到 Know（score）边上，并为 Know 边的 score 属性赋值。

以下是一个建立上述数据集映射的 mapping.json 样例，各字段的含义如下：

```
{
  "content": [
    {
      "source": {
        "sourceType": "CSV",//文件类型
        "encoding": "utf-8",//编码格式
        "filePath": "data/socialNetwork/person.csv",//数据源文件在 Galaxybase docker 中的路径
        "delimiter": ",",//列分隔符
        "enclosingChar": "\"",//本例数据源文件中不包含封闭符，所以不发挥作用
        "hasHeader": true   //本例数据源文件第一行是表头
      },
      "vertexes": [
        {
          "type": "Person",//点的类型名
          "primaryId": "id",//主键属性名，创建图模型时定义
          "pkColumnIndex": 0,//本例数据源文件中，点的数据唯一标识位于第 0 列
          "hasDuplicatePrimaryKey": false,//本例数据源中，点的主键在不同行不会重复
          "property": [//点的属性信息列表
            {
```

```
                "columnIndex": 1,
                "alias": "name" //本例数据源中，第 1 列记录的是点类型 Person 的 name 属性
              },
              {
                "columnIndex": 2,
                "alias": "sex"
              },
              {
                "columnIndex": 3,
                "alias": "age"
              },
              {
                "columnIndex": 4,
                "alias": "country"
              }
            ]
          }
        ]
      },
      {
        "source": {
          "sourceType": "CSV",
          "encoding": "utf-8",
          "filePath": "data/socialNetwork/know.csv",
          "delimiter": ",",
          "enclosingChar": "\"",
          "hasHeader": true
        },
        "edges":[
          {
            "type": "Know",//边的类型名，创建图模型时定义
            "fromKeyColumnIndex": 0,//本例数据源文件中，起始点数据唯一标识位于第 0 列
            "toKeyColumnIndex": 1,//本例数据源文件中，终止点数据唯一标识位于第 1 列
            "property": [//边的属性信息列表
              {
                "columnIndex": 2,
                "alias": "score" //本例数据源文件中，第 2 列记录的是边类型 Know 的 score 属
性
              }
            ]
          }
        ]
      }
    ]
```

```
}
```

## 2.3 Galaxybase-console-buildgraph.jar 图构建样例与讲解

1、工具介绍

Galaxybase-console-buildgraph.jar 是创邻科技所提供的 GalaxyBase 建图 Java 架包，选手可以通过使用此架包达到自动化建图的意图。

Galaxybase-console-buildgraph.jar 会将 schema.json 、mapping.json 发送给 Galaxybase，配置启动参数后便可完成图构建。下面是各启动参数的含义。

```
java -jar Galaxybase-console-buildgraph.jar
        [-a|--address]
                图库主节点 ip 和端口，eg：127.0.0.1:18088。（必填字段)
                此处的端口是图库端口，默认值为 18088，如果部署时修改过该端口，此处也需相应修改。
        [-f|--portFrontend]
                数据可视化服务端口，默认为 8888，如果修改过该端口，则需设置该参数。（选填字段）
        [-u|--userName]
                选手名。（必填字段）
        [-p|--password]
                密码。（必填字段）
        [-e|--onlyGetSchema]
                填 true 表示只执行一项操作：读取当前的图模型。（选填字段）
        [-g|--graphName]
                图名称，注意要与 schema.json 中的图名字一致。（必填字段）
        [-s|--schema Path]
                schema.json 文件的存储路径。（选填字段）
                创建图模型和修改图模型时必须填写该字段，其他时候不能填写该字段。
        [-m|--mapping Path]
                mapping.json 文件的存储路径。（选填字段）
                如果不需要导入数据，可以不填该字段。
```

2、工具操作示例

创建图模型并导入数据。以下语句表示以选手名 admin、密码 password 连接位于 IP 127.0.0.1 的图服务，读取当前目录下的文件 schema.json 作为图模型发送给图服务，并根据图模型创建名称为 SocialNetwork 的图，然后读取当前目录下的文件 mapping.json 作为映射配置发送给图服务，命令图服务根据该映射配置往名称为 SocialNetwork 的图中加载数据。

```
java -jar Galaxybase-console-buildgraph.jar -a 127.0.0.1:18088 -u admin -p password -g SocialNetwork -s schema.json -m mapping.json
```

3、工具获取

工具地址：附件中 tools/ Galaxybase-console-buildgraph.jar