1. Datagen
   1. 简介

测试数据的生成是进行benchmark测试的第一步，LDBC为其制定的测试需求设计了一个数据生成器：ldbc\_snb\_datagen ([github仓库地址](https://github.com/ldbc/ldbc_snb_datagen))，本实验使用该数据生成器进行数据生成。

* 1. ldbc\_snb\_datagen
     1. 环境需求

1. **Apache Hadoop v2.6.0**

ldbc\_snb\_datagen (以下简称datagen)使用Apache Hadoop v2.6.0进行数据生成，实验机条件有限，我们按照以下的方式进行数据生成：



1. **Apache Maven**

ldbc\_snb项目主要使用Apache Maven工具进行项目管理和自动构建，在datagen的安装中，需要配置好maven工具以支持安装的进行。

1. **Python 2.7**

datagen通过python实现对生成数据属性的自定义设置，官方给出的版本需求是python 2.7。

* + 1. 参数配置

将datagen文件下载到本地后，根据datagen官方教程配置运行环境。参数配置主要是配置~/\*/ldbc\_snb\_datagen-master/run.sh文件中的路径：

* + - DEFAULT\_HADOOP\_HOME=/usr/local/java/hadoop-2.6n.0
    - $LDBC\_SNB\_DATAGEN\_HOME/test\_params.ini

第一条参数设置对应Hadoop的实际安装路径；

第二条参数设置对应datagen进行数据生成时参数文件位置。

运行命令./run.sh即可通过maven工具自动构建项目生成jar包。我们将利用这个jar包生成所需的测试数据集。

1. datagen参数

datagen通过参数文件的形式对生成数据的属性（图属性、数据规模、部分数据格式等）进行自定义设置，下面就实验中用到的几个重要的参数进行介绍。

* **ldbc.snb.datagen.serializer.personSerializer**
  1. 该参数主要用于设置社交网络中person节点和knows边数据序列化的模式；
  2. 本次实验设置

[*ldbc.snb.datagen.serializer.snb.interactive.CSVPersonSerializer*](https://github.com/ldbc/ldbc_snb_datagen/blob/master/src/main/java/ldbc/snb/datagen/serializer/snb/interactive/CSVPersonSerializer.java)

* **ldbc.snb.datagen.serializer.invariantSerializer**

1. 该参数主要用于设置社交网络中person节点和knows边数据序列化的模式；
2. 本次实验设置

[*ldbc.snb.datagen.serializer.snb.interactive.CSVInvariantSerializer*](https://github.com/ldbc/ldbc_snb_datagen/blob/master/src/main/java/ldbc/snb/datagen/serializer/snb/interactive/CSVInvariantSerializer.java)

* **ldbc.snb.datagen.serializer.personActivitySerializer**

1. 该参数主要用于设置社交网络中person节点和knows边数据序列化的模式；
2. 本次实验设置

*ldbc.snb.datagen.serializer.snb.interactive.CSVPersonActivitySerializer*

* **ldbc.snb.datagen.generator.scaleFactor**

1. 该参数主要用于设置社交网络数据的规模。本实验测试的数据用于snb.interactive测试，该条件下的数据规模从0.1GB~1000GB共有9个可用的参数选项，对于不同的数据规模设置，图属性大致如下：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Scale Factor | 1 | 3 | 10 | 30 | 100 | 300 | 1000 |
| # of Persons | 11K | 27K | 73K | 182K | 499K | 1.25M | 3.6M |
| # of Years | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| Start Year | 2010 | 2010 | 2010 | 2010 | 2010 | 2010 | 2010 |

1. 本次实验设置

*snb.interactive.0.1*

* **ldbc.snb.datagen.serializer.dateFormatter**

1. 本次实验基于Janus graph实例的基础上完成，该实例实现了importer工具完成图数据的导入，要求数据及文件中日期格式必须为Long型，因此在生成数据时要对日期格式进行设置
2. 本次实验设置

*ldbc.snb.datagen.serializer.formatter.LongDateFormatter*

通过上述的参数设置，本次实验将生成标准的大小为100MB左右的测试数据，基本上每个不同entity的数据内容分别存储在不同的csv文件中。

* 1. 数据

datagen的数据生成利用Hadoop的Mapreduce方法实现，最后生成的数据将构成一张完整的社交网络图。每次执行datagen将生成3类数据文件，主要分布在2个文件夹中。下面将对datagen产生的数据位置、格式和属性进行介绍。

* + 1. 数据分布

datagen生成的与benchmark测试相关的数据分布在2个文件夹中，一个命名为social\_network，一个命名为substitution\_parameters。

1. social\_network

该文件夹存储了datagen生成的社交网络图的数据，数据格式是csv。这些文件由mapreduce任务生成，因此被存储在hdfs上。本次实验需要将文件从虚拟机上传到服务器上进行测试，因此在完成数据生成后，需要将该文件夹通过get方法下载到虚拟机本地上便于之后的文件传输。

1. substitution\_parameters

该文件针对benchmark中设计的25个BI query和14个IC(interactive) query设置了参数，这些参数是在进行查询时必须提供的。参数文件的格式为txt文件。

* + 1. 数据属性

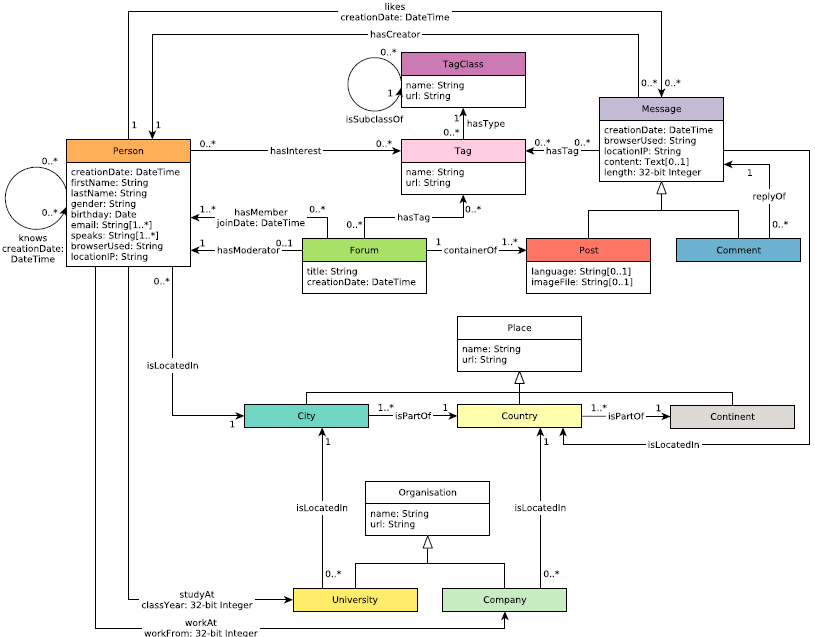
1. 文件属性

datagen将生成三类文件：

* 数据集：用于benchmark测试的主要数据集，大约占生成数据的90%；
* 更新流（update stream）：用于update query的数据，大约占生成数据的10%；
* 参数（substitution parameters）：用于BI query和IC query的参数。

1. data schema

datagen生成的所有文件构成完整的社交网络图。根据官方文件的说明，下图展示了这个图数据的schema：



* 1. 数据使用

本次实验的实验数据存储在~/janusgraph/test-data-100m/下，包括social\_network和substitution\_parameters文件夹，在进行数据测试时，对workload的参数进行如下设置：

* ldbc.snb.interactive.parameters\_dir=/home/user26/test-data-100m/substitution\_parameters
* ldbc.snb.interactive.updates\_dir=/home/user26/test-data-100m/social\_network