# LinkML to MySQL Schema Conversion

Robert Petryszak

12 March 2024

# Functionality delivered:

- Generation of **gk_central.sql**, including DDL and DataModel table insert statement
- Based on comparison of **previous** and **newly generated** gk_central.sql:
  - Generation of the following ‘new vs previous’ difference reports:
    - **Raw**: which tables/attributes were created and dropped. Doesn't make clear which attributes were moved or changed from single-valued to multi-valued or vice versa
    - **Processed**: makes explicit which tables/attributes are truly new/dropped, and lists precisely which attributes have been moved (e.g. from subclass to a superclass) and/or changed from single-valued to multi-valued or vice versa
  - Based on Processed difference report, generation of **gk_central.update.sql**, including:
    - DDL update and create statements
    - Data population statements reflecting updated DDL
    - DataModel table insert statement
    - DDL drop table statements

# Usage and Testing:

- generate.py mysql → gk_central.sql (from schema.yaml)
- Make changes in schema.yaml to reflect test cases → schema.yaml.changed
- generate.py mysql → gk_central.sql.changed (from schema.yaml.changed)
- compare_sql.py gk_central.sql gk_central.sql.changed <output_dir> generate_update_ddl →
  - gk_central.update.sql
  - Print out raw and processed difference reports (see: gk_central.diffs.txt)

# DDL template files used for SQL generation:

- gk_central.sql

```
/*!40101 SET @OLD_CHARACTER_SET_CLIENT=@@CHARACTER_SET_CLIENT */;
/*!40101 SET @OLD_CHARACTER_SET_RESULTS=@@CHARACTER_SET_RESULTS */;
/*!40101 SET @OLD_COLLATION_CONNECTION=@@COLLATION_CONNECTION */;
/*!40101 SET NAMES utf8mb4 */;
/*!40103 SET @OLD_TIME_ZONE=@@TIME_ZONE */;
/*!40103 SET TIME_ZONE='+00:00' */;
/*!40014 SET @OLD_UNIQUE_CHECKS=@@UNIQUE_CHECKS, UNIQUE_CHECKS=0 */;
/*!40014 SET @OLD_FOREIGN_KEY_CHECKS=@@FOREIGN_KEY_CHECKS, FOREIGN_KEY_CHECKS=0 */;
/*!40101 SET @OLD_SQL_MODE=@@SQL_MODE, SQL_MODE='NO_AUTO_VALUE_ON_ZERO' */;
/*!40111 SET @OLD_SQL_NOTES=@@SQL_NOTES, SQL_NOTES=0 */;

@SCHEMA_CONTENT@

/*!40101 SET character_set_client = @saved_cs_client */;
/*!40103 SET TIME_ZONE=@OLD_TIME_ZONE */;
/*!40101 SET SQL_MODE=@OLD_SQL_MODE */;
/*!40014 SET FOREIGN_KEY_CHECKS=@OLD_FOREIGN_KEY_CHECKS */;
/*!40014 SET UNIQUE_CHECKS=@OLD_UNIQUE_CHECKS */;
/*!40101 SET CHARACTER_SET_CLIENT=@OLD_CHARACTER_SET_CLIENT */;
/*!40101 SET CHARACTER_SET_RESULTS=@OLD_CHARACTER_SET_RESULTS */;
/*!40101 SET COLLATION_CONNECTION=@OLD_COLLATION_CONNECTION */;
/*!40111 SET SQL_NOTES=@OLD_SQL_NOTES */;
```

# DDL template files used for SQL generation:

- gk_table.sql

```
--
-- Table structure for table `@TABLE_NAME@`
--
DROP TABLE IF EXISTS `@TABLE_NAME@`;
/*!40101 SET @saved_cs_client     = @@character_set_client */;
/*!40101 SET character_set_client = utf8 */;
CREATE TABLE `@TABLE_NAME@` (
@TABLE_CONTENT@
) ENGINE=InnoDB@AUTO_INCREMENT@ DEFAULT CHARSET=utf8 COLLATE=utf8_unicode_ci;
/*!40101 SET character_set_client = @saved_cs_client */;
```

# DDL template files used for SQL generation:

- gk_class2non_instance_attr_table.sql

```
DROP TABLE IF EXISTS `@CLAZZ@_2_@ATTRIBUTE@`;
/*!40101 SET @saved_cs_client     = @@character_set_client */;
/*!40101 SET character_set_client = utf8 */;
CREATE TABLE `@CLAZZ@_2_@ATTRIBUTE@` (
  `DB_ID` int(10) unsigned DEFAULT NULL,
  `@ATTRIBUTE@_rank` int(10) unsigned DEFAULT NULL,
  `@ATTRIBUTE@` @MYSQL_TYPE@,
  KEY `DB_ID` (`DB_ID`),
  KEY `@ATTRIBUTE@` (`@ATTRIBUTE@`@DISPLAY_WIDTH@)
) ENGINE=InnoDB DEFAULT CHARSET=utf8 COLLATE=utf8_unicode_ci;
/*!40101 SET character_set_client = @saved_cs_client */;
```

# DDL template files used for SQL generation:

- gk_class2instance_attr_table.sql

```
DROP TABLE IF EXISTS `@CLAZZ@_2_@ATTRIBUTE@`;
/*!40101 SET @saved_cs_client     = @@character_set_client */;
/*!40101 SET character_set_client = utf8 */;
CREATE TABLE `@CLAZZ@_2_@ATTRIBUTE@` (
  `DB_ID` int(10) unsigned DEFAULT NULL,
  `@ATTRIBUTE@_rank` int(10) unsigned DEFAULT NULL,
  `@ATTRIBUTE@` int(10) unsigned DEFAULT NULL,
  `@ATTRIBUTE@_class` varchar(64) COLLATE utf8_unicode_ci DEFAULT NULL,
  KEY `DB_ID` (`DB_ID`),
  KEY `@ATTRIBUTE@` (`@ATTRIBUTE@`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8 COLLATE=utf8_unicode_ci;
/*!40101 SET character_set_client = @saved_cs_client */;
```

# SQL generation-specific schema changes needed:

- Various attributes/slots added to effect various curation-specific overrides (e.g. constraints), but had to be excluded from both DDL (via config in generate.py) and from graph-core classes (i.e. removed explicitly in schema.web.diff.yaml)

- Mysql-specific types introduced:
  - mysql_signed_int_type (most integer values int(10) unsigned in mysql, but some just int(10))
  - mysql_inverse_slot - needed for DataModel content only

- Lines added or removed:
  - schema.yaml, schema.web.diff.yaml → 325
  - generate.py → 595
  - compare_sql.py → 462

# Assumptions and notes:

- Assumption: it's the user's responsibility to make sure that an attribute that is moved from a **superclass** down the hierarchy in linkml **is always moved to all subclasses** of that superclass. When generating update DDL, compare_sql.py just follows the exact changes made in schema.yaml and doesn't try to second-guess the user's intentions.

- Note that the SQL-specific changes to schema.yaml have not been reflected in graph-core-curator java classes and hence those classes should not be used.