DQ-MeeRKat – Getting Started

# External Prerequisites

To facilitate an easy start with DQ-MeeRKat and no version conflicts, we provide the files for the three required external tools within the GitHub project.

1. **InfluxDB** (required)

We used InfluxDB to store the CDQM time series data. Since the Java API was not working by the time of the creation, we used the InfluxDB browser version. DQ-MeeRKat does not offer an embedded mode (like Derby or GraphDB), but runs InfluxDB outside the Java runtime to persist CDQM results over time.

Run on Linux:

startInflux.sh (for starting InfluxDB server)

startInfluxConsole.sh (for querying InfluxDB using the console)

Run on Windows:

InfluxDB\influxdb-1.8.5-1\_windows\influxd.exe

1. **Grafana** (required)

Grafana is a browser-based dashboard for visualization. After the start, a browser window is opened with the dashboard URL (http://localhost:3000 by default).

Run on Linux:

startGrafana.sh

Run on Windows:

Grafana\grafana-7.5.5\_windows\bin\grafana-server.exe

In order to simplify the starting of the two required external tools, we provide a batch file startWinEnv.bat for Windows 10 and a shell script startLinuxEnv.sh for Linux Mint (Tina).

1. **GraphDB** (optional since embedded version can be used too; this is only for managing the repositories with a GUI and visualize them)

Run on Windows:

GraphDB\GraphDB\_Free-9.0.0.exe

OpenGraphDB Server and Workbench in browser:

<http://localhost:7200/>

Quick start guide for further usage:

<http://graphdb.ontotext.com/free/quick-start-guide.html>

Open Demo Dashboard:

Setup - Create new repository:

* Repository ID: DQM
* Repository title: DQ-MeeRKat-repo
* Storage folder: <local-path-to-repo>\kg-repo\repositories\kg-repo\storage
* Leave default settings for remaining options

Restart GraphDB and explore graph via Explore - Graphs overview.

# Use DSD Connectors to Connect to Data Sources

Currently, there are implementations for

* CSV files
* Cassandra DBs
* DSD files
* MySQL DBs
* Oracle DBs (still need to be verified and comprehensively tested)

To each of those data sources, it is possible to create a connection using a DSConnector(only accessing the schema, e.g., sufficient for DSD files), or DSInstanceConnector(for accessing schema + instances of a data source).

ConnectorCSVconn1 = new ConnectorCSV("path/DataCoSupplyChainDataset.csv", ",",

"\n", "SupplyChain");

ConnectorMySQLconn2 = ConnectorMySQL.getInstance("jdbc:mysql://localhost:port",

"dbname", "user", "pw");

In order to transform the schema description of the local data source to virtual DSD elements in the Java runtime environment, it is necessary to load the schema information.

DatasourcedsSC= conn1.loadSchema();

DatasourcedsCE= conn2.loadSchema();

# Creation of a Knowledge Graph to Manage all Datasources

To conveniently handle multiple DSD data source objects, we developed the class DSDKnowledgeGraph, which holds all data sources and their connections.

DSDKnowledgeGraph kg = new DSDKnowledgeGraph("automotive");

kg.addDatasourceAndConnector(dsSC, conn1);

kg.addDatasourceAndConnector(dsCE, conn2);

The KG class holds an instance of an embedded GraphDB, which persists the triples for later reuse and offers methods for export the KG or to annotate all contained data sources (with their concepts and attributes) with RDPs.

# Creation of a Reference Data Profile

To create and annotate a reference data profile to a DSD element, a corresponding method .annotateProfile(RecordSet)is provided, which needs to be called for each DSD element that should be annotated. In the following example, each attribute of each concept in the Northwind (NW) is annotated with a data profile that uses the first 5,000 records of the respective concept (table).

for (Concept c : ds.getConceptsAndAssociations()) {

RecordListrs = conn.getPartialRecordSet(c, 0, 5000);

for (Attribute a : c.getAttributes()) {

a.annotateProfile(rs);

}

}

Instead of annotating the RDPs for single DSD elements, it is possible to traverse all elements in the KG and to initiate the creation of an RDP for each automatically.

kg.addDataProfile(1000);

The parameter denotes the number of records used to calculate the RDP. Here, we used the first 1000 records.

# Continuously Monitor the Data Quality

For this step, InfluxDB and Grafana needs to be running (see Section 1).

Open Grafana in your browser: <http://localhost:3000/>

To execute the two demos presented in our paper, the following steps need to be taken:

Grafana

1. Menu 🡪 Configuration 🡪 Data Sources 🡪 Add data source 🡪InfluxDB
   1. URL: <http://localhost:8086>
   2. With Credentials (use your local credentials)
   3. Database name is testSeries
2. Menu 🡪Create🡪 Import 🡪 Upload .json File 🡪 Select …
   1. Grafana\Dashboard-Exports\Supplychain-periodic-dashboard.json
   2. Grafana\Dashboard-Exports\Test-streaming-data-dashboard.json
3. Execute the demo from Eclipse:
   1. src\main\java\dqm\jku\trustkg\demos\repeatability\ DemoPeriodicData.java
   2. src\main\java\dqm\jku\trustkg\demos\repeatability\ DemoStreamingData.java

Now, the data is stored in the InfluxDB and you can visualize and observe it with Grafana.