Enhancing Statistical Discovery with Oracle RDF on Oracle Cloud

Shoki Nishimura, National Statistics Center of Japan
Yusuke Takeyoshi, Senior Principal Consultant, Oracle Japan
Agenda

✓ Background
  • Introduction of e-Stat System
  • Why We Developed LOD
  • How We Configured e-Stat LOD
  • Integrating GeoSpatial RDF Data in e-Stat LOD
  • Sample Application (Demo)

✓ Technical Details of e-Stat LOD
  • LOD System Architecture
  • Size and Scale of e-Stat LOD
  • SPARQL Performance Concerns
  • Database Design to Improve Performance
Agenda

✓ Background
  • Introduction of e-Stat System
  • Why We Developed LOD
  • How We Configured e-Stat LOD
  • Integrating GeoSpatial RDF Data in e-Stat LOD
  • Sample Application (Demo)

✓ Technical Details of e-Stat LOD
  • LOD System Architecture
  • Size and Scale of e-Stat LOD
  • SPARQL Performance Concerns
  • Database Design to Improve Performance
Introduction of e-Stat System - Portal Site for Official Statistics of Japan

- In 2008, e-Stat started to publish statistical data of government agencies (Format: Excel)
- In 2014, API service started (Format: XML, JSON, CSV)
- In 2016, LOD (Linked Open Data) service started (Format: RDF)

File download

- 2008
  - Excel
  - 610 statistics
  - 1.4 million tables
- 2014
  - API
  - XML
  - 230 statistics
  - 150 thousand datasets
- 2016
  - LOD
  - JSON
  - CSV
  - 9 statistics
  - 87 datasets
  - 2.1 billion triples

e-Stat (https://www.e-stat.go.jp/)
What is RDF and LOD?

RDF (Resource Description Framework)

- RDF is a standard model for data interchange on the web.
- RDF uses URIs to name the relationship as well as the two ends of the link (this is usually referred to as a Triple).

Structure of Triple

Subject ——— Predicate ——— Object

Structure of RDF Graph

LOD (Linked Open Data)

- LOD is structured open data interlinked with other data.
- LOD builds on RDF technologies.

TIM BERNERS-LEE'S FIVE-STAR SCHEME

- OL: open licence
- RE: machine readable
- OF: open format
- URI: uniform resource identifier
- LD: linked data
Why We Developed LOD
From “Link to File” To “Link to Data”

Clarify semantics and origins for data

**Link to File**
- Assign URI to page (http://www.e-stat.go.jp/pages.html)
- Assign URI to file (http://www.e-stat.go.jp/xls0001.xls)

**Link to Data**

<table>
<thead>
<tr>
<th>Standard area code</th>
<th>Sex Age</th>
<th>Total (Sex)</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>44 years [Person]</td>
<td>45 years [Person]</td>
<td>44 years [Person]</td>
</tr>
<tr>
<td>Saitama-city</td>
<td>-</td>
<td>16,130</td>
<td>19,245</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>8,293</td>
<td>9,938</td>
</tr>
<tr>
<td>Kawaguchi-city</td>
<td>-</td>
<td>6,582</td>
<td>8,022</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>3,526</td>
<td>4,289</td>
</tr>
</tbody>
</table>

Assign URI to each data
(http://data.e-stat.go.jp/lof/.../obs00001)
(http://data.e-stat.go.jp/lof/.../C11201)
Why We Developed LOD

Metadata for statistical data in Japan is not standardized, which makes it hard to process data.

Define standardized metadata as **RDF** to make it **machine-readable**

**Conventional statistical metadata**

- Prefecture
- Municipality
- Age (5 years)
- Municipality-A
- Age-2010
- Age (10 years)

**Statistical DB**

**Standardized RDF metadata**

- Region 2010
- Prefecture
- Region
- Municipality
- Municipality-A
- Age (5 years)
- Age (15 years)
- Age
- Age (10 years)
- Age-2010
How We Configured e-Stat LOD

Example: The population of 44-year-old men in Kawaguchi City in 2010

e.g. Population of 2010 Population census

<table>
<thead>
<tr>
<th>Standard area code</th>
<th>Total (Sex)</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>44 years</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Unit Of Person]</td>
<td>44 years</td>
<td>45 years</td>
</tr>
<tr>
<td></td>
<td>16,130</td>
<td>8,293</td>
<td>9,938</td>
</tr>
<tr>
<td></td>
<td>19,245</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saitama-city</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kawaguchi-city</td>
<td>6,582</td>
<td>3,526</td>
<td>4,289</td>
</tr>
<tr>
<td></td>
<td>8,022</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Original Excel format

Converted RDF format (R2RML)

convert each observation values (cells) into RDF data

Analytics and Data Summit 2020
Integrating GeoSpatial RDF Data in e-Stat LOD

Major statistics such as Population Census are integrated with more fine-grained location data, such as 250m x 250m square grids, which are defined as Polygons.

- **Cell_id**
- **GRID_code**
- **Population in target area**
- **Value of Population**
- **String of World Grid Code**
- **geo:hasGeometry**
- **geo:asWKT**
- **Polygon**

- Target Location Info
- Population in target area
- Value of Population
- String of World Grid Code
- geo:hasGeometry
- geo:asWKT
- Polygon

✓ Statistical data can be obtained with GeoSPARQL.

**Example of GeoSPARQL FILTER expression**

```
FILTER ( ogc:sfWithin( ?wkt , "POLYGON(...)"^^geo:wktLiteral )
```

Query population info WITHIN a search polygon
Sample Application


Linked Open Data

LOD sample applications

Search LOD on map

Select area on map

Search LOD on map data

Easy LOD search

Analytics and Data Summit 2020
Sample Application

Click on the map to display a list of applicable region names on the right. 'Grid code is using the JIS X 0208:2016 system.'

Region name

GeoSPARQL

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX ogcfl: <http://www.opengis.net/def/function/geosparql/>
PREFIX dcterms: <http://purl.org/dc/terms/>

SELECT DISTINCT ?area ?areaName ?areaType ?areaid ?areaWKT
WHERE {
  ?area geo:hasGeometry / geo:asWKT ?areaWKT;
  a ?areaType;
  dcterms:identifier ?areaid;
  rdfs:label ?areaName.
FILTER (ogcfl:sfContains(?areaWKT, 'Point(139.78247469054514 35.73443005624056)^^geo:wtLiteral'))
} LIMIT 10
Sample Application

select dataset

- Population and Households
  - Population
    - Population by sex - 5th level grid squares
    - Population by age and sex - 5th level grid squares
  --households
    - Households by level of household - 5th level grid squares
    - Private households by size of household - 5th level grid squares
    - Private households by family type of household - 5th level grid squares
  - Private households by division of household member's age - 5th level grid squares
  - Private households by presence of household member under 15 years of age - 5th level grid squares
  - Private households by presence of aged household members - 5th level grid squares

select measure and dimension

- Dataset: Population by sex - 5th level grid squares
- Measure: Population
- Dimension:
  - All (Sex)
  - Male
  - Female

View observation value

- Search result
  - Sex: All (Sex)
  - Time period: 2015
  - Area: 5th grid: 205339468221
  - Population (Unit of person): 2178

- Sex: Male
  - Time period: 2015
  - Area: 5th grid: 205339468221
  - Population (Unit of person): 1046

- Sex: Female
  - Time period: 2015
  - Area: 5th grid: 205339468221
  - Population (Unit of person): 1132

SPARQL

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX qs: <http://purl.org/linked-data/cube#>
PREFIX qb: <http://purl.org/linked-data/cube#>
PREFIX dcterms: <http://purl.org/dc/terms/>

SELECT DISTINCT ?selectedMeasure1 ?selectedMeasureUnit1
WHERE { dbp:population ?selectedMeasure1 ?selectedMeasureUnit1

Download
Agenda

✔ Background
  • Introduction of e-Stat System
  • Why We Developed LOD
  • How We Configured e-Stat LOD
  • Integrating GeoSpatial RDF Data in e-Stat LOD
  • Sample Application (Demo)

✔ Technical Details of e-Stat LOD
  • LOD System Architecture
  • Size and Scale of e-Stat LOD
  • SPARQL Performance Concerns
  • Database Design to Improve Performance
SPARQL Performance Concerns

- In 2018, triples increased to 1,300 million (including GeoSpatial data).
  - At that time, the e-Stat LOD was running on on-premise Exadata 12cR2.

- SPARQL performance became no longer acceptable…
  - Tested 138 different SPARQL queries (including several GeoSPARQL)

<table>
<thead>
<tr>
<th></th>
<th>Exadata 12cR2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. SPARQL response time</td>
<td>65.74 secs</td>
</tr>
<tr>
<td>Number of queries running over 10 sec.</td>
<td>108</td>
</tr>
<tr>
<td>Number of queries running over 300 sec.</td>
<td>13</td>
</tr>
</tbody>
</table>

➤ We needed a drastic measures to improve SPARQL performance!
Database Design to Improve Performance

The following tuning drastically improved query performance.

1. Using Database In-Memory features
2. Partitioning RDF table by triple "predicate"
3. Optimizing Optimizer Statistics based on actual SPARQLs

*Tuning 1, 2 are only available from Oracle 18c.*
Database Design to Improve SPARQL Performance

1. Using Database In-Memory Features

DB In-Memory (DBIM) improved SPARQL performance by 10 to 60 times.

- Enabling DBIM for RDF is very simple.
  
  ```java
  exec SEM_APIs.ENABLE_INMEMORY(TRUE);
  ```

- To reduce the amount of data accessed:
  - The populated data is automatically compressed in memory
  - In-Memory Indexes automatically prunes the data accessed

**SPARQL**

```sparql
SELECT ?s ?o
WHERE {
  ?s rdfs:label ?o
} LIMIT 10
```
In many cases, graph patterns in SPARQL queries specify "predicate" URI and query "object" values.

```
select ?year ?population
where {
  ?s estat-measure:population ?population;
  sdmx-dimension:refArea / rdfs:label 'Kyoto-shi'@en ;
  cd-dimension:timePeriod ?year ;
  cd-dimension:sex cd-code:sex-all ;
  cd-dimension:nationality cd-code:nationality-japan ;
  g00200521-dimension-2010:area g00200521-code-2010:area-all ;
  cd-dimension:age cd-code:age-all .
}
```
To reduce the amount of data accessed, we partitioned the RDF_LINK$ table using hash values of the RDF predicate IDs.

**SPARQL**

```sparql
SELECT ?s ?o
WHERE {
    ?s rdfs:label ?o
} LIMIT 10
```

**TABLE:**
RDF_LINK$ (partitioned by predicate ID hash)

**EXAMPLE:**
Each partition stores triples with the following predicates:
- rdfs:subClassOf
- estat-measure:population
- rdfs:label
- cd-dimension:age
- rdf:type

Access only a partition containing “rdfs:label” predicate
3. Optimizing Optimizer Statistics

In Oracle Database, RDF triples are stored in relational tables (RDF_LINK$, RDF_VALUE$, ...), so SPARQLs are translated and executed as semantically the same SQLs.

- Optimizer Statistics are very important to generate optimal execution plans.

Examples of Optimizer Statistics

- **RDF_LINK$**
  - subject, predicate, object IDs
  - Number of rows
  - Avg. row length

- **RDF_VALUE$**
  - value, type, ID to value mapping
  - Distinct num of values in each column

Optimizer makes optimal execution plans considering statistics

- Full Table Scan or Index Scan?
- JOIN order
- JOIN method (LOOP or HASH?)
  
  e.t.c.
3. Optimizing Optimizer Statistics

For the optimizer to make a good execution plan against complex SPARQL queries, we gathered **column group statistics**, which enables optimizer to consider a correlation relationship between different columns. Which column group statistics are useful was determined using **SPARQLs actually executed so far in the e-Stat LOD**.

**STEP1:** Tell the database to monitor column group usage for the specified seconds.
```sql
exec DBMS_STATS.SEED_COL_USAGE(NULL, NULL, 600);
```

**STEP2:** Execute as many SPARQLs as possible, which are executed so far, within the specified time

**STEP3:** Mark the useful column groups detected during the monitoring
```sql
SELECT DBMS_STATS.CREATE_EXTENDED_STATS('MDSYS', 'RDF_LINKS') FROM DUAL;
```

**STEP4:** Gather statistics by a pre-built procedure **SEM_PERF.GATHER_STATS**.
The marked column group statistics are automatically gathered.
```sql
exec SEM_PERF.GATHER_STATS(...);
```
Database Design to Improve SPARQL Performance

How Much SPARQL Performance Improved?

<table>
<thead>
<tr>
<th>Tested 138 different SPARQL queries against 1.3 billion triples (including several GeoSPARQL)</th>
<th>On-prem Exadata Oracle 12cR2 (former platform)</th>
<th>Oracle Gen2 Cloud Oracle 18c with tunings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. SPARQL response time</td>
<td>65.74 sec.</td>
<td>1.27 sec.</td>
</tr>
</tbody>
</table>

Former platform  →  Current platform

Avg. SPARQL response time

300M of triples  500M of triples  1,300M of triples  2,100M of triples

2015 - 2017 - 2018 - 2019 -

The performance remained constant, even after triples increased to 2.1 billion!
Summary / Key Takeaways

✓ Publication of the 1st statistical LOD in Japan
  - 9 major statistics are published as LOD with Oracle Cloud
  - RDF triples are generated by use of R2RML from relational tables
  - GeoSpatial triples are integrated and published as LOD

✓ Performance improvement for SPARQL queries
  - We achieved 50 times faster performance applying the following changes:
    - Migrating the entire LOD platform to Oracle Gen2 Cloud
    - Utilizing DBIM features
    - Partitioning a RDF table by triple predicate
    - Gathering column group statistics
Questions & Answers
ANALYTICS AND DATA SUMMIT 2020
All Analytics. All Data. No Nonsense.
February 25-27, 2020
How We Configured e-Stat LOD

Data in e-Stat LOD is defined using RDF Data Cube Vocabulary (W3C)

- The RDF Data Cube Vocabulary provides a way to publish multi-dimensional statistics in such a way that it can be linked to related data sets and concept, [https://www.w3.org/TR/vocab-data-cube/](https://www.w3.org/TR/vocab-data-cube/).
- Each observation, or data in each cell, is described by dimensions, measures, and attributes.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Measure</th>
<th>2010 Population Census</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total (Sex)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>44 years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45 years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
</tr>
<tr>
<td>44 years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45 years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attribute</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saitama-city</td>
<td>16,130</td>
<td>19,245</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8,293</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9,938</td>
</tr>
<tr>
<td>Kawaguchi-city</td>
<td>6,582</td>
<td>8,022</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3,526</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4,289</td>
</tr>
</tbody>
</table>

Analytics and Data Summit 2020
How We Configured e-Stat LOD

RDF data was generated from statistics tables in our database with R2RML (R2RML = RDB to RDF Mapping Language).

Statistics Tables (Relational tables)

Transform tables to RDF format (R2RML)

Example of R2RML mapping file

```xml
@prefix rr: <http://www.w3.org/ns/r2rml#>.
@prefix ex: <http://example.com/ns#>.

<#TriplesMap1>
  rr:logicalTable [ rr:tableName "EMP" ];
  rr:subjectMap [
    rr:template "http://data.example.com/employee/{EMPNO}";
    rr:class ex:Employee;
  ];
  rr:_predicateObjectMap [ rr:predicate ex:name;
    rr:objectMap [ rr:column "ENAME" ];
  ].
</#TriplesMap1>
```

R2RML defines:
- Base table(s) / view(s)
- Rules that map each row in the base table to RDF triples

Analytics and Data Summit 2020
Integrating GeoSpatial RDF Data in e-Stat LOD

Administrative Divisions

Japan → prefectural divisions (e.g., Tokyo) → municipal divisions (e.g., Shinjuku-ward)

standard area codes

Statistics

world grid square codes

Grid Squares

3rd grid square (1km)
4th grid square (500m)
5th grid square (250m)

Polygon

Small Areas

Subdivided administrative units of a municipal divisions

Analytics and Data Summit 2020
Size and Scale of e-Stat LOD (as of today)

AP Tier
Fuseki on Jetty (x2 VM instances)
- CPU: 8 cores
- RAM: 120 GB

DB Tier
Oracle Database Extreme Performance
- CPU: 12 cores
- RAM: 754 GB
- BareMetal Instance
- Number of Triples: 2.1 billion
1. Using Database In-Memory Features

DB In-Memory Settings for e-Stat LOD

- 380GB In-Memory Area (SGA = 600GB)
- Set the RDF semantic network indexes to INVISIBLE for the optimizer

Two semantic network indexes are created on RDF_LINK$ table by default.
- Index 1: Predicate - Object – Subject
- Index 2: Predicate - Subject - Object

- Minimize the area for METADATA to maximize the In-Memory area size for DATA
  - Set "_inmemory_64k_percent"=1 to reduce the metadata area to 1%

<table>
<thead>
<tr>
<th>POOL</th>
<th>ALLOC GB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1MB POOL</td>
<td>356.8</td>
</tr>
<tr>
<td>64KB POOL</td>
<td>22.9</td>
</tr>
</tbody>
</table>

99% for DATA
1% for METADATA
Database Design to Improve SPARQL Performance

1. Using Database In-Memory Features

Example: Tested SPARQL Query
Queries population census data in Kyoto City

```
select ?year ?population
where {
  ?s estat-measure:population ?population;
  sdmx-dimension:refArea / rdfs:label 'Kyoto-shi'@en ;
  cd-dimension:timePeriod ?year ;
  cd-dimension:sex cd-code:sex-all ;
  cd-dimension:nationality cd-code:nationality-japan ;
  g00200521-dimension-2010:area g00200521-code-2010:area-all ;
  cd-dimension:age cd-code:age-all .
}
```

<table>
<thead>
<tr>
<th>year</th>
<th>population</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;2015&quot;^^xsd:gYear</td>
<td>&quot;1412924&quot;^^xsd:decimal</td>
</tr>
<tr>
<td>&quot;2010&quot;^^xsd:gYear</td>
<td>&quot;1408039&quot;^^xsd:decimal</td>
</tr>
</tbody>
</table>

WITHOUT DB In-Memory
7.1 sec

WITH DB In-Memory
0.58 sec

12x faster
2. Partitioning RDF Table by Triple “predicate”

Partitioning Settings for e-Stat LOD

- Hash-Partitioning RDF_LINKS can be done when creating a semantic network.

```
BEGIN
  SEM_API.CREASEM_NETWORK(
    '< tablespace name for semantic network >',
    options=>>'model_partitioning=BY_HASH_P model_partitions=64 '
  );
END;
```

- The number of partitions = 64
  - Partitioning by Hash should be done by a power of 2 (2, 4, 8, 16, 32, 64, 128, …) to equally distribute the number of triples in each partition.
  - In e-Stat LOD, the distinct number of predicates is 144.