GDPR & You

The Nightmare of Business Analytics

Gianni Ceresa

Managing Director of DATAlysis GmbH (Switzerland)

Working with BI and EPM tools for about 10 years

Oracle ACE

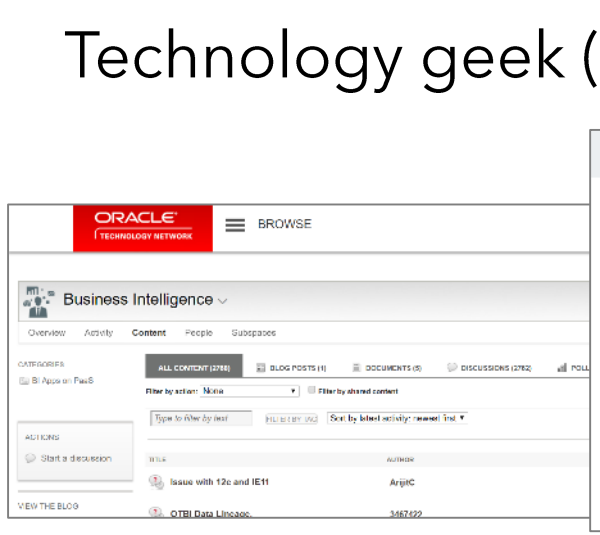
Part-time blogger on **gianniceresa.com**

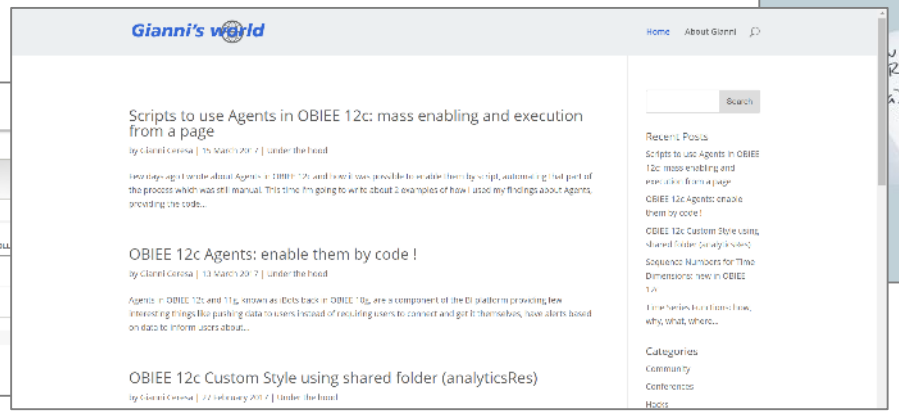
Full-time IRC (freenode | #obihackers) resident

Same group on Telegram **http://telegram.me/obihackers**

ODC (ex OTN) forums addict

Technology geek (or just geek in general)





GDPR

BRACE YOURSELF
GDPR IS COM
imgflip.com
imgflip.com
GDPR
ARE YOU TAKING IT SERIOUSLY?
S BUT...
DU MEAN GDPR IS COMING?
WE ARE STILL WAITING FOR WINTER
imgflip.com
IN
DOUBLE DARE YOU
imgflip.com

25 May 2018

GDPR

General ***D***ata ***P***rotection ***R***egulation

- Approved by EU Parliament on April 2016
- ***It is already in place !!***
- Enforcement date: 25 May 2018 (fines will start from that date)

Few points:

- The same across the European Union
- "Personal data": any information relating to a person who can be identified (directly or indirectly)
- Fines: lot of money! Up to €20 million or 4% global revenue *(the greater of the two)*
- Data Protection Officer
- Privacy management
- Breach & Notification
- Data subject access requests
- Data retention
- Right to be forgotten

GDPR

Trying to keep it simple:

- Know where the data is stored in your company
- Who has access (can't allow full DB access anymore)

Over the last 12-24 month GDPR has been a key topic at conferences … in the database track mainly

- Which DB store what?
- Who has access to the DB?

ERP/CRM systems also covered the topic as they often are the entry point where data is gathered

Overconfidence

What DB store data? Easy-peasy!!

"I master my DB, I master security in my DB, I can do auditing on it, nothing to worry about"

GDRP & Analytics – the Theory

The two simple questions for GDPR from an Analytics point of view:

- Where the data is stored in your company?
  - Everyone with access to the enterprise analytical platform can potentially export, and therefor store, data locally on his workstation
  - Self-service BI: Excel files, local departmental DBs with data inside out of the control of IT
- Who has access to data?
  - Half of the company has access to the enterprise analytical platform, who can access what?
  - Scheduled reports sent to a group mailbox or published on a shared folder
  - Inheritance side effects in the security model
    - Users will complain when they don't have access, they will hardly say something when having too much (also because not aware maybe)

GDRP & Analytics – the Theory

How can Analytics compliance from a GDPR point of view be assessed?

data lineage (what)

+

real security (who)
(resolving inheritance to discover "holes")

Apply to OBIEE and Oracle Analytics Cloud

GDRP & Analytics – Out of the Box Solutions

Data lineage:

- Catalog reports (via Catalog Manager)
  Hierarchical columns missing and few other pieces

- Repository Documentation (via Administration Tool)
  Flat "static" file which need to be generated all the time. Can be a bit confusing to read when having lot of mappings on the same logical objects. Missing aliases!

- Usage Tracking
  Only tracks what actually happened, not what could happen (executed analysis vs having access to it but not clicking on it)

Real security:

- nothing *(except for a Sample App script for catalog permissions)*

GDRP & Analytics – Out of the Box Solutions

You ***can't*** do a GDPR assessment based on these tools!

Two options:

A) Do the ostrich and put your head in the sand (and cross your fingers nothing happen and nobody asks anything)

B) Find something else …

GDRP & Analytics – The "Something Else"

Data Lineage on Steroids

GDRP & Analytics – Data Lineage on Steroids

1) Collect ***all*** the metadata

- WSDL web services for RPD, Catalog, security
- LDAP query access to Weblogic embedded LDAP (users & groups)
- DB access to OPSS (Oracle Platform Security Services) with mappings between groups/users and application roles (external LDAP providers groups/users included)
- LDAP query access to external corporate LDAP (Active Directory etc.)

2) Store and analyse data easily

- Graph database
  - vertices (nodes): any kind of object (analysis, physical column, application role etc.)
  - edges: the relations between edges (permissions, "child of", "uses/contains/references" etc.)

GDRP & Analytics – Data Lineage on Steroids

Looking a bit more in detail these aspects …

Data Lineage on Steroids – WSDL Web Services

OBIEE has a main WSDL linking various web services for the different aspects

The WSDL has various versions as it is extended / updated over time

/analytics-ws/saw.dll/wsdl/v12 (for OBIEE 12c / OAC)

OBIEE 11.1.1.7.130605 has up to v8

OBIEE 12c has up to v12

OAC 17.4 has up to v12

Documentation for OBIEE 12.2.1.3:

https://docs.oracle.com/middleware/12213/biee/BIEIT/soa_overview.htm

(beware the doc isn't really up to date, it looks more like a mix of 11g/12c versions)

Data Lineage on Steroids – WSDL Web Services

Key services and methods:

- XMLViewService Service
  - executeSQLQuery() : Runs a SQL query.
- WebCatalogService Service
  - readObjects() : Reads an object from the catalog.
  - getSubItems() : Retrieves the collection of child subitems for an object in the catalog.
  - getItemInfo() : Retrieves catalog information for an object.
- SecurityService Service
  - getGlobalPrivileges() : Gets the list of all global privileges.
  - getAccounts() : Searches for Oracle BI EE user accounts. (mix of users/groups/app roles)

Data Lineage on Steroids – WSDL Web Services: RPD metadata

Get the RPD metadata from the running OBIEE server by calling a NQS function

(more detail on NQS functions https://www.slideshare.net/ChristianBerg8/oracle-open-world-neos-voyage-2014)

call NQSQueryMetadataObjects('objectId', '', '', '', 'false', '', '')

where objectId is a number representing each type of object in the RPD

=> return some XML with the metadata

```
1500 : ObjectPrivilege     3003 : PhysicalColumn      3058 : AWLevel
1501 : User                3006 : PhysicalForeignKey  3060 : AWColumn
2000 : BusinessModel       3008 : PhysicalKey         3064 : PhysicalDimension
2004 : LogicalComplexJoin  3012 : ComplexJoin         3900 : ## Unknown ##
2006 : LogicalColumn       3023 : Database            4004 : PresentationCatalog (SubjectArea)
2008 : LogicalKey          3025 : PhysicalCatalog     4008 : PresentationTable
2010 : ## Unknown ##       3027 : SchemaName          4010 : PresentationColumn
2019 : Dimension           3029 : ConnectionPool      4014 : TargetLevel
2025 : LogicalLevel        3031 : Variable            4016 : ListCatalog
2027 : ## Unknown ##       3033 : InitBlock           4018 : QualifiedListItem
2033 : LogicalForeignKey   3039 : CubeTable           4020 : QualifyingKey
2035 : LogicalTable        3041 : PhysicalHierarchy   4026 : SegCatalog
2037 : LogicalTableSource  3043 : CubeColumn          4028 : PresentationHierarchy
2046 : ## Unknown ##       3045 : PhysicalLevel       4030 : PresentationLevel
2048 : MeasureDefn         3049 : AnalyticWorkspace   4204 : QueryPrivilege
2050 : AggrRule            3052 : AWDimension         4206 : Group
2054 : OverrideAggrRule    3054 : AWCube              4209 : PrivilegePackage
3001 : PhysicalTable       3056 : AWHierarchy         4213 : Project
```

Data Lineage on Steroids – WSDL Web Services: RPD metadata

PresentationColumn example:

```
<?xml version="1.0" encoding="UTF-8" ?>
<Repository xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <DECLARE>
        <PresentationColumn hasDispDescription="false" hasDispName="false" id="4010:386" name="Product" overrideLogicalName="false" parentId="4008:372"
        parentName="&quot;Sample Targets Lite&quot;..&quot;Products&quot;" parentUid="80cb6c15-0fa8-0000-714b-e31d00000000" uid="
        80cb6d6f-0faa-0000-714b-e31d00000000">
            <Description>
                <![CDATA[Returns the product description from the Product dimension. This column is logically sorted by BIEE server according to the Product
                Sequence column value, and is filtered by BIEE server using the Product Number column value. Naturally drills into Office Column.]]>
            </Description>
            <RefLogicalColumn id="2006:281" qualifiedName="&quot;SampleApp Lite&quot;.&quot;D1 Products&quot;.&quot;Product&quot;" uid="
            80cb6809-07d6-0000-714b-e31d00000000"/>
        </PresentationColumn>
        <PresentationColumn hasDispDescription="false" hasDispName="false" id="4010:387" name="Product Number" overrideLogicalName="false" parentId="4008:372"
        parentName="&quot;Sample Targets Lite&quot;..&quot;Products&quot;" parentUid="80cb6c15-0fa8-0000-714b-e31d00000000" uid="
        80cb6d70-0faa-0000-714b-e31d00000000">
            <Description>
                <![CDATA[Product Key is an integer that uniquely identifies each record in the Product dimension table.]]>
            </Description>
            <RefLogicalColumn id="2006:280" qualifiedName="&quot;SampleApp Lite&quot;.&quot;D1 Products&quot;.&quot;Product Number&quot;" uid="
            80cb6808-07d6-0000-714b-e31d00000000"/>
        </PresentationColumn>
        <PresentationColumn hasDispDescription="false" hasDispName="false" id="4010:388" name="Product Type" overrideLogicalName="false" parentId="4008:372"
        parentName="&quot;Sample Targets Lite&quot;..&quot;Products&quot;" parentUid="80cb6c15-0fa8-0000-714b-e31d00000000" uid="
        80cb6d71-0faa-0000-714b-e31d00000000">
            <Description>
                <![CDATA[Returns the Product Type description from the Product dimension. This column is logically filtered by BIEE server using the Product
                Type Key column value. Naturally drills into Product Column.]]>
            </Description>
            <RefLogicalColumn id="2006:282" qualifiedName="&quot;SampleApp Lite&quot;.&quot;D1 Products&quot;.&quot;Product Type&quot;" uid="
            80cb680a-07d6-0000-714b-e31d00000000"/>
        </PresentationColumn>
```

Data Lineage on Steroids – WSDL Web Services

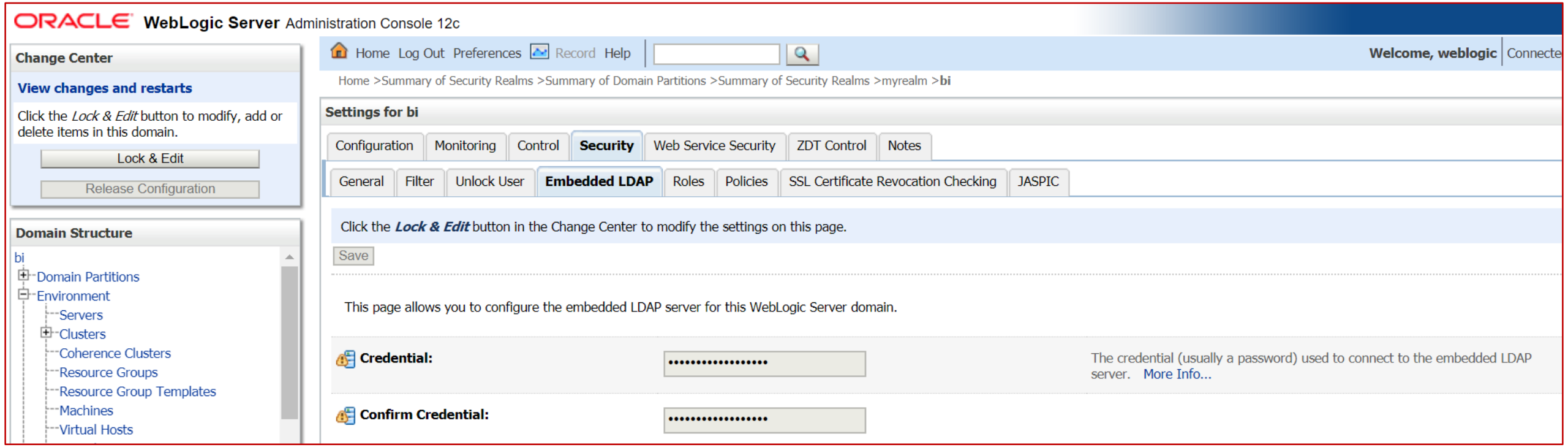
Some considerations:

- Not for a manual usage (you don't call a Web Service by hand)
- Easy to script as a real standard WSDL Web Service
  - Python, Java etc. : all will master it just fine
- Results are XML (always)
  - Arbitrary data structures
  - Hard to process by hand
  - Easy to manage by script
- Requires a valid OBIEE account with the widest access possible
  - Web Services return only what the used credentials can see
  - "weblogic" is the kind of account working the best (system admin)

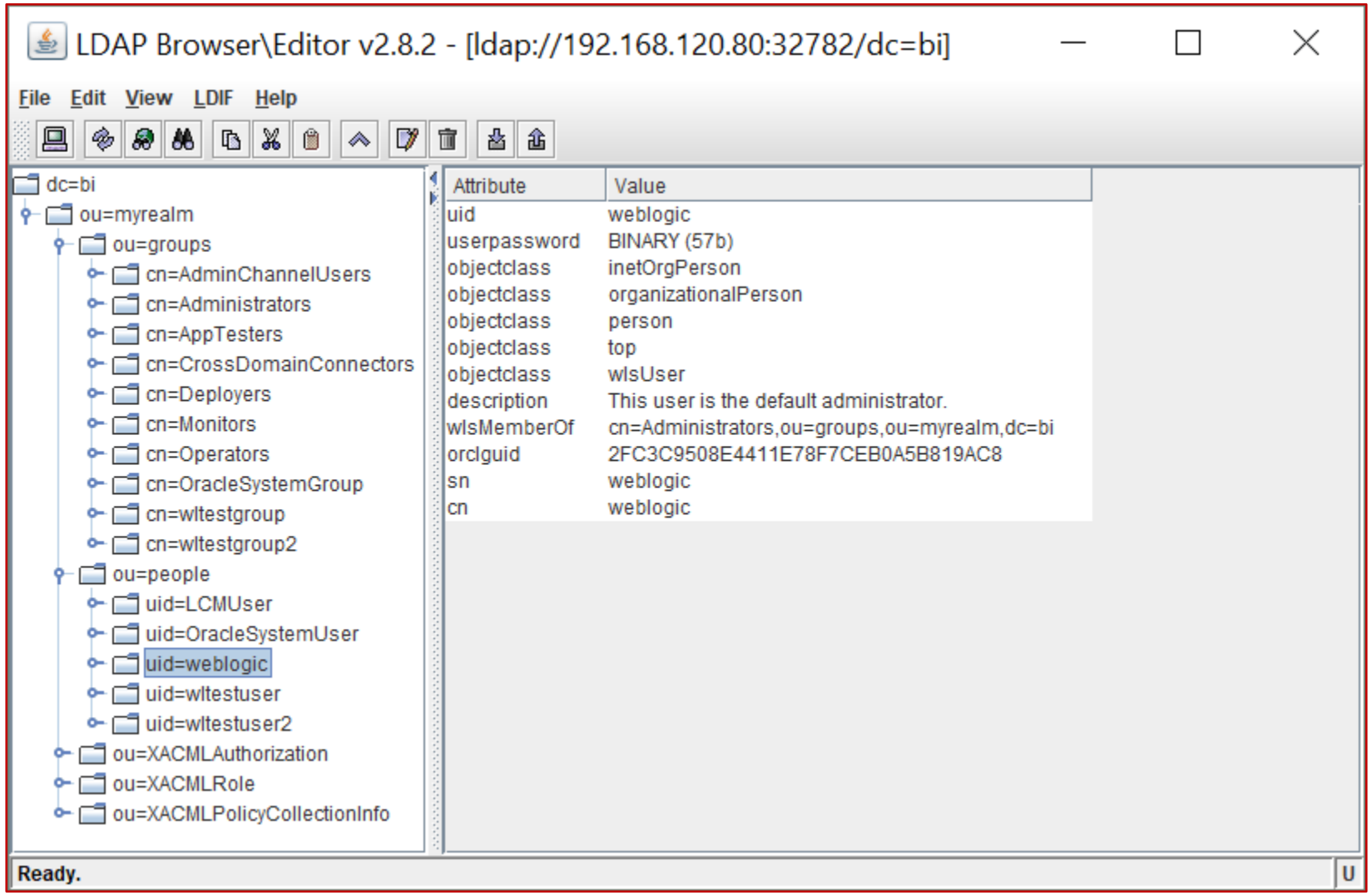
Data Lineage on Steroids – Embedded LDAP

Enable LDAP access with a LDAP client

- Weblogic Console
- Domain > Security > Embedded LDAP
- Set “Credential” to the password of your choice
- Restart



Data Lineage on Steroids – Embedded LDAP



Data Lineage on Steroids – Oracle Platform Security Services

In OBIEE 12c the Fusion Middleware security isn't only in files anymore, a copy is stored in the database

As part of the various schemas created by the RCU one is called <prefix>_OPSS

DO NOT ALTER THE CONTENT OF THE DB !!!

A kind of LDAP-like structure / naming

Interesting tables:

- JPS_DN : "looks like" it stores the "distinguished name" (unique) of each element (app roles, groups, users)
- JPS_ATTRS : "looks like" it stores attributes / tree structure of the security model

because it isn't documented and we aren't supposed to put our hands there

Data Lineage on Steroids – Oracle Platform Security Services

table: JPS_DN

	ENTRYID	RDN	PARENTDN
1	7963	cn=biserviceadministrator	cn=opssroot,cn=jpscontext,cn=opsssecuritystore,cn=obi,cn=roles,
2	7964	cn=bicontentauthor	cn=opssroot,cn=jpscontext,cn=opsssecuritystore,cn=obi,cn=roles,
3	7965	cn=biconsumer	cn=opssroot,cn=jpscontext,cn=opsssecuritystore,cn=obi,cn=roles,

table: JPS_ATTRS

	JPS_ATTRS_ID	JPS_DN_ENTRYID	ATTRNAME	ATTRVAL	ATTRKIND
1	50007	7963	uniquemember	cn=LdapAdmins+orcljaznjavaclass=weblogic.security.principal.WLSGroupImpl	0
2	53007	7963	uniquemember	cn=wltestgroup+orcljaznjavaclass=weblogic.security.principal.WLSGroupImpl	0
3	29962	7963	uniquemember	cn=weblogic+orcljaznjavaclass=weblogic.security.principal.WLSUserImpl	0
4	50003	7964	uniquemember	cn=LdapAuthors+orcljaznjavaclass=weblogic.security.principal.WLSGroupImpl	0
5	53005	7964	uniquemember	cn=wltestuser+orcljaznjavaclass=weblogic.security.principal.WLSUserImpl	0
6	31448	7964	uniquemember	cn=BIServiceAdministrator,cn=Roles,cn=obi,cn=opssSecurityStore,cn=JPSContext,cn=opssRoot	0
7	53003	7965	uniquemember	cn=Test LDAPUser+orcljaznjavaclass=weblogic.security.principal.WLSUserImpl	0
8	31450	7965	uniquemember	cn=BIContentAuthor,cn=Roles,cn=obi,cn=opssSecurityStore,cn=JPSContext,cn=opssRoot	0
9	31452	7965	uniquemember	cn=authenticated-role+orcljaznjavaclass=oracle.security.jps.internal.core.principals.JpsAuthenticatedRoleImpl	0
10	50005	7965	uniquemember	cn=LdapConsumers+orcljaznjavaclass=weblogic.security.principal.WLSGroupImpl	0

Data Lineage on Steroids – Oracle Platform Security Services

In OBIEE 11g similar information can be extracted from system-data-jazn.xml file

A clean and well structured XML:

- application roles
- members
- types

```
<application locale="en_US">
    <name>obi</name>
    <app-roles>
        <app-role>
            <name>BISystem</name>
            <display-name>BI System Role</display-name>
            <guid>D8507DC05C1111E3BF5191B4076B84C0</guid>
            <class>oracle.security.jps.service.policystore.ApplicationRole</class>
            <members>
                <member>
                    <class>weblogic.security.principal.WLSUserImpl</class>
                    <name>BISystemUser</name>
                </member>
            </members>
        </app-role>
        <app-role>
            <name>BIAdministrator</name>
            <display-name>BI Administrator Role</display-name>
            <guid>D8507DC15C1111E3BF5191B4076B84C0</guid>
            <class>oracle.security.jps.service.policystore.ApplicationRole</class>
            <members>
                <member>
                    <class>weblogic.security.principal.WLSGroupImpl</class>
                    <name>BIAdministrators</name>
                </member>
            </members>
        </app-role>
        <app-role>
            <name>BIAuthor</name>
            <display-name>BI Author Role</display-name>
            <guid>D8507DC25C1111E3BF5191B4076B84C0</guid>
            <class>oracle.security.jps.service.policystore.ApplicationRole</class>
            <members>
                <member>
                    <class>weblogic.security.principal.WLSGroupImpl</class>
                    <name>BIAuthors</name>
                </member>
                <member>
                    <class>oracle.security.jps.service.policystore.ApplicationRole</class>
                    <name>BIAdministrator</name>
                    <guid>D8507DC15C1111E3BF5191B4076B84C0</guid>
                </member>
```

Data Lineage on Steroids – Oracle Platform Security Services

The alternative, supported and documented, way to extract mappings between users / groups / application roles and application roles is WLST

- ListAppRoles : list all the application roles defined in the system
- ListAppRolesMembers : return a list of the members of a given application role and their type (user, group or application role)

Can be scripted and automated

Require to be executed on the OBIEE server itself or another system with WLST available and able to connect to OBIEE

Data Lineage on Steroids – Graph Database

Why a graph database to store metadata?

- Most of the objects retrieved will have references to others
  - Number of references is variable between 1 and N
  - Nature of the reference is variable depending on the kind of relationship
  - References can have a set of properties (between 0 and N)
- Most of the objects retrieved will have a set of attributes (properties)
  - Number of attributes is variable between 0 and N
  - Nature of the attribute is variable depending on the kind of object

A relational database can perfectly store that kind of data with mapping tables and storage in rows (instead of columns)

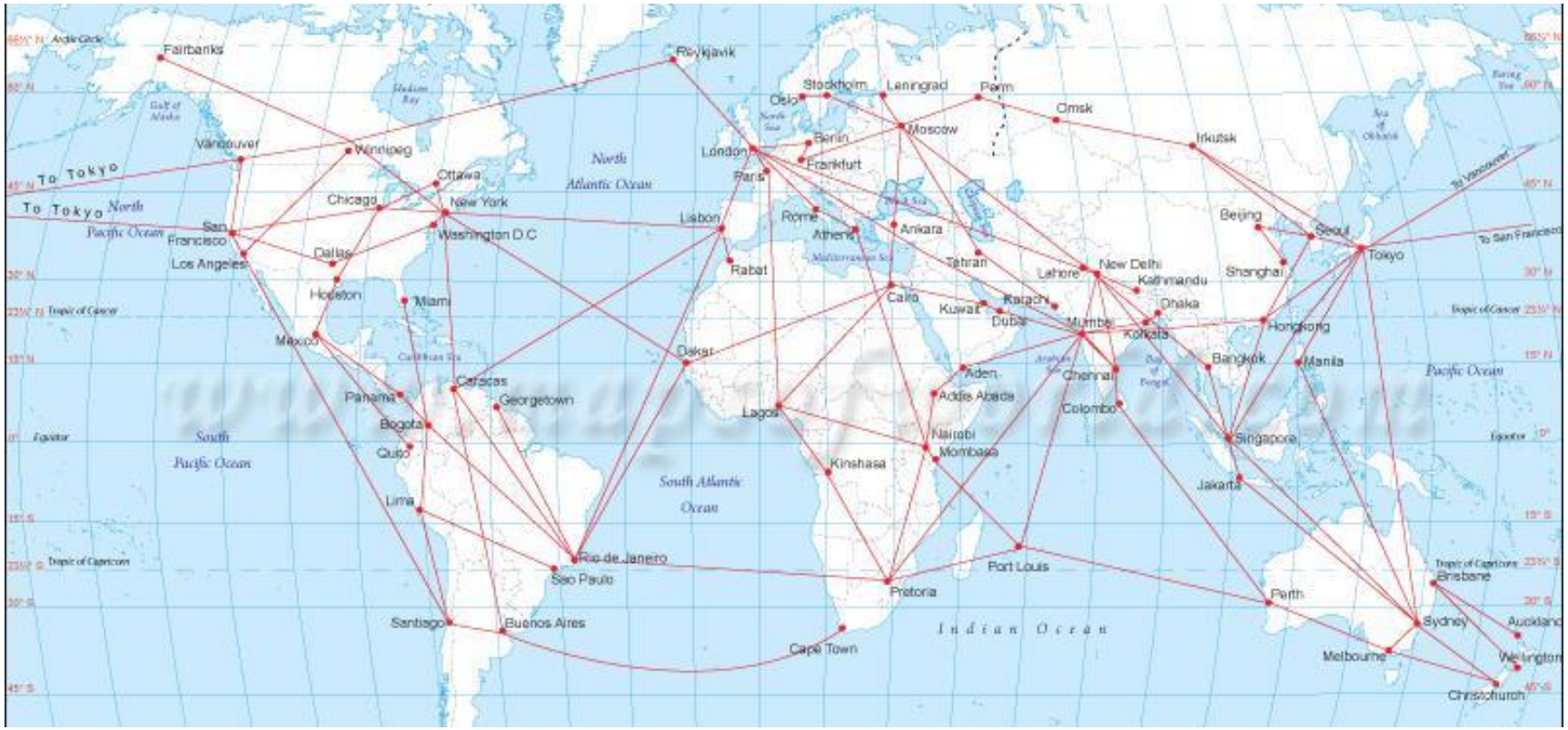
A graph database provides the best flexibility, easy of use, power of analysis

(and it's a lot more fun)

Data Lineage on Steroids – Graph Database

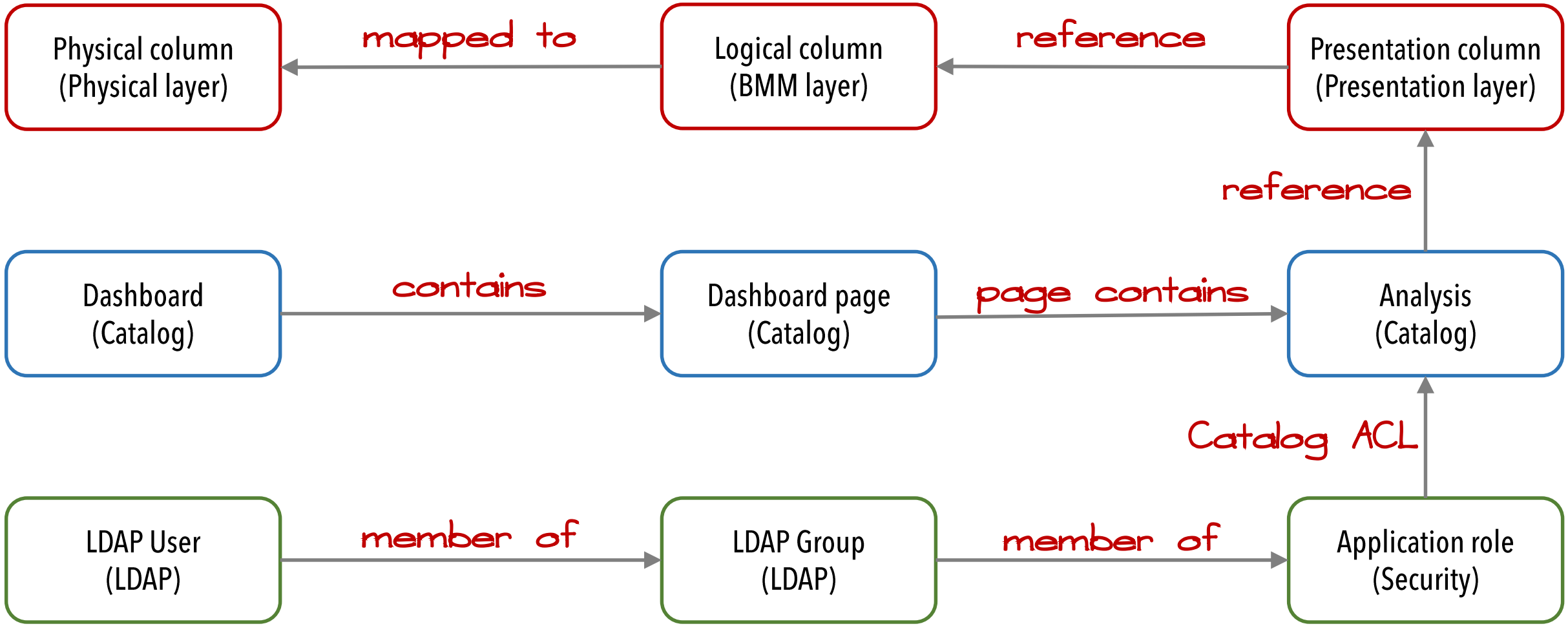
Data lineage can be navigate from a point A to a point B, just like you would do to travel from a location A to a location B : that's what a graph database is also really good with

From a physical column in a table of a database, to an analysis on the front-end



Data Lineage on Steroids – Graph Database: the Idea

A simple OBIEE system can be seen as … a graph

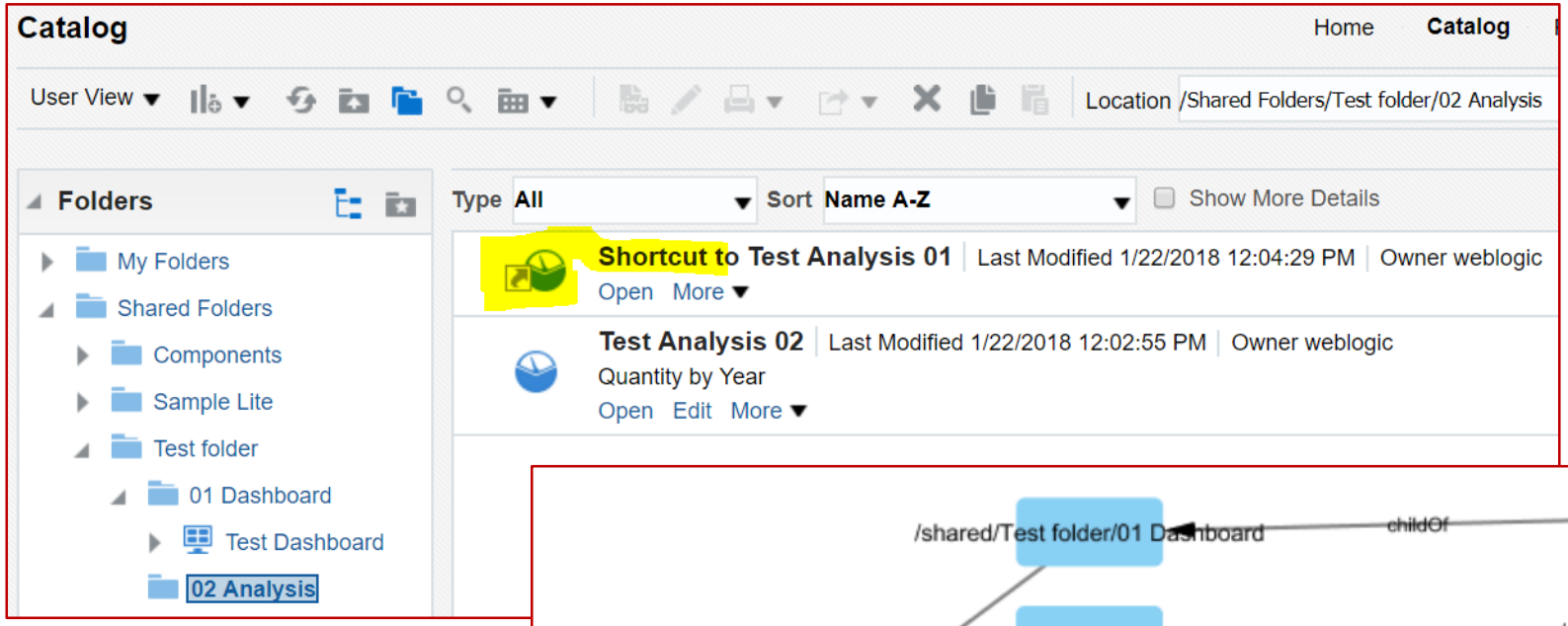


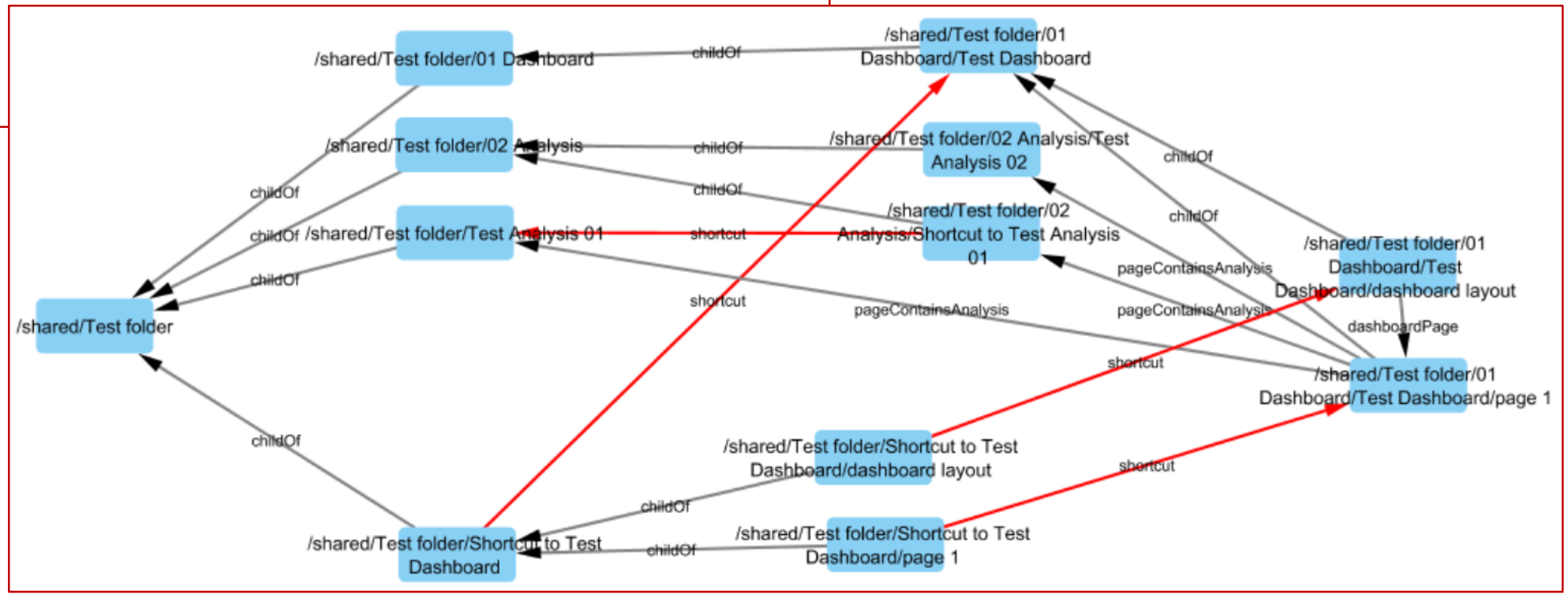
Data Lineage on Steroids

Some of the cases ***Data Lineage on Steroids*** solves and which can't be done with the out of the box tools:

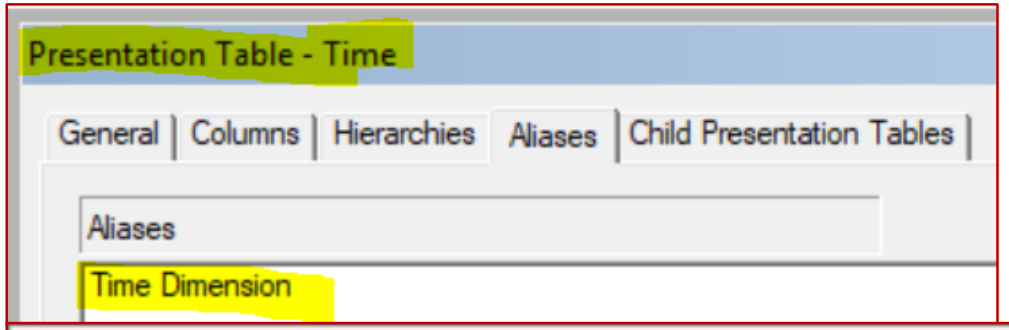
- Shortcuts in the catalog
- Aliases in the Presentation layer of the RPD
- Security inheritance

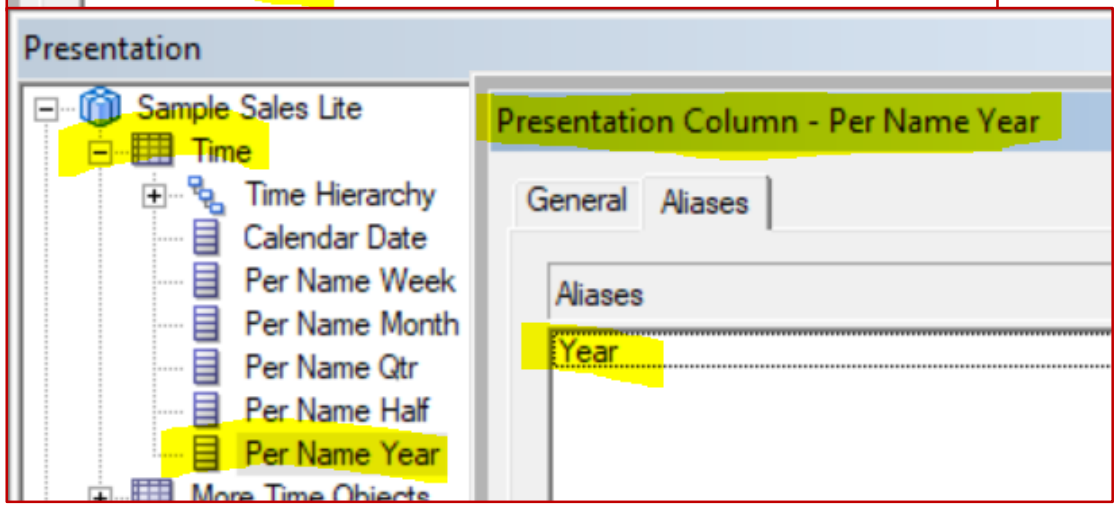
Data Lineage on Steroids – Catalog Shortcuts



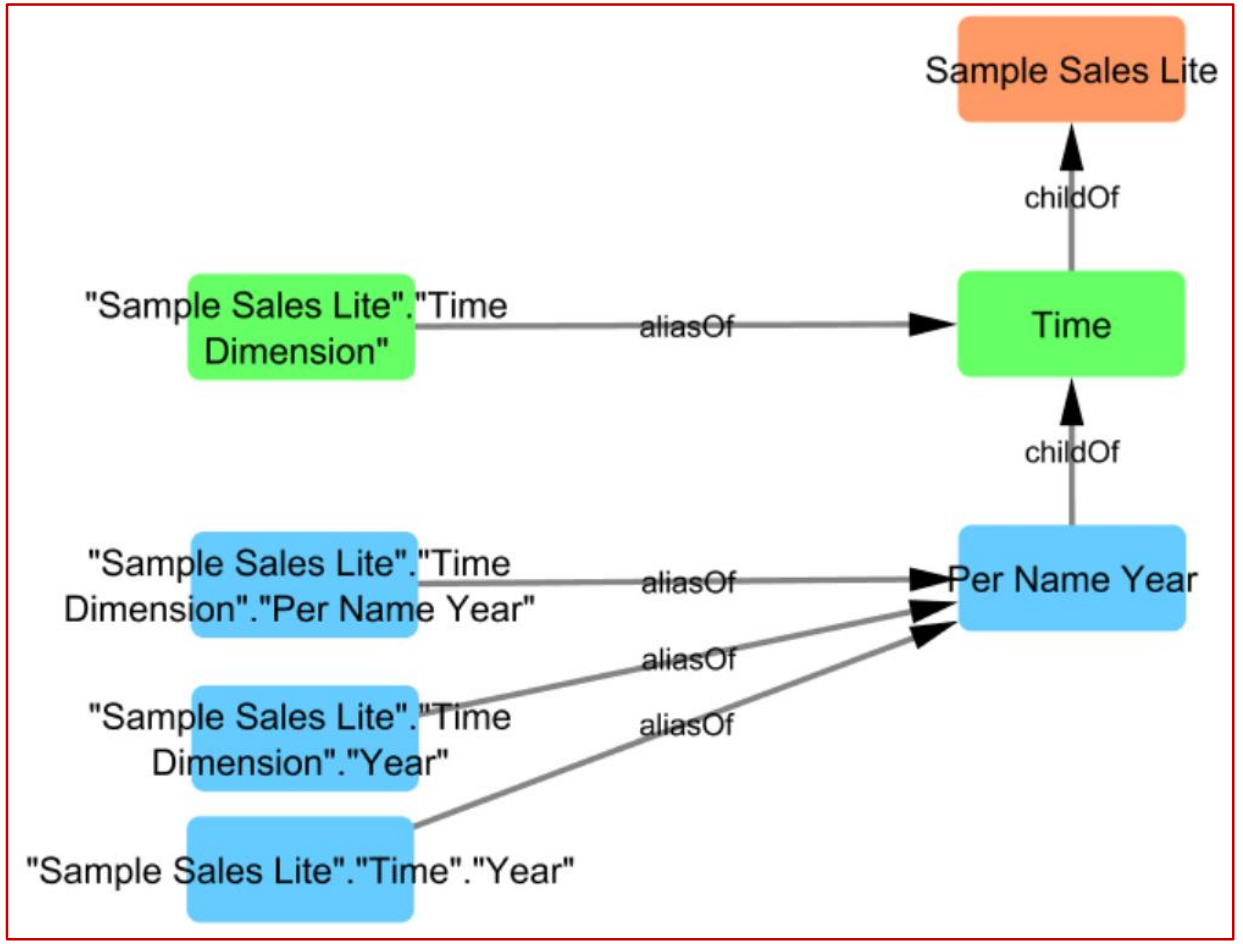


Data Lineage on Steroids – RPD Aliases



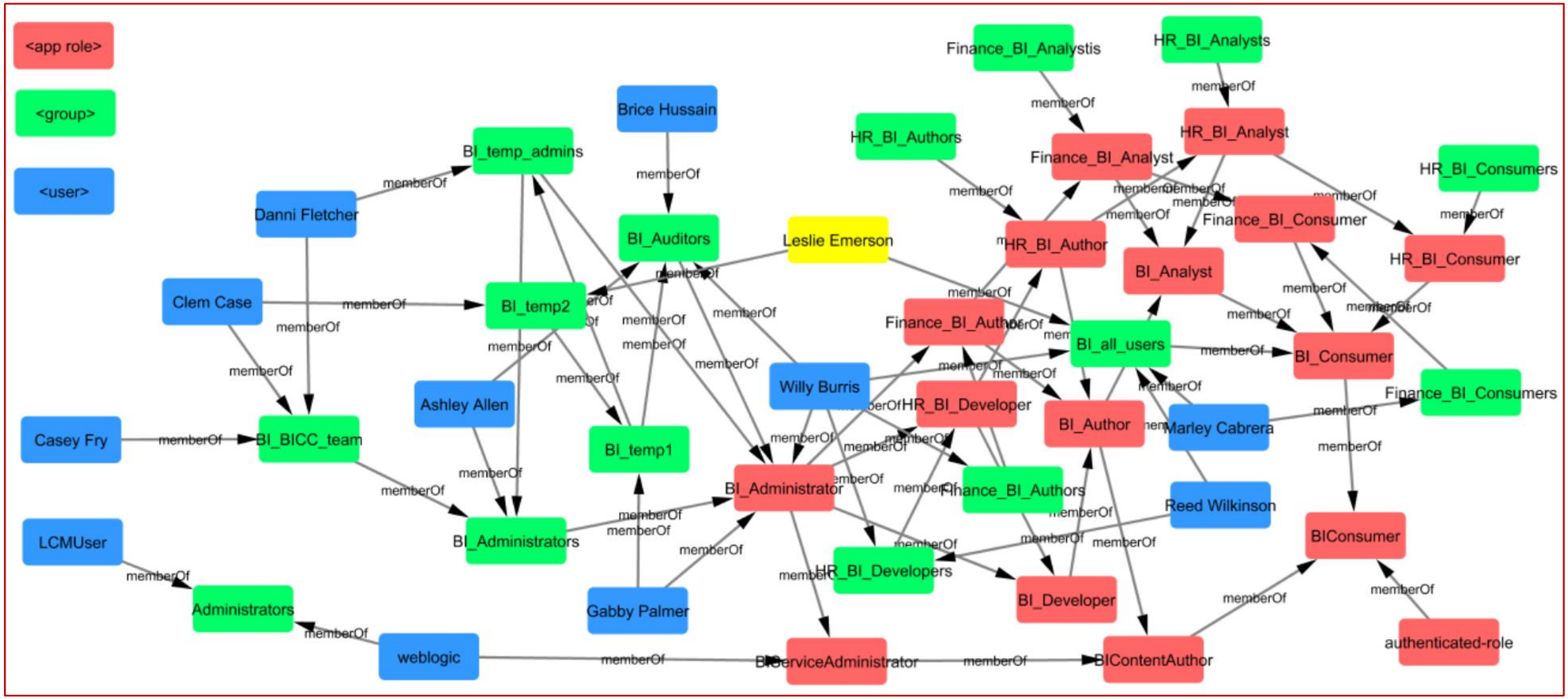


"Time"."Per Name Year"	"Time"."Year"	"Time Dimension"."Per Name Year"	"Time Dimension"."Year"
2006	2006	2006	2006
2007	2007	2007	2007
2008	2008	2008	2008
2009	2009	2009	2009
2010	2010	2010	2010
2011	2011	2011	2011

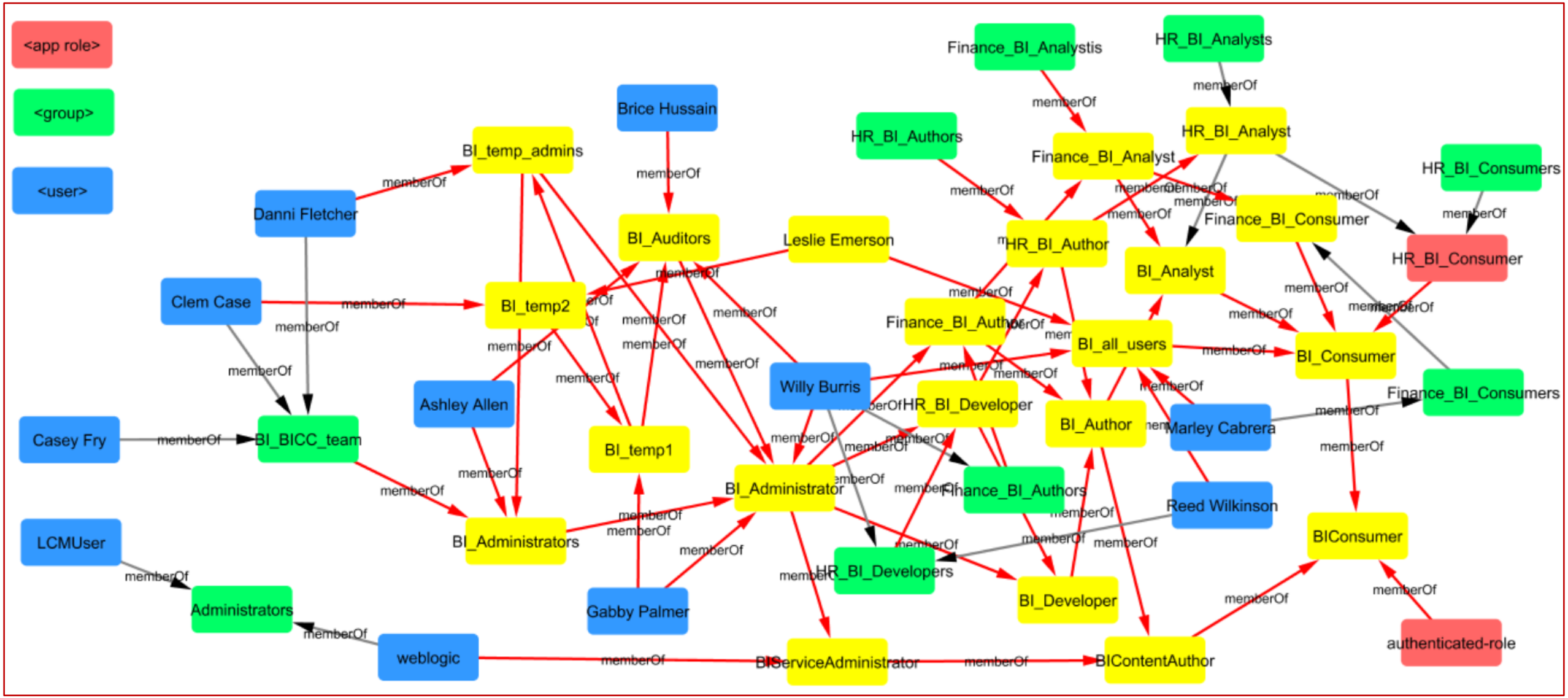


Data Lineage on Steroids – Security Inheritance

What groups/application roles is “Leslie Emerson” part of directly or indirectly?



Data Lineage on Steroids – Security Inheritance



Data Lineage on Steroids – Swiss Army Knife for GDPR assessment

Data Lineage on Steroids – Swiss Army Knife for GDPR assessment

My journey going through this concept of “Data Lineage on Steroids”

1. Started creating a graph based on the RPD, layer by layer
2. Continued by adding the Catalog
3. Finished by Security

4. Analysis of the collected metadata later

Data Lineage on Steroids – Swiss Army Knife for GDPR assessment

- A single tool performing all the steps and generating the graph database for analysis at later time
- For easy of use and multi-platform support (Windows and Linux) Java looked like the best solution
- The tool can be packaged as a single JAR which can be executed manually or scheduled on any system with Java 8
- A config file allows to set all the settings like hostnames, credentials, attributes for the extraction of metadata, LDAP properties etc.
- Output as:
  - XGMML file (eXtensible Graph Markup and Modelling Language)
  - PGX graph on remote server or PGB locally (binary Oracle graph engine file)
  - Neo4j
  - Any other graph format/engine accessible by Java

Data Lineage on Steroids – Swiss Army Knife for GDPR assessment

```
[23:07:16,129] INFO GDPRObiee - ################################
[23:07:16,145] INFO GDPRObiee - # (c) DATAlysis GmbH, 2017    #
[23:07:16,145] INFO GDPRObiee - # - - - - - - - - - - - - - - #
[23:07:16,145] INFO GDPRObiee - # Contact details:            #
[23:07:16,145] INFO GDPRObiee - #    https://datalysis.ch     #
[23:07:16,145] INFO GDPRObiee - #    info@datalysis.ch        #
[23:07:16,145] INFO GDPRObiee - ################################
[23:07:16,298] INFO GDPRObiee -
[23:07:16,298] INFO GDPRObiee - Welcome to the DATAlysis GmbH GDPR Obiee tool
[23:07:16,298] INFO GDPRObiee - ------------------------------------------
[23:07:16,298] INFO GDPRObiee -
[23:07:16,298] INFO GDPRObiee -  OBIEE server: http://192.168.120.80:19502
[23:07:16,298] INFO GDPRObiee -  OBIEE username: weblogic
[23:07:16,298] INFO GDPRObiee -
[23:07:16,298] INFO GDPRObiee - ------------------------------------------
[23:07:16,298] INFO GDPRObiee - Supported Graph engines:
[23:07:16,298] INFO GDPRObiee -   - NONE
[23:07:16,298] INFO GDPRObiee -   - PGX
[23:07:16,298] INFO GDPRObiee -   - NEO4J
[23:07:16,298] INFO GDPRObiee - ------------------------------------------
[23:07:16,298] INFO GDPRObiee - Graph type set to NONE
[23:07:20,654] INFO GDPRObiee - Connect to OBIEE (session ID '5n6bdj4viq4tabjjov22tekbb677oq1n9ps8hkdgl9rqgfa1')
[23:07:20,654] INFO GDPRObiee - Parsing RPD ...
[23:07:23,737] INFO GDPRObiee - Parse security ...
[23:07:29,972] INFO GDPRObiee - Parse Web Catalog path '/shared' recursively ...
[23:07:37,579] INFO GDPRObiee - Disconnect from OBIEE (session ID '5n6bdj4viq4tabjjov22tekbb677oq1n9ps8hkdgl9rqgfa1')
[23:07:37,587] INFO Graph - (balance) 0 vertex added
[23:07:37,588] INFO Graph - # Graph summary
[23:07:37,588] INFO Graph -   graph has 1348 vertices, 3118 edges
[23:07:37,981] INFO GraphNone - Graph saved as XGMML: C:\Users\Gianni\Documents\graph_sandbox01.xgmml
[23:07:37,981] INFO GDPRObiee - ---###--- Execution completed ----------###---
```

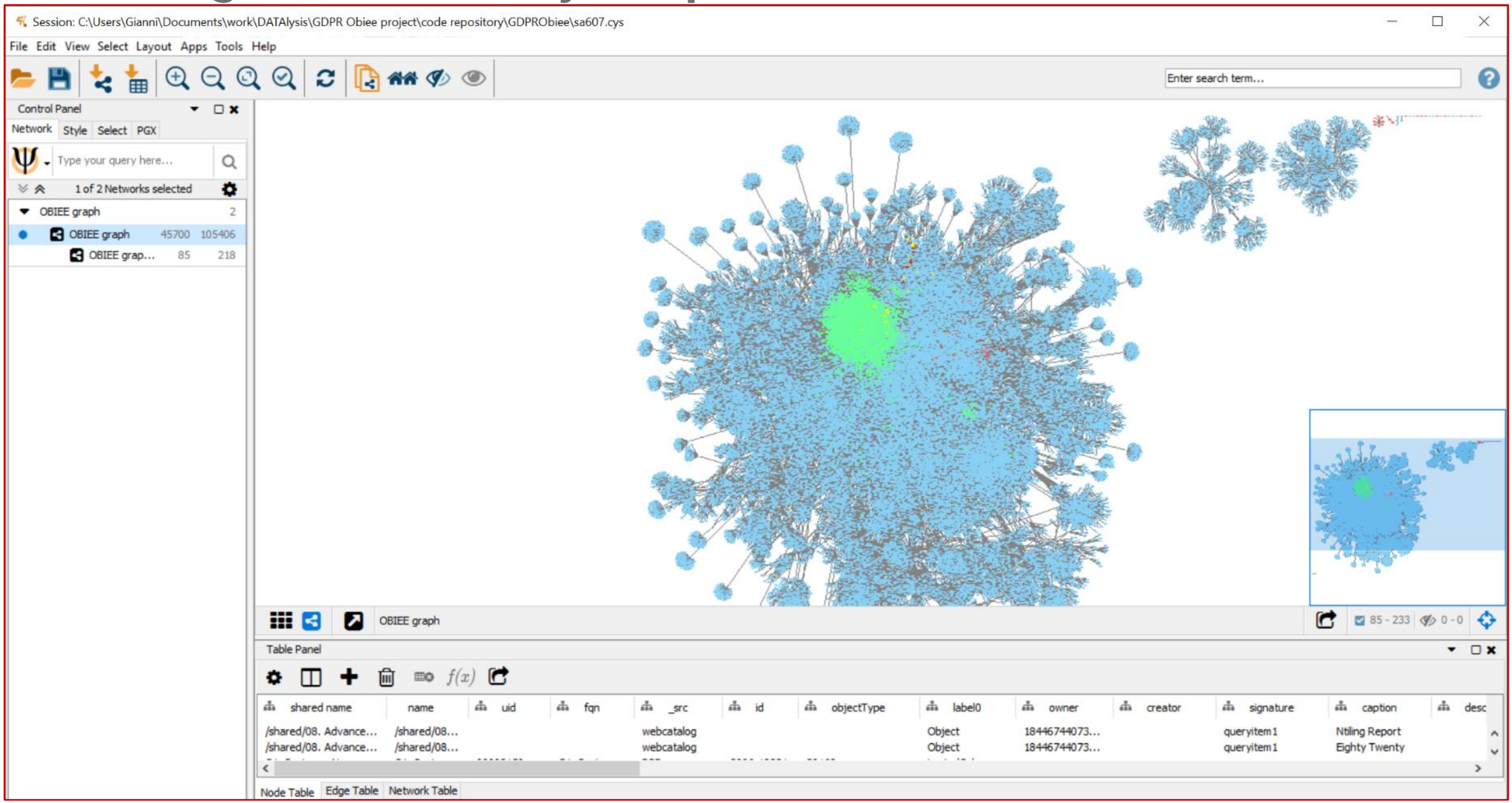
Data Lineage on Steroids – Graph Database Analysis

Two possible way to make analysis:

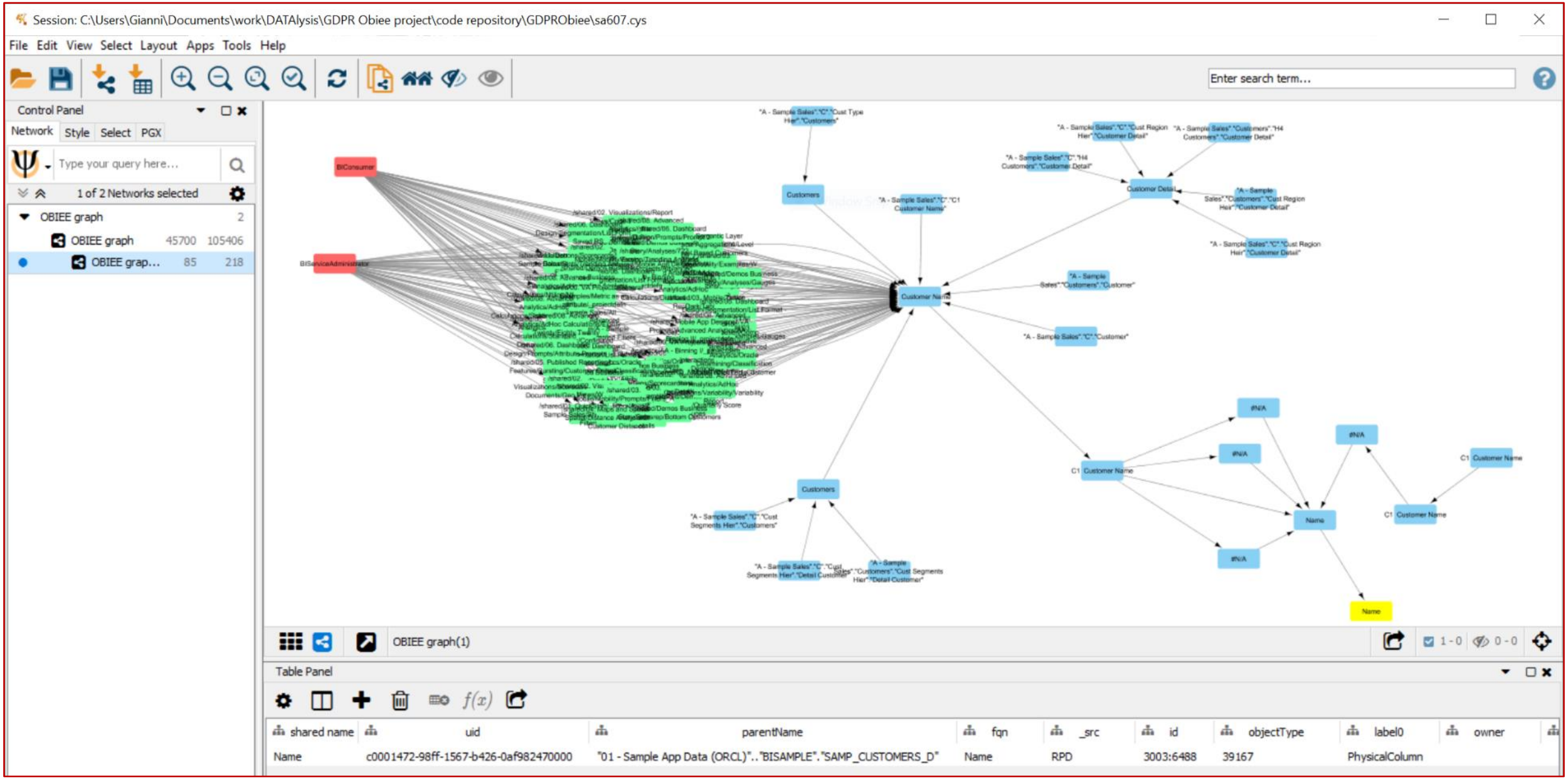
- Visually by Cytoscape
  - Free tool for graph analysis with some integrations for PGX available
  - Full support of XGMML files for “offline” analysis of the graph
  - Search for source node (sensitive database columns)
  - Get full path up to the catalog objects connected to that node
  - Create a sub-graph for more detailed analysis

- By Green-Marl built-in scripts in PGX
  - For example “Hop Distance” to calculate the distance between a given source node and all the connected nodes
  - Query nodes with a distance value set and with properties saying it’s a catalog object
  - Return the list of catalog objects connected to the source node

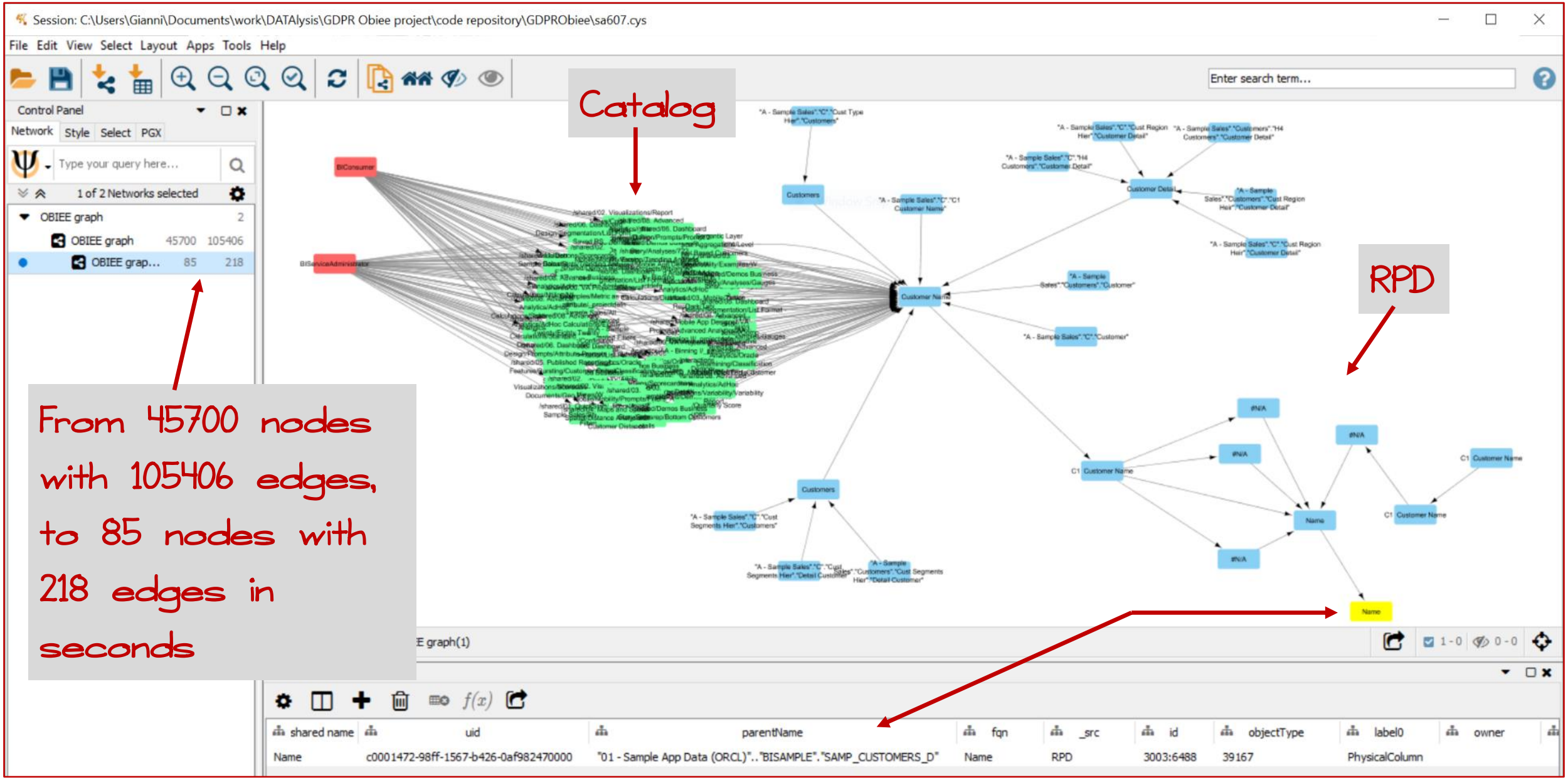
Data Lineage on Steroids – Cytoscape



Data Lineage on Steroids – Graph Database Analysis



Data Lineage on Steroids – Graph Database Analysis



Data Lineage on Steroids – Cytoscape PGX connection

So far there isn't a "nice" visual interface for the Oracle Parallel Graph AnalytiX engine (PGX) on-premises

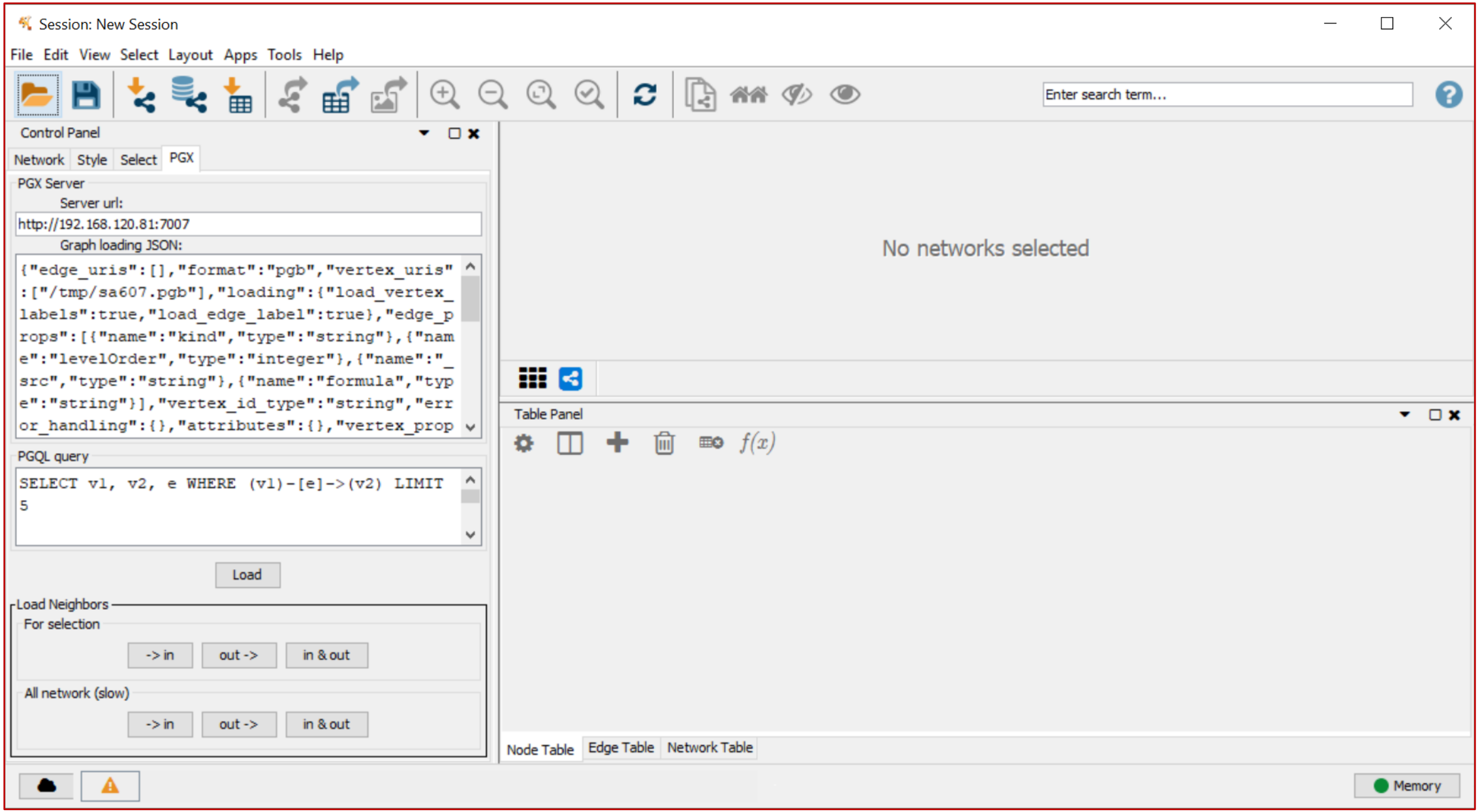
- command line
- Java API
- Python interface to the Java API (limited and with issues)
- Apache Zeppelin notebook integration only textually (no visual rendering)
  - Custom visualization via JavaScript (D3js) plugin works only for a small subsets of a graph

- On cloud Oracle is about to release a nicely integrated notebook solution (March 2018 or later)

The solution is "do it myself" : Cytoscape can be extended by plugins

- Custom plugin to connect to a PGX server, load a graph and run PGQL queries

Data Lineage on Steroids – Cytoscape PGX connection



Data Lineage on Steroids – Extra details

Data Mashup in OBIEE and Self-Service (Excel files loaded and used in DV projects or analysis)

- The XML of the Catalog objects allow to extract the references to Excel files
- Extra Web Services allows to get all the metadata of the Excel file source
- Being Excel the content is difficult to evaluate, require a "human" decision

Keep in mind that Self-Service and GDPR can hardly coexist, the human factor (providing the data for Self-Service Analytics) is always the main point of failure

Evolution of the Analytical platform over time:

- Graphs still have a weakness: evolution over time
- PGX provide a solution by keeping versions of the graphs and be able to load updates
- External versioning of the export XGMML or PGB file can also solve that aspect

Conclusion

- Do not underestimate the impact of GDPR on your enterprise analytical platform
- Out of the box no real solution (not covering the required details)
- Can be achieved by adopting a “Data Lineage on Steroids”
  - Using existing Web Service
  - Using all the possible accesses to security setup
  - Not easy by hand
  - Can be developed into a small app
  - Graph database for analysis
- Not a “click here and done” solution
  - Helps for a GDPR assessment of the situation
  - Doesn’t replace a human brain analysing results