

Preserving Relational Databases – Hungarian Use Case

Zoltán Lux
Zoltán Szatucsek
József Mezei
National Archives of Hungary

Zsolt Makovi
Csaba Kovács
XperTeam Zrt.



E-ARK final conference
Hungarian National Archives, Budapest
6-8 December 2016



Agenda

- Introduction to E-ARK (Hungarian) Pilot 7
- Achievements/solution using Data Warehouse approach in E-ARK Pilot 7
 - E-ARK WEB - Data Exploration Solution
 - Data Warehouse concept & components usage in E-ARK WEB - Data Exploration
 - Presentation data by Oracle BI
 - Presentation data by Data Explorer (Oracle APEX application)

Some Facts about the Preserved Databases

- Record management system(+) of the Hungarian Prosecution Offices
- From 1993, ...to 2000
- App. 300 databases
- All have the same structure, containing 19 tables
- But,
 - No documentation was available
 - Unclear and - from today's perspective - illogical data model
- Data originally in DBASE files
- Application software written in Clipper (no documentation)

E-ARK Pilot 7 Access to Databases - Scenario 3

Scope of the pilot:

- Extract data from a relational database (SIARD 2.0)
- Create submission information package (SIP)
- Ingest SIP package into a long-term archival repository (SIP to AIP)
- Searching AIPs according to the users' requests
- Examine the applicability of data warehouse concept in an archival environment
- Create user friendly web-based application for search and presentation



E-ARK final conference
Hungarian National Archives, Budapest
6-8 December 2016



E-ARK Pilot 7 Access to Databases

- Main goals

- Improving the presentation of the data to the user
 - Improving search facilities (filtering, faceted search, interactive reports)
 - Providing tools for analyzing the data
 - Visualize the data
- What has to be archived in order to be able to present the data to the user this way
- What can be automated, what has to be done manually.

Tools used in Pilot 7

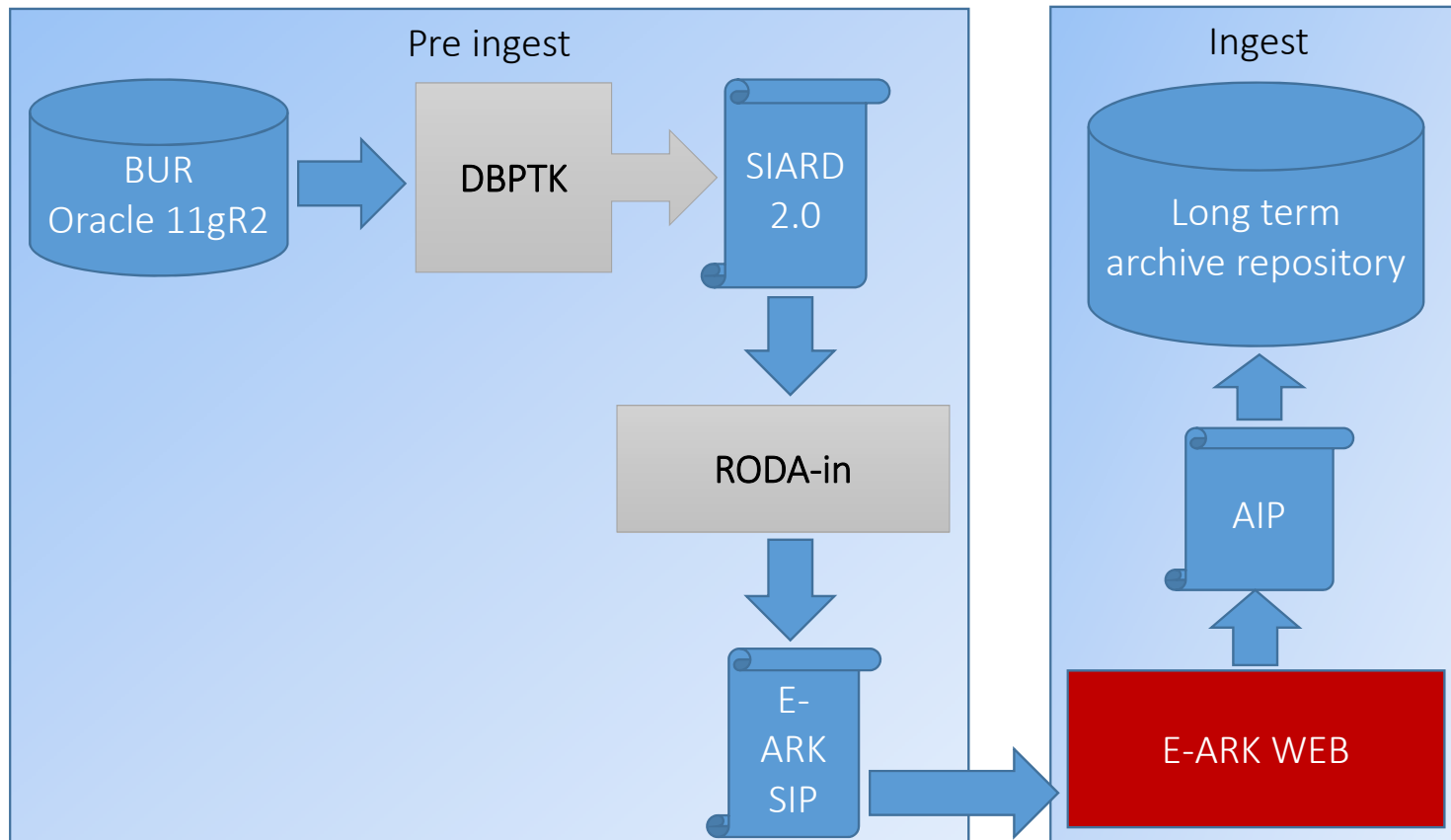
- Database Preservation Toolkit (DBPTK)
- SIP creator (RODA-in)
- E-ARK WEB (Repository of Authentic Digital Objects - Hadoop, HBASE, SolR, Lily)
- Database Visualization Toolkit
- Oracle Database Enterprise Edition with OLAP option
- Oracle Data Integrator
- Oracle SQL Developer
- Oracle Analytic Workspace Manager
- Oracle Application Express
- Oracle Business Intelligence
- Oracle Maps Viewer



E-ARK final conference
Hungarian National Archives, Budapest
6-8 December 2016



The Archival Process



E-ARK final conference
Hungarian National Archives, Budapest
6-8 December 2016

Archival Process Steps

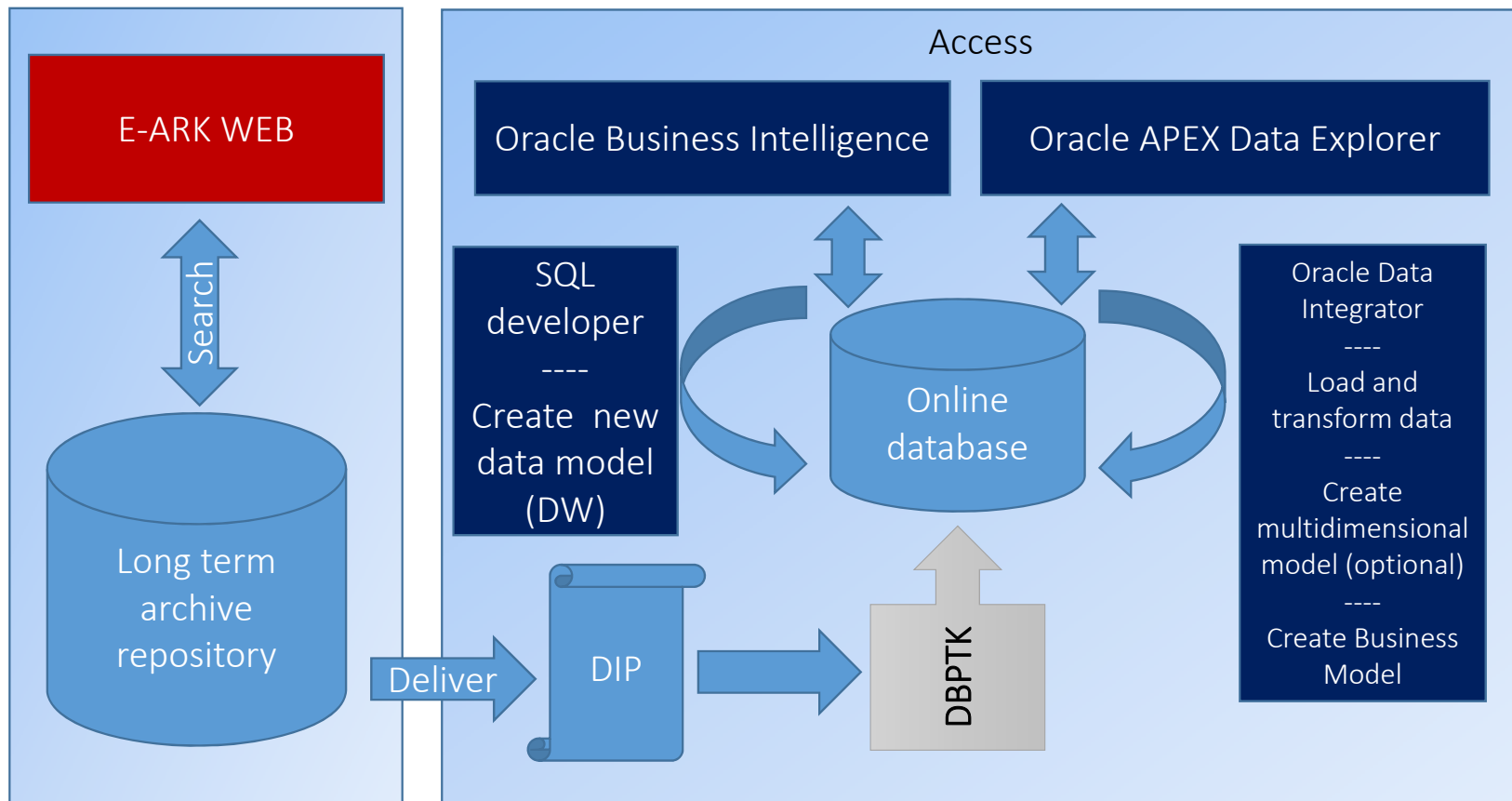
1. Extract data from a relational database into SIARD format (XML) using DBPTK
2. Create SIP (pack the SIARD file and add extra meta data) with RODA-in
3. Ingest the SIP and convert to AIP into the long term archive repository with E-ARK WEB
4. Index the archive with EARK WEB to enable full text search



E-ARK final conference
Hungarian National Archives, Budapest
6-8 December 2016



Accessing Archived Data (RDBMS)



E-ARK final conference
Hungarian National Archives, Budapest
6-8 December 2016

Data Access Steps 1/2

1. Search the repository and create from AIPs the DIP with EARK WEB
2. Load the DIP into live database (for example Oracle) with DBPTK
3. Create new data model (using data warehouse concepts) with SQL Developer to represent the data in a meaningful format
4. Load and transform the data with Oracle Data Integrator into the new data model

Data Access Steps 2/2

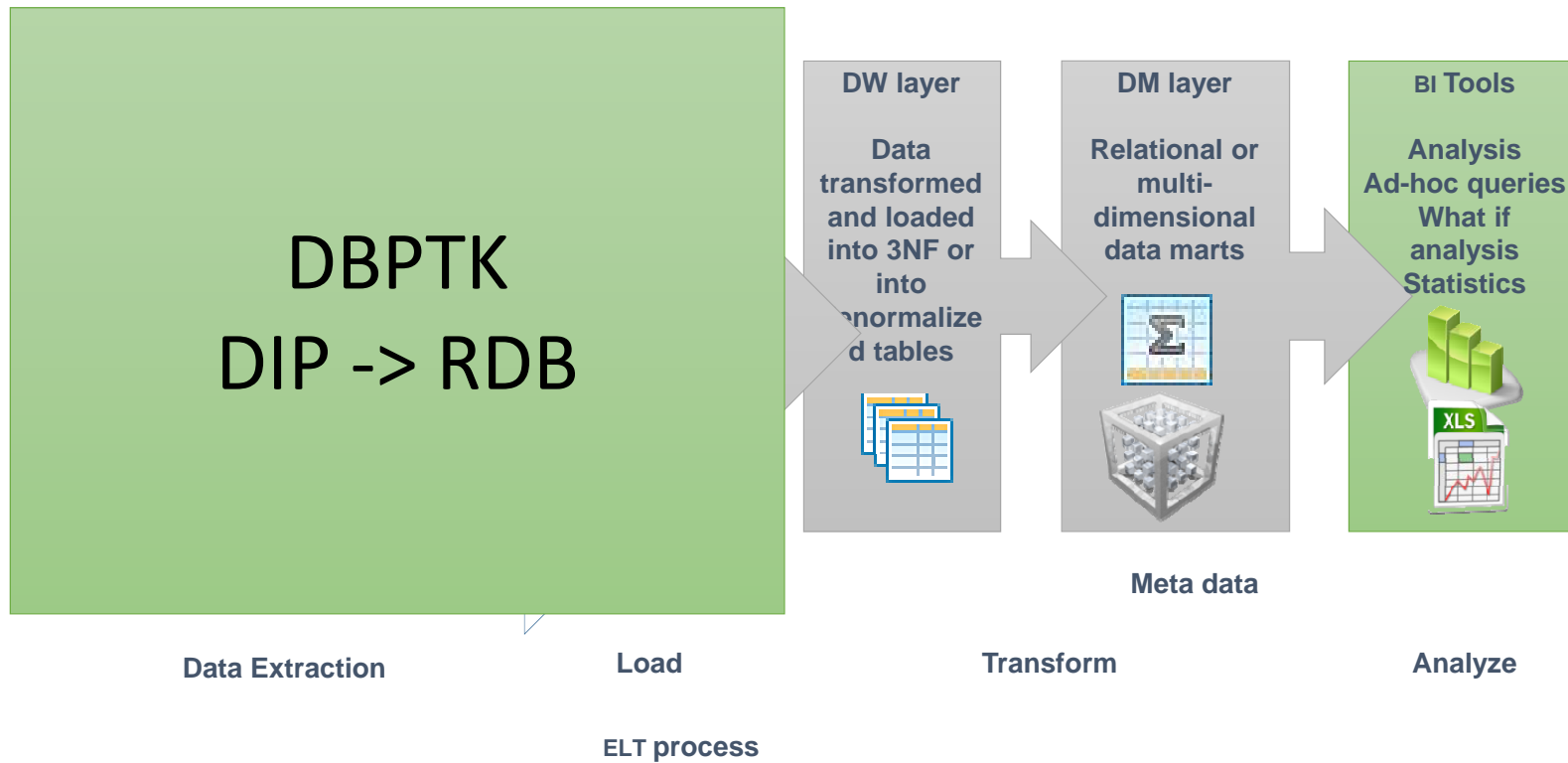
5. Optionally create multidimensional model, to pre-aggregate data (OLAP)
6. Create Business Model for Oracle Business Intelligence (identify fact and dimensions, join tables, create measures and hierarchies, add meaningful names to tables and columns, etc.)
7. Load the business model into BI Catalog
8. Create reports, analysis, charts in Oracle BI and Oracle APEX Data Explorer
9. Grant access on the reports and analysis to the requester

The Data Warehouse Concept

Why we are using data warehouse in this pilot:

- (Data from multiple data sources can be converted, transformed and ingested into a centralized database)
- Tables can be de-normalized, no complex joins needed when querying data
- The database structure is hidden by the business model, when analyzing the data, the archivist/consumer only knows the „business term” not the structure of the database
- Powerful analytic tools exists, capable to integrate, analyze and visualize the data from various sources

Data Warehouse Components

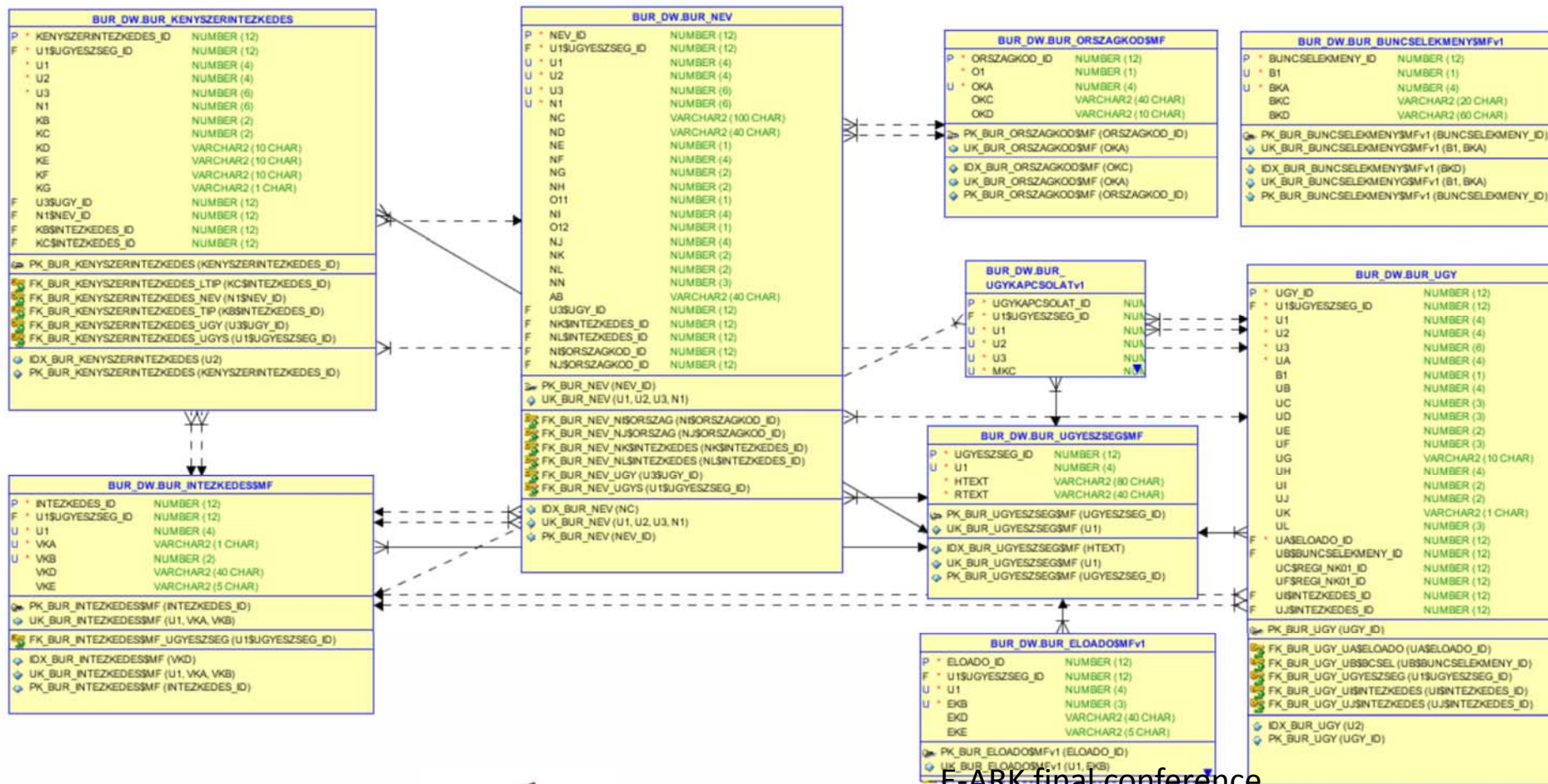


Querying Relational Database

Problems when querying data from normalized relational database:

- Multiple table must be joined, even for simply data analysis
- The table and column names hold the business information
- If foreign keys are missing, hard to identify the relation between the tables
- Complex queries can be slow
- The archivist must know the database structure

Archived Database Data Model



E-ARK final conference

Hungarian National Archives, Budapest

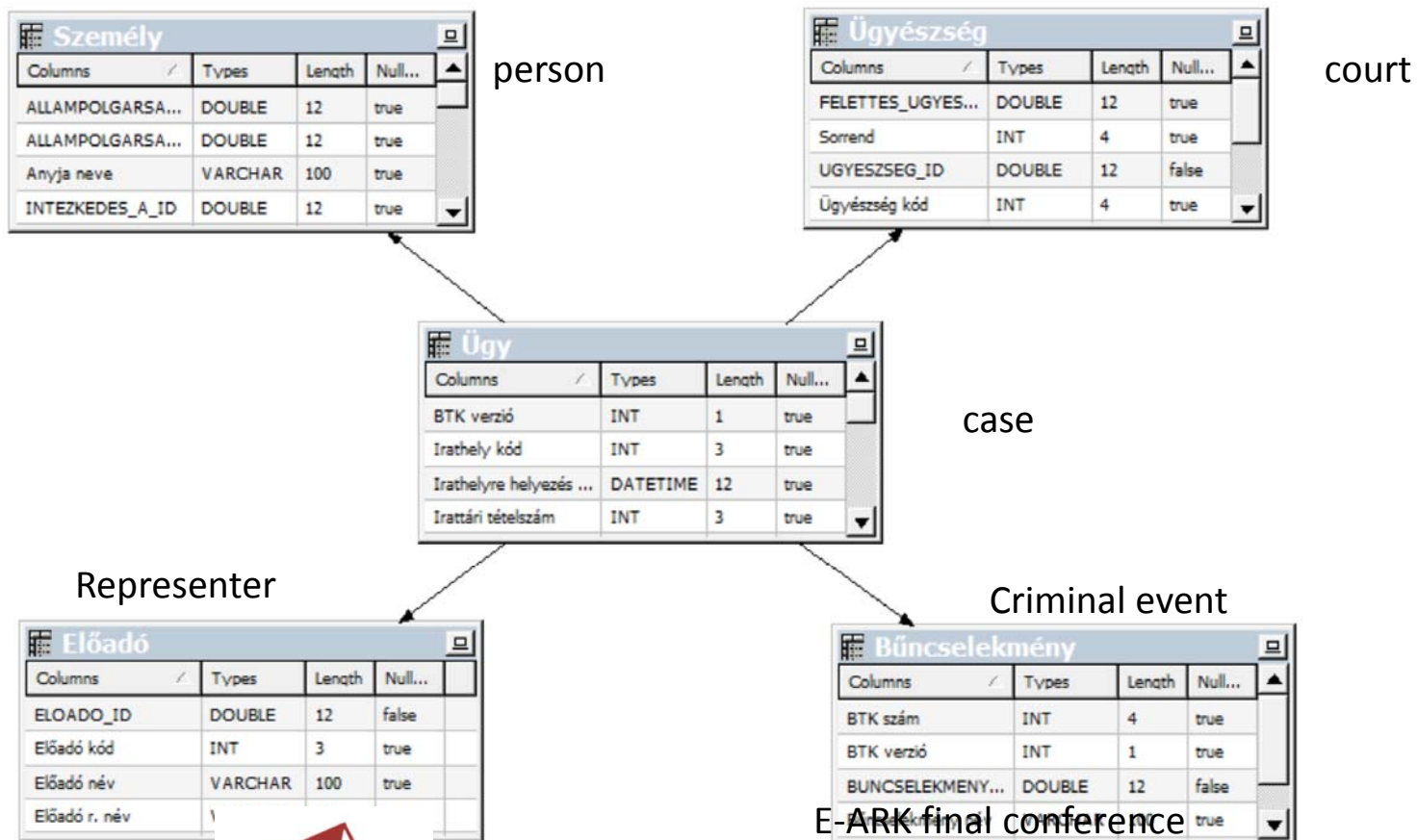
6-8 December 2016

Moving data closer to the analytical needs of the requester

The data can be analyzed more easily, when it is prepared, transformed, de-normalized:

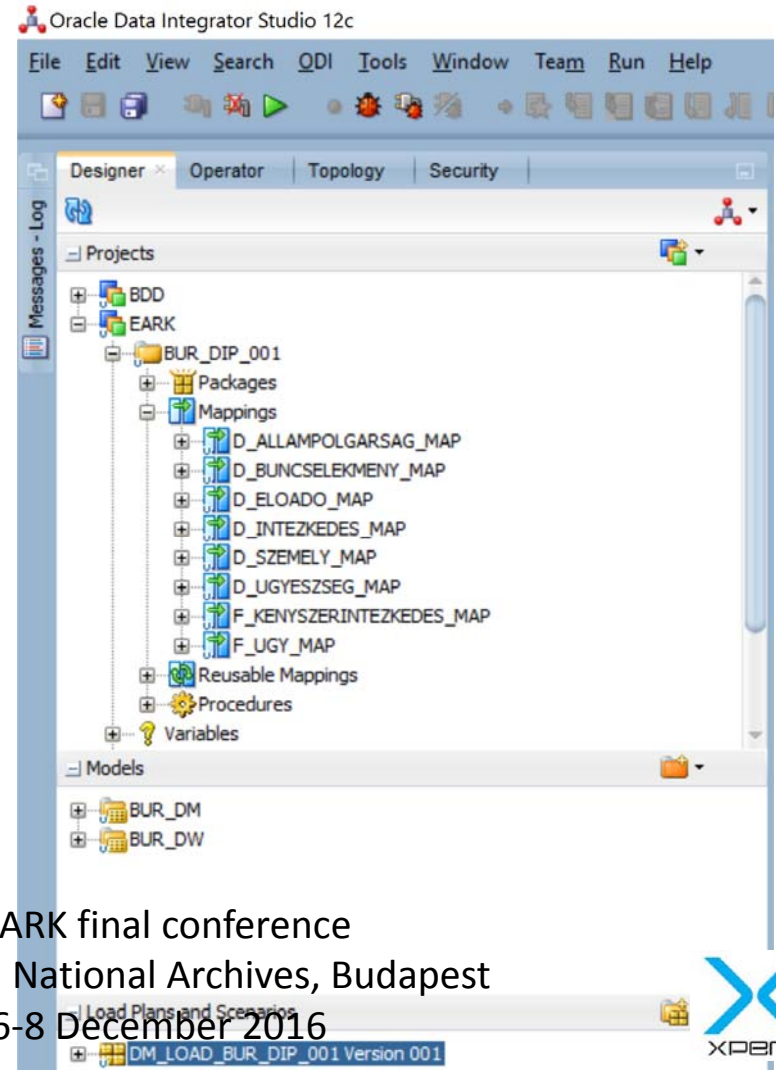
- Build data warehouse and data marts (simplified data model)
- Load and transform the data with ODI
- Create business model (hides the complexity of the database)
- Pre-create reports, analysis and dashboards

Simplified Business Model



Oracle Data Integrator

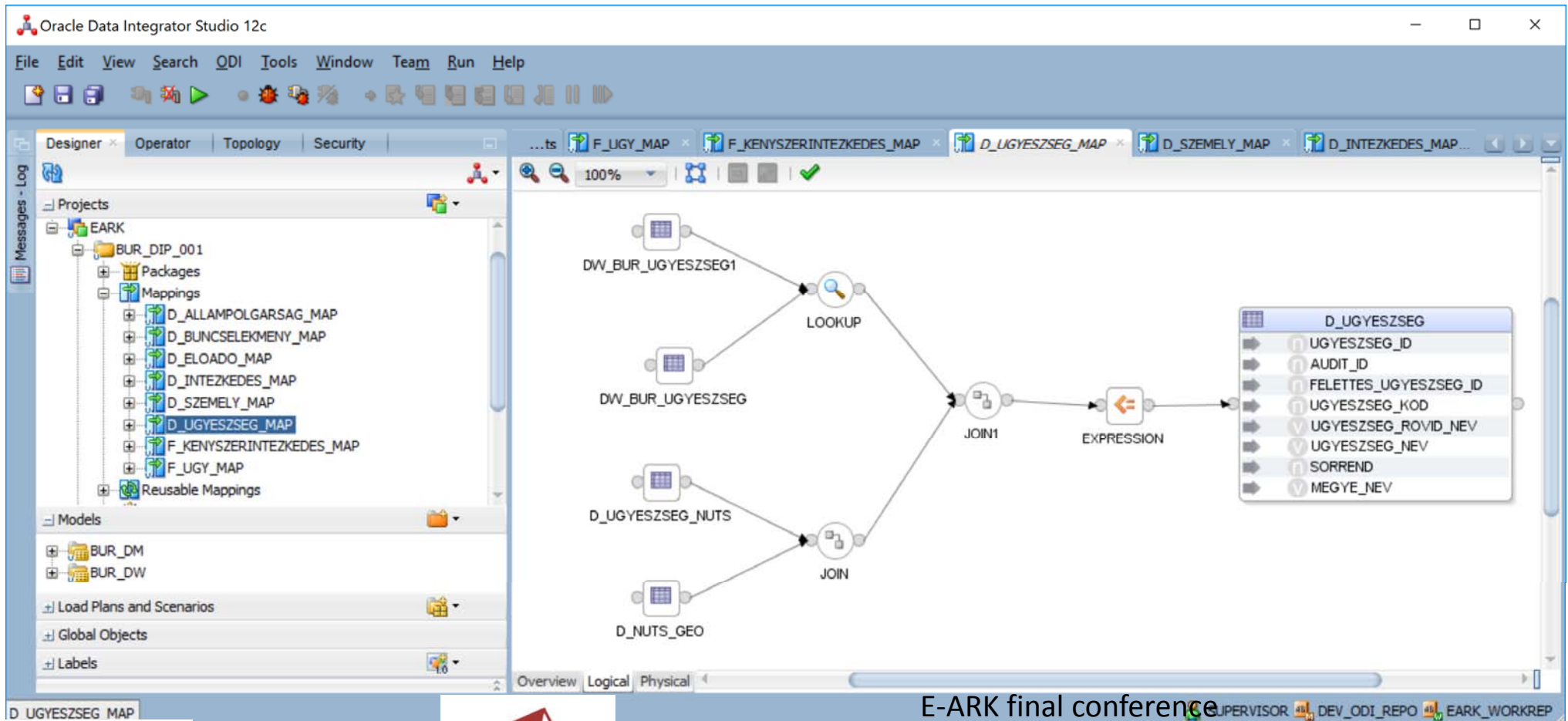
- comprehensive data integration platform
- declarative user interface
- flexible and high-performance architecture
- parallelism when executing data integration processes
- big data support



E-ARK final conference
Hungarian National Archives, Budapest
6-8 December 2016



Data transformation

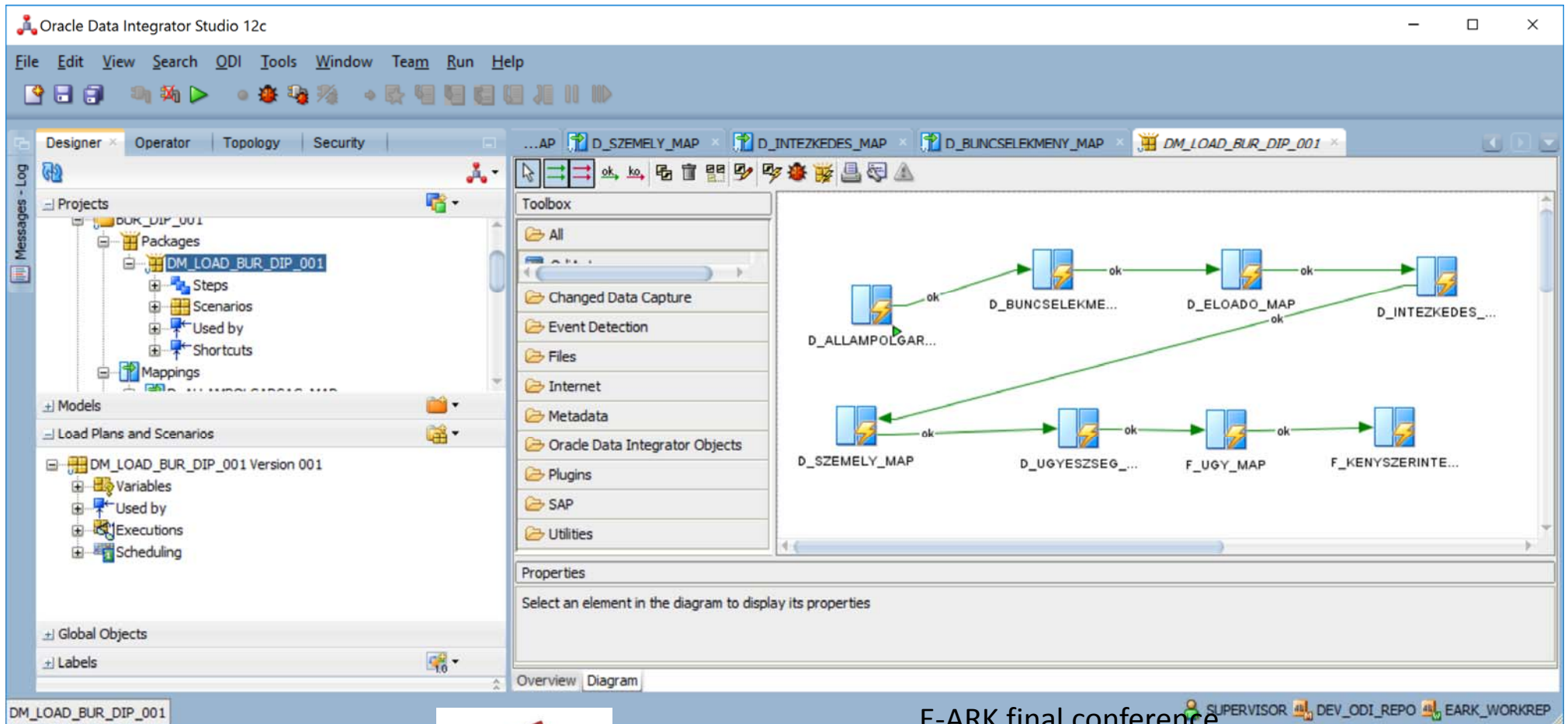


E-ARK final conference

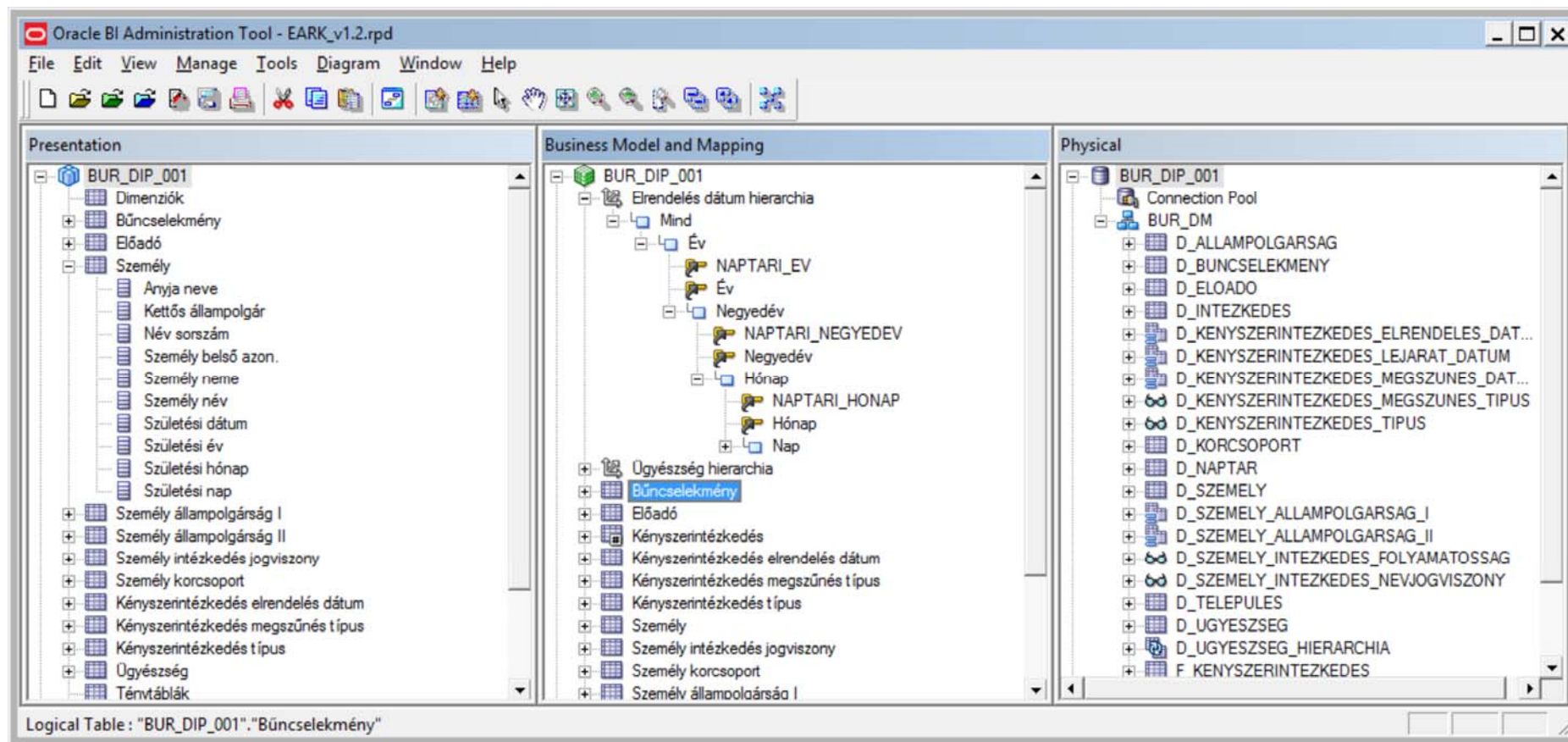
Hungarian National Archives, Budapest

6-8 December 2016

Data load



BI model



E-ARK final conference
Hungarian National Archives, Budapest
6-8 December 2016

Data Access with BI

BI Analysis and Interactive Reporting

- self-service, ad-hoc queries
- analysis, reports
- dashboards, dashboard prompts
- charts, maps
- drill down, hierarchy navigation
- built-in analytic functions
- data export into CSV, Excel

BI Catalog

ORACLE Business Intelligence






Search All

Catalog Home Catalog Favorites Dashboards

User View Location /Shared Folders/EARK/BUR/BUR

Folders My Folders Shared Folders EARK BUR BUR

Type All Sort Name A-Z Show More Details

 Kényszerintézkedés Last Modified 7/19/2016 2:58:05 PM Owner weblogic Open Edit More
 Korcsoport Last Modified 7/19/2016 2:58:05 PM Owner weblogic Open Edit More
 Ügyek Last Modified 7/19/2016 2:58:05 PM Owner weblogic Open Edit More
 Ügyek OLAP Last Modified 7/19/2016 2:58:05 PM Owner weblogic Open Edit More
 Ügyek térképen Last Modified 7/19/2016 2:58:05 PM Owner weblogic Open Edit More

BI Dashboard and Analysis

ORACLE Business Intelligence Search All [] Advanced Administration Help Sign Out

BUR Home Catalog Favorites Dashboards New Open Signed In As weblogic

Ügyek Korcsoport Kényszerintézkedés Ügyek térképen Ügyek OLAP

* Ügyszám év
 (All Column Values)
 1993
 1994
 1995
 1996
 1997
 1998
 1999

Apply Reset

* Büncselekmény r. név
 (All Column Val.
 Adócsalás
 Alk.rend er.meg
 Alkrend ell.szerv

Büncselekmények száma éves bontásban
 Time run: 8/29/2016 1:07:16 PM

Büncselekmény név	1993		1994		1995		1996		1997		1998		1999		Darab	%
	Darab	%	Darab	%	Darab	%	Darab	%	Darab	%	Darab	%	Darab	%		
A hulladékgazdálkodás rendjének megsértése									14	27.5%	24	47.1%	13	25.5%	51	100.0%
A polgári szolgálat megtagadása												3	100.0%	3	100.0%	
Adócsalás	30	1.1%	358	12.7%	224	7.9%	170	6.0%	382	13.5%	1019	36.1%	642	22.7%	2825	100.0%
Alkotmányos rend erőszakos megváltoztatása	1	25.0%	1	25.0%	2	50.0%									4	100.0%
Apartheid	1	33.3%	1	33.3%						1	33.3%				3	100.0%
Az alkotmányos rend elleni szervezkedés			2	100.0%											2	100.0%
Banktitok megsértése					1	20.0%			1	20.0%	2	40.0%	1	20.0%	5	100.0%
Becsületsértés			102	28.2%	71	19.6%	25	6.9%	44	14.9%	12	19.9%	31	8.6%	362	100.0%



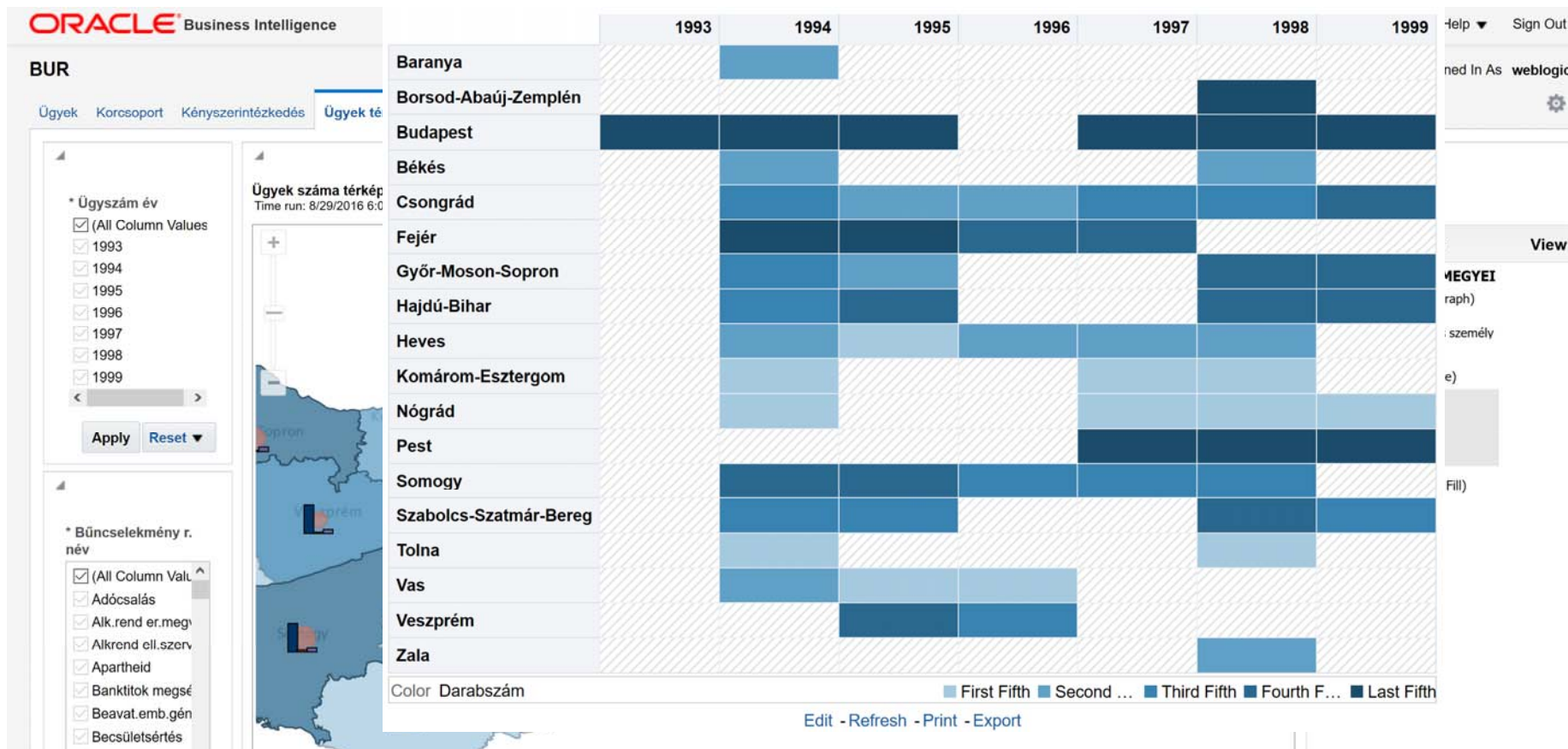
E-ARK final conference
 Hungarian National Archives, Budapest
 6-8 December 2016



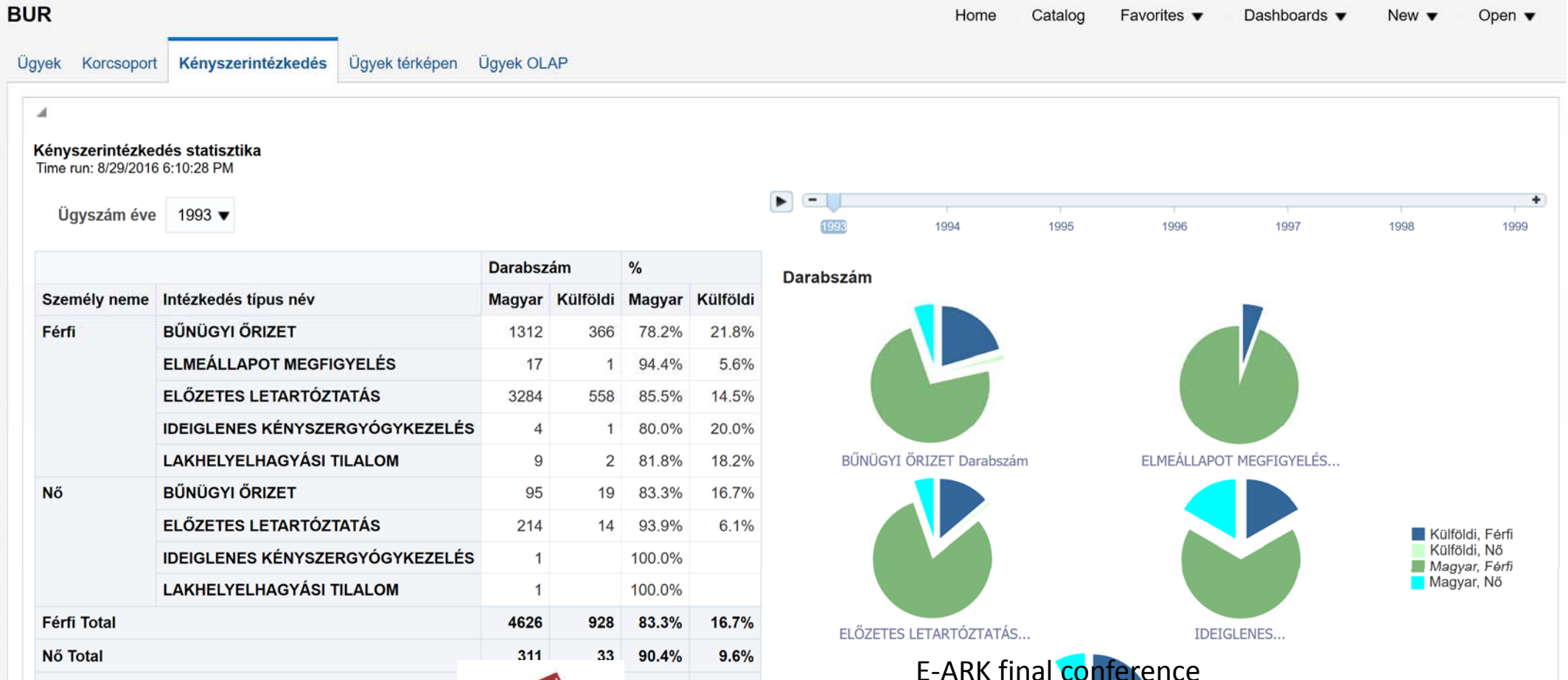
Distribution of criminal cases over age group and gender



Criminal cases on map



Interactive charts



E-ARK final conference
Hungarian National Archives, Budapest
6-8 December 2016



Data Access with Data Explorer

- User friendly web-based application (built with Oracle APEX)
- Register of data sources and data sets (tables, views, custom queries)
- Interactive reports on data sets
- Filter column data
- Aggregations, custom calculations, pivot
- Charts
- Data export in CSV

Data Explorer Interactive Report

Data Explorer admin ▾

Search: 1. Riport 1

- ▼ Saved Report = "Riport 1" ×
- Edit Pivot ×
- Buncselekmény Nev in 'Bűnpártolás, Csalás, Egyesülési és gyülekezési szabadság megsértése, Egyesülési joggal visszaélés' ×

	1993	1994	1995	1996	1997	1998	1999
Buncselekmény Nev	Sum Ugy Darab	Sum Ugy Darab	Sum Ugy Darab	Sum Ugy Darab	Sum Ugy Darab	Sum Ugy Darab	Sum Ugy Darab
Bűnpártolás	30	86	79	18	56	124	43
Csalás	3,533	10,527	5,896	1,827	3,503	8,628	3,217
Egyesülési joggal visszaélés	-	-	1	-	-	-	-
Egyesülési és gyülekezési szabadság megsértése		3				1	

Conclusions – Open Issues

- The requirements on presentation react to the requirements of the preservation.
 - What and how should be archived.
 - Data - in how many and what kind of representations should be archived?
 - Documentation - what kind of documentations should be attached?
 - How can we somehow „standardize” the archivation of very specific objects like OLAP cubes?
 - Which metadata should be used?
 - Emulation
 - Preservation planning

Thank You for your Attention!



E-ARK final conference
Hungarian National Archives, Budapest
6-8 December 2016

