Information Integration Systems
Chapter 2. Federated Databases (2: CSV, XLSx format)

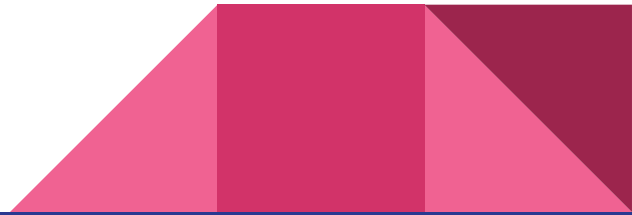
SIA & SDBIS

CSV and XLSx Data Source Access

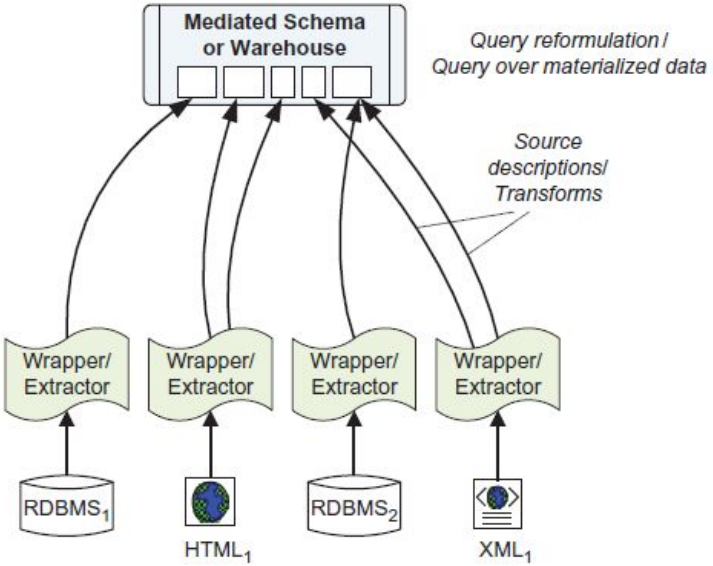
Local Data File Integration

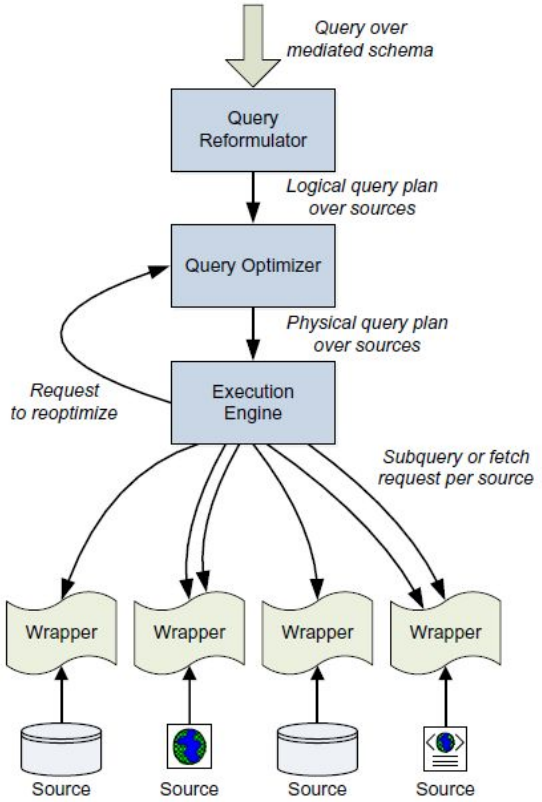
Federated Database Systems. Concepts (Review)

- Federation: “disparate data(bases) integration into a unified logical structure”.
- Federated MDBS
  - MDBS: Multidatabase System: “A distributed DBMS in which each site maintains complete autonomy”
  - F-MDBS: “... is a cross between distributed DBMS and centralized DBMS; it is a distributed system for global users and a centralized system for local users”.
- Federated Database: “The sources are independent, but one source can call on others to supply information.”



Data-based *perspective* [1, 10..14] (virtual database) (Review)

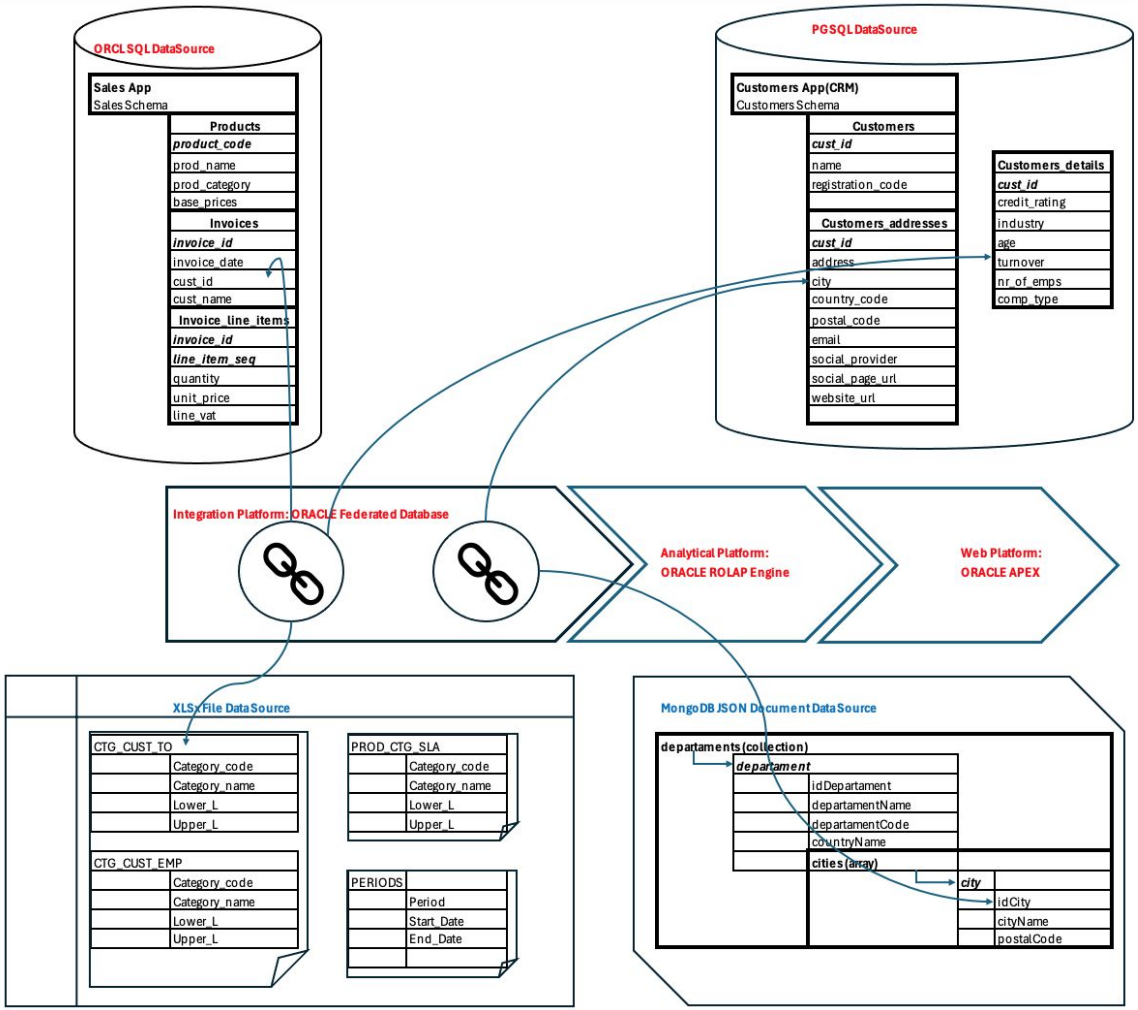




Case Study (Review)

- Data Sources
  - SQL: Oracle DB Database 12c/18c/19c/21c,
  - SQL: PostgreSQL 9/10/12,
  - **CSV/XLSx from Local FileSystem, Virtual File System (FTP)**
  - XML: Local FileSystem or (Web)REST Data Services
  - JSON: Local FileSystem or (Web)REST Data Services

- Data source Access Model
  - External Tables
  - Remote Views
    - SQL Remote Views
    - XML Remote Views
    - REST Remote Views
  - Local Tables (ETL)
- Integration Model
  - Consolidation Views
  - Analytical Views
- Web Model:
  - ORDS REST Views,
  - APEX Reports and Charts


ORCL SQL DataSource
Sales App
Sales Schema
Products
product_code
prod_name
prod_category
base_prices
Invoices
invoice_id
Invoice_date
cust_id
cust_name
Invoice_line_items
invoice_id
line_item_seq
quantity
unit_price
line_vat
PGSQL DataSource
Customers App(CRM)
Customers Schema
Customers
cust_id
name
registration_code
Customers_addresses
cust_id
address
city
country_code
postal_code
email
social_provider
social_page_url
website_url
Customers_details
cust_id
credit_rating
industry
age
turnover
nr_of_emps
comp_type
Integration Platform: ORACLE Federated Database
Analytical Platform:
ORACLE ROLAP Engine
Web Platform:
ORACLE APEX
XLSx File Data Source
CTG_CUST_TO
Category_code
Category_name
Lower_L
Upper_L
CTG_CUST_EMP
Category_code
Category_name
Lower_L
Upper_L
PROD_CTG_SLA
Category_code
Category_name
Lower_L
Upper_L
PERIODS
Period
Start_Date
End_Date
MongoDB JSON Document Data Source
departaments (collection)
departament
idDepartament
departamentName
departamentCode
countryName
cities (array)
city
idCity
cityName
postalCode

2.2 Architecture and components of FDB

- 2.2.1 Federated Database System Concept
- 2.2.2 Data Source Model and Access Components
- 2.2.3 Integration and Analytical Model
- 2.2.4 Integration Web Model

2.2.2 Data Source Model and Access Components

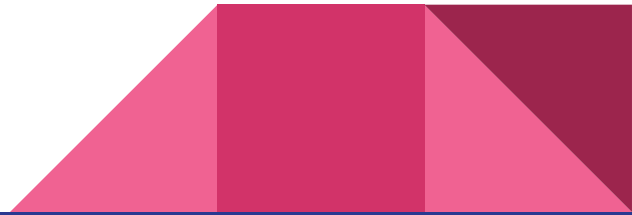
- SQL Data Source Access Components
- **CSV, XLS Data Source Access Components**
- XML, JSON Data Source Access Components

Data Source Access Components for External data files (local filesystem): CSV and XLSX

- Integration Strategy: access to tabular rows from CSV/XLSx local files.
  - Model|Schema Matching:
    - Tabular Doc Data Source -> Oracle.SQL FDB SQL Table Schema
- Oracle DB Infrastructure:
  - DIRECTORY objects,
  - Oracle Loader tool,
  - EXTERNAL TABLE objects,
  - SELECT TABLE clause/operator.
- PL/SQL Library to work with XLSX format:
  - ExcelTable.lib open source project (dependent_link1, depedent_link2).

Data Source Access Components for External data files: CSV

- Define DIRECTORY to access external files from disk.
  - Grant necessary privileges.
- Define EXTERNAL TABLE:
  - CREATE TABLE with ORGANIZATION EXTERNAL clause
  - using ORACLE_LOADER TYPE clause to invoke data loader tool
    - with column definition clause to implement type matching rules.
- Define local VIEWS on EXTERNAL TABLEs to implement operations:
  - Type matching
  - Table Schema matching
  - Data cleaning
  - Data filtering



Data Source Access Components for External data files: **CSV**

- CREATE TABLE statement version
- ORGANIZATION EXTERNAL main clause
  - TYPE subclause with variants:
    - ORACLE_LOADER
    - ORACLE_DATAPUMP to unload the load data from/to Oracle tables using Oracle Data Pump tool (as for expdp and impdp shell commands).
- ORACLE_LOADER parameters:
  - DEFAULT DIRECTORY and LOCATION
  - ACCESS PARAMETERS
    - RECORDS and FIELDS delimiters;
    - *Columns definitions*

Data Source Access Components for XLSX External File Data Source

- Defining DIRECTORY to access external files from disk.
  - Grant necessary privileges to federated schema owner.
- Install ExcelTable.library by (or download the complete kit from portal.feaa.uaic.ro):
  - Download ExcelTable.zip from https://github.com/mbleron/ExcelTable
  - Unzip and run *install.sql*.
- Defining external VIEWS using:
  - CREATE OR REPLACE VIEW command
  - TABLE clause to convert in SQL structure the output (DATASET) produced by:
    - ExcelTable.getRows() function from ExcelTable.lib
      - Having column-definition parameter to implement type matching rules
    - ExcelTable.getFile() function from ExcelTable.lib to read content from data file as BLOB.

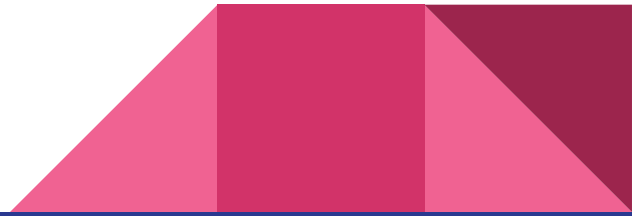
TABLE function/operator of SELECT phrase

- (Pipelined) Table Functions:
  - functions that return (nested) TABLE Type,
  - functions that return ANYDATASET Type,
  - functions that return SYS_REFCURSOR Type.
- Table functions could be invoked from SELECT-SQL using TABLE operator.

CASE STUDY: ORCL FDB CSV/Excel Table

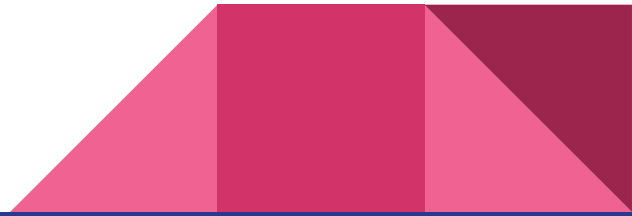
Case Study: Text data CSV file access

- SQL Script Example:
  - 23_AM_CSV_ExternalTable_View.sql
- External Data Source Files
  - 13_DS_CSV_CTG_CUST_TO.csv
  - 13_DS_CSV_CTG_CUST_EMP.csv
  - 13_DS_CSV_Periods.csv
- Internal Data Source Model: external tables
  - CTG_CUST_TO
  - CTG_CUST_EMP
  - Periods



Case Study: Office XLSx document access

- SQL Script Example:
  - 23_AM_XLS_ExcelTable_View.sql
- External Data Source Files
  - 13_DS_XLS_CustProdCateg.xlsx
- Internal Data Source Model: external views
  - CTG_CUST_TO_VIEW
  - CTG_CUST_EMP_VIEW
  - Periods_RAW_VIEW
  - Periods_VIEW



References [External CSV docs Access]

- Oracle-Base: oracle-external-table
- Oracle-base: XML external tables
- Oracle-Doc_1, Oracle-doc_2

References [XLSx Access]

- Oracle-Base: TABLE() function, REF CURSORs
- PL/SQL Excel Table Library
  - Doc: ExcelTable&blog
- Other PL/SQL XLS libraries:
  - XML_Spreadsheet
  - Alexandria PL/SQL
  - Ora_excel
  - XLSX_writer: tutorial
- Oracle-base Table Functions
- Oracle-doc Table Functions
- Oracle-doc ANYDATASET
- Oracle-base Ref-cursors