POZNAN UNIVERSITY OF TECHNOLOGY

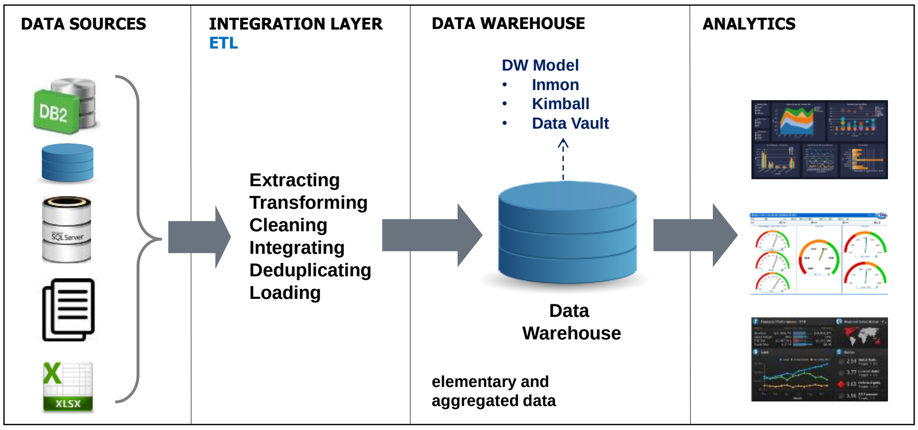
# Traditional Data Warehouse Architectures

Robert Wrembel
Poznan University of Technology
Institute of Computing Science
Poznań, Poland
Robert.Wrembel@cs.put.poznan.pl
www.cs.put.poznan.pl/rwrembel

## Outline

- ➲ Data Warehouse architectures
- ➲ Data integration and loading: ETL vs. ELT

# DW Architecture 1 (basic)



DATA SOURCES

INTEGRATION LAYER
ETL

Extracting
Transforming
Cleaning
Integrating
Deduplicating
Loading

DATA WAREHOUSE

DW Model
- Inmon
- Kimball
- Data Vault

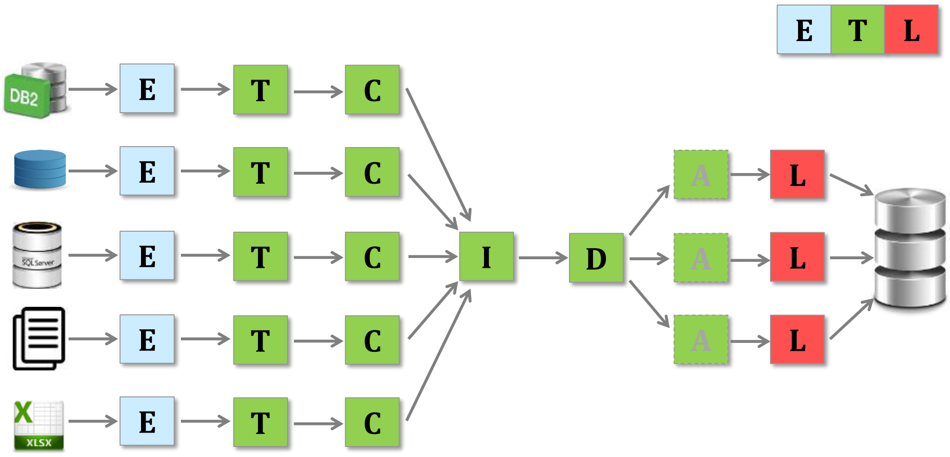
Data
Warehouse

elementary and
aggregated data

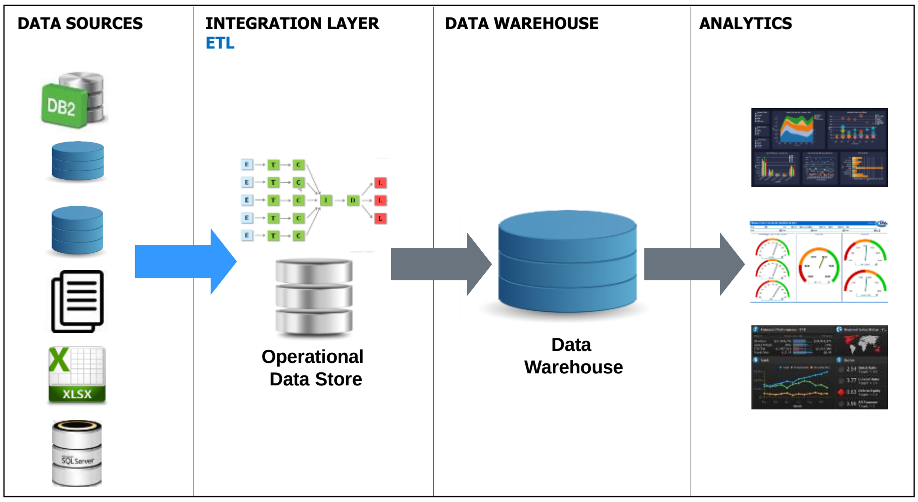
ANALYTICS

typically OLTP data sources

# ETL workflow



# DW Architecture 2



DATA SOURCES

INTEGRATION LAYER
ETL

DATA WAREHOUSE

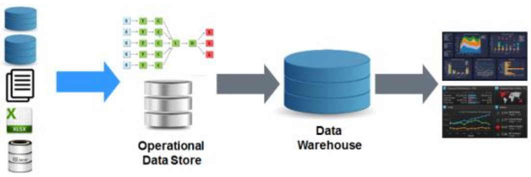
ANALYTICS

Operational
Data Store

Data
Warehouse

# Operational Data Store

- ➲ = Staging Area = (Data) Stage
- ➲ A repository for an ETL engine
- ➲ To separate normal processing at DSs from data ingest
  - to separate transactional from batch processing
- Disk storage for processing large data volumes that will not fit in RAM
- To provide means for data provenance

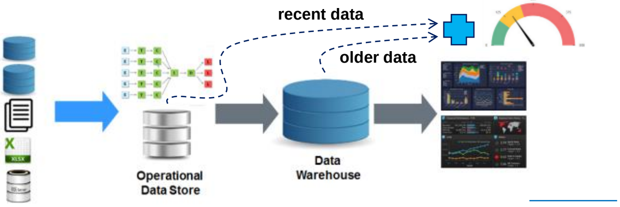


Operational
Data Store

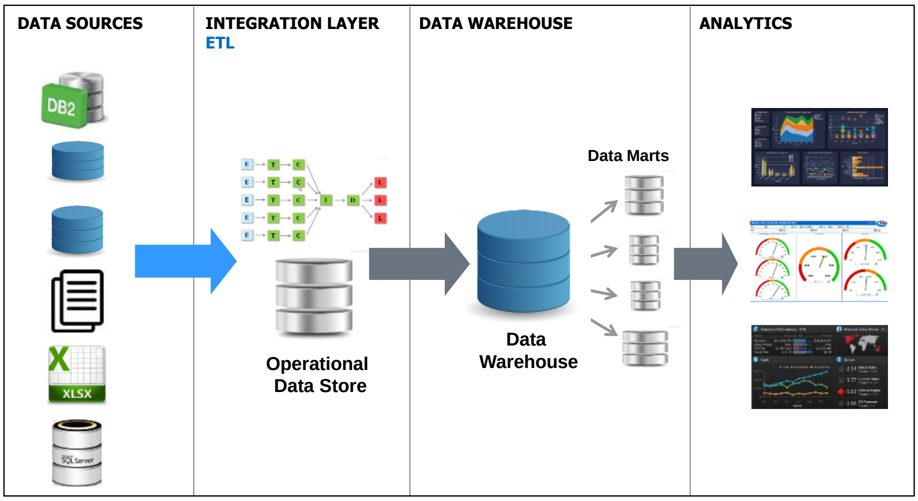
Data
Warehouse

# Operational Data Store

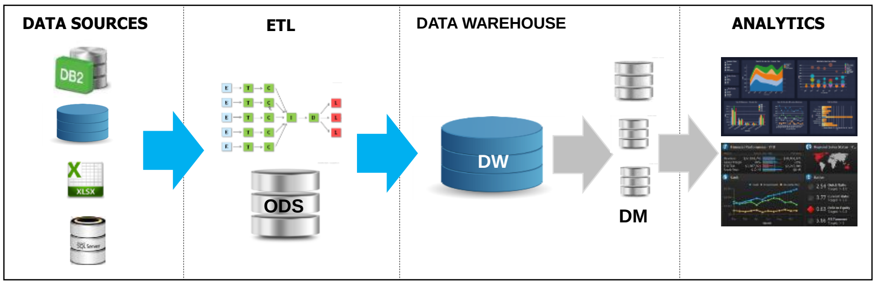
- ➲ **To store intermediate results → to be shared (used) by multiple ETL tasks**
  - **re-using the same result datasets by multiple processes (optimization)**
  - **for recovery after crash of an ETL process**
    - **re-executing a stopped process from a failed phase**
- ➲ **Recent data can be accessed before a DW is refreshed**
- ➲ **Implementation**
  - **database**
  - **(distributed) file system**



# DW Architecture 3



# DW Architecture (cd.)

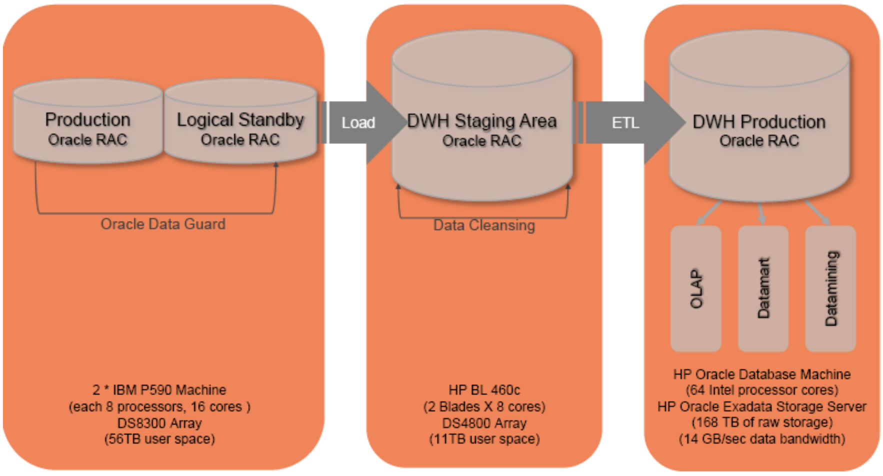


single server or cluster

main memory appliance
(superserver)

single dedicated server
(typically for MOLAP)
or
the same server as for DW

# Allegro DW



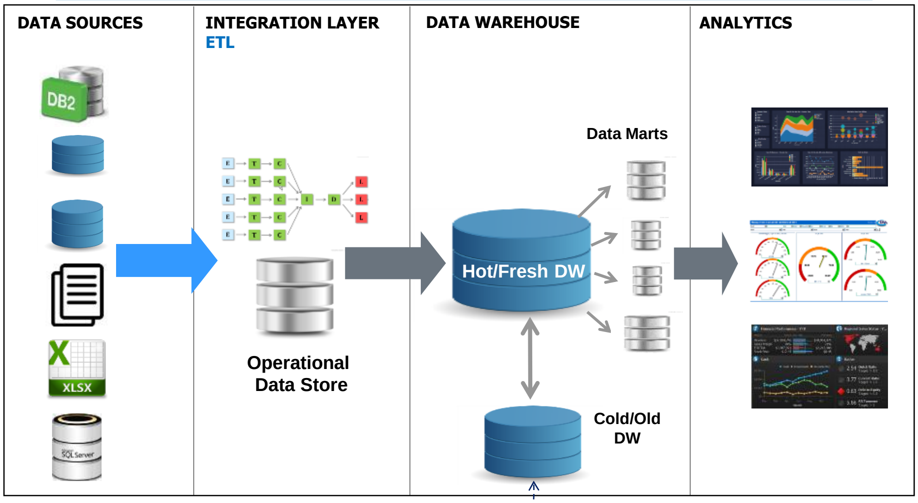
2 * IBM P590 Machine
(each 8 processors, 16 cores )
DS8300 Array
(56TB user space)

HP BL 460c
(2 Blades X 8 cores)
DS4800 Array
(11TB user space)

HP Oracle Database Machine
(64 Intel processor cores)
HP Oracle Exadata Storage Server
(168 TB of raw storage)
(14 GB/sec data bandwidth)

**C. Maar, R. Kudliński: Allegro on the way from XLS based controlling to a modern BI environment. National conference on Data Warehousing and Business Intelligence, Warsaw, 2008**

# DW Architecture 4



DATA SOURCES

INTEGRATION LAYER
ETL

DATA WAREHOUSE

ANALYTICS

Operational Data Store

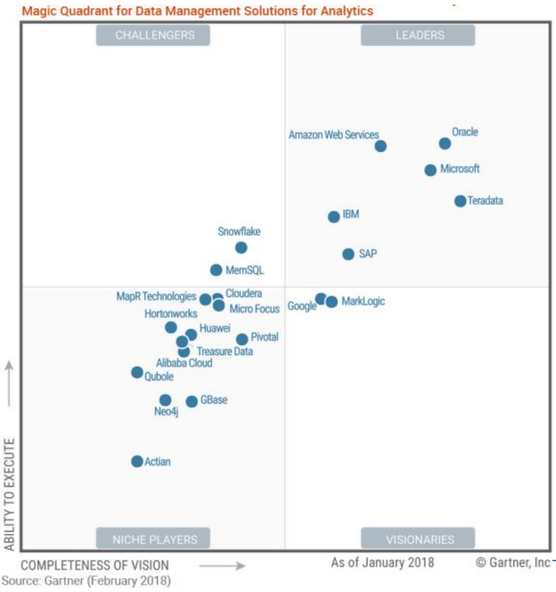
Hot/Fresh DW

Data Marts

Cold/Old DW

data older than n years
large pool of HDDs
small RAM

# Gartner Report: DW servers



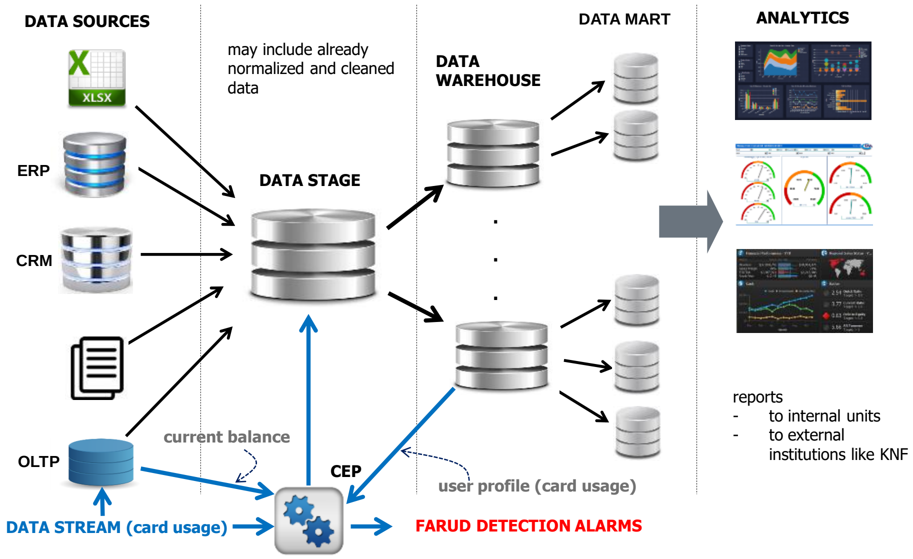
Magic Quadrant for Data Management Solutions for Analytics

Source: Gartner (February 2018)

# Large DW Architectures

- # data sources: 100 - 200
- Fact table: nn * $10^9$ rows
- Fact table: n TB
- Multiple relational DWs in an organization
  - DW size: nn TB
- Multiple data marts
- n * $10^3$ to nnn * $10^3$ ETL workflows
- DW composed of 100+ tables
  - on average 50+ attributes/table

# DW in Bank



DATA SOURCES

may include already normalized and cleaned data

DATA STAGE

DATA WAREHOUSE

DATA MART

ANALYTICS

current balance

CEP

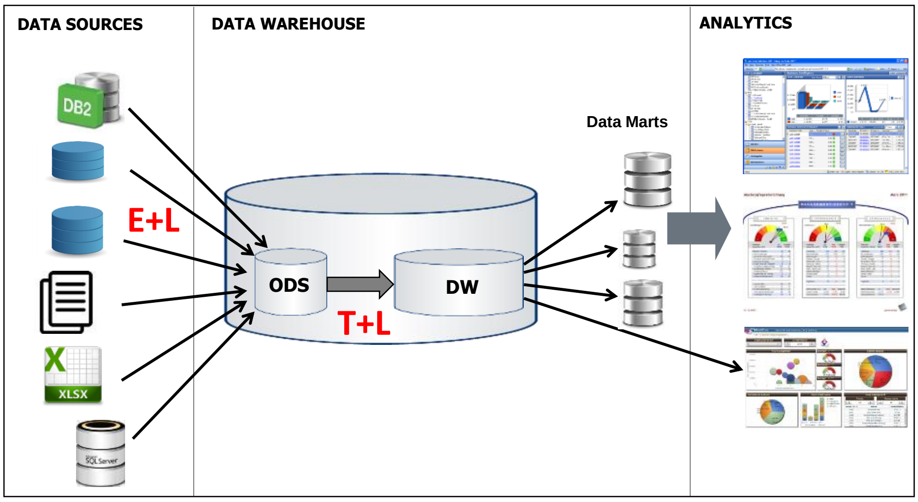
user profile (card usage)

DATA STREAM (card usage)

FARUD DETECTION ALARMS

reports
- to internal units
- to external institutions like KNF

# DW Architecture 5: ELT (ELTL)



# ELT Architecture

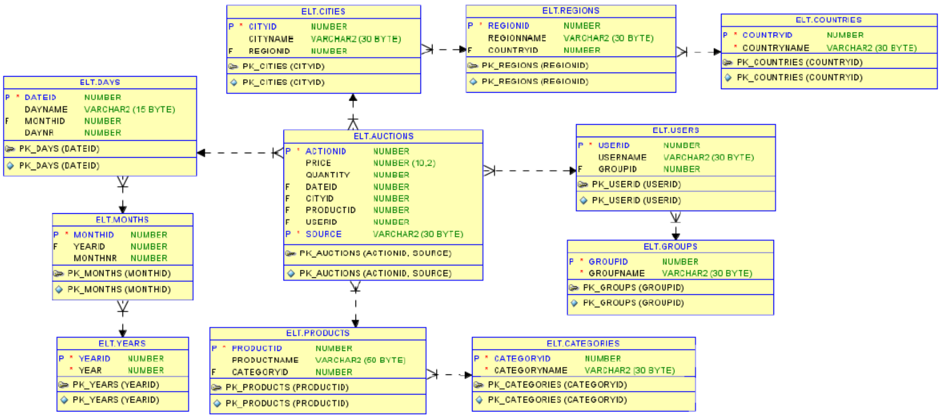
- Performance
  - data stored in a DB ⇨ processing by means of: SQL, PL/SQL, SQL PL, Transact SQL
  - data processed in a DB buffer cache ⇨ native DB environment
  - advanced query optimization offered by DBMS
  - single server for ELT and HD ⇨ heavier workload
- Functionality
  - data provenance
  - drill through
- Costs
  - single DW server
  - less software licences (OS, DBMS)

# ETL vs. ELT (experiment 1)

- Data sources
  - topic: Internet auctions
  - storage:
    - Oracle11g (Object-Relational model)
    - MySQL
    - PostgreSQL
    - XML
- Data warehouse: Oracle11g

# ETL vs. ELT (experiment 1)

- DW schema



ELT.CITIES
P * CITYID NUMBER
CITYNAME VARCHAR2 (30 BYTE)
F REGIONID NUMBER
PK_CITIES (CITYID)
PK_CITIES (CITYID)

ELT.REGIONS
P * REGIONID NUMBER
REGIONNAME VARCHAR2 (30 BYTE)
F COUNTRYID NUMBER
PK_REGIONS (REGIONID)
PK_REGIONS (REGIONID)

ELT.COUNTRIES
P * COUNTRYID NUMBER
* COUNTRYNAME VARCHAR2 (30 BYTE)
PK_COUNTRIES (COUNTRYID)
PK_COUNTRIES (COUNTRYID)

ELT.DAYS
P * DATEID NUMBER
DAYNAME VARCHAR2 (15 BYTE)
F MONTHID NUMBER
DAYNR NUMBER
PK_DAYS (DATEID)
PK_DAYS (DATEID)

ELT.AUCTIONS
P * ACTIONID NUMBER
PRICE NUMBER (10,2)
QUANTITY NUMBER
F DATEID NUMBER
F CITYID NUMBER
F PRODUCTID NUMBER
F USERID NUMBER
P * SOURCE VARCHAR2 (30 BYTE)
PK_AUCTIONS (ACTIONID, SOURCE)
PK_AUCTIONS (ACTIONID, SOURCE)

ELT.USERS
P * USERID NUMBER
USERNAME VARCHAR2 (30 BYTE)
F GROUPID NUMBER
PK_USERID (USERID)
PK_USERID (USERID)

ELT.MONTHS
P * MONTHID NUMBER
F YEARID NUMBER
MONTHNR NUMBER
PK_MONTHS (MONTHID)
PK_MONTHS (MONTHID)

ELT.GROUPS
P * GROUPID NUMBER
* GROUPNAME VARCHAR2 (30 BYTE)
PK_GROUPS (GROUPID)
PK_GROUPS (GROUPID)

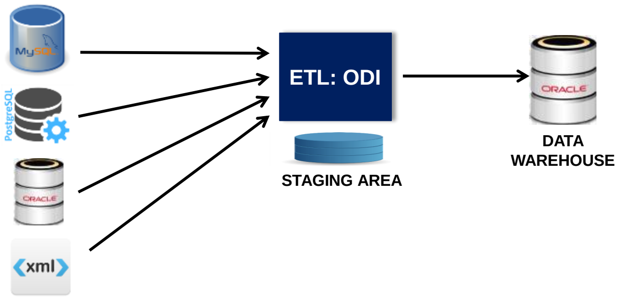
ELT.YEARS
P * YEARID NUMBER
* YEAR NUMBER
PK_YEARS (YEARID)
PK_YEARS (YEARID)

ELT.PRODUCTS
P * PRODUCTID NUMBER
PRODUCTNAME VARCHAR2 (50 BYTE)
F CATEGORYID NUMBER
PK_PRODUCTS (PRODUCTID)
PK_PRODUCTS (PRODUCTID)

ELT.CATEGORIES
P * CATEGORYID NUMBER
* CATEGORYNAME VARCHAR2 (30 BYTE)
PK_CATEGORIES (CATEGORYID)
PK_CATEGORIES (CATEGORYID)

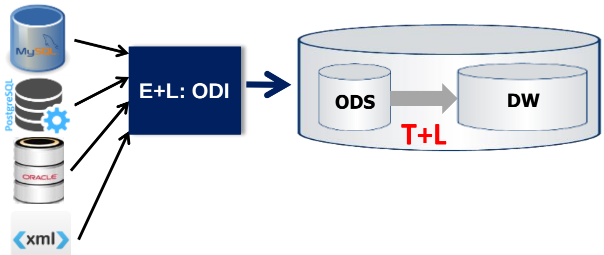
# ETL vs. ELT (experiment 1)

- Transformations of data for:
  - dimensions
  - fact table
- ETL architecture ⇨ Oracle Data Integrator (ODI)
  - ETL in a staging area on a separate server

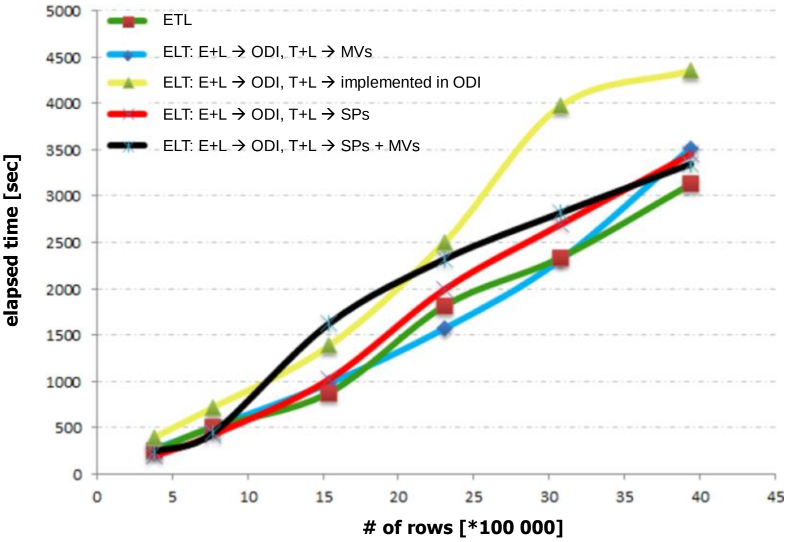


# ETL vs. ELT (experiment 1)

- ELT architecture
  - T+L in a staging area on the same server as a DW
  - variant 1: E+L → ODI, T+L → implemented in ODI
  - variant 2: E+L → ODI, T+L → implemented as materialized views (MVs)
  - variant 3: E+L → ODI, T+L → implemented as stored packages (SPs)
  - variant 4: E+L → ODI, T+L → SPs + MVs

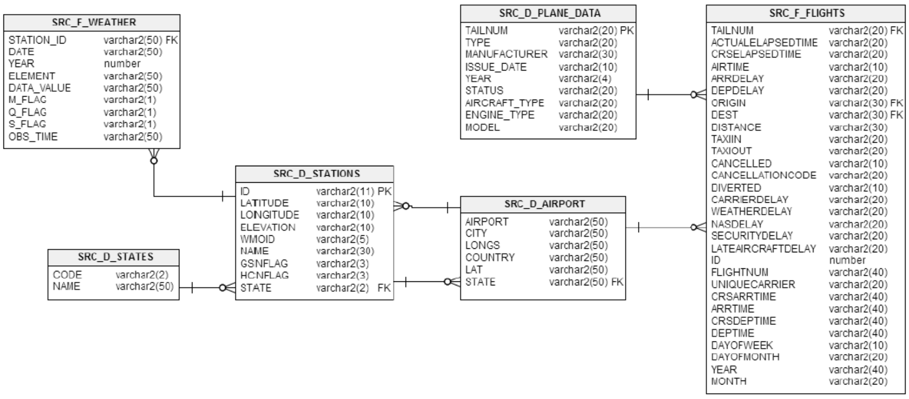


# ETL vs. ELT (experiment 1)



# ETL vs. ELT (experiment 2)

- **Data source**
  - **flight and weather data in the US, from 1986 until 2008**
  - **6 tables in Oracle11g**
- **Data warehouse: Oracle11g**
- **ETL/ELT: Informatica**



| SRC_F_WEATHER | |
|---|---|
| STATION_ID | varchar2(50) FK |
| DATE | varchar2(50) |
| YEAR | number |
| ELEMENT | varchar2(50) |
| DATA_VALUE | varchar2(50) |
| M_FLAG | varchar2(1) |
| Q_FLAG | varchar2(1) |
| S_FLAG | varchar2(1) |
| OBS_TIME | varchar2(50) |

| SRC_D_STATES | |
|---|---|
| CODE | varchar2(2) |
| NAME | varchar2(50) |

| SRC_D_STATIONS | |
|---|---|
| ID | varchar2(11) PK |
| LATITUDE | varchar2(10) |
| LONGITUDE | varchar2(10) |
| ELEVATION | varchar2(10) |
| WMOID | varchar2(5) |
| NAME | varchar2(30) |
| GSNFLAG | varchar2(3) |
| HCNFLAG | varchar2(3) |
| STATE | varchar2(2) FK |

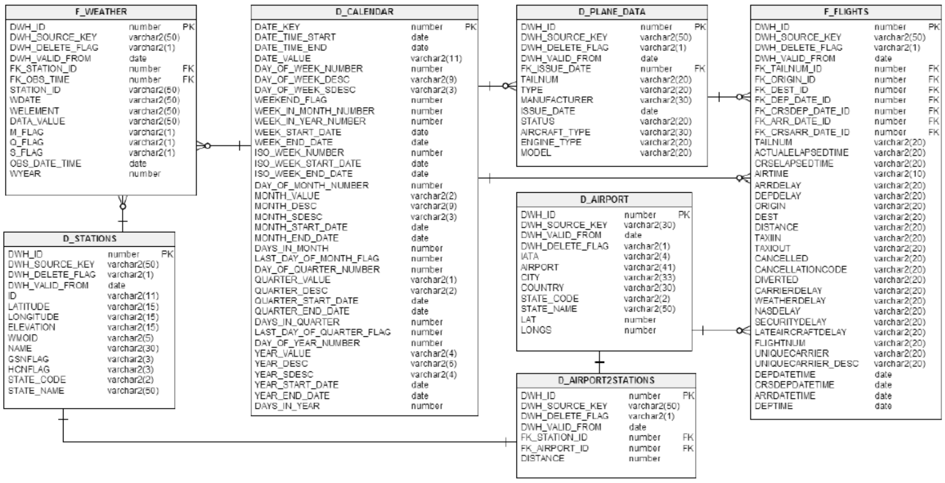
| SRC_D_PLANE_DATA | |
|---|---|
| TAILNUM | varchar2(20) PK |
| TYPE | varchar2(20) |
| MANUFACTURER | varchar2(30) |
| ISSUE_DATE | varchar2(10) |
| YEAR | varchar2(4) |
| STATUS | varchar2(20) |
| AIRCRAFT_TYPE | varchar2(20) |
| ENGINE_TYPE | varchar2(20) |
| MODEL | varchar2(20) |

| SRC_D_AIRPORT | |
|---|---|
| AIRPORT | varchar2(50) |
| CITY | varchar2(50) |
| LONGS | varchar2(50) |
| COUNTRY | varchar2(50) |
| LAT | varchar2(50) |
| STATE | varchar2(50) FK |

| SRC_F_FLIGHTS | |
|---|---|
| TAILNUM | varchar2(20) FK |
| ACTUALELAPSEDTIME | varchar2(20) |
| CRSELAPSEDTIME | varchar2(20) |
| AIRTIME | varchar2(10) |
| ARRDELAY | varchar2(20) |
| DEPDELAY | varchar2(20) |
| ORIGIN | varchar2(30) FK |
| DEST | varchar2(30) FK |
| DISTANCE | varchar2(30) |
| TAXIIN | varchar2(20) |
| TAXIOUT | varchar2(20) |
| CANCELLED | varchar2(10) |
| CANCELLATIONCODE | varchar2(20) |
| DIVERTED | varchar2(10) |
| CARRIERDELAY | varchar2(20) |
| WEATHERDELAY | varchar2(20) |
| NASDELAY | varchar2(20) |
| SECURITYDELAY | varchar2(20) |
| LATEAIRCRAFTDELAY | varchar2(20) |
| ID | number |
| FLIGHTNUM | varchar2(40) |
| UNIQUECARRIER | varchar2(20) |
| CRSARRTIME | varchar2(40) |
| ARRTIME | varchar2(40) |
| CRSDEPTIME | varchar2(40) |
| DEPTIME | varchar2(40) |
| DAYOFWEEK | varchar2(10) |
| DAYOFMONTH | varchar2(20) |
| YEAR | varchar2(40) |
| MONTH | varchar2(20) |

# ETL vs. ELT (experiment 2)

- ➲ DW schema (augmented with: calendar, airplane data, airport data)



| F_WEATHER | | |
|---|---|---|
| DWH_ID | number | PK |
| DWH_SOURCE_KEY | varchar2(50) | |
| DWH_DELETE_FLAG | varchar2(1) | |
| DWH_VALID_FROM | date | |
| FK_STATION_ID | number | FK |
| FK_OBS_TIME | number | FK |
| STATION_ID | varchar2(50) | |
| WDATE | varchar2(50) | |
| WELEMENT | varchar2(50) | |
| DATA_VALUE | varchar2(50) | |
| M_FLAG | varchar2(1) | |
| Q_FLAG | varchar2(1) | |
| S_FLAG | varchar2(1) | |
| OBS_DATE_TIME | date | |
| WYEAR | number | |

| D_STATIONS | | |
|---|---|---|
| DWH_ID | number | PK |
| DWH_SOURCE_KEY | varchar2(50) | |
| DWH_DELETE_FLAG | varchar2(1) | |
| DWH_VALID_FROM | date | |
| ID | varchar2(11) | |
| LATITUDE | varchar2(15) | |
| LONGITUDE | varchar2(15) | |
| ELEVATION | varchar2(15) | |
| WMOID | varchar2(5) | |
| NAME | varchar2(30) | |
| GSNFLAG | varchar2(3) | |
| HCNFLAG | varchar2(3) | |
| STATE_CODE | varchar2(2) | |
| STATE_NAME | varchar2(50) | |

| D_CALENDAR | | |
|---|---|---|
| DATE_KEY | number | PK |
| DATE_TIME_START | date | |
| DATE_TIME_END | date | |
| DATE_VALUE | varchar2(11) | |
| DAY_OF_WEEK_NUMBER | number | |
| DAY_OF_WEEK_DESC | varchar2(9) | |
| DAY_OF_WEEK_SDESC | varchar2(3) | |
| WEEKEND_FLAG | number | |
| WEEK_IN_MONTH_NUMBER | number | |
| WEEK_IN_YEAR_NUMBER | number | |
| WEEK_START_DATE | date | |
| WEEK_END_DATE | date | |
| ISO_WEEK_NUMBER | number | |
| ISO_WEEK_START_DATE | date | |
| ISO_WEEK_END_DATE | date | |
| DAY_OF_MONTH_NUMBER | number | |
| MONTH_VALUE | varchar2(2) | |
| MONTH_DESC | varchar2(9) | |
| MONTH_SDESC | varchar2(3) | |
| MONTH_START_DATE | date | |
| MONTH_END_DATE | date | |
| DAYS_IN_MONTH | number | |
| LAST_DAY_OF_MONTH_FLAG | number | |
| DAY_OF_QUARTER_NUMBER | number | |
| QUARTER_VALUE | varchar2(1) | |
| QUARTER_DESC | varchar2(2) | |
| QUARTER_START_DATE | date | |
| QUARTER_END_DATE | date | |
| DAYS_IN_QUARTER | number | |
| LAST_DAY_OF_QUARTER_FLAG | number | |
| DAY_OF_YEAR_NUMBER | number | |
| YEAR_VALUE | varchar2(4) | |
| YEAR_DESC | varchar2(6) | |
| YEAR_SDESC | varchar2(4) | |
| YEAR_START_DATE | date | |
| YEAR_END_DATE | date | |
| DAYS_IN_YEAR | number | |

| D_PLANE_DATA | | |
|---|---|---|
| DWH_ID | number | PK |
| DWH_SOURCE_KEY | varchar2(50) | |
| DWH_DELETE_FLAG | varchar2(1) | |
| DWH_VALID_FROM | date | |
| FK_ISSUE_DATE | number | FK |
| TAILNUM | varchar2(20) | |
| TYPE | varchar2(20) | |
| MANUFACTURER | varchar2(30) | |
| ISSUE_DATE | date | |
| STATUS | varchar2(20) | |
| AIRCRAFT_TYPE | varchar2(30) | |
| ENGINE_TYPE | varchar2(20) | |
| MODEL | varchar2(20) | |

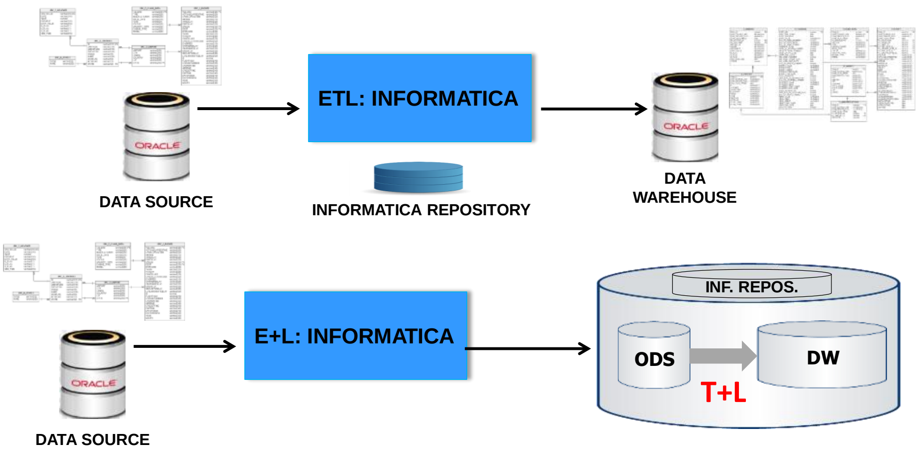
| D_AIRPORT | | |
|---|---|---|
| DWH_ID | number | PK |
| DWH_SOURCE_KEY | varchar2(30) | |
| DWH_VALID_FROM | date | |
| DWH_DELETE_FLAG | varchar2(1) | |
| IATA | varchar2(4) | |
| AIRPORT | varchar2(41) | |
| CITY | varchar2(33) | |
| COUNTRY | varchar2(30) | |
| STATE_CODE | varchar2(2) | |
| STATE_NAME | varchar2(50) | |
| LAT | number | |
| LONGS | number | |

| D_AIRPORT2STATIONS | | |
|---|---|---|
| DWH_ID | number | PK |
| DWH_SOURCE_KEY | varchar2(50) | |
| DWH_DELETE_FLAG | varchar2(1) | |
| DWH_VALID_FROM | date | |
| FK_STATION_ID | number | FK |
| FK_AIRPORT_ID | number | FK |
| DISTANCE | number | |

| F_FLIGHTS | | |
|---|---|---|
| DWH_ID | number | PK |
| DWH_SOURCE_KEY | varchar2(50) | |
| DWH_DELETE_FLAG | varchar2(1) | |
| DWH_VALID_FROM | date | |
| FK_TAILNUM_ID | number | FK |
| FK_ORIGIN_ID | number | FK |
| FK_DEST_ID | number | FK |
| FK_DEP_DATE_ID | number | FK |
| FK_CRSDEP_DATE_ID | number | FK |
| FK_ARR_DATE_ID | number | FK |
| FK_CRSARR_DATE_ID | number | FK |
| TAILNUM | varchar2(20) | |
| ACTUALELAPSEDTIME | varchar2(20) | |
| CRSELAPSEDTIME | varchar2(20) | |
| AIRTIME | varchar2(10) | |
| ARRDELAY | varchar2(20) | |
| DEPDELAY | varchar2(20) | |
| ORIGIN | varchar2(20) | |
| DEST | varchar2(20) | |
| DISTANCE | varchar2(20) | |
| TAXIIN | varchar2(20) | |
| TAXIOUT | varchar2(20) | |
| CANCELLED | varchar2(20) | |
| CANCELLATIONCODE | varchar2(20) | |
| DIVERTED | varchar2(20) | |
| CARRIERDELAY | varchar2(20) | |
| WEATHERDELAY | varchar2(20) | |
| NASDELAY | varchar2(20) | |
| SECURITYDELAY | varchar2(20) | |
| LATEAIRCRAFTDELAY | varchar2(20) | |
| FLIGHTNUM | varchar2(20) | |
| UNIQUECARRIER | varchar2(20) | |
| UNIQUECARRIER_DESC | varchar2(20) | |
| DEPDATETIME | date | |
| CRSDEPDATETIME | date | |
| ARRDATETIME | date | |
| DEPTIME | date | |

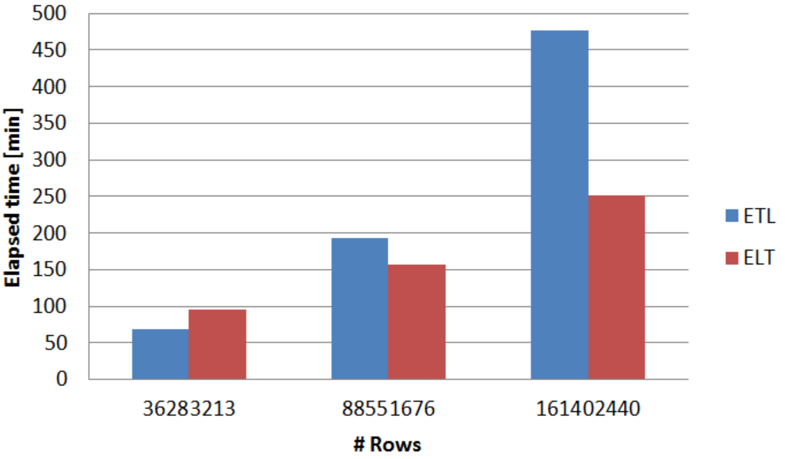
# ETL vs. ELT (experiment 2)

- ➲ ETL ⇨ Informatica
- ➲ ELT ⇨ Informatica (E+L), DB views (T+L)

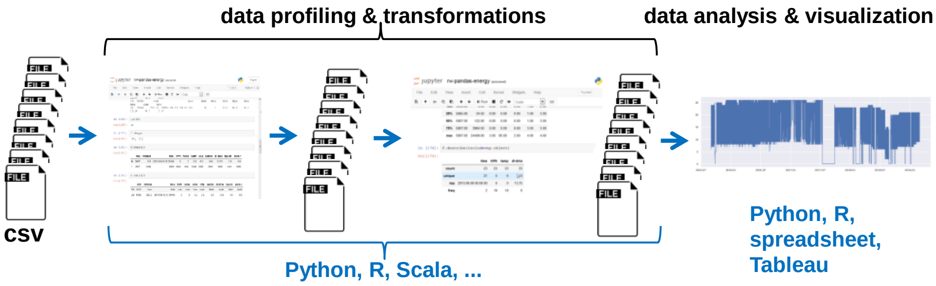


# ETL vs. ELT (experiment 2)

- ETL ⇨ Informatica
- ELT ⇨ Informatica (E+L), DB views (T+L)



# Architecture for Data Science



data profiling & transformations

data analysis & visualization

CSV

Python, R, Scala, ...

Python, R, spreadsheet, Tableau

- Data stored in files
  - performance problem
  - no backup & recovery
  - no access control
- Data & code sharing is difficult
  - re-usability problem
  - low programming productivity