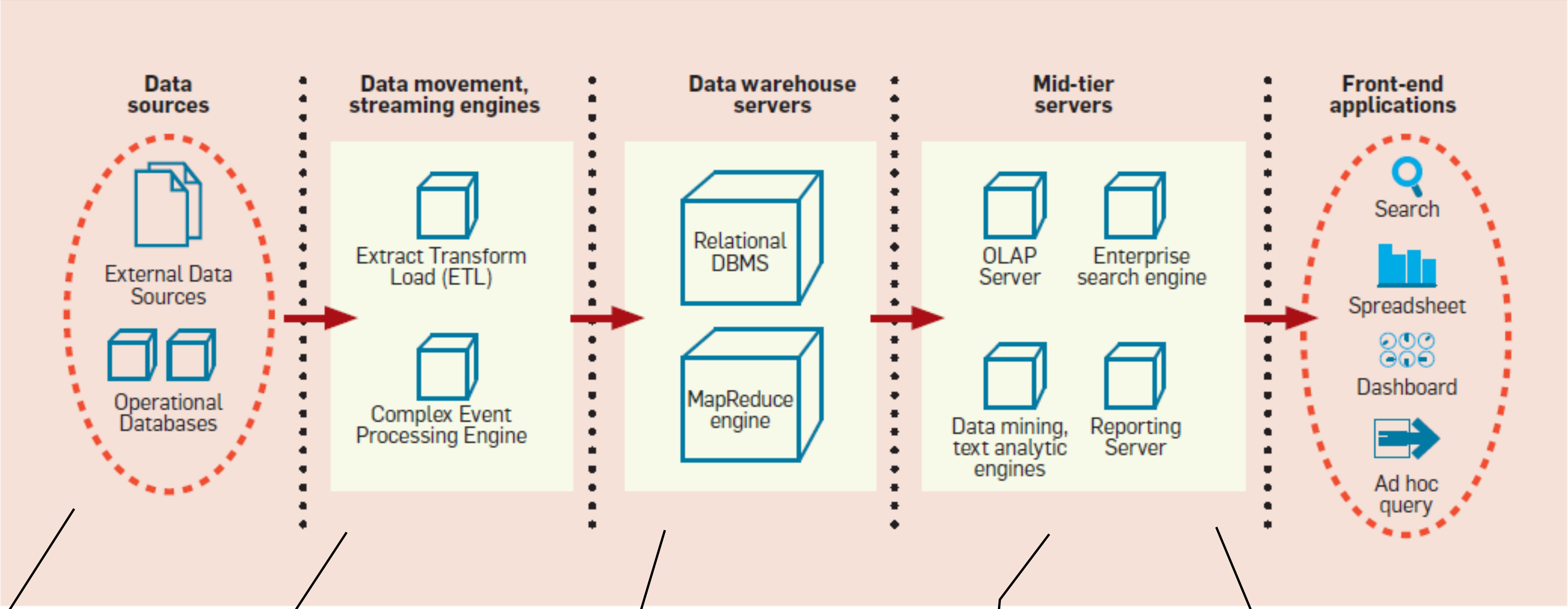
LABORATORY OF DATA SCIENCE

ETL – Extract, Transform and Load

Data Science & Business Informatics Degree

BI Architecture

Figure 1. Typical business intelligence architecture.



6 lessons – data access

4 lessons – data quality & ETL

1 lessons – analytic SQL

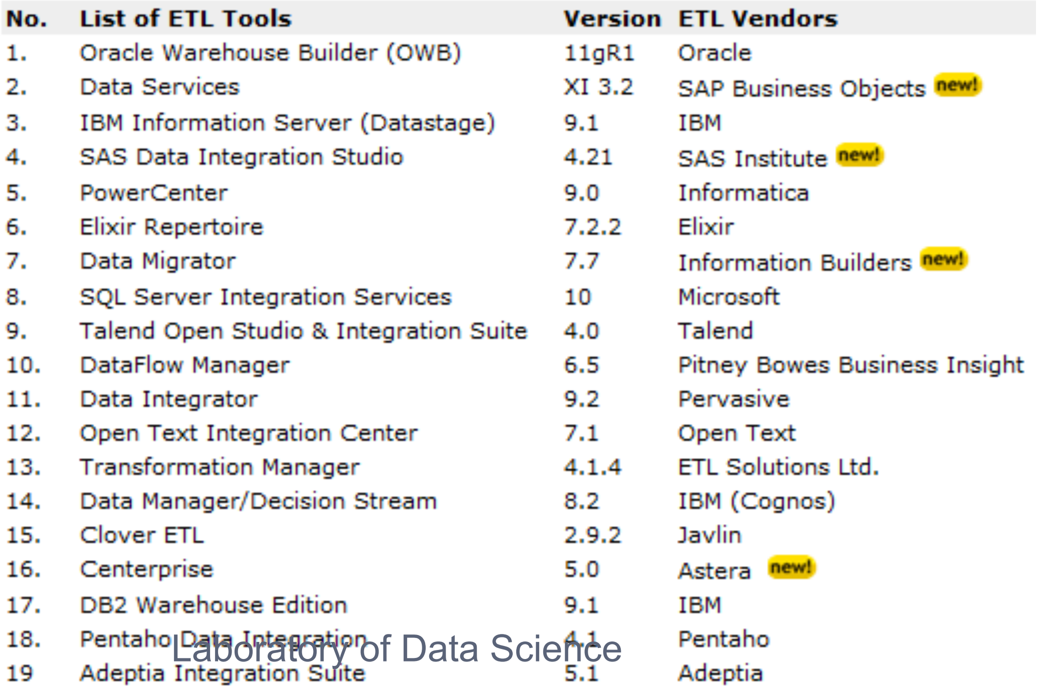
6 lessons – OLAP and reporting

5 lessons – data mining

Extract, Transform and Load

ETL (extract transform and load) is the process of extracting, transforming and loading data from heterogeneous sources in a data base/warehouse.

- Typically supported by (**visual**) tools.



No.	List of ETL Tools	Version	ETL Vendors
1.	Oracle Warehouse Builder (OWB)	11gR1	Oracle
2.	Data Services	XI 3.2	SAP Business Objects new!
3.	IBM Information Server (Datastage)	9.1	IBM
4.	SAS Data Integration Studio	4.21	SAS Institute new!
5.	PowerCenter	9.0	Informatica
6.	Elixir Repertoire	7.2.2	Elixir
7.	Data Migrator	7.7	Information Builders new!
8.	SQL Server Integration Services	10	Microsoft
9.	Talend Open Studio & Integration Suite	4.0	Talend
10.	DataFlow Manager	6.5	Pitney Bowes Business Insight
11.	Data Integrator	9.2	Pervasive
12.	Open Text Integration Center	7.1	Open Text
13.	Transformation Manager	4.1.4	ETL Solutions Ltd.
14.	Data Manager/Decision Stream	8.2	IBM (Cognos)
15.	Clover ETL	2.9.2	Javlin
16.	Centerprise	5.0	Astera new!
17.	DB2 Warehouse Edition	9.1	IBM
18.	Pentaho Data Integration	4.1	Pentaho
19	Adeptia Integration Suite	5.1	Adeptia

ETL tasks

- **Extract**: access data sources
  - Local, distributed, file format, connectivity standards

- **Transform**: data manipulation for quality improvm
  - Selecting data
    - remove unnecessary, duplicated, corrupted, out of limits (ex., age=999) rows and columns, sampling, dimensionality reduction
  - Missing data
    - fill with default, average, filter out
  - Coding and normalizing
    - to resolve format (ex., CSV, ARFF), measurement units (ex., meters vs inches), codes (ex., person id), times and dates, min-max norm, …
  - Attribute Splitting/merging
    - of attributes (ex., address vs street+city+country)

ETL tasks

- Managing surrogate key
  - generation and lookup
- Aggregating data
  - At a different granularity. Ex., grain “orders” (id, qty, price) vs grain“customer” (id, no. orders, amount), discretization into bins, …
- Deriving calculated attributes
  - Ex., margin = sales – costs
- Resolving inconsistencies – record linkage
  - Ex., Dip. Informatica Via Buonarroti 2 is (?) Dip. Informatica Largo B. Pontecorvo 3
- Data merging-purging
  - from two or more sources (ex., sales database, stock database)

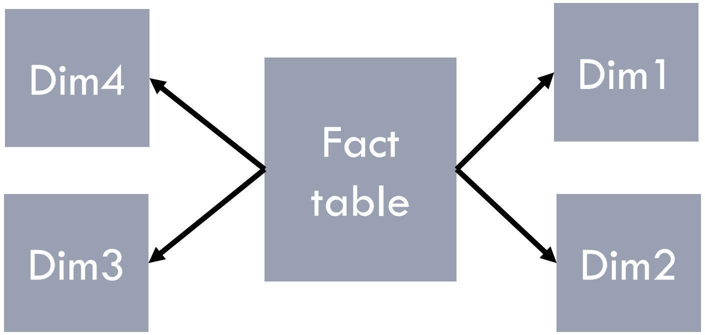
ETL tasks

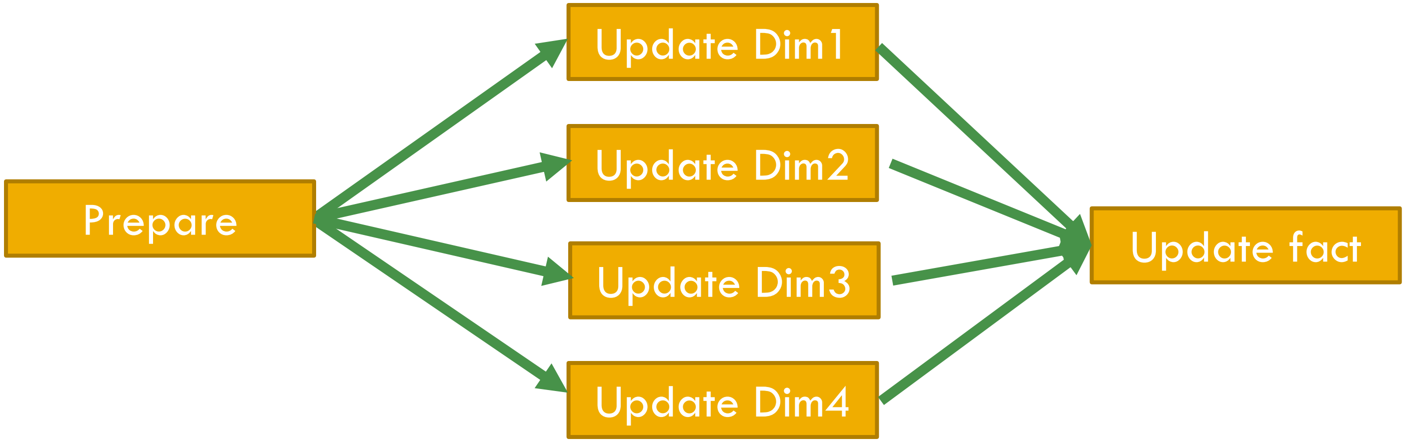
- **Load**
  - Data staging area
    - Area containing intermediate, temporary, partially processed data
  - Types of loading:
    - Initial load (of the datawarehouse)
    - Incremental load
      - Types of updates: append, destructive merge, constructive merge
    - Full refresh

ETL process management

- **Control flow** of ETL tasks
  - Task precedence
- **Data flow** ETL tasks
  - Access data source, transform data, load
- Error and warnings management
- Scheduling
- Metadata
- Required infrastructure
  - HW, SW, Personnel

ETL process management





LABORATORY OF DATA SCIENCE

SSIS - SQL Server Integration Services

Data Science & Business Informatics Degree

Background

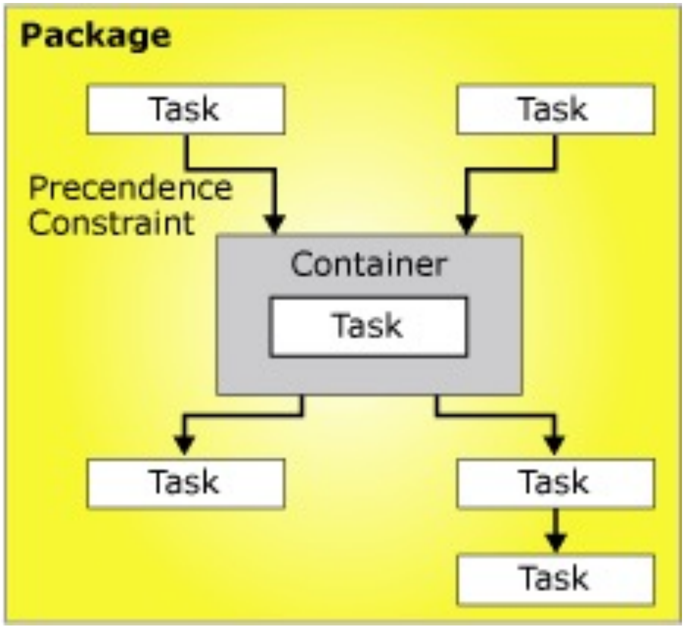
- **SSIS** is a tool for ETL
  - It can be used independently from SQL Server
  - Formerly called Data Transformation Services (in SQL Server 2000)
- Docs and samples
  - Tutorial from Books on Line
    - http://msdn.microsoft.com/en-us/library/ms141026.aspx
  - CodePlex samples
    - http://www.codeplex.com/SqlServerSamples#ssis
  - On-line community
    - http://sqlis.com

Developing SSIS projects

- Developer framework
  - Integrated within SSDT/BIDS
    - Solution = collection of projects
    - Project = developer project (C++, C#, IS, …)
- Demo
  - File → New Project → Integration Services
  - Panels: solution explorer, server explorer, others
  - SSIS packages (.dtsx extension)
    - Panels: control flow, data flow

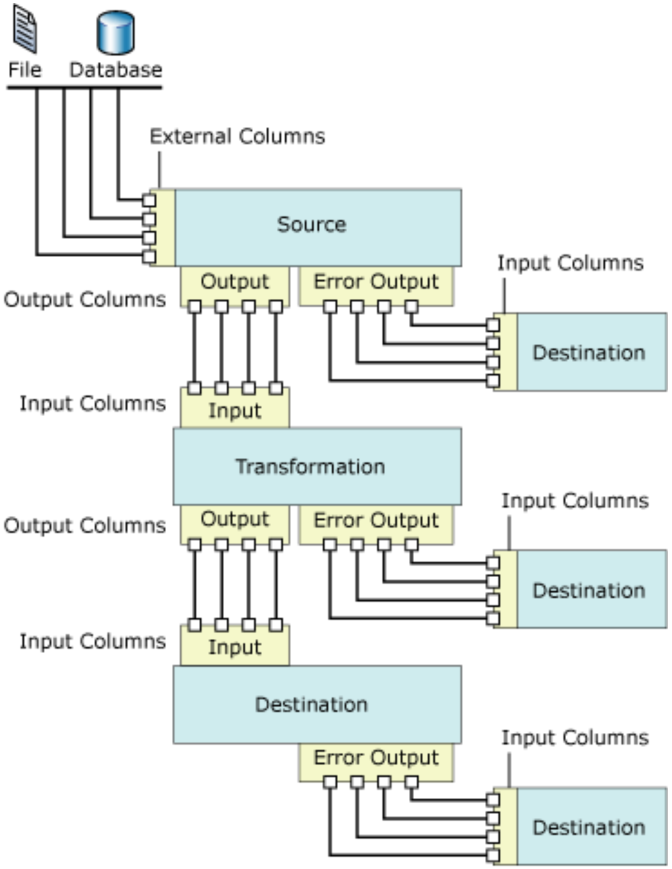
Control flow / Jobs

- Tasks, Containers & Precedence
  - Tasks
    - ETL tasks (list in the Toolbox panel)
  - Container
    - Iteration
  - Precedence
    - Arrows connecting tasks specify precedence type

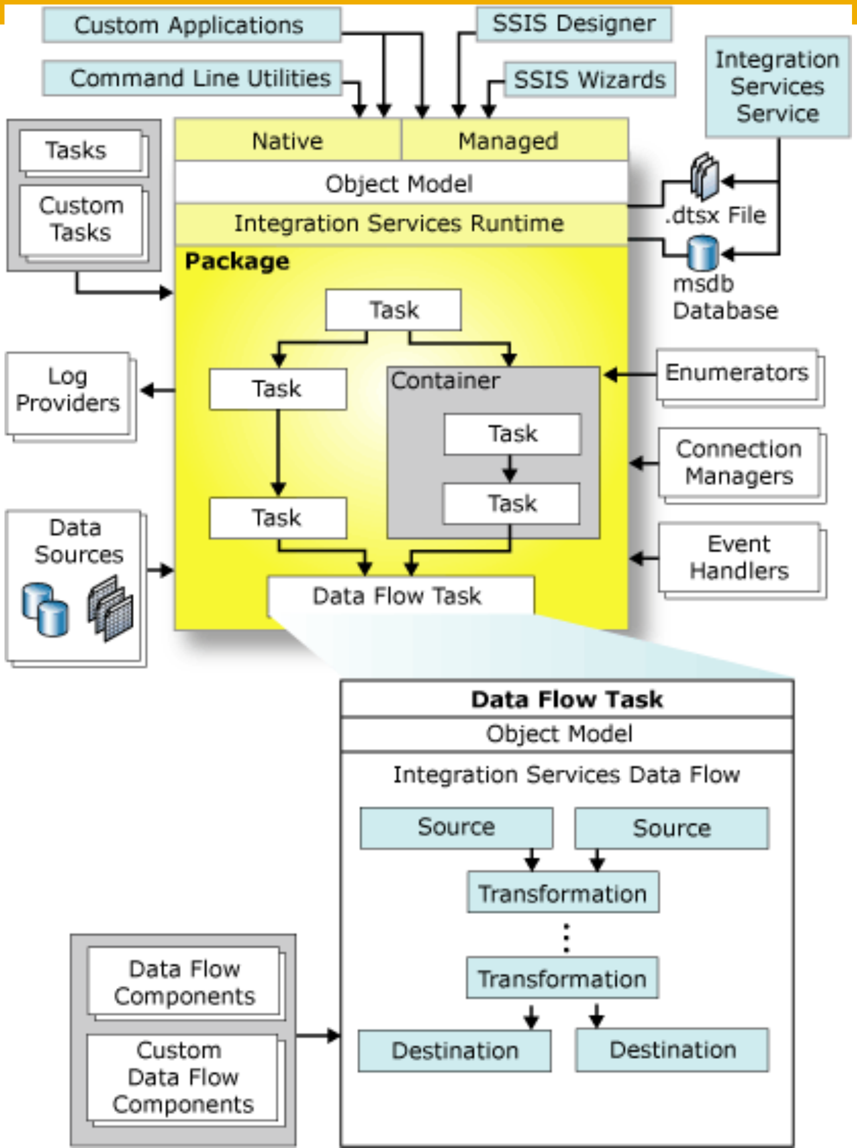


Data flow / Transformations

- Special tasks
- Define pipelines of data flows from sources to destination
  - Data flow sources
  - Data flow transformation
  - Data destination
  - Toolbox panel for list



SSIS projects structure



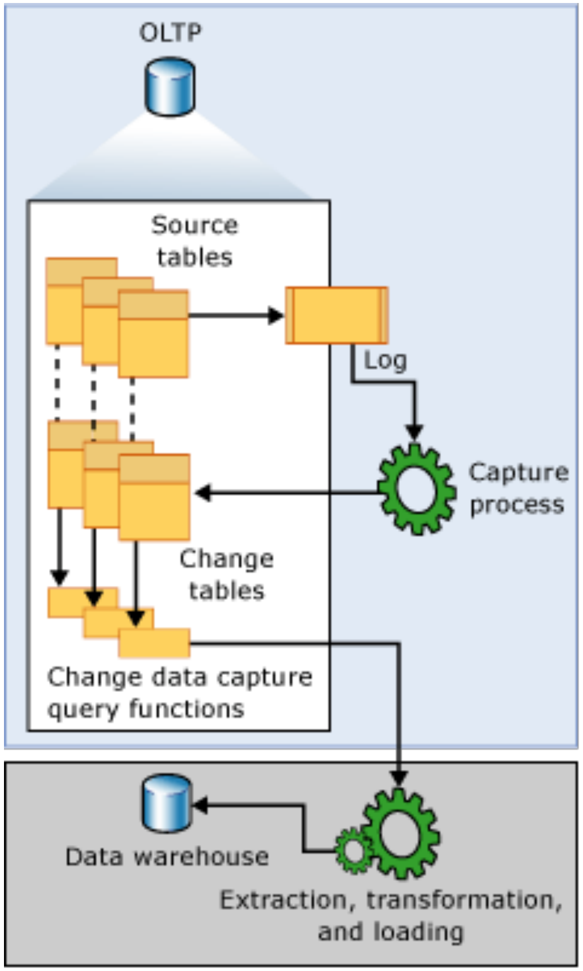
SSIS data types

- SSIS defines a set of reference data types
  - As seen for connectivity standards (ODBC, JDBC, OLE DB)
  - http://msdn.microsoft.com/en-us/library/ms141036.aspx
- Data type from sources are mapped into SSIS types
- SSIS transformations works on SSIS types
- SSIS types are mapped to destination data types

Debug, deployment, scheduling

- Debug
  - Data viewers
- Deployment
  - Save project on file
  - Save project on remote SSIS server
    - Project->Deploy
  - Load project from remote SSIS server
    - File->Add new project->Integration Services Import Project Wizard
- Launch
  - Local run
    - From Visual Studio
    - From command line: dtexec
    - From explorer: double click on .dtsx files
  - Remote run on SSIS servers
    - On demand / scheduled

Change data capture


OLTP
Source tables
Log
Capture process
Change tables
Change data capture query functions
Data warehouse
Extraction, transformation, and loading

BUSINESS INTELLIGENCE LABORATORY

ETL Demo: Pipeline, Sampling and Surrogate Keys

Business Informatics Degree

Pipeline

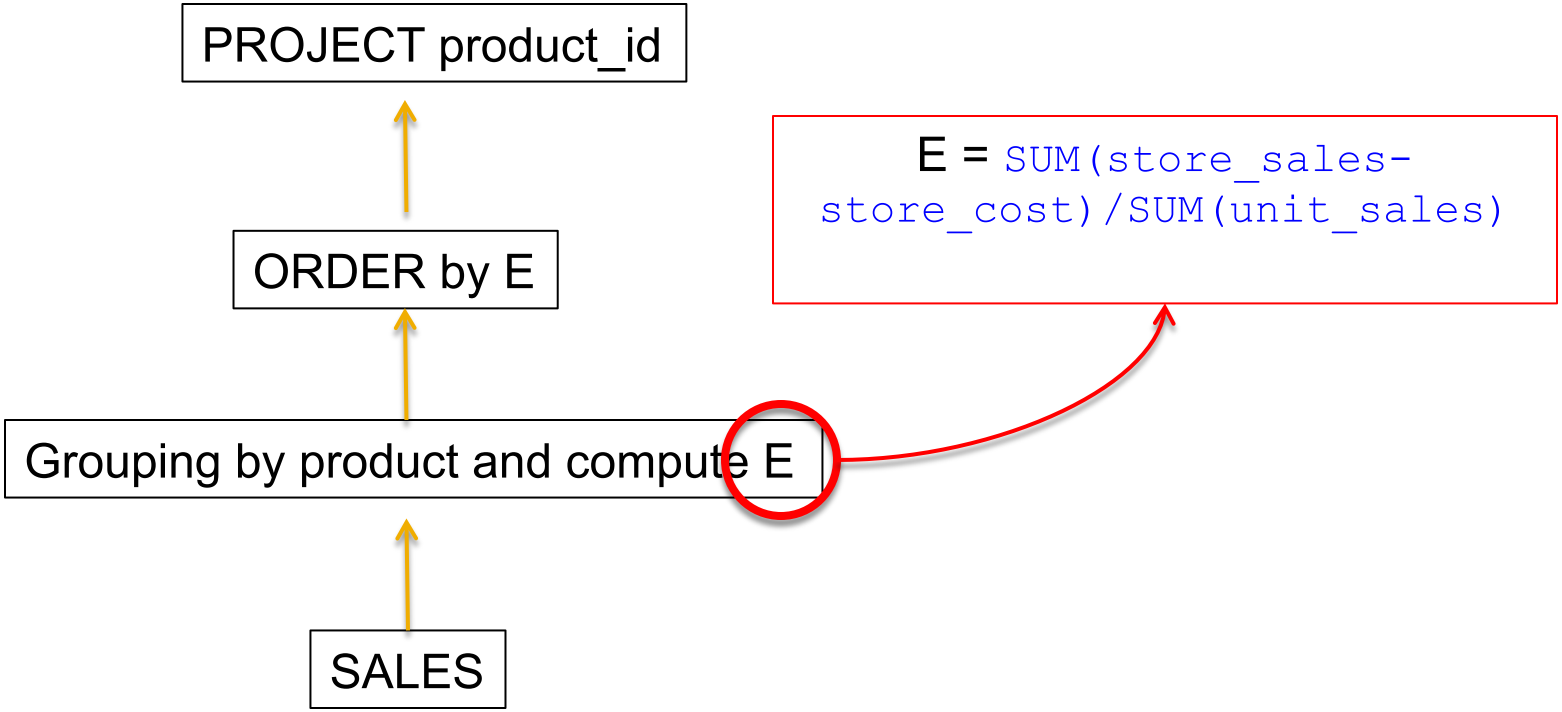
- Consider the Foodmart sales database
- Design a SSIS project for writing to a CSV file the list of products ordered descending by avg gain
  - Gain of a single sale in sales_fact table is defined as store_sales-store_cost
  - Avg gain of a product is the sum of gains of sales of the product divided by the total units_sales sold
- Do not use views! Do all work in SSIS.

SQL SOLUTION

```
SELECT product_id
FROM Sales
GROUP BY product_id
ORDER BY SUM(store_sales-store_cost)/
         SUM(unit_sales)
```

… and what about adding Product_name?

BASIC IDEA OF SISS SOLUTION



PROJECT product_id

ORDER by E

Grouping by product and compute E

SALES

E = `SUM(store_sales-store_cost)/SUM(unit_sales)`

Stratified subsampling

- Consider the census table on the Lbi database
- Design a SSIS project for writing to a CSV a random sampling of 30% stratified by sex
    - 30% of males plus 30% of females
- Do not use views! Do all work in SSIS.

BUSINESS INTELLIGENCE LABORATORY

Lab exercise on ETL: SCD

Business Informatics Degree

SCD: background

- **Slowly Changing Dimensions**
  - Datawarehouse dimensions members updates
  - Three types:
    - Type 1: overwrite previous value
    - Type 2: keep all previous values
    - Type 3: keep last N previous values (N ~ 1, 2, 3)
  - Each attribute of the dimension can have its own type
    - Type 1: name, surname, …
    - Type 2: address, …

SCD: input and output tables

- Database FoodMart in SQL Server
- Input
  - table **customer**
- Output in Lbi database
  - create a table **customer_dim**
    - columns
      - surrogate_key (PK), customer_id, customer_name, address, date_start, date_end
    - with
      - surrogate_key being a surrogate key, customer_name including name and surname, address made of address1-city-zip-province-country, date_start and date_end are dates

Preliminary step

- □ Develop a SSIS package that adds to **customer_dim** the customers in **customer** that are not already in it

SCD: type 1 updates

- Overwrite previous value
- Changes on the input table **customer**
  - On 10/3/2007
    - 231, Mario Rosi, Via XXV Aprile Pisa
  - On 12/3/2007
    - 231, Mario Rossi, Via XXV Aprile Pisa
  - Surname has been corrected

SCD: type 1 updates

- The DW **customer_dim** table looks as:
  - On 10/3/2007, and up to 12/3/2007

surrogate_key, customer_id, name, address, date_start, date_end

874, 231, Mario Rosi, Via XXV Aprile Pisa, 10/3/2007, NULL

  - On 12/3/2007

surrogate_key, customer_id, name, address, date_start, date_end

874, 231, Mario Rossi, Via XXV Aprile Pisa, 10/3/2007, NULL

SCD: type 2 updates

- Keep all previous values
- Changes on the input table **customer**
  - On 12/3/2007
    - 231, Mario Rossi, Via XXV Aprile Pisa
  - On 25/9/2008
    - 231, Mario Rossi, Via Risorgimento Pisa
  - Customer has changed his address

SCD: type 2 updates

- The DW **customer_dim** table looks as:
  - On 12/3/2007, and up to 25/9/2008

surrogate_key, customer_id, name, address, date_start, date_end

874, 231, Mario Rossi, Via XXV Aprile Pisa, 10/3/2007, NULL

  - On 25/9/2008

surrogate_key, customer_id, name, address, date_start, date_end

874, 231, Mario Rossi, Via XXV Aprile Pisa, 10/3/2007, 25/9/2008

987, 231, Mario Rossi, Via Risorgimento Pisa, 25/9/2008, NULL

Lab exercise

- Design a SSIS project to update **customer_dim** starting from **customer** as follows:
  - Customers in **customer** that are not in **customer_dim** are added to it
  - Updates of **customer_name** are of Type 1
  - Updates of **address** are of Type 2

BUSINESS INTELLIGENCE LABORATORY

Other lab exercises on ETL

Business Informatics Degree

Sales during travels

- A sale in *sales_fact* was done during a travel if the store of the sale was not in the city of residence of the customer. Develop a SSIS package which produces a CSV file with a row for every customer with:
  - the customer full name
  - the total sales to the customer
  - the ratio of sales done during travels

Sales in weekends of previous month

- For a given customer and month, the frequency of purchases in weekends (FPW) is the number of distinct weekend days (Saturdays or Sundays) of the **previous** month in which the customer made a purchase. Develop a SSIS package which produces a CSV file with a row for every customer and month with:
  - the customer full name
  - the month and year
  - the customer FPW