Basic Data Warehousing Concepts and Methodology

A Presentation by

- Shalini Guha

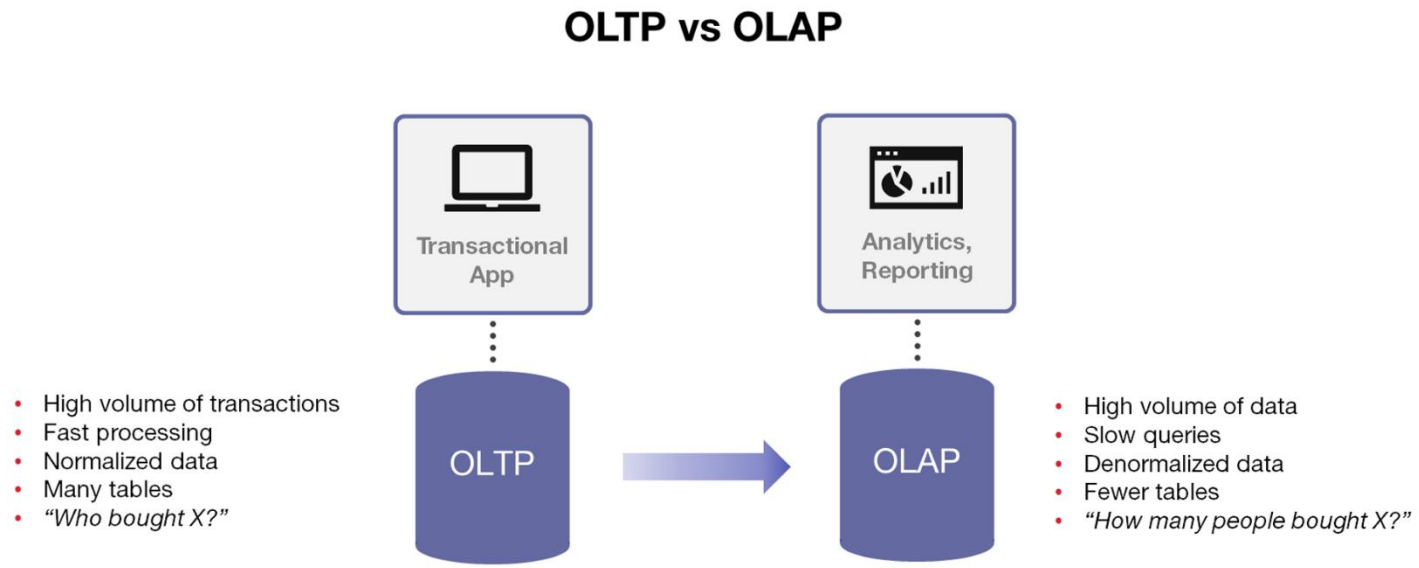
13.07.2018

Data Warehouse

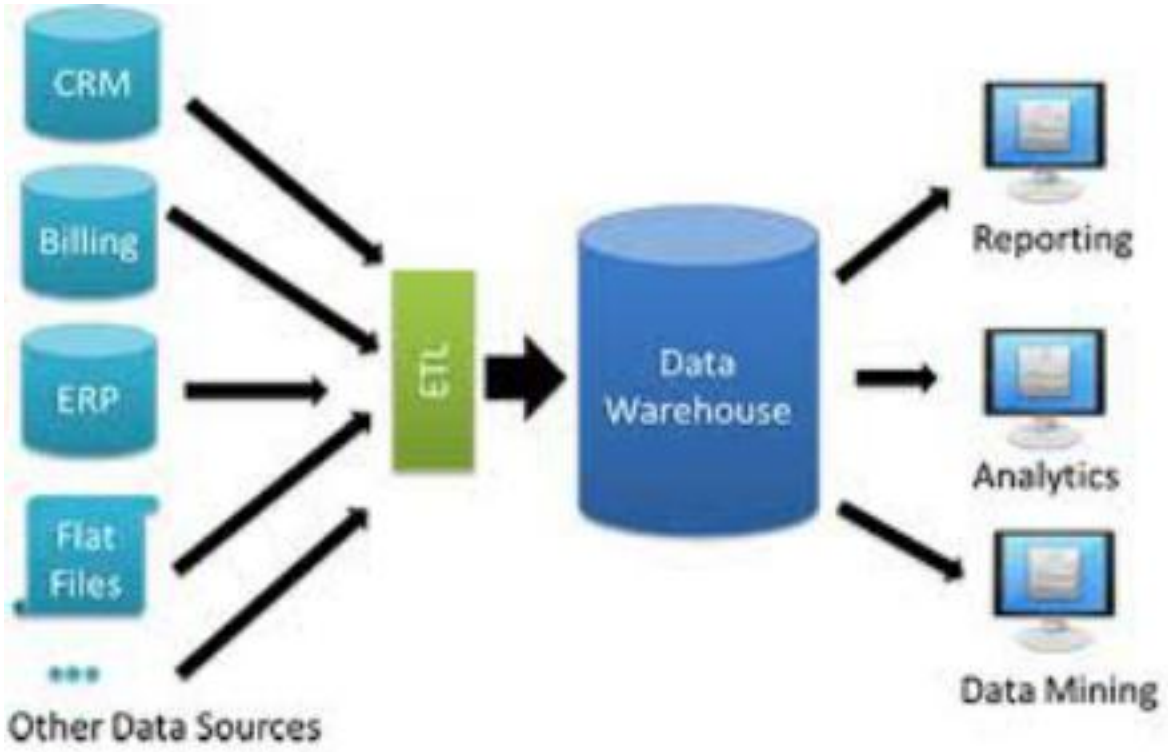
- **Data warehousing** is a technology that aggregates structured data from one or more sources so that it can be compared and analyzed for greater business intelligence.
  - **Subject Oriented**
  - **Integrated**
  - **Time Variant**
  - **Non-volatile**

Data Warehouse(OLAP) vs Operational Databases (OLTP)

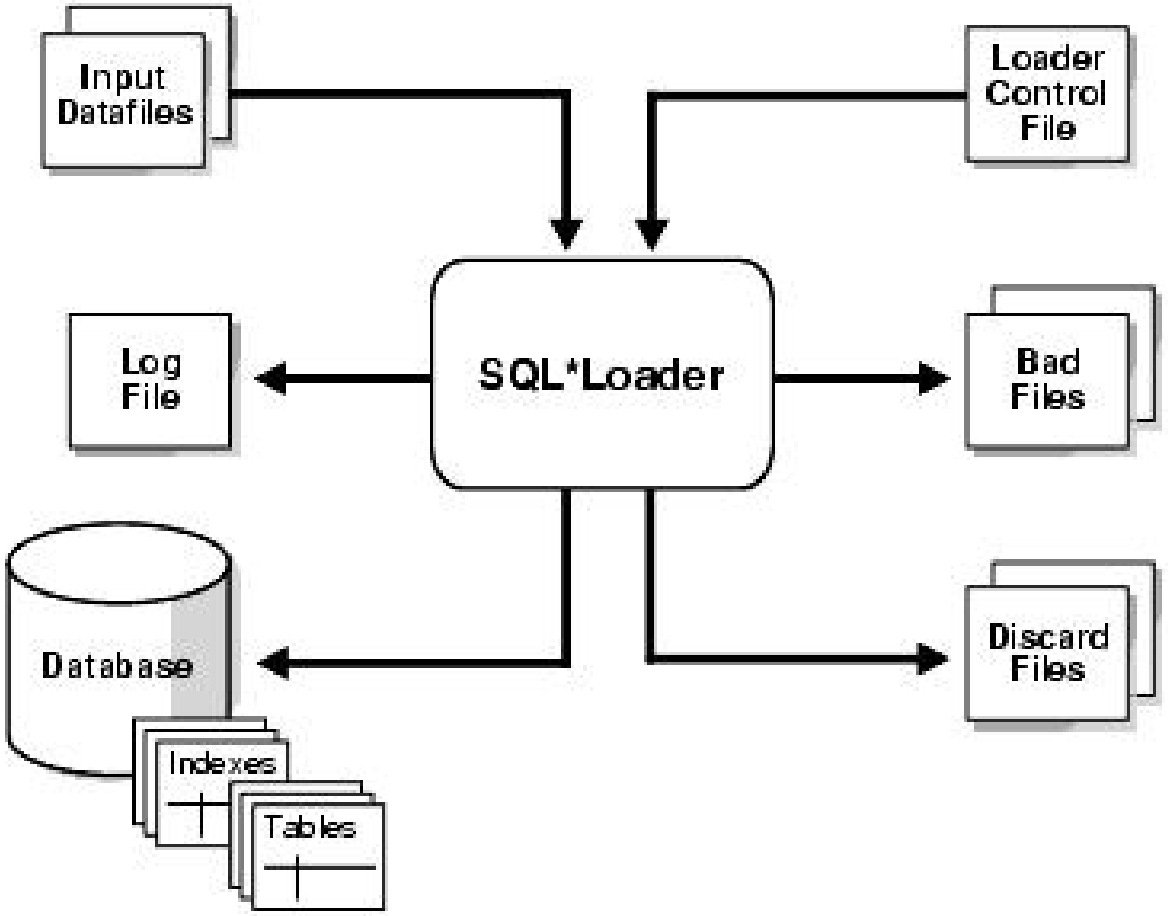
Data warehouses use a different design from standard operational databases. The latter are optimized to maintain strict accuracy of data in the moment by rapidly updating real-time data. Data warehouses, by contrast, are designed to give a long-range view of data over time. They trade off transaction volume and instead specialize in data aggregation.



ETL : Extract, Transform, Load



Preparing Dimension Tables using SQL Loader

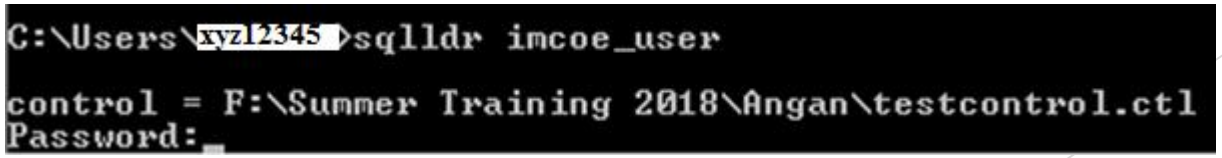


Truncate-Load data using Control Files

- Create a new Table in Oracle.
- Create the Following Control File

```
load data
infile  <CSV File Path>
BADfile  <BAD File Path>
truncate into table ORDER_ITEMS
fields terminated by ","
(
        ITEM_ID ,
        ORDER_ID ,
        QUANTITY ,
        DISCOUNT
)
```

- Run the SQL Loader utility through the Command Prompt



Loading Data From Multiple Files

Here, all the data of the CSV files present in a folder are imported together in a table in Oracle.
It is done so by creating a **batch file(.bat)**.
The batch file is implemented to create a CSV file where all the data of the multiple CSV files are imported.

```
@echo off
copy <File Path>\all_files1.csv
sqlldr '<Oracle user name>' control='testcontrol.ctl' log='Results.log'
pause
```

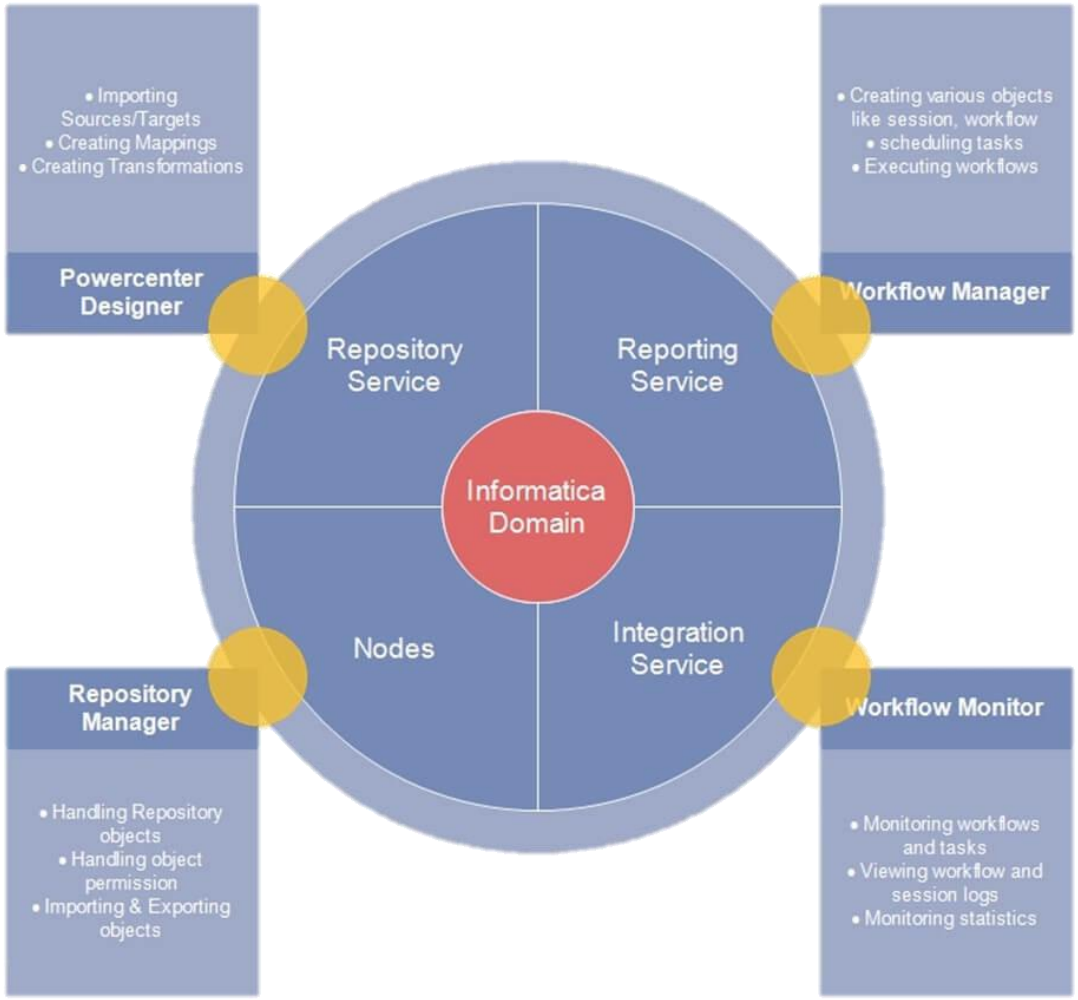
After implementing the batch file, a new CSV file is created which contains the data of all the CSV files.
Here a new CSV file of name **all_files1** is created

ETL Using Informatica Powercenter

- Informatica is a Software development company, which offers data integration products. It offers products for ETL, data masking, data Quality, data replica, data virtualization, master data management, etc.

- Informatica Powercenter ETL/Data Integration tool is a most widely used tool and in the common term when we say Informatica, it refers to the Informatica PowerCenter tool for ETL.

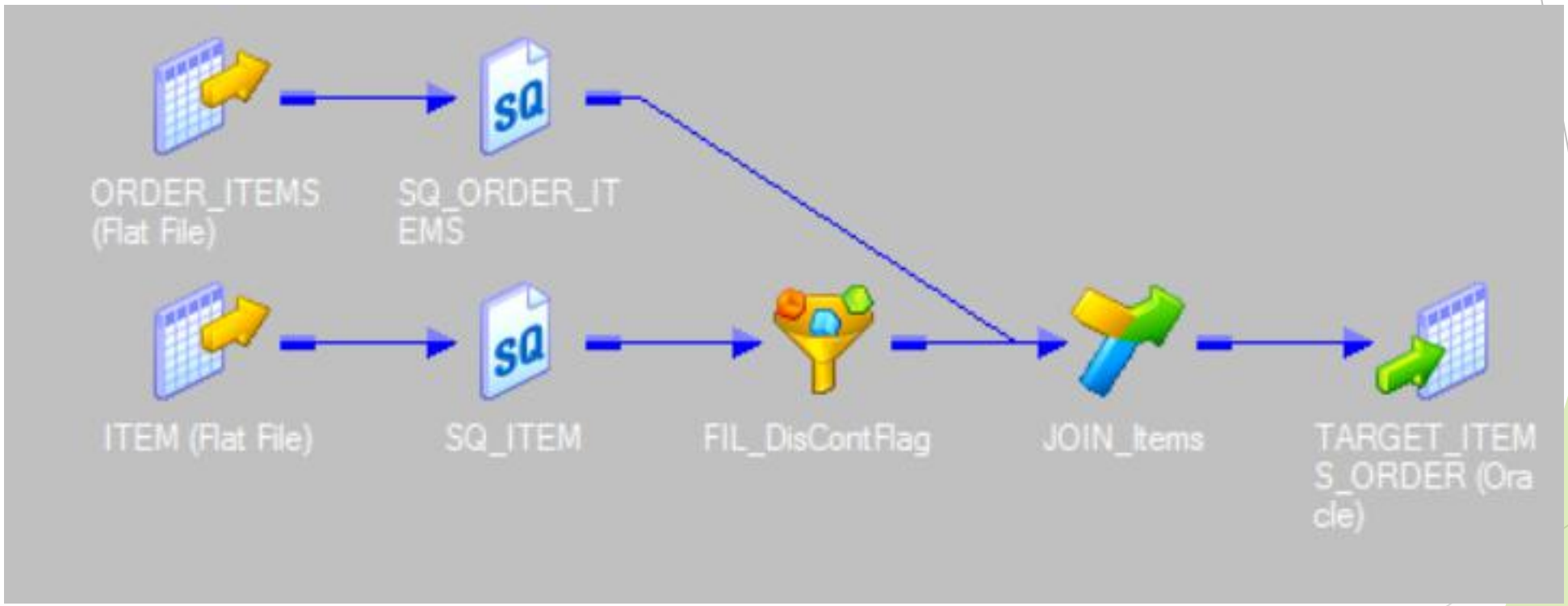
Informatica Architecture



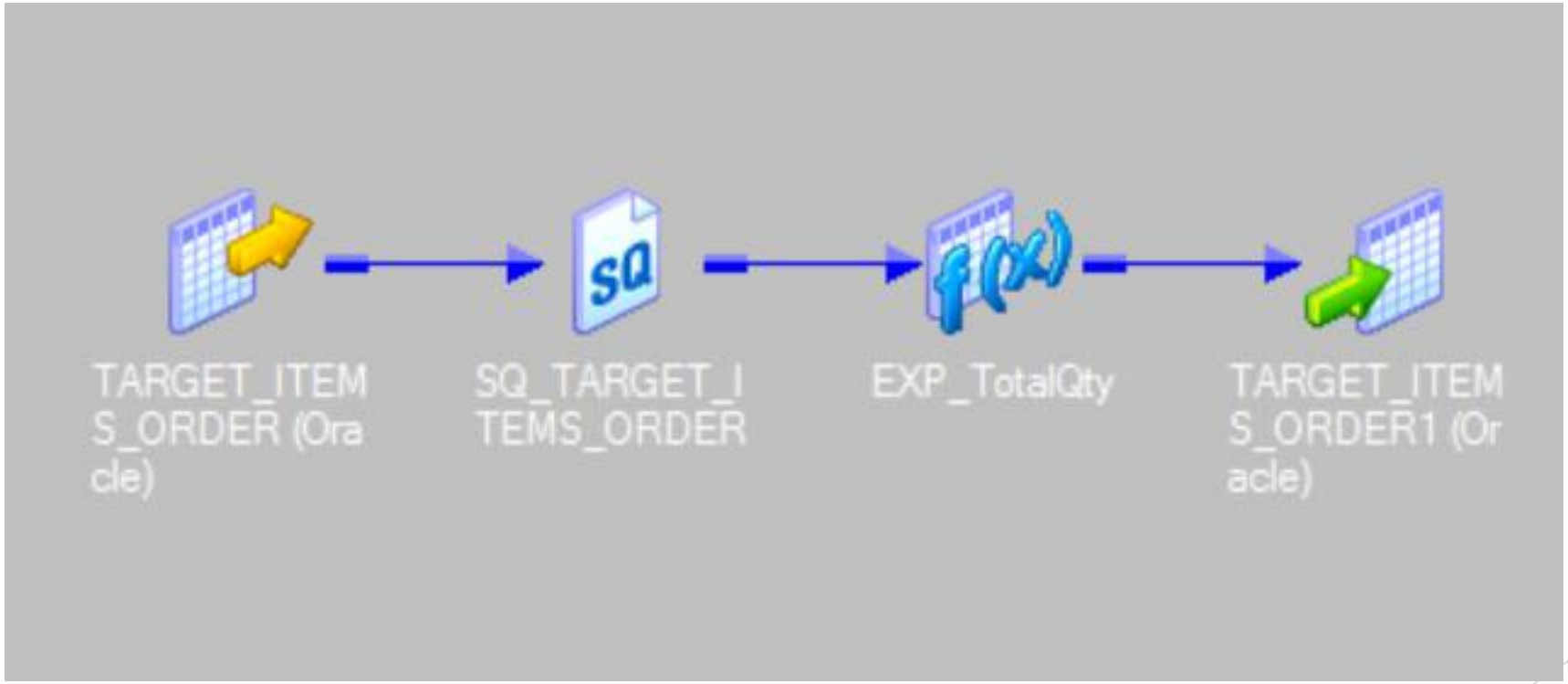
Scenario

- Our objective here is to get the information of the amount of sales done by the chain of departmental stores.
- We try to know the ‘Total Cost’ of items sold for each Order (i.e. the amount of transactions done in each order) along with other Order details.
- The ‘Order cost’ is the sum of all the ‘items cost’ in the order with discount adjustments if any.
- The final target table should reflect the Total Cost against each Order_ID with other Order details. The intermediate target tables can be used as the source for subsequent mappings.

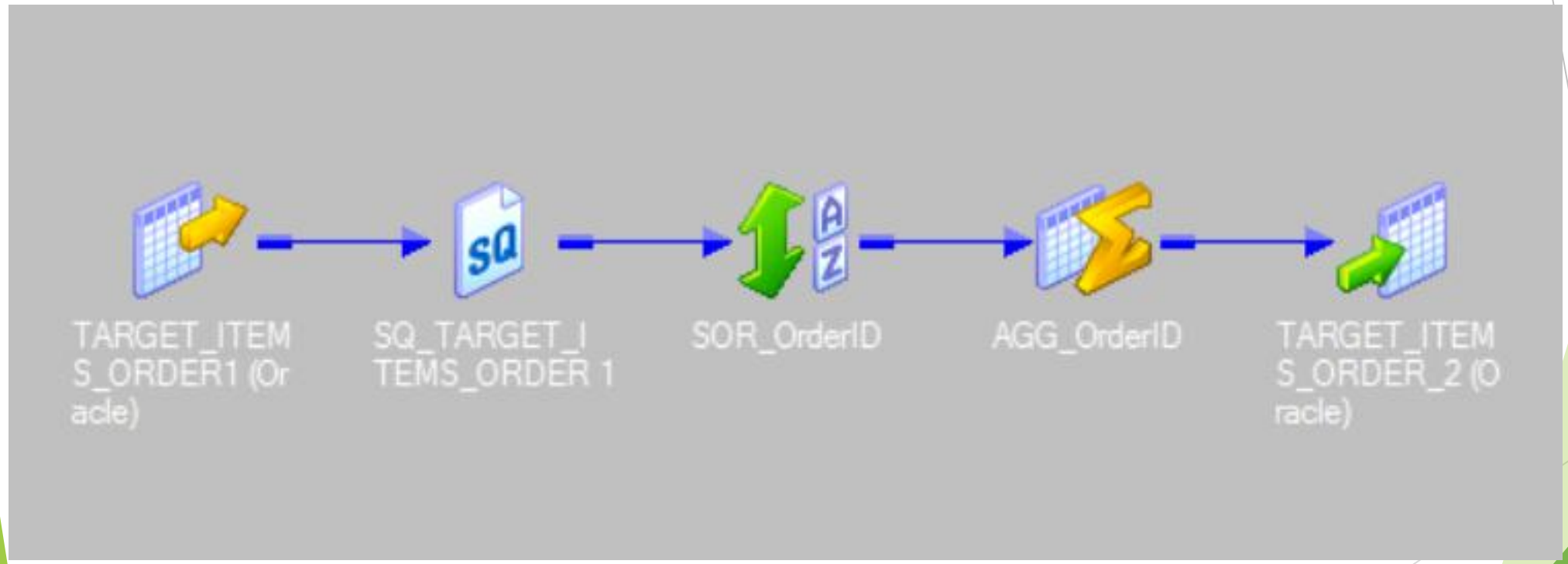
Part 1 and 2: Joining Filtered Records



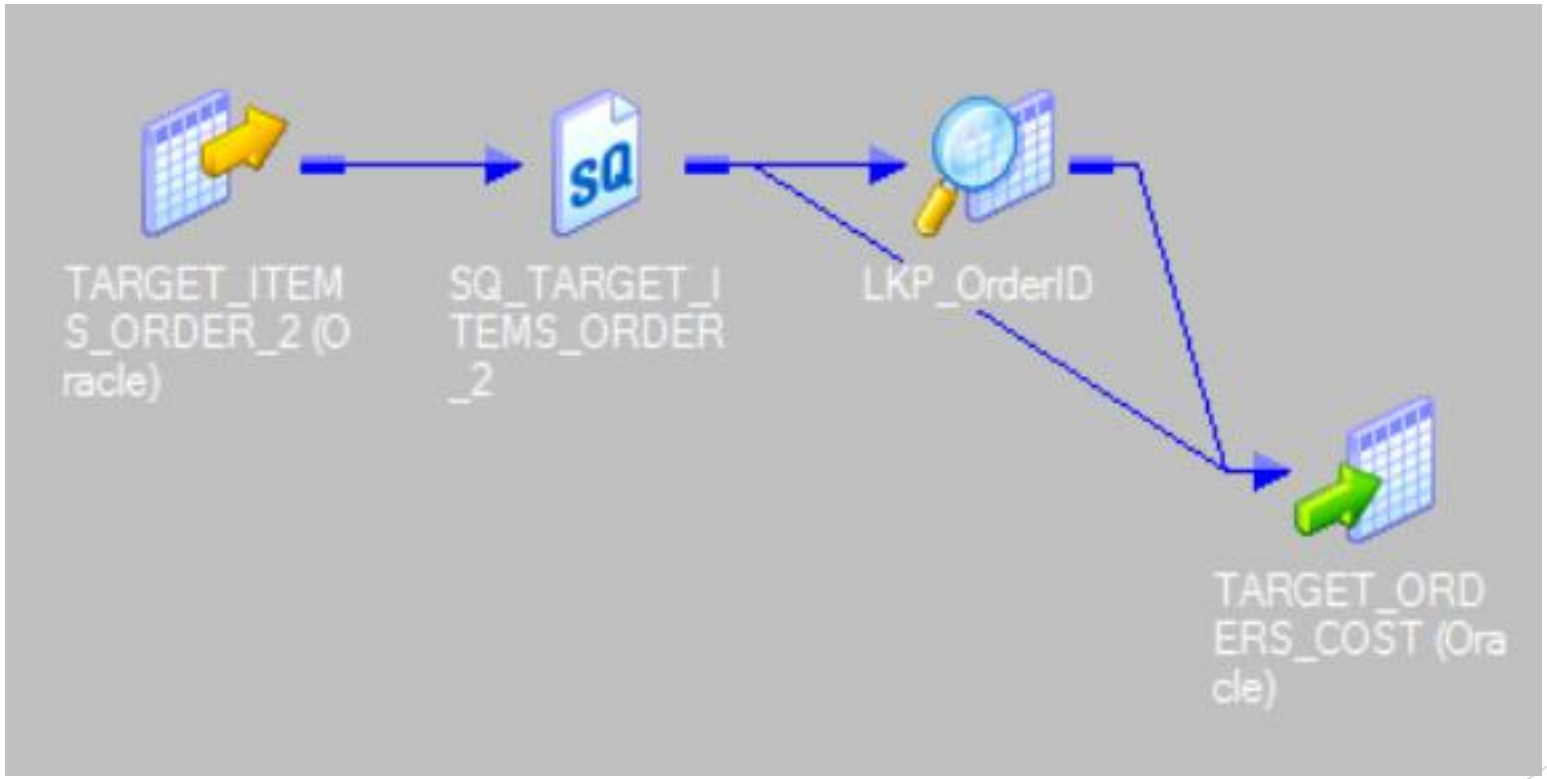
Part 3: Calculating Total Cost for each Item



Part 4: Calculating total cost for each Order



Part 5: Obtain Order Details using Order ID



Conclusion

- Excellent GUI interfaces for Administration, ETL Design, Job Scheduling, Session monitoring, Debugging, etc.
- Access to wide range of enterprise data sources
  - Mainframe and file-based data
  - Relational data
  - Message queues
  - XML and unstructured data
  - Third party application data
- Can easily adopt and integrate with vendor supplied data handling utility (Oracle Parallel processing and Load balancing)
- Single point of control (web based) for enterprise wide applications ensuring high degree of security with reduced administration overhead

Any Questions?