Introduction to ETL

David Amadi

Outline

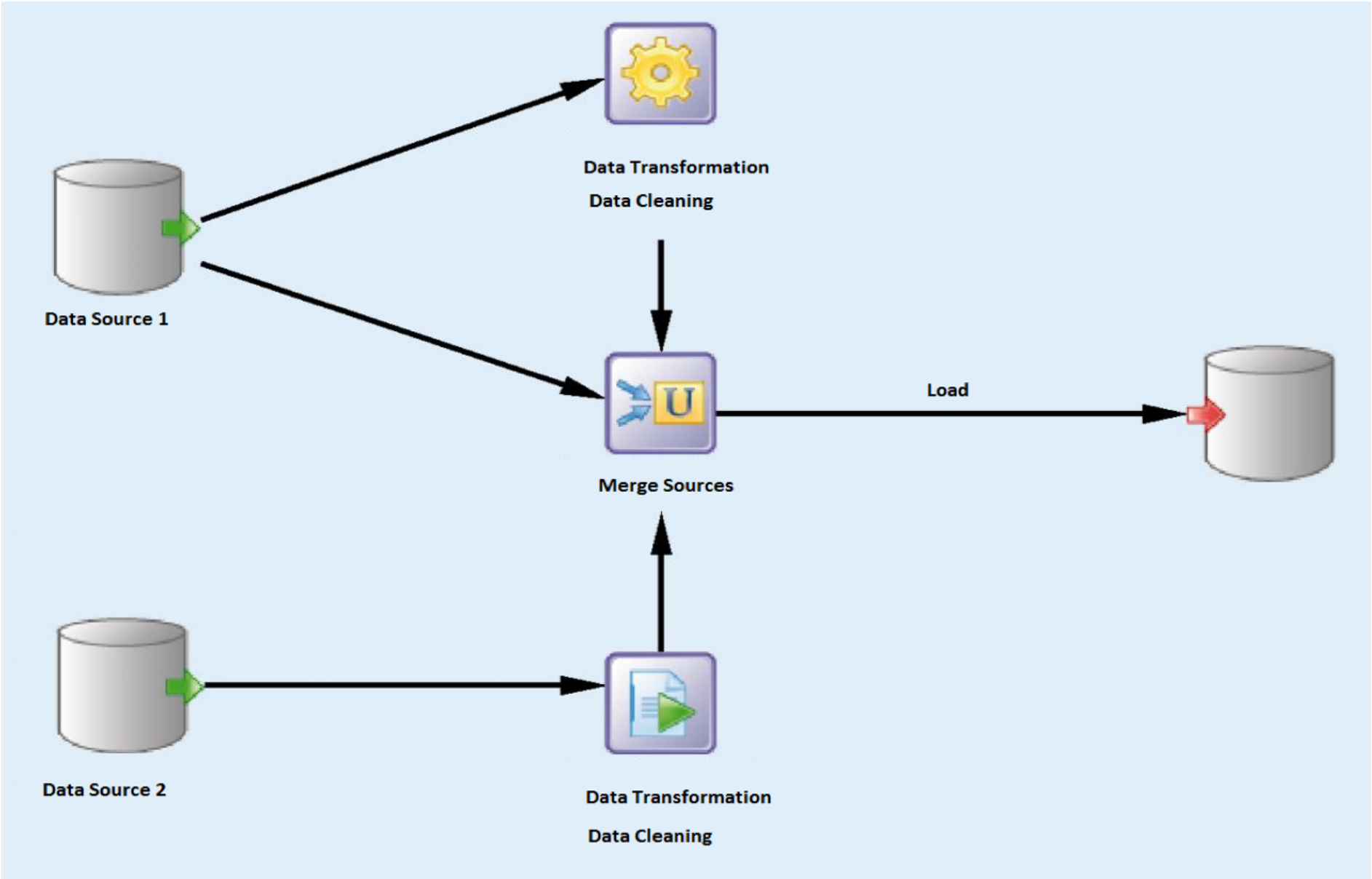
- **Understanding ETL**
- **Direct connection using IDE for DB's**
- **ODBC connection**
- **ETL using Pentaho Kettle**

What is ETL?

- **Extract:**Extract data from single or multiple data sources
- **Tranform:**Check on the quality of data and do necessary cleaning and convert to standards (making the data fit for the specific purpose)
- **Load:**Data loaded into the target warehouse / repository which will be used for reporting / analysis purpose.

Focus: Preparing the data for reporting / analysis

Schematic Representation of ETL



ETL Tools

Few popular commercial and freeware (open-sources) ETL Tools

- IBM Infosphere DataStage
- Informatica PowerCenter
- Oracle Warehouse Builder
- Oracle Data Integrator
- SAS ETL Studio
- Business Objects Data Integrator
- Microsoft SQL Server Integration Services
- Ab Initio
- Pentaho Data Integration (PDI) - Kettle
- Talend Integrator Suite
- CloverETL
- Jasper ETL

Direct connection using IDE for DB's

- Workbench, Download and install MySQL GUI tool
  - Sequel Pro, phpMyAdmin, SQLyog etc
- Configuration of the interface
  - Use of SQL editor to run SQL scripts
  - Run SQL scrips on a connected database
  - Export data as .csv,xml,SQL...
- Terminal
  - To connect to database and run SQL scripts

Select Statement

Select components (MySQL)

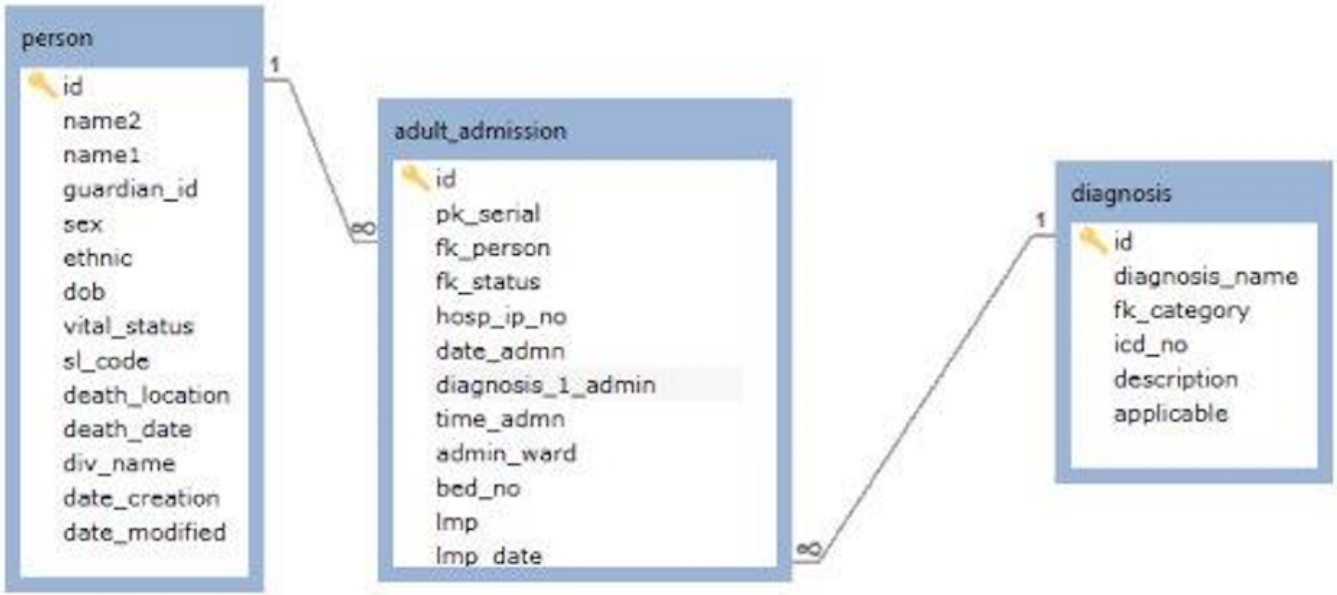
- Select (Required)
- Expression (required)
- From Table (Only optional if table data is not used)
- Where Condition (optional)
- Group By (optional)
- Having (optional)
- Order By (optional)
- Limit (optional)

Connecting to MySQL Database

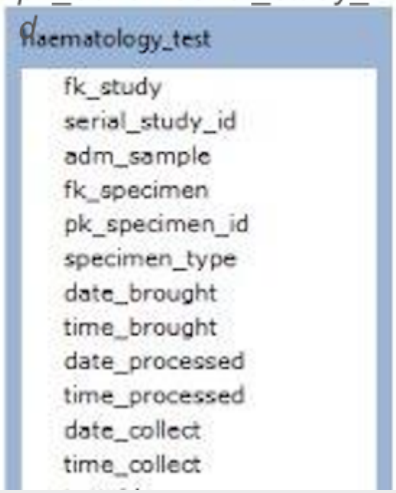
Credentials

- Connection name: tiba_workshop
- mysql_host_address: keklf-mysqluat
- username: tiba_usr
- password: tiba2019
- database: tiba

ER Diagram of Sample Database



pk_serial=serial_study_i
d



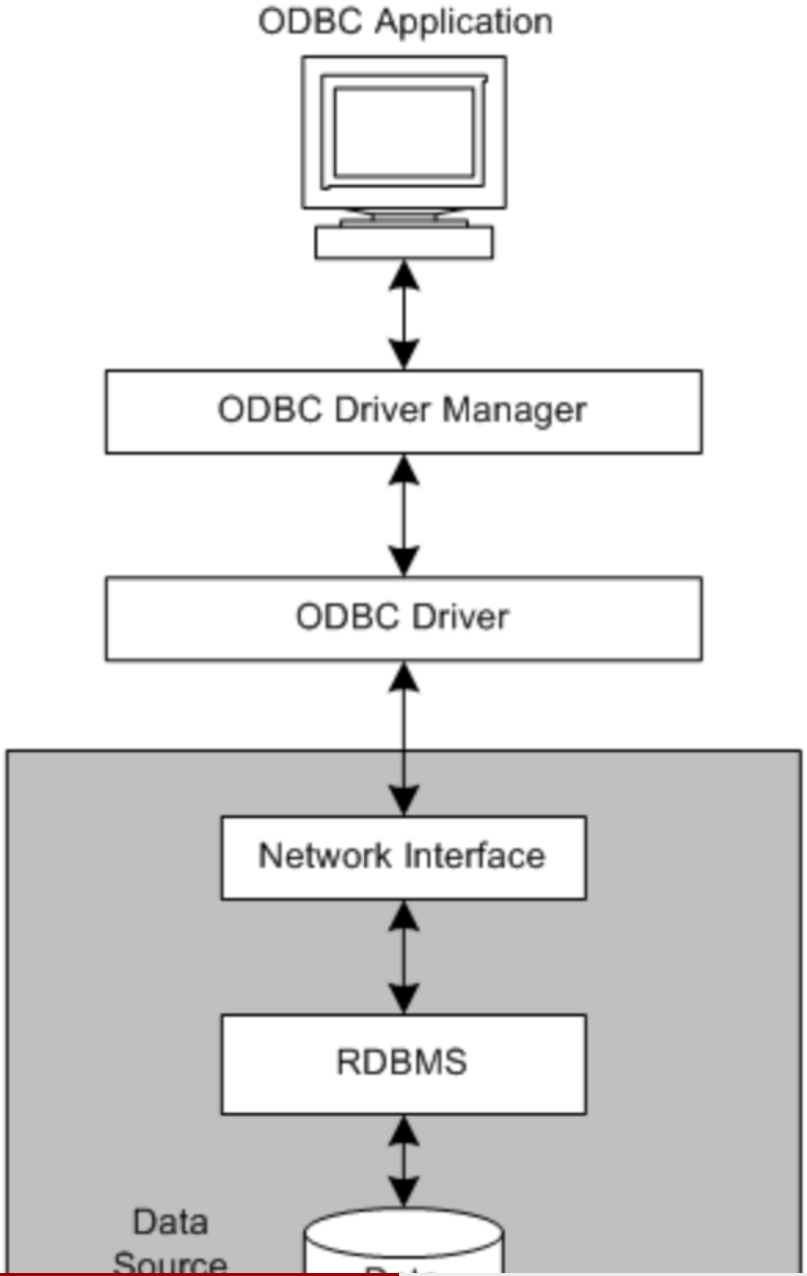
Practicum

Session 1: Duration 15 minutes

ODBC

- API for database access
  - Access any data from any application, regardless of which DBMS is handling data
- ODBC Components:
  - ODBC Application (R, STATA e.t.c)
  - ODBC Driver (processes ODBC functions)
  - ODBC Driver manager
  - ODBC data source(DBMS)
- Runs in most O/S (Windows, Unix, Mac OS X)

ODBC Architecture



ODBC Connection: R

- Configuration of ODBC Drivers
- Installation of R Packages:

```
install.packages("RMysQL") # driver for MySQL
install.packages("RPostgresSQL") # driver for PostgresSQL
```

ODBC: Connect to database

```
#load R packages
library(DBI) # provides the interface
library(RMySQL) #for connecting to the MySQL database

# cretae an ODBC connection to MySQL database
con <- dbConnect(MySQL(),
                 user="username",
                 password="password",
                 dbname="databasename",
                 host="hostname")
```

Sample Data Request

4. DESCRIPTION TYPE OF DATA REQUIRED (INCLUDE DETAILS OF STUDY THAT HOLDS THE DATA YOU REQUIRE IF APPLICABLE i.e. study number, study title and PI name)

Dataset to contain mortality information from KHDSS and adult ward admission at two time periods I.e. 2015 and 2016 for women at reproductive age.
Variable of interest include:

- Sex, age, date of admission, discharge diagnosis,admin_ward

5. OBJECTIVES OF DATA USE INCLUDING TIME FRAME

- Determine risk factors associated with mortality

6.ANY RISKS/SENSITIVITIES OR BENEFITS OF ANALYSIS, including to Data Subjects and groups of Data Subjects

- The results will allow policy makers to generate actionable recommendation, prediction and generate new hypothesis.

7. PLANNED OUTPUTS OF ANALYSIS

- Presentation at conferences and publication

Practicum

Session 2: Duration 30 minutes

ETL Using Pentaho Kettle

- Pentaho Data Integration is also known as Kettle
- Pentaho Data Integration is a powerful ETL tool using an innovative, metadata-driven approach
- With graphical, easy-to-use drag and drop design environment for building ETL jobs and transformations, results in faster development and simplified deployment
- DownLoad:

Pentaho Data Integration Community Edition (PDI CE)

Heterogeneity Support in Kettle

- Kettle supports a wide range of database (35+) and file systems for input and output

- CSV
- XML
- Test File
- XML
- RSS
- MySQL
- Oracle
- Microsoft Access
- MS SQL
- Microsoft Excel

Common Uses of Kettle

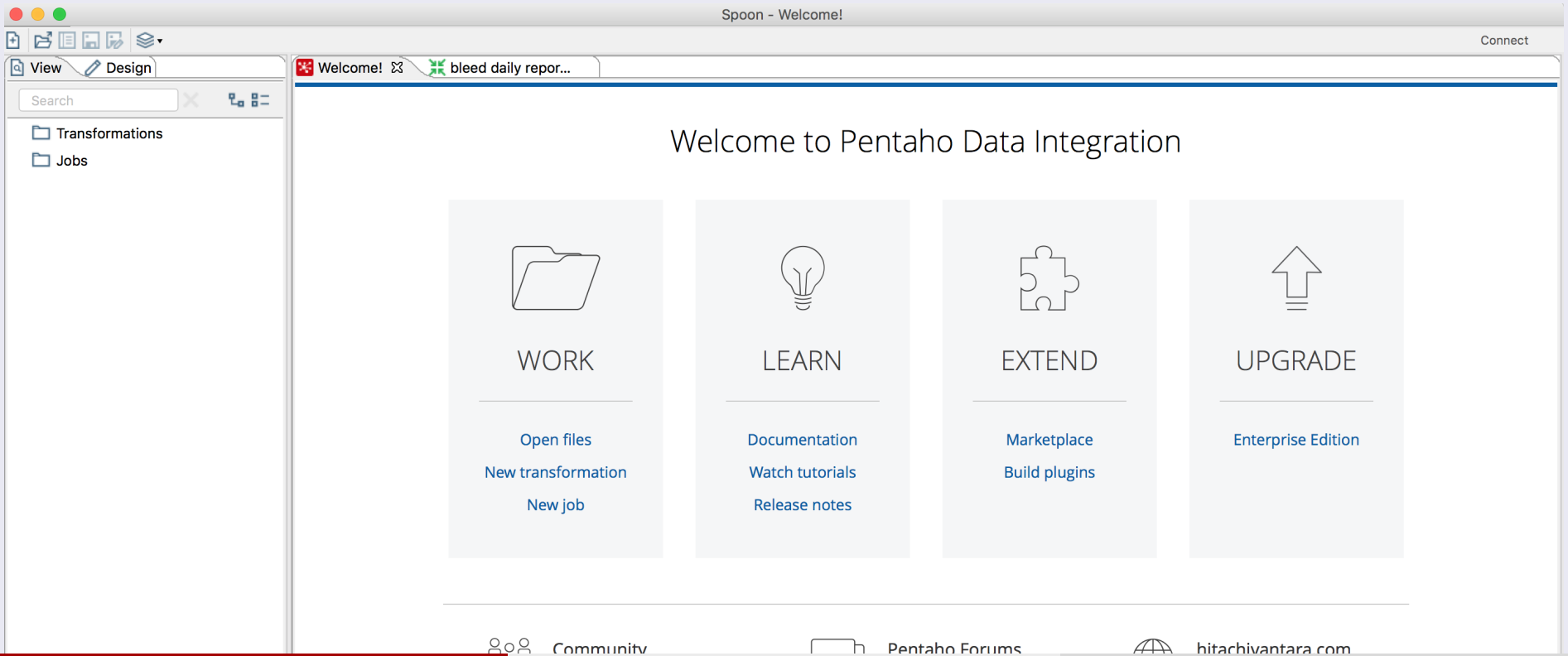
Pentaho Data Integration (Kettle) is a flexible ETL tool with the following uses:

- Data warehouse population with built-in support for Data Anonymization / Data Masking
- Data migration between different databases and applications
- Loading huge data sets into databases
- Data Cleansing with steps ranging from very simple to very complex transformations

First look at Kettle

- Spoon is the design interface for building ETL jobs and transformations.
- Provides a drag and drop interface to graphically describe what you want to take place in your transformations

Create complex jobs without having to read or write much of the code



Kettle Transformations and Jobs

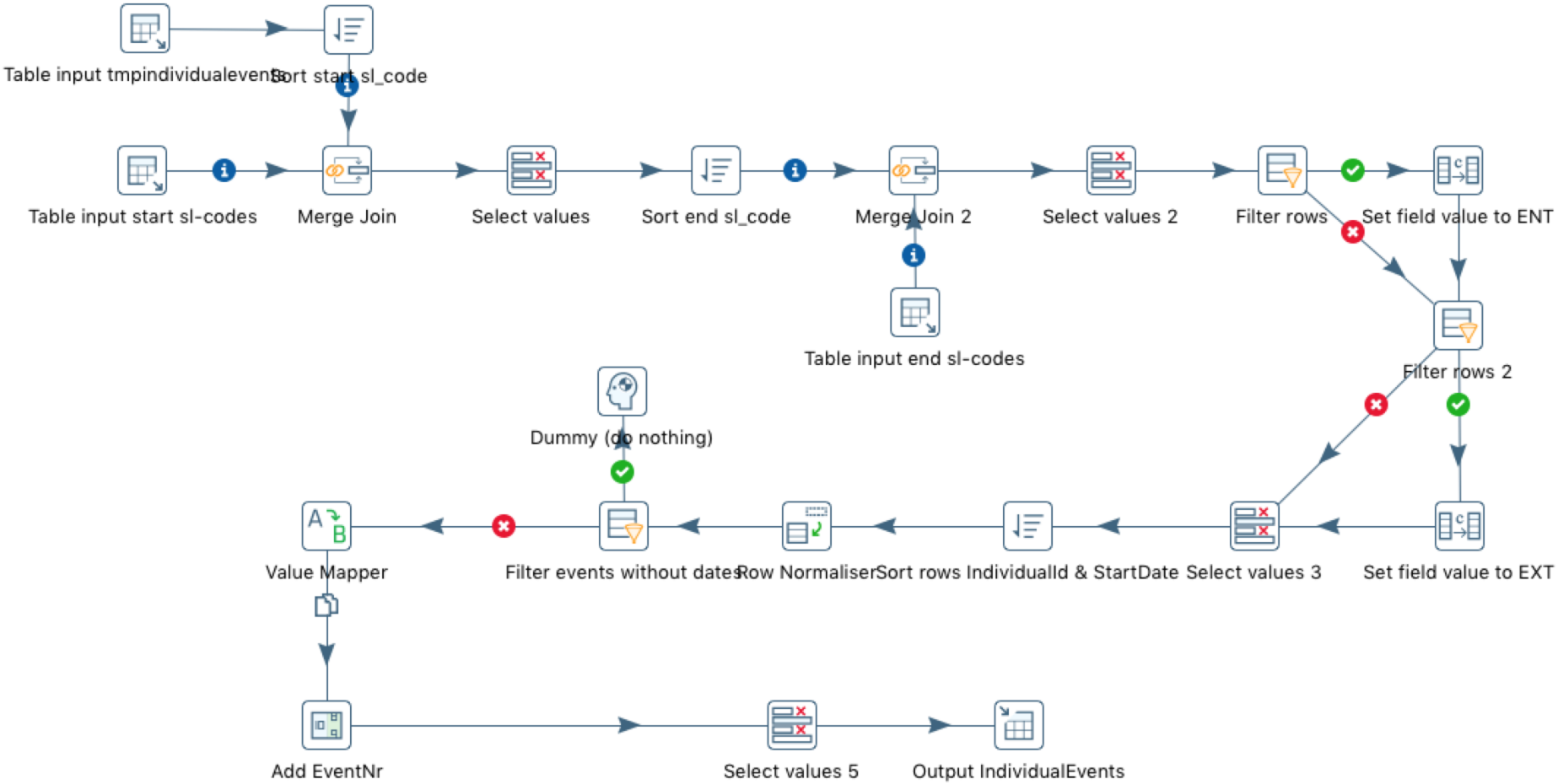
The Data Integration perspective of Spoon allows to two basic document types:

- **Transformations** are used to describe the data flows for ETL like, reading from a source, transforming data and loading it into a target location.
- **Jobs** are used to coordinate ETL activities such like, defining the flow and dependencies for what order transformations should be run, or prepare for execution by checking conditions like, "Is the source file available?," or "Does a table exist in the database?"

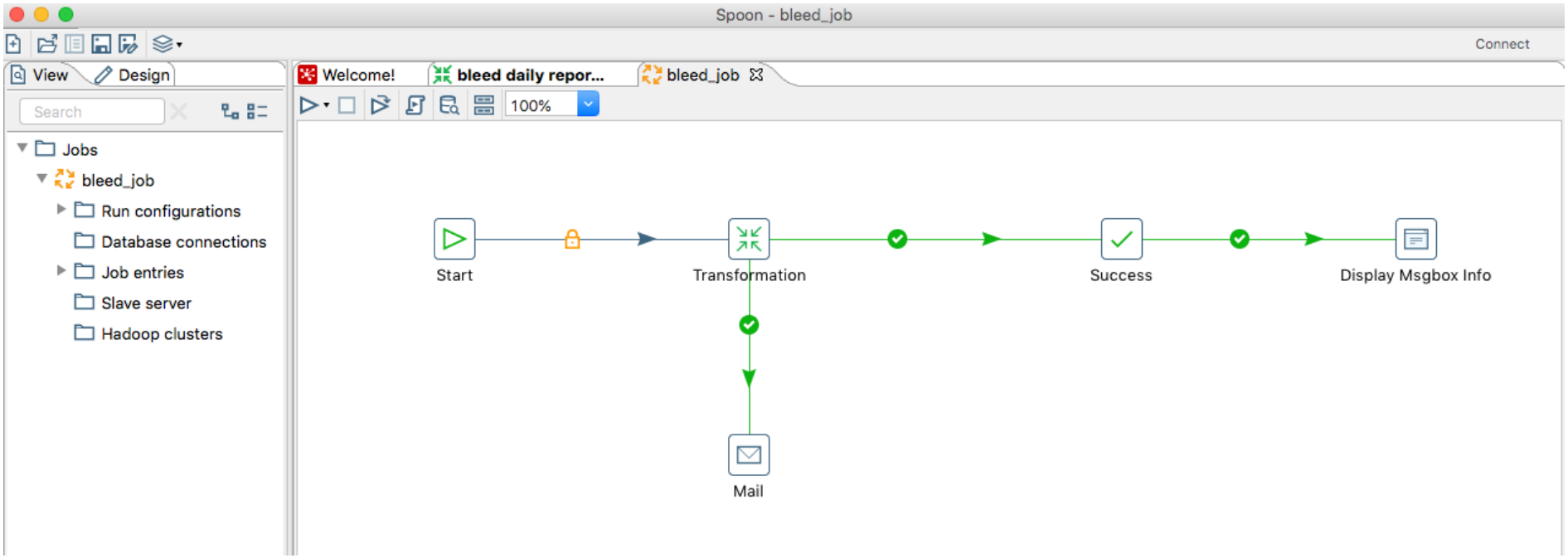
Transformation Steps

1. Create a connection to the data source
2. Input: Select the Input objects, e.g., database table
3. Transformation: sort, filter, merge, clean ...
4. Output: Select the Output destination, e.g., database, flat file ...

Sample Transformation



Sample Job


Spoon - bleed_job
Connect
View
Design
Search
Jobs
bleed_job
Run configurations
Database connections
Job entries
Slave server
Hadoop clusters
Welcome!
bleed daily repor...
bleed_job
100%
Start
Transformation
Success
Display Msgbox Info
Mail

KEMRI | Wellcome Trust

Acknowledgements

KEMRI | Wellcome Trust