Step 1 - Preparation

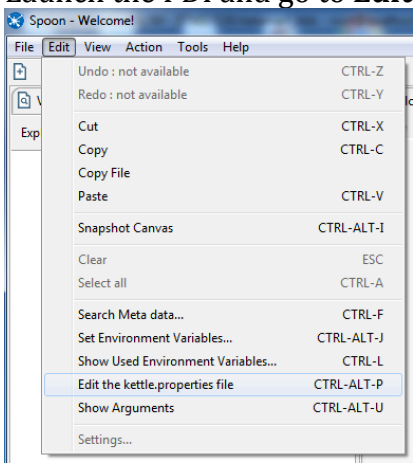
Download the zipped file, **w4-exercise.zip**, from the Blackboard (**Content → Week 04 Dimensional Modeling… → Assignments**).

Extract the files into a folder.

Run the DDL scripts in the **00-w4-ddl-scripts.sql** via your choice of a MySQL client (SQLYog, MySQL Workbench, Sequel Pro, etc). Make sure that there are 5 new tables created as below.

1. `datamart_kbb.dim_customer`
2. `datamart_kbb.dim_customer_scd1`
3. `datamart_kbb.dim_customer_scd2`
4. `source_db.city_state`
5. `source_db.customer`
6. `target_db.stg_customer`

Launch the PDI and go to **Edit → Edit the kettle.properties file**.



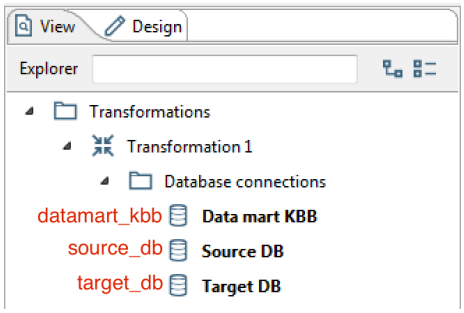
Add two variables **ETL_USER_NAME** and **ETL_USER_PASS** as the screenshot below (You can add a new variable by right-clicking on any grid and then selecting either the **Insert before this row** or **Insert after this row**). Please make sure that you enter your own database account information in the **Value** column.

Variable name	Value
ETL_USER_NAME	your_db_username
ETL_USER_PASS	your_db_password

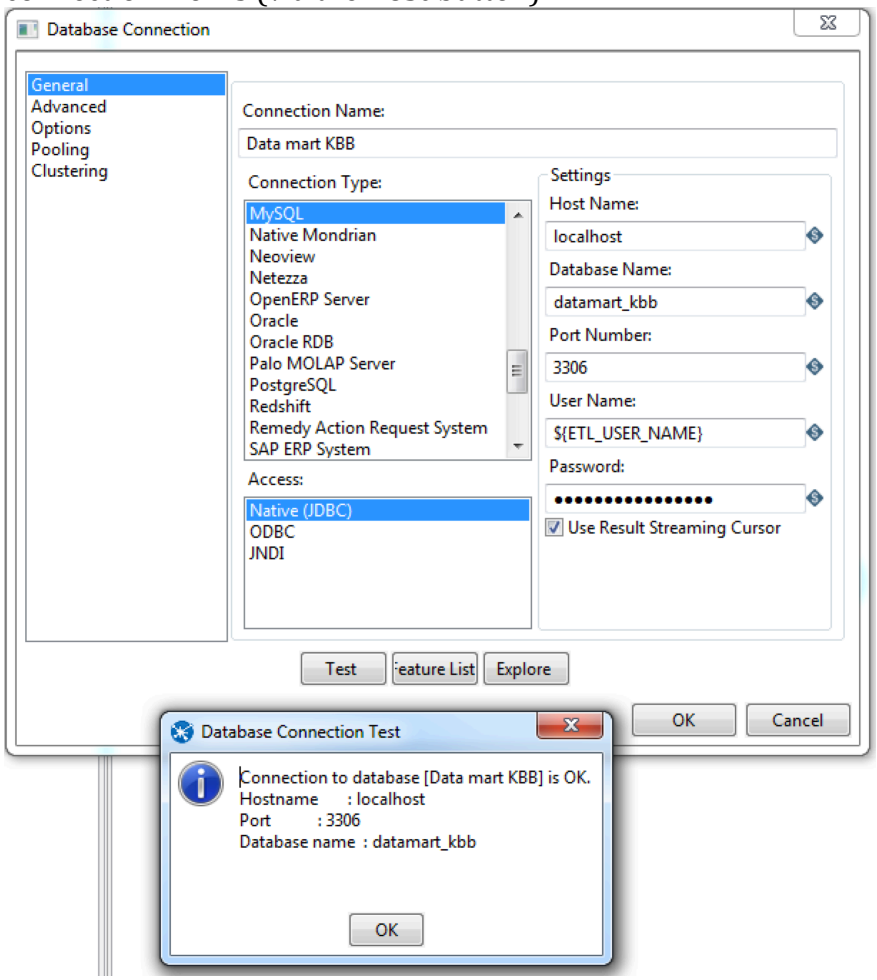
Click **OK** at the bottom.

Now create a new transformation and go to the **Database connections** on the left where you can see your already shared database connections while working on the

Week 01 and Week 02 assignments. There you should see three connections as below:



Modify all the connections (right click or double click on each connection) to use the variables you defined in the previous section (**ETL_USER_NAME** & **ETL_USER_PASS**). You can do that by entering CTRL+SPACE in the **User Name** and the **Password** and selecting the variables. Once you did, please make sure that the connection works (via the **Test** button).



Step 2 - Loading DIM_CUSTOMER (10 points)

This section is to load the customer information in the source excel file, **customer_raw.xls**, into the **dim_customer** table via both the Inmon's and the Kimball's approaches. This should be very similar to what we did for the Week 02 class exercise. If you have not participated in the week 02 class, please check out the video (the second part in particular).

The contents of the source excel file are as below:

	A	B	C	D	E	F
1	customer_id	first_name	last_name	date_of_birth	city	state
2	C40540	Tairian	Adelsen	5/30/94	Pittsburgh	Pennsylvania
3	C65558	David	Philbin	1/18/92	Pittsburgh	Pennsylvania
4	C90576	Lorren	Runner	12/26/84	Pittsburgh	Pennsylvania
5	C15595	Sarah	Cotton	5/21/93	Pittsburgh	Pennsylvania
6	C40613	Marco	Hussie	1/29/71	Pittsburgh	Pennsylvania
7	C65631	Brent	Paldino	12/28/91	Pittsburgh	Pennsylvania
8	C90650	Michael	Wilson	3/17/90	Pittsburgh	Pennsylvania
9	C15668	Gina	Harper	9/25/92	Pittsburgh	Pennsylvania
10	C40686	Molly	Worrell	10/31/92	Pittsburgh	Pennsylvania
11	C65705	Lauryn	Abid	9/27/77	Morgantown	West Virginia
12	C90723	Tairan	Adelson	6/30/93	Pittsburgh	Pennsylvania

FOR THE INMON'S APPROACH, YOU WILL CREATE…

1. **tr-inmon-load-city-state-yourID.ktr**: This transformation is to populate the **source_db.city_state** table (normalized). Please note that the **city** and the **state** columns are the composite natural key as you can see from the DDL of the table below:

```
CREATE TABLE `city_state` (
  `city_state_id` bigint(20) NOT NULL AUTO_INCREMENT,
  `city` varchar(64) DEFAULT '-',
  `state` varchar(64) DEFAULT '-',
  PRIMARY KEY (`city_state_id`),
  UNIQUE KEY `nkey_city_state` (`city`,`state`)
) ENGINE=MyISAM DEFAULT CHARSET=utf8;
```

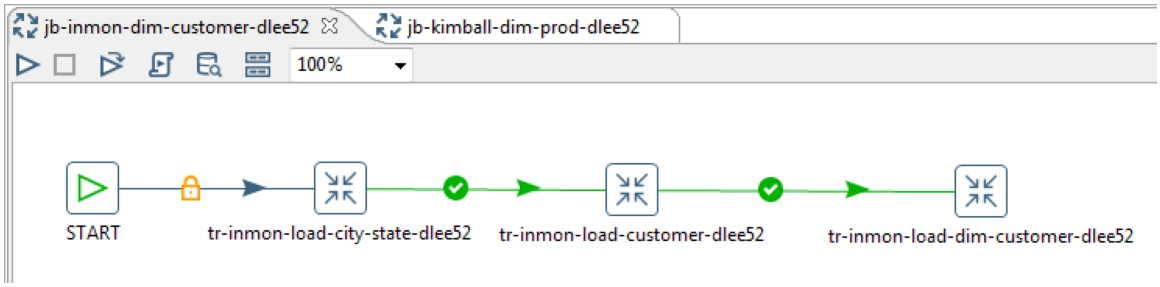
2. **tr-inmon-load-customer-yourID.ktr**: This transformation is to populate the **source_db.customer** table (normalized). The **customer_id** is the

natural key. Here is the DDL of the table:

```
CREATE TABLE `customer` (
  `id` bigint(20) unsigned NOT NULL AUTO_INCREMENT,
  `customer_id` varchar(32) NOT NULL DEFAULT '-',
  `first_name` varchar(32) DEFAULT NULL,
  `last_name` varchar(32) DEFAULT NULL,
  `date_of_birth` datetime DEFAULT NULL,
  `city_state_id` bigint(20) unsigned DEFAULT NULL,
  PRIMARY KEY (`id`),
  UNIQUE KEY `idx_customer` (`customer_id`)
) ENGINE=MyISAM DEFAULT CHARSET=utf8;
```

3. **tr-inmon-load-dim-customer-yourID.ktr**: This transformation is to populate the **datamart_kbb.dim_customer**. The two normalized tables (source_db.city_state & source_db.customer) in the previous steps are the sources in this transformation.

4. **jb-inmon-dim-customer-yourID.kjb**: This is a job to include all those three transformations you created above. For your reference, this is how my job looks like:

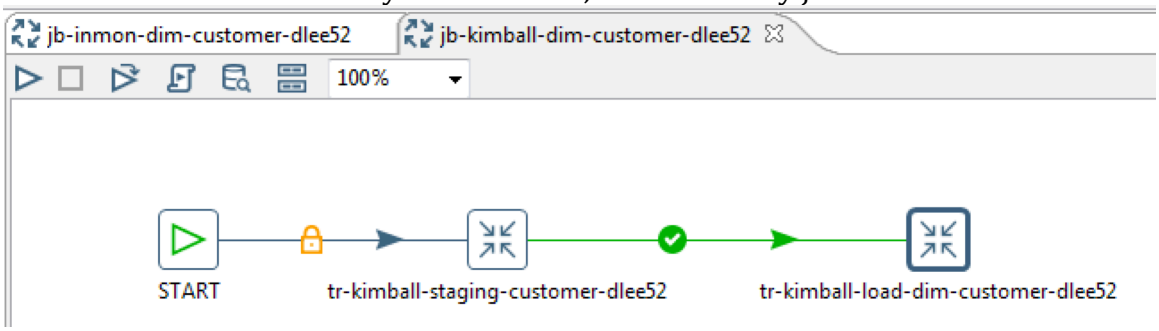


FOR THE KIMBALL'S APPROACH, YOU WILL CREATE…

1. **tr-kimball-staging-customer-yourID.ktr**: This transformation is to stage the source excel file into the **target_db.stg_customer**.

2. **tr-kimball-load-dim-customer-yourID.ktr**: This transformation is to populate the **datamart_kbb.dim_customer**. The **target_db.stg_customer** is the source input in this transformation.

3. **jb-kimball-dim-customer-yourID.kjb**: This is a job to include those two transformations above. For your reference, this is how my job looks like:



Step 3 - Loading DIM_CUSTOMER_SCD1 and DIM_CUSTOMER_SCD2 (10 points)

In this section, you are to load the customer information into the SCD Type 1 and 2 dimensions. If you have not participated in the week 04 class, please check out the class exercise files or the video (the second part in particular).

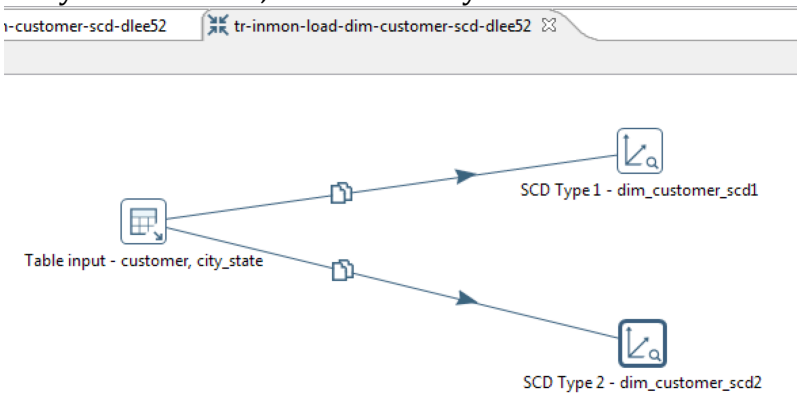
You will create…

1. **tr-inmon-load-dim-customer-scd-yourID.ktr**: This transformation uses the two normalized tables (**source_db.city_state** and **source_db.customer**) as the input sources and loads into two different SCD type tables as below:

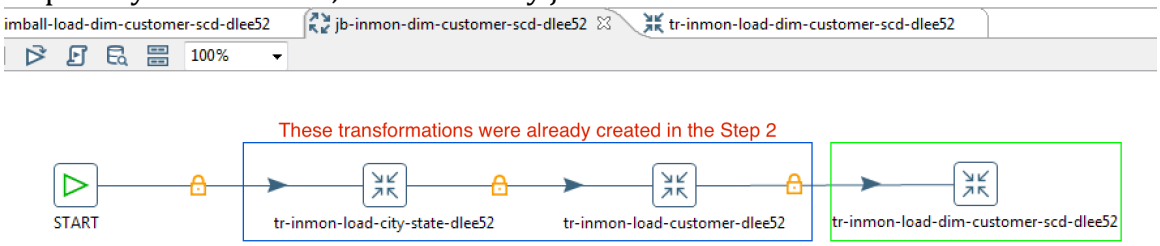
    a. **datamart_kbb.dim_customer_scd1**: treats all the attributes other than the surrogate and natural keys (**first_name, last_name, date_of_birth, city**, and **state**) as SCD Type 1 (**Punch Through** in PID's terminology).

    b. **datamart_kbb.dim_customer_scd2**: treats those **first_name, last_name, city**, and **state** attributes as SCD Type 2 (**Insert** in PDI's terminology)

For your reference, this is how my transformation looks like:

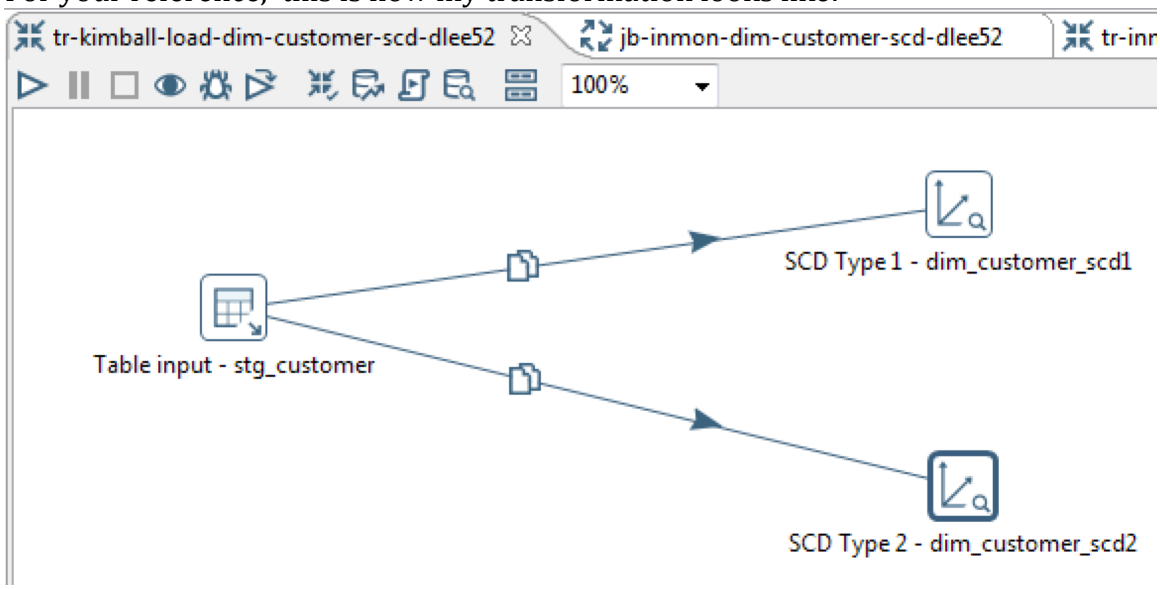


2. **jb-inmon-dim-customer-scd-dlee52.kjb**: This job includes the two transformations created in the Section 2 and the one created in the previous step. For your reference, this is how my job looks like:

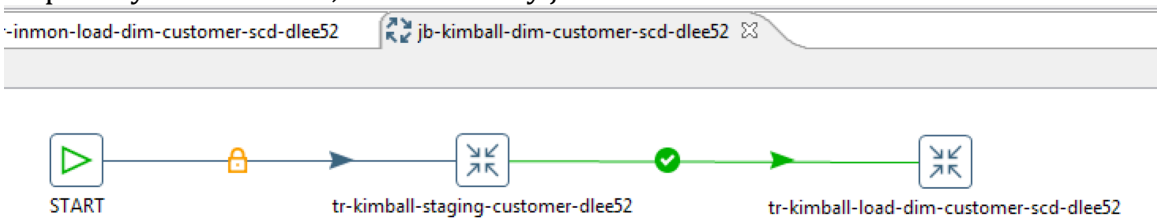


3. **tr-kimball-load-dim-customer-scd-yourID.ktr**: This transformation uses the staging table (**target_db.stg_customer**) as the input source and loads into the same SCD type tables (**datamart_kbb.dim_customer_scd1** and **datamart_kbb.dim_customer_scd2**) .

For your reference, this is how my transformation looks like:



4. **jb-kimball-dim-customer-scd-dlee52.kjb**: This job includes the transformation created in the Section 2 and the one created in the previous step. For your reference, this is how my job looks like:



Step 4 - Submit your Week 04 homework

*Submit those eight transformations, the two jobs, and your **screenshots in a single zip file**. As for the screenshots, **no snapshot is needed for the Step 2**. Just include the snapshots of the Step 3 results only. To be more specific, the snapshots must demonstrate how the changes in the **customer_raw.xls** file got reflected in the **dim_customer_scd1** and the **dim_customer_scd2** tables.*

The zipped file must follow the naming convention below:

yourID_week04.zip
(e.g. dlee52_week04.zip)