**Due**

Saturday, March 14, 2020 by 11:59 PM

**Goals**

Understanding of the structure of star-schema and concept of data cubes, differences of star schema from a normalized database schema, and how to create an example star-schema in Excel using PowerPivot along with the use of slicers in Excel pivot tables for slicing and dicing data cubes. Additionally, the assignment aims to provide developing skills in using Talend, an ETL tool, and SQL to interact, extract, transform and load data from a relational database into a data warehouse.

You may partner with up to 2 other classmates.

**Objectives**

After completion of this assignment students should be able to:

* List at least one difference between a star and a normalized schema
* Articulate the differences between measures and dimensions in data
* Identify slow changing dimensions
* Design and draw a star schema structure by inspecting different data sources (mainly tables) and structures
* Create a star-schema structure in Excel PowerPivot and use PowerPivot and pivot tables for slicing and dicing that schema
* Create aggregate measures and KPIs and implement them in MySQL and PowerPivot
* Develop SQL queries to address strategic, operational, and analytical data requests
* Insert data into a data warehouse with Talend
* Create various types of dashboards (strategic, operational, and analytical) in Tableau using the data from the dimensional models

**Preparatory Work**

Download [all the files](https://lmu.box.com/v/dimmodeling) associated with the tutorial that we completed during the lecture.

They are:

* A Tutorial on creating Star Schema with PowerPivot.docx
* Real Estate Data for PowerPivot.xlsx
* Completed PowerPivot work with RE Data.xlsx
* Understand and create date in PowerPivot.pdf

Complete the tutorials on PowerPivot and star schema.

Also complete the tutorials on Power Pivot using MySQL.

If you want additional practice, you can also try the data from the Northwind database. Instructions for that one can be obtained from [this link](https://lmu.box.com/v/starschema-practice-nwdb).

**TASKS**

1. **Initial Data Exploration**

Install [MySQL Workbench](https://dev.mysql.com/downloads/workbench/) to explore and run queries against the database you will be querying for the assignment.

If you’re on Windows, you may have to install:

* [Microsoft .NET Framework (4.6)](https://www.microsoft.com/en-us/download/details.aspx?id=53344)
* [Visual C++ Redistributable for Visual Studio 2015](https://www.microsoft.com/en-us/download/details.aspx?id=48145)

Using MySQL Workbench, log in to the Dillard’s relational database by clicking on the plus sign to create a new connection.



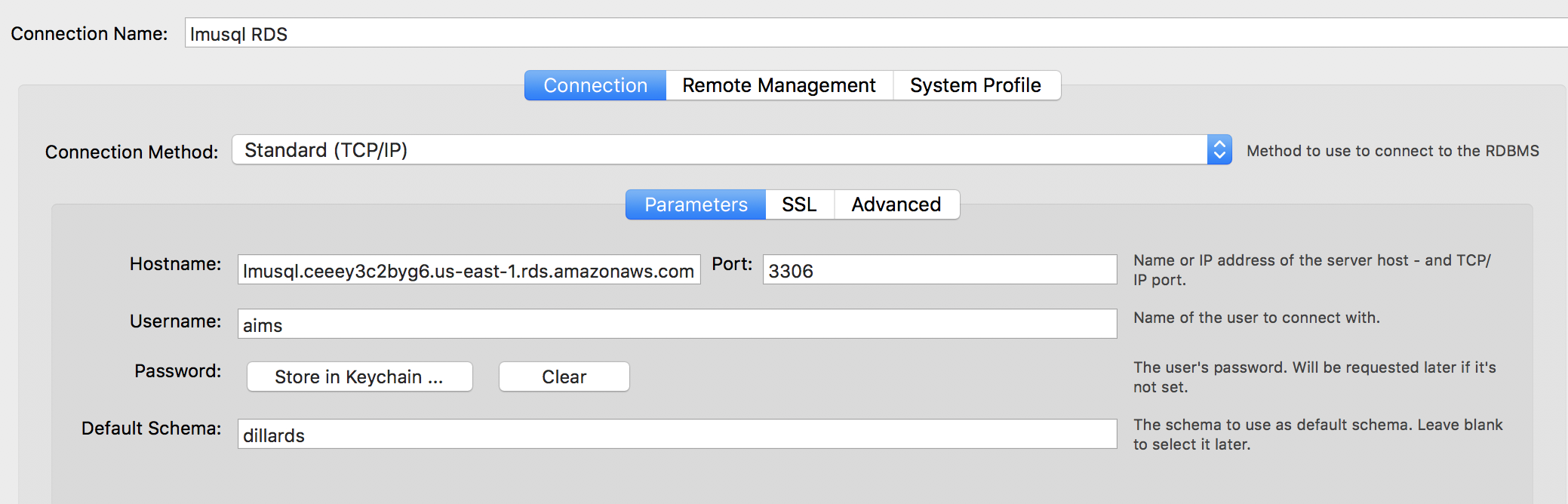
**Database Login Credentials:**

Host: lmusql.ceeey3c2byg6.us-east-1.rds.amazonaws.com

Username: analyst (The analyst username only has SELECT privileges)

Password: lmumba630!

Database: dillards

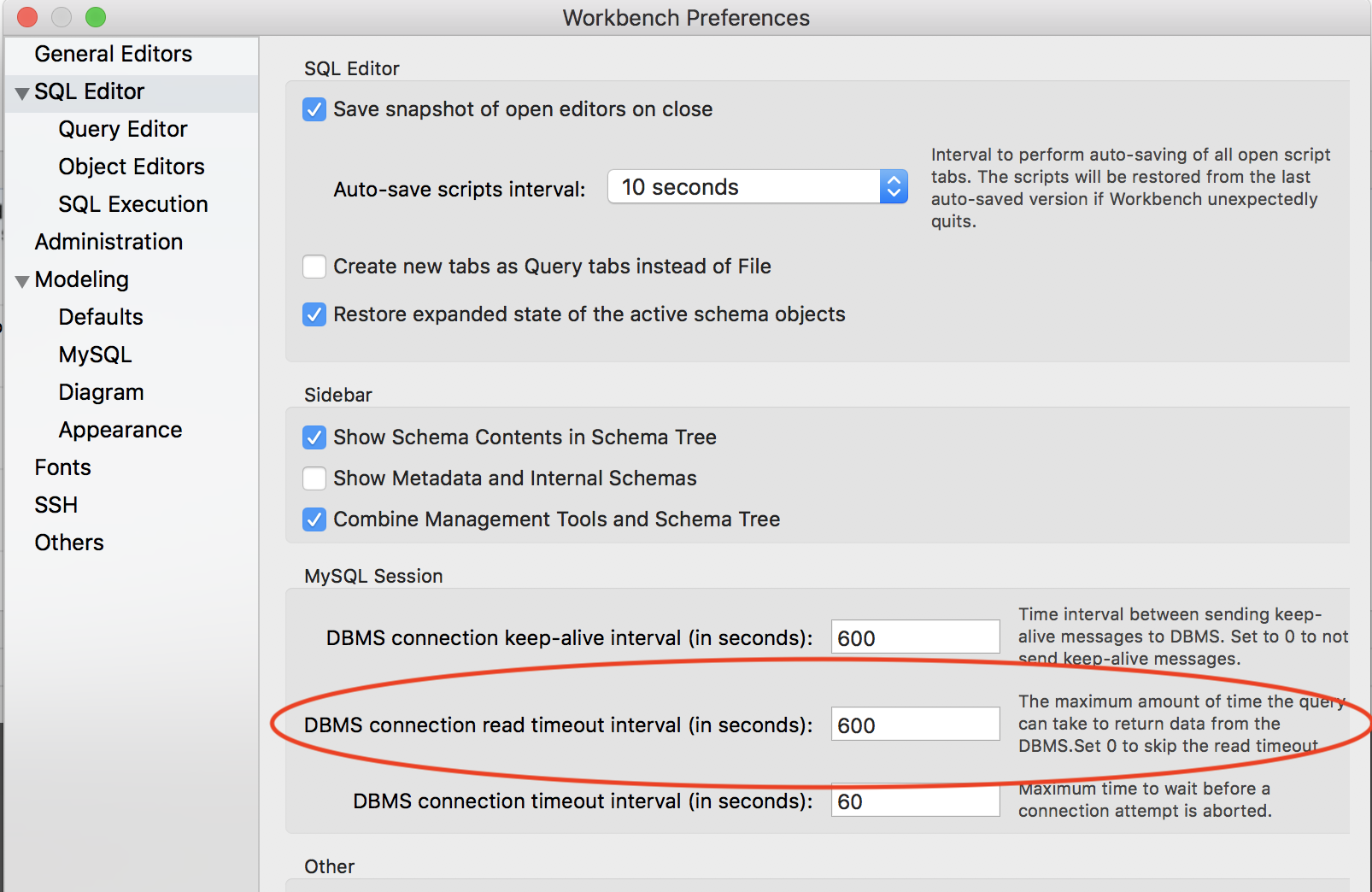


To add the password, click "Store in Keychain"

You can refer to the Dillard’s Relational Database Entity Relationship Diagram at the end of this document or you can load the Model into MySQL Workbench with [this file](https://lmu.box.com/s/4qdj1zyq4ilo5npoz9cg1sch9ijdssbo).

If you run into this error when running a query, update the timeout under Preferences from 30 seconds to 600 seconds.

“Error Code: 2013. Lost connection to MySQL server during query”



1. **Data Warehouse Purpose**

The next step is to build a dimensional model and a data warehouse using the data from Dillard’s. To be able to do that, you have to identify the facts and dimensions which are dependent on the information needs of a decision-maker. For this assignment, you are the decision-maker and assume that you will need to answer the following data requests by querying the data warehouse that you will create. Take time to fully understand the Strategic, Operational, and Analytical data requests below before designing your data warehouse as it will determine the tables and columns you will create.

**Strategic:**  
1. Total revenue per month for April, May, and June in 2005. Return the month name and the month's total revenue.  
2. Total purchase count per month for April, May, and June in 2005. Return the month name and the month's total purchase count.  
3. Total profit per month for April, May, and June in 2005. Return the month name and the month's total profit.  
  
**Operational:**  
1. Average revenue per transaction from April 1, 2005 to April 30, 2005 for stores in Texas. Return the date and the average revenue per transaction for the date.

- To get average revenue per transaction (column names would change based on the schema you created):

SUM(purchase\_revenue) / SUM(purchase\_transaction\_count) AS average\_revenue\_per\_transaction

2. Daily purchase count for the ESPRIT department from April 7, 2005 to April 14, 2005. Return the date and the date's purchase count.  
3. The 5 lowest performing stores for April 1, 2005 to April 30, 2005 based on purchase revenue. Return the store ID and the store's total revenue for the entire date range.  
  
**Analytical:**  
1. Top 10 SKUs based on quantity sold for May 7, 2005 to May 14, 2005. Return the SKU and the quantity sold for the SKU.  
2. Top 3 department and city combinations based on revenue for December 1, 2004 to December 31, 2004. Return the department and city and the revenue for the department and city combination.  
3. The number of returned items (STYPE = 'R') for each day of the week for June 2005. Return the day of the week name, i.e. Monday, Thursday, etc and its returned items total.

3. **Data Warehouse Design**

Before you create the data warehouse, identify the following based on the data requests required in part 2. Make a copy of [this spreadsheet](https://lmu.box.com/v/BSAN-6060-star-schema-Excel) to submit your answers. Save the file as ***star\_schema\_details.xlsx***

Step 1. Specify the fact table. For this assignment, you will need only 1 fact table.

1. Table name
   * 1. Should have a "\_fact" suffix
   1. Field Name
   2. Mark with a Y next to the field name
      1. Primary key?
      2. Foreign key?
      3. Measure?
      4. Calculated?
   3. Aggregate Function Used - SUM, COUNT, AVERAGE, etc
   4. The grain level should be set to date.

Step 2. Specify the dimension tables. For this assignment, you will need 3 dimension tables, which includes the date\_dim table that we would create in class. For each table, specify:

1. Table name
   * 1. Should have a "\_dim" suffix
   1. Field Name
      1. List the fields in the order of the hierarchy if one exists. Start from high to low.
   2. Hierarchy? - Mark Y next to the field name if it's part of a hierarchy
   3. Primary key? - Mark with a Y next to the field name

Step 3. Draw a Star Schema with the 3 dimension tables and the one fact table identified from the steps above. You can use any of the drawing tools available to you such as Visio, Google Drawings, [Lucidchart](https://www.lucidchart.com/pages/templates/er-diagram/database-er-diagram-template), etc. Save the file as ***data\_warehouse\_erd.pdf***

5. **Data Warehouse Implementation**

Create the fact tables and the dimensions tables in your lmu.build MySQL database. Please note that by default, date is a dimension and thus must be added to the entire schema as its own dimension table. The transaction date is used as a way to link the date dimension with the fact tables. Date dimension also has hierarchies and you need to add that. Stop at the Quarter level.

6. Download and install [Talend Open Studio for Data Integration](https://www.talend.com/products/data-integration/data-integration-open-studio/). After installing Talend, you will need to install additional modules to support connecting to MySQL:

Help -> Install Additional Packages -> Check the box next to Required third-party libraries then accept all of the licenses.

If you run into Java related errors when trying to open Talend, make sure you have Java version 8 installed for the [Java Development Kit](https://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html) and the [Java Runtime Environment](https://www.oracle.com/technetwork/java/javase/downloads/jre8-downloads-2133155.html).

Using Talend, create jobs to populate the fact and dimension tables identified above by extracting the data from the Dillard’s relational database that you explored in #1. Your input component can be tFileInputDelimited (exported csv file) or tDBInput(MySQL) (SQL query from the Dillard’s relational database).

1. Fact table
   1. Use the tAggregateRow Talend component
2. Dimension tables
   1. Use the tMap Talend component
   2. Create a job for each of the 3 dimension tables
      1. The date\_dim table was covered in class. Be sure to include the job in your deliverable.
3. Export the jobs to provide your deliverable.
   1. Selecting all of the jobs under Job Designs
   2. Right click on the selected jobs
   3. Click Export items
   4. Click “Select archive file:”
   5. Click Browse
   6. Set Save As: to talend
   7. Update the destination directory as needed
   8. Check Export Dependencies
   9. The resulting archive zip file will be named ***talend.zip***

7. Implement the star schema in PowerPivot including relevant hierarchies (some examples of hierarchies are State -> City -> Store; Department -> SKU, etc) and the aggregate measures that you have identified earlier. Remember that aggregate measures should be informative about the performance of the business (i.e. they are typically the KPIs). In PowerPivot, create KPIs for the aggregate measures. You can use an absolute value of your choice as target. Save the file as ***dillards.xlsx***

8. **Using the Data Warehouse**

Provide the following for each data request specified in #2 above:

1. SQL query to SELECT from your data warehouse
   1. Save all SQL statements in a single file named ***assignment\_01.sql***. Add a comment above the SQL to denote the task, i.e. # Strategic 2, # Analytical 3.
2. Return the results in a Pivot Table using the PowerPivot data model from #7 below. All data requests should be saved to one Excel file, ***dillards.xlsx***
3. Visualize the results with Tableau
   1. Create a dashboard for each data request category (Strategic, Operational, Analytical). Save the dashboard and its visualizations into its own Tableau Packaged Workbook. (File -> Export Packaged Workbook). You will submit 3 workbooks: ***strategic.twbx, operational.twbx, analytical.twbx***

**Deliverables:**

***star\_schema\_details.xlsx***

***data\_warehouse\_erd.pdf***

***assignment\_01.sql***

***strategic.twbx***

***operational.twbx***

***analytical.twbx***  
***dillards.xlsx***

***talend.zip***

**SQL Hints:**

\*\*\* All of the SQL statements you will provide will have a GROUP BY. \*\*\*

**To group the details for a single transaction (purchase or return):**

FROM transactions

GROUP BY STORE, REGISTER, TRANNUM, SALEDATE

**Purchase:**

FROM transactions

WHERE STYPE = 'P'

**Return:**

FROM transactions

WHERE STYPE = 'R'

**Revenue per purchase transaction:**  
SELECT SUM(AMT)   
FROM transactions

WHERE STYPE = 'P'

GROUP BY STORE, REGISTER, TRANNUM, SALEDATE;  
  
**Profit:**  
Subtract (transactions.QUANTITY \* sku\_store.COST) from transactions.AMT WHERE STYPE = ‘P’.  
Profit should be stored in the fact table as a measure.



**Dillard’s Relational Database Entity Relationship Diagram**

