

Assignment 1: Data Platforms

This assignment covers data platforms, including “classic” data warehousing, and consists of two parts:

- a) In **Part A**, which is mostly conceptual and creative, you will be put in charge of planning an analytics platform for an organization of your choice. Specifically, you will address a set of design questions and respond to management’s inquiries.
- b) **Part B** is a “hands-on” exercise where you will create a “classic” data mart for a (fictional) company, develop ETL processes to load data from the OLTP environment, set up a data cube, and run a set of queries that revolve around important business questions on it.

Instructions

Deadline

Make sure to upload **all** your results **before December 1, 2019, 23:55**.

Submission guidelines

Compress your solution folder and upload it on TUWEL using the following naming convention:

`<group no>_VU_BI_2019_Assignment_1.zip/.tgz`

Within this ZIP-file, make sure to place everything into a top-level folder named

`<group no>_VU_BI_2019_Assignment_1`

That folder should contain all your deliverables as follows:

1. A text file `group.txt` that lists your group members (Name, student id).
2. A single PDF document that describes your solutions for Part A of the assignment.

Filename: `PartA_DWH_<group no>.pdf`

3. The solution directory for part B (cf. the deliverables listed in each section).

The final directory layout of the extracted solution should hence look as follows (obviously, replacing `<groupNo>` with your group number):

```
<group no>_VU_BI_2019_Assignment_1
  <group no>_group.txt
  PartA_DWH_<group no>.pdf
  PartB_DWH_<groupNo>/
```

Please make sure to upload only one solution per group.

Assignment 1: Data Platforms

Questions

Please post general questions in the TUWEL discussion forum. You can also discuss problems and issues you are facing there. We appreciate if you help out other students to tackle general problems or questions (and may take that into account in case you are short a few points for a better grade). For obvious reasons, however, please do not post any solutions there.

For specific questions regarding the assignments, you can also contact Cornelia Michlits (cornelia.michlits@tuwien.ac.at), the tutor for this course.

Assignment 1: Data Platforms

Part A: Data Warehousing Basics and Modeling [50 points]

Pick a medium to large organization¹ that you are closely familiar with. If you are not familiar with the internals of any such organization, assume that you execute the following tasks for a typical organization in a sector that you are familiar with (or that you have researched):

1. Intro [2 points]

Describe the organization and its mission in three sentences or less.

2. OLTP [12 points]

Enumerate 3 operational systems used by the organization to run its daily business. For each of these operational (OLTP) systems:

- Describe the operations that it performs (bullet point list).
- Characterize the data that the system stores (bullet point list).
- List 5 typical questions that can be answered using that data.
- List 5 strategically important questions that cannot be answered based (only) on the data that this system holds.

3. Data Warehouse [12 points]

Assume that you have been tasked with developing a Data Warehouse for the organization. The management team has instructed you that this DWH should become a central “analytics hub” for the organization and that it should support them in their strategic decisions. To prepare planning its development, you are tasked with creating a concept for a DWH project by addressing the following questions.

- Provide 5 examples for important strategic questions that will be answered using the data in the DWH.
- Define 5 key subject areas. Then pick one of these areas and describe the data that will be stored on it in the DWH. List the source systems that will provide the various pieces of data on the subject area.
- Propose an architecture and development approach for the DWH project.² Motivate your choice and highlight the advantages and drawbacks of your proposal.
- List 5 potential challenges that you foresee in the development process (e.g., with respect to data governance, data quality, ETL procedures, value delivery etc.).

4. Data Mart [12 points]

Choose one of the subject areas you identified in step 2 and design a dimensional (i.e., Star or Snowflake) schema for a Data Mart that will cover the subject area. Make sure to include appropriate facts and choose a set of dimensions that facilitates meaningful queries.

¹ I.e., any organization with a large enough operational systems infrastructure and therefore a need for a DWH.

² Architecture: e.g., Kimball's data warehouse bus, Inmon's Corporate Information Factory, Virtual Data Warehouse, HTAP etc.

Development approach: e.g., Kimball, Inmon, Agile...

Assignment 1: Data Platforms

5. Data Lake [12 Points]

The management team of the organization has heard about recent technological developments and approached you with a few questions about data lakes.

Specifically, the management team wants to know:

- a) Whether (or not) a data lake would be a cost-effective alternative to the data warehouse solution you proposed.
- b) What use cases (business questions or entirely new business opportunities) a data lake could address that the proposed DWH cannot handle and what benefits you foresee.
- c) How a data lake would fit into your proposed architecture.
- d) Whether the introduction of a data lake would require any organizational changes.

Expected results:

- Report on each of these questions in your solution document.
- Be concise and answer succinctly (bullet point lists will typically be sufficient)!
- Include figures where appropriate (architecture, schema of your data mart design etc.)

Deliverables:

- A PDF document describing your solution. Page limit: **max.** 4 pages total, including figures; ACM two-column proceedings template (sigconf option under LaTeX).
Link: <https://www.acm.org/publications/proceedings-template>
- Note that using a wrong document format or exceeding the page limit will incur a **penalty of 10 Points!**

Assignment 1: Data Platforms

Part B: Data Warehouse Implementation [50 points]

Preliminaries

You will use a set of open-source, cross-platform tools to implement a data mart and use it to address a range of business questions. Specifically, you will need the following software:

- Java Runtime Environment
- MySQL (or MariaDB)
- MySQL Connector/J
- Pentaho Data Integration
- Pentaho Schema Workbench

The complete software stack is available for Linux, Windows, and MacOS. Compared to more full-fledged commercial BI solutions, the individual elements can be set up fairly quickly on any machine with minimal system requirements and good enough performance for basic needs. Our focus in this exercise is on developing an understanding of concepts rather than learning a particular vendor's graphical tools – the stack serves this purpose well. Please refer to the Lab Setup section of this document for installation instructions.

Introduction

Bikes&More is an online bike and bike accessories reseller operating from the U.S. that ships its products to customers in the US and Canada as well as Great Britain and Australia. The company is aggressively expanding by targeting their customers directly and offering their products on line to reduce distribution cost.

Since its inception in 2000, Bikes&More has used a reliable, but increasingly outdated order transaction processing system, which management intends to upgrade to a new, more powerful system that not only supports daily business processes, but also facilitates strategic analysis of the collected business data. In later development stages, all relevant data will be collected in a central data warehouse which will feed the data into subject-specific analytic applications.

In a pilot project, you have been tasked with developing a data mart that provides detailed access to sales data, allows the management team to analyze key business questions, and supports their decision-making.

In this pilot project, you will:

1. Implement a new schema for a transactional database and migrate legacy data to the new system.
2. Implement a star schema and create a data mart for sales analysis.
3. Define and execute an ETL process to extract, transform, and load data from the transactional database into the sales data mart.
4. Create an OLAP cube for multidimensional sales data analysis.
5. Implement a set of queries to inform management about critical business metrics.

Assignment 1: Data Platforms

Lab Setup [0 points]

1. Make sure to have a Java runtime environment properly installed
2. Set up MySQL Community (using your package manager of choice or download it from <http://dev.mysql.com/downloads/mysql/>)
3. Download and extract Pentaho Data Integration Community Edition (now part of Hitachi) from <https://sourceforge.net/projects/pentaho/>
4. Download and extract Pentaho Schema Workbench Community Edition from <https://sourceforge.net/projects/mondrian/files/schema%20workbench/>
5. Download the MySQL JDBC driver (MySQL Connector/J) from <http://dev.mysql.com/downloads/connector/j/>
6. Copy the MySQL JDBC driver to the Pentaho Data integration lib folder (i.e., data-integration/lib) and to the Pentaho Schema Workbench drivers folder (i.e., schema-workbench/drivers).
7. You can run the Pentaho tools using the provided shell scripts/batch files, i.e., data-integration/spoon.sh and schema-workbench/workbench.sh, respectively).
8. Download the csv data dump of the transactional database (needed in Step 1) from TUWEL.

Notes:

- If you are on MacOS or Linux, make sure that your class path does not contain any spaces or Pentaho Schema Workbench may fail to start (cf. <http://jira.pentaho.com/browse/PSW-95>)
- Pentaho Schema Workbench tends to have some (fairly annoying) GUI issues. Using the up and down arrows or saving and reopening the schema usually helps in case the tree hierarchy disappears.
- Pentaho Data Integration is quite powerful and straightforward; the Pentaho Schema Workbench provides an open source, cross-platform, lightweight, and self-contained OLAP engine and serves our purposes here (even though it is not the prettiest piece of software).

Assignment 1: Data Platforms

1. OLTP [5 points]

Business consultants have proposed a redesign of the logical schema of the company's operational database as part of the migration to a new system. The redesign process resulted in the database table layout depicted in Figure 1.

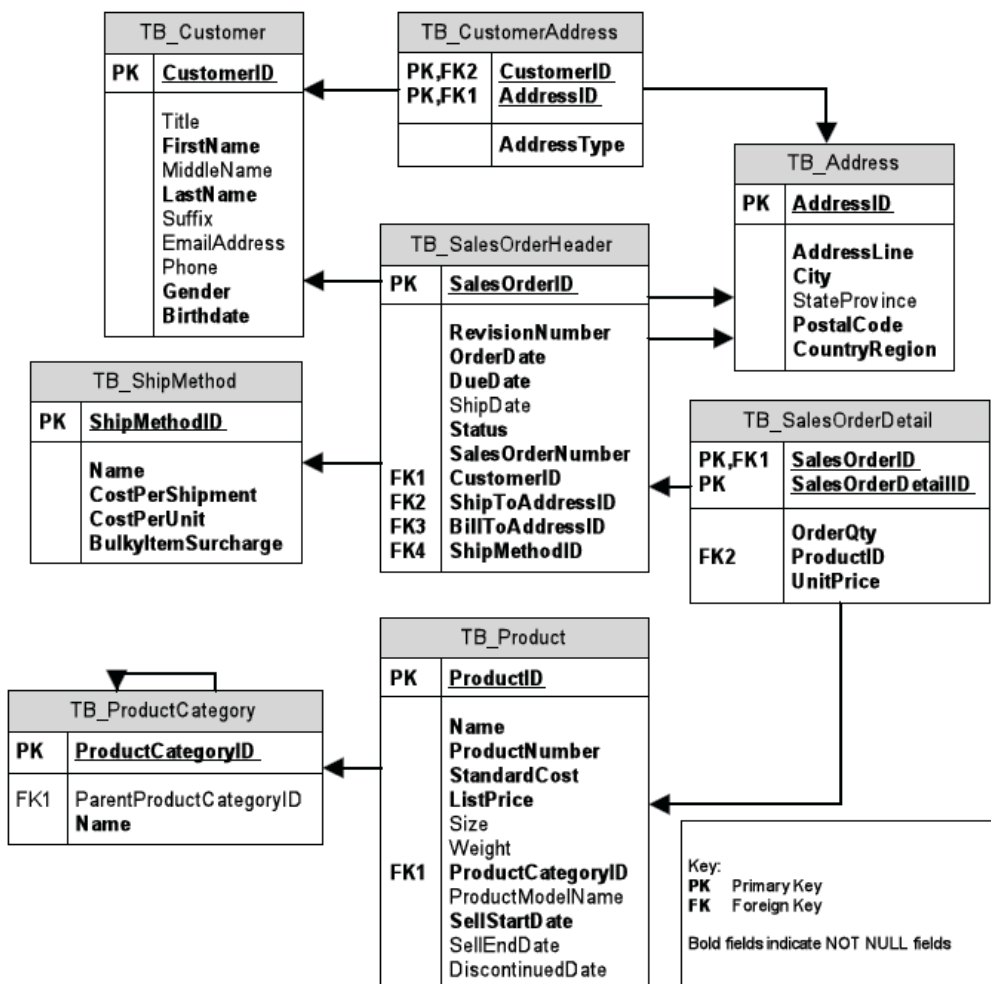


Figure 1 Logical OLTP Schema

Task description

Based on the logical relational data model depicted in Figure 1, your team is tasked with implementing the database (see Appendix: Data Dictionary) as follows:

1. Create OLTP DB:

Name your MySQL transactional database **BI_OLTP_<groupNo>**
(where **<groupNo>** is your group number)

2. Table create scripts:

For each table, create an sql script **TB_<table_name>.sql** that creates the table according to the definition specified in Figure 1.

Assignment 1: Data Platforms

a. Primary key constraints:

Define the necessary primary keys as table constraints and name them PK_<table_name>, where <table_name> is replaced with the respective name of the table, e.g. for table TB_Address the primary key is named PK_Address.

b. Foreign Key Constraints:


Name the foreign key constraints FK_<column_name>_<tableName>, where <column_name> is the name of the column without the 'ID' suffix that represents the reference and <tableName> is the name of the table that stores the foreign key – for example, the references between TB_SalesOrderHeader and TB_Address are represented by the columns BillToAddressID and ShipToAddressID. The corresponding foreign key constraints should be named FK_BillToAddress_SalesOrderHeader and FK_ShipToAddress_SalesOrderHeader, respectively (this should ensure that the foreign key names are unique).

Place the scripts in your solution folder **task 1/oltp/TB_<tableName>.sql**.

3. OLTP create job:

Create a Data Integration job in Pentaho Spoon to invoke the table create scripts from step 1 in the appropriate sequence.

Detailed steps:

- a. Create a Data Integration Repository
 - i. Tools → Repository → Connect..
 - ii. Click the + to add a repository
 - iii. Choose Kettle File repository
 - iv. Select a base directory and assign BI_<groupNo> as a name (where <groupNo> is your group number)
- b. Create a Database connection
 - i. Click the  symbol and select Database connection
 - ii. Make sure you have installed the JDBC driver (MySQL connector/J)
 - iii. Set the connection parameters
 1. Access: "Native (JDBC)"
 2. Host name: "localhost"
 3. Database Name: "BI_OLTP"
 - iv. Name the connection BI_OLTP_<groupNo>
- c. Create an integration job and name it "OLTP create". Include it in your solution folder as OLTP create.kjb. (should work automatically if you create a Kettle file repository in your solution folder).
- d. Add the SQL create scripts to the integration job (using execute SQL steps) in an appropriate sequence.
- e. Execute the integration job to create your OLTP database

4. Load data

The data of the legacy transactional database has already been converted to the new schema and dumped into tab-separated files (download them from TUWEL). To load the

Assignment 1: Data Platforms

data into your OLTP database, create a new integration job named OLTP load in Spoon, add the database connection, and use “MySQL bulk” load steps to load the data dumps for table in the appropriate sequence (i.e., ensure that the database is in a consistent state after loading).

Note: The tsv files are tab-delimited, i.e., when setting up the MySQL bulk load steps, you need to remove the comma from the delimiter field and click “Insert TAB”. Make sure to load the data into the correct columns (the sequence of columns in the csv file is specified in the Appendix).

Deliverables

Put all the files created in this task into the task1 folder.

Resources

- Pentaho Data Integration (Kettle) Tutorial
[http://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+\(Kettle\)+Tutorial](http://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+(Kettle)+Tutorial)

2. OLAP [25 points]

Your team has been tasked with the implementation of a sales analysis data mart as the first step towards the development of a comprehensive business intelligence system for the company. To this end, you have conducted extensive requirements elicitation interviews with the responsible managers of the sales, logistics and products departments. Your task now in this step is to implement the corresponding star schema in a separate database as well as the process that extracts, transforms and loads the data from the transactional database.

In the following, you will find the definition of the required information package that should be implemented in a star schema (Table 1), as well as a specification of the transformations that need to be performed (Table 2).

Time	Customer	Product	Location
<u>Date</u>	<u>CustomerID</u>	<u>ProductID</u>	<u>AddressID</u>
Day Number of Month	Name*	Product Number	Postal Code
Month Number of Year	Birth Date	Name	City
Calendar Year	Age*	Model Name	State Province
	Gender*	Standard Cost	Country Region
	Email	List Price	
	Phone	Product Sub Category	
		Product Top Category	
		<i>Sell Start Date</i>	
		<i>Sell End Date</i>	
		<i>Discontinued Date</i>	
		Size	
		Weight	
		Is Bulky Item*	

FACTS: Sales Order Number, Sales Order Line Number*, CustomerID, ProductID, ShipToAddressID, BillToAddressID, Shipment Method*, Unit Price, Discount*, Order Quantity, Order Line Total*, Order Line Profit*, Tax Amount*, Order Line Freight Cost*, Order Status, Order Date, Due Date, Ship Date, Is Late Shipment*

Table 1 Information Package: Bikes & More sales. Underlined items denote the primary keys, italic items denote foreign keys, and items with an asterisk (*) require transformations as specified in Table 2

Assignment 1: Data Platforms

Attribute Name	Transformation Instruction
Customer Name VARCHAR(150)	concatenate first, middle, last name and name suffix to form the customer's name in the following format: <code>FirstName [MiddleName.] LastName[, Suffix]</code> . Note that NULL values should be omitted.
Gender VARCHAR(10)	Convert 'M' to 'Male' and 'F' to 'Female'
Age INTEGER	Calculate the age from the customer's birth date. Use '2016---01---01' as end date.
Is Bulky Item (INTEGER: 0 = false <>0 = true)	<i>true</i> if the Product is a bike, i.e. its top product category is 'Bikes', or if the Product is a bike frame, i.e. its product category ID is either 16, 18, or 20, <i>false</i> otherwise.
Sales Order Line Number VARCHAR(100)	the Sales Order Line Number is built by concatenating 'SOL' + SalesOrderID + '-' + SalesOrderDetailID Note that the resulting string must not contain any whitespaces.
Ship Method VARCHAR(100)	Use the shipment method's name for this field.
Discount NUMERIC(10,4)	Calculate the discount according to the following formula: $Discount = OrderQuantity * UnitPrice * DiscountRate$ cf. Table 3 for the applicable discount rates by Top Product Category. Note: You may truncate data after the fourth decimal place.
Order Line Total NUMERIC(10,4)	calculate the total payable amount by order line with the formula $LineTotal = OrderQuantity * UnitPrice - Discount$
Order Line Profit NUMERIC(10,4)	calculate the profit per order line with the formula $Profit = LineTotal - OrderQuantity * StandardCost$
Tax Amount NUMERIC(10,4)	calculate the tax amount of an order line with the formula $TaxAmount = VAT * LineTotal$ cf. Table 4 for the applicable country-specific VAT rate values (use the country of the shipping address to look up the applicable tax rate)
Order Line Freight Cost NUMERIC(10,4)	calculate the freight cost for each line as follows (the BulkyItemSurcharge should be applied only to bulky items): $Freight = (OrderQuantity * PerItemCost) + BulkyItemSurcharge$ cf. Table 5 for the <i>per shipment</i> and <i>per item costs</i> as well as the <i>bulky item surcharge</i> based on the chosen shipment method. <i>Note:</i> Freight cost per order line refers to order lines, i.e., subsets of a shipment. The order line freight cost does therefore not include the per-shipment cost.
Is Late Shipment (INTEGER: 0 = false <>0 = true)	<i>true</i> if the Ship Date lies past the Due Date, <i>false</i> otherwise

Table 2 OLTP to OLAP Transformations

Assignment 1: Data Platforms

Top Product Category		Minimum Order Quantity	Discount Rate
1	Bikes	---	none
2	Components	---	none
3	Clothing	5	5%
		10 and above	10%
4	Accessories	5	4%
		10 and above	11%

Table 3 Discount Rates per Top Product Category

Country	VAT Rate
United States	8%
Canada	13%
United Kingdom	17.5%
Australia	10%

Table 4 VAT rates by country/region

Shipment Method	Per Shipment Cost	Per Item Cost	Bulky Item Surcharge
Standard	\$4	\$2	+\$5
Cargo (International)	\$10	\$6	+\$10
Overseas Deluxe	\$30	\$7	+\$12

Table 5 Freight Cost by Shipment Method

Assignment 1: Data Platforms

Task descriptions

2a) Data Mart creation

In this step, your team will design a snowflake schema and implement the data mart.

1. Create a database named BI_OLAP_<groupNo>.
2. Write the SQL scripts that create the necessary fact and dimension tables and place them in a folder named task2/DM/DM_<table name>.sql.
 - Name the fact table DM_FactSales
 - Name the dimension tables DM_<dimension_name>
 - Omit the white space characters for the column names, i.e. *'Is Bulky Item'* becomes *'IsBulkyItem'*
 - Make sure to define all foreign key constraints.
3. Create a data integration job that executes the create scripts in an appropriate sequence.
4. Run your job to create the data mart

Notes:

- "Date" is a keyword in SQL that you need to escape properly (i.e., `Date` in MySQL; "Date" can only be used if Ansi SQL mode is enabled)

2b) ETL

In this step, your team will develop the required ETL processes that extract the data from the operational database, transform the data according to transformation instructions specified in Table 2, and load the data into the data mart's fact and dimension tables.

You will use Pentaho Data Integration's Spoon to define the ETL process. For the actual transformations, you have two options: you can either define the transformations in plain SQL or define the ETL jobs directly in Spoon, using the built-in transformation steps. In either case, you should create a data integration job that loads the data into the data mart in an appropriate sequence to ensure consistency.

1. Implement the transformations in plain SQL or as Kettle transformations in spoon. Name the transformations or ETL scripts according to the fact/dimension name. Place the resulting files in task 2/etl in your solution folder.
2. Create a data integration job that executes the sql scripts or transformation in an appropriate sequence and loads all data into the data mart.
3. Execute your ETL process to populate the data mart from the transactional database.

Notes: If you implement your SQL process in plain SQL, use INSERT INTO <table_name> SELECT ... commands.

2c) ROLAP Cube definition

Once the data mart has been populated, you will use Pentaho Schema Workbench to define an OLAP cube, including appropriate hierarchies for rollup and drill down analyses. In the final part of this hands on (task 3), you will then use the created OLAP cube to formulate Multidimensional Expression Query language (MDX) queries that you will report to management.

Assignment 1: Data Platforms

1. Start Pentaho Schema Workbench and establish a connection to your OLAP database (Options → Connection...; make sure the MySQL JDBC driver is placed in the schema workbench drivers folder)
2. Create a schema (File → New → Schema) and save it as task2/OLAP/bike_sales.xml in your solution folder
3. Add a cube and name it bike_sales. Add the following measures:
 - a. Profit (column OrderLineProfit)
 - b. Revenue (column OrderLineTotal)
 - c. Quantity (column OrderQuantity)
 - d. Discount
 - e. FreightCost (column OrderLineFreightCost)
 - f. TaxAmount
 - g. IsLateShipment

Choose appropriate aggregators for your measures and "Currency" formatString where appropriate.

4. Add the Date dimension
 - a. Add the dimension on the top level as a direct subnode of Schema
 - b. Make sure to set it up as a time dimension (i.e., choose type TimeDimension).
 - c. Add a single hierarchy named Days and set the primaryKey field to Date) and add the DM_Time dimension to it.
 - d. Add the three temporal levels Year, Month, and Day. For each level, choose the respective table (CalendarYear, MonthNumberOf, DayNumberOfMonth) and column and choose the correct levelType (TimeYears, TimeMonths, TimeDays).

5. Define the Location dimension:

Add Location as a subnode of Schema and add a hierarchy Areas with levels CountryRegion, State, and City. Like in the previous step, set the respective tables, columns and keys.

6. Define the Product dimension:

- a. Add Product as a subnode of Schema
- b. Add a hierarchy ProductCategory with levels TopCategory, and SubCategory. Again, set the respective tables, columns and keys.
- c. Add a second hierarchy Name with a single level Name

7. Define the Customer dimension with the following three single-level hierarchies and set the respective tables, columns and keys:
 - a. Age
 - b. Gender
 - c. Name

8. Add dimension usages to the bike_sales cube:
 - Customer (with foreign key CustomerID)
 - Product (with foreign key ProductID)

Assignment 1: Data Platforms

- ShippedTo (using Location with foreign key ShipToAddressID)
- BilledTo (using Location with foreign key BillToAddressID)
- ShipDate (using Date with foreign key ShipDate)
- OrderDate (using Date with foreign key OrderDate)

Make sure you set the foreignKey and source fields of each dimension usage.

Notes

Schema Workbench does not warn you about missing definitions (e.g., table, primary key for dimensions) – if you don't get any values for your MDX query, check if you have set everything correctly.

Deliverables

Put the file(s) created in this task into the *task2* folder.

Resources

- Mondrian Documentation: <http://mondrian.pentaho.com/documentation/index.php>
- Youtube playlist: Pentaho Workbench OLAP Cubes – Using Pentaho Schema Workbench <http://bit.ly/2gpEcP7>

3. Queries [20 points]

A number of users have contacted your team and requested various custom queries, which have been collected and are specified below. Your team is assigned the task to deliver the requested information using either plain SQL or MDX query language. You can choose which language is more appropriate to answer each of the questions, but you should formulate and **hand in at least three MDX queries** executed in Schema workbench.

Task description

Implement each of the following queries using either plain SQL or MDX targeting the data mart or the OLAP cube, respectively. For each query, example result set tables with sample data are provided. For MDX queries, the (albeit less useful) default result set format is fine. To develop your MDX queries, use File → New → MDX Query in the Schema Workbench.

A) Sales

1. What was the profit per top product category for calendar year 2015, sorted by profit in descending order?

Top Product Category	Profit
Bikes	90000
Components	85000
...	...

Assignment 1: Data Platforms

2. What was the total revenue per region for calendar year 2014 (use *ShipToAddress* and *Order Date*, sort by *Top Region*)?

Region	Revenue
United States	70000.00
Canada	80000.00
...	...

3. What is the revenue trend by the top product categories over the last available 24 month by month (use the *Bill to Address* and *Order Date*, sort by *Year*, *Month*, *Top Region*)?

Year	Month	Top Product Category	Monthly Revenue
2013	1	United States	3450
2013	1	Canada	2960
2013	2	United States	3120
...

4. What are the top 5 most profitable customers and their 4 top most frequently purchased products (sort by *Customer Rank*, *Product Rank*)?

Customer Rank	Customer Name	Product Rank	Product Name
1	Joe Short	1	Touring Rear Wheel
1	Joe Short	2	Touring Front Wheel
...
2	Sara North	1	Women's Tights, S
...

5. What are the top 10 most profitable customers in the first half year of 2015 (use the *Order Date*, sort by *Profit*)?

Customer Name	Profit
Joe Short	2430.00
Sara North	2350.00
Thomas Wayne	2160.00
...	...

B) Production

1. What are the 3 least selling products of each of the top product categories (sort by *TopProduct Category*, *Product Rank*) overall?

Top Product Category	Product Rank	Product Name
Bikes	n-1	Mountain-100 Silver, 44
Bikes	n	Road-250 Red, 44
...
Components	n-1	LL Crank set
...

Assignment 1: Data Platforms

C) Logistics

1. What is the percentage of late shipments compared to total shipments with respect to the shipment method by country/region (use the *Ship to Address*, sort by *Country/Region, Shipment Method*)?

Shipment Method	Country Region	% of Late Shipments
Cargo International	Australia	1.8000
Overseas Deluxe	Australia	0.5400
Standard Ground	Canada	1.2900
...

D) Accounting

1. What is the total of charged taxes for January through July of 2014 by month?

Month	Tax Payments
1	21000.00
2	19500.00
...	...

Deliverables

Create a file for each query solution named according to the schema `<department_numbering><#>.<query_type>` where `<department_numbering>` is either A, B, C, D representing the Sales (A), Production (B), Logistics (C) or Accounting (D) department, respectively, `<#>` the query number, and `<query_type>` the query format (SQL or MDX) chosen to solve the query. E.g. if the second query of the Sales team was solved with a MDX query, the file should be named `A2.mdx`, on the other hand, if it was solved with SQL, it should be named `A2.sql`.

Note: Do not put any comments in these files!

Instead, if necessary, create a separate text file named `comments.txt`.

Put all your files (including your `comments.txt` if present) into a subfolder named `task3`.

Resources

- A Brief MDX Tutorial Using Mondrian <http://cobb.typepad.com/files/mdx.pdf>

Assignment 1: Data Platforms

Appendix: Data Dictionary for Part B

Introduction/Instructions

This is a simple set of sample data files comprised of 8 files to be used to perform the lab examples. This data has already been extracted from the company's legacy order transaction systems and is ready to be loaded into the newly installed OLTP system. Each file is briefly described below. Most of the field names are self-explanatory, but in cases additional information is provided.

All source files were created as text flat files using:

- Code page 1252 (ANSI – Latin I)
- English Locale (United States)
- Data fields are Tab delimited

File: Address.csv

Fields	Key	Type	Relates to (File/Field)
AddressID	yes	Integer	CustomerAddress.csv/AddressID SalesOrderHeader.csv/ShipToAddressID SalesOrderHeader.csv/BillToAddressID
AddressLine		String	
City		String	
StateProvince		String	
CountryRegion		String	
PostalCode		String	

File: Customer.csv

Fields	Key	Type	Relates to (File/Field)	
CustomerID	yes	Integer	CustomerAddress.csv/CustomerID SalesOrderHeader.csv/CustomerID	
Title		String		
FirstName		String		
MiddleName		String		
LastName		String		
Suffix		String		
EmailAddress		String		
Phone		String		
Gender		String		'M' for 'Male', 'F' for 'Female'
Birthdate		Date (YYYY---MM---DD)		

Assignment 1: Data Platforms

File: CustomerAddress.csv

Fields	Key	Type	Relates to (File/Field)
CustomerID	yes	Integer	Customer.csv/CustomerID
AddressID		Integer	Address.csv/AddressID
AddressType		String	

File: SalesOrderHeader.csv

Fields	Key	Type	Relates to (File/Field)	
SalesOrderID	yes	Integer	SalesOrderDetail.csv/SalesOrderID	
RevisionNumber		Integer		Incremental number to track changes to the sales order over time.
OrderDate		Timestamp		Dates the sales order was created.
DueDate		Timestamp		Date the order is due to the customer.
ShipDate		Timestamp		Date the order was shipped to the customer.
Status		Integer		Order current status. 1 = In process 2 = Approved 3 = Back ordered 4 = Rejected 5 = Shipped 6 = Canceled
SalesOrderNumber		String		Unique sales order identification number.
CustomerID		Integer	Customer.csv/CustomerID	
ShipToAddressID		Integer	Address.csv/AddressID	Customer shipping address.
BillToAddressID		Integer	Address.csv/AddressID	Customer billing address.
ShipMethodID		Integer	ShipMethod.csv/ShipMethodID	

File: SalesOrderDetail.csv

Fields	Key	Type	Relates to (File/Field)	
SalesOrderID	yes	Integer	SalesOrderHeader.csv/SalesOrderID	
SalesOrderDetailID		Integer		
OrderQty		Integer		Quantity ordered per product.
ProductID		Integer	Product.csv/ProductID	
UnitPrice		Numeric		Selling price of a single product.

Assignment 1: Data Platforms

File: Product.csv

Fields	Key	Type	Relates to (File/Field)	
ProductID	yes	Integer	SalesOrderDetail.csv/ProductID	
Name		String		Name of the product.
ProductNumber		String		Unique product identification number.
StandardCost		Numeric		Standard cost of the product.
ListPrice		Numeric		Selling price.
Size		String		Product size.
Weight		Numeric		Product weight.
ProductCategoryID		Integer	ProductCategory.csv/ProductCategoryID	Product is a member of this product category.
ProductModelName		String		Product is a member of this product model.
SellStartDate		Timestamp		Date the product was available for sale.
SellEndDate		Timestamp		Date the product was no longer available for sale.
DiscontinuedDate		Timestamp		Date the product was discontinued.

File: ProductCategory.csv

Fields	Key	Type	Relates to (File/Field)	
ProductCategoryID	yes	Integer	ProductCategory.csv/ProductCategoryID Product.csv/ProductCategoryID	
ParentProductCategoryID		Integer		
Name		String		Category description.

File: ShipMethod.csv

Fields	Key	Type	Relates to (File/Field)
ShipMethodID	yes	Integer	SalesOrderHeader.csv/ShipMethodID
Name		String	
CostPerShipment		Numeric	
CostPerUnit		Numeric	
BulkItemSurcharge		Numeric	