Project 1: Data Warehousing (Descriptive analytics)

Brief Summary

Given a concrete Use Case (i.e., ACME-Flying Use Case), the aim of this project is to design a Data Warehouse that allows to perform descriptive analysis (i.e., OLAP). The project consists of two tasks, that need to be submitted as two separate submissions and will be evaluated separately:

1. Design a multidimensional model (i.e., star schema) that complies with the requirements defined in the statement (see Section 1). In this step you will use the Indyco Builder software.
2. Develop an ETL process that allows to extract, transform and load the data from existing databases (i.e., AMOS and AIMS) to the schema developed in the first step (see Section 2). Here you will use the Talend Open Studio for Data Integration software.

1 Multidimensional Modeling for the ACME-Flying Use Case

Note: *This part of the project is supposed to start on the second week, but you can plan the work the way it suits you best.*

You must create a multidimensional model, potentially consisting of different stars, that allows to easily retrieve the KPIs about aircraft utilization (namely FH, TO, ADIS, ADOS, ADOSS, ADOSU, DU, DC, DYR, CNR, TDR and ADD) and logbook reporting (namely RRh, RRc, PRRh, PRRc, MRRh and MRRc). All these metrics would be obtained from the available data in the databases of AIMS and AMOS. These data will need to be complemented with the following sources:

- A file containing for every aircraft registration code, its manufacturer registration code, the aircraft model and manufacturer.
- Another file containing the maintenance personnel, their identifier and the code of the airport where they work.

Finest temporal granule for FH and TO is the day, while the ADIS and ADOS are calculated per month or year, like the DYR, CNR, TDR, ADD, and all Report Rates. The latest can also be analysed per person and airport of the reporting person (just in case of MAREP), as well. Thus, required queries are:

a) Give me FH and TO per aircraft (also per model) per day (also per month and per year).

b) Give me ADIS, ADOS, ADOSS, ADOSU, DYR, CNR, TDR, ADD per aircraft (also per model) per month (also per year).

c) Give me the RRh, RRc, PRRh, PRRc, MRRh and MRRc per aircraft (also per model and manufacturer) per month (also per year).

d) Give me the MRRh and MRRc per airport of the reporting person per aircraft (also per model).

Deliverables

1. Indyco folder (in a single zip file) with all the files of the conceptual design
2. Commented SQL file with the CREATE TABLE statements and auxiliary views (if any) in PostgreSQL
3. PDF file (one single A4 page, 2.5cm margins, font size 12, inline space 1.15) with all assumptions made and justifying the decisions you made (if any)

Assessment criteria

i) Conciseness of explanations (only first page will be considered in the evaluation)

ii) Understandability

iii) Coherence

iv) Soundness

Evaluation

- 60% Deliverables
- 40% Exercises related to the project done individually in the day of the partial exam

2 Extract-Transform-Load (ETL) Process Design for the ACME-Flying Use Case

Note: *This part of the project is supposed to start on the third week, but you can plan the work the way it suits you best.*

You must create an Extract-Transform-Load (ETL) process that can be executed in order to **extract** data from the AIMS and AMOS operational databases and additionally provided data sources, **transform** these data to conform to the star schemas previously defined in the lab on Data Warehouse design, and **load** the data into the created star schemas.

The designed ETL process should adhere to the following instructions:

Extraction

- Connect to the operational databases AIMS and AMOS for extracting the base operational data.
- In addition, you should use additional data sources (aircraft-manufaturerinfo-lookup.csv and maintenance-personnel-airport-lookup.csv) and thus extract them.

Transformation

- Integrate data coming from AIMS and AMOS data sources. You should consider integrating these two sources having in mind the two common attributes that they share, i.e., flightID (in tables AIMS $->$ Flights and AMOS $->$ OperationInterruption), and aircraftRegistration (in tables AIMS $->$ Slots and AMOS $->$ MaintenanceEvents, WorkOrders).
- Complement the operational data coming from AIMS and AMOS by means of performing a lookup to the external data sources about:
  - **Aircraft manufacturer information** (aircraft-manufacturerinfo-lookup.csv) such that with each aircraft registration code, your ETL also provides its manufacturer registration code, the aircraft model and manufacturer.
  - **Maintenance personnel employment place** (maintenance-personnel-airport-lookup.csv) such that for each person from the maintenance personnel (i.e., reporteurID from table TechnicalLogBookOrders), your ETL also provides information at which airport this person works.

- Improve the quality of the source data by means of but not limited to removing duplicates/overlaps, removing incomplete records, correcting attribute consistency problems (by means of fixing/removing affected records), in order to guarantee the **business rules** presented earlier in the Data Warehouse design session but also attached as an appendix to this document (see Appendix A).
  - Indicate clearly where in the ETL process each rule is checked/acted upon (e.g., put a comment with the ID of the rule or name the task using the ID of the rule)
  - In the case you propose removing the affected records from the data flow, be sure you make them available for further offline analysis of the possible errors.
    - Otherwise, in the case you propose fixing the affected values, elaborate the decision and the assumptions taken (e.g., briefly refer to it in the report).
- Derive additional attributes, by means of, but not limited to value conversion and formula calculation, in order to enable the calculation of the requested KPIs (see the lab on Data Warehousing design).
  - For example, to calculate Flight Hours (FH) you should subtract actualDeparture from actualArrival times, and for Flight cycles (TO) you should count only the non-cancelled flights in table Flights.

Loading

- Load dimension tables of your star schemas, paying special attention to enable navigation through different aggregation levels (i.e., roll-up and drill-down operations).
  - For example, aircraft dimension table with information about the corresponding aircraft model.
- Load fact tables of your star schemas, enabling the calculation of all the metrics needed to retrieve the required KPIs.

Deliverables

1. Talend transformation(s) and job(s) inside a single zip file. Please remember to empty the recycle bin in Talend before exporting the project.
2. PDF file (one single A4 page, 2.5cm margins, font size 11, inline space 1.15) with all assumptions made and justifying the decisions you made (if any).

Assessment criteria

i) Conciseness of explanations (only first page will be considered in the evaluation)

ii) Understandability

iii) Coherence

iv) Soundness

Evaluation

- 60% Deliverables
- 40% Exercises related to the project done individually in the classroom the day of the partial exam

A Appendix

Business Rules

In the following you can find the business rules that are supposed to be true in the data. Nevertheless, neither the processes nor the DBMS enforced them. Thus, they may have been violated giving rise to quality problems. In the ETL process, you are expected to enforce them, that is, check if they are violated and act upon them.

`AMOS` database

Identifiers

BR-1 `WorkPackageID` is an identifier of `WorkPackage`.

BR-2 `workOrderID` is an identifier of `WorkOrders/ForecastedOrders/TechnicalLogBookOrders`.

BR-3 `maintenanceID` is an identifier of `MaintenanceEvents/OperationInterruption`.

Datatypes/Domains

BR-4 `subsystem` of `MaintenanceEvents` should be a 4 digits ATA code[1]

BR-5 `delayCode` in `OperationInterruption` should be a 2 digits IATA code[2]

BR-6 `WorkPackageID/workOrderID/maintenanceID` should be simply SERIAL numbers generated by an autoincrement[3] mechanism.

BR-7 `ReportKind` values "PIREP" and "MAREP" refer to pilot and maintenance personnel as reporters, respectively.

BR-8 `MELCathegory` values `A,B,C,D` refer to `3,10,30,120` days of allowed delay in the repairing of the problem in the aircraft, respectively.

BR-9 `airport` in `MaintenanceEvents` must have a value.

Other business rules

BR-10 In `OperationInterruption`, departure must coincide with the date of the `FlightID` (see below how it is composed).

BR-11 The `Flight` registered in `OperationInterruption`, must exist in the `Flights` of `AIMS` database, and be marked as "delayed" (i.e., `delayCode` is not null) with the same IATA delay code.

BR-12 In `MaintenanceEvents`, maintenance duration must have the expected length according to the kind of maintenance (Delay – minutes, Safety – undetermined/unlimited, `AircraftOnGround` - hours, `Maintenance` – hours to max 1 day, `Revision` – days to 1 month).

`AIMS` database

Identifiers

BR-13 `FlightID` is an identifier of `Flights`.

Datatypes/Domains

BR-14 `FlightID` is derived by concatenating the following values:
`Date-Origin-Destination-FlightNumber-AircraftRegistration`
(lengths: 6+1+3+1+3+1+4+1+6=26).

Other business rules

BR-15 In a `Slot`, `scheduledArrival` must be posterior to the `scheduledDeparture`.

BR-16 A `Flight` is not longer than 24 hours.

[1] ATA codes for commercial aircrafts: https://en.wikipedia.org/wiki/ATA_100

[2] IATA delay codes: https://en.wikipedia.org/wiki/IATA_delay_codes

[3] https://www.postgresql.org/docs/9.1/datatype-numeric.htmlDATATYPE-NUMERIC-TABLE

BR-17 All the hours of a `Flight` are imputed to the date of its `scheduledDeparture`.

BR-18 In `Flights`, departure and arrival airports must be those in the `FlightID` (unless this `Flight` has been diverted).

BR-19 In a `Flight`, `actualArrival` is posterior to `actualDeparture`.