Basi di dati II – Homework su OLAP e ETL – 21 marzo 2016

Si richiede di costruire e alimentare un *data mart* relativo alle carriere degli studenti del nostro Ateneo. In particolare, il progetto consiste nell'implementazione di alcuni schemi dimensionali discussi in aula e nella realizzazione dei flussi di ETL per alimentarli a partire dai dati di input che vengono forniti.

parte 1 – OLAP Si richiede la realizzazione di una versione semplificata di alcuni schemi dimensionali discussi in aula. In particolare sono di interesse INGRESSI, ESAMI e LAUREE, i cui dettagli sono riportati nel documento contenente la progettazione dimensionale, illustrato in aula e disponibile sul sito del corso (link, lucidi 29, 36, 40). Per ogni schema dimensionale, sono di interesse solamente le dimensioni indicate nel seguente elenco:

INGRESSI : Studente, Anno Accademico, Corso di Studi, Tipo di Immatricolazione, Tempo

ESAMI : Studente, Anno Accademico, Corso di Studi, Corso, Tempo

LAUREE : Studente, Anno Accademico, Corso di Studi, Tempo

Le specifiche di dettaglio della realtà di interesse sono state discusse in aula e sono riportate nel documento pure disponibile sul sito del corso (link). È inoltre disponibile la *matrice DWH bus* (link), che evidenzia, per ogni schema dimensionale, quali sono le misure e quali le dimensioni. Notare che la matrice copre l'intero data mart, ma gli schemi dimensionali da realizzare sono esclusivamente quelli evidenziati in rosso, con le sole dimensioni di interesse riportate (in sostanza non si considera la dimensione Docente).

Indicazioni operative

- Creare tutti gli schemi relazionali a supporto di quelli dimensionali in un unico schema *Postgres* (nome del database DWH; nome dello schema AVA).
- Rispettare le convenzioni illustrate a lezione e disponibili nel documento sugli strumenti OLAP (link, lucido 18).
- Configurare opportunamente i metadati in *Pentaho Business Analytics*, realizzando un data source per ogni schema dimensionale richiesto (quindi tre data source in totale).

parte 2 – ETL Occorre realizzare i flussi di ETL necessari per alimentare il data mart costruito nella parte 1. Come input, verrà fornito in un archivio zip (link, disponibile da martedì 22), con un insieme di file di dati, estratti dai sistemi dell'Ateneo e anonimizzati, contenenti gli eventi accaduti nella carriera di ciascuno studente dal 2012 al 2014. L'archivio contiene una serie di *file evento* con nome `WS_ZIP_A7_YYYY_N_....out.TXT`, dove `YYYY` rappresenta l'anno di riferimento e `N` il tipo di evento che il file tratta. Nella directory `CODICI GENERALI` sono invece presenti dei file testuali ed Excel, con le informazioni necessarie a decodificare i codici utilizzati nei file evento. Il file `ans_guida.pdf` (contenuto nell'archivio) è una guida dettagliata del tracciato record dei file evento. Infine, la matrice *matrice DWH bus* riporta per ogni misura e dimensione, un'indicazione di quali file evento è necessario considerare, con i dettagli relativi ai record di interesse e agli eventuali file di decodifica necessari.

Indicazioni operative

- Creare in *Pentaho Data Integration* una trasformazione `<nome dimensione>.ktr` per ogni dimensione.
- Creare un job `dimensioni.kjb` che alimenta tutte le dimensioni.
- Creare una trasformazione `<nome fact table>.ktr` per ogni fact table.
- Creare un job `fact_tables.kjb` che alimenta tutte le fact table.

- Dopo aver eseguito tutti i job creati, verificare il corretto caricamento dei dati eseguendo alcune interrogazioni o prove di navigazione in Pentaho Business Analytics.

Modalità di lavoro e di consegna Il lavoro può essere svolto individualmente o in gruppi di due o tre persone. Gli esperimenti devono però essere, almeno in parte, individuali. La consegna è individuale.
Ogni gruppo crea un repository *GitHub* con url `https://github.com/<nome gruppo>/<nome repo>` (ad esempio `https://github.com/myGroup/myRepo`). Il repository dovrà avere il seguente contenuto:

- directory **OLAP**
  - `schema_relazionale.sql`
  - `Ingressi.xmi`
  - `Ingressi.xml`
  - `Esami.xmi`
  - `Esami.xml`
  - `Lauree.xmi`
  - `Lauree.xml`
- directory **ETL**
  - `dimensioni.kjb`
  - `fact_tables.kjb`
  - `<nome fact table>.ktr` (uno per ogni fact table)
  - `<nome dimensione>.ktr` (uno per ogni dimensione)
- `README.md`

Il file `schema_relazionale.sql` è l'export dei DDL del solo schema del database Postgres creato, come descritto nel documento sulla modellazione dimensionale (link, lucido 6). I file con estensione `xmi` e `xml` sono ottenuti effettuando l'export di ciascun data source da Business Analytics, come descritto nello stesso documento (link, lucido 7). I file con estensione `kjb` e `ktr` sono ottenuti salvando rispettivamente i vari job e le varie trasformazioni da Pentaho Data Integration. Il file `README.md` contiene cognome e nome di tutti i membri del gruppo (senza virgole, su righe diverse e in ordine alfabetico).
Al momento della consegna ciascuno studente (N.B. la consegna è individuale), indica su Moodle il repository dove il proprio gruppo ha rilasciato il progetto:

```
Consegna homework:  <URL del repository del proprio gruppo>
```

ad esempio:

```
Consegna homework:  https://github.com/rossi/AVAProject
```