Esercitazione ad integrazione
Corso di modelli dei dati di nuova generazione e Labo. basi di dati II
Anno Accademico 2004/2005

Analisi dei crimini negli stati uniti per età, sesso e razza nel 1997

Descrizione della sorgente dei dati

Il *National Archive of Criminal Justice Data* (`http://www.icpsr.umich.edu/NACJD`) degli stati uniti detiene e distribuisce dati sulla giustizia e crimini verificati negli stati uniti in determinati periodi di tempo. Questi dati sono forniti in file testuali che rispettano una grammatica ben definita (code book).

Tra i vari dataset, in questa esercitazione siamo interessati a quelli relativi a:
"*Uniform Crime Reporting Program [United States]: Arrests by Age, Sex and Race for Police Agencies in Metropolitan Statistical Areas, 1960-1997*"

Questo dataset fornisce informazioni sul numero di arresti effettuati dal FBI ogni anno da agenzie locate in aree metropolitane. I dataset sono stati realizzati sulla base di modulistica in cui vengono riportati gli arresti effettuati sulla base del sesso, età e razza dei criminali e il tipo di violazione compiuta (sono considerate 43 violazioni). Per questa collezione, gli arresti riportai da ogni agenzia sono aggregati in base all'anno, per ognuno degli anni dal 1960 al 1997, e la modulistica originaria è stata ristrutturata per creare 2 documenti separati per ogni anno: header file e detail file. I record negli header files sono legati ai record dei detail files attraverso l'identificatore dell'agenzia che ha riportato i dati (*originating agency identifier* - ORI). Altre variabili comuni ai due files sono: stato, census group, anno, divisione, e agenzia metropolitana statistica (*metropolitan statistical agency* - MSA). Il file header contiene anche il nome dell'agenzia e la popolazione di cui l'agenzia è responsabile. Il file details contiene anche il codice della violazione e l'età, il sesso, e la razza dell'arrestato.

Di questi dati siamo interessati al header e detail file corrispondenti all'anno 1997.

La grammatica che seguono i dati in questi due file viene descritta nel documento allegato (**Allegato A**).

Obiettivo dell'esercitazione

Lo scopo dell'esercitazione è la creazione di un data warehouse per questo dominio applicativo in Oracle. Ogni gruppo dovrà:

1. Effettuare uno studio concettuale del dominio applicativo utilizzando l'allegato A. Lo scopo di questo studio è la realizzazione di uno schema concettuale e logico del dominio.
2. Sviluppare un foglio di mappatura con ORACLE LOADER per caricare tali dati in una base di dati relazionale il cui schema relazionale è stato sviluppato al passo 1 (viene fornita la documentazione necessaria per effettuare questa operazione).
3. Caricare i dati in tale schema.
4. Sviluppare i modelli concettuali dei data marts necessari a rispondere alle seguente analisi (verificare sulla base di dati la sensatezza di queste interrogazioni)
    a. Analizzare la quantità di crimini minorili nel New England che riguardano abusi per vendita o possesso di droga.
    b. Confrontare i crimini per scommesse (gambling) sulla base dell'età e del sesso di giovani tra i 18 e 25 anni negli stati del sud de gli stati uniti.

I data marts sviluppati devono consentire di effettuare roll-up e drill down lungo le varie dimensioni. Le dimensioni devono essere costruite considerando le informazioni contenute nell'*Allegato A*. Inoltre, i data marts devono essere generali in modo da poter rispondere sia alle interrogazioni proposte che ad interrogazioni analoghe (senza compromettere l'efficienza del sistema).

5. Sviluppare i modelli logici dei data marts e popolarli utilizzando la base di dati operazionale (quella realizzata dai punti 1 al 3).
6. Sviluppare le interrogazioni SQL per rispondere alle interrogazioni del punto 4.
7. Ogni gruppo dovrà poi affrontare uno dei seguenti problemi (questi problemi possono essere l'oggetto del seminario da sviluppare):
    a. Uso di strutture ausiliarie di accesso per migliorare i tempi di risposta del sistema.
    b. Uso di partizionamento orizzontale o verticale dei dati per migliorare le prestazioni del sistema.
    c. Uso di viste materializzate per migliorare le prestazioni del sistema.

Documenti da consegnare

Per la valutazione del progetto è necessario consegnare un documento contenente:

1. Modello ER che descrive il contenuto dell'informazione della sorgente e schema relazionale sviluppato.
2. Script per il caricamento dei dati nella base di dati operazionale.
3. Modello concettuale dei fatti, dimensioni e misure sviluppati a partire dal modello ER del punto 1 al fine di rispondere alle interrogazioni richieste al punto 4.
4. Modello logico dei fatti, dimensioni e misure con gli script di creazione degli oggetti dimensione.
5. Discussione sulla possibilità di utilizzare viste materializzate per migliorare le prestazioni del sistema (se utilizzate, riportare il codice di creazione delle viste).
6. Script/interrogazioni sviluppate per popolare le tabelle dei fatti e delle dimensioni.
7. Codice delle interrogazioni SQL sviluppate.

Consegna

Il progetto deve essere terminato entro il 30 Settembre 2005. Dopo tale data occorrerà concordare con il docente una nuova prova di esame.

Per monitorare l'andamento del lavoro dei vari gruppi, in modo da evitare dispersioni, verranno fissate 2 consegne intermedie. Le scadenze delle consegne intermedie sono fissate dai gruppi. Il docente ha quindi la possibilità di calibrare il testo dell'esercitazione in base ai tempi "effettivi" richiesti per svolgere e produrre i documenti di ogni singola fase. Alla prima consegna intermedia, il gruppo deve anche consegnare un documento che indica per ognuno dei documenti da consegnare il membro del gruppo responsabile per quel documento e chi ha collaborato principalmente nella sua realizzazione.

	Prima consegna intermedia	**Seconda consegna intermedia**
Documenti e script da consegnare	Punti 1,3	Punti 4,7

La valutazione verrà effettuata sul documento finale. I suggerimenti del docente sulle consegne intermedie, non verranno considerati nella valutazione finale.