Progetto per il corso di Basi di Dati

Istruzioni Generali

Il progetto deve essere svolto solo da gruppi di 2 persone.

> Gruppi di 1 persona possono essere ammessi in casi di gravi problemi. È necessario inviare una e-mail al docente spiegando i motivi che non consentono di formare un gruppo di 2 persone.

Data di Consegna: Giovedì 7 Gennaio 2016 ore 07:59 (GMT +01:00)

[Entro l'8 Dicembre] Creare la cartella `query-assignment3` all'interno della cartella condivisa precedentemente su Google Drive con account di Ateneo (leggi sotto!).

Il progetto prevede lo studio, la realizzazione e l'ottimizzazione di un database, la conversione e il caricamento di dati nel database progettato e l'analisi dei dati attraverso la realizzazione di query.

Il progetto è quindi diviso in 3 sezioni:

1. Creazione del database
2. Inserimento dei dati (e loro eventuale normalizzazione e conversione di formato)
3. Progettazione delle Query SQL per l'analisi dei dati

I criteri di valutazione terranno conto di parametri quali:

1. La corretta importazione dei dati, tale da non causare una perdita di informazione o un'inserimento erroneo.
2. Il tempo necessario al caricamento e alla formattazione dei dati
3. Lo spazio occupato su disco dal database, considerando sia i dati che le strutture ausiliare (e.s. indici)
4. La correttezza delle risposte delle query richieste
5. La velocità di esecuzione delle query

Tutti i dati iniziali devono essere presenti nel database una volta importati (seppure in formato o con schema diverso).

> **Esempio:** se nel file dei dati è presente un certo numero di aziende distinte con un certo numero di valutazioni, lo stesso numero di aziende e valutazioni deve essere presente nel database importato (come risultato di una opportuna query o vista).

File da Consegnare & Metodo di Consegna

Il progetto finale deve essere consegnato in un file compresso `.zip` con nome `db2016_matricola1_matricola2.zip`, dove `matricola1` e `matricola2` sono i numeri di matricola dei due membri del gruppo.

> **Esempio:** `db2016_167890_198765.zip`

L'archivio deve contenere solamente

1. Un file `creazione.sql` che assume l'esistenza di un database con nome `db2016` ed esegue la creazione degli **schemi**, delle **tabelle**, delle **viste** e delle **stored procedures** necessarie.
2. Un file `importazione.sql` che assume l'esistenza dei file `.csv` nella cartella `/tmp/dati` ed esegue tutte le operazioni necessarie per importare i dati dai file alle tabelle del punto precedente. Inoltre contiene ogni necessaria operazione di pulizia e formattazione delle tabelle e dei dati.
3. Un file `query_XX.sql` per ogni query assegnata (dove `XX` è il numero della query)
4. Un file `members.txt` contenente matricola, nome, cognome e email di entrambi i membri del gruppo

I file `query_XX.sql` al punto **(3) non possono contenere** creazioni di viste, tabelle temporanee o procedure. Possono però fare uso della clausola `WITH` [*].

> NB. L'archivio contenente i file del progetto consegnato non deve contenere nessun altro file, nemmeno i file di dati!
>
> I dati utilizzati durante la valutazione saranno diversi da quelli forniti e verranno creati separatamente. La cartella `/tmp/dati` verrà rimossa una volta la fase di importazione verrà terminata [*].

Il progetto deve essere inserito compresso nella cartella condivisa, utilizzata su Google Drive con account di Ateneo per i precedenti assignment, e non deve essere modificato dopo la data di consegna (altrimenti verrà considerato consegnato fuori data).

Il progetto deve essere condiviso con soli permessi di lettura.

Correzione e Valutazione

I file SQL verrano eseguiti su due macchine ed i tempi di esecuzione misurati su entrambe.

> NB. I file verrano eseguiti attraverso il comando `psql < nome_file.sql`, oppure all'interno di postgres con il comando `\i nome_file.sql`

Clotho

```
PostgreSQL 9.3.6
compiled by gcc (Ubuntu 4.8.2-19ubuntu1) 4.8.2, 64-bit
on Intel(R) Pentium(R) D CPU 3.40GHz — No of Cores:2
with 8GB of RAM
```

Snowwhite

```
PostgreSQL 9.3.6
compiled by gcc (Ubuntu 4.8.2-19ubuntu1) 4.8.2, 64-bit
on Intel(R) Xeon(R) CPU E5-2440 0 @ 2.40GHz — No of Cores:24
with 198GB of RAM
```

> NB. È responsabilità degli studenti assicurarsi che i file possano essere eseguiti correttamente sulle macchine qui riportate, verificando con attenzione le versioni del sistema operativo e di Potsgres. **Nessun programma esterno, libreria aggiuntiva o non installata di default può essere usata.**

Note sul Voto

1. Ogni progetto che non rispetta le modalità di consegna e presenta data di ultima modifica successiva alla scadenza non verrà valutato (0 punti);
2. Ogni progetto che contiene file diversi da quelli richiesti non verrà valutato (0 punti);
3. Ogni progetto in cui almeno 1 file genera errori verrà valutato interamente 0 punti;
4. Ogni progetto in cui almeno 1 query non ritorna il risultato esatto o nell'esatto formato verrà valutato interamente 0 punti;
5. Ogni progetto che rispetta tutte le condizioni di consegna in grado di eseguire completamente senza errori e in cui ogni query ritorna il risultato esatto verrà valutato `>= 18/30`.
6. Nella valutazione finale del progetto lo spazio occupato su disco avrà peso per il `30%` del voto mentre la velocità di esecuzione dei file `.sql` influirà per il restante `70%`.

Contesto & Dati

Si assuma di dover rispondere alle esigenze di un'agenzia di trading che vuole condurre un'analisi sulle azioni in borsa delle maggiori compagnie. A questo scopo ha acquistato dei dati da un rivenditore. In questo modo ha ottenuto un set di dati che contiene informazioni sugli elenchi delle azioni in vari mercati degli Stati Uniti. Si tratta di dati storici giornalieri negli anni in cui un'azienda è stata quotata, e include solo le informazioni di base sul prezzo ed il volume (quantità) delle azioni. Alcuni dati possono essere mancanti per errori nei sistemi di rilevazione. I mercati di interesse sono due: lo NYSE e il NASDAQ.

Assieme ai dati delle azioni sono state raccolte informazioni generali di contatto, come ad esempio gli indirizzi e-mail dell'amministratore delegato, per un set di imprese importanti, anche se non quotate sui mercati. In particolare, i dati raccolti sono stati suddivisi nei seguenti file:

1. `companies.csv` : ogni linea contiene i dettagli di un'azienda e le sue informazioni di contatto, se disponibili.
    - `name` nome esteso dell'azienda
    - `address` indirizzo/via della sede principale
    - `city` città della sede principale
    - `state` stato (USA) della sede principale
    - `zipcode` codice postale della sede principale
    - `phone` numero di telefono dell'ufficio di riferimento
    - `website` sito web
    - `general_email` e-mail di contatto dell'ufficio di riferimento
    - `ceo_name` nome e cognome dell'amministratore delegato
    - `ceo_email` e-mail dell'amministratore delegato
2. `stock_symbols.csv` : ogni linea contiene l'informazione identificativa delle azioni di un'azienda quotata in borsa
    - `market` il mercato in cui l'azione viene venduta `[NYSE| NASDAQ]`
    - `company_name` il nome esteso con cui l'azienda è stata quotata in borsa (**non è garantito sia uguale a** `name` in `companies.csv`)
    - `symbol` il codice con cui l'azione dell'azienda viene identificata sul mercato
3. `stock_price.csv` : ogni linea contiene le informazioni di valore di un'azione per un singolo giorno
    - `symbol` il codice con cui l'azione dell'azienda viene identificata sul mercato (`symbol` in `stock_symbols.csv`)
    - `date` data della valutazione
    - `price` prezzo ufficiale dell'azione per la giornata
    - `volume` numero di azioni sul mercato
    - `open` valore dell'azione all'apertura del mercato
    - `low` valore minimo raggiunto dall'azione nella giornata
    - `high` valore massimo raggiunto dall'azione nella giornata

File Speciali

I seguenti file contengono informazioni speciali relative a specifiche richieste dell'agenzia di trading

1. `market.csv` : il file contiene un'unica linea relativa al mercato di interesse
    - `market` il mercato di interesse `[NYSE| NASDAQ]`
2. `states.csv` : il file contiene una lista di stati di interesse (`state` in `companies.csv`)
    - `state` stato (USA)

3. `symbols.csv` : il file contiene una lista di codici di azioni di interesse (`symbol` in `stock_symbols.csv`)
    - `symbol` il codice con cui l'azione dell'azienda viene identificata sul mercato
4. `date.csv` : il file contiene una lista di date di interesse e il loro tipo
    - `date` data in formato `YYYY-MM-DD`
    - `type` tipo in `[ST|MN|MX|ED]`

Il significato dei codici `[ST|MN|MX|ED]` è una informazione non disponibile e protetta da `segreto industriale` .

Snapshot dei Dati

Per svolgere il progetto e le query, viene fornito un campione di circa `0.01%` dei dati totali disponibili per quanto riguarda il file `stock_price` . Mentre degli altri due file, `companies.csv` e `` `, vengono forniti il circa il `` 50%` dei dati.

Il campione di dati è scaricabile al link: http://disi.unitn.it/~lissandrini/files/sample.tar.gz

Query da Realizzare

[Entro l'8 Dicembre] Creare la cartella `query-assignment3` all'interno della cartella condivisa precedentemente su Google Drive con account di Ateneo. E garantire su quella cartella permessi di scrittura agli account

`d.papadimitriou@unitn.it` , `matteo.lissandrini@unitn.it` e `velgias@unitn.it`

NON inserire alcun file in questa directory

Fatta eccezione per la `query_00` tutte le altre query devono contenere un unico statement `SELECT` , `DELETE` o `UPDATE` (a seconda della richiesta).

Fatta eccezione per la `query_00` , a meno che non sia richiesto esplicitamente, tutte le altre query devono stampare i risultati a schermo e non salvare su file

- [query_00] Ricostruire i file `.csv` utilizzati in input ordinandoli su `date` , `market` e `name` (a seconda dei campi contenuti) e salvati nella cartella `/tmp/output` .

> Il resto delle query verrà caricato su Google Drive nella cartella `query-assignment3`