Università degli Studi di Modena e Reggio Emilia

Facoltà di Ingegneria – Sede di Modena
Corso di Laurea Specialistica in Ingegneria Informatica

BART: Uno strumento di analisi di business e reportistica

Relatore:
Prof. Sonia Bergamaschi

Candidato:
Mattia Bonacorsi

Anno Accademico 2005-2006

Introduzione

Def: Un *cubo* è una struttura che contiene dati multidimensionali

All'interno della tesi si sono approfondite tematiche che riguardano:

- definizione e creazione di un cubo
- tecniche di alimentazione di un data warehouse
- tecniche per migliorare le prestazioni del data warehouse
- interrogazione ed estrazioni di dati da un cubo
- progetto e implementazione di un prototipo che permetta la definizione di cubi e la loro interrogazione

Strumenti utilizzati

Gli strumenti utilizzati per la realizzazione del prototipo sono:

- Mondrian
  - OLAP Server
  - Permette l'interrogazione di sorgenti dati relazionali
  - Gestione del caching dei dati
- JPivot
  - Interfaccia Web
  - Visualizzazione dei risultati di interrogazioni multidimensionali
  - Navigazione dei risultati con operazioni di drill-down, roll-up, slice o drill-position
- Struts
  - Framework per lo sviluppo di applicazioni Web basate su Java

Drill-down, Roll-up

Tempo inizio		Measures	
(All)	ANNO_INIZIO	DURATA_MEDIA (gg)	DURATA_MAX (gg)
−Tutte		157.08	860
Tutte	+2001	175.09	693
	+2002	167.76	739
	+2003	152.84	860
	+2004	155.39	540
	+2005	158.07	379
	+2006	87.73	157

Drill-down

Tempo inizio			Measures	
(All)	ANNO_INIZIO	TRIMESTRE_INIZIO	DURATA_MEDIA (gg)	DURATA_MAX (gg)
−Tutte			157.08	860
Tutte	+2001		175.09	693
	+2002		167.76	739
	+2003		152.84	860
	+2004		155.39	540
	−2005		158.07	379
	2005	+T1	160.91	379
		+T2	164.64	339
		+T3	153.85	224
		+T4	132.27	157
	+2006		87.73	157

Roll-up

Tempo inizio			Measures	
(All)	ANNO_INIZIO	TRIMESTRE_INIZIO	DURATA_MEDIA (gg)	DURATA_MAX (gg)
−Tutte			157.08	860
Tutte	+2001		175.09	693
	+2002		167.76	739
	+2003		152.84	860
	+2004		155.39	540
	−2005		158.07	379
	2005	+T1	160.91	379
		+T2	164.64	339
		+T3	153.85	224
		+T4	132.27	157
	+2006		87.73	157

Tempo inizio		Measures	
(All)	ANNO_INIZIO	DURATA_MEDIA (gg)	DURATA_MAX (gg)
−Tutte		157.08	860
Tutte	+2001	175.09	693
	+2002	167.76	739
	+2003	152.84	860
	+2004	155.39	540
	+2005	158.07	379
	+2006	87.73	157

Slice, Drill-position

Tempo inizio			Measures	
(All)	ANNO_INIZIO	TRIMESTRE_INIZIO	DURATA_MEDIA (gg)	DURATA_MAX (gg)
−Tutte			157.08	860
Tutte	+2001		175.09	693
	+2002		167.76	739
	+2003		152.84	860
	−2004		155.39	540
	2004	+T1	143.02	430
		+T2	147.93	463
		+T3	157.00	540
		+T4	179.35	418
	−2005		158.07	379
	2005	+T1	160.91	379
		+T2	164.64	339
		+T3	153.85	224
		+T4	132.27	157
	+2006		87.73	157

Slice

Tempo inizio			Measures	
(All)	ANNO_INIZIO	TRIMESTRE_INIZIO	DURATA_MEDIA (gg)	DURATA_MAX (gg)
Tutte	−2004		155.39	540
	2004	+T1	143.02	430
		+T2	147.93	463
		+T3	157.00	540
		+T4	179.35	418

↑Tempo inizio		Measures	
↑(All)	↑ANNO_INIZIO	DURATA_MEDIA (gg)	DURATA_MAX (gg)
Tutte	↓2001	175.09	693
	↓2002	167.76	739
	↓2003	152.84	860
	↓2004	155.39	540
	↓2005	158.07	379
	↓2006	87.73	157

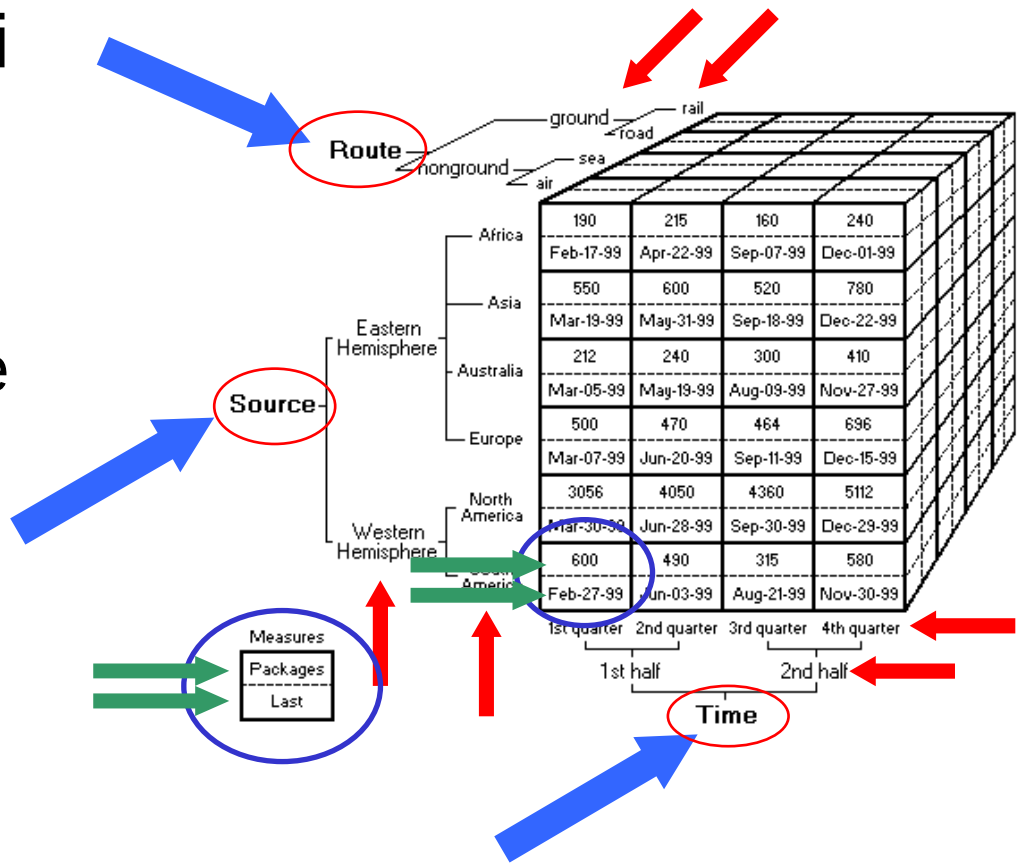
Drill-position

↑Tempo inizio			Measures	
↑(All)	↑ANNO_INIZIO	↑TRIMESTRE_INIZIO	DURATA_MEDIA (gg)	DURATA_MAX (gg)
Tutte	↓2004		155.39	540
	2004	↓T1	143.02	430
		↓T2	147.93	463
		↓T3	157.00	540
		↓T4	179.35	418

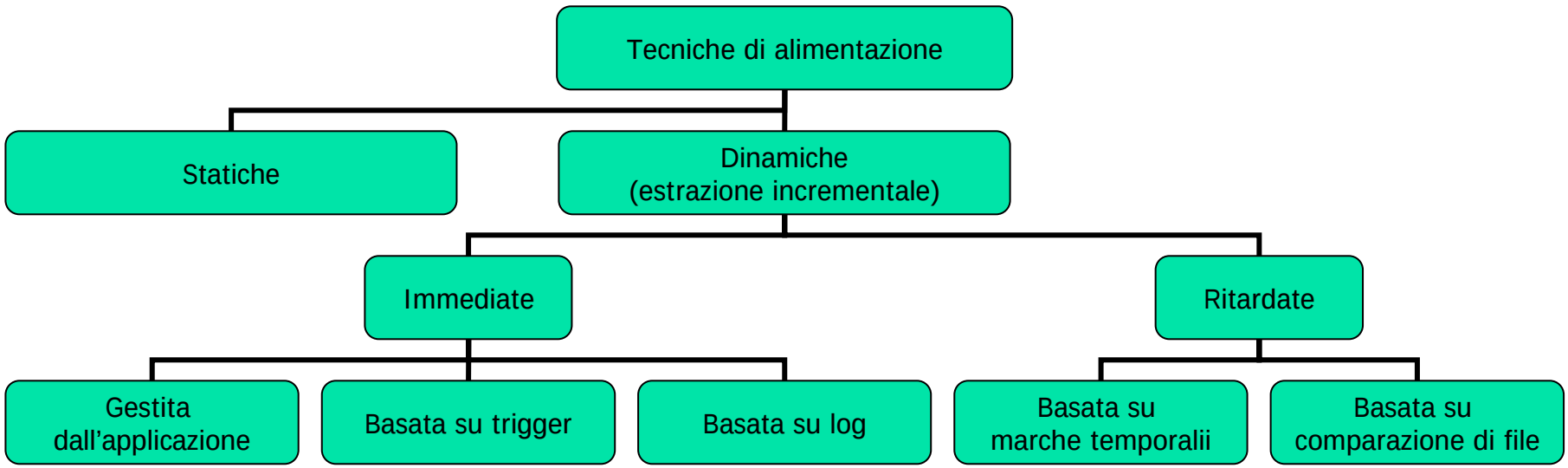
Definizione e creazione di un cubo

La creazione di un cubo comporta la definizione di:

- Dimensioni di analisi
  - Nome
  - Livelli di aggregazione
  - Gerarchie alternative di navigazione
- Misure
  - Nome
  - Tipo di dato
  - Funzione di aggregazione



Tecniche di alimentazione di un cubo



Per la valutazione di quale tecnica utilizzare è necessario tenere presente fattori quali:

- ✓ L'onerosità computazionale richiesta da ogni tecnica a seconda della mole di dati
- ✓ La natura dei dati (transitoria, semi-storicizzata, storicizzata)
- ✓ La dipendenza dal DBMS
- ✓ Impatto prestazionale sul sistema.

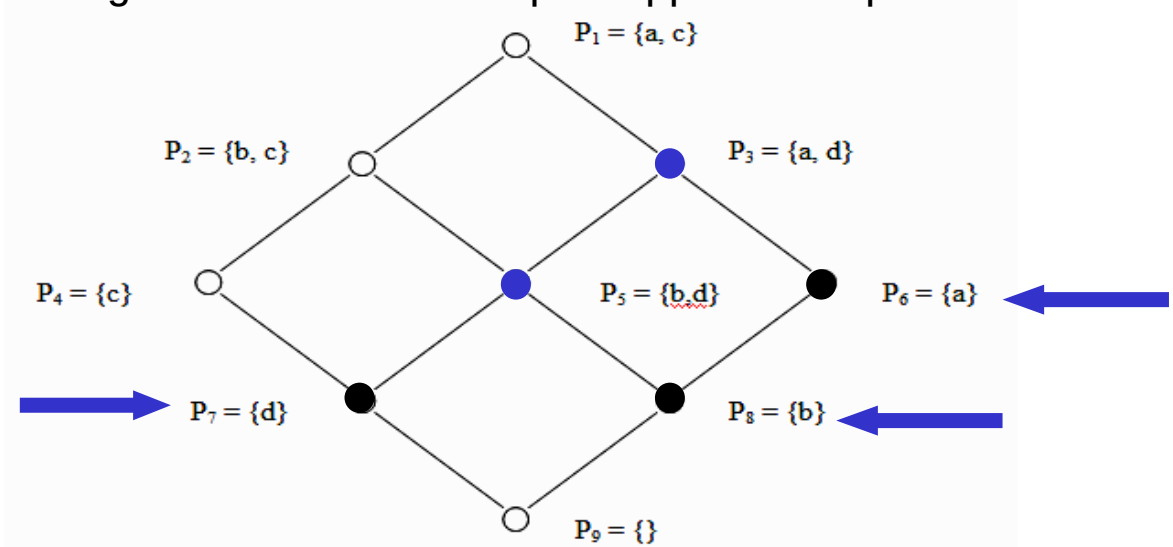
Tecniche per il miglioramento delle prestazioni

Algoritmo di selezione delle viste candidate a essere materializzate

Def: materializzare una vista di un cubo significa creare una nuova tabella all'interno dello schema relazionale che contiene le misure del cubo aggregate a un livello maggiore della fact table

Input:

- Dipendenze funzionali tra gli attributi delle tabelle dimensionali
- Insieme di interrogazioni caratteristiche per l'applicazione presa in esame



Output: Insieme delle viste che se materializzate produrranno un vantaggio prestazionale.

- Possibilità di utilizzare un criterio euristico per ridurre ulteriormente la cardinalità dell'insieme delle viste candidate considerando la cardinalità delle tabelle dimensionali.

Interrogazione di un data warehouse

L'interrogazione di un data warehouse avviene attraverso l'uso del linguaggio MDX

All'interno del prototipo è possibile utilizzare per l'interrogazione con MDX

- Misure calcolate
- Insiemi calcolati
- Filtri sulle dimensioni non coinvolte
- Prodotti cartesiani di insiemi
- Altre funzioni (come quelle di ordinamento, ranking, ...)

Prototipo Realizzato (1/4)

L'applicazione realizzata implementa le seguenti funzionalità:

- Definizione degli schemi che descrivono cubi basati su schemi relazionali
- Definizione di diversi tipi di elementi per la visualizzazione dei risultati di un'interrogazione
- Navigazione dei risultati di un'interrogazione con operazioni di:
  - drill-down
  - roll-up
  - slice
  - drill-position
- Costruzione in modo visuale delle interrogazioni a partire da un'interrogazione predefinita
- Definizione di procedure notturne di alimentazione

Prototipo realizzato (2/4)

Per definire una sorgente dati è necessario indicare:

- I parametri della connessione JDBC alla base di dati
- Un file di catalogo
- Una interrogazione di default

```
<Cube name="AccessLog">
<Table name="BART_ALA_FACT_TABLE"/>
  <Dimension foreignKey="IDDATETIME" name="DateTime">
    <Hierarchy allMemberName="Tutte" hasAll="true"
                                          primaryKey="ID">
      <Table name="BART_ALA_DATE"/>
       <Level column="ANNO" name="ANNO"
                                           uniqueMembers="true"/>
       <Level column="TRIMESTRE" name="TRIMESTRE"
                                          uniqueMembers="false"/>
       <Level column="MESE" name="MESE"
                                          uniqueMembers="false"/>
       <Level column="GIORNO" name="GIORNO"
                                           uniqueMembers="false"/>
    </Hierarchy>
  </Dimension>
<Dimension foreignKey="IDUTENTE" name="User">
    <Hierarchy allMemberName="Tutte" hasAll="true"
                                          primaryKey="ID">
       <Table name="BART_ALA_USER"/>
       <Level column="UTENTE" name="NOME"
                                           uniqueMembers="true"/>
    </Hierarchy>
  </Dimension>
  <Measure aggregator="sum" column="CLICK" name="CLICK"/>
 </Cube>
```

Il file di catalogo:

- È un file XML
- Specifica il nome della tabella dei fatti e delle tabelle dimensionali
- Specifica gli attributi corrispondenti a ciascun livello di ogni dimensione
- Specifica eventuali gerarchie alternative per una dimensione
- Specifica le chiavi primarie ed esterne di ogni tabella coinvolta
- Specifica il nome, il tipo e l'aggregatore da utilizzare per ogni misura contenuta nella tabella dei fatti
- Specifica eventuali viste aggregate da utilizzare per risolvere le interrogazioni

Prototipo realizzato (3/4)

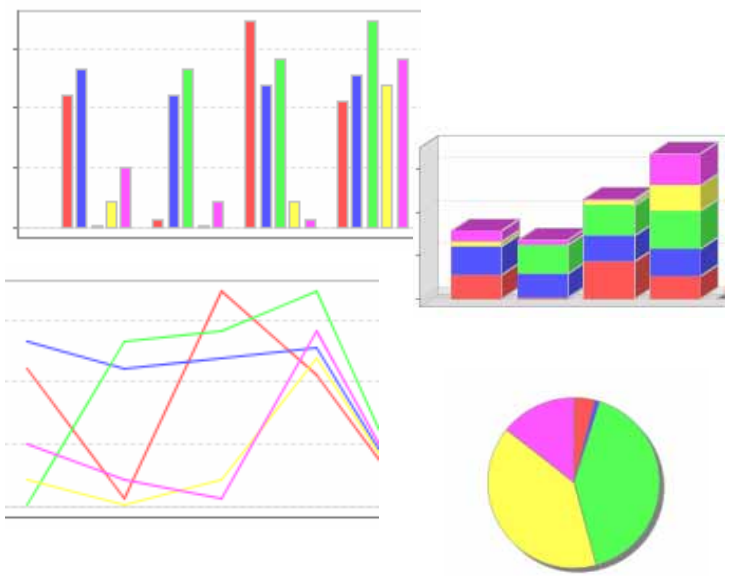
Possibilità di definire diversi tipi di elementi

- Report
- Tabelle
- Grafici
- Pagine

Ogni elemento può essere creato

- Con la corrispondente query MDX
- Per semplici interrogazioni, con un editor visuale

			Tempo	
Tipo	TipoAu		Tutte	
(All)	(All)	TIPOAU	+2005	+2006
+Tutte	Tutte	+Autorizzazione	282	58
		+Diniego		
		+Modifica	24	17
		+Normale		
		+Revoca		1
		+Rinnovo		1
		+Sospensione		



Prototipo realizzato (4/4)

- Primo contesto
  - Analisi dei dati generati da un'applicazione esistente
  - Alimentazione di tipo statico
  - Processi di trasformazione e pulizia dei dati durante l'alimentazione
- Secondo contesto
  - Analisi dei file di log di accesso generati da Tomcat per ricavare le azioni svolte su un'applicazione sviluppata con Struts
  - Alimentazione di tipo incrementale, ritardata
  - Complesso sistema di regole definibili dall'amministratore per capire le azioni dell'utente

Conclusioni (1/2)

- Buone prestazioni anche in presenza di considerevoli moli di dati
  - Diversi milioni di record all'interno della fact table
- Alta configurabilità di Mondrian
  - Sia per l'accesso alle sorgenti dati
  - Sia per i parametri di gestione della cache
- Funzioni di composizione delle query visuali
- È stato evidenziato qualche problema legato alla riscrittura delle query MDX da parte di JPivot
- Mancanza in MDX di funzioni molto utili come now()

Conclusioni (2/2)

Il prototipo:

- È stato realizzato durante un periodo di tirocinio di 6 mesi
- ~230 classi Java
- ~90 pagine JSP

Sviluppi futuri del prototipo

- Creazione e modifica dei cataloghi on-line
- Procedure di salvataggio automatico periodico dei risultati di alcune interrogazioni