OLAPBIRCH

INTEGRAZIONE DI UN ALGORITMO DI CLUSTERING GERARCHICO A SUPPORTO DELLE TECNOLOGIE OLAP

Relatore:
Prof. Michelangelo Ceci

Laureando:
M. Cristina Tarantino

Correlatori:
Prof. Alfredo Cuzzocrea
Prof. Donato Malerba

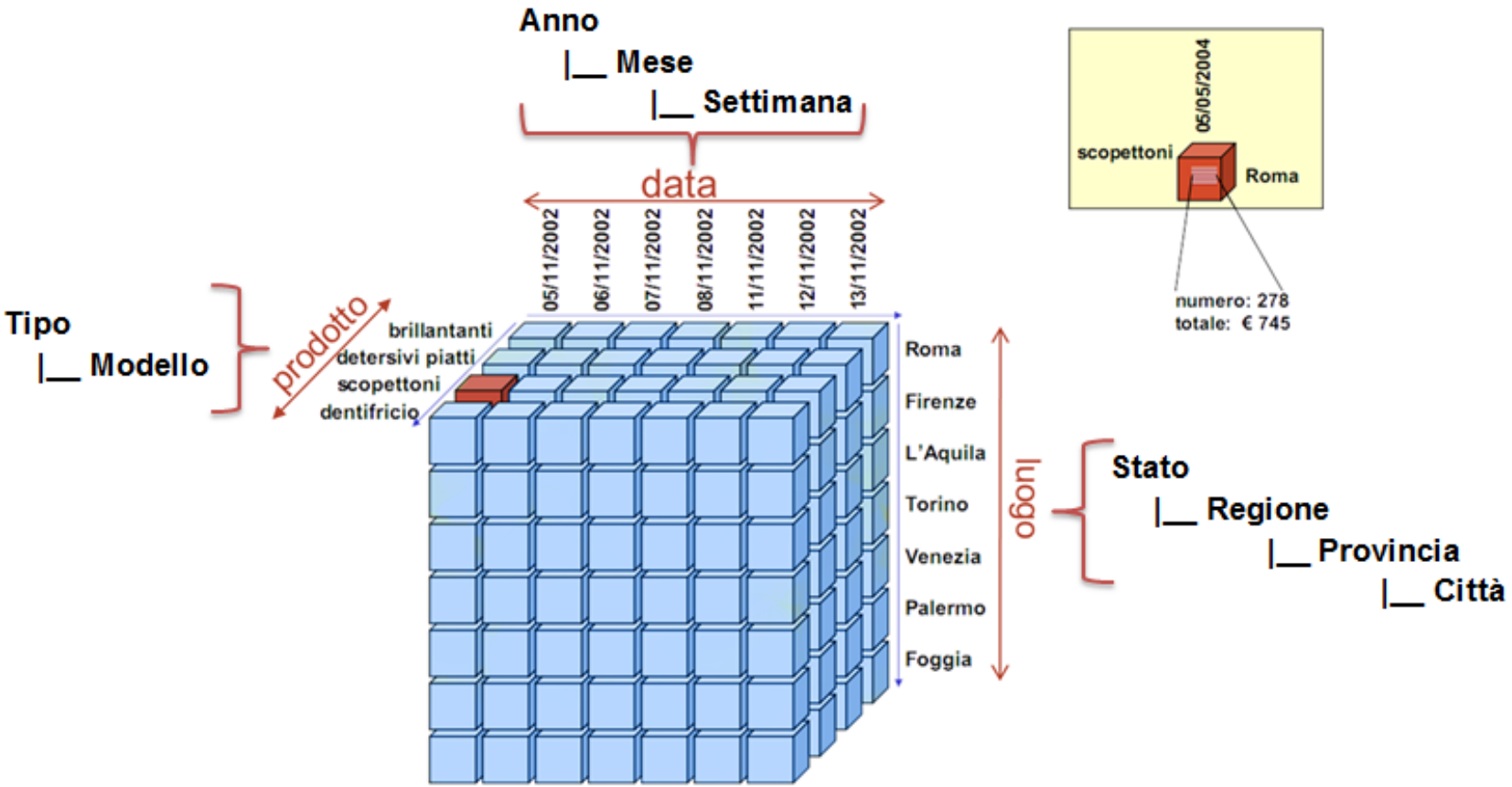
Corso di Laurea in Informatica e Tecnologie per la Produzione del Software
Facoltà di Scienze MM. FF. NN.
Università degli Studi di Bari "Aldo Moro"

24 Febbraio 2011

Data Warehouse

Rappresentazione concettuale

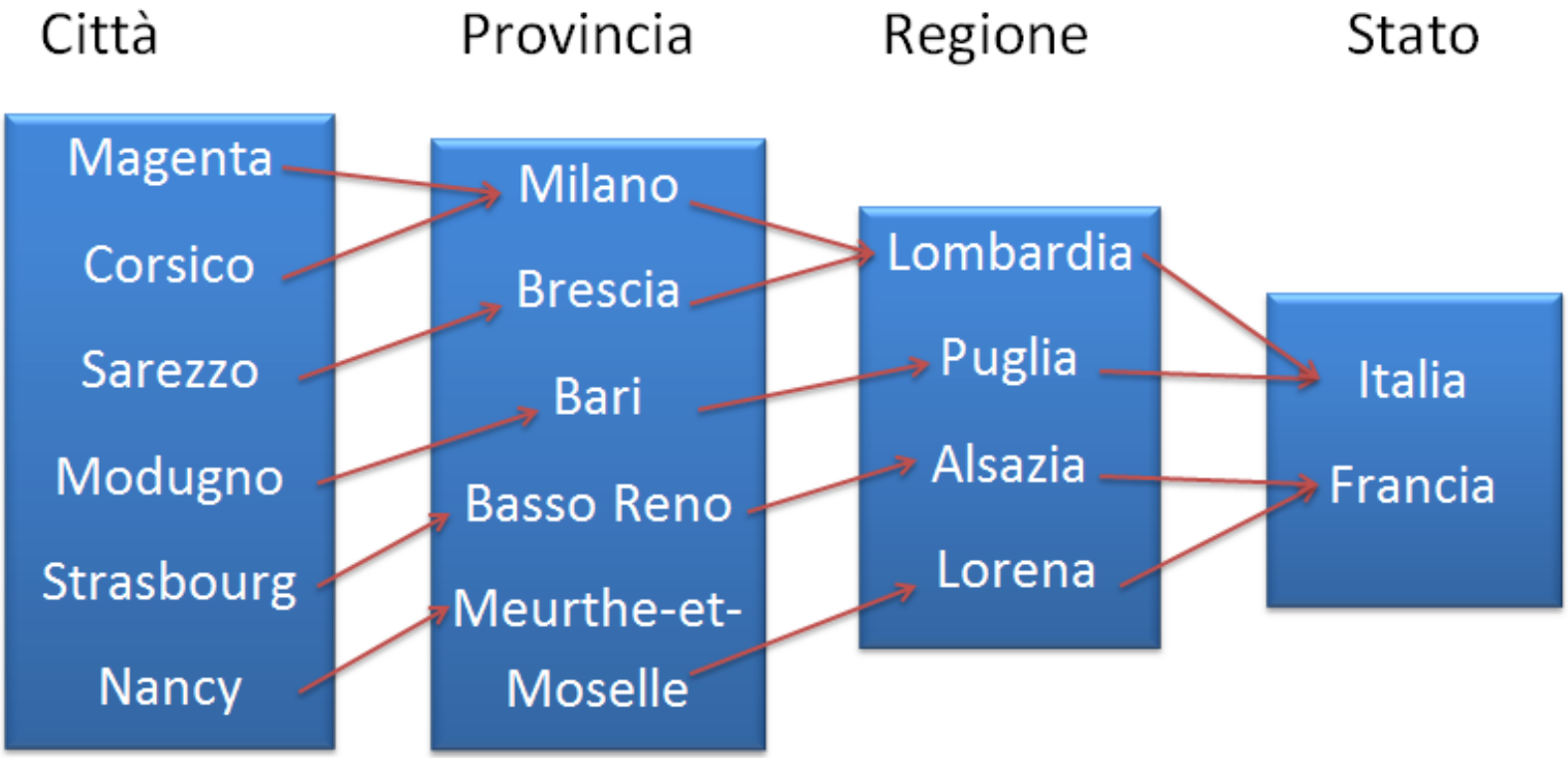
Un processo decisionale è influenzato da un **fatto**, ovvero un insieme di **eventi**. Per poterli agevolmente analizzare si usa una rappresentazione concettuale **multidimensionale** i cui assi, chiamati dimensioni di analisi, definiscono diverse prospettive per la loro identificazione.



Data Warehouse

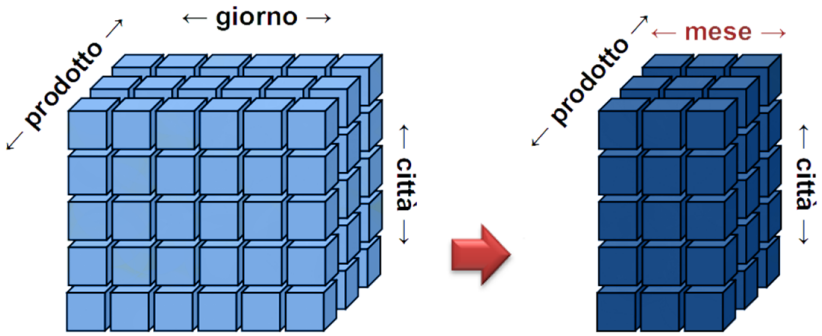
Rappresentazione concettuale

Ciascuna dimensione è associata ad una **gerarchia di attributi dimensionali** di aggregazione che ne raggruppa i valori in diversi modi.

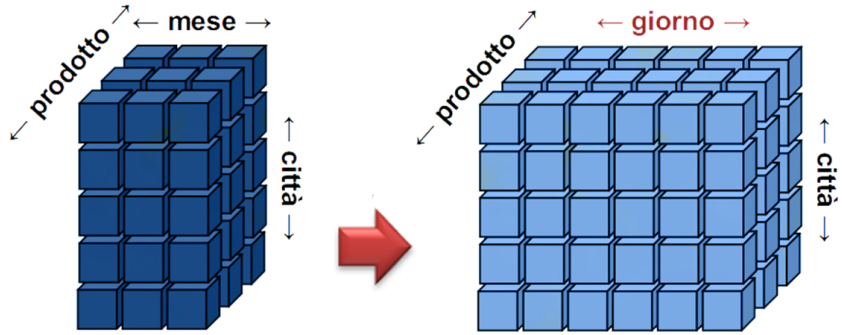


Analisi OLAP

Operatore Roll-Up

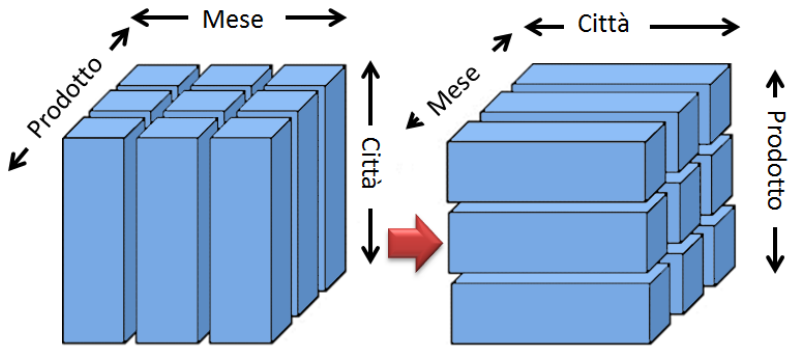


Operatore Drill-Down



Operatore Slice-and-Dice

Operatore Pivoting



Clustering Gerarchico

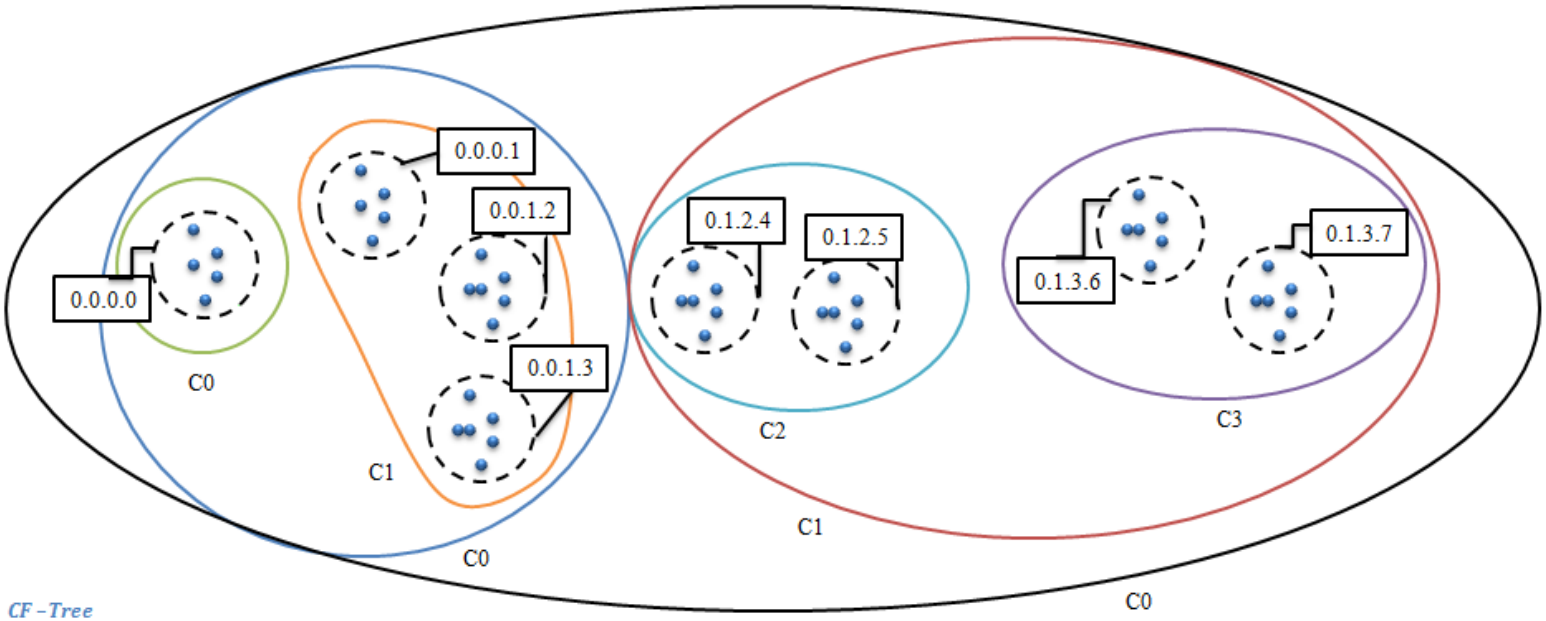
Il clustering gerarchico è un approccio di clustering che mira a costruire una gerarchia di cluster, sulla base di relazioni non note a priori.
A differenza degli operatori OLAP, esso permette di avere gerarchie anche su attributi continui.

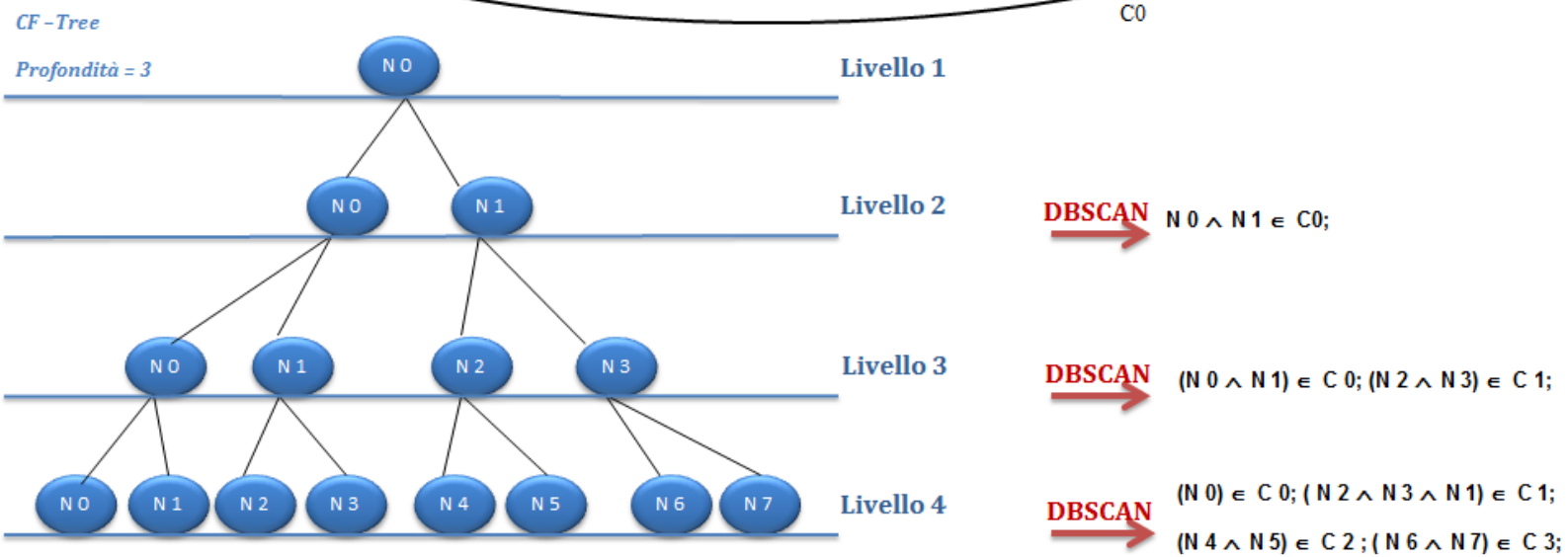
Scopo della Tesi

Permettere l'integrazione e l'interazione di un algoritmo di clustering gerarchico *incrementale* con le tecnologie OLAP

- Riconoscere classi di dati organizzati in un modello multidimensionale (OLAP Data Warehouse) in base ad attributi continui.
- Migliorare le prestazioni in termini di *efficacia* dei sistemi OLAP.
- Fornire un migliore supporto ai processi decisionali di un'organizzazione, accrescendo, la capacità dei *decision maker* di esplorare aspetti che normalmente le tecnologie OLAP non permettono di analizzare.

OLAPBIRCH


0.0.0.0
0.0.0.1
0.0.1.2
0.0.1.3
0.1.2.4
0.1.2.5
0.1.3.6
0.1.3.7
C0
C1
C0
C2
C1
C3
C0


CF – Tree
Profondità = 3
N0
Livello 1
N0
N1
Livello 2
N0
N1
N2
N3
Livello 3
N0
N1
N2
N3
N4
N5
N6
N7
Livello 4
DBSCAN
N 0 ∧ N 1 ∈ C0;
DBSCAN
(N 0 ∧ N 1) ∈ C 0; (N 2 ∧ N 3) ∈ C 1;
DBSCAN
(N 0) ∈ C 0; (N 2 ∧ N 3 ∧ N 1) ∈ C 1;
(N 4 ∧ N 5) ∈ C 2 ; (N 6 ∧ N 7) ∈ C 3;

OLAPBIRCH

Interazione con motore OLAP Mondrian

```
<!ELEMENT Hierarchy ((%Relation;)?,(Level)*,
        (MemberReaderParameter)*, (Attribute)+, (Depth)  )>
    <!ATTLIST Hierarchy
    hasAll (truelfalse) #REQUIRED
    allMemberName CDATA #IMPLIED
    allMemberCaption CDATA #IMPLIED
    primaryKey CDATA #IMPLIED
    primaryKeyTable CDATA #IMPLIED
    defaultMember CDATA #IMPLIED
    memberReaderClass CDATA #IMPLIED>

    <!ELEMENT Attribute EMPTY>
    <!ATTLIST Attribute
    name CDATA #IMPLIED
    table CDATA #REQUIRED
    column CDATA #REQUIRED
    nameColumn CDATA #REQUIRED
    type (Numeric) Numeric #REQUIRED>

    <!ELEMENT Depth EMPTY>
    <!ATTLIST Depth
    value (Numeric) Numeric #REQUIRED>
```

```
      <Attribute name="totalprice" table="orders" column="
        o_totalprice" nameColumn="o_totalprice" type="Integer"/>
      <Attribute name="orderpriority" table="orders" column="
        o_orderpriority" nameColumn="o_orderpriority" type="
        Integer" />
      <Depth value="10"/>

    </Hierarchy>
  </Dimension>

  <Measure name="quantity" column="l_quantity" datatype="Numeric"
      aggregator="distinct-count" visible="true">
    </Measure>
    <Measure name="extendedprice" column="l_extendedprice"
       datatype="Numeric" aggregator="distinct-count" visible="
       true">
    </Measure>
    <Measure name="discount" column="l_discount" datatype="Numeric
       " aggregator="distinct-count" visible="true">
    </Measure>
    <Measure name="tax" column="l_tax" datatype="Numeric"
       aggregator="distinct-count" visible="true">
    </Measure>

  </Cube>
</Schema>
```

OLAPBIRCH

Interazione con motore OLAP Mondrian

Attraverso il nuovo file XML, che descrive le dimensioni di interesse, OLAPBIRCH ha tutte le informazioni per procede:

- nel recupero dei dati numerici dal database, attraverso una query SQL codificata dal file XML
- nella costruzione dell'abero dei CF, poichè nel file XML è automaticamente definito lo spazio dimensionale dei dati
- nel controllo della crescita della struttura gerarchica
- nel salvataggio delle etichette dei cluster per ciascun punto, in quanto nel file XML sono specificate le dimension table primarie per una particolare sessione di analisi

OLAPBIRCH

Interazione con motore OLAP Mondrian

Attraverso il nuovo file XML, che descrive le dimensioni di interesse, OLAPBIRCH ha tutte le informazioni per procede:

- nel recupero dei dati numerici dal database, attraverso una query SQL codificata dal file XML
- nella costruzione dell'abero dei CF, poichè nel file XML è automaticamente definito lo spazio dimensionale dei dati
- nel controllo della crescita della struttura gerarchica
- nel salvataggio delle etichette dei cluster per ciascun punto, in quanto nel file XML sono specificate le dimension table primarie per una particolare sessione di analisi

OLAPBIRCH

Interazione con motore OLAP Mondrian

Attraverso il nuovo file XML, che descrive le dimensioni di interesse, OLAPBIRCH ha tutte le informazioni per procede:

- nel recupero dei dati numerici dal database, attraverso una query SQL codificata dal file XML
- nella costruzione dell'abero dei CF, poichè nel file XML è automaticamente definito lo spazio dimensionale dei dati
- nel controllo della crescita della struttura gerarchica
- nel salvataggio delle etichette dei cluster per ciascun punto, in quanto nel file XML sono specificate le dimension table primarie per una particolare sessione di analisi

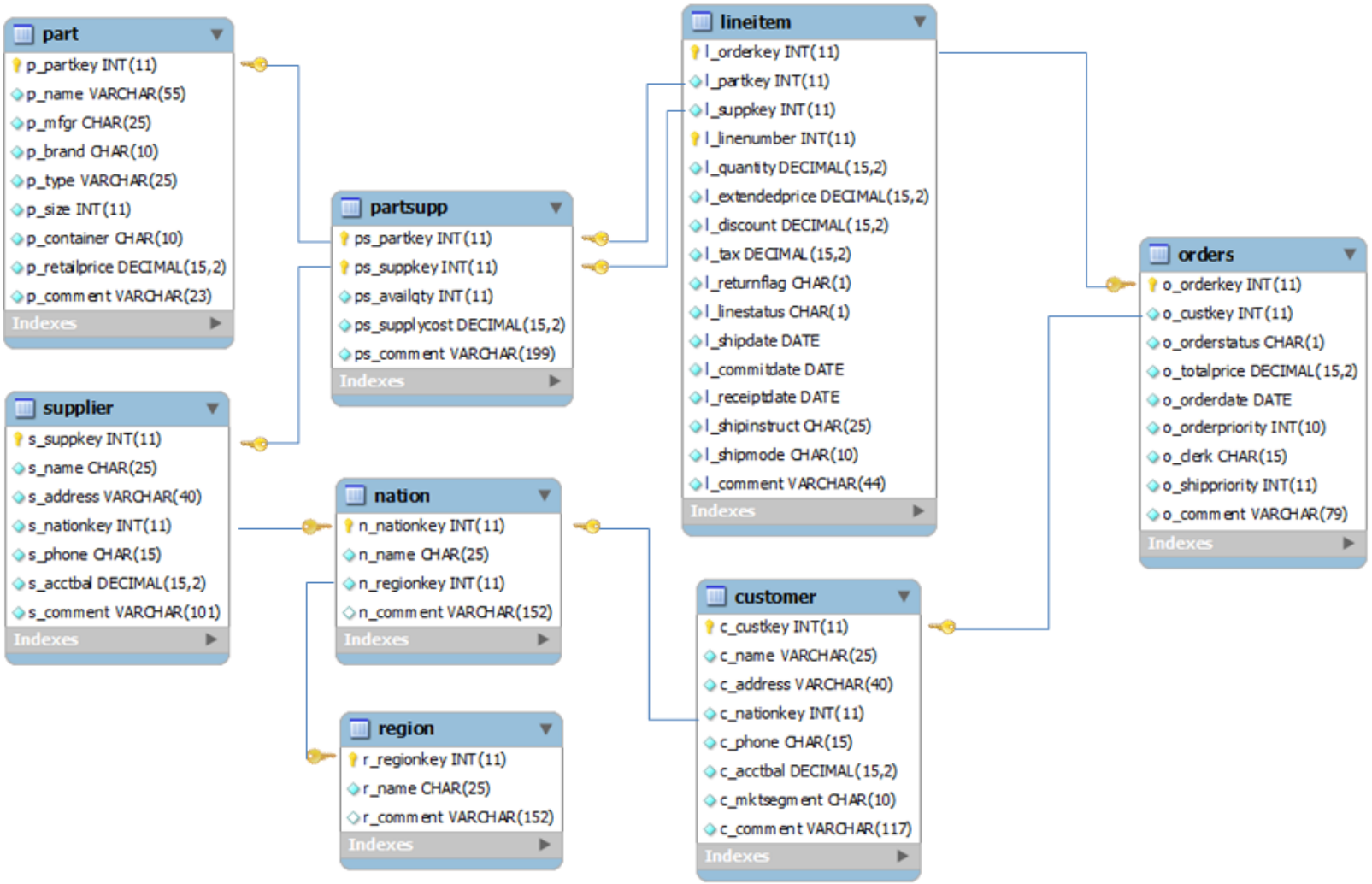
OLAPBIRCH

Interazione con motore OLAP Mondrian

Attraverso il nuovo file XML, che descrive le dimensioni di interesse, OLAPBIRCH ha tutte le informazioni per procede:

- nel recupero dei dati numerici dal database, attraverso una query SQL codificata dal file XML
- nella costruzione dell'abero dei CF, poichè nel file XML è automaticamente definito lo spazio dimensionale dei dati
- nel controllo della crescita della struttura gerarchica
- nel salvataggio delle etichette dei cluster per ciascun punto, in quanto nel file XML sono specificate le dimension table primarie per una particolare sessione di analisi

TPC Benchmark™H



Risultati sperimentali

Diagramma Livello 7 *CF-Tree*

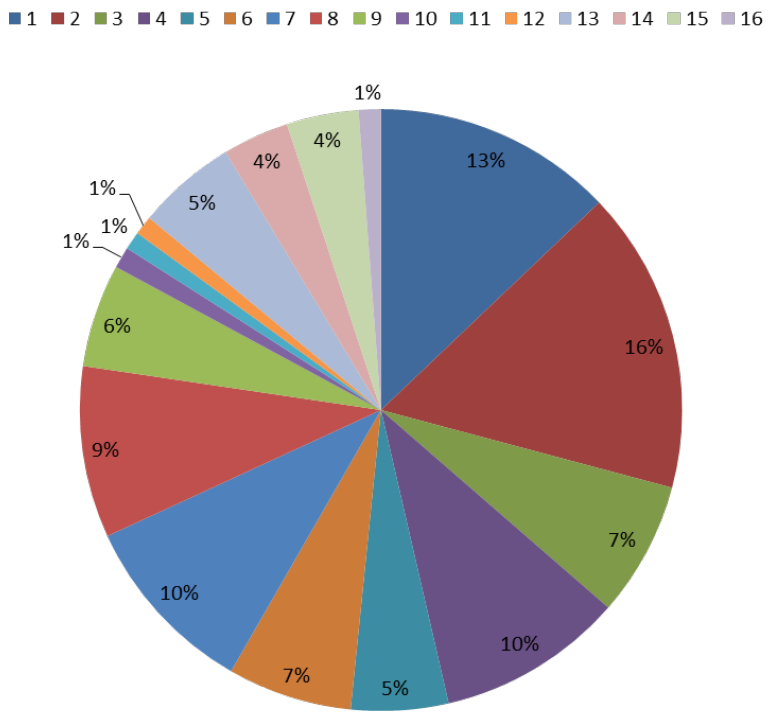
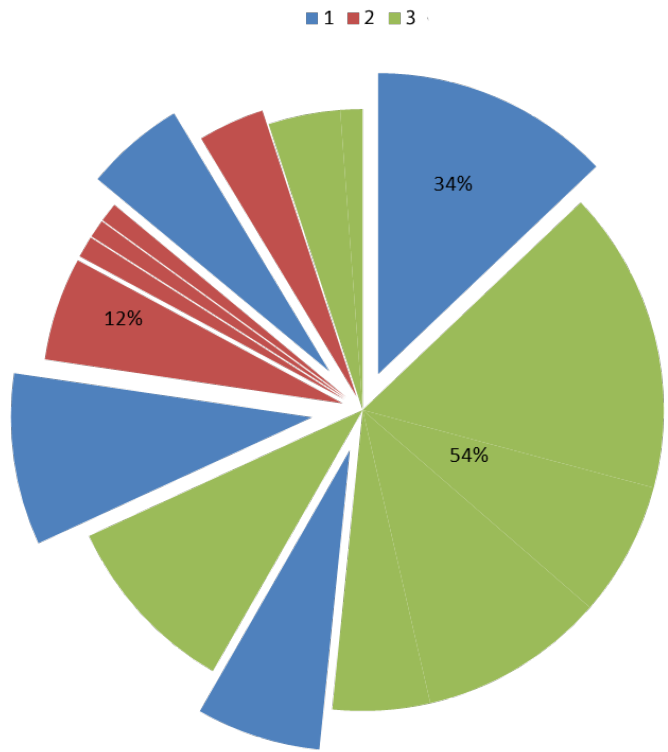
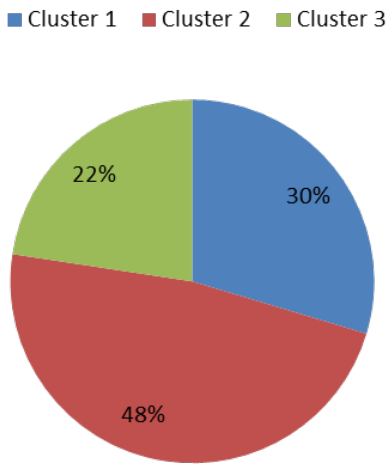


Diagramma Livello 7 DBSCAN

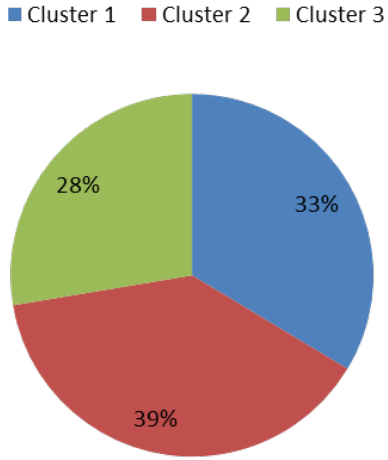


Risultati sperimentali

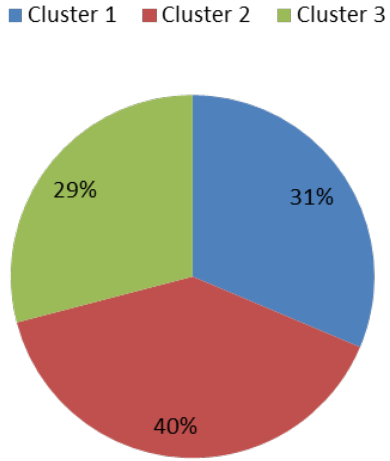
Distribuzione di Totalprice



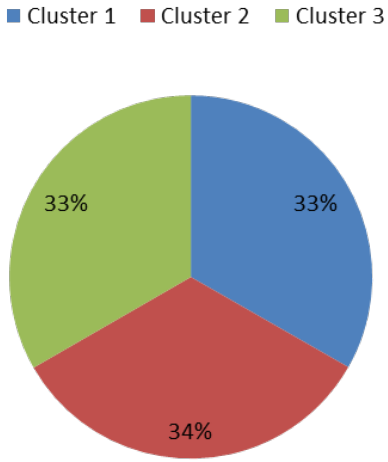
Distribuzione di Linenumber



Distribuzione di Quantity

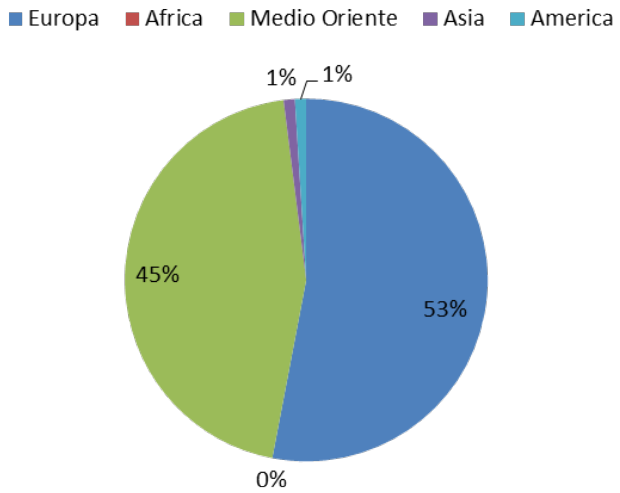


Distribuzione di Acctbal

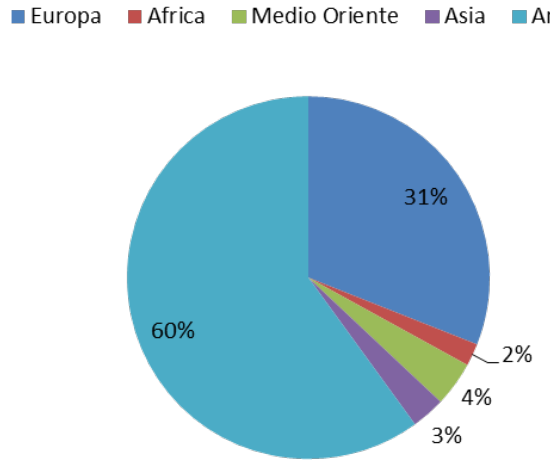


Risultati sperimentali

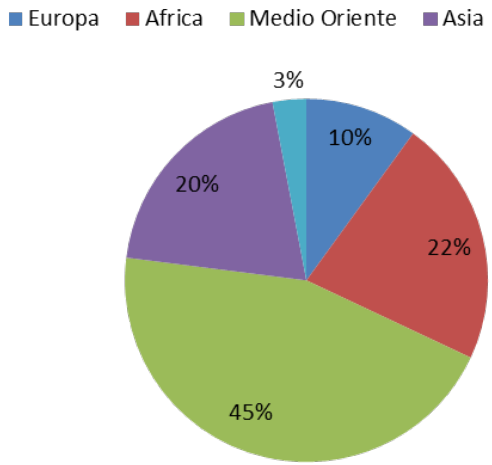
Distribuzione della proprietà Region sul *cluster* 1



Distribuzione della proprietà Region sul *cluster 2*



Distribuzione della proprietà Region sul *cluster 3*



Conclusioni

Il sistema permette, dunque, di:

- mantenere il CF-Tree sempre aggiornato, essendo l'algoritmo di costruzione dello stresso incrementale
- effettuare operazioni di roll-up e drill-down sul CF-Tree in qualsiasi momento e senza lunghe attese
- operare sessioni di analisi su grandi database, indipendentemente dalla piattaforma utilizzata
- identificare cluster qualitativamente soddisfacenti
- scegliere le dimensioni di interesse per una particolare sessione di analisi
- analizzare un DW OLAP senza particolari conoscenze dello stesso, facilitando notevolmente, il processo deduttivo dell'analista

Conclusioni

Il sistema permette, dunque, di:

- mantenere il CF-Tree sempre aggiornato, essendo l'algoritmo di costruzione dello stresso incrementale
- effettuare operazioni di roll-up e drill-down sul CF-Tree in qualsiasi momento e senza lunghe attese
- operare sessioni di analisi su grandi database, indipendentemente dalla piattaforma utilizzata
- identificare cluster qualitativamente soddisfacenti
- scegliere le dimensioni di interesse per una particolare sessione di analisi
- analizzare un DW OLAP senza particolari conoscenze dello stesso, facilitando notevolmente, il processo deduttivo dell'analista

Conclusioni

Il sistema permette, dunque, di:

- mantenere il CF-Tree sempre aggiornato, essendo l'algoritmo di costruzione dello stresso incrementale
- effettuare operazioni di roll-up e drill-down sul CF-Tree in qualsiasi momento e senza lunghe attese
- operare sessioni di analisi su grandi database, indipendentemente dalla piattaforma utilizzata
- identificare cluster qualitativamente soddisfacenti
- scegliere le dimensioni di interesse per una particolare sessione di analisi
- analizzare un DW OLAP senza particolari conoscenze dello stesso, facilitando notevolmente, il processo deduttivo dell'analista

Conclusioni

Il sistema permette, dunque, di:

- mantenere il CF-Tree sempre aggiornato, essendo l'algoritmo di costruzione dello stresso incrementale
- effettuare operazioni di roll-up e drill-down sul CF-Tree in qualsiasi momento e senza lunghe attese
- operare sessioni di analisi su grandi database, indipendentemente dalla piattaforma utilizzata
- identificare cluster qualitativamente soddisfacenti
- scegliere le dimensioni di interesse per una particolare sessione di analisi
- analizzare un DW OLAP senza particolari conoscenze dello stesso, facilitando notevolmente, il processo deduttivo dell'analista

Conclusioni

Il sistema permette, dunque, di:

- mantenere il CF-Tree sempre aggiornato, essendo l'algoritmo di costruzione dello stresso incrementale
- effettuare operazioni di roll-up e drill-down sul CF-Tree in qualsiasi momento e senza lunghe attese
- operare sessioni di analisi su grandi database, indipendentemente dalla piattaforma utilizzata
- identificare cluster qualitativamente soddisfacenti
- scegliere le dimensioni di interesse per una particolare sessione di analisi
- analizzare un DW OLAP senza particolari conoscenze dello stesso, facilitando notevolmente, il processo deduttivo dell'analista

Conclusioni

Il sistema permette, dunque, di:

- mantenere il CF-Tree sempre aggiornato, essendo l'algoritmo di costruzione dello stresso incrementale
- effettuare operazioni di roll-up e drill-down sul CF-Tree in qualsiasi momento e senza lunghe attese
- operare sessioni di analisi su grandi database, indipendentemente dalla piattaforma utilizzata
- identificare cluster qualitativamente soddisfacenti
- scegliere le dimensioni di interesse per una particolare sessione di analisi
- analizzare un DW OLAP senza particolari conoscenze dello stesso, facilitando notevolmente, il processo deduttivo dell'analista

Grazie

GRAZIE PER LA CORTESE ATTENZIONE!