

Esercizi d'esame

DW design

Schema concettuale 1

Una piattaforma di streaming di film che offre lo streaming on demand vuole effettuare delle analisi sui film e i suoi di utenti.

Si vogliono effettuare le analisi per film. Ogni film è identificato da un nome univoco. Ogni film è caratterizzato dal regista che lo ha diretto. Ogni film è diretto da un/a solo/a regista e un regista può dirigere più film.

Ogni film è caratterizzato da uno o più generi cinematografici. I generi cinematografici memorizzati da sistema sono 6: commedia, thriller, animazione, avventura, fantasy e musical.

La società vuole effettuare le analisi rispetto ad alcune informazioni degli utenti. In particolare, vuole effettuare le analisi rispetto alla città, provincia e regione dell'utente, rispetto alla fascia di età ('<25', '25-40', '40-60', '70+') e rispetto al genere.

Il sistema memorizza la data di visione (streaming) del film. In particolare, le analisi saranno effettuate rispetto alla data, tenendo conto del giorno della settimana, mese, e mese dell'anno.

La società vuole analizzare le interazioni degli utenti quali il numero di streaming e il numero medio di streaming per utente

Schema concettuale 1

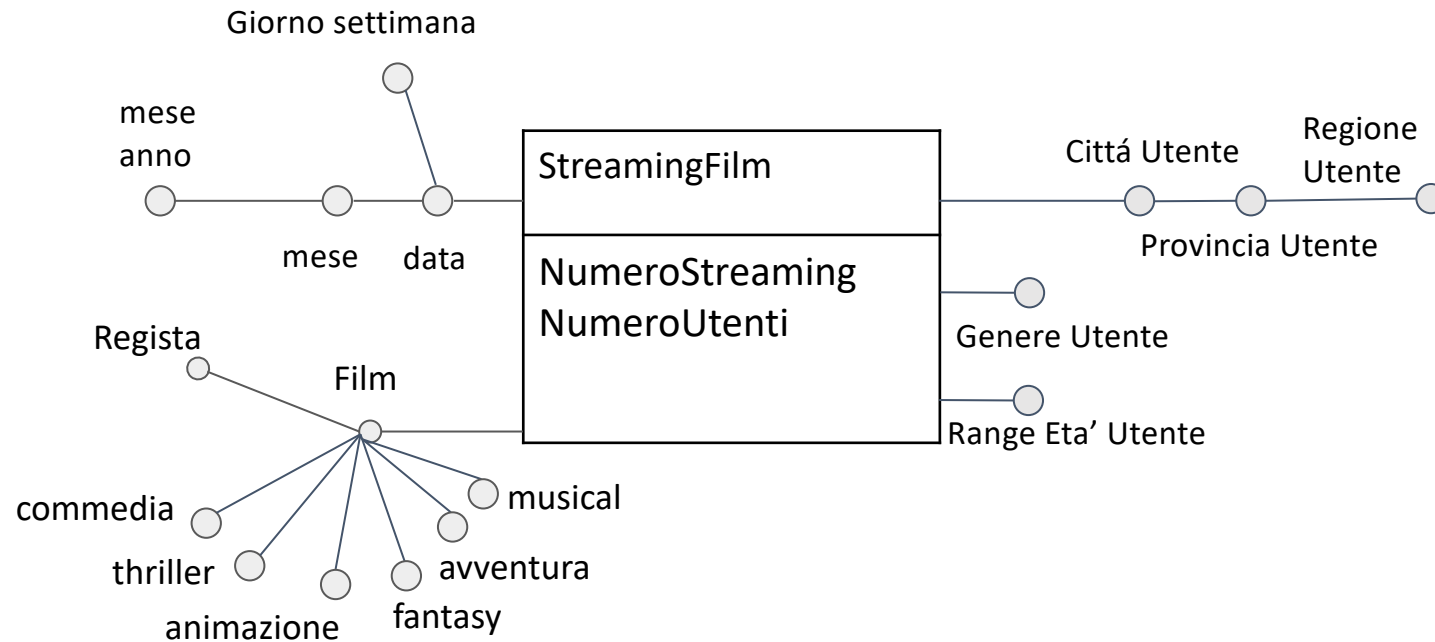
Le analisi devono essere effettuate al variare delle seguenti condizioni:

- film
- Regista del film
- Genere cinematografico
- Città, provincia e regione dell'utente
- Fascia di età dell'utente
- Genere dell'utente
- Data, giorno della settimana, mese e anno di visione.

Esempi di analisi di interesse sono il numero medio di streaming per utente separatamente per fascia di età,, il numero di streaming separatamente per regista.

Definire e caratterizzare il Dimensional Fact Model (modello concettuale) del data warehouse che soddisfi i requisiti applicativi descritti. Per la risoluzione dell'esercizio è possibile utilizzare sia la notazione testuale sia la drawing box

Soluzione schema concettuale 1 - SOL



Schema concettuale 2

Un'azienda che si occupa di gestione delle prenotazioni e servizi di centri medici vuole effettuare delle analisi sulle prenotazioni e sui ricavi dei centri medici.

Si vogliono effettuare le analisi per centro medico. Ogni centro medico è identificato da un nome ed e' caratterizzato dalla sua collocazione geografica (citta', provincia e regione).

L'azienda vuole analizzare le visite rispetto al loro tipo (e.g., visita oculistica, dermatologica, cardiologica). Ogni prenotazione e' associata ad un solo tipo di visita.

Le analisi devono essere effettuate rispetto alla data della visita, il mese e anno e la fascia oraria (8-10, 10-12, 12-14, 14-16).

Schema concettuale 2

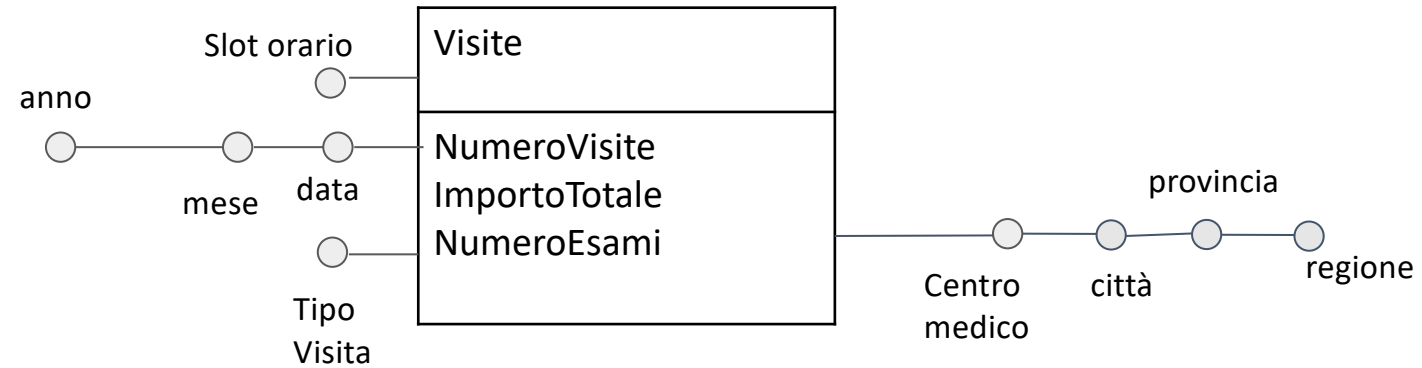
L'azienda è interessata ad analizzare il numero di visite, il numero di esami medio che sono effettuati ad ogni visita e l'importo totale speso, al variare delle seguenti condizioni:

- Data, mese, anno della visita
- Fascia oraria della visita
- Nome del centro medico
- Città', provincia, regione del centro medico
- Tipo di visita

Esempi di analisi di interesse sono il numero medio di esami effettuati per tipo di visita, e l'importo medio per numero di visite.

Definire e caratterizzare il Dimensional Fact Model (modello concettuale) del data warehouse che soddisfi i requisiti applicativi descritti. Per la risoluzione dell'esercizio è possibile utilizzare sia la notazione testuale sia la drawing box

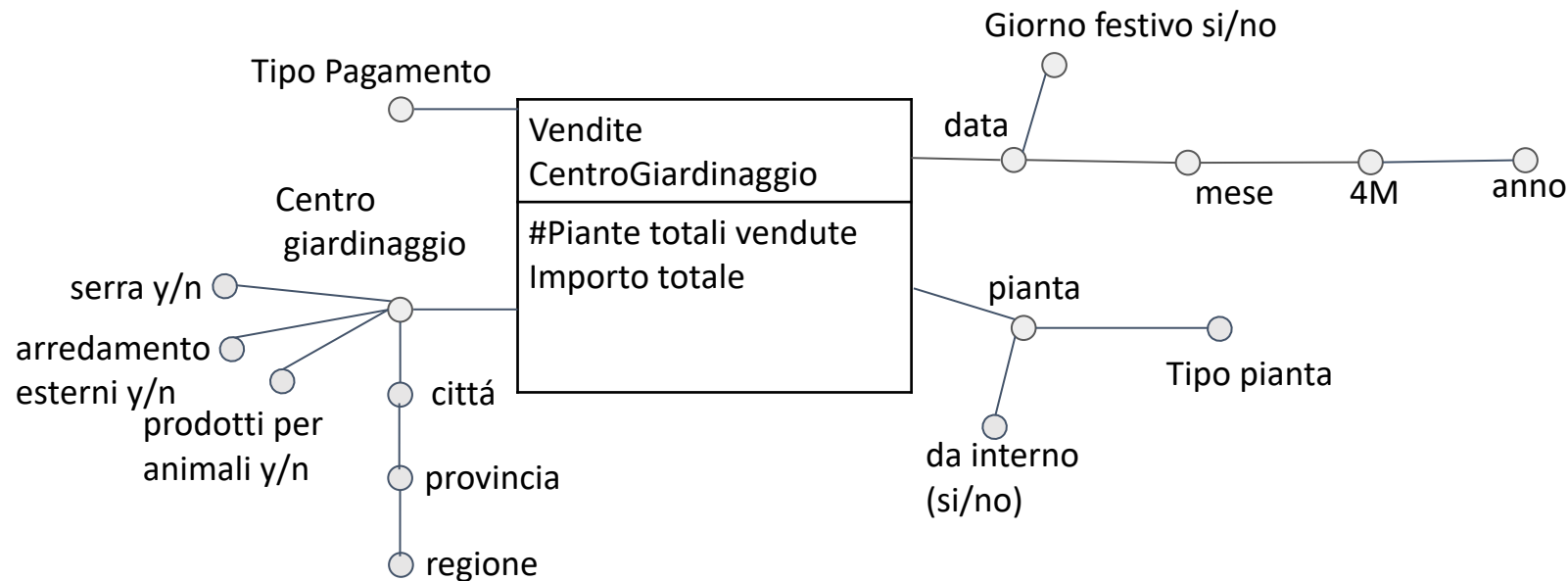
Soluzione schema concettuale 2 - SOL



Esercizio Progettazione Logica

Una catena di Centri di giardinaggio con sedi in tutta Italia è interessata ad analizzare le vendite di piante nei suoi centri.

Dato il Dimensional Fact Model riportato in Figura progettato per effettuare le analisi di interesse, definire il corrispondente modello logico relazionale (evidenziando le chiavi primarie di ogni tabella).



Tipo Pagamento può assumere i seguenti valori: “Contante”, “Carta”, e “Satispay ”

Esercizio Progettazione Logica - SOL

Pianta(**CodPianta**, Pianta, TipoPianta, DalInterno)

Tempo(**CodTempo**, Data, GiornoFestivo, Mese, 4M, anno)

CentroGiardinaggio(**CodCentro**, NomeCentroGiardinaggio, citta',
provincia, regione, serra, arredamentoEsterni, prodottiPerAnimali)

Vendite(**CodTempo**, **CodPianta**, **CodCentro**, **TipoPagamento**,
#piantevendute, importo totale,)

SQL Esteso

Dato il seguente schema logico relazione di un sistema di data warehouse (le chiavi primarie sono sottolineate e in grassetto)

CARATTERISTICHE-CONSULENTI (IDCarConsulenti, Fascia-Età, Ruolo, Categoria-Consulenti)

SEDE-AZIENDA (IDSedeAzienda, Sede-Azienda, Azienda, Dimensione-Azienda, Tipologia-Azienda, Categoria-Azienda, Città-Sede-Azienda, Regione-Sede-Azienda, Stato-Sede-Azienda)

TEMPO (IDTempo, Mese, 2-Mesi, 3-Mesi, 4-Mesi, 6-Mesi, Anno)

TIPOLOGIA_CONSULENZA

(IDTipologiaConsulenza, TipologiaConsulenza, CategoriaInnovazione, CategoriaTecnologia, CategoriaStrategia&Business)

ATTIVITA'-DI-CONSULENZE (IDCarConsulenti, IDSedeAzienda, IDTempo, IDTipologiaConsulenza, CostoTotale, NumeroOre, NumeroConsulenti)

Scrivere la seguente interrogazione in SQL esteso

SQL Esteso – Query 1

Separatamente per anno e regione della sede azienda, considerando le attività di consulenza svolte da consulenti afferenti alla fascia di età (35-50] e di categoria *specialist*, visualizzare

- il costo medio mensile delle consulenze
- la percentuale del costo totale delle consulenze rispetto al costo totale complessivo per anno e stato della sede azienda
- la posizione in una graduatoria (rank) in ordine decrescente di numero medio di consulenti mensili

Si effettui l'analisi separatamente per Tipologia Consulenza.

SQL Esteso – Query 1 - SOL

```
SELECT Anno, Regione-Sede-  
Azienda, TipologiaConsulenza, SUM(CostoTotale)/COUNT(DISTINCT Mese),  
100* SUM(CostoTotale)/ SUM(SUM(CostoTotale)) OVER(PARTITION BY Anno,  
Stato-Sede-Azienda, TipologiaConsulenza),  
RANK () OVER (PARTITION BY TipologiaConsulenza ORDER BY  
SUM(NumeroConsulenti)/COUNT(DISTINCT Mese) DESC)  
FROM ATTIVITA'-DI-CONSULENZE AC, CARATTERISTICHE-CONSULENTI CC, SEDE-  
AZIENDA SA, TEMPO T, TIPOLOGIA_CONSULENZA TC  
WHERE AC.IDCarConsulenti= CC.IDCarConsulenti AND  
AC.DSedeAzienda=SA.DSedeAzienda  
AND TC.IDTipologiaConsulenza=AC.IDTipologiaConsulenza AND Fascia-Età='35-  
50' AND Categoria-Consulenti= 'Specialist' AND AC.IDTempo=T.IDTempo  
GROUP BY Anno, Regione-Sede-Azienda, Stato-Sede-Azienda,  
TipologiaConsulenza
```

SQL Esteso – Query 2

Scrivere la seguente interrogazione in SQL esteso

Separatamente per azienda e semestre (6-Mesi), considerando solo le aziende di dimensioni medie e gli anni dal 2021 al 2023, visualizzare:

- la percentuale del costo delle consulenze rispetto all'importo annuale per azienda
- l'importo medio orario delle consulenze
- la posizione in una graduatoria (rank), separatamente per tipologia azienda, sui costi totali delle consulenze in ordine decrescente di importo

SQL Esteso – Query 2 - SOL

```
SELECT 6-Mesi, Azienda, SUM(CostoTotale)/SUM(NumeroOre),  
100* SUM(CostoTotale)/ SUM(SUM(CostoTotale)) OVER(PARTITION BY  
Anno, Azienda),  
RANK () OVER (PARTITION BY TipologiaAzienda ORDER BY  
SUM(CostoTorale) DESC)  
FROM ATTIVITA'-DI-CONSULENZE AC, SEDE-AZIENDA SA, TEMPO T  
WHERE AC.DSedeAzienda=SA.DSedeAzienda  
AND AC.IDTempo=T.IDTempo AND Dimensione= 'Medie AND (Anno  
>='2021' AND Anno<='2023')  
GROUP BY 6-Mesi, Azienda, Anno, Tipologia-Azienda
```

SQL Esteso – Query 3

Scrivere la seguente interrogazione in SQL esteso

Separatamente per tipologia consulenza e quadrimestre (attributo 4-Mesi), visualizzare:

- il numero medio bimestrale di ore di consulenza
- la percentuale del numero di ore di consulenza rispetto al totale annuale (per tipologia di consulenza)
- il totale cumulativo del costo di consulenza al trascorrere dei quadrimestri, separatamente per anno (e per tipologia di consulenza).

Si effettui l'analisi separatamente per categoria azienda.

SQL Esteso – Query 3 - SOL

```
SELECT TipologiaConsulenza,  
Quadrimestre, CategoriaAzienda, SUM(NumeroOre)/COUNT(DISTINCT 2-Mesi),  
    100* SUM(NumeroOre)/SUM(SUM(NumeroOre)) OVER(PARTITION BY Anno,  
TipologiaConsulenza, CategoriaAzienda),  
SUM(SUM(CostoTotale)) OVER (PARTITION BY Anno,  
TipologiaConsulenza, CategoriaAzienda ORDER BY Quadrimestre ROWS  
UNBOUNDED PRECEDINGS)  
FROM ATTIVITA'-DI-CONSULENZE AC, TIPOLOGIA_CONSULENZA TC, TEMPO T, SEDE-  
AZIENDA SA  
WHERE AC.IDTipologiaConsulenza=TC.IDTipologiaConsulenza AND  
AC.IDTempo=T.IDTempo AND SA.IDSedeAzienda=AC.IDSedeAzienda  
GROUP BY TipologiaConsulenza, Quadrimestre, Anno, CategoriaAzienda
```

Vista materializzata

Dato lo schema logico precedente, considerare le seguenti query di interesse:

1. Considerando solo le aziende con dimensione maggiore di 50 dipendenti site in Piemonte, visualizzare il costo totale e il numero totale di ore di consulenza per ogni città dell'azienda e anno.
2. Visualizzare il costo totale per ogni categoria dei consulenti, anno, regione.
3. Per le tipologie di consulenza di categoria innovazione, visualizzare il costo medio per consulente, separatamente per ogni semestre (attributo 6-Mesi) e fascia d'età.

Dato lo schema logico precedente, si svolgano le seguenti attività

1. Definire una vista materializzata con CREATE MATERIALIZED VIEW, in modo da ridurre il tempo di risposta delle query di interesse da (1) a (3) sopra riportate. In particolare si specifichi la query SQL associata al **Blocco A** nella seguente istruzione:

```
CREATE MATERIALIZED VIEW ViewCorsi  
BUILD IMMEDIATE  
REFRESH FAST ON COMMIT  
AS  
  Blocco A
```

Vista materializzata - SOL

Blocco A - Query per vista materializzata

```
SELECT CittàSedeAzienda, RegioneSedeAzienda, Dimensione-Azienda, 6-Mesi,  
Anno, CategoriaConsulenti, Fascia-Età, CategoricalInnovazione, SUM(CostoTotale) AS  
TotCosto, SUM(NumeroOre) AS TotOre, SUM(NumeroConsulenti) AS TotConsulenti  
FROM CARATTERISTICHE-CONSULENTI CC, SEDE-AZIENDA SA, TIPOLOGIA-  
CONSULENZA TC, TEMPO T, ATTIVITA-DI-CONSULENZE AC  
WHERE CC.IDCarConsulenti = AC.IDCarConsulenti AND SA.IDSedeAzienda =  
AC.IDSedeAzienda AND T.IDTempo = AC.IDTempo AND TC.IDTipologiaConsulenza =  
AC.IDTipologiaConsulenza  
GROUP BY CittàSedeAzienda, RegioneSedeAzienda, 6-Mesi, Anno,  
CategoriaConsulenti, Fascia-Età, CategoricalInnovazione, Dimensione-Azienda
```

Vista materializzata - SOL (supporting)

Query 1

GB: Citta;SedeAzienda, Anno

Where: DimensioneAzienda>50,
RegioneSedeAzienda='Piemonte'

Aggr: SUM(Costo), Sum(NumeroOre)

Tab: SedeAzienda, Tempo

Query 2:

GB: CategoriaConsulenti, Anno, RegioneSedeAzienda

Where: /

Aggr: SUM(Costo)

Tab: CaratterisitchConsulenti, Tempo, SedeAzienda

Query 3:

GB: 6-Mesi, FasciaEta'

Where: CategoricalInnovazione = 'Yes'

Aggr: SUM(Costo),
SUM(NumeroConsulenti)

Tab: Tempo, CaratterisitchConsulenti

Vista materializzata – SOL (supporting)

Query 1

GB: **CittaSedeAzienda**, Anno

Where: **DimensioneAzienda**>50,
RegioneSedeAzienda=‘Piemonte’

Aggr: SUM(Costo), Sum(NumeroOre)

Tab: SedeAzienda, Tempo

Query 2:

GB: **CategoriaConsulenti**, Anno, RegioneSedeAzienda

Where: /

Aggr: SUM(Costo)

Tab: CaratterisitchConsulenti, Tempo, SedeAzienda

Query 3:

GB: **6-Mesi**, **FasciaEta’**

Where: **CategoriaInnovazione** = ‘Yes’

Aggr: SUM(Costo),
SUM(NumeroConsulenti)

Tab: Tempo, CaratterisitchConsulenti

Identificatore: CittaSedeAzienda, DimensioneAzienda, CategoriaConsulenti, 6-Mesi, FasciaEta, CategoriaInnovazione

Domanda 1

Quale delle seguenti affermazioni sulla Support Vector Machine non è corretta?

- a. Nessuna affermazione è errata
- b. Non può essere usata per problemi multiclasse
- c. E' robusta al rumore
- d. Ha un'interpretazione geometrica
- e. Può essere adattata a classi non separabili linearmente

Domanda 1 - SOL

Non può essere usata per problemi multiclasse

Domanda 2

- Stiamo lavorando su un dataset composto da 20 campioni contenente 4 differenti classi. Il vettore seguente rappresenta la ground truth per le classi da 0 a 3:

$$gt = [2, 3, 2, 0, 1, 3, 0, 3, 0, 3, 3, 2, 1, 1, 0, 0, 0, 3, 3, 3]$$

Un classificatore predice il seguente vettore di classi:

$$pr = [2, 0, 0, 3, 2, 0, 0, 0, 0, 2, 0, 2, 3, 1, 2, 3, 3, 2, 0, 3]$$

Quale delle seguenti affermazioni è vera?

- a. La recall della classe 0 è più bassa della precision
- b. L'accuratezza supera lo 0.4
- c. La precisione media supera lo 0.4
- d. La precisione per la classe 0 è la più bassa
- e. Nessuna risposta è corretta

Domanda 2 - SOL

La precisione media supera lo 0.4

Precisione(C) = predizione corrette classe C rispetto predizione classe C

Recall(C) = predizione corrette classe C rispetto punti appartengono a classe C

$$\text{accuracy} = \mathbf{6/20} = 0.3$$

$$\text{precisions} = [2/8, 1/1, \mathbf{2/6}, 1/5]$$

$$\text{recall_0} = 2/6 = \text{TP_0} / (\text{TP_0} + \text{FN_0})$$

$$\text{f1_score} = 2 * \text{precision}[0] * \text{recall_0} / (\text{precisions}[0] + \text{recall_0})$$

$$\text{mean_prec} = \text{sum}(\text{precisions})/4 = 0.44583333333333333336$$

Domanda 3

Abbiamo a disposizione 5 punti con le seguenti coordinate (x, y):

- A (0, 0)
- B (0, 4)
- C (6, 2)
- D (4, 2)
- E (2, 4)

Vogliamo applicare il clustering di tipo K-Means per 2 iterazioni con i seguenti centroidi iniziali: (0, 0) e (5, 5). La metrica di distanza è

$$d(a, b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$$

dove a e b sono due punti. Quali cluster si ottengono?

- a. Nessuna risposta è corretta
- b. {A, B} {C, D, E}
- c. {A, E} {B, C, D}
- d. {A, E} {B, C, D}
- e. {A, B, E} {C, D}

Domanda 3 - SOL

$\{A, B\} \{C, D, E\}$

Step1

- $c_0 = (0, 0)$
- $c_1 = (5, 5)$
- $\{A, B\} \{C, D, E\}$
- $d(A, c_0)^2 = 0;$
- $d(B, c_0)^2 = 16;$
- $d(C, c_1)^2 = 1 + 9;$
- $d(D, c_1)^2 = 1 + 9;$
- $d(E, c_1)^2 = 9 + 1$

Domanda 3 - SOL

Step2

- $c_0 = (0, 2)$
- $c_1 = (4, 8/3)$
- $\{A, B\} \{C, D, E\}$
- $d(A, c_0)^2 = 4$
- $d(B, c_0)^2 = 4$
- $d(C, c_1)^2 = 4 + 4/9$
- $d(D, c_1)^2 = 0 + 4/9$
- $d(E, c_1)^2 = 4 + 16/9$

Domanda 4

Dato il seguente dataset di transazioni:

- {A, C}
- {B, D, E}
- {A, B, D, E}
- {C, D}
- {A, B, C, E}
- {A, D}
- {B, C, D, E}
- {A, B, C, D, E}
- {C, E}
- {A, B, C, D, E}

Applicare l'algoritmo APriori con un $\text{minsup} > 3$. Quali sono gli itemset di lunghezza 1, 2 e 3 che vengono generati da Apriori dopo i passi di join e prune (con principio Apriori), prima del conteggio del supporto nella base dati?

Domanda 4 - SOL

Livello 1

A 6

B 6

C 7

D 7

E 7

Livello 2

AB 4

AC 4

AD 4

AE 4

BC 4

BD 5

BE 6

CD 4

CE 5

DE 5

Livello 3

ABC

ABD

ABE

ACE

ACD

ADE

BCD

BCE

BDE

CDE

Domanda 4+

Dato il seguente dataset di transazioni:

- C D
- A C E
- C D E
- A B
- A B D E
- D E
- A C D E
- A E
- D E
- A B C

Applicare l'algoritmo APriori con un minsup ≥ 2 . Quali sono gli itemset di lunghezza 1, 2 e 3 che vengono generati da Apriori dopo i passi di join e prune (con principio Apriori), prima del conteggio del supporto nella base dati?

Domanda 4* - SOL

Livello 1

A 6

B 3

C 5

D 6

E 7

Livello 2

AB 3

AC 3

AD 2

AE 4

BC 1 ----

BD 1 ----

BE 1 ----

CD 3

CE 3

DE 4

Livello 3

ABC ----

ABD ----

ABE ----

ACD

ACE

ADE

CDE