

Progettazione Logica

30/05/2014 9.37

A.GORI - 2011

1

Progettazione Logica

- La versione multidimensionale dei dati usata nel DW può essere realizzata usando modelli logici diversi:

ROLAP

- Modello Relazionale: realizza la visione multidimensionale basandosi su un R-DBMS

MOLAP

- Modello Multi-Dimensionale: realizza la versione multi-dimensionale basandosi su un MD-DBMS

30/05/2014 9.37

A.GORI - 2011

2

Progettazione Logica

- Obiettivo della fase di progettazione logica è la generazione di uno schema logico che minimizzi il tempo di risposta delle query presenti nel workload nel rispetto dei vincoli di spazio e di tempo
- I passi principali della progettazione logica e fisica sono:
 - Traduzione degli schemi di fatto in schemi logici (Star Schema, Snowflake Schema, Constellation Schema)
 - Materializzazione delle viste
 - Creazione delle tabelle
 - Partizionamento verticale
 - Partizionamento orizzontale

30/05/2014 9.37

A.GORI - 2011

3

Star Schema

30/05/2014 9.37

A.GORI - 2011

4

Star Schema

- La modellazione multidimensionale su sistemi relazionali è basata sul cosiddetto SCHEMA A STELLA o STAR SCHEMA e sulle sue varianti

30/05/2014 9.37

A.GORI - 2011

5

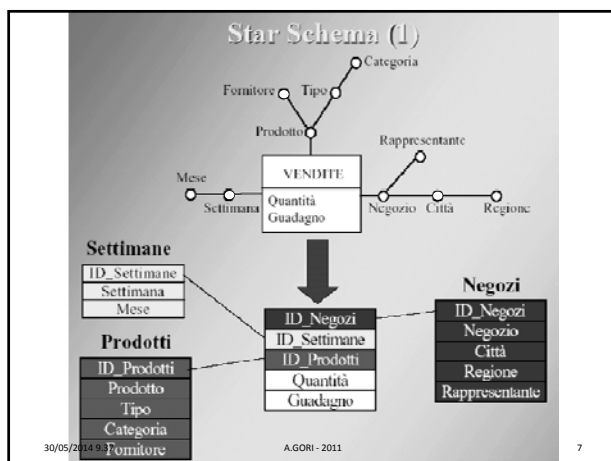
Star Schema

- Uno schema a stella è composto da:
 - Una relazione centrale (la **Fact Table**), che individua l'oggetto dello studio. Essa contiene tutte le misure e gli attributi descrittivi direttamente collegati al fatto.
 - Un insieme di relazioni $DT_1, \dots, DT_n$ (**Dimension Table**), ognuna corrispondente ad una dimensione.
In pratica, per ogni gerarchia, viene costruita una DT che ne contiene gli attributi.

30/05/2014 9.37

A.GORI - 2011

6



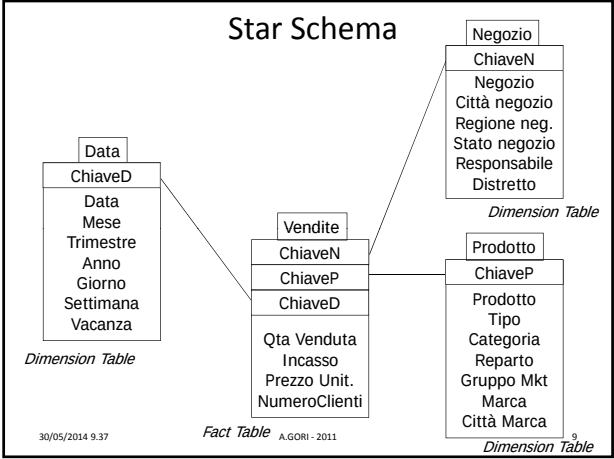
Star Schema

- Ogni DT_i è caratterizzata da una chiave primaria d_i e da un insieme di attributi che descrivono le dimensioni di analisi a diversi livelli di aggregazione.
- La chiave della FT è composta dalle chiavi delle varie DT , ossia dall'insieme $d_1, \dots, d_n$. Inoltre FT contiene un attributo per ogni misura.
- L'accesso ai dati e la visione multidimensionale dei dati avviene tramite join tra la Fact Table e le Dimension Table.

30/05/2014 9.37

A.GORI - 2011

8



Star Schema

La visione multidimensionale dei dati si ottiene facendo il join tra la fact table e le dimension table.

```
SELECT *
FROM VENDITE AS FT, PRODOTTO AS DT1, NEGOZIO AS
DT2, DATA AS DT3
WHERE FT.ChiaveP=DT1.ChiaveP AND
FT.ChiaveN=DT2.ChiaveN AND
FT.ChiaveD=DT3.ChiaveD
```

30/05/2014 9.37 A.GORI - 2011 10

Star Schema

VENDITE

chiaveN	chiaveD	chiaveP	Qta V.	Incaso	
1	1	1	170	85	
2	1	2	320	460	
3	2	3	412	412	
.....					

NEGOZIO

chiaveN	negozio	città	regione	...
1	COOP1	Bologna	Em.Rom.	
2	COOP2	Roma	Lazio	
3	COOP3	Roma	Lazio	
.....				

Le DT non sono in terza forma normale per la presenza di dip.funzionali

Una possibile istanza dello schema a stella (1/2).

30/05/2014 9.37 A.GORI - 2011 11

Star Schema

PRODOTTO

chiaveP	prodotto	tipo	Categ.	marca	
1	Latte A	Latticino	Alimentari	Slurp	
2	Latte B	Latticino	Alimentari	Gnam	
3	Yogurt X	Latticino	Alimentari	Gnam	
.....					

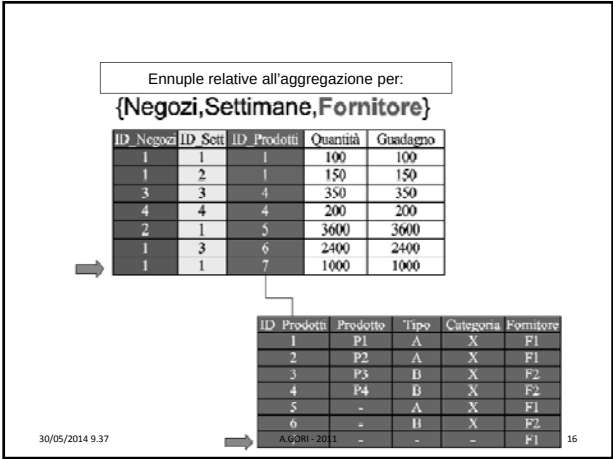
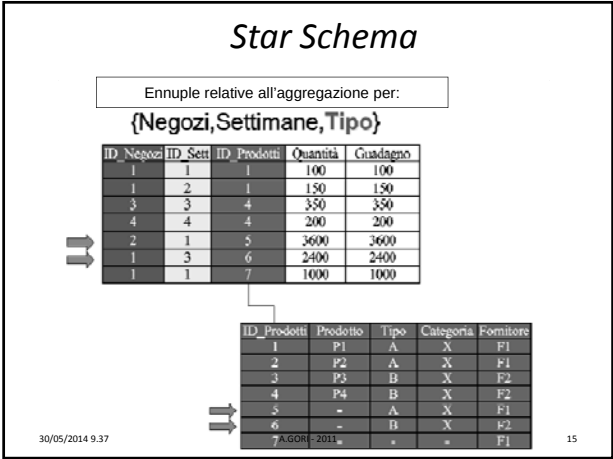
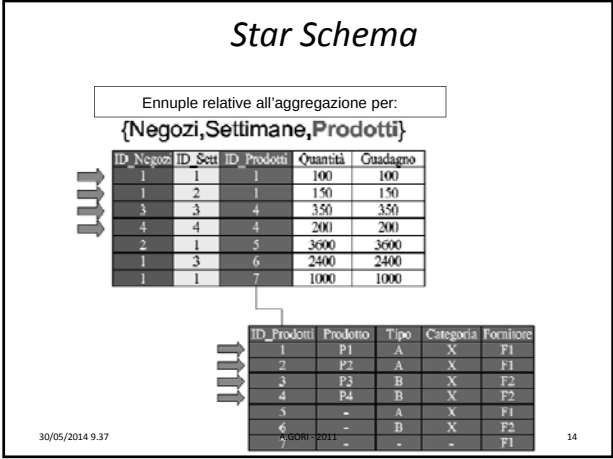
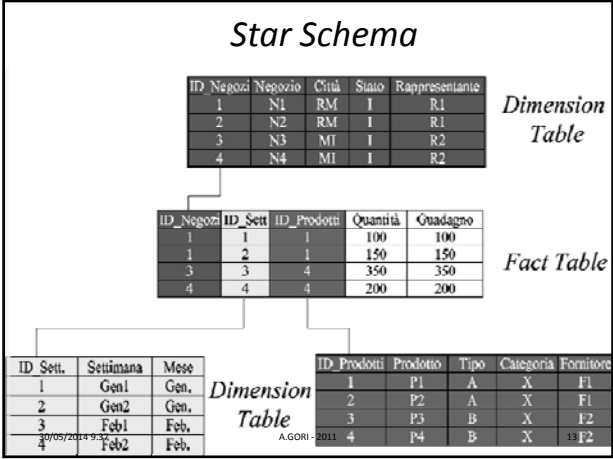
Ridondanza

DATA

chiaveD	data	Mese	anno	
1	02/09/2002	09/2002	2002	
2	08/09/2002	09/2002	2002	
3	03/10/2002	10/2002	2002	
.....				

Una possibile istanza dello schema a stella (2/2).

30/05/2014 9.37 A.GORI - 2011 12



Star Schema

- La caratteristica più evidente di questa soluzione è la denormalizzazione delle DT:
 - Riduzione del costo di reperimento dati grazie alla riduzione del numero di Join necessari
 - Aumento della dimensione delle tabelle
- La FT ha un'elevata dimensione e contiene ennuple relative a diversi livelli di aggregazione
- L'elevata dimensione può incidere sui tempi di accesso
- Non ci sono problemi di sparsità, in quanto vengono memorizzate solo le ennuple corrispondenti a punti dello spazio multi-dimensionale per cui esistono dati.

30/05/2014 9.37

A.GORI - 2011

17

Snowflake Schema

30/05/2014 9.37

A.GORI - 2011

18

- La denormalizzazione dello Star Schema è evidenziata dalla presenza di dipendenze funzionali nelle DT.

Negoziò → Città negozio
Città negozio → Regione

30/05/2014 9.37

A.GORI - 2011

19

Snowflake Schema

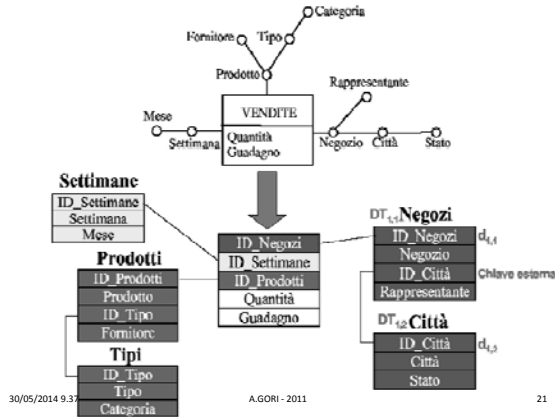
- Lo Schema a fiocco di neve (Snowflake Schema) consiste nella riduzione del livello di denormalizzazione al fine di ottenere uno schema logico più vicino ai dettami della teoria relazionale
 - Quanto si deve spingere la "normalizzazione"?
- La normalizzazione che viene fatta è di solito **parziale** sulle Dimension Table.

30/05/2014 9.37

A.GORI - 2011

20

Lo snowflake schema



Snowflake Schema

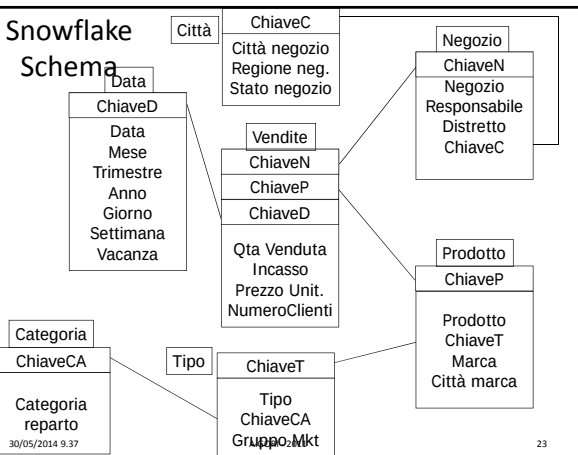
- Uno schema Snowflake si ottiene da uno a stella decomponendo una o più DT_i in più tabelle $DT_{i,1}, \dots, DT_{i,m}$ in modo da eliminare alcune delle dipendenze funzionali presenti in esse.
- Ogni DT sarà caratterizzata da:
 - Una chiave primaria $d_{i,j}$
 - Un sottoinsieme degli attributi di DT_i che dipendono funzionalmente da $d_{i,j}$
 - Zero o più chiavi esterne riferite ad altre $DT_{i,j}$ necessarie a garantire la ricostruibilità del contenuto di DT_i .
- Chiamiamo primarie le DT agganciate direttamente sulla fact table, secondarie le rimanenti

30/05/2014 9.37

A.GORI - 2011

22

Snowflake Schema



Snowflake Schema

- Nell'esempio precedente, si ottiene una parziale normalizzazione tramite l'introduzione delle tabelle CITTA', TIPO e CATEGORIA.
- Ogni passo di normalizzazione lavora su di un "percorso" dello schema di fatto ed individua una sotto-gerarchia che viene separata in una DT secondaria

30/05/2014 9.37

A.GORI - 2011

24

Snowflake Schema

- Conseguenze (1/2):

Spazio richiesto per la memorizzazione complessivamente si riduce
Es.



- Le corrispondenze tra i valori degli attributi Città e Regione vengono memorizzate una sola volta
- Si utilizza una chiave surrogata ChiaveC per indicare l'abbinamento tra negozio e città, che occuperà un numero di byte (es. 4 byte) inferiore rispetto al numero di byte necessari per il campo Città (es. 30 byte)



È necessario introdurre nuove chiavi che consentano di determinare le corrispondenze tra DT primarie e secondarie (es. ChiaveT chiave del Tipo inserita come chiave esterna nella tabella Prodotto)

30/05/2014 9.37

A.GORI - 2011

25

Snowflake Schema

- Conseguenze (2/2):



L'esecuzione di query che riguardano solo gli attributi contenuti nella FT e nelle **DT primarie** è avvantaggiata poiché i join coinvolgono tabelle di dimensioni inferiori



Al contrario, il tempo di esecuzione delle query che coinvolgono attributi delle **DT secondarie** aumenta poiché è necessario un maggior numero di join.

30/05/2014 9.37

A.GORI - 2011

26

Snowflake Schema

- Nella decomposizione di uno schema a stella per il passaggio ad uno schema a fiocco di neve, occorre porre particolare **ATTENZIONE** a:

- Non ridurre il potere informativo
- Contemplare tutti gli attributi
- Rispettare le dipendenze funzionali
- Suddividere nelle relazioni il giusto insieme di attributi
- Affinchè lo **snowflaking** sia efficace, tutti gli attributi del sottoalbero dell'attributo da cui ha origine la normalizzazione devono essere spostati nella nuova relazione.

30/05/2014 9.37

A.GORI - 2011

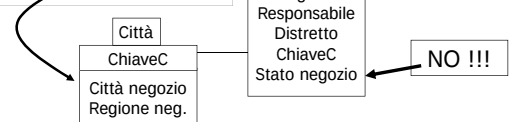
27

Snowflake Schema

Esempio (dipendenze funzionali transitive in cascata):

chiaveN → negozio, negozio → città neg.,
città neg. → regione neg., regione neg. → stato neg.

Occorre che tutti gli attributi che dipendono (funzionalmente) transitivamente dall'attributo di snowflaking siano posti nella nuova relazione.

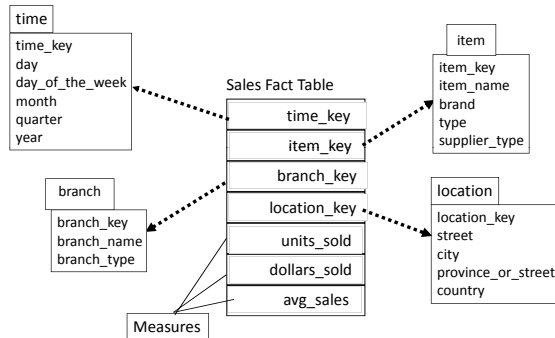


30/05/2014 9.37

A.GORI - 2011

28

Esempio di Star Schema



30/05/2014 9.37

A.GORI - 2011

29

Star VS Snowflake

- **Esistono pareri contrastanti sull'utilità dello snowflaking:**
 - ✓ Contrasta con la filosofia del data warehousing
 - ✓ Rappresenta un inutile "abbellimento" dello schema
- **Può essere utile**
 - ✓ Quando il rapporto tra le cardinalità della dimension table primaria e secondaria è elevato, poiché determina un forte risparmio di spazio

30/05/2014 9.37

A.GORI - 2011

30

Occupazione di memoria

ESEMPIO:

Stima dell'occupazione di memoria di una base di dati dimensionale di esempio

- tempo: 2 anni di 350 giorni, ovvero 700 giorni
 - negozi: 300 negozi
 - prodotti: 30.000 prodotti
 - fatti relativi alle vendite
- ipotizziamo un livello di sparsità del 10% delle vendite giornaliere dei prodotti nei negozi ovvero, che ogni negozio venda giornalmente 3.000 diversi prodotti
 $700 \times 300 \times 3000 = 630.000.000$ record

30/05/2014 9.37

A.GORI - 2011

31

Resistere allo snowflaking

- Lo snowflaking può rivelarsi molto svantaggioso
– vano per l'occupazione di memoria
- ad esempio, supponiamo che la dimensione prodotto contenga 30.000 record, di circa 2.000 byte ciascuno occupando quindi 60MB di memoria primaria
- la tabella fatti contiene invece 630.000.000 record, di circa 10 byte ciascuno occupando quindi 6.3GB di memoria primaria
- le tabelle fatti sono sempre molto più grandi delle tabelle dimensione associate anche riducendo l'occupazione di memoria della dimensione prodotto del 100%, l'occupazione di memoria complessiva è ridotta di meno dell'1% può peggiorare decisamente le prestazioni

30/05/2014 9.37

A.GORI - 2011

32

Snowflake schema

- **Vantaggi del modello "a fiocco di neve" rispetto al modello "a stella"**
- Miglioramento nelle prestazioni delle interrogazioni attraverso **una minore occupazione di spazio** dovuta all'eliminazione di dati ridondanti
- l'utilizzo, in occasione dei "joins", di **piccole tabelle normalizzate invece di grosse tabelle non normalizzate**
- **Svantaggi del modello a fiocco di neve rispetto al modello "a stella"**
- Struttura più complessa
- Creazione e gestione di un maggior numero di tabelle
- Maggiore difficoltà nella gestione delle queries a causa di molteplici "joins"
- Maggiore difficoltà nel decidere quale tabella deve essere usata in una query

30/05/2014 9.37

A.GORI - 2011

33

Considerazioni sugli Schemi a Stella e a Fiocco di neve

- Lo schema a fiocco di neve è ottimizzato per occupare meno spazio. Ottimale in presenza di DW enormi
- Fattorizzando maggiormente, impiega più tempo per risolvere le join. Inoltre, impedisce l'utilizzo di indici bitmapped, a discapito delle prestazioni di ricerca.

30/05/2014 9.37

A.GORI - 2011

34

Constellation Schema

30/05/2014 9.37

A.GORI - 2011

35

Constellation Schema

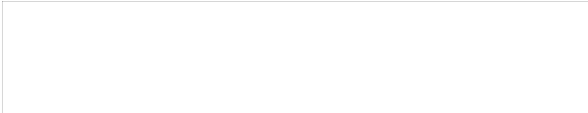
- Una parziale eliminazione delle inefficienze dello Star Schema si ottiene suddividendo le ennuple della FT a seconda del livello di aggregazione
- Si memorizzano in Fact Table differenti, i dati relativi a pattern di aggregazione diversi.
- Le Fact Table che corrispondono a pattern di aggregazione in cui una o più dimensioni sono completamente aggregate non presenteranno le relative chiavi esterne.
 - Fact Constellation: più Fact Table che condividono le Dimension Table;
sono viste come una collezione di "stelle", detta anche galaxy schema o constellation schema

30/05/2014 9.37

A.GORI - 2011

36

Schemi relazionali e viste



- ✓ L'accesso alle fact table è ottimizzato, quello alle dimension table no.
- ✓ La dimensione delle fact table è di molto superiore a quella delle dimension table e conseguentemente la loro ottimizzazione gioca un ruolo fondamentale.

Settimane		
ID_Settimane		
Settimana		
Mese		

Prodotti		
ID_Prodotti		
Prodotto		
Tipo		
Categoria		
Fornitore		

Negozi		
ID_Negozi		
Negozio		
Città		
Stato		
Rappresentante		

30/05/2014 9:37

37

Confronto tra Fact e Dimension Tables

Fact Tables:

- n° righe elevatissimo (10E6-10E9)
- n° di campi ridotto
- Dati principalmente numerici
- Molteplici chiavi esterne (FK) alle tabelle dimensionali
- Sono tabelle "strette e lunghe" (ridurre al max i campi)

Dimension Tables

- n° di righe contenuto (10E2-10-3)
- Numero di campi elevato, per gestire le gerarchie
- Dati testuali o temporali
- Sono tabelle "larghe e corte" (ridurre al max le righe)

30/05/2014 9:37

A.GORI - 2011

38

Star Schema

ID_Negozi	Negozio	Città	Stato	Rappresentante
1	N1	RM	I	R1
2	N2	RM	I	R1
3	N3	MI	I	R2
4	N4	MI	I	R2

Dimension Table

ID_Negozi	ID_Sett	ID_Prodotti	Quantità	Guadagno
1	1	1	100	100
1	2	1	150	150
3	3	4	350	350
4	4	4	200	200

Fact Table

ID_Sett.	Settimana	Mese
1	Gen1	Gen.
2	Gen2	Gen.
3	Feb1	Feb.
4	Feb2	Feb.

Dimension Table

ID_Prodotti	Prodotto	Tipo	Categoria	Fornitore
1	P1	A	X	F1
2	P2	A	X	F1
3	P3	B	X	F2
4	P4	B	X	F2

30/05/2014 9:37

A.GORI - 2011

39

Star Schema

Ennuple relative all'aggregazione per:
{Negozi,Settimane,Prodotti}

ID_Negozi	ID_Sett	ID_Prodotti	Quantità	Guadagno
1	1	1	100	100
1	2	1	150	150
3	3	4	350	350
4	4	4	200	200
2	1	5	3600	3600
1	3	6	2400	2400
1	1	7	1000	1000

ID_Prodotti	Prodotto	Tipo	Categoria	Fornitore
1	P1	A	X	F1
2	P2	A	X	F1
3	P3	B	X	F2
4	P4	B	X	F2
5	-	A	X	F1
6	-	B	X	F2
7	-	-	-	F1

30/05/2014 9:37

A.GORI - 2011

40

Star Schema

Ennuple relative all'aggregazione per:

{Negozio, Settimane, Tipo}

ID_Negozi	ID_Sett	ID_Prodotti	Quantità	Guadagno
1	1	1	100	100
1	2	1	150	150
3	3	4	350	350
4	4	4	200	200
2	1	5	3600	3600
1	3	6	2400	2400
1	1	7	1000	1000

ID_Prodotti	Prodotto	Tipo	Categoria	Fornitore
1	P1	A	X	F1
2	P2	A	X	F1
3	P3	B	X	F2
4	P4	B	X	F2
5	-	A	X	F1
6	-	B	X	F2
7	A.GORI-2011	-	-	F1

30/05/2014 9.37

41

Ennuple relative all'aggregazione per:

{Negozio, Settimane, Fornitore}

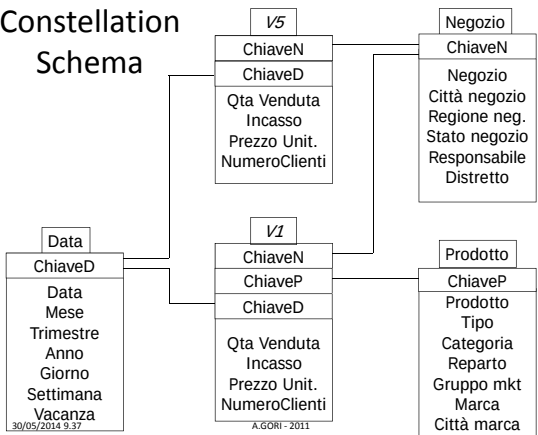
ID_Negozi	ID_Sett	ID_Prodotti	Quantità	Guadagno
1	1	1	100	100
1	2	1	150	150
3	3	4	350	350
4	4	4	200	200
2	1	5	3600	3600
1	3	6	2400	2400
1	1	7	1000	1000

ID_Prodotti	Prodotto	Tipo	Categoria	Fornitore
1	P1	A	X	F1
2	P2	A	X	F1
3	P3	B	X	F2
4	P4	B	X	F2
5	-	A	X	F1
6	-	B	X	F2
7	A.GORI-2011	-	-	F1

30/05/2014 9.37

42

Constellation Schema

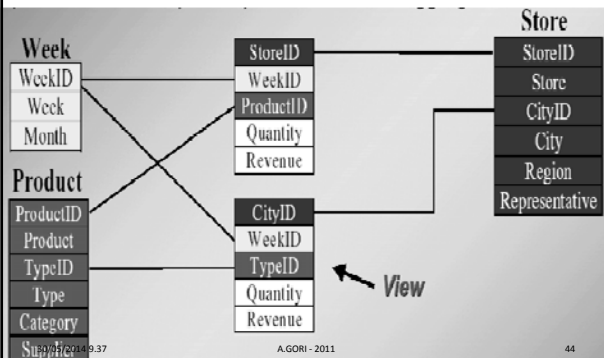


30/05/2014 9.37

A.GORI-2011

43

Constellation Schema



30/05/2014 9.37

A.GORI-2011

44