# Progetto DW 2020/2021

# Introduzione

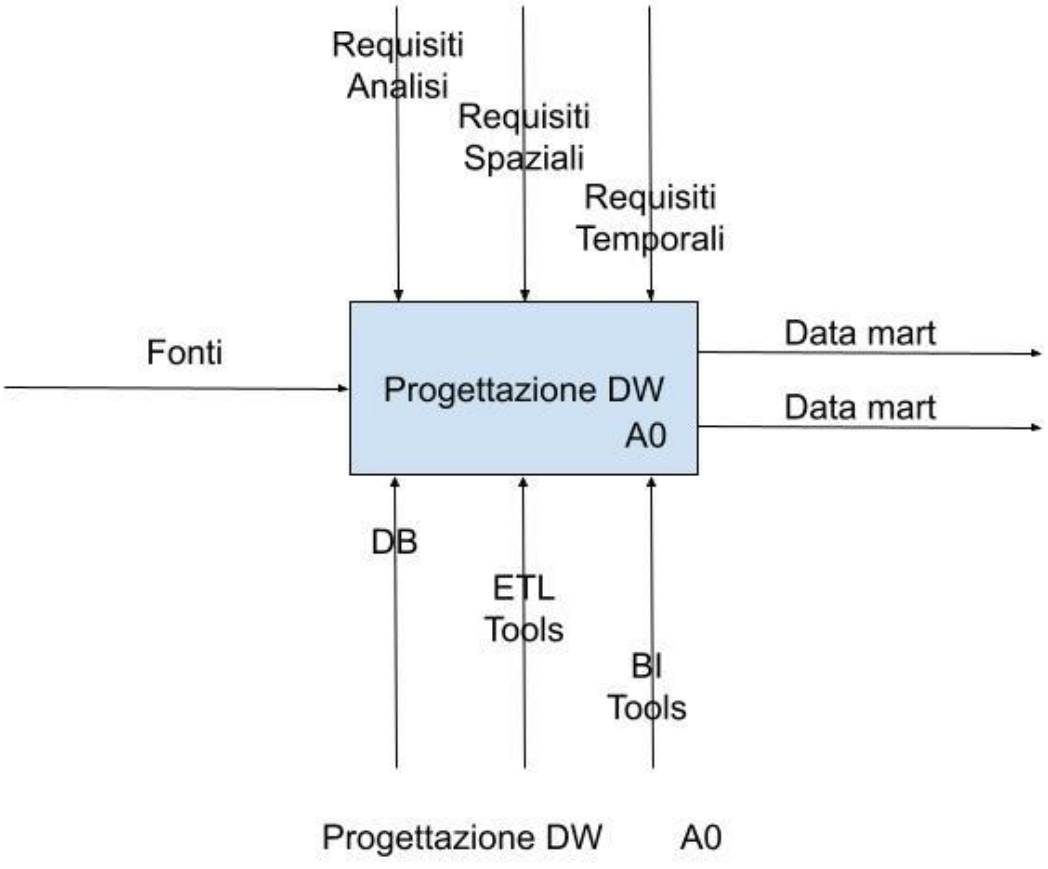
Il contesto di riferimento è quello manifatturiero tessile. Vengono considerate due fonti, un e-commerce online gestito da terzi ed un db operazionale (IBM DB2) di proprietà.
L'obiettivo di questo progetto di DW è quello di generare due data mart, uno per le giacenze di magazzino ed uno per la marginalità delle vendite.
Il server in cui installare il DW sarà di proprietà, il DB utilizzato sarà lo stesso del db operazionale (IBM DB2), gli strumenti per fare ETL saranno software custom scritti in java ed infine lo strumento di BI sarà IBM Cognos.

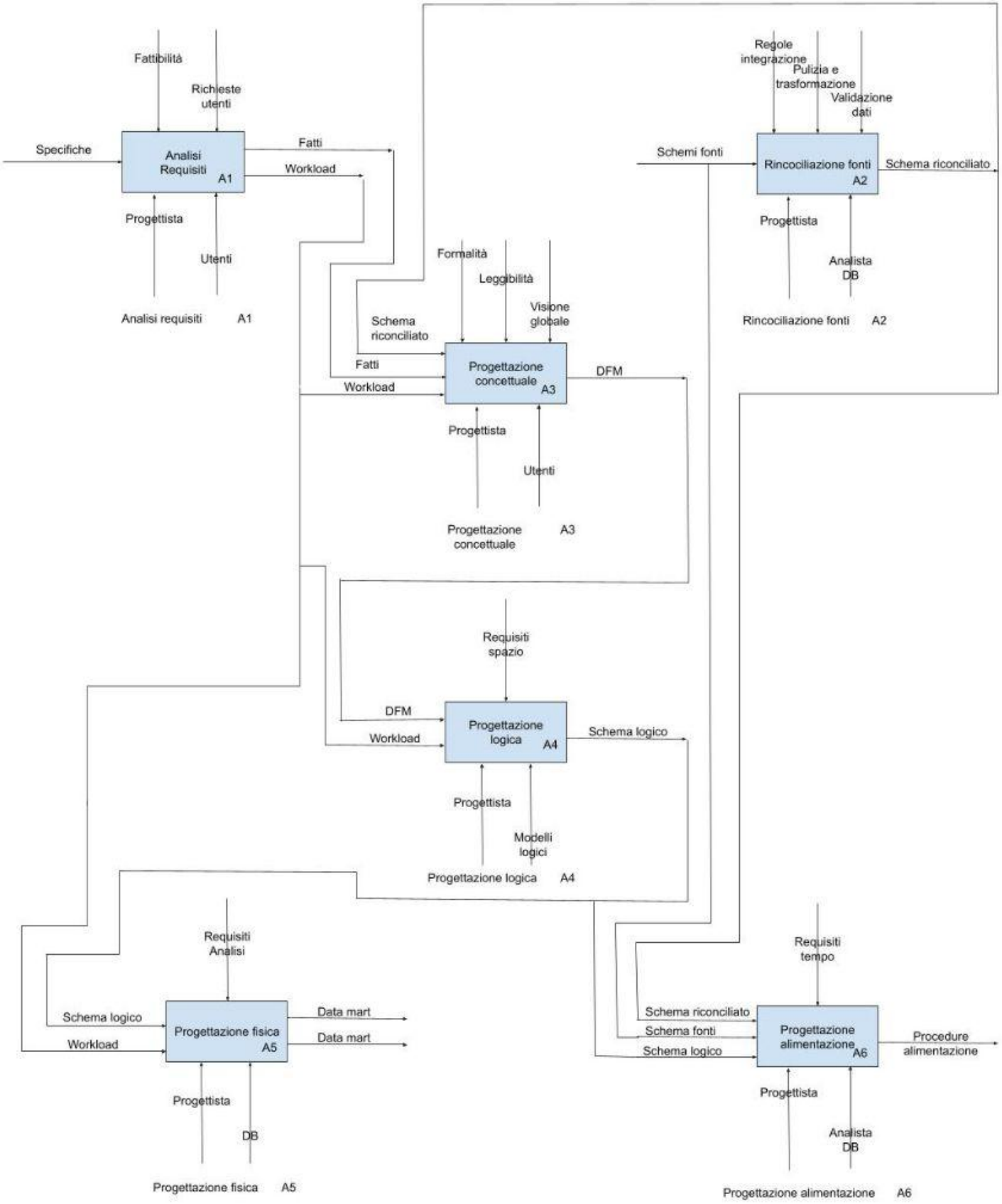
La produzione aziendale tratta materie prime, filati e subbi per produrre tessuti, sciarpe e foulard.
Il DB operazionale tratta la parte delle vendite B2B, mentre l'e-commerce tratta la parte B2C, che viene contabilizzata da un'azienda collegata.
Il DW dovrà unificare le vendite di entrambe le fonti.

Verranno analizzati i seguenti processi:

1) Analisi dei requisiti
2) Riconciliazione delle fonti
3) Progettazione concettuale
4) Progettazione logica
5) Progettazione fisica
6) Progettazione alimentazione

### Modellazione per processi con IDEF0




Fattibilità
Richieste utenti
Specifiche
Analisi Requisiti
A1
Fatti
Workload
Progettista
Utenti
Analisi requisiti A1
Regole integrazione
Pulizia e trasformazione
Validazione dati
Schemi fonti
Rincociliazione fonti
A2
Schema riconciliato
Progettista
Analista DB
Rincociliazione fonti A2
Formalità
Leggibilità
Visione globale
Schema riconciliato
Fatti
Workload
Progettazione concettuale
A3
DFM
Progettista
Utenti
Progettazione concettuale A3
Requisiti spazio
DFM
Workload
Progettazione logica
A4
Schema logico
Progettista
Modelli logici
Progettazione logica A4
Requisiti Analisi
Schema logico
Workload
Progettazione fisica
A5
Data mart
Data mart
Progettista
DB
Progettazione fisica A5
Requisiti tempo
Schema riconciliato
Schema fonti
Schema logico
Progettazione alimentazione
A6
Procedure alimentazione
Progettista
Analista DB
Progettazione alimentazione A6

## Analisi dei requisiti

In questa fase si instaura un colloquio con gli utenti utilizzatori allo scopo di raccogliere e definire le esigenze e gli utilizzi che si intende fare con i data mart.

### Data mart giacenze

Con questo data mart si vuole analizzare:

- quantità in magazzino ad una certa data;
- valore di magazzino ad una certa data;
- giorni di fermo di ciascun lotto ad una certa data;
- scostamento tra valore effettivo di produzione e valore stimato ad una certa data.

Il fatto che si intende analizzare è quello delle giacenze di magazzino.
Le dimensioni di analisi sono quelle relative ai lotti di produzione e alla data.
La granularità dei dati riportati sarà a livello di lotto di produzione e giorno di fotografia, in modo da poter analizzare nel massimo dettaglio le anomalie dei costi di produzione.
La storicizzazione dei dati sarà giornaliera; i dati relativi alle estrazioni più vecchie di 30 giorni verranno eliminati, mantenendo solamente i dati relativi alle date di fine mese.

L'esposizione dei dati richiesta è un elenco tabellare navigabile sulle gerarchie di analisi, che rappresenti una fotografia per ciascun giorno dei lotti di produzione presenti in magazzino.

### Data mart vendite

Con questo data mart si vuole analizzare:

- clienti e relativo fatturato in un certo periodo;
- agenti e relativo fatturato in un certo periodo;
- articoli più redditizi, in termini di scostamento tra costo di produzione e prezzo di vendita, in un certo periodo;
- andamento delle vendite mensile e trimestrale;
- confronto delle vendite con gli anni precedenti.

Il fatto che si intende analizzare è quello delle vendite.
Le dimensioni di analisi sono quelle relative ai clienti, agli agenti, ai lotti di produzione e alla data.
La granularità dei dati riportati sarà a livello di lotto di produzione venduto e documento di vendita, in modo da poter analizzare nel massimo dettaglio la marginalità di ciascun lotto.
La storicizzazione dei dati sarà giornaliera, senza nessun tipo di archiviazione.
L'esposizione dei dati richiesta è sia un elenco tabellare navigabile sulle gerarchie di analisi, che una serie di infografiche relative agli andamenti e ai confronti annuali.

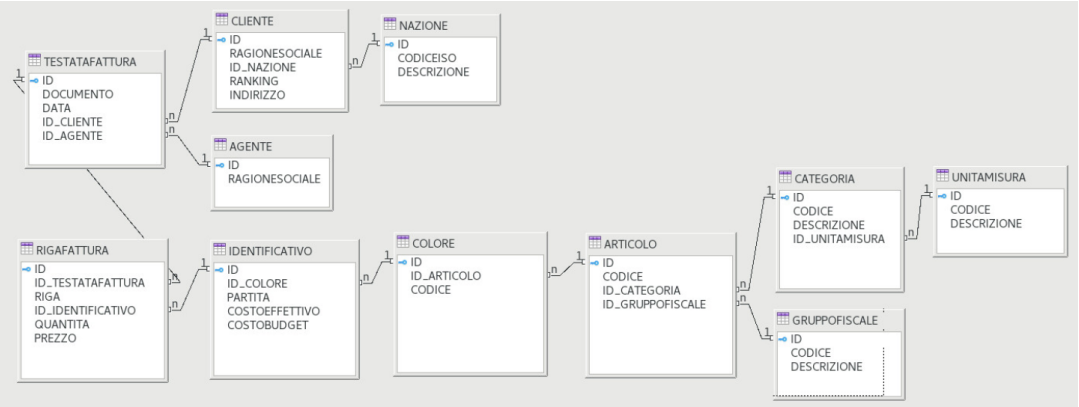
# Riconciliazione delle fonti

## Fonti giacenze

Per quanto riguarda il primo data mart, quello delle giacenze, verrà utilizzata un'unica fonte, quella del db operazionale.
Ecco lo schema del database


## Fonti vendite

Per quanto riguarda il secondo data mart, quello delle vendite verranno utilizzate due fonti: la prima è quella del db operazionale, la seconda un web service generato da parte del fornitore dell'e-commerce che, ricevendo una data come parametro, restituisce un file xml con le vendite fatte a partire da quella data.
Ecco lo schema del database




Ecco un esempio di xml

```
<?xml version="1.0" encoding="UTF-8"?>
<ecommerce>
	<invoice>
		<customer>
			<name>Mario Bianchi</name>
			<address>Via Rossi, Prato</address>
			<nation>ITA</nation>
		</customer>
		<date>2021-12-20</date>
		<document>FR 100</document>
		<item>
			<article>ART_100</article>
			<color>COL_100</color>
			<lot>PTA 1</lot>
			<quantity>10.0</quantity>
			<price>1.0</price>
```

```
            </item>
            <item>
                  <article>ART_200</article>
                  <color>COL_200</color>
                  <lot>PTA 200</lot>
                  <quantity>5.0</quantity>
                  <price>10.0</price>
            </item>
      </invoice>
      <invoice>
            ...
      </invoice>
</ecommerce>
```

I dati del db operazionale sono da considerarsi corretti e consistenti, in quanto il db e l'applicazione sono manutenuti dalla stessa figura che sviluppa il DW.
Per quanto riguarda l'e-commerce i prodotti ed i lotti venduti provengono da una precedente integrazione tra sistemi e sono quindi da considerarsi corretti; per quanto riguarda i clienti ci si affida al sistema sottostante all'e-commerce.
Non si hanno casi di omonimia, in quanto il file xml contiene i tag in lingua inglese; si hanno però casi di sinonimia e sarà utilizzato il seguente vocabolario di traduzione:

- il tag `<invoice>` corrisponde ad un record della tabella TESTATAFATTURA
    - il tag `<date>` corrisponde al campo DATA della tabella TESTATAFATTURA
    - il tag `<document>` corrisponde al campo DOCUMENTO della tabella TESTATAFATTURA
- il tag `<customer>` corrisponde ad un record della tabella CLIENTE
    - il tag `<name>` corrisponde al campo RAGIONESOCIALE della tabella CLIENTE
    - il tag `<address>` corrisponde al campo INDIRIZZO della tabella CLIENTE
    - il tag `<nation>` corrisponde ad un record della tabella NAZIONE
- il tag `<item>` corrisponde ad un record della tabella RIGAFATTURA
    - il tag `<article>` corrisponde ad un record della tabella ARTICOLO
    - il tag `<color>` corrisponde ad un record della tabella COLORE
    - il tag `<lot>` corrisponde ad un record della tabella IDENTIFICATIVO
    - il tag `<quantity>` corrisponde al campo QUANTITA della tabella RIGAFATTURA
    - il tag `<price>` corrisponde al campo PREZZO della tabella RIGAFATTURA

Non si presentano conflitti tra i due schemi logici.
Con questa traduzione siamo riusciti a riconciliare i due schemi, allineandoli al db operazionale, da considerare come schema riconciliato.

# Progettazione concettuale

A fronte dell'analisi dei requisiti sono emersi due fatti da analizzare: il fatto relativo alle giacenze ed il fatto relativo alle vendite.

## DFM giacenze

Basandosi sul carico di lavoro valutato in fase di analisi, le dimensioni di analisi sono quelle relative ai lotti di produzione e alla data.
Le gerarchie di interesse sono:

- per quanto riguarda i lotti di produzione si considerano i relativi articoli, le categorie ed i gruppi fiscali;
- per quanto riguarda la data non sono necessarie particolari gerarchie.

Le misure necessarie sono:

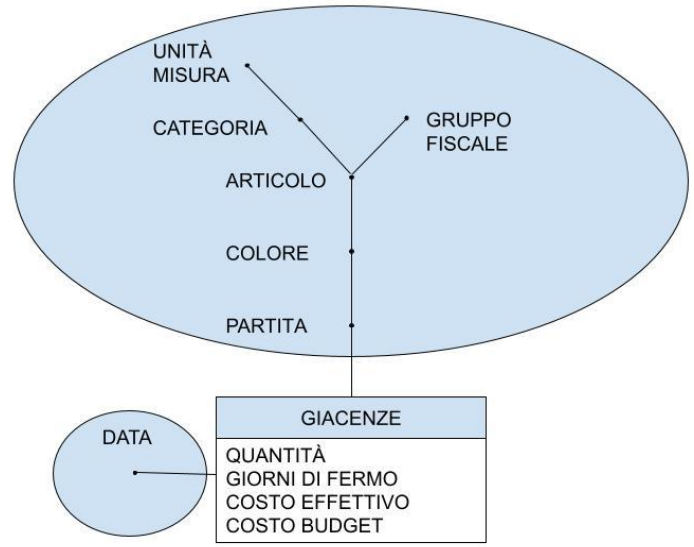
- la quantità di giacenza, misura di livello;
- il costo di produzione, misura unitaria;
- il costo stimato, misura unitaria;
- i giorni passati dall'ultima movimentazione, misura unitaria.

Per la dimensione temporale, nessuna delle misure elencate è additiva; per la gerarchia dei lotti di produzione invece la quantità può essere sommata (semi additiva), mentre per le altre misure unitarie, si può calcolare la media o i picchi di minimo o di massimo.
Per rispettare le richieste analizzate è necessario introdurre delle misure derivate:

- valore, ottenuto come prodotto tra quantità e costo effettivo; misura che risulta semi additiva sulla gerarchia dei lotti di produzione;
- valore di scostamento tra valore effettivo e valore stimato, ottenuto come prodotto tra la quantità e la differenza tra costo effettivo e il costo stimato; misura che risulta semi additiva sulla gerarchia dei lotti di produzione.

Viene anche introdotta una misura di supporto che conteggia gli elementi aggregati, in modo da poter calcolare le medie dei costi.

In base a queste esigenze è stato elaborato il seguente dimensional fact model.



**DFM vendite**

Basandosi sul carico di lavoro valutato in fase di analisi, le dimensioni di analisi sono quelle relative ai clienti, agli agenti, ai lotti di produzione e alla data; si utilizza anche l'attributo descrittivo relativo al documento.
La dimensione agenti è opzionale.
Le gerarchie di interesse sono:

- per quanto riguarda i clienti si considerano la nazione ed il ranking aziendale;
- per quanto riguarda gli agenti non sono necessarie particolari gerarchie;
- per quanto riguarda i lotti di produzione si considerano i relativi articoli e le categorie;
- per quanto riguarda la data si considera il mese, il trimestre e l'anno.

Le misure necessarie sono:

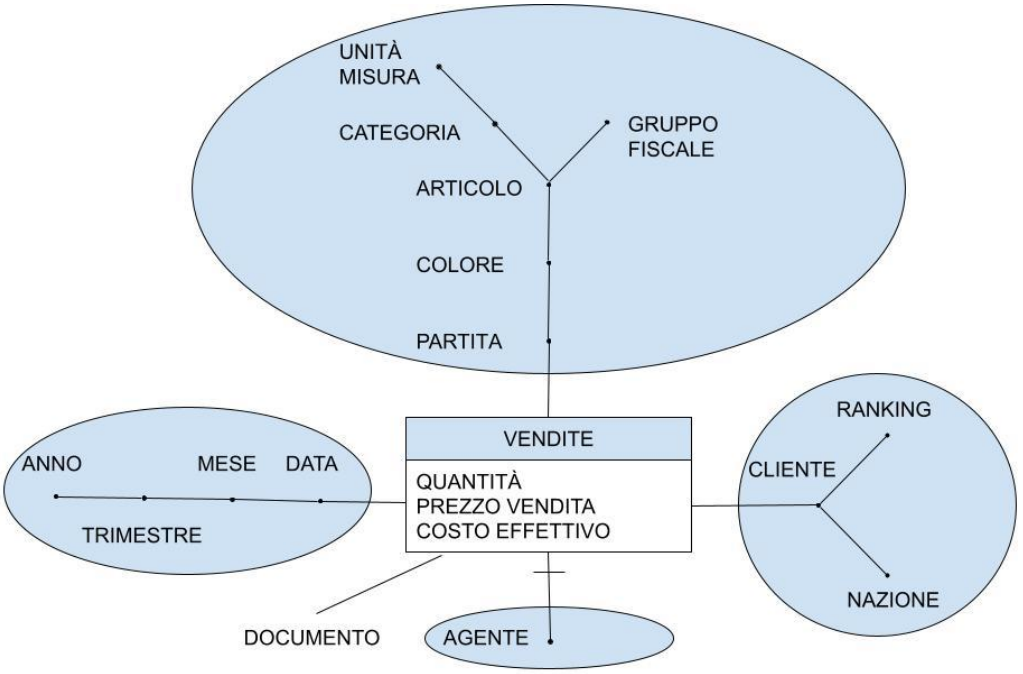
- la quantità venduta, misura di flusso;
- il prezzo di vendita, misura unitaria;
- il costo di produzione del lotto, misura unitaria.

La misura quantità è additiva su tutte le dimensioni, mentre per le altre misure unitarie, si può calcolare la media o i picchi di minimo o di massimo.
Per rispettare le richieste analizzate è necessario introdurre delle misure derivate:

- valore, ottenuto come prodotto tra quantità e prezzo di vendita; misura che risulta additiva su tutte le dimensioni;
- margine di profitto, ottenuto come prodotto tra la quantità e la differenza tra prezzo di vendita e costo effettivo; misura che risulta additiva su tutte le dimensioni.

Viene anche introdotta una misura di supporto che conteggia gli elementi aggregati, in modo da poter calcolare le medie del prezzo e del costo.

In base a queste esigenze è stato elaborato il seguente dimensional fact model.



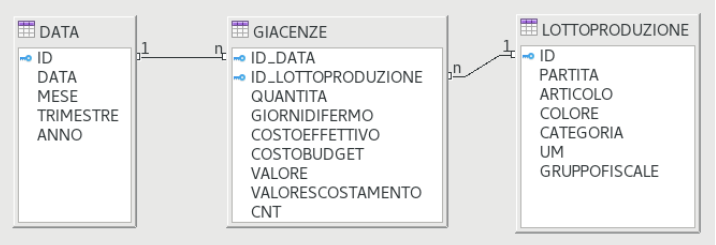

Entrambi i modelli concettuali sono stati sviluppati basandosi sul carico di lavoro, quindi sono in grado di rispondere alle interrogazioni previste in fase di analisi.

## Progettazione logica

Il modello logico scelto è quello ROLAP, sul quale verrà implementato uno star schema.
Le gerarchie verranno appiattite sulle tabelle delle dimensioni, applicando una forte denormalizzazione.
In entrambi i data mart per le dimension table sono state utilizzate chiavi surrogate, in modo da poter agevolare i piani di accesso e indicizzazione del DBMS; in fase di alimentazione si dovrà tenere conto di questa scelta.
In entrambi i data mart è presente un'unica fact table, che contiene le varie misure che caratterizzano il fatto; sono presenti varie dimension table, una per dimensione, dove ciascuna tabella contiene gli attributi che caratterizzano l'intera gerarchia della dimensione.
Tutte le dimension table hanno come chiave primaria una chiave surrogata, denominata ID.
La fact table contiene vari attributi che referenziano le dimension table.
L'accesso ai dati e la visione multidimensionale dei dati avviene tramite join tra la fact table e le dimension table.
Nelle dimension table saranno presenti dei record che rappresenteranno le aggregazioni date dalle gerarchie dimensionali, con solo alcuni campi valorizzati (p.e. nella tabella LOTTOPRODUZIONE saranno presenti dei record aggiuntivi con il solo campo UM valorizzato, in modo da poter aggregare per la sola unità di misura).

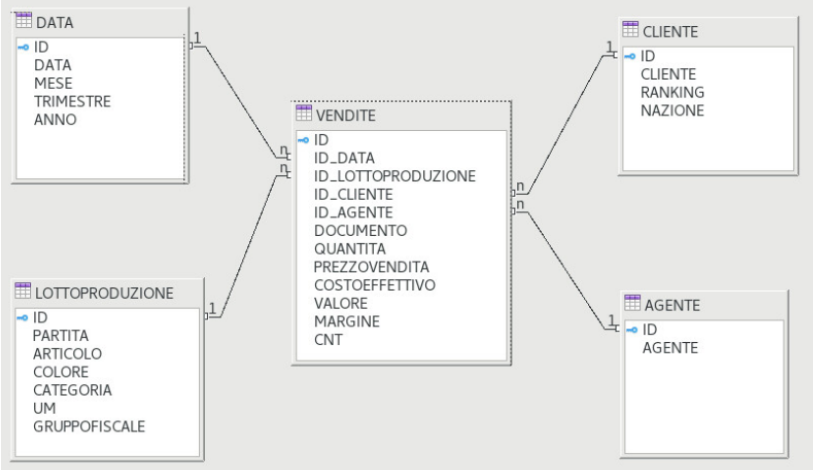
### Giacenze

In questo data mart la fact table è GIACENZE, mentre le dimension table sono DATA e LOTTOPRODUZIONE.
La chiave primaria della fact table è costituita dagli attributi che individuano le dimensioni.




### Vendite

In questo data mart la fact table è VENDITE, mentre le dimension table sono DATA, LOTTOPRODUZIONE, CLIENTE e AGENTE.
La chiave primaria della fact table è una chiave surrogata, perché le dimensioni non individuano univocamente una vendita, in quanto la stessa partita potrebbe essere venduta con più documenti.
La relazione con la dimension table AGENTE è opzionale, come analizzato in fase di progettazione concettuale.
Su questa fact table è presente anche l'attributo descrittivo DOCUMENTO.



# Progettazione fisica

Per quanto riguarda la progettazione fisica verrà utilizzato un unico DB e un unico schema per gestire i due data mart, in modo da cercare di ottimizzare lo spazio su disco necessario a gestire l'intero DW, in quanto alcune gerarchie dimensionali (data e lotto produzione) sono comuni ai due data mart.
Per cercare di diminuire al massimo i tempi di risposta per l'utente, verranno precalcolate delle viste materializzate e generati una serie di indici appropriati.

## Viste precalcolate

### Giacenze

La vista primaria è data da:

- data, partita;

Le viste secondarie necessarie a soddisfare il carico di lavoro sono:

- data, um;
- data, categoria;
- data, gruppo fiscale;

La cardinalità delle aggregazioni delle viste secondarie è piuttosto limitata:

- le unità di misura sono KG, MT, NR;
- le categorie (che indicano il tipo di articolo) sono materie prime, filati e subbi per produrre tessuti, sciarpe e foulard;
- i gruppi fiscali rappresentano delle aggregazioni di articoli fiscalmente omogenei.

Da questa osservazione si decide di materializzare tutte le viste secondarie.

### Vendite

La vista primaria è data da:

- data, partita;

Le viste secondarie necessarie a soddisfare il carico di lavoro sono:

- data, cliente;
- data, agente;
- data, articolo;
- mese, anno;
- trimestre, anno;

La cardinalità dei clienti, degli agenti e degli articoli può crescere rapidamente, quindi le viste secondarie che riguardano questi attributi dimensionali non verranno materializzate, ma calcolate partendo dalla vista primaria facendo aggregazione.
Al contrario le ultime due (mese, anno e trimestre, anno) verranno materializzate, dato che la cardinalità è piuttosto limitata.

## Indici

La fase iniziale di progettazione degli indici è puramente teorica. Una volta che il DW sarà reso operativo è necessaria una fase di messa a punto e ottimizzazione degli indici basata sull'effettivo carico di lavoro.

### Giacenze

Sulla tabella dei fatti, GIACENZE, è già presente la chiave primaria composta con i due attributi ID_DATA e ID_LOTTOPRODUZIONE, che rappresentano le chiavi surrogate delle due dimensioni; questa chiave primaria permette anche di ottimizzare l'esecuzione delle join sulle tabelle delle dimensioni; la scelta dell'indice composto con ID_DATA in prima posizione è stata fatta in quanto la data sarà presente in tutte le interrogazioni.
Sulla tabella della dimensione DATA, verrà indicizzato l'attributo DATA.
Sulla tabella della dimensione LOTTOPRODUZIONE invece verranno creati due indici distinti sugli attributi ARTICOLO e PARTITA, in quanto le interrogazioni più frequenti li riguarderanno.
Inoltre sempre sulla tabella dei fatti saranno indicizzate in maniera distinta anche le misure GIORNIDIFERMO e VALORESCOSTAMENTO, in quanto sono probabili condizioni che le riguarderanno.

### Vendite

Sulla tabella dei fatti, VENDITE, oltre alla chiave primaria surrogata ID, saranno quattro indici sugli attributi ID_DATA, ID_LOTTOPRODUZIONE, ID_CLIENTE e ID_AGENTE, che rappresentano le chiavi surrogate delle quattro dimensioni; questa serie di chiavi permettono di ottimizzare l'esecuzione delle join sulle tabelle delle dimensioni.
Sulla tabella della dimensione DATA, verranno generati tre indici, il primo sull'attributo DATA, dal momento in cui il filtro del periodo riguarda praticamente tutte le interrogazioni; il secondo sarà un indice composto sugli attributi ANNO e TRIMESTRE ed il terzo sugli attributi ANNO e MESE; anche in questo caso l'ordinamento degli indici composti è pensato in base ai filtri che verranno richiesti.
Sulla tabella della dimensione LOTTOPRODUZIONE invece verranno creati due indici distinti sugli attributi ARTICOLO e PARTITA, in quanto le interrogazioni più frequenti li riguarderanno.
Sulla tabella della dimensione CLIENTE verranno creati tre indici, il primo sull'attributo CLIENTE, in quanto è molto probabile che un filtro in tal senso venga spesso richiesto, il secondo sulla coppia di attributi RANKING e CLIENTE ed il terzo sulla coppia NAZIONE e CLIENTE, sempre per agevolare le risposte ai filtri.
Sulla tabella della dimensione AGENTE è presente un unico attributo, AGENTE, che verrà indicizzato.
Inoltre sempre sulla tabella dei fatti saranno indicizzate in maniera distinta anche le misure VALORE e MARGINE, in quanto sono probabili condizioni che le riguarderanno.

## Progettazione alimentazione

Per entrambi i data mart verrà eseguita un'elaborazione notturna, appena terminati i backup del DB.
Per primo verrà elaborato il data mart delle giacenze, poi quello delle vendite.
In entrambi i casi verrà eseguita un'estrazione basata su marche temporali.

### Giacenze

Il data mart si basa solamente sul DB operazionale, in particolare sulla movimentazione del magazzino, quindi tutti i dati sono storicizzati.
Ogni estrazione avviene in modalità statica, in quanto viene fatta una fotografia ai saldi dei quantitativi dei lotti presenti in magazzino alla data di estrazione; i saldi sono ottenuti eseguendo la seguente query sul DB operazionale:

```
SELECT
	MM.DATA,
	IDE.PARTITA,
	A.ARTICOLO,
	C.COLORE,
	CAT.CODICE,
	UM.CODICE,
	GF.CODICE,
	MM.QUANTITA,
	IDE.COSTOEFFETTIVO,
	IDE.COSTOBUDGET,
	MM.QUANTITA * IDE.COSTOEFFETTIVO AS VALORE,
	MM.QUANTITA * (IDE.COSTOEFFETTIVO - IDE.COSTOBUDGET) AS VALORESCOSTAMENTO
FROM
	DBO.MOVIMENTOMAGAZZINO MM
	JOIN DBO.IDENTIFICATIVO IDE ON MM.ID_IDENTIFICATIVO = IDE.ID
	JOIN DBO.COLORE C ON IDE.ID_COLORE = C.ID
	JOIN DBO.ARTICOLO A ON C.ID_ARTICOLO = A.ID
	JOIN DBO.GRUPPOFISCALE GF ON A.ID_GRUPPOFISCALE = GF.ID
	JOIN DBO.CATEGORIA CAT ON A.ID_CATEGORIA = CAT.ID
	JOIN DBO.UNITAMISURA UM ON CAT.ID_UNITAMISURA = UM.ID
WHERE
	MM.DATA = :DATA
```

Partendo dai risultati della query devono essere fatte le seguenti operazioni:

- valorizzazione le tabelle dimensionali DATA e LOTTOPRODUZIONE con i valori mancanti;
- valorizzazione della tabella dei fatti GIACENZE, recuperando le chiavi esterne delle tabelle dimensionali;
- precalcolo delle viste, aggregando i dati secondo le modalità definite in precedenza.

## Vendite

In questo caso invece il data mart si basa su due fonti: il db operazionale ed il web service; in entrambi i casi si può supporre che i dati siano storicizzati.
La prima estrazione avviene in modalità statica, sia per il db operazionale che per il web service, ed in entrambi i casi si andrà a richiedere le fatture con data del documento maggiore o uguale al 01/01/1970.
Le successive estrazioni avverranno in modalità incrementale, andando a richiedere le fatture con data del documento uguale al giorno precedente, supponendo che i backup terminino dopo mezzanotte.
Per primo viene esaminato il db operazionale eseguendo la seguente query:

```
SELECT
	TF.DATA,
	TF.NUMERO || '/' || RF.RIGA AS DOCUMENTO,
	IDE.PARTITA,
	A.ARTICOLO,
	C.COLORE,
	CAT.CODICE,
	UM.CODICE,
	GF.CODICE,
	CL.RAGIONESOCIALE,
	CL.RANKING,
	NAZ.CODICEISO,
	AGENTE.RAGIONESOCIALE,
	RF.QUANTITA,
	RF.PREZZO,
	IDE.COSTOEFFETTIVO,
	RF.QUANTITA * RF.PREZZO AS VALORE,
	RF.QUANTITA * (RF.PREZZO - IDE.COSTOEFFETTIVO) AS MARGINE
FROM
	DBO.TESTATAFATTURA TF
	JOIN DBO.AGENTE AG ON TF.ID_AGENTE = AG.ID
	JOIN DBO.CLIENTE CL ON TF.ID_CLIENTE = CL.ID
	JOIN DBO.NAZIONE NAZ ON CL.ID_NAZIONeE = NAZ.ID
	JOIN DBO.RIGAFATTURA RF ON TF.ID = RF.ID_TESTATAFATTURA
	JOIN DBO.IDENTIFICATIVO IDE ON RF.ID_IDENTIFICATIVO = IDE.ID
	JOIN DBO.COLORE C ON IDE.ID_COLORE = C.ID
	JOIN DBO.ARTICOLO A ON C.ID_ARTICOLO = A.ID
	JOIN DBO.GRUPPOFISCALE GF ON A.ID_GRUPPOFISCALE = GF.ID
	JOIN DBO.CATEGORIA CAT ON A.ID_CATEGORIA = CAT.ID
	JOIN DBO.UNITAMISURA UM ON CAT.ID_UNITAMISURA = UM.ID
WHERE
	TF.DATA >= :DATA
```

In seconda battuta viene interrogato l'url del web service, il quale risponderà con un file xml nel formato definito inizialmente.
L'url richiamato sarà qualcosa del tipo *https://ecommerce.org/invoice?dateStart=:**DATA***; il file xml prodotto conterrà l'elenco delle fatture generate a partire dalla data richiesta.

Partendo dai risultati della query e del web service devono essere fatte le seguenti operazioni:

- valorizzazione le tabelle dimensionali DATA, LOTTOPRODUZIONE, AGENTE (solo per il db operazionale) e CLIENTE con i valori mancanti;
- valorizzazione della tabella dei fatti VENDITE, recuperando le chiavi esterne delle tabelle dimensionali;
- precalcolo delle viste, aggregando i dati secondo le modalità definite in precedenza.

## Presentazione dati