Analisi dell'utilizzo di biciclette nella città di Bologna in relazione alle condizioni meteo.

Emma Villa 885906

Contents

1 Introduzione **2**
1.1 Contesto della mobilità urbana sostenibile . . . 2
1.2 Obiettivo del progetto . . . 2

2 Analisi e valutazione di qualità delle sorgenti **2**
2.1 Analisi delle sorgenti . . . 2
2.2 Valutazione della qualità delle sorgenti . . . 3
2.3 Scelta dell'approccio di integrazione: ELT . . . 3

3 Progettazione dello schema riconciliato **4**
3.1 Comparazione degli schemi . . . 4
3.2 Schema Concettuale Riconciliato . . . 5
3.3 Schema Logico Riconciliato (ODS) . . . 5

4 Analisi dei Requisiti e Carico di Lavoro **6**
4.1 Glossario dei Requisiti . . . 6
4.2 Stima dei Volumi di Dati . . . 6
4.3 Carico di Lavoro Preliminare . . . 6

5 Progettazione Concettuale e Logica **7**
5.1 Schema Concettuale (DFM) . . . 7
5.2 Schema Logico . . . 8
5.3 Gestione delle Dimensioni Dinamiche (SCD) . . . 8

6 Processo ELT **9**
6.1 Architettura Logica del Data Warehouse . . . 9
6.2 Fase 1: Estrazione e Caricamento . . . 9
6.3 Fase 2: Livello ODS . . . 9
6.4 Fase 3: Data Mart . . . 10

7 Dashboard **11**
7.1 Trend Temporali . . . 11
7.2 Impatto Meteorologico . . . 12
7.3 Analisi Oraria e Direzionale . . . 12
7.4 Distribuzione territoriale . . . 13
7.5 Flussi per quartiere . . . 13

8 Conclusioni **14**
8.1 Sintesi dei Risultati . . . 14

1 Introduzione

1.1 Contesto della mobilità urbana sostenibile

Negli ultimi anni, la città di Bologna ha investito in modo significativo nella mobilità sostenibile, promuovendo l'utilizzo della bicicletta come alternativa di trasporto ecologico. Questo cambiamento è supportato da una rete di piste ciclabili in espansione e da una serie di politiche urbane che hanno l'obiettivo di ridurre le emissioni di CO_2.

A differenza di molti servizi di sharing che monitorano il flusso ciclistico attraverso GPS mobili installati sulle biciclette, l'amministrazione comunale di Bologna monitora il traffico attraverso una rete di sensori fissi, delle colonnine conta-bici, posizionate in vari punti della città.

Tuttavia, il semplice conteggio dei passaggi non è sufficiente per comprendere ed analizzare i flussi ciclistici. Per effettuare un'analisi significativa bisogna tenere in considerazione variabili esterne che incidono sul comportamento delle persone. In particolare si vuole prendere in considerazione:

1. Fattore spaziale: La distribuzione dei flussi tra il centro storico e la periferia.
2. Fattore meteorologico: L'influenza di fattori meteorologici come la pioggia, la neve o le temperature estreme sulla scelta di muoversi in bicicletta.
3. Fattore temporale: La variabilità dei flussi in base all'ora del giorno ed in base a se il giorno è feriale oppure festivo.

1.2 Obiettivo del progetto

Lo scopo del progetto è la progettazione e la realizzazione di un Data Warehouse che integri i dati degli ultimi 5 anni dei passaggi ciclistici rilevati dalle colonnine, le condizioni meteorologiche orarie e la suddivisione dei quartieri di Bologna.

L'obiettivo finale è quello di fornire un sistema di supporto alle decisioni basato sui dati. In particolare l'analisi dei flussi potrà contribuire ad orientare le scelte dell'amministrazione comunale di Bologna riguardo:

- l'individuazione delle aree in cui risulta più sensato realizzare nuove piste ciclabili.
- la valutazione dell'opportunità di introdurre o potenziare un servizio di bike sharing.
- la gestione di un eventuale sistema di bike sharing, come la scelta di dove posizionare le rastrelliere oppure nella ridistribuzione delle biciclette sul territorio.

2 Analisi e valutazione di qualità delle sorgenti

2.1 Analisi delle sorgenti

In questa fase vengono analizzate le tre sorgenti dati che alimenteranno il Data Warehouse.

2.1.1 Sorgente 1: Flussi delle bici

- Formato: CSV con separatore TAB
- Volume: circa 350.000 record

Questa sorgente è resa disponibile dal portale Open Data del comune di Bologna tramite API. Il dataset non è statico ma viene aggiornato periodicamente e la risposta del servizio contiene le rilevazioni storiche dei sensori.

2.1.2 Sorgente 2: Dati Meteorologici

I dati meteorologici sono acquisiti tramite Open-Meteo Historical API. Questo è un servizio che richiede il passaggio di parametri specifici come latitudine, longitudine e le variabili richieste.

- Formato: CSV con separatore la virgola
- Volume: circa 44.000 record

Nome Campo	Tipo dato	Note
data	String	Include offset (es: 2024-03-04T12:00:00+01:00)
direzione_centro	Integer	Passaggi verso il centro
direzione_periferia	Integer	Passaggi verso la periferia
totale	Integer	Somma calcolata dei passaggi
colonnina	String	Nome della stazione (es: "Sabotino")
geo_point_2d	String	Coordinate separate da virgola (es: 44.49, 11.32)

Table 1: Schema grezzo sorgente "Flussi"

Nome Campo	Tipo dato	Note
time	Timestamp	YYYY-MM-DDTHH:MM
temperature_2m	Double	Temperatura reale
apparent_temperature	Double	Temperatura percepita
weather_code	Integer	Codice numerico della condizione meteorologica

Table 2: Schema grezzo sorgente "Meteo"

2.1.3 Sorgente 3: Divisione in quartieri

I dati geospaziali relativi ai confini dei quartieri a Bologna sono recuperati tramite il portale Open Data del comune, permettendo di contestualizzare le colonnine all'interno dei vari quartieri.

- Formato: JSON
- Volume: 6 record

Nome Campo	Tipo dato	Note
geo_point_2d	JSON Object	Coordinate del centro del quartiere
geo_shape	JSON Object	Struttura GeoJSON contenente le coordinate del poligono del quartiere
cod_quar	String	ID univoco del quartiere
quartiere	String	Nome del quartiere (es: "Naville")
area	Null	Dato presente ma privo di valore (null)

Table 3: Schema grezzo sorgente "Quartieri"

Per quanto riguarda il parametro area, che dovrebbe rappresentare la superficie di ogni quartiere, questa verrà calcolata geometricamente in fase di ELT.

2.2 Valutazione della qualità delle sorgenti

Prima di integrare le diverse fonti di dati, è importante analizzare ogni singola sorgente.

1. Flussi delle bici: il campo `data` include l'offset del fuso orario, questo potrebbe portare a incoerenze nell'analisi temporale o nell'aggregazione per fasce orarie. Bisognerebbe separare l'informazione della data da quella del fuso orario, che non è necessaria nel nostro caso di studio.
2. Dati Meteorologici: il campo `weather_code` utilizza una codifica numerica standard che è corretta tecnicamente ma non è facilmente interpretabile per fini analitici.
3. Quartieri: il campo `area` sebbene sia presente nello schema, contiene il valore `null` per la totalità dei record. Rendendo necessario derivare il valore in base al perimetro del quartiere.

2.3 Scelta dell'approccio di integrazione: ELT

Date le dimensioni dei dataset e la necessità di operare trasformazioni complesse (parsing JSON, calcoli geometrici), si è optato per un approccio ELT (Extract, Load, Transform).

Questa tecnica prevede di caricare i dati as-is nello storage di destinazione, delegando al motore del database

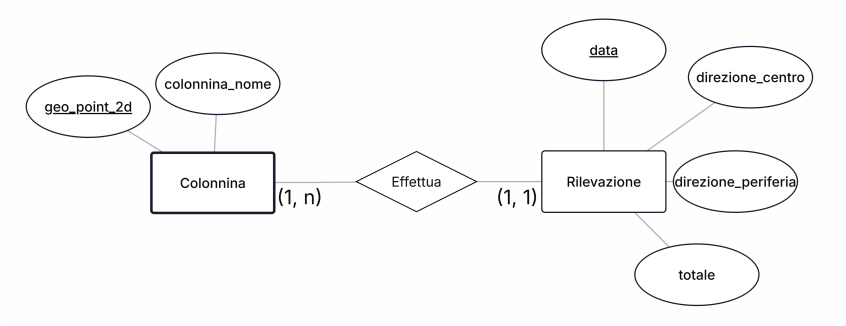
il carico computazionale delle trasformazioni. Ciò permette di mantenere una copia fedele dei dati grezzi (Staging Area).

La selezione degli strumenti utilizzati è stata guidata dai formati dei dati rilevati nella fase di analisi:

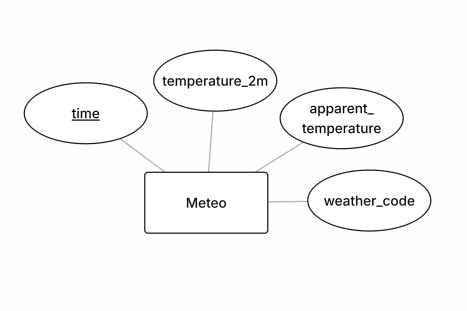
- Gestione dati Spaziali (DuckDB): Dato che sia la sorgente "Quartieri" che la sorgente "Flussi" hanno dei dati relativi a coordinate geografiche, è necessario utilizzare un DBMS con estensioni spaziali native, come DuckDB che ha l'estensione `spatial`, in grado di gestire formati GeoJSON e supportare operazioni geometriche come i join spaziali.
- Gestione delle Trasformazioni (DBT): Per gestire la logica di riconciliazione, si è scelto lo strumento DBT (Data Build Tool). Permette di definire le regole di trasformazione in SQL, garantendo la tracciabilità del dato (data lineage) e la gestione delle dipendenze tra modelli.

3 Progettazione dello schema riconciliato

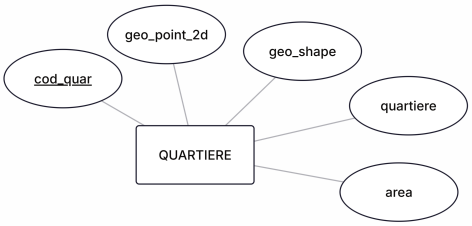
In questa fase del progetto si vuole passare dalle singole sorgenti ad uno schema globale riconciliato. Partendo dai dati disponibili bisogna derivare i modelli E/R delle sorgenti, per poi arrivare al modello concettuale riconciliato e poi al modello logico dello schema riconciliato.



(a) Schema Locale Flussi (Sorgente 1)



(b) Schema Locale Meteo (Sorgente 2)



(c) Schema Locale Quartieri (Sorgente 3)

Figure 1: Modelli E/R delle sorgenti locali prima dell'integrazione

Come si evince dalla Figura 1, lo schema dei flussi (a) modella la relazione tra la colonnina e rilevazioni, mentre meteo (b) e quartieri (c) descrivono rispettivamente le dimensioni temporale e spaziale in isolamento.

3.1 Comparazione degli schemi

Il confronto tra gli schemi locali ha evidenziato una serie di conflitti:

1. Conflitti di Eterogeneità Riguardano l'uso di formalismi differenti con diverso potere espressivo.

L'attributo temporale in Rilevazione (data) e in Meteo (time) presenta un conflitto di rappresentazione, infatti in Rilevazione rappresenta anche il fuso orario.

2. Conflitti sui Nomi (Sinonimie) É stata rilevata una discrepanza nella denominazione di concetti semanticamente equivalenti.

La dimensione temporale è identificata come `data` nella sorgente Colonnine e `time` nella sorgente Meteo.

3. Conflitto Semantico Si ha un livello diverso di astrazione.

La sorgente Flussi modella la posizione come un punto geometrico puntuale (latitudine/longitudine della colonnina), mentre la sorgente Quartieri modella l'area geografica come un poligono complesso. La relazione tra le due entità non è 1:1, ma una relazione spaziale di inclusione.

3.2 Schema Concettuale Riconciliato

A seguito della risoluzione dei conflitti, gli schemi locali sono stati fusi in un unico Schema Concettuale Riconciliato.

- Relazione Spaziale: È stata definita la relazione Contiene tra Quartiere e Colonnina. La cardinalità è 1:N (un Quartiere contiene molte Colonnine, una Colonnina si trova in un solo Quartiere).
- Relazione Temporale: L'entità Meteo è stata collegata a Rilevazione tramite la relazione Avviene basata sulla sincronizzazione temporale oraria.

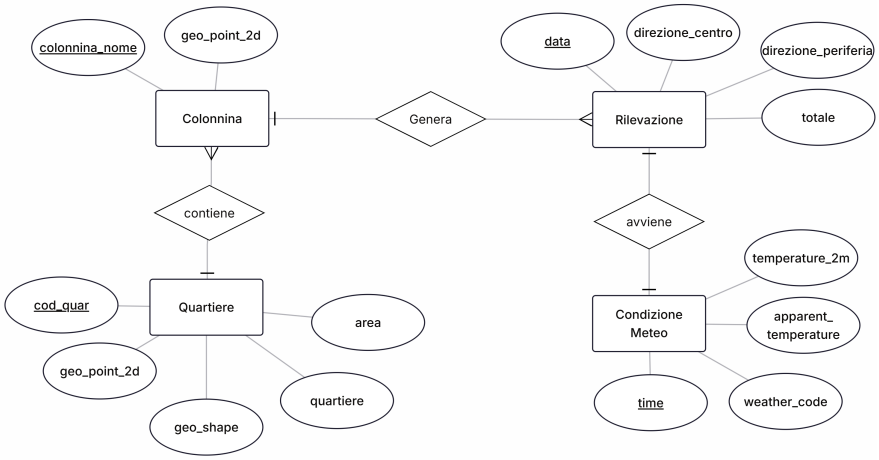


Figure 2: Schema Concettuale Riconciliato

3.3 Schema Logico Riconciliato (ODS)

Il modello concettuale è stato poi tradotto in uno schema logico, che corrisponde al livello ODS (Operational Data Store) implementato nel Data Warehouse. In questa fase sono stati definiti i tipi di dato finali e le chiavi primarie/esterne.

L'ODS è la sorgente unica per l'alimentazione del datamart, la sua progettazione influenza la complessità dei processi di alimentazione del datamart.

Avendo progettato un ODS semi-storicizzato, è stata inserita in ogni tabella una coppia di date, `update_time` e `insert_time` per tenere traccia dell'inserimento e dell'aggiornamento.

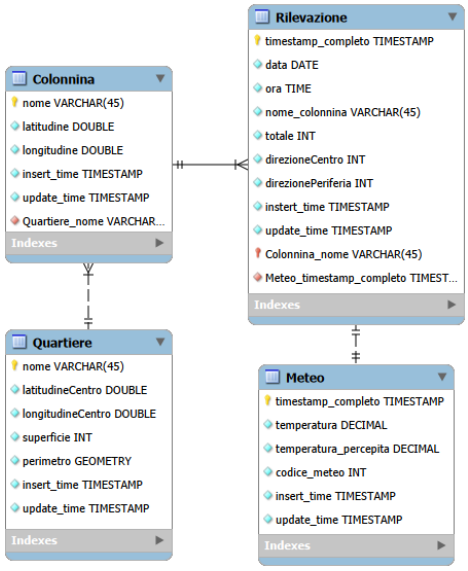


Figure 3: Schema logico ODS

4 Analisi dei Requisiti e Carico di Lavoro

Prima di procedere alla modellazione dimensionale, è necessario formalizzare i requisiti del sistema e stimare il carico di lavoro previsto. Questa fase guida le scelte di progettazione fisica e di aggregazione dei dati.

4.1 Glossario dei Requisiti

Il glossario definisce le entità e le metriche di interesse per l'analisi, mappandole sugli attributi derivati dalla fase di riconciliazione.

Concetto	**Tipologia**	**Descrizione Semantica**
Rilevazione	Fatto	L'evento di rilevazione di biciclette transitate in una specifica ora davanti a una determinata colonnina
Totale	Misura (Additiva)	Conteggio complessivo dei transiti (somma delle due direzioni).
DirezioneCentro	Misura (Additiva)	Conteggio dei ciclisti diretti verso il centro.
DirezionePeriferia	Misura (Additiva)	Conteggio dei ciclisti in uscita dal centro.
Colonnina	Dimensione	Il sensore fisico installato sulla strada. Permette l'analisi granulare del traffico.
Quartiere	Dimensione	Riferimento geografico (poligono), per permettere aggregazioni territoriali.
Data	Dimensione	Riferimento al calendario (Giorno, Mese, Anno).
Ora	Dimensione	Riferimento orario (0-23).
Meteo	Dimensione	Descrizione delle condizioni atmosferiche associate all'ora della rilevazione.

Table 4: Glossario dei termini di business

4.2 Stima dei Volumi di Dati

Dimensionamento Teorico:

- Finestra Temporale: 5 anni (2021-2025) $\approx$ 1.800 giorni.
- Granularità Temporale: Oraria (24 rilevazioni al giorno).
- Dimensione Spaziale: 24 Colonnine distribuite su 5 Quartieri amministrativi.

Il numero di righe della tabella dei fatti è dato dal prodotto tra le ore totali e il numero di sensori:

$$\text{Max Teorico} = 1.800 \text{ giorni} \times 24 \text{ ore} \times 224 \text{ colonnine} \approx \mathbf{1.000.000 \text{ righe}}$$

Tuttavia, analizzando la Sorgente 1, che riguarda i flussi di bici (in base alle varie colonnine e all'orario degli ultimi 5 anni) si hanno 352.622 righe. Questa differenza è dovuta alla crescita progressiva delle colonnine presenti sul territorio nel corso degli anni.

4.3 Carico di Lavoro Preliminare

Il carico di lavoro è stato definito attraverso un insieme di interrogazioni in linguaggio naturale che il sistema deve essere in grado di soddisfare. L'analisi di queste interrogazioni è fondamentale per valutare la granularità dei fatti e le dimensioni.

Il fatto centrale identificato è il flusso delle bici cioè la Rilevazione, e può essere analizzata rispetto a varie dimensioni.

Interrogazione	Dimensioni Coinvolte	Operazioni OLAP
Qual è l'andamento del traffico ciclabile totale per ogni mese dell'anno?	Dimensione Data	Somma sulla misura *Totale*, raggruppando per mese.
Qual è il quartiere della città con il maggior volume di passaggi?	Dimensione Quartiere	Somma sulla misura *Totale* e ordinamento decrescente dei risultati.
Quali sono le ore di punta in cui si registrano più passaggi?	Dimensione Ora	Media o somma sulla misura *Totale* in base all'attributo *Ora*.
Come varia l'utilizzo della bici in caso di pioggia rispetto al tempo sereno?	Dimensione Meteo	Filtrare sull'attributo *Descrizione Meteo*, e confronto tra i passaggi medi aggregati secondo la *Descrizione Meteo*
Durante la mattina, il flusso è prevalentemente verso il centro o verso la periferia?	Dimensione Ora	Filtrare sulla *Fascia Oraria* e fare le somme sulle misure *Dir. Centro* e *Dir. Periferia*.

Table 5: Carico di lavoro preliminare e operazioni richieste

L'analisi delle interrogazioni conferma la necessità di mantenere granularità oraria del fatto, per analizzare correttamente la mobilità urbana.

5 Progettazione Concettuale e Logica

Seguendo la metodologia di progettazione di Data Warehousing, si procede prima alla definizione dello schema concettuale e successivamente alla sua traduzione nello schema logico (orientato al DBMS relazionale).

5.1 Schema Concettuale (DFM)

La progettazione concettuale ha portato alla definizione del *Dimensional Fact Model* (DFM), rappresentato in Figura 4. Il modello ha come fulcro il fatto **Rilevazione**, che permette di misurare il flusso delle biciclette.

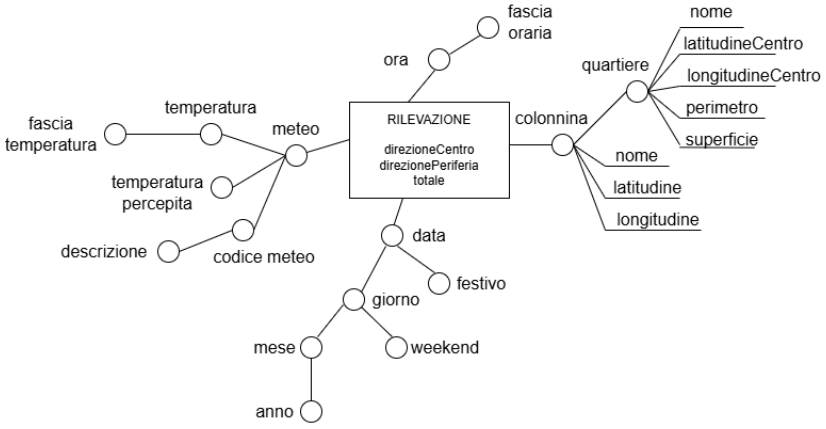


Figure 4: Dimensional Fact Model

Gli elementi presenti nello schema sono:

- **Misure:** Sono le proprietà numeriche del fatto, su cui verranno eseguite le aggregazioni (Sum, Avg).
  - Totale: Il volume complessivo dei transiti.
  - DirezioneCentro e DirezionePeriferia: I dettagli direzionali del flusso.
- **Dimensioni:**
  - Dimensione Spaziale: Il nodo colonnina determina univocamente il quartiere di appartenenza. Gli attributi descrittivi (rappresentati dalle linee nello schema) includono le coordinate geografiche

per la colonnina e le proprietà geometriche per il quartiere.
- Dimensione Temporale: Presenta una gerarchia standard *Data* → *Mese* → *Anno*, arricchita da attributi utili all'analisi come Weekend e Festivo.
- Dimensione Oraria: Separata dalla data per analizzare le fasce orarie (es: Mattina, ...).
- Dimensione Meteorologica: Permette di correlare i flussi alle condizioni atmosferiche, consente di aggregare per descrizione (es. Pioggia) o per fascia di temperatura.

5.2 Schema Logico

La traduzione dello schema concettuale verso il modello relazionale ha portato all'adozione di uno schema di tipo **Snowflake**.

La **Fact Table** si trova al centro dello schema e contiene le misure numeriche e le chiavi esterne verso tutte le dimensioni.

La caratteristica distintiva rispetto a un classico schema a Stella è nella **dimensione spaziale**. Infatti questa dimensione è normalizzata, la tabella Colonnina è separata dalla tabella Quartiere. Questa scelta architetturale è dettata dal fatto che i dati geometrici sono complessi, come il poligono del quartiere, e in questo modo si evita ridondanza in memorizzazione.

Le altre dimensioni (**Temporale**, **Oraria**, **Meteorologica**) invece hanno una struttura denormalizzata, permettendo di effettuare query più efficienti.

Figure 5: SnowFlake Schema

5.3 Gestione delle Dimensioni Dinamiche (SCD)

La gestione dell'evoluzione temporale degli attributi dimensionali (*Slowly Changing Dimensions*) è stata affrontata analizzando ciascuna dimensione rispetto al dominio applicativo.

- **Data e Ora:** Sono dimensioni strutturali e immutabili per definizione.
- **Meteo:** I dati meteorologici, una volta acquisiti e storicizzati, non possono cambiare (es. "Il 15 agosto pioveva").
- **Colonnina:** È soggetta a modifiche negli attributi descrittivi, come il nome, oppure può essere spostata, cambiando il valore di latitudine e longitudine.
- **Quartiere:** I confini amministrativi possono essere ridefiniti nel tempo, alterando l'associazione tra una coordinata spaziale e il quartiere di appartenenza.

5.3.1 Strategia adottata

Per le gerarchie geografiche sarebbe corretto applicare in via teorica o un SCD tipo 2, per preservare la veridicità storica oppure SCD tipo 3 per confrontare i dati mantenendo sia il valore corrente che quello precedente.

Nel progetto è stata però adottata una strategia SCD tipo 1, che permette l'aggiornamento del dato sovrascrivendo il valore passato. Questa scelta è stata presa per vari motivi;

1. La frequenza di spostamento fisico delle colonnine o di ridefinizione dei quartieri a Bologna è storicamente molto bassa, rendendo l'impatto dell'errore di sovrascrittura trascurabile.
2. L'obiettivo primario del sistema è supportare decisioni sulla rete infrastrutturale attuale. Quindi si preferisce avere visione dei flussi storici ma sulla configurazione geografica odierna.

3. L'implementazione tramite viste in dbt garantisce che il Data Mart sia sempre allineato all'ultima versione, riducendo la complessità e rendendo le interrogazioni più veloci.

6 Processo ELT

Il popolamento del Data Warehouse è stato realizzato implementando una pipeline di tipo ELT (Extract, Load, Transform).

A differenza del tradizionale approccio ETL, in cui le trasformazioni avvengono prima del caricamento, l'architettura scelta sfrutta la potenza di DuckDB per eseguire le trasformazioni direttamente sui dati caricati, orchestrate e versionate tramite il framework **dbt** (data build tool).

6.1 Architettura Logica del Data Warehouse

Per garantire modularità, consistenza e tracciabilità del dato, l'architettura del sistema è stata organizzata secondo un modello logico a tre livelli. Questa stratificazione permette di separare nettamente le fasi di acquisizione, integrazione e presentazione del dato.

Il flusso dei dati attraversa i seguenti livelli logici:

1. **Raw Layer:** Qui vengono caricati i dati grezzi provenienti dalle sorgenti nel loro formato originale, senza alcuna trasformazione strutturale. Questo livello funge da archivio e permette di riprocessare i dati in caso di errori nelle fasi successive senza dover interrogare nuovamente le fonti esterne.
2. **Dati riconciliati / ODS:** Rappresenta il livello intermedio di integrazione. In questa fase avvengono la pulizia (data cleaning), la normalizzazione dei tipi di dato e la gestione delle logiche spaziali.
3. **Data Mart Layer:** È il livello finale esposto agli strumenti di analisi. Contiene le tabelle dimensionali e i fatti modellati secondo lo schema logico definito in precedenza. Qui i dati sono organizzati per rispondere efficientemente alle query.

6.2 Fase 1: Estrazione e Caricamento

Questa fase ha l'obiettivo di trasferire i dati dalle sorgenti eterogenee all'interno del database DuckDB.

Vengono utilizzati due script python differenti.

1. **Extractor:** Questo script gestisce l'interazione con le sorgenti dati esterne. Si occupa di:
   - Eseguire le chiamate API verso i fornitori dati (Portale Open Data Bologna, Open-Meteo).
   - Salvare i payload ricevuti (JSON o CSV) in locale, preservando il formato originale.
2. **Loader:** Questo script si occupa esclusivamente dell'interazione con DuckDB. Legge i file grezzi salvati dall'Extractor e li carica nelle tabelle del livello `raw` come raw_colonnine, raw_meteo e raw_quartieri. L'utilizzo delle funzioni native di DuckDB (come `read_csv_auto` o `read_json_auto`) consentono di ottimizzare la procedura di caricamento, delegando al database l'inferenza dei tipi di dato e della struttura tabellare.

Avendo separato il Loader dall'extractor, teoricamente è possibile ri-eseguire il caricamento nel database senza dover necessariamente effettuare nuove chiamate API.

6.3 Fase 2: Livello ODS

Per gestire la complessità delle trasformazioni nell'ODS, è stata utilizzata una separazione tra logica e materializzazione, dividendo ogni ODS in due modelli distinti:

6.3.1 1. Il Modello di Trasformazione (Ephemeral)

I file con suffisso `_transform` contengono esclusivamente la logica SQL di manipolazione del dato grezzo. Questi modelli sono configurati come `materialized='ephemeral'`: ciò significa che dbt non crea una tabella fisica nel database, ma inietta il codice SQL nel modello successivo, ottimizzando lo storage.

Esempio ods_colonnina_transform.sql

```
{{ config(materialized='ephemeral') }}
SELECT DISTINCT
    colonnina as nome,
    CAST(split_part(geo_point_2d, ',', 1) AS DOUBLE) as latitudine,
    CAST(split_part(geo_point_2d, ',', 2) AS DOUBLE) as longitudine,
    CAST(data AS TIMESTAMP) AS timestamp_completo,
    current_timestamp AS insert_time,
    current_timestamp AS update_time
FROM {{ source('staging', 'raw_colonnine') }}
```

6.3.2 2. Il Modello di Storicizzazione (Incremental)

Il modello principale si occupa esclusivamente della strategia di persistenza del dato. Configurato come `materialized='incremental'`, questo modello legge dalla sua versione *transform* e applica la logica di aggiornamento, inserendo solo i nuovi record o aggiornando quelli esistenti.

Esempio: ods_colonnina.sql

```
{{ config(materialized='incremental', unique_key='nome') }}
SELECT *
FROM {{ ref('ods_colonnina_transform') }}
{% if is_incremental() %}
  WHERE timestamp_completo > (SELECT IFNULL(MAX(timestamp_completo), '0001-01-01 00:00:00')
  FROM {{ this }})
{% endif %}
```

6.4 Fase 3: Data Mart

Il livello finale, il Data Mart, è stato implementato attraverso una strategia di Materializzazione Incrementale. Questa scelta permette di stabilizzare le chiavi surrogate.

6.4.1 Generazione Chiavi per le Dimensioni Standard

Per le dimensioni che non dipendono da altre entità (come `dm_data`, `dm_ora`, `dm_meteo`), viene utilizzata una `SEQUENCE` di DuckDB per assegnare un ID univoco progressivo a ogni nuovo record inserito.

```
{% set initialize %}
    CREATE SEQUENCE IF NOT EXISTS seq_dm_meteo;
{% endset %}
SELECT
    {% if is_incremental() %}
        IFNULL(target.id_meteo, nextval('seq_dm_meteo'))
    {% else %}
        nextval('seq_dm_meteo')
    {% endif %} as id_meteo,
    ...,
```

6.4.2 Gestione delle Gerarchie Spaziali (Colonnina-Quartiere)

Una logica più complessa è necessaria per la dimensione `dm_colonnina`, in cui ogni colonnina deve contenere la chiave esterna del quartiere in cui si trova.

Poiché il dato grezzo della colonnina possiede solo le coordinate geografiche (Latitudine/Longitudine), c'è stato bisogno di implementare un passaggio intermedio tramite il modello `dm_colonnina_fk_lookup`. Inizialmente ho utilizzato la semplice inclusione geometrica (ST_Within) che però ha portato a dei problemi quando una colonnina si trovava vicino al confine, in quanto risultava in più quartieri.

Per garantire un'assegnazione corretta, invece di cercare un'inclusione esatta è stata inserita una tolleranza. `ST_Within(ST_Point(c.longitudine, c.latitudine), oq.geom_perimetro, 0.0002)` è una funzione specifica di DuckDB spatial che permette di verificare che il punto dato come primo argomento si trovi

fisicamente dentro il poligono dato come secondo argomento con una tolleranza di circa 20 metri.

Tra i quartieri candidati, alla colonnina viene assegnato il quartiere il cui centroide è più vicino alla colonnina.

```
SELECT
    c.nome,
    ...
    ST_Distance(
            ST_Point(c.longitudine, c.latitudine),
            ST_Point(dq.longitudineCentro, dq.latitudineCentro)
        ) as distanza
FROM {{ ref('ods_colonnina') }} c
    LEFT JOIN {{ ref('ods_quartiere') }} oq
        ON ST_Within(ST_Point(c.longitudine, c.latitudine), oq.geom_perimetro, 0.0002)
    LEFT JOIN {{ ref('dm_quartiere') }} dq
        ON oq.nome = dq.nome_quartiere
```

6.4.3 Popolamento della Fact Table

La tabella dei fatti (`dm_rilevazione`) è l'unica che necessita di associare (lookup) tutte le chiavi dimensionali contemporaneamente. Utilizzando un modello intermedio (`dm_rilevazione_fk_lookup`), le chiavi naturali presenti nei dati vengono sostituite con le rispettive chiavi surrogate recuperate dalle dimensioni già popolate.

Inoltre il caricamento incrementale è ottimizzato tramite un meccanismo basato su una tabella di controllo (`last_execution_times`), che garantisce l'elaborazione dei soli dati giunti dopo l'ultima esecuzione:

```
{% if is_incremental() %}
    WHERE timestamp_completo > (
        SELECT COALESCE(MAX(time), '1900-01-01')
        FROM last_execution_times
        WHERE target_table = '{{ this.identifier }}'
    )
{% endif %}
```

7 Dashboard

L'obiettivo del progetto, è fornire uno strumento di supporto alle decisioni per l'amministrazione comunale, in modo che in base alle analisi effettuate possano decidere dove collocare nuove piste ciclabili o se iniziare un servizio di bikesharing.

Dopo aver alimentato i Data Mart, i dati sono stati visualizzati tramite una dashboard interattiva, realizzata utilizzando Tableau Public, che permette di rispondere alle interrogazioni sviluppate durante l'analisi del carico di lavoro preliminare.

7.1 Trend Temporali

Sono stati elaborati due grafici distinti in modo da analizzare il flusso sia stagionale che storico.

Il grafico mostra la stagionalità dell'utilizzo delle biciclette a Bologna. Aggregando i dati storici per mese, emerge che:

- I picchi di utilizzo si registrano nei mesi primaverili come Maggio e Giugno e nei mesi autunnali come Settembre ed Ottobre
- Si nota una discesa dei flussi con l'intensificazione dell'inverno che raggiunge il suo minimo nel mese di Gennaio.
- Si ha un calo netto nel mese di Agosto, che coincide con il periodo tipicamente di ferie e la chiusura di scuole e

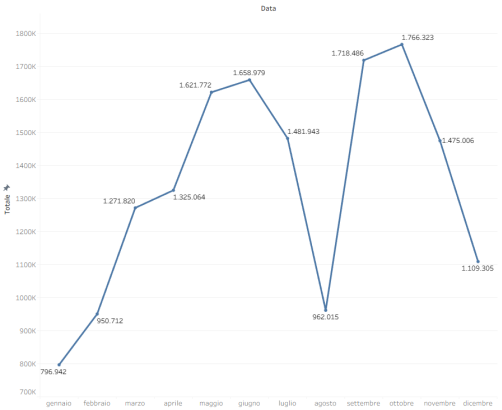


Figure 6: Andamento mensile

uffici.

Il grafico 7 invece mostra l'andamento di tutto il flusso ciclistico
rilevato dal 2021 alla fine del 2025.

La curva mostra tendenzialmente di crescere, con picchi sempre più alti. Per l'interpretazione del grafico però si deve tener conto che l'altezza degli ultimi picchi potrebbero essere dovuta oltre che all'effettivo incremento dell'utilizzo delle biciclette, anche all'installazione di un numero maggiore di colonnine sul territorio.

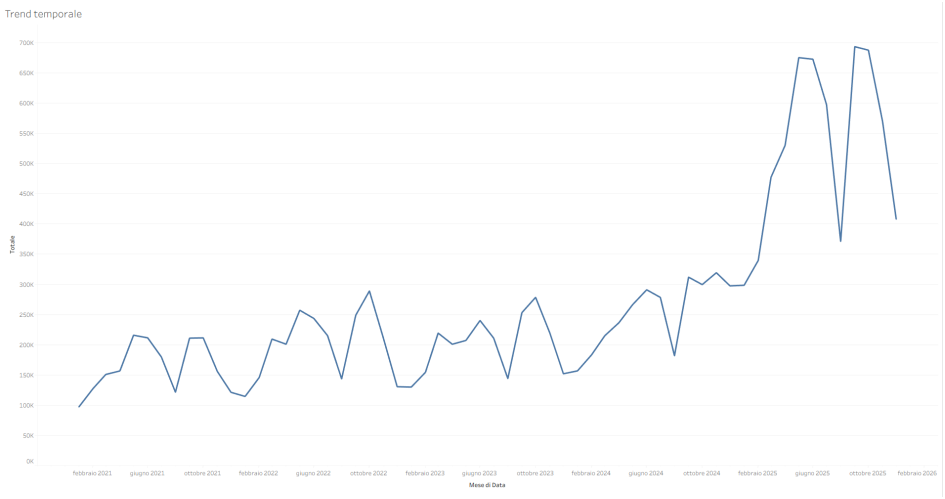


Figure 7: Andamento storico

7.2 Impatto Meteorologico

Uno degli obiettivi specifici del progetto era valutare quanto le condizioni meteorologiche potessero influenzare la decisione di utilizzare la bicicletta. Il seguente grafico mette in relazione i passaggi medi in un'ora in cui non ci sono precipitazioni rispetto a un'ora in cui piove o nevica.

Si osserva una riduzione significativa dell'uso della bicicletta in presenza di precipitazioni, in cui il traffico è quasi la metà rispetto a quando il cielo è sereno o solo nuvoloso. Per il caso specifico della neve, si osserva che il calo non è dovuto solo alla precipitazione in sé, ma anche dovuto alle temperature rigide associate.

Un'analisi condotta sulla fascia di temperatura ha evidenziato infatti una correlazione diretta tra i gradi e l'utilizzo della bicicletta.

- Freddo (< 10 gradi): si registra una media di 31,56 passaggi orari.
- Mite (10-25 gradi): il valore sale significativamente a 50,27 passaggi orari medi.
- Caldo (> 25 gradi): si raggiunge il picco con una media di 62,74 passaggi orari.

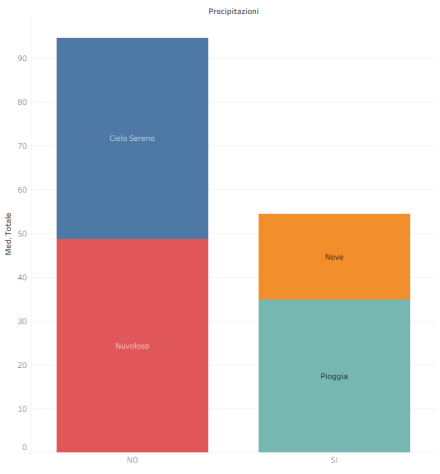


Figure 8: Precipitazioni

7.3 Analisi Oraria e Direzionale

Per rappresentare i picchi di traffico è stata analizzata la distribuzione oraria, tenendo però conto anche della direzione di marcia e dei quartieri in cui viene generato il flusso.

7.3.1 Identificazione ore di punta

Il grafico 9 mostra il volume medio dei passaggi per ogni ora del giorno, segmentato per quartiere.

Si possono notare due picchi principali in corrispondenza degli orari in cui solitamente si fanno spostamenti per andare o tornare da scuola/lavoro. Il picco mattutino è netto e si raggiunge alle 8 di mattina, con quasi 400 passaggi orari. Il rientro, e quindi il picco pomeridiano, è più distribuito tra le 16 e le 19.

Grazie alla visualizzazione a barre impilate, si può notare come i quartieri con San Donato-San Vitale e Porto-Saragozza sono quelli che producono più flusso ciclistico in qualsiasi orario.

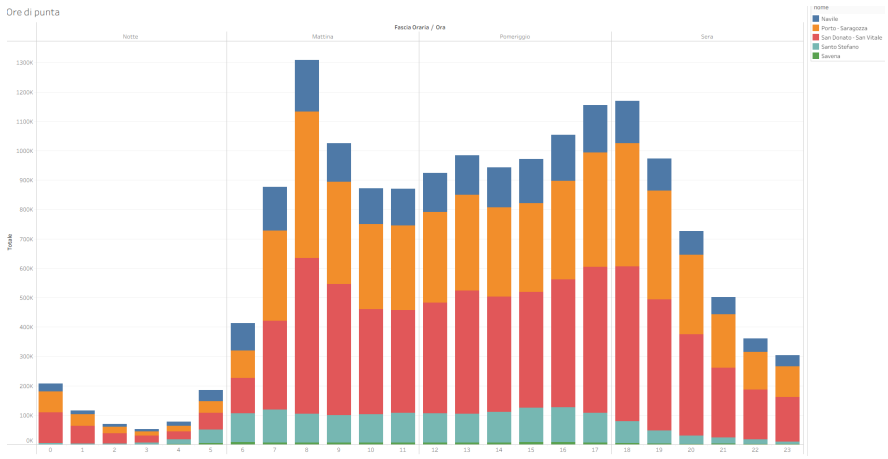


Figure 9: Ore di punta

7.3.2 Analisi flussi direzionali

Una volta identificati gli orari di punta, si è voluto analizzare come questo flusso si dividesse tra persone che si spostano verso il centro della città e persone che si spostano verso la periferia. In figura il flusso in direzione centro è identificato dal verde acqua mentre il flusso verso la periferia è rappresentato dal verde scuro.

- Mattina (7:00-9:00): in queste ore prevale nettamente il flusso diretto verso il centro. Suggerendo una correlazione con gli orari tipici di ingresso a scuole e posti di lavoro.
- Pomeriggio/Sera (17:00-19:00): resta sempre maggiore il flusso diretto verso il centro, ma la differenza tra i due flussi diminuisce molto, e ci sono più persone che tornano verso la periferia.

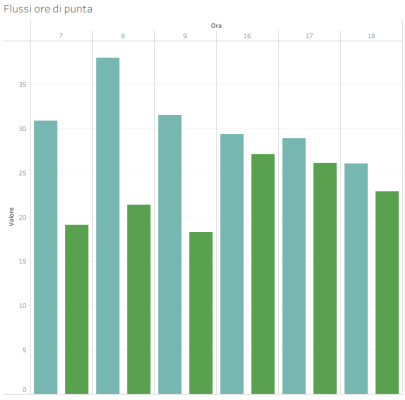


Figure 10: Direzione ore di punta

7.4 Distribuzione territoriale

L'ultimo livello di analisi ha l'obiettivo di contestualizzare i flussi all'interno del comune di Bologna.

7.5 Flussi per quartiere

Dall'analisi emerge che uno dei quartieri con il volume di passaggi maggiore è San Donato - San Vitale, che è storicamente conosciuto per la zona delle fiere e la città della salute.

Un altro quartiere con un grande flusso ciclistico è Porto-Saragozza, che contiene parte del centro storico di Bologna insieme al quartiere Santo Stefano.

Nel quartiere Borgo Panigale-Reno non vengono registrati flussi ciclistici, questo quartiere è quello geograficamente più lontano rispetto al centro di Bologna quindi probabilmente il comune non ha ancora installato delle colonnine conta-bici in questo quartiere.

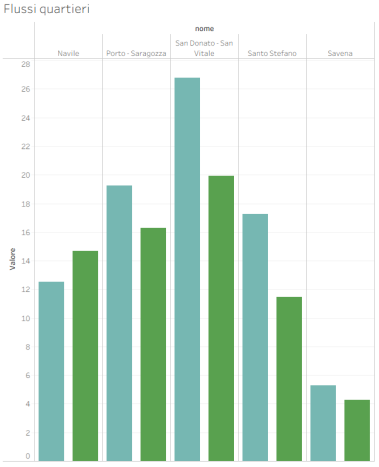


Figure 11: Flusso nei quartieri

Inoltre si nota come il divario tra flusso direzione centro (Verde Acqua) e quello direzione periferia (Verde scuro) sia più elevato nei quartieri in cui c'è un maggior volume di passaggi.

7.5.1 Mappa densità passaggi

Per visualizzare come le colonnine fossero distribuite sul territorio, e in quali punti precisi venisse rilevato un numero maggiore di passaggi in bici è stata realizzata una mappa dei passaggi (Figura 12), dove ogni colonnina è rappresentata da un cerchio, colorato in base al quartiere in cui si trova e con la dimensione proporzionale al flusso che rileva.

La mappa conferma che San Donato-San Vitale e Porto-Saragozza sono i quartieri più trafficati e che, in particolare, siano due colonnine specifiche a rilevare la maggior parte dei passaggi.

Si nota inoltre come le colonnine siano state posizionate nelle direttrici verso il centro città, ma non ce ne siano nel centro preciso di Bologna.

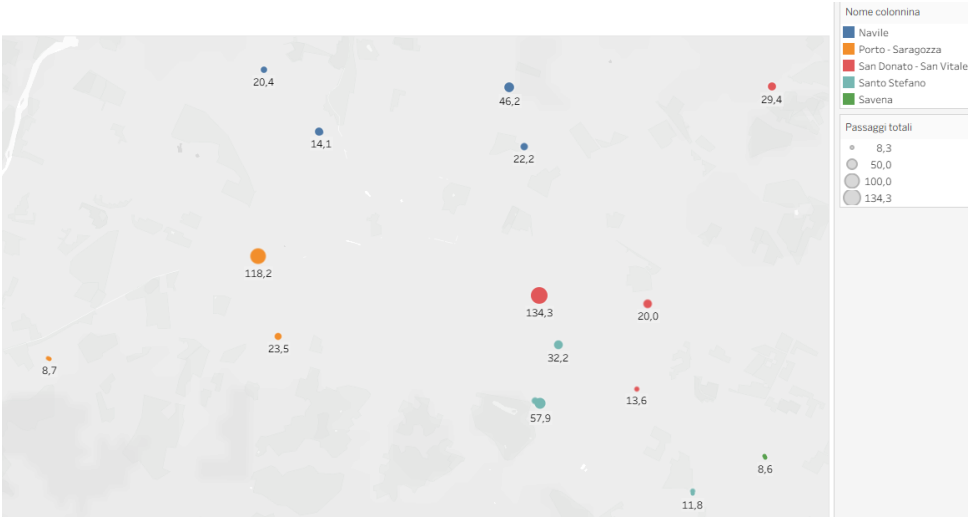


Figure 12: Mappa colonnine

8 Conclusioni

Il progetto ha portato alla realizzazione di un Data Warehouse, in grado di integrare dati eterogenei provenienti da sensori, servizi meteorologici e dati geografici. L'architettura ELT basata su DuckDB e dbt ha dimostrato di poter gestire efficacemente la complessità delle trasformazioni spaziali e temporali, fornendo una base dati solida.

Il processo di trasformazione ha permesso di portare i dati da grezzi a dati pronti per effettuare delle analisi. L'analisi condotta attraverso la dashboard finale ha permesso di rispondere alle interrogazioni che sono state redatte nella fase iniziale, offrendo all'amministrazione comunale o ad aziende di bikesharing indicazioni concrete per la pianificazione urbana.

8.1 Sintesi dei Risultati

Dall'incrocio delle dimensioni spaziali, temporali e meteorologiche sono emerse tre evidenze principali che possono guidare le future politiche di mobilità:

8.1.1 Pendolarismo

L'analisi oraria ha rivelato picchi di traffico estremamente concentrati alle ore 08:00 (direzione centro) e tra le 17:00 e le 19:00 in entrambe le direzioni.

Questo suggerisce che un eventuale servizio di Bike Sharing comunale non può basarsi su una distribuzione statica delle bici. Sarebbe invece necessario ridistribuire le bici in base agli orari e alle zone di utilizzo.

8.1.2 Meteo

I dati hanno mostrato un crollo dell'utilizzo non solo in caso di precipitazioni, ma anche in presenza di temperature rigide (< 10°C), dove la media dei passaggi scende a 31,56 rispetto ai 62,74 registrati con temperature calde.

Questo sottolinea il fatto che la bicicletta a Bologna è ancora percepita prevalentemente come mezzo stagionale. Per incentivare l'uso tutto l'anno, bisognerebbe introdurre misure per la manutenzione invernale delle piste ciclabili, evitando possibili allagamenti delle piste e rimuovendo la neve o il ghiaccio.

8.1.3 Quartieri

I quartieri San Donato-San Vitale e Porto-Saragozza rappresentano le aree con la maggiore concentrazione di passaggi.

Tuttavia l'analisi del flusso potrebbe non rappresentare alla perfezione i dati reali, in quanto la mappa delle colonnine ha evidenziato una carenza di sensori nel cuore del centro storico e nella parte nord-ovest della città, lasciando scoperta una parte importante del territorio.

Per completare il quadro conoscitivo, si raccomanda l'installazione di nuovi punti di rilevazione anche nel centro di Bologna e nel quartiere Borgo Panigale-Reno. Questo permetterebbe di tracciare in modo più completo gli spostamenti e comprendere le zone di destinazione finale, dati cruciali per pianificare il posizionamento di nuove rastrelliere per bici o la costruzione di nuove piste ciclabili.