Data warehousing per applicazioni ambientali

Obiettivi di apprendimento

Al termine di questo argomento, gli studenti saranno in grado di:

- Spiegare il concetto e il ruolo dell'archiviazione dei dati nelle applicazioni ambientali.
- Configurare un data warehouse e gestirne le funzionalità principali.
- Cercare, scaricare e caricare dati ambientali da fonti esterne documentate.
- Applicare interrogazioni SQL al data warehouse ambientale realizzato.

Introduzione

I dati ambientali vengono raccolti mediante diverse misurazioni, ognuna delle quali descrive soltanto un aspetto del complesso sistema ecologico. Alcuni dataset riguardano le proprietà fisiche dell'aria, altri le condizioni meteorologiche, altri ancora lo stato di salute degli ecosistemi o i livelli di attività umana. Sebbene ciascuno fornisca informazioni preziose, nessuno offre da solo un quadro completo dei processi che influenzano l'ambiente. Per questo motivo, l'analisi dei dati ambientali inizia quasi sempre con la loro integrazione e armonizzazione in una visione unificata delle condizioni ambientali.

In questa attività emergono due sfide principali: individuare i dati pertinenti, spesso distribuiti tra più istituzioni, e conservarli in un formato adatto alle analisi successive. I dati possono provenire da formati, periodi temporali e fasi di elaborazione differenti; devono quindi essere integrati in un sistema che consenta operazioni affidabili di unione, aggregazione e confronto. Questa è precisamente la funzione di un data warehouse: riunire dati eterogenei in una struttura unificata, offrendo un ambiente stabile per interrogazioni complesse, analisi delle tendenze e interpretazione dei risultati.

Un data warehouse si distingue dai tradizionali database transazionali per scopo e modalità d'uso. Mentre i sistemi transazionali sono ottimizzati per inserimenti rapidi e frequenti, i data warehouse sono orientati alla conservazione e all'analisi di grandi quantità di dati storici. L'attenzione si concentra sulle interrogazioni analitiche, sull'evoluzione delle tendenze nel tempo e sull'integrazione di fonti diverse, utile per comprendere fenomeni complessi e le loro interrelazioni.

Nelle scienze ambientali questo approccio è fondamentale, poiché misurazioni isolate raramente permettono di cogliere i cambiamenti ecologici. La qualità dell'aria, ad esempio,

dipende dalla combinazione di condizioni meteorologiche, fonti di emissione, caratteristiche geografiche e variazioni stagionali. Solo osservando congiuntamente questi dati per periodi prolungati possiamo valutare accuratamente l'evoluzione delle condizioni ambientali, l'efficacia degli interventi e i possibili rischi.

La Figura 4.1 illustra un'architettura semplificata di data warehouse che funge da punto centrale per diversi formati di dati, dai tradizionali file CSV e dai record semistrutturati JSON/XML ai file Parquet e ai dati recuperati tramite API o database di altri sistemi. Tutte queste fonti vengono consolidate in un unico sistema, nel quale i dati sono organizzati, collegati e preparati per l'analisi. Il data warehouse diventa così un'infrastruttura affidabile per ricercatori, istituzioni e decisori, favorendo aggiornamenti continui e politiche basate su evidenze.

Una volta caricati nel data warehouse, i dati permettono numerose analisi: dalle semplici statistiche descrittive e visualizzazioni all'esplorazione delle relazioni tra variabili e all'impiego di tecniche avanzate di machine learning e intelligenza artificiale. In ambito ambientale, questo può aiutare a prevedere episodi di inquinamento, valutare l'influenza delle condizioni meteorologiche sulla qualità dell'aria e analizzare le tendenze di lungo periodo dei processi ecologici.

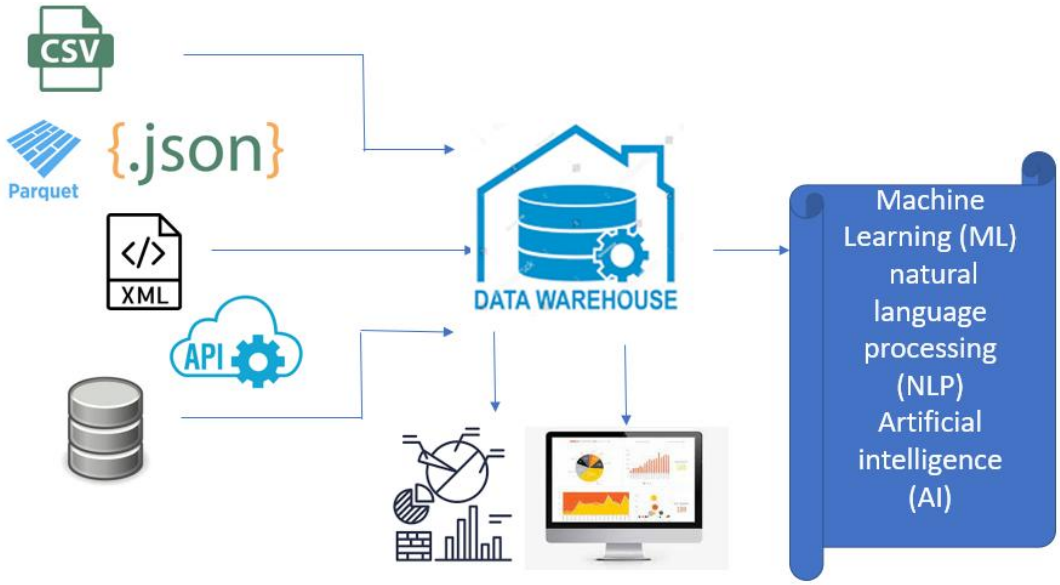

Figura 4.1 Panoramica semplificata dell'architettura di un data warehouse

Contesto tecnologico

Oggi esiste una grande varietà di tecnologie per costruire data warehouse, differenti per scalabilità, complessità di gestione, funzionalità e costi complessivi. La scelta dipende dalle

esigenze degli utenti, dalle risorse disponibili, dai tipi di dati e dai metodi di analisi previsti. A un estremo vi sono installazioni locali di database relazionali tradizionali, che offrono pieno controllo ma richiedono supporto tecnico e manutenzione continua. All'altro estremo vi sono piattaforme cloud commerciali, caratterizzate da elevata automazione, buone prestazioni e capacità di elaborare grandi dataset, ma con modelli di tariffazione e complessità operative differenti.

Per questo materiale è stato scelto Oracle Autonomous Database. Oracle Autonomous Database [1] consente di creare data warehouse in un ambiente stabile, affidabile e altamente automatizzato. Nell'ambito di Oracle Cloud [2] sono disponibili risorse Always Free, accessibili anche ai principianti senza ulteriori costi infrastrutturali. I sistemi integrati per la gestione automatica della sicurezza, della configurazione e degli aggiornamenti eliminano molte barriere tecniche, consentendo agli utenti meno esperti di concentrarsi sull'analisi e sull'interpretazione dei dati.

La scelta di questa piattaforma non esclude comunque altre soluzioni. I principi presentati si applicano anche ad altre tecnologie: Amazon Redshift [3], Google BigQuery [4], Microsoft Azure Synapse Analytics [5] oppure sistemi open source come Apache Hive [6], a seconda delle esigenze e delle risorse. Ognuna adotta un approccio diverso all'archiviazione dei dati, con propri modelli di scalabilità e vantaggi, ma i concetti fondamentali dell'integrazione dei dati e dell'elaborazione analitica rimangono invariati.

Creazione e configurazione di un data warehouse

Per creare un data warehouse con Oracle Autonomous Database si accede innanzitutto al proprio account Oracle Cloud. È possibile creare un account gratuito tramite il programma Oracle Academy [7] oppure seguendo le istruzioni riportate in [8]. Dopo l'accesso, selezionare il menu Oracle AI Database e successivamente Autonomous AI Database, come illustrato nella Figura 4.2. Viene mostrata una panoramica dei database associati all'account.

Fare clic su Create Autonomous Database per configurare un nuovo database. Il modulo richiede le informazioni essenziali sull'istanza, tra cui nome del database, account cloud utilizzato e ambiente di lavoro. Si consiglia di scegliere un ambiente Lakehouse, pensato per realizzare data warehouse e dotato di strumenti aggiuntivi per amministrare il database, eseguire query SQL e applicare tecniche di machine learning ai dataset.

Abilitare quindi l'accesso Always Free e selezionare la versione del database. La versione più comune è proposta per impostazione predefinita, con la possibilità di passare a quella più recente. Definire poi le credenziali dell'amministratore impostando la password dell'utente ADMIN, che

inizialmente ha accesso completo al sistema e si occupa di creare utenti, caricare dati e configurare il database. La password deve soddisfare i requisiti di sicurezza. Infine, scegliere le modalità di accesso alla rete dell'istanza.

In un ambiente Always Free è generalmente possibile accedere in modo sicuro via Internet da qualsiasi dispositivo, usando un wallet digitale e i meccanismi di autenticazione integrati nel database autonomo. Questo approccio semplifica notevolmente l'attività didattica, evitando la necessità di una rete dedicata o di ulteriori componenti infrastrutturali, come tunnel VPN o restrizioni sugli indirizzi IP.

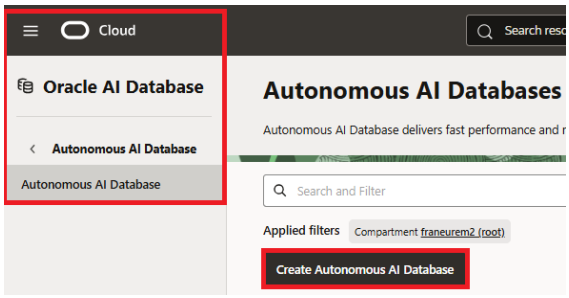

Figura 4.2 Accesso alla piattaforma Oracle Autonomous Database

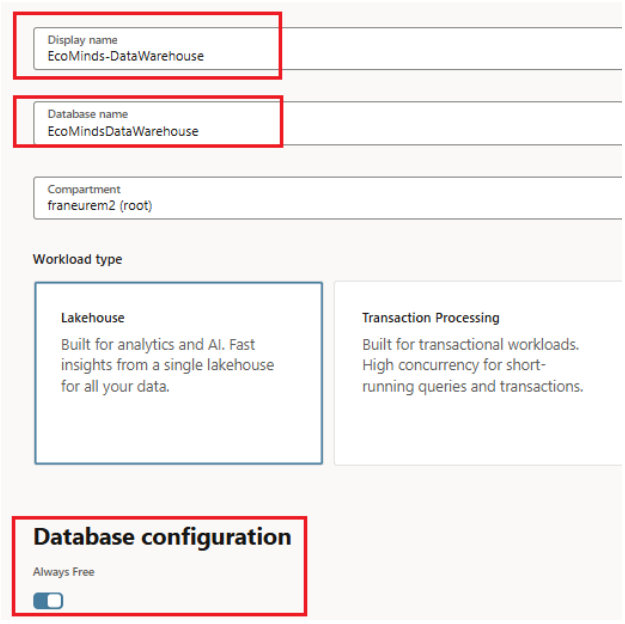

Figura 4.3 Scelta del nome del data warehouse e del tipo di carico di lavoro

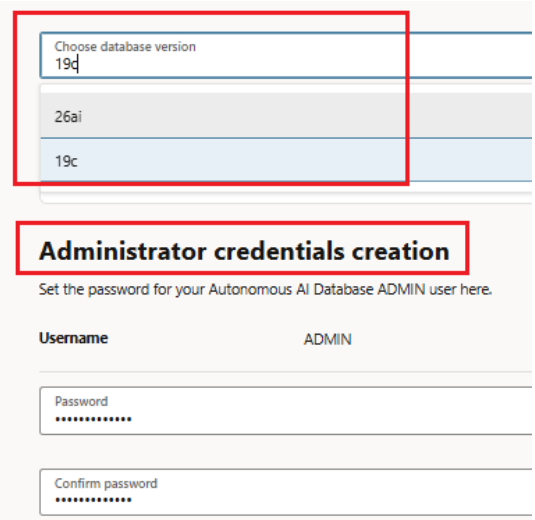

Figura 4.4 Configurazione finale del data warehouse

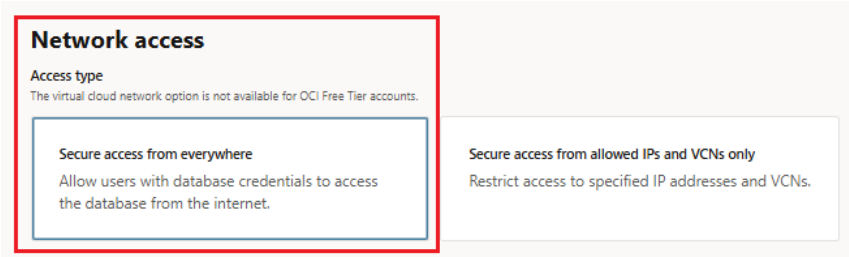

Figura 4.5 Selezione dell'accesso alla rete

Autenticazione e ambiente di sviluppo

Tutte le analisi di dati ambientali nel data warehouse si basano su SQL per recuperare, filtrare, aggregare e collegare i dati, creando viste analitiche complete. Per eseguire correttamente queste interrogazioni, i ricercatori devono poter accedere al data warehouse e disporre di un ambiente di sviluppo affidabile.

Esistono due modalità principali per connettersi a Oracle Autonomous Database. La prima, e più semplice, utilizza l'interfaccia web di Oracle Cloud. Durante l'accesso viene utilizzata l'autenticazione a due fattori (2FA), che permette di accedere alle risorse necessarie. Una volta effettuato l'accesso, selezionare Database Actions e poi SQL per aprire l'ambiente di sviluppo web integrato in cui eseguire istruzioni SQL (Figure 4.6 e 4.7). Questo metodo non richiede ulteriori installazioni e può essere usato da qualsiasi dispositivo e browser, risultando particolarmente comodo per i principianti.

Un'alternativa è l'applicazione desktop Oracle SQL Developer, scaricabile gratuitamente da [9]. SQL Developer offre strumenti avanzati per gestire script, supporta progetti più complessi,

consente di esaminare la struttura del database e facilita le attività amministrative. Per connettersi è necessario scaricare un wallet digitale dalla console Oracle Cloud tramite Database Connection (Figura 4.8). Il pacchetto comprende certificati e file di configurazione necessari per stabilire una connessione mTLS sicura al data warehouse, con autenticazione reciproca tra client e server. Dopo aver caricato il wallet in SQL Developer (Figura 4.9), è possibile connettersi e lavorare con un normale ambiente desktop (Figura 4.10).

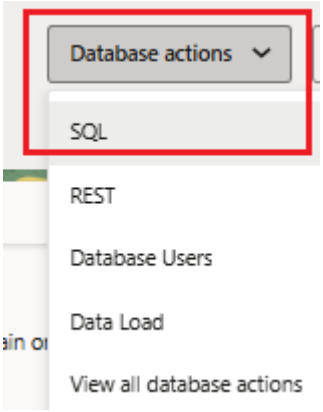

Figura 4.6 Avvio dell'applicazione web SQL Developer

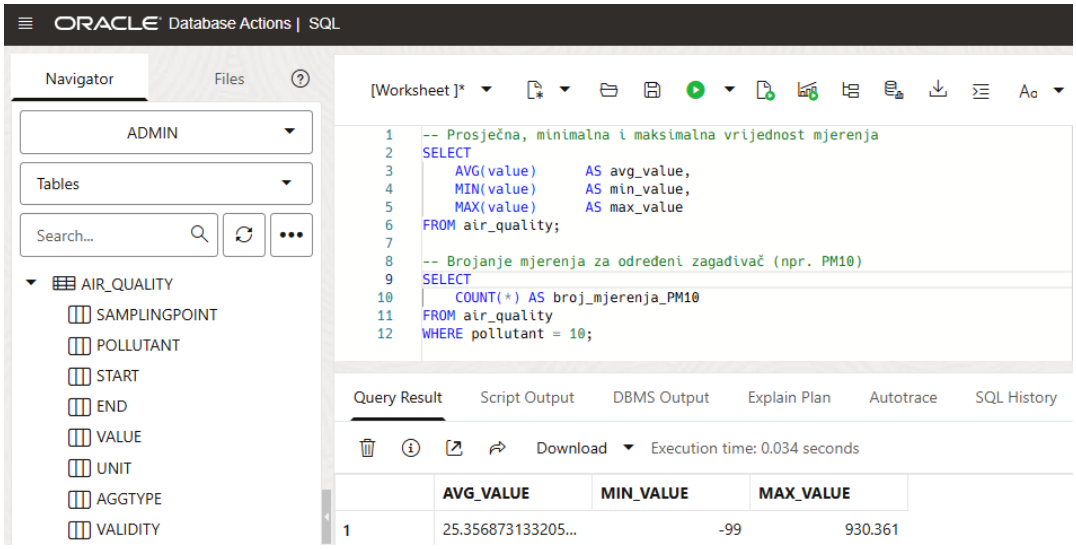

Figura 4.7 Applicazione web SQL Developer

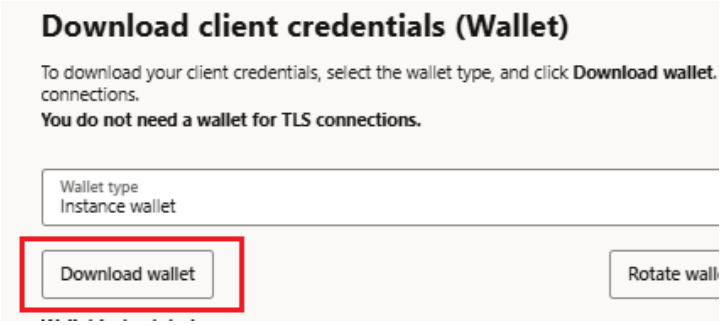

Figura 4.8 Download del wallet digitale

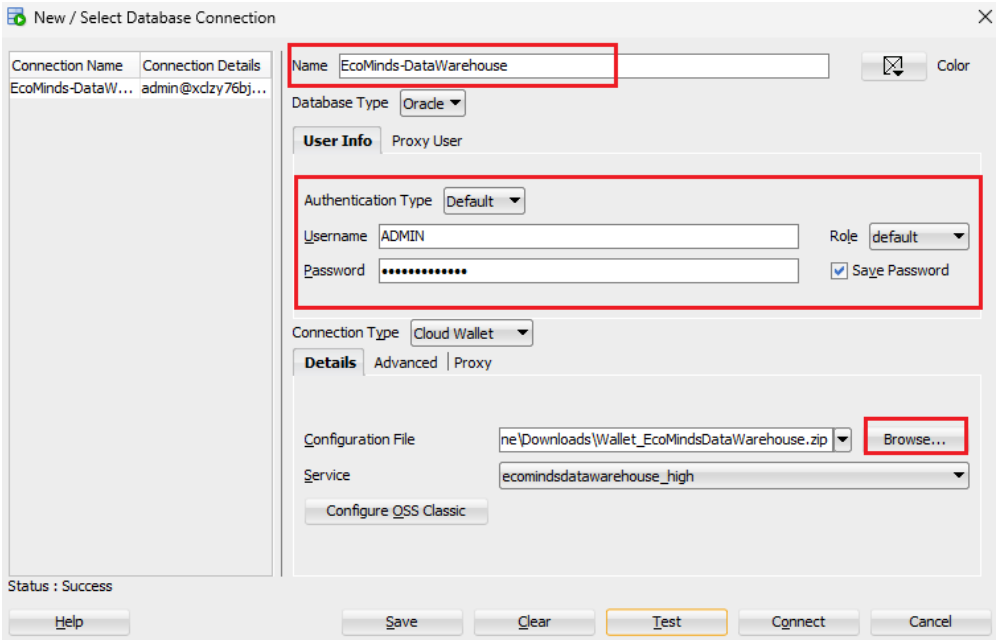

Figura 4.9 Configurazione della connessione al data warehouse da SQL Developer desktop

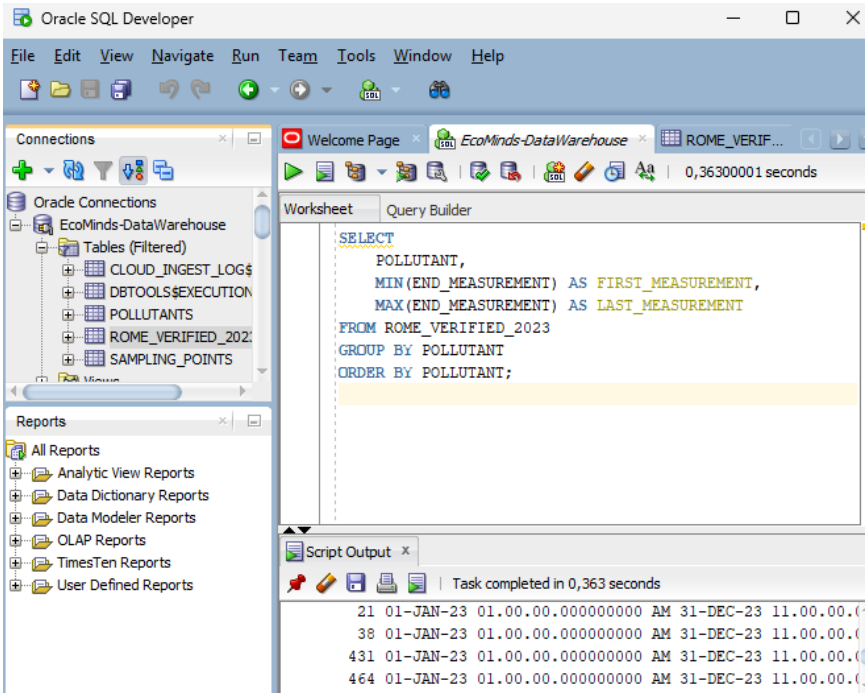

Figura 4.10 Lavorare nell'applicazione desktop SQL Developer

Gestione dei permessi degli utenti

Dopo la creazione del database, il primo utente disponibile è ADMIN. In ambito didattico, soprattutto durante le esercitazioni introduttive, si può utilizzare temporaneamente l'account ADMIN per semplificare il caricamento dei dati e l'esecuzione di query SQL di base. Sebbene questa scelta sia utile per prendere familiarità con il sistema, non è considerata appropriata nelle applicazioni professionali.

Negli ambienti di produzione, il ruolo ADMIN deve essere riservato alle attività amministrative, come la creazione di utenti, l'assegnazione di privilegi e la configurazione del database. Poiché l'account ADMIN dispone di privilegi completi, il suo utilizzo per le operazioni quotidiane comporta rischi per la sicurezza. Per le attività di analisi e sviluppo si creano quindi account o ruoli dedicati con accesso limitato, in accordo con il principio del privilegio minimo.

Si osservi che ADMIN non ha accesso a tutti gli strumenti specializzati di Autonomous Database. Alcune funzionalità, come Oracle Machine Learning (OML), l'analisi dei grafi e specifici componenti web, richiedono account utente dotati di particolari privilegi. Per lavorare con questi strumenti avanzati vengono quindi creati utenti separati, con livelli di accesso definiti, come mostrato nelle Figure 4.11–4.13.

Figura 4.11 Selezione della gestione utenti

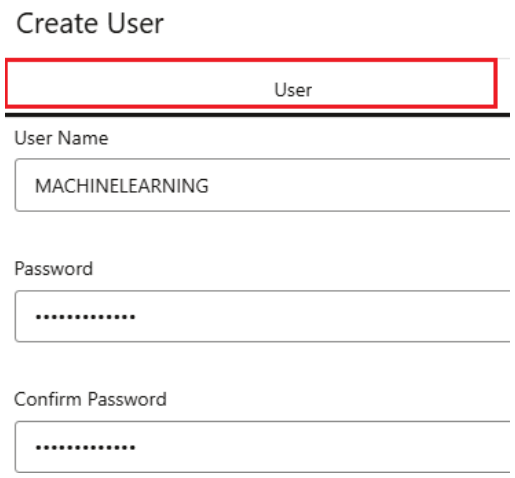

Figura 4.12 Creazione di un nuovo utente

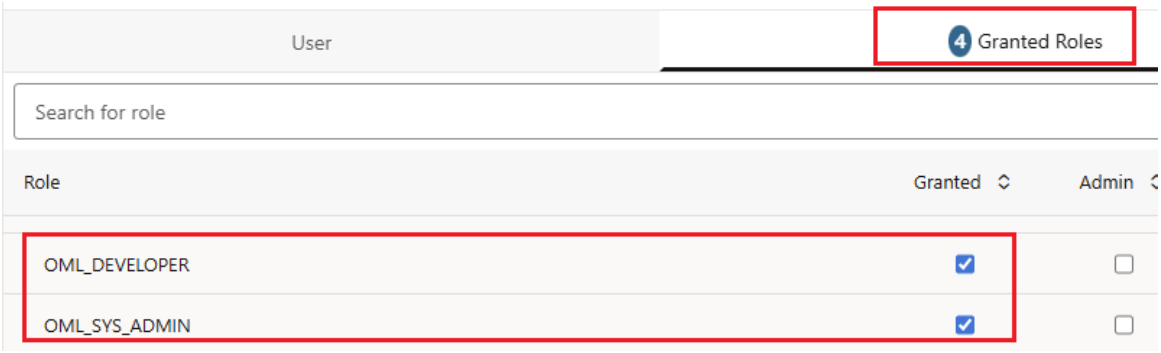

Figura 4.13 Definizione dei permessi degli utenti

Caricamento dei dati

Una preparazione adeguata dei dati è essenziale per il corretto funzionamento del data warehouse. Dopo aver configurato l'istanza di Autonomous Database, i dati vengono caricati attraverso l'interfaccia Data Load, accessibile da Database Actions (Figure 4.14 e 4.15).

Il modo più semplice per importare dati consiste nel caricare file locali mediante trascinamento (drag and drop) nell'ambiente Data Load. Il sistema riconosce automaticamente la struttura, suggerisce nomi delle colonne e tipi di dati e permette di creare una nuova tabella. Questo metodo funziona bene per dataset di piccole dimensioni in formato CSV, JSON o simili e non richiede strumenti aggiuntivi.

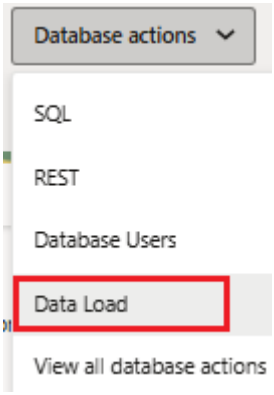

Figura 4.14 Avvio di Data Load

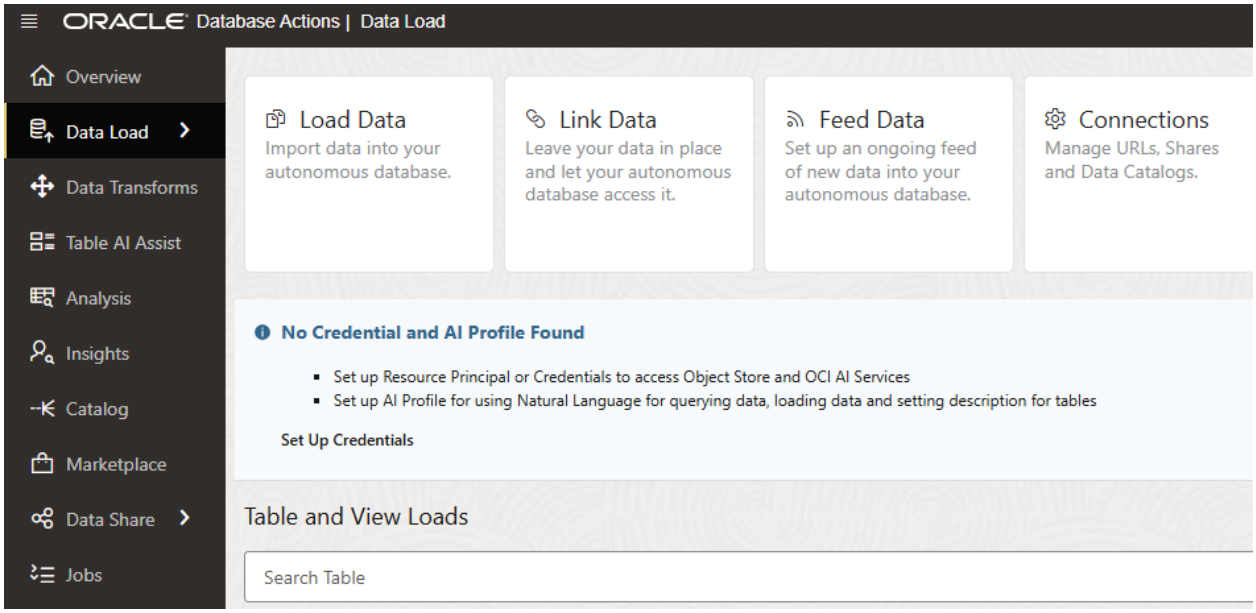

Figura 4.15 Interfaccia Data Load

Per i file di grandi dimensioni, soprattutto in formato Parquet, si consiglia di caricarli preliminarmente nel servizio Oracle Object Storage di Oracle Cloud Infrastructure. Nel repository i dati vengono conservati in contenitori chiamati bucket; nella modalità permanentemente gratuita sono disponibili fino a 20 GB di spazio, sufficienti per molte applicazioni analitiche.

Un bucket viene creato dalla sezione Storage – Object Storage, impostandone il nome (Figure 4.16 e 4.17). Successivamente i file preparati possono essere caricati e scaricati, per poi essere importati nel data warehouse (Figura 4.18).

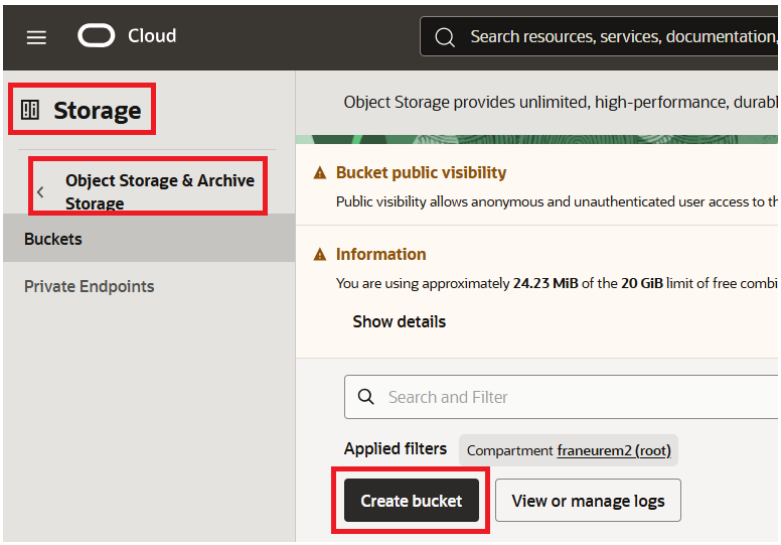

Figura 4.16 Creazione di un bucket Oracle Object Storage

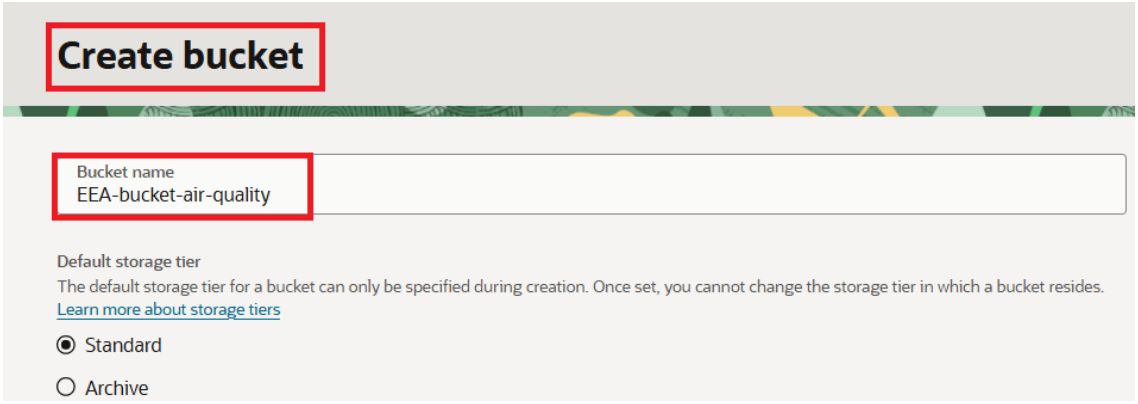

Figura 4.17 Creazione di un bucket Oracle Object Storage

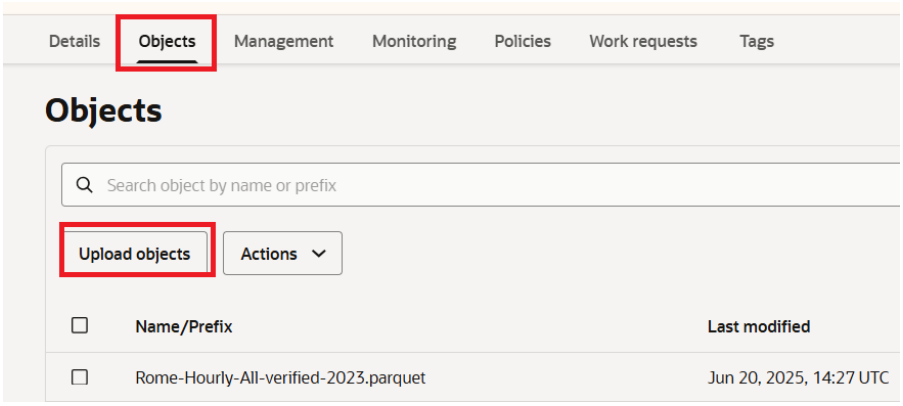

Figura 4.18 Caricamento di file in un bucket

Quando i file sono presenti nel bucket, l'importazione nel database viene avviata dalla stessa interfaccia Data Load. Al posto di un file locale, si seleziona la sorgente Cloud Store (Figura 4.19)

e si inserisce l'URL dell'oggetto presente nel bucket. Il sistema associa automaticamente i campi e crea la tabella corrispondente.

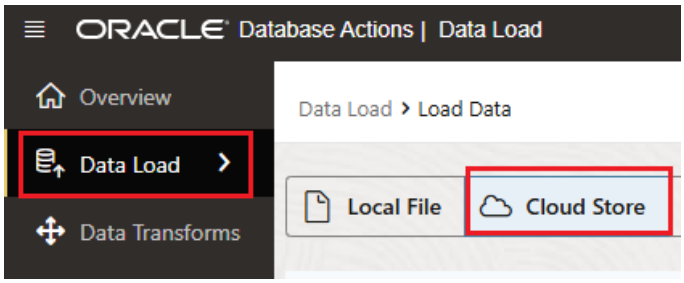

Figura 4.19 Caricamento da Object Storage con il modulo Data Load

Modellazione dei dati

Dopo il caricamento, il passo successivo consiste nel modellare i dati in funzione delle analisi future. È importante progettare tabelle chiare e comprensibili per garantire l'esecuzione efficiente delle query SQL.

Inizialmente le tabelle vengono in genere create direttamente dai file caricati. Ogni dataset, come le misurazioni della qualità dell'aria o i dati meteorologici, dovrebbe essere organizzato in una tabella distinta, in cui ogni riga rappresenta una singola misurazione. Queste sono spesso chiamate tabelle grezze (raw tables), perché conservano la struttura dei dati sorgente senza ulteriori trasformazioni. Pur consentendo analisi di base, possono essere successivamente riorganizzate per rendere le interrogazioni più semplici e affidabili.

Per analisi più complesse è utile standardizzare preventivamente i nomi delle colonne, i formati delle date e gli identificatori delle località o delle stazioni di misura. Una struttura coerente facilita l'unione delle tabelle e riduce il rischio di errori interpretativi. In alcuni casi si può ricorrere a una semplice denormalizzazione, aggiungendo direttamente alla tabella delle misure informazioni sulla stazione provenienti da una tabella ausiliaria, così da semplificare query che altrimenti richiederebbero più join.

Uno dei modi più comuni e concettualmente semplici di organizzare i dati è il cosiddetto schema a stella (star schema), illustrato nella Figura 4.20. Esso prevede una tabella centrale, ad esempio quella delle misurazioni, come fonte principale dei dati, mentre le informazioni descrittive su luoghi, stazioni e categorie vengono conservate in tabelle più piccole, dette tabelle delle dimensioni. Queste possono essere collegate alla tabella centrale quando necessario, semplificando la scrittura delle query analitiche.

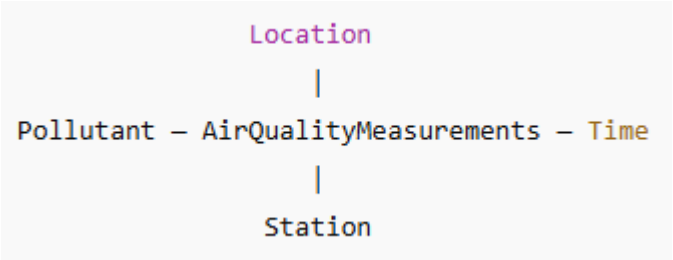

Figura 4.20 Semplice schema a stella

Visualizzazione dei dati

Dopo aver eseguito una query SQL è possibile rappresentarne i risultati mediante grafici, usando il modulo integrato Charts. Per accedervi, selezionare View all database actions, quindi la scheda Data Studio, che contiene Charts, come illustrato nella Figura 4.21. Lo strumento consente di creare semplici visualizzazioni direttamente dai risultati SQL, senza applicazioni esterne né installazioni aggiuntive. Permette di interpretare rapidamente i dati, in particolare per individuare tendenze, confrontare dataset e riconoscere anomalie.

Si consideri, ad esempio, una tabella di misurazioni orarie della concentrazione di una specifica massa d'aria nell'arco di un periodo prolungato. In questo caso può essere utile concentrarsi sulle misurazioni di un singolo giorno. L'esempio utilizza un grafico a barre: sull'asse X sono riportati gli orari delle misurazioni della giornata selezionata, mentre sull'asse Y compaiono i valori corrispondenti (Figura 4.22).

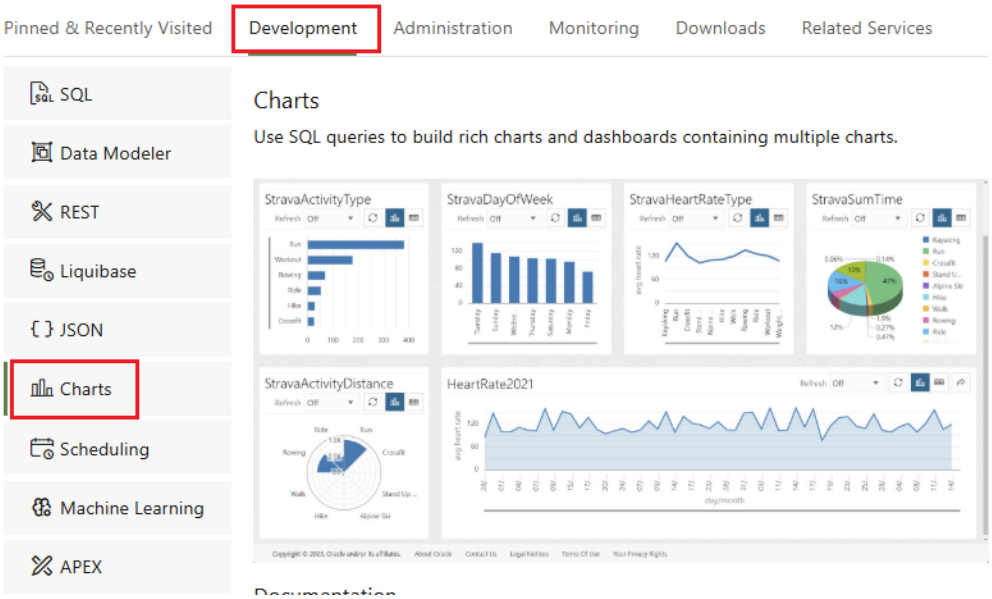

Figura 4.21 Avvio del modulo Charts

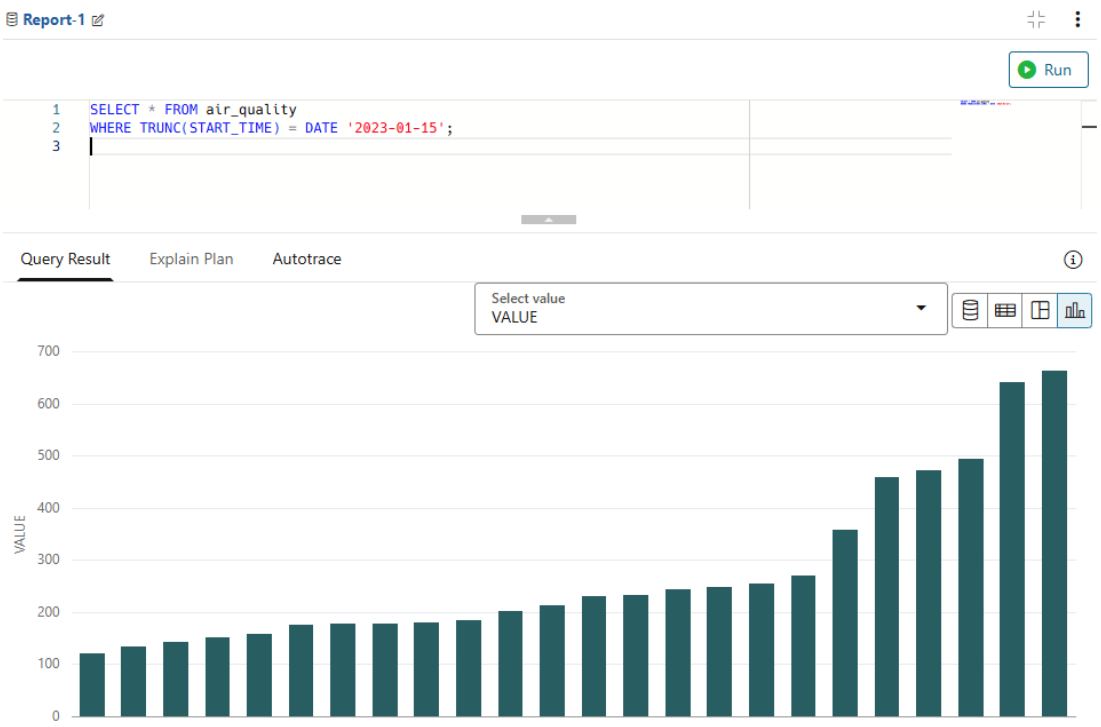

Figura 4.22 Semplice visualizzazione dei dati

Machine learning all'interno del database

Un data warehouse basato su Oracle Autonomous Database comprende Oracle Machine Learning (OML), un sistema di machine learning integrato che consente di addestrare e distribuire modelli direttamente sui dati memorizzati nel database. A differenza degli approcci che esportano i dati verso strumenti esterni, OML permette di prepararli, modellizzarli e valutarli all'interno del database, evitando infrastrutture aggiuntive e trasferimenti di grandi volumi di dati.

Per aprire OML, selezionare Database Actions – View all database actions, quindi la scheda Development e il menu Machine Learning (Figura 4.23). Si apre l'ambiente di lavoro OML, illustrato nella Figura 4.24.

Il modo più semplice di utilizzare OML è tramite OML Notebooks (Figura 4.24), che consente di scrivere ed eseguire script SQL e Python direttamente nel browser. Offre funzionalità di classificazione, regressione, clustering, rilevamento delle anomalie e addestramento automatico dei modelli.

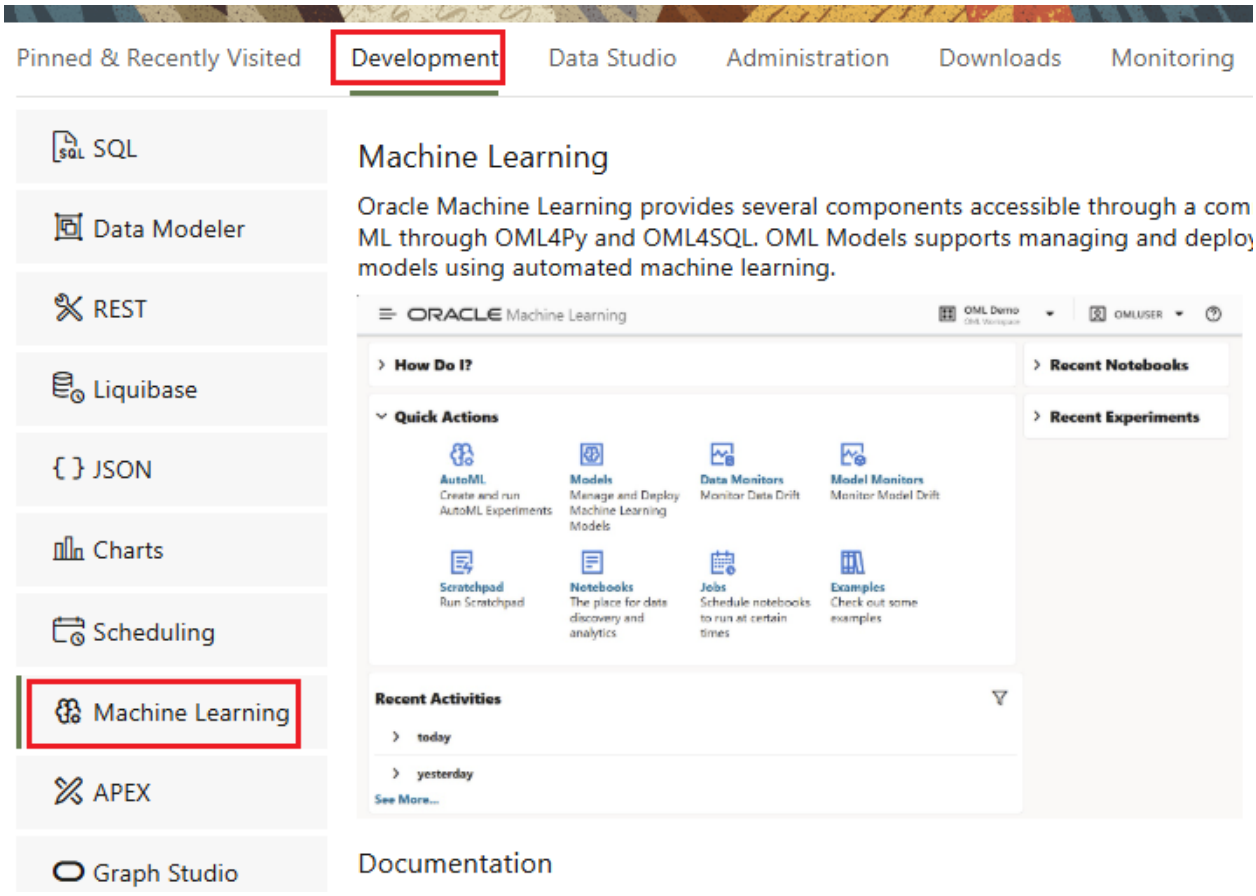

Figura 4.23 Avvio di OML

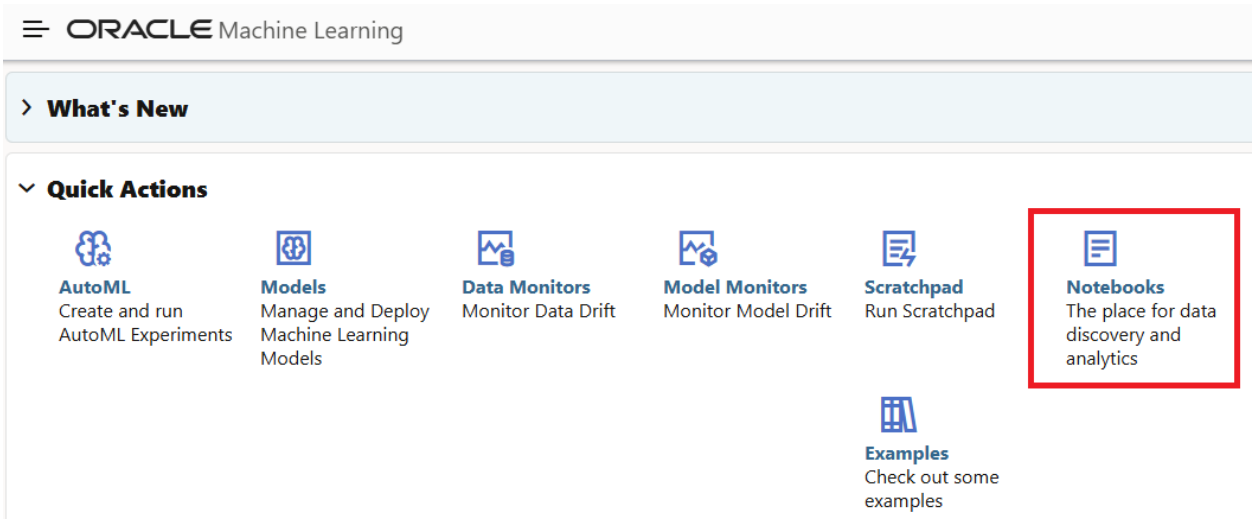

Figura 4.24 Interfaccia OML

Caso di studio: analisi dell'inquinamento atmosferico

Questo caso di studio presenta dati reali sulla qualità dell'aria utilizzati da agenzie ambientali europee, istituti di ricerca e sistemi di analisi. Le misurazioni orarie sono fornite dall'Agenzia europea dell'ambiente (EEA).

L'esempio illustra l'intero processo di lavoro con dataset ambientali. La prima parte riguarda le modalità di raccolta e conservazione dei dati nel data warehouse; le fasi successive comprendono analisi di fenomeni ambientali, come il monitoraggio del biossido di azoto (NO_2) o del particolato ($PM_{2,5}$).

Fonti europee di dati sulla qualità dell'aria

I dati sulla qualità dell'aria si possono scaricare dal portale EEA [10], che raccoglie, verifica e pubblica le misurazioni delle reti nazionali di monitoraggio ambientale. Il sistema EEA permette di accedere a serie dettagliate di concentrazioni orarie, giornaliere e annuali di inquinanti come biossido di azoto (NO_2), ozono, PM2.5 e altre sostanze, essenziali per valutare la qualità dell'aria.

Le stazioni di misura sono distribuite in tutta Europa e adottano metodi standardizzati conformi alle direttive europee sul monitoraggio della qualità dell'aria. Prima della pubblicazione, i dati vengono sottoposti a validazione per eliminare errori, incoerenze logiche e record mancanti. I dati elaborati sono disponibili tramite il servizio ufficiale EEA [11], anche nei formati Parquet e CSV.

Dopo aver scelto l'area geografica (Figura 4.25), il periodo e l'intervallo temporale delle misurazioni, il sistema genera un pacchetto con i dati richiesti. Il file scaricato contiene generalmente più record Parquet, uno per ciascuna stazione dell'area selezionata. Questa organizzazione consente analisi dettagliate; tuttavia, nelle attività analitiche può essere utile riunire i record in una sola tabella. A tale scopo si possono usare strumenti di unione di file Parquet, come Agents for Data [12], che permettono di trascinare e combinare più file in un unico dataset (Figura 4.26).

È inoltre importante comprendere la struttura dei dati scaricati, il significato dei loro attributi e i collegamenti alla documentazione pubblicata: metadati, dizionari dei dati, vocabolari e documentazione tecnica delle API.

.

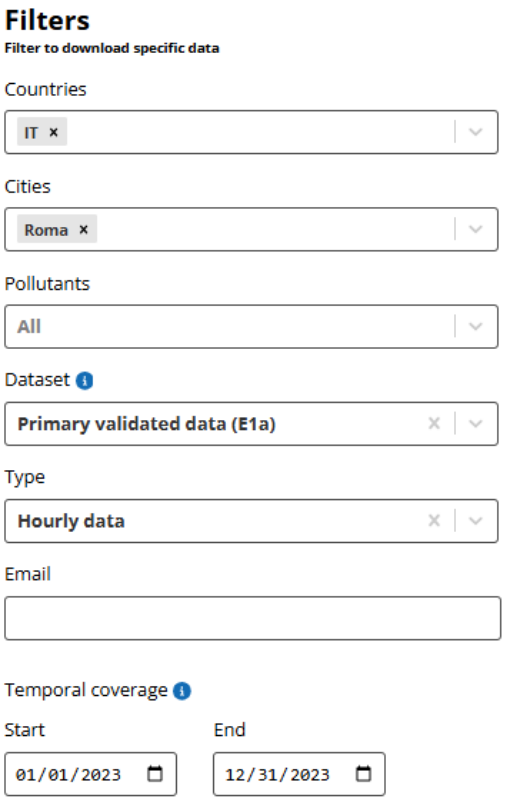

Figura 4.25 Download dei dati EEA

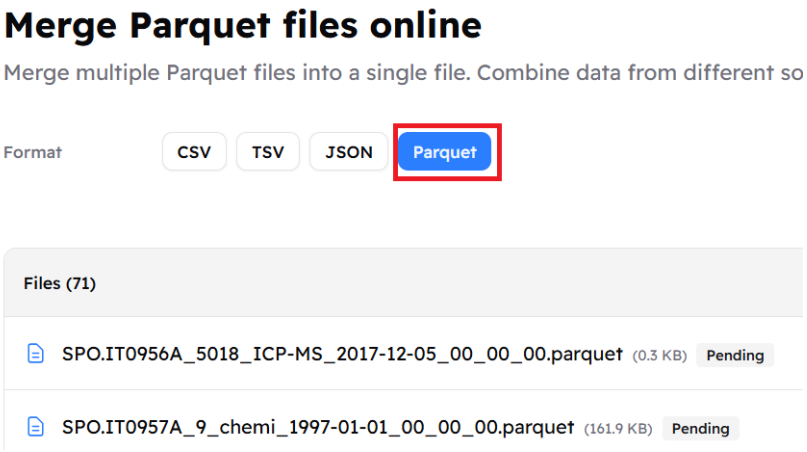

Figura 4.26 Unione di più file Parquet sulla piattaforma Agents for Data

Archiviazione delle misurazioni EEA in Oracle Object Storage

Dopo aver scaricato i dati EEA, il passo successivo consiste nel salvarli in un contenitore preconfigurato del servizio Oracle Object Storage. All'interno del bucket selezionare Objects, fare clic su Upload Objects e caricare il file Parquet delle misurazioni (Figura 4.27).

Ogni oggetto caricato riceve un proprio URL, visualizzabile nelle relative proprietà (Figura 4.28). Questo indirizzo verrà usato successivamente per importare i dati nel data warehouse.

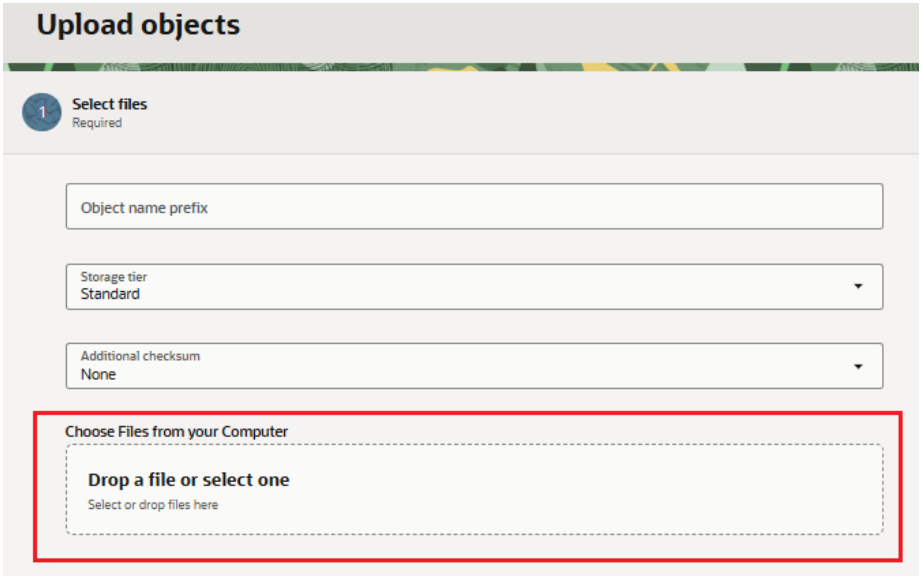

Figura 4.27 Caricamento del file EEA in un bucket

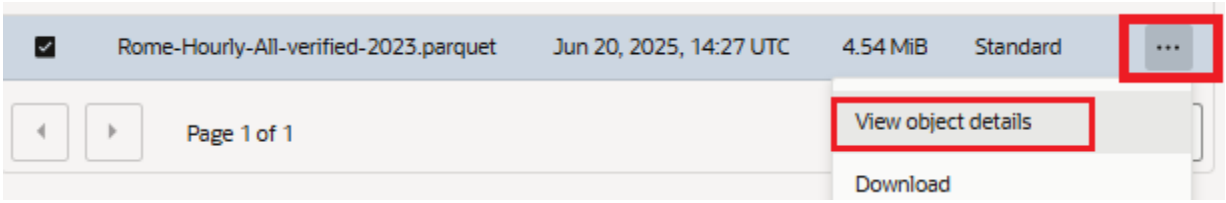

Figura 4.28 Generazione degli URL degli oggetti

Importazione delle misurazioni nel data warehouse

Aprire quindi l'interfaccia Data Load di Autonomous Database e selezionare la scheda Cloud Store. Nel campo della sorgente dati inserire l'URL, copiato in precedenza, dell'oggetto Parquet memorizzato nel bucket (Figura 4.29).

Nel pannello di sinistra viene mostrata la struttura dei dati disponibili, che possono essere selezionati e trascinati nell'area di lavoro per l'importazione. Il caricamento inizia facendo clic sul pulsante Start (Figura 4.30).

Prima di importare si consiglia di aprire Settings per verificare la struttura del file e impostare le opzioni principali: creare una nuova tabella o aggiornare una tabella esistente, associare le colonne e scegliere lo schema dell'utente (Figura 4.31).

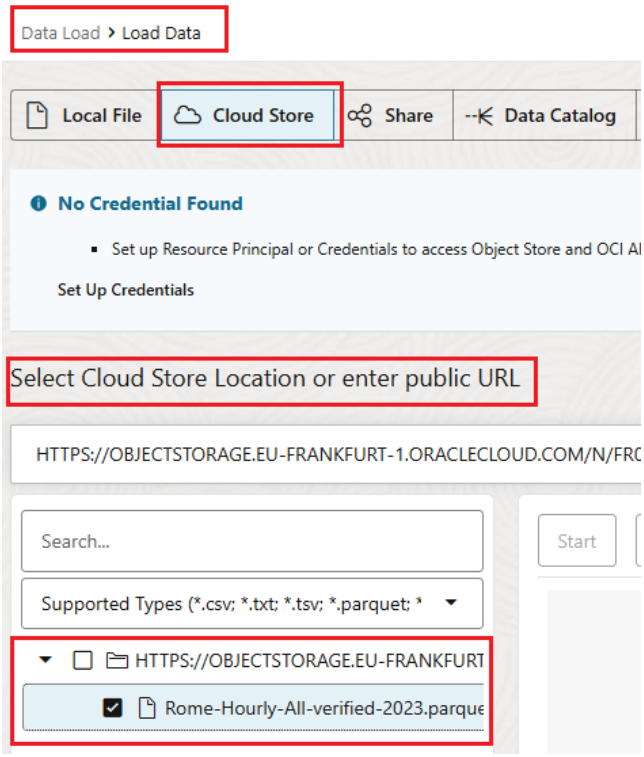

Figura 4.29 Preparazione del caricamento delle misurazioni

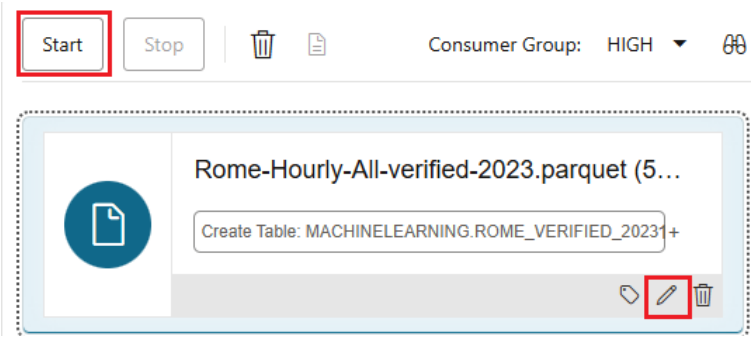

Figura 4.30 Preparazione del caricamento delle misurazioni

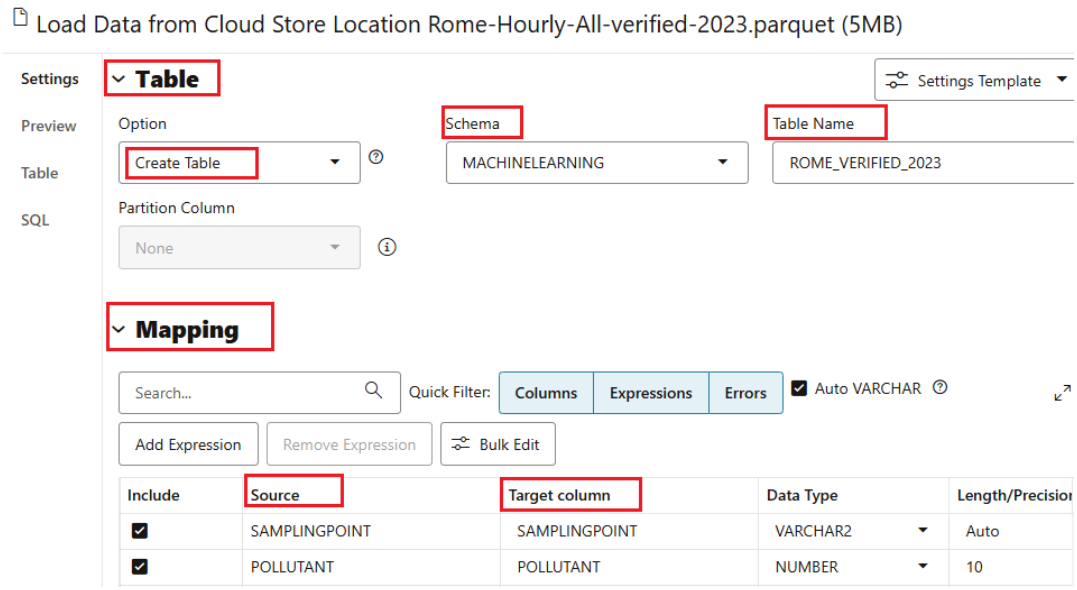

Figura 4.31 Impostazione dei parametri di trasferimento

Esempi di analisi preliminare dei dati

Di seguito sono riportate alcune interrogazioni di base che consentono di visualizzare il contenuto delle tabelle e comprenderne gli attributi disponibili. Questa esplorazione iniziale permette di chiarire la struttura dei dati, gli intervalli delle misurazioni e gli identificatori degli inquinanti e delle stazioni.

La seguente query SQL consente di visualizzare tutti gli inquinanti presenti nella tabella delle misurazioni:

```
-------------------------------------------------------------------------------
-- 1. List all pollutants that were measured
--    Each pollutant is identified by its numeric code.
-------------------------------------------------------------------------------
SELECT DISTINCT POLLUTANT
FROM ROME_VERIFIED_2023
ORDER BY POLLUTANT;
```

È inoltre utile compilare l’elenco completo delle stazioni di misura:

```
-------------------------------------------------------------------------------
-- 2. List all measurement stations (sampling points)
--    Each SAMPLINGPOINT represents a unique monitoring location.
-------------------------------------------------------------------------------
SELECT DISTINCT SAMPLINGPOINT
FROM ROME_VERIFIED_2023
ORDER BY SAMPLINGPOINT;
```

È possibile anche individuare gli intervalli temporali delle misurazioni per tutti gli inquinanti registrati:

```
-------------------------------------------------------------------------------
-- 3. Determine the measurement range for each pollutant
--    Shows the first and last measurement timestamp per pollutant.
-------------------------------------------------------------------------------
SELECT
    POLLUTANT,
    MIN(END_MEASUREMENT) AS FIRST_MEASUREMENT,
    MAX(END_MEASUREMENT) AS LAST_MEASUREMENT
FROM ROME_VERIFIED_2023
GROUP BY POLLUTANT
ORDER BY POLLUTANT;
```

Analogamente alla query precedente, è possibile visualizzare gli intervalli temporali delle misurazioni raggruppati per stazione.

```
-------------------------------------------------------------------------------
-- 4. Determine the measurement range for each sampling point
--    Shows when each station started and stopped reporting measurements.
-------------------------------------------------------------------------------
SELECT
    SAMPLINGPOINT,
    MIN(END_MEASUREMENT) AS FIRST_MEASUREMENT,
    MAX(END_MEASUREMENT) AS LAST_MEASUREMENT
FROM ROME_VERIFIED_2023
GROUP BY SAMPLINGPOINT
ORDER BY SAMPLINGPOINT;
```

Collegamento delle misurazioni alle descrizioni degli inquinanti

Purtroppo, dalla sola tabella delle misurazioni di qualità dell'aria (ROME_VERIFIED_2023) non è possibile sapere a quali inquinanti corrispondano i dati, perché sono riportati soltanto codici numerici. Per risolvere il problema si collega questa tabella a un'altra contenente informazioni dettagliate sugli inquinanti e sulle loro denominazioni. I dati necessari sono disponibili nel vocabolario EEA [13]. L'elenco degli inquinanti può essere copiato in un file Excel e caricato nel data warehouse attraverso Data Load. In questo esempio i dati vengono importati nella tabella POLLUTANTS, come illustrato nella Figura 4.32.

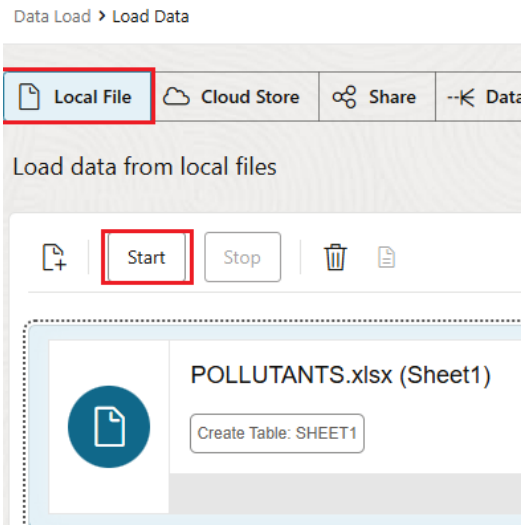

Figura 4.32 Importazione dei codici degli inquinanti atmosferici

È ora possibile collegare misurazioni e tipi di inquinanti in un'unica query SQL.

```
SELECT
    DISTINCT m.POLLUTANT AS POLLUTANT_CODE,    -- numeric code as stored in
measurements
    p.LABEL,                                    -- official pollutant name
    p.NOTATION,                                 -- notation (e.g., SO2, PM10)
    p.MANDATORYUNIT                            -- recommended measurement unit
FROM ROME_VERIFIED_2023 m
LEFT JOIN POLLUTANTS p
       ON m.POLLUTANT = p.POLLUTANT_ID
ORDER BY POLLUTANT_CODE;
```

È inoltre possibile calcolare la concentrazione media annua di NO_2 per tutte le stazioni di misura.

```
-------------------------------------------------------------------------------
-- Annual NO2 averages (pollutant = 8), excluding negative values.
-- Joined with POLLUTANTS for pollutant name, symbol and unit.
-------------------------------------------------------------------------------
SELECT m.SAMPLINGPOINT, m.POLLUTANT AS POLLUTANT_CODE, p.LABEL AS POLLUTANT_NAME,
       p.NOTATION AS POLLUTANT_SYMBOL, p.MANDATORYUNIT AS UNIT,
       EXTRACT(YEAR FROM m.END_MEASUREMENT) AS YEAR, AVG(m.VALUE) AS AVG_VALUE,
       COUNT(*) AS COUNT_MEASUREMENTS, MIN(m.END_MEASUREMENT) AS FIRST_MEASUREMENT,
       MAX(m.END_MEASUREMENT) AS LAST_MEASUREMENT
FROM ROME_VERIFIED_2023 m
LEFT JOIN POLLUTANTS p ON m.POLLUTANT = p.POLLUTANT_ID
WHERE m.POLLUTANT = 8          -- filter for NO2
  AND m.VALUE >= 0             -- exclude negative measurements
GROUP BY m.SAMPLINGPOINT, m.POLLUTANT, p.LABEL, p.NOTATION, p.MANDATORYUNIT,
         EXTRACT(YEAR FROM m.END_MEASUREMENT)
ORDER BY m.SAMPLINGPOINT, YEAR;
```

L'esecuzione della query precedente indica che la stazione identificata come IT/SPO".IT1837A_8_chemi_2006-12-17_00:00:00 registra una concentrazione media annua di NO_2 di 47,074 µg/m³, superiore al limite di 40,00 µg/m³ e potenzialmente dannosa per la salute.

Collegamento delle misurazioni alle località delle stazioni

Dopo aver verificato che alcune stazioni registrano superamenti dei livelli di NO_2, occorre ottenere informazioni aggiuntive sulla loro posizione. La tabella ROME_VERIFIED_2023 contiene soltanto gli identificatori delle stazioni, quindi non permette di stabilirne direttamente la localizzazione esatta. Si aggiunge pertanto al data warehouse una tabella con dati descrittivi delle stazioni, fra cui nomi, coordinate geografiche e tipologie.

Questi dati provengono dal portale EEA [10], che offre un elenco completo delle stazioni europee. L'elenco si scarica in formato CSV selezionando Metadata e poi Download CSV (Figura 4.29). Il file viene successivamente caricato nel data warehouse tramite Data Load (Figura 4.30), creando una nuova tabella, qui denominata SAMPLING_POINTS, con le informazioni rilevanti sulle stazioni di misura (Figura 4.31).

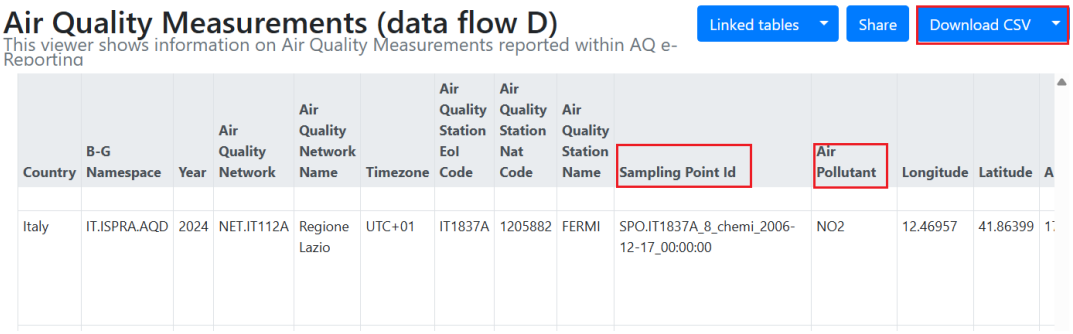

Figura 4.29 Download dei dati delle stazioni di misura

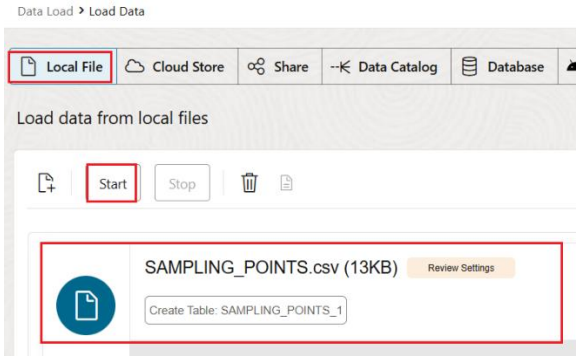

Figura 4.30 Caricamento dei dati delle stazioni di misura

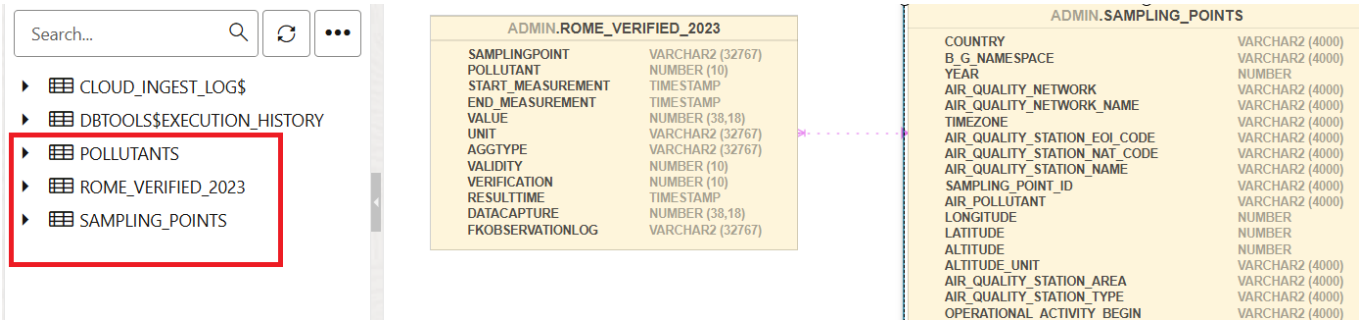

Figura 4.31 Tabelle sull'inquinamento atmosferico

Dall'analisi delle tabelle importate risulta che ROME_VERIFIED_2023 può essere collegata a POLLUTANTS tramite la colonna POLLUTANT della prima tabella e la colonna POLLUTANT_ID della seconda.

Inoltre, ROME_VERIFIED_2023 può essere collegata a SAMPLING_POINTS mediante SAMPLINGPOINT in ROME_VERIFIED_2023 e SAMPLING_POINT_ID in SAMPLING_POINTS.

Tuttavia, si presenta un problema: i valori della colonna SAMPLINGPOINT di ROME_VERIFIED_2023 non coincidono direttamente con quelli della tabella SAMPLING_POINTS. Ad esempio, in ROME_VERIFIED_2023 un identificatore può contenere il prefisso IT/, mentre lo stesso identificatore nella tabella SAMPLING_POINTS compare senza tale prefisso.

In altre parole, il record di ROME_VERIFIED_2023 contiene un prefisso nazionale (IT/) che manca in SAMPLING_POINTS. Poiché una JOIN SQL richiede la corrispondenza esatta dei valori, le due tabelle non possono essere collegate direttamente.

Il problema si risolve usando la funzione SUBSTR(SAMPLINGPOINT, 4) nella clausola JOIN: questa elimina i primi tre caratteri (IT/) dal valore presente in ROME_VERIFIED_2023. I valori diventano così confrontabili, consentendo il collegamento corretto tra SAMPLINGPOINT e SAMPLING_POINT_ID e l'unione delle tre tabelle in un'unica interrogazione.

```
SELECT
    -- measurement data
    r.start_measurement, r.end_measurement, r.value, r.unit,

    -- pollutant information
    p.label AS pollutant_label,

    -- sampling point location data
    sp. country, sp.municipality, sp.air_quality_station_name,
    sp.air_quality_station_area, sp.air_quality_station_type,
    sp.latitude, sp.longitude
```

```
FROM rome_verified_2023 r
JOIN pollutants p
      ON p.pollutant_id = r.pollutant
JOIN sampling_points sp
      ON sp.sampling_point_id = SUBSTR(r.samplingpoint, 4)
WHERE r.value >= 0;
```

Si può quindi completare una query aggregata che collega le tre tabelle e calcola le concentrazioni medie annuali di NO_2 (codice 8). La query mostra contemporaneamente le misurazioni di inquinamento e gli attributi geografici delle stazioni, facilitando l'interpretazione dei risultati.

```
SELECT
    -- pollutant info
    p.label AS pollutant_label,

    -- sampling point (location data)
    sp.country, sp.municipality, sp.air_quality_station_name,
    sp.air_quality_station_area, sp.air_quality_station_type,
    sp.latitude, sp.longitude,

    -- year and aggregated value
    EXTRACT(YEAR FROM r.start_measurement) AS measurement_year,
    AVG(r.value) AS avg_pollutant_value

FROM rome_verified_2023 r
JOIN pollutants p ON p.pollutant_id = r.pollutant
JOIN sampling_points sp ON sp.sampling_point_id = SUBSTR(r.samplingpoint, 4)

WHERE r.value >= 0
  AND r.pollutant = 8

GROUP BY
    p.label,
    sp.country, sp.municipality, sp.air_quality_station_name,
    sp.air_quality_station_area, sp.air_quality_station_type,
    sp.latitude, sp.longitude,
    EXTRACT(YEAR FROM r.start_measurement)

ORDER BY measurement_year, sp.country, sp.municipality;
```

I risultati della query precedente sono riportati nella Tabella 1. Il valore limite della concentrazione media annuale di NO_2 è pari a 40 µg/m³ (microgrammi per metro cubo); l'analisi riguarda, tra le altre, la stazione Fermi di Roma, situata in piazza Enrico Fermi. Le coordinate della stazione (41,86399° N, 12,46957° E) permettono di visualizzarla con Google Maps e Street View (Figura 4.32).

Tabella 1: panoramica delle concentrazioni medie annuali di NO_2 per stazione di misura a Roma (2023)

POLLUTANT	COUNTRY	MUNICIPALITY	AIR QUALITY STATION	AREA	AIR QUALITY STATION TYPE	LATITUDE	LONGITUDE	YEAR	VALUE
Nitrogen (air)	Italy	CORSO FRANCIA - Roma (RM)	C.SO FRANC	urban	car	41,94745	12,46956	2023	35,61
Nitrogen (air)	Italy	LARGO MAGNA GRECIA - Roma (RM)	L.GO MAGNA GRECIA	urban	car	41,88307	12,50895	2023	33,96
Nitrogen (air)	Italy	LARGO PERESTRELLO - Roma (RM)	L.GO PERESTRELLO	urban	background	41,88602	12,54162	2023	26,22
Nitrogen (air)	Italy	LGO BELOLLI - Roma (RM)	CINECITTA	urban	background	41,85772	12,56865	2023	25,77
Nitrogen (air)	**Italy**	**P.ZZA E. FERMI, FRONTE CIVICO16 - Roma (RM)**	**FERMS**	**urban**	**car**	**41,86399**	**12,46957**	**2023**	**47,07**
Nitrogen (air)	Italy	Piazza Cairoli adiacente via dei Giubbonari - Roma (RM)	ARENUL	urban	background	41,89397	12,47539	2023	33,00
Nitrogen (air)	Italy	VIA AURELIA KM 20 CASTEL DI GUIDO - Roma (RM)	CASTEL DI GUIDÓ	rural	background	41,88945	12,26633	2023	8,17
indicates that the NO_2 mean annual limit value is 40 µg/m³ (micrograms per cubic meter) at the Fermi measuring station located in	Italy	VIA DELLA MELORIA, FRONTE CIVICO 27 - Roma (RM)	CIPRO	urban	background	41,90636	12,44759	2023	30,04
Nitrogen (air)	Italy	VIA MONTE CANDA, ADIACENTE CIVICO 12 - Roma (RM)	BUFALOTTA	urban	background	41,94765	12,53367	2023	29,94
Nitrogen (air)	Italy	VIA PAOLO EMILIO PAVOLINI 69/71 - Roma (R)	MALAGROT	suburban	background	41,87478	12,34559	2023	16,90
Nitrogen (air)	Italy	VIA SALARIA - Roma (RM)	VILLA ADA	urban	background	41,93289	12,50695	2023	19,76
Nitrogen (air)	Italy	VIA TENUTA DEL CAVALIERE 102 - Roma (RM)	TENUTA DEL CAVALIER	suburban	background	41,92947	12,65848	2023	23,89
Nitrogen (air)	Italy	VIA TIBURTINA, ADIACENTE CIVICO 654 - Roma (RM)	TIBURNIA	urban	car	41,91027	12,54885	2023	35,50



Figura 4.32 Google Street View presso la stazione Fermi

Esercitazioni pratiche

- Accedere alla console Oracle Cloud.
- Creare una nuova istanza di Autonomous Database per realizzare il data warehouse.
- Creare un bucket in Oracle Object Storage.
- Accedere al portale EEA e consultare la documentazione e le descrizioni dei modelli di dati (collegamenti ai metadati e ai vocabolari).
- Scaricare il file Parquet delle misurazioni correnti dell'inquinamento atmosferico del proprio Paese e archiviarlo nel contenitore dati.
- Caricare i dati nel data warehouse (misurazioni dal bucket, elenco degli inquinanti predisposto nel file Excel POLLUTANTS e file CSV SAMPLING_POINTS).
- Analizzare le misurazioni del particolato PM2.5. Quali sono le fonti di questo inquinamento e quali effetti ha sulla salute?
- Individuare le località in cui sono stati registrati i livelli più elevati di PM2.5.
- Confrontare i valori misurati con i limiti previsti.

Valutazione

Quali sono le principali caratteristiche e i vantaggi dei data warehouse?

A. Consentono l'analisi di grandi quantità di dati storici
B. Sono ottimizzati per transazioni rapide in tempo reale
C. Unificano e collegano dati provenienti da più fonti
D. Sostituiscono le attività di reporting dei database transazionali

Per organizzare dati meteorologici in un data warehouse mediante un semplice schema a stella, qual è l'insieme minimo di tabelle necessario? (Una risposta corretta.)

A. Una tabella delle misurazioni e una tabella delle località di misura
B. Una tabella delle misurazioni, una dei sensori e una dell'operatore del sistema
C. Soltanto una tabella delle misurazioni, poiché contiene già tutte le informazioni
D. Tutte le precedenti

Quale proprietà descrive meglio le cosiddette tabelle grezze (raw tables) nel modello dei dati ambientali?

A. Rappresentano i dati sorgente provenienti da sistemi esterni, senza ulteriori trasformazioni
B. Contengono dati temporanei scaricati da sistemi esterni
C. Memorizzano dati provenienti da sistemi esterni, organizzati per accesso e analisi rapidi
D. Sono ottimizzate per eseguire query aggregate e calcolare statistiche

Quali passaggi descrivono meglio la modellazione del data warehouse nel caso di studio sui dati ambientali?

A. Scaricare i dati da fonti esterne
B. Esaminare e comprendere la struttura dei dati (metadati e vocabolari), individuando gli attributi principali e le dipendenze
C. Collegare le tabelle
D. Tutte le precedenti

Quali sono le principali funzionalità dell'interfaccia Data Load durante l'utilizzo di un data warehouse Oracle Autonomous Database?

A. Generazione automatica di query SQL per l'analisi
B. Importazione di dati da un bucket Object Storage
C. Gestione degli account utente e delle impostazioni di sicurezza
D. Ottimizzazione delle query SQL analitiche
E. Importazione di file locali

Riferimenti bibliografici e letture di approfondimento

[1] "Autonomous AI Database", Oracle. [Online]. Available: https://www.oracle.com/autonomous-database/ (accessed Dec. 9, 2025).

[2] "Oracle Cloud Infrastructure (OCI)", Oracle. [Online]. Available: https://www.oracle.com/cloud/ (accessed Dec. 9, 2025).

[3] "Amazon Redshift," AWS. [Online]. Available: https://aws.amazon.com/redshift/ (accessed Dec. 9, 2025).

[4] "BigQuery | AI data platform | Lakehouse | EDW", Google Cloud. [Online]. Available: https://cloud.google.com/bigquery (accessed Dec. 9, 2025).

[5] "Azure Synapse Analytics", Microsoft Azure. [Online]. Available: https://azure.microsoft.com/en-us/products/synapse-analytics (accessed Dec. 9, 2025).

[6] "Apache Hive", Apache Software Foundation. [Online]. Available: https://hive.apache.org/ (accessed Dec. 9, 2025).

[7] "Solutions Cloud Program", Oracle Academy. [Online]. Available: https://academy.oracle.com/en/solutions-cloud-program.html (accessed Dec. 9, 2025).

[8] "Oracle Cloud Free Tier", Oracle. [Online]. Available: https://www.oracle.com/cloud/free (accessed Dec. 9, 2025).

[9] "Oracle SQL Developer – Downloads". Available: https://www.oracle.com/database/sqldeveloper/technologies/download (accessed Dec. 9, 2025).

[10] European Environment Agency, Available: https://www.eea.europa.eu/en (accessed Dec. 9, 2025).

[11] European Environment Agency, Data Download Service, All Available: https://eeadmz1-downloads-webapp.azurewebsites.net (accessed Dec. 9, 2025).

[12] Agents for Data, "Parquet Merge Tool", Available: https://www.agentsfordata.com/parquet/merge (accessed Dec. 9, 2025).

[13] European Environment Information and Observation Network, Air Quality Pollutant Vocabulary, Available: https://dd.eionet.europa.eu/vocabulary/aq/pollutant (accessed Dec. 9, 2025).