Podatkovno skladiščenje za okoljske aplikacije

Učni izidi

Po zaključku te teme bodo študenti sposobni:

- Pojasniti koncept in vlogo shranjevanja podatkov pri uporabi okoljskih podatkov.
- Vzpostaviti podatkovno skladišče in upravljati njegove temeljne funkcije.
- Prenesti, poiskati in naložiti okoljske podatke iz dokumentiranih zunanjih virov.
- Uporabljati poizvedbe SQL v ustvarjenem podatkovnem skladišču okoljskih podatkov.

Uvod

Okoljski podatki se zbirajo z različnimi meritvami, od katerih vsaka zajema le en vidik kompleksnega ekološkega sistema. Nekateri podatkovni nizi se osredotočajo na fizikalne lastnosti zraka, drugi na meteorološke razmere, tretji pa na zdravje ekosistemov ali raven človekove dejavnosti. Čeprav vsak zagotavlja dragocene informacije, nobeden ne ponuja celovitega pregleda procesov, ki vplivajo na okolje. Zato se analitično delo z okoljskimi podatki skoraj vedno začne z njihovim povezovanjem in usklajevanjem v enoten prikaz stanja okolja.

Pri tem se pojavita dva glavna izziva: iskanje ustreznih podatkov, ki so pogosto razpršeni med več ustanovami, in njihovo shranjevanje v obliki, primerni za prihodnje analize. Podatki so lahko v različnih oblikah zapisa, časovnih okvirih in stopnjah obdelave, zato jih je treba povezati v sistem, ki omogoča zanesljivo združevanje, agregiranje in primerjavo. Prav to je naloga podatkovnega skladišča: združiti heterogene podatke v enotno strukturo. S tem zagotavlja stabilno analitično okolje, ki podpira kompleksne poizvedbe, analizo trendov in poglobljeno razlago.

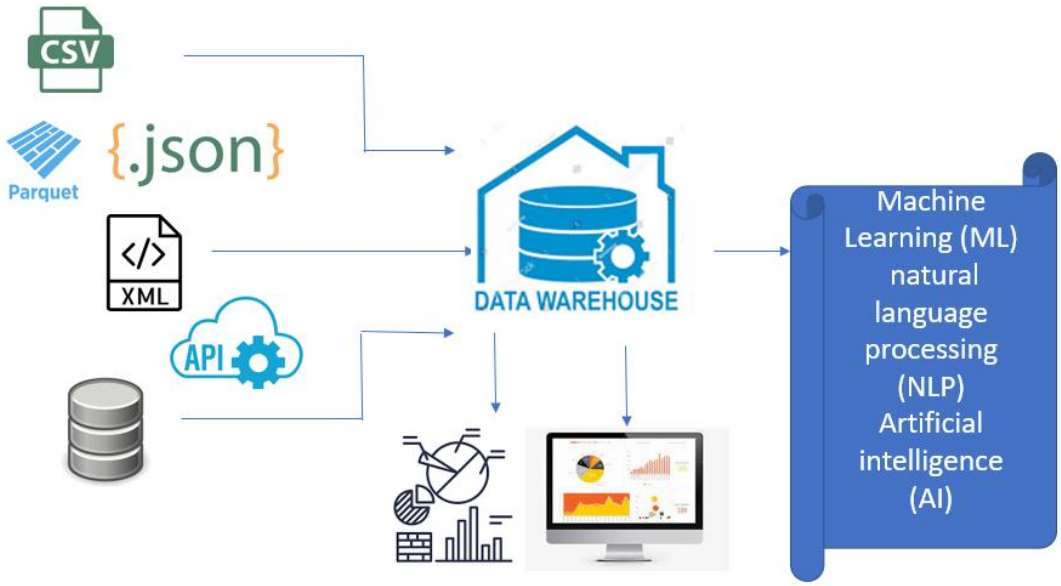
Podatkovno skladišče se po namenu in uporabi razlikuje od običajnih transakcijskih podatkovnih zbirk. Medtem ko so transakcijski sistemi optimizirani za hiter in pogost vnos podatkov, so podatkovna skladišča namenjena dolgoročnemu shranjevanju in analiziranju velikih zbirk preteklih podatkov. Poudarek je na analitičnih poizvedbah, analizi časovnih trendov in povezovanju raznovrstnih podatkovnih virov, ki pomagajo razumeti kompleksne pojave in njihove medsebojne povezave.

V okoljskih znanostih je ta pristop bistven, saj posamezne meritve le redko zajamejo ekološke spremembe. Kakovost zraka je denimo posledica kombinacije meteoroloških razmer, virov emisij, geografskih značilnosti in sezonskih sprememb. Le če te podatke opazujemo skupaj v

daljšem obdobju, lahko natančno ocenimo napredek na področju okolja, učinkovitost izvedenih ukrepov in morebitne nevarnosti.

Slika 4.1 prikazuje poenostavljeno arhitekturo podatkovnega skladišča, ki služi kot osrednje vozlišče za različne oblike podatkov, od običajnih datotek CSV in polstrukturiranih zapisov JSON/XML do formata Parquet ter podatkov, pridobljenih prek vmesnikov API ali iz podatkovnih zbirk drugih sistemov. Vsi ti podatkovni viri so združeni v enoten sistem, v katerem so podatki organizirani, povezani in pripravljeni za analizo. Tako podatkovno skladišče predstavlja zanesljivo infrastrukturo za raziskovalce, ustanove in odločevalce ter podpira stalne nadgradnje in razvoj politik, ki temeljijo na dokazih.

Ko so podatki naloženi v skladišče, omogočajo različne analize, od osnovnih statističnih povzetkov in vizualizacij do raziskovanja povezav med spremenljivkami ter uporabe naprednih metod strojnega učenja in umetne inteligence (AI). Na okoljskem področju lahko to pomaga pri napovedovanju dogodkov onesnaženja, ocenjevanju vpliva vremenskih razmer na kakovost zraka in analiziranju dolgoročnih trendov ekoloških procesov.



Slika 4.1 Poenostavljen pregled arhitekture podatkovnega skladišča

Tehnološki okvir

Danes je za izgradnjo podatkovnih skladišč mogoče uporabiti širok nabor tehnologij, ki se razlikujejo po razširljivosti, zahtevnosti upravljanja, funkcionalnostih in skupnih stroških. Izbira ustrezne tehnologije je odvisna od konkretnih potreb uporabnikov, razpoložljivih virov, vrst

podatkov in načrtovanih analitičnih metod. Na eni strani so lokalne namestitve običajnih relacijskih podatkovnih zbirk, ki zagotavljajo popoln nadzor, vendar zahtevajo obsežno tehnično podporo in stalno vzdrževanje. Na drugi strani so komercialne platforme v oblaku, ki zagotavljajo visoko stopnjo avtomatizacije, zmogljivo delovanje in možnost obdelave velikih podatkovnih nizov, vendar imajo različne obračunske modele in različno zahtevnost upravljanja.

Za to izdajo je bila izbrana Oracle Autonomous Database. Oracle Autonomous Database [1] omogoča ustvarjanje podatkovnih skladišč v stabilnem, zanesljivem in visoko avtomatiziranem okolju. Kot del Oracle Cloud [2] je vključena med stalno brezplačne vire, zato je začetnikom dostopna brez dodatnih stroškov infrastrukture. Vgrajeni sistemi za samodejno upravljanje varnosti, konfiguracije in nadgradenj odpravljajo tehnične ovire, ki bi sicer lahko ovirale manj izkušene uporabnike, ter jim omogočajo, da se osredotočijo na analizo in razlago podatkov.

Seveda uporaba te platforme ne izključuje drugih rešitev. Načela iz tega besedila veljajo tudi za različne tehnologije, zato lahko bralci glede na svoje potrebe in vire raziščejo možnosti, kot so Amazon Redshift [3], Google BigQuery [4], Microsoft Azure Synapse Analytics [5] ali odprtokodni sistemi, kot je Apache Hive [6]. Vsaka od teh tehnologij ponuja svojstven pristop k shranjevanju podatkov, različne modele prilagajanja zmogljivosti in lastne prednosti. Temeljni koncepti povezovanja podatkov in analitične obdelave pa ostajajo enaki.

Ustvarjanje in nastavitev podatkovnega skladišča

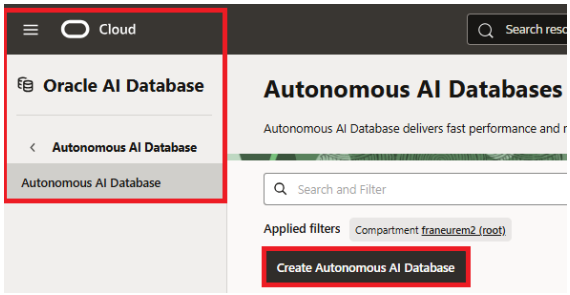
Postopek namestitve podatkovnega skladišča Oracle Autonomous Database se začne s prijavo v račun Oracle Cloud. Brezplačen račun lahko ustvarite v okviru programa Oracle Academy [7] ali sledite navodilom v viru [8]. Po prijavi izberite meni Oracle AI Database in nato možnosti Autonomous AI Database, kot prikazuje slika 4.2. Prikaže se pregled vseh podatkovnih zbirk v uporabnikovem računu.

Za vzpostavitev nove podatkovne zbirke kliknite povezavo Create Autonomous Database. V obrazec vnesite osnovne podatke o primerku, med drugim ime podatkovne zbirke, uporabniški račun v oblaku, ki se uporablja za namestitev, in delovno okolje. Priporočljivo je izbrati okolje Lakehouse, ki je prilagojeno gradnji podatkovnih skladišč ter ponuja dodatna orodja za upravljanje podatkovne zbirke, izvajanje poizvedb SQL in uporabo metod strojnega učenja na podatkovnih nizih.

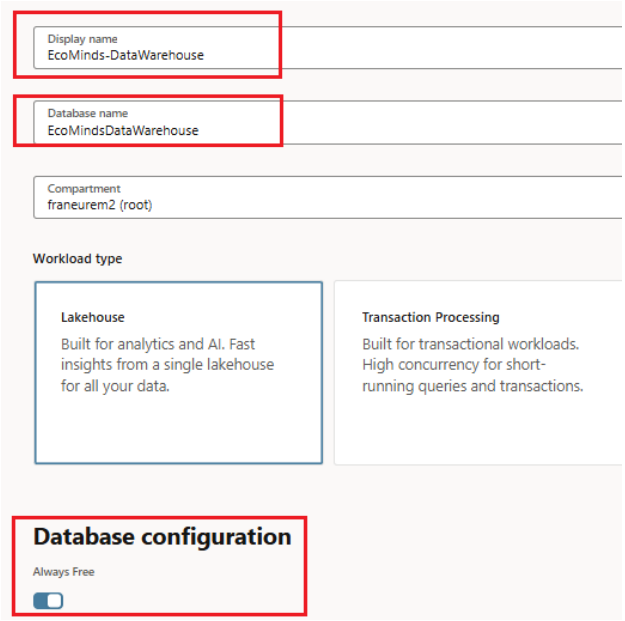
Nato omogočite stalno brezplačen dostop in izberite različico podatkovne zbirke. Najpogosteje uporabljena različica je privzeto vključena, na voljo pa je tudi nadgradnja na najnovejšo različico. Zatem določite poverilnice skrbnika tako, da nastavite geslo za

sistemskega skrbnika (uporabniško ime ADMIN), ki ima sprva popoln dostop do sistema ter je odgovoren za ustvarjanje novih uporabnikov, nalaganje podatkov in konfiguriranje podatkovne zbirke. Geslo mora izpolnjevati varnostne zahteve. V zadnjem koraku izberite način dostopa do omrežja primerka.

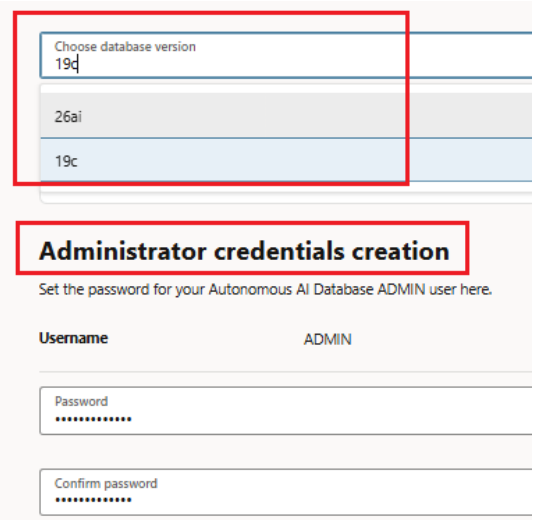
V stalno brezplačnem okolju se najpogosteje uporablja možnost varnega dostopa prek interneta s katere koli naprave. Dostop s katere koli lokacije je zavarovan z digitalno denarnico in mehanizmi za preverjanje pristnosti, vgrajenimi v Autonomous Database. Ta pristop močno poenostavi delo v izobraževalnih okoljih, saj odpravlja potrebo po namenskem omrežju ali dodatnih infrastrukturnih komponentah, kot so tuneli VPN ali omejitve naslovov IP.



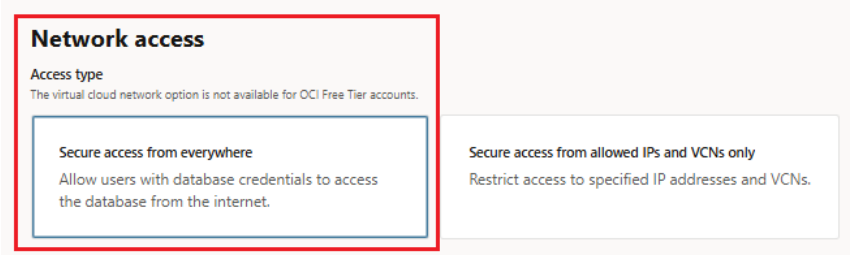
Slika 4.2 Dostop do platforme Oracle Autonomous Database



Slika 4.3 Ime podatkovnega skladišča in izbira delovne obremenitve



Slika 4.4 Končna konfiguracija podatkovnega skladišča



Slika 4.5 Izbira omrežnega dostopa

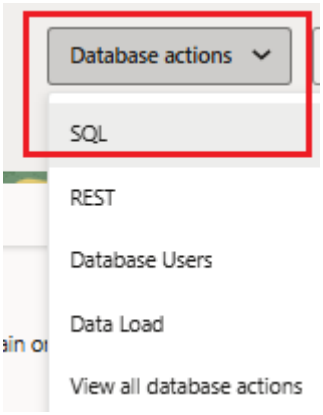
Preverjanje pristnosti in razvojno okolje

Vse analize okoljskih podatkov v podatkovnem skladišču so odvisne od jezika SQL, ki se uporablja za pridobivanje, filtriranje, agregiranje in združevanje podatkov v celovite analitične prikaze. Za nemoteno izvajanje teh poizvedb raziskovalci potrebujejo dostop do podatkovnega skladišča in zanesljivo razvojno okolje.

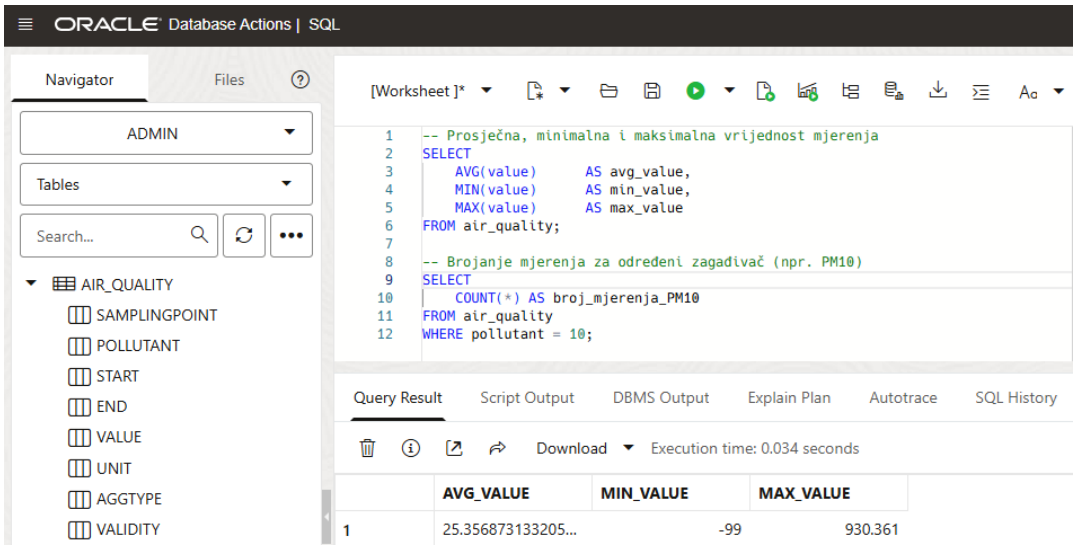
Obstajata dva glavna načina povezave z Oracle Autonomous Database. Prvi in najpreprostejši je uporaba spletnega vmesnika Oracle Cloud. Med prijavo je omogočeno dvostopenjsko preverjanje pristnosti (2FA), ki zagotovi dostop do vseh potrebnih virov. Po uspešni prijavi izberite Database Actions in nato SQL, da odprete integrirano spletno razvojno okolje za izvajanje stavkov SQL (sliki 4.6 in 4.7). Prednost tega načina je, da ne zahteva dodatne namestitve; ker je dostopen iz katere koli naprave in brskalnika, je posebej priročen za začetnike.

Drugi način dostopa je prek namizne aplikacije Oracle SQL Developer, ki jo je mogoče brezplačno prenesti z naslova [9]. SQL Developer omogoča napredno upravljanje skriptov,

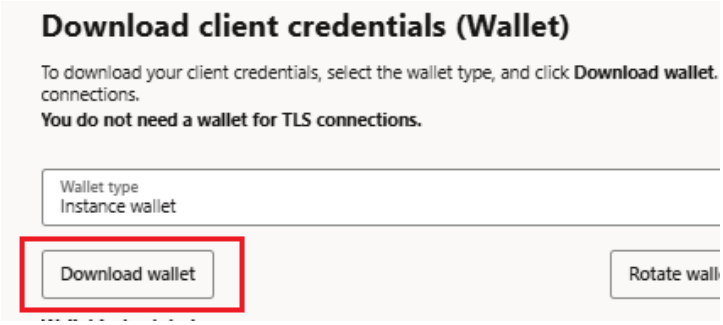
podpira zahtevnejše projekte, omogoča pregled strukture podatkovne zbirke in pomaga pri skrbniških opravilih. Za povezavo s to aplikacijo morate iz konzole Oracle Cloud prenesti digitalno denarnico, tako da kliknete Database Connection (slika 4.8). Paket vsebuje potrdila in konfiguracijske datoteke, potrebne za vzpostavitev varne povezave mTLS s podatkovnim skladiščem, s čimer je zagotovljeno vzajemno preverjanje pristnosti odjemalca in strežnika. Ko digitalno denarnico naložite v SQL Developer (slika 4.9), se lahko povežete s podatkovnim skladiščem in z njim delate v običajnem namiznem okolju (slika 4.10).



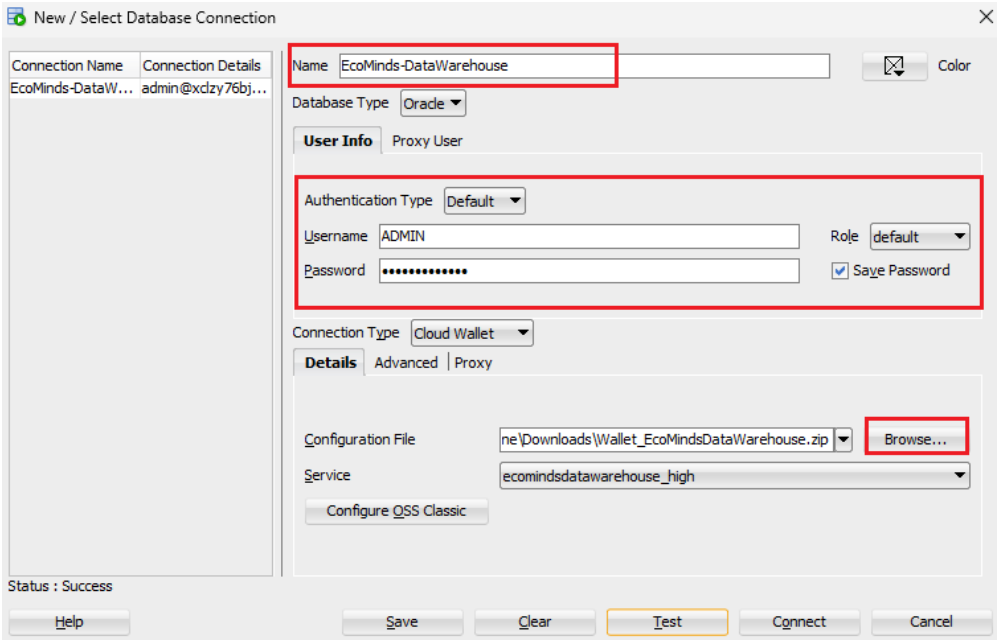
Slika 4.6 Zagon spletne aplikacije SQL Developer



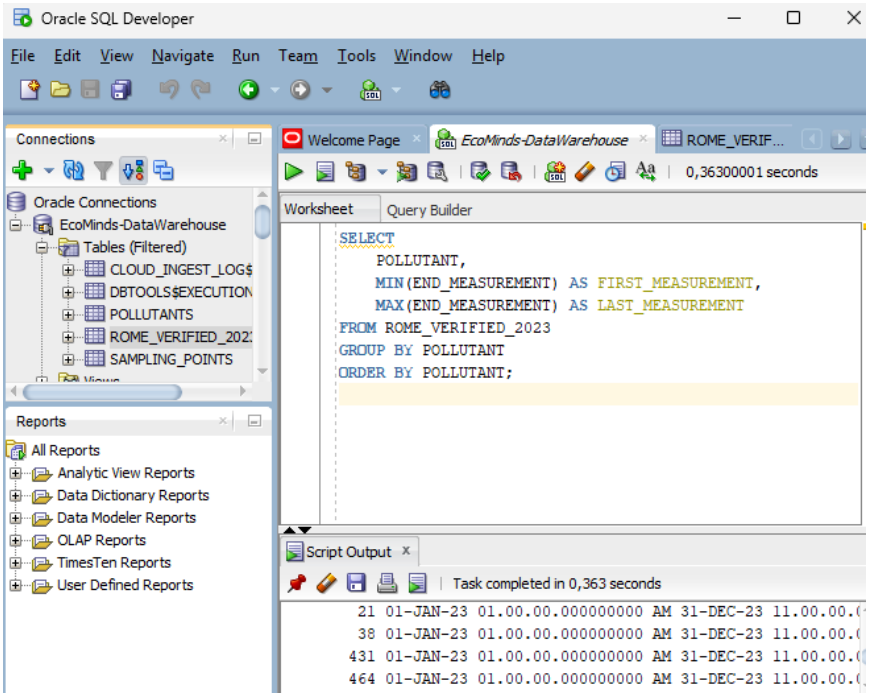
Slika 4.7: Spletna aplikacija SQL Developer



Slika 4.8 Prenos digitalne denarnice



Slika 4.9 Nastavitev povezave s podatkovnim skladiščem v namizni aplikaciji SQL Developer



Slika 4.10: Delo v namizni aplikaciji SQL Developer

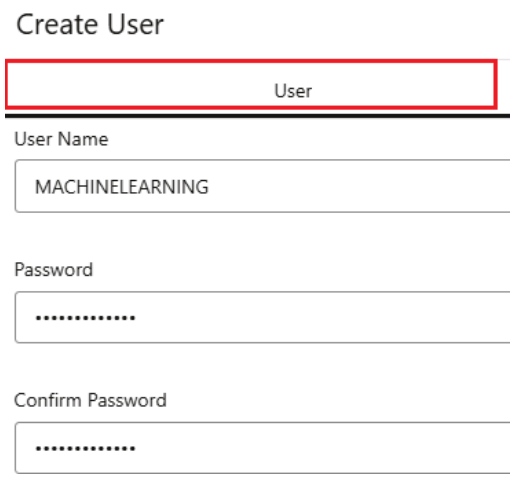
Upravljanje uporabniških pravic

Po ustvarjanju podatkovne zbirke je prvi razpoložljivi uporabnik v sistemu ADMIN. V izobraževalnih okoljih, zlasti pri uvodnih vajah, je zaradi poenostavitve nalaganja podatkov in izvajanja osnovnih poizvedb SQL začasno dovoljena uporaba računa ADMIN. Čeprav je ta pristop priročen za začetno spoznavanje sistema, v profesionalnih okoljih ni sprejemljiv.

V produkcijskih okoljih mora biti vloga ADMIN rezervirana za skrbniška opravila, kot so ustvarjanje uporabnikov, dodeljevanje privilegijev in konfiguriranje podatkovne zbirke. Ker ima račun ADMIN vse pravice, njegova uporaba za vsakodnevna opravila predstavlja varnostno tveganje. Zato se v skladu z načelom najmanjših privilegijev za uporabnike, ki opravljajo analitične ali razvojne naloge, ustvarijo namenski računi oziroma vloge z omejenim dostopom.

Upoštevajte, da ADMIN nima dostopa do vseh specializiranih orodij v Autonomous Database. Nekatere funkcionalnosti, kot so Oracle Machine Learning (OML), grafovna analitika in določene spletne komponente, zahtevajo privilegirane uporabniške račune. Zato se za delo s temi naprednimi orodji običajno ustvarijo ločeni uporabniki z jasno določenimi ravnmi dostopa, kot prikazujejo slike 4.11–4.13.

Slika 4.11 Izbira upravljanja uporabnikov



Slika 4.12 Vnos novega uporabnika

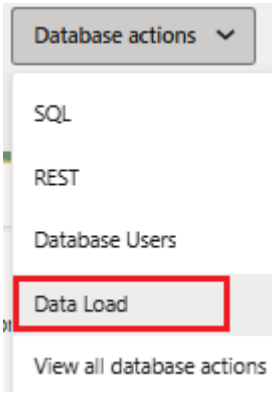


Slika 4.13 Določanje uporabniških pravic

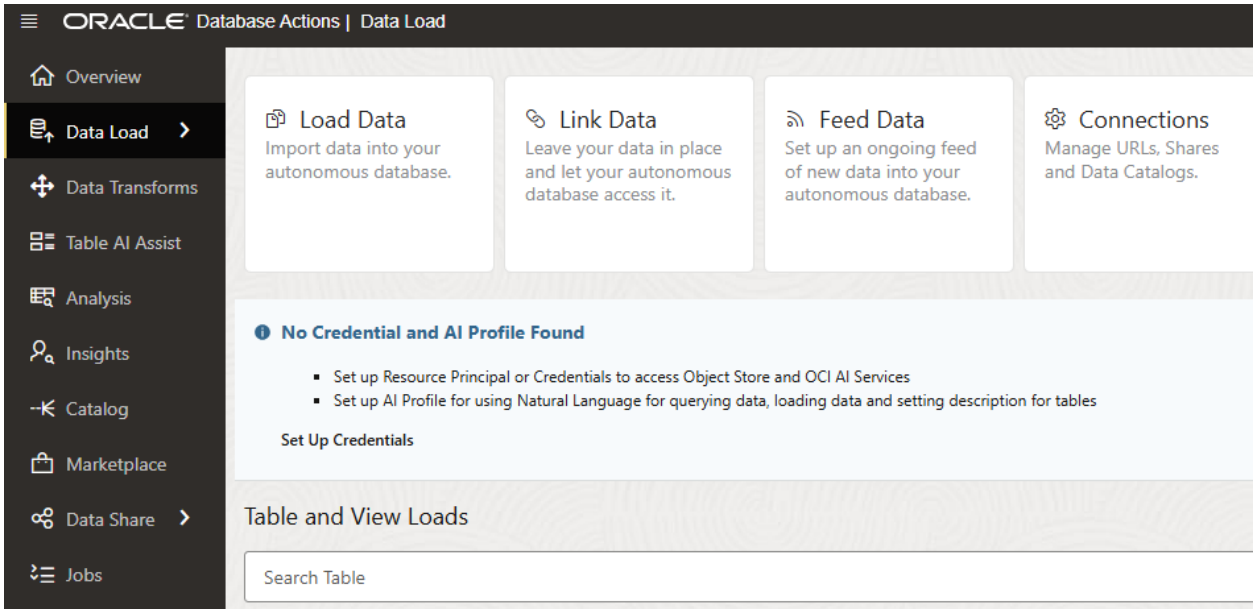
Nalaganje podatkov

Pravilna priprava podatkov je bistvena za učinkovito delovanje podatkovnega skladišča. Po konfiguriranju podatkovnega skladišča v primerku Autonomous Database se vsi podatki nalagajo prek enotnega vmesnika Data Load, ki je dostopen v Database Actions (sliki 4.14 in 4.15).

Najpreprostejši način za uvoz podatkov v skladišče je nalaganje lokalnih datotek s preprostim postopkom vlečenja in spuščanja na delovno površino vmesnika Data Load znotraj delovnega okolja. Sistem samodejno prepozna strukturo, predlaga imena stolpcev in podatkovne tipe ter ponudi možnost ustvarjanja nove tabele. Ta pristop je primeren za manjše podatkovne nize v formatih CSV, JSON ali podobnih oblikah in ne zahteva dodatnih orodij.



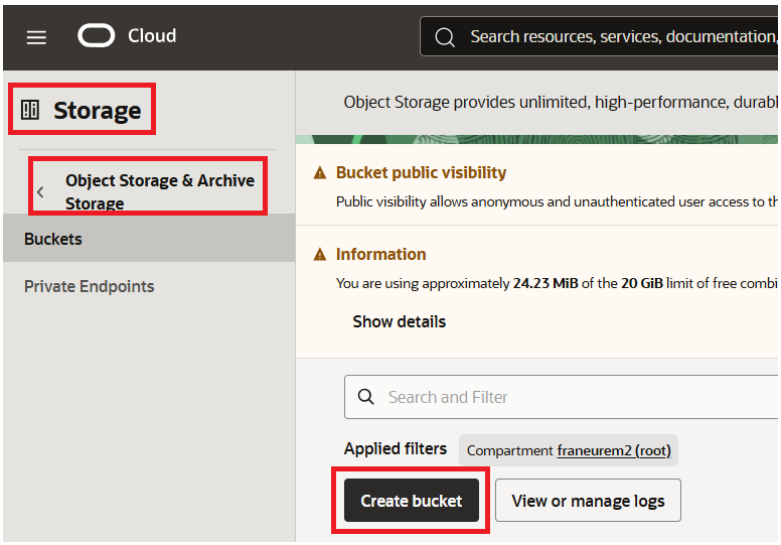
Slika 4.14 Zagon orodja Data Load



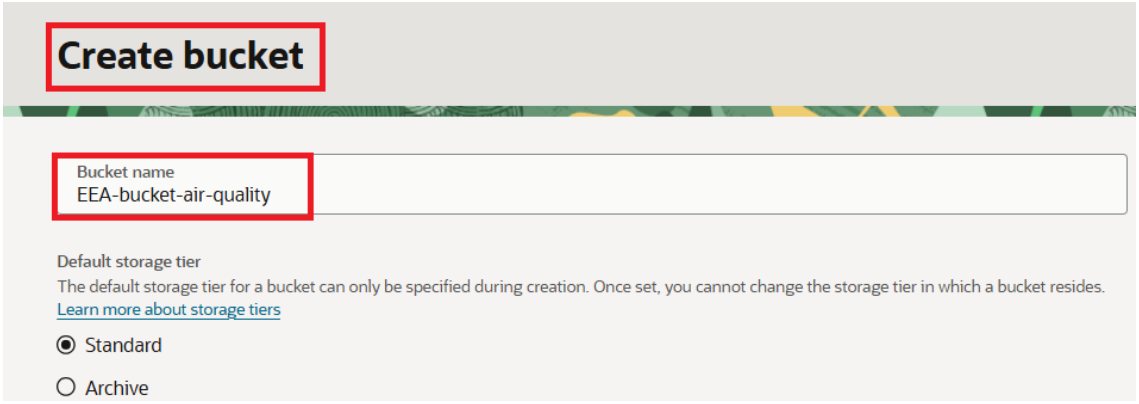
Slika 4.15 Vmesnik Data Load

Pri velikih datotekah, zlasti v formatu Parquet, je priporočljivo, da jih najprej naložite v storitev Oracle Object Storage v infrastrukturi Oracle Cloud Infrastructure. V navedenem repozitoriju se podatki hranijo v vsebnikih (angl. buckets), v stalno brezplačnem načinu pa je na voljo do 20 GB prostora, kar zadostuje za različne analitične namene.

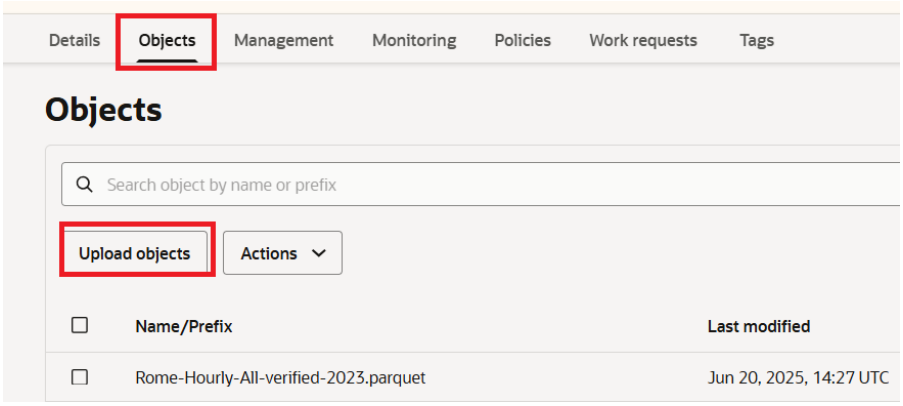
V razdelku Storage - Object Storage ustvarite vsebnik in mu nastavite ime default (sliki 4.16 in 4.17). Nato lahko pripravljene datoteke naložite v podatkovno skladišče ali jih iz njega prenesete (slika 4.18).



Slika 4.16 Ustvarjanje novega vsebnika v Oracle Object Storage.



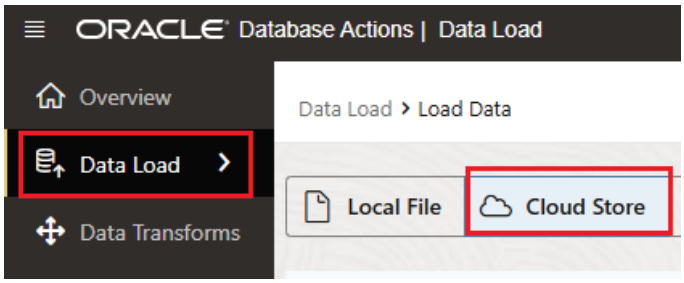
Slika 4.17 Ustvarjanje novega vsebnika v Oracle Object Storage.



Slika 4.18 Nalaganje datotek v vsebnik

Ko so datoteke shranjene v vsebniku, se njihov uvoz v podatkovno zbirko znova izvede prek istega vmesnika Data Load. Namesto lokalne datoteke izberemo vir Cloud Store (slika 4.19) in

vnesemo URL objekta iz vsebnika. Sistem nato samodejno preslika podatke in ustvari ustrezno tabelo.



Slika 4.19 Nalaganje podatkov iz Object Storage z modulom Data Load

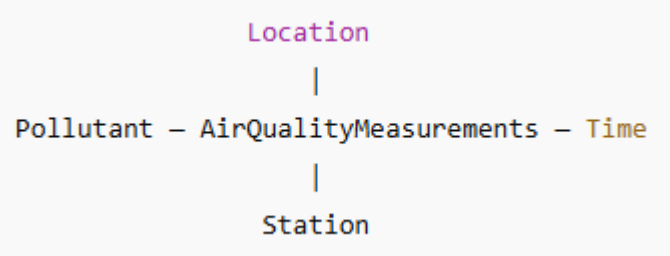
Modeliranje podatkov

Po nalaganju podatkov sledi njihovo modeliranje, ki omogoča nadaljnje analize. Poudarek je na zasnovi jasnih in razumljivih struktur tabel v podatkovni zbirki, da se zagotovi učinkovito izvajanje poizvedb SQL.

Na začetku so tabele navadno ustvarjene neposredno iz naloženih datotek. Vsak podatkovni niz, denimo meritve kakovosti zraka ali meteorološki zapisi, mora biti urejen v ločeni tabeli, pri čemer vsaka vrstica predstavlja posamezno meritev. Te tabele pogosto imenujemo surove tabele, ker prikazujejo strukturo izvornih podatkov brez nadaljnjih pretvorb. Čeprav je na njih mogoče izvajati osnovne analize, je mogoče njihovo strukturo dodatno izboljšati, da se omogoči lažje in zanesljivejše izvajanje poizvedb.

Za zahtevnejše analize je koristno vnaprej poenotiti imena stolpcev, oblike datumov ter identifikatorje lokacij ali merilnih postaj. Dosledna struktura olajša združevanje tabel in zmanjšuje tveganje napak pri razlagi. V nekaterih primerih je mogoče uporabiti osnovno denormalizacijo, denimo tako, da se podatki o merilni postaji iz pomožne tabele dodajo neposredno v tabelo meritev, kar poenostavi poizvedbe, ki bi sicer zahtevale več združevanj.

Eden najpogostejših in konceptualno preprostih načinov organiziranja podatkov je tako imenovana zvezdna shema, prikazana na sliki 4.20. Osrednjo tabelo, denimo tabelo meritev, uporablja kot glavni vir podatkov, medtem ko so opisni podatki o merilnih lokacijah, postajah ali drugih kategorijah shranjeni v manjših tabelah. Te tabele, imenovane dimenzijske tabele, se po potrebi povezujejo z osrednjo tabelo, zato je analitične poizvedbe lažje zapisati.

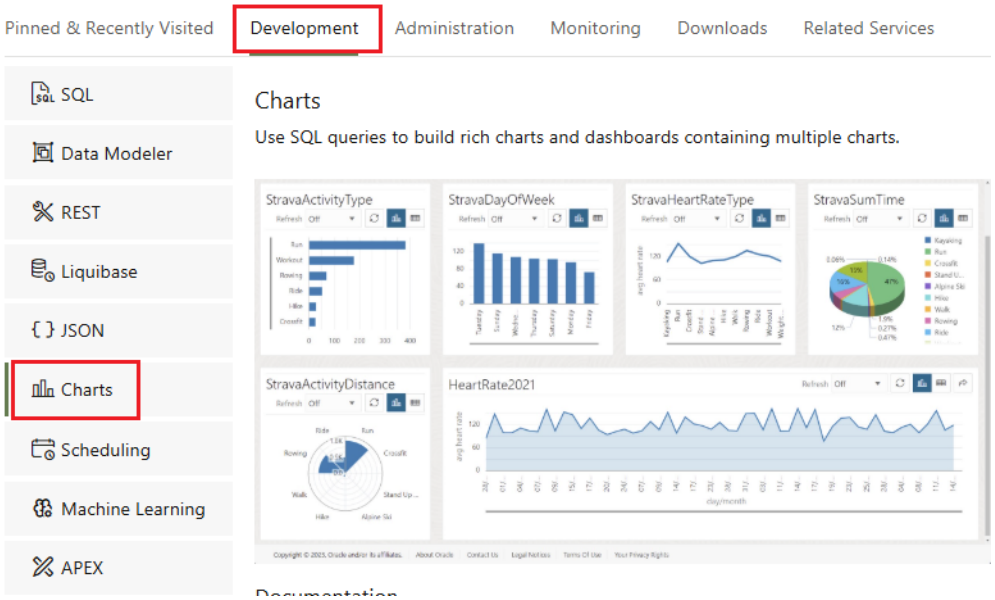


Slika 4.20 Preprosta zvezdna shema

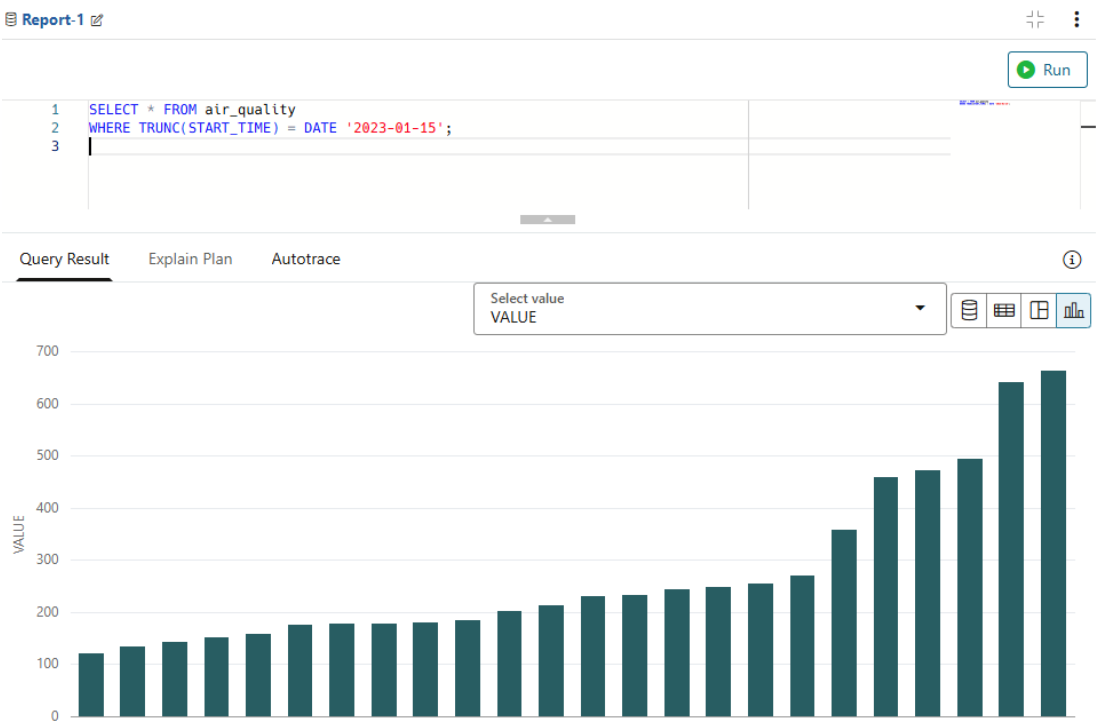
Vizualizacija podatkov

Po izvedbi poizvedbe SQL lahko rezultate prikažete v obliki grafikonov z vgrajenim modulom Charts. Do modula dostopate prek možnosti View all database actions; nato odprite zavihek Data Studio, ki vsebuje modul Charts, kot je prikazano na sliki 4.21. To orodje omogoča ustvarjanje preprostih vizualizacij neposredno iz rezultatov poizvedb SQL, ne da bi potrebovali zunanje aplikacije ali dodatne namestitve. Omogoča hitro in preprosto razlago podatkov, zlasti pri analizi trendov, primerjavi podatkovnih nizov ali prepoznavanju anomalij.

Predstavljajmo si na primer tabelo urnih meritev koncentracije določene snovi v zraku v daljšem obdobju. V tem primeru se je pogosto treba osredotočiti le na meritve za posamezen dan. V prikazanem primeru je uporabljen stolpčni grafikon: na osi X so prikazani časi meritev za izbrani dan, na osi Y pa pripadajoče vrednosti (slika 4.22).



Slika 4.21 Zagon modula Charts



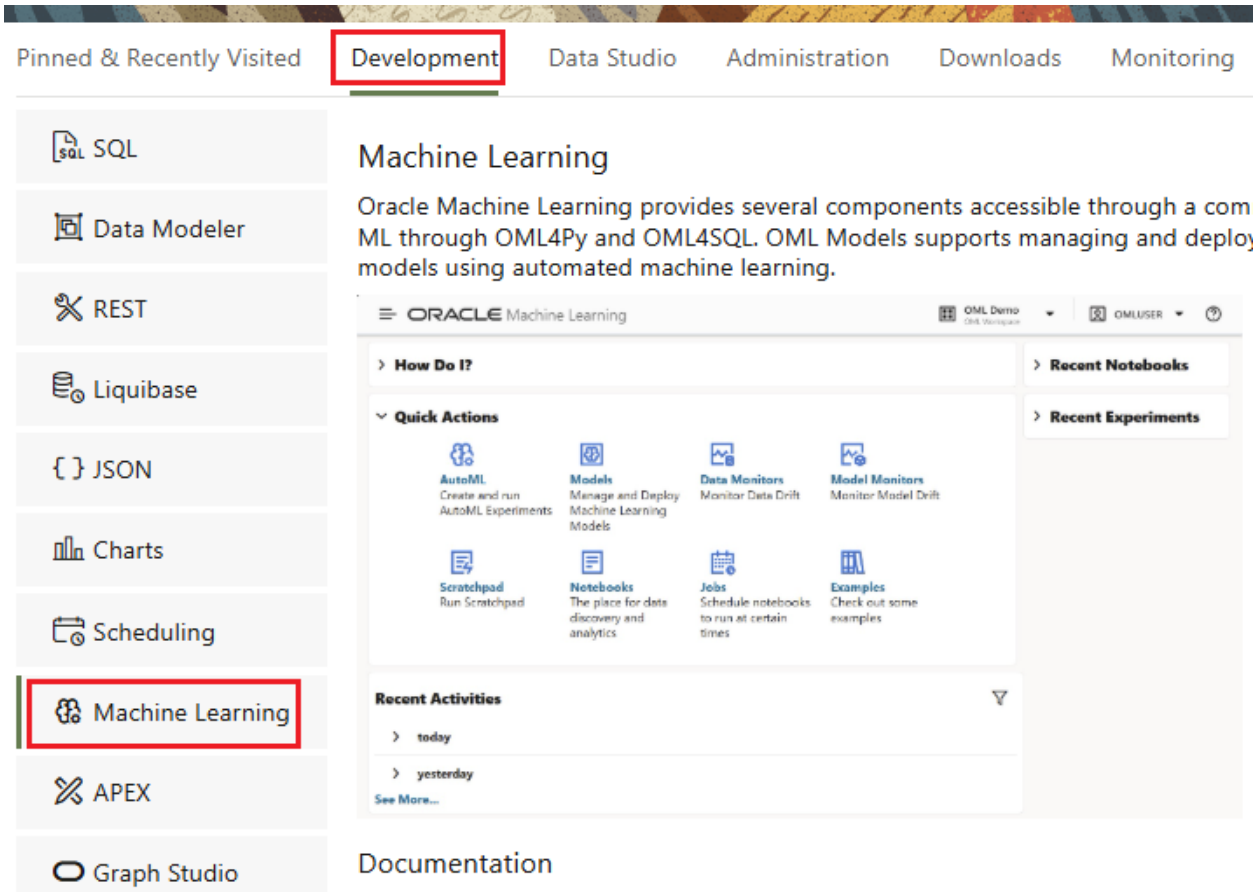
Slika 4.22: Preprosta vizualizacija podatkov

Strojno učenje v podatkovni zbirki

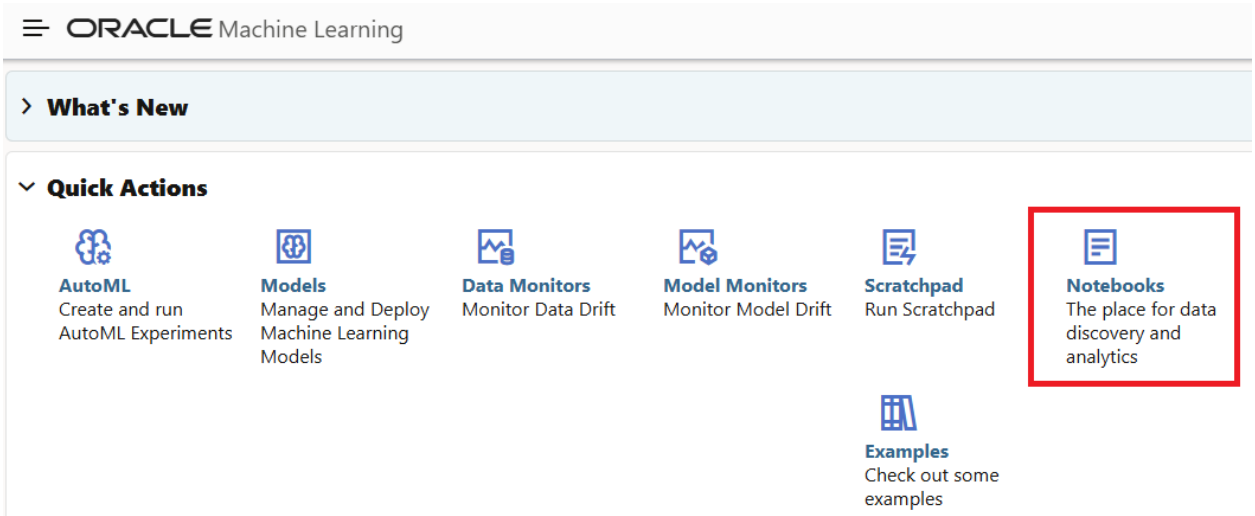
Podatkovno skladišče z Oracle Autonomous Database vključuje vgrajen sistem strojnega učenja Oracle Machine Learning (OML), ki omogoča učenje in uvajanje modelov neposredno na podatkih, shranjenih v podatkovni zbirki. Za razliko od pristopov, pri katerih se podatki izvozijo v zunanja orodja, OML omogoča pripravo podatkov, modeliranje in vrednotenje znotraj podatkovne zbirke, s čimer odpravlja potrebo po dodatni infrastrukturi in obsežnih prenosih podatkov.

Do okolja OML dostopate tako, da izberete Database Actions - View all database actions, nato kliknete zavihek Development in meni Machine Learning (slika 4.23). S tem se odpre delovni prostor OML, prikazan na Sliki 4.24.

Najpreprostejši način dela v OML je prek okolja OML Notebooks (glejte sliko 4.24), v katerem lahko skripte SQL in Python pišete ter izvajate neposredno v spletnem brskalniku. Ponuja funkcionalnosti za klasifikacijo, regresijo, gručenje, odkrivanje anomalij in samodejno učenje modelov.



Slika 4.23 Zagon OML



Slika 4.24 Vmesnik OML

Študija primera: analiza onesnaženosti zraka

Ta študija primera predstavlja podatke o kakovosti zraka iz resničnega sveta, ki jih uporabljajo evropske okoljske agencije, raziskovalne ustanove in analitični sistemi. Te urne meritve zagotavlja Evropska agencija za okolje (EEA).

Primer prikazuje celoten postopek dela z okoljskimi podatkovnimi nizi. Prvi del obravnava načine zbiranja in shranjevanja podatkov v podatkovnem skladišču. Poznejši koraki vključujejo analizo podatkov, povezano z okoljskimi vprašanji, kot je spremljanje dušikovega dioksida (NO_2) ali delcev $PM_{2.5}$.

Evropski viri podatkov o kakovosti zraka

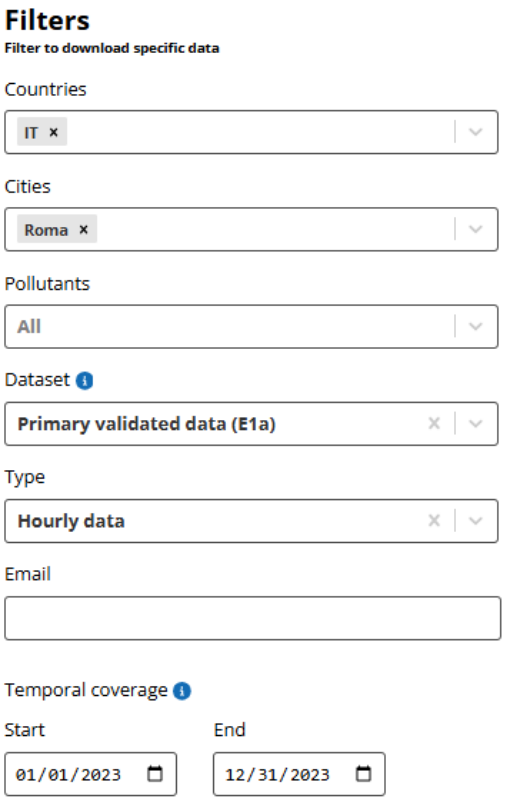
Podatke o kakovosti zraka je mogoče prenesti s portala EEA [10], ki zbira, preverja in objavlja meritve nacionalnih mrež za spremljanje okolja. Sistem EEA omogoča dostop do podrobnih urnih, dnevnih in letnih zapisov koncentracij onesnaževal, kot so dušikov dioksid (NO_2), ozon, $PM_{2.5}$ in druge snovi, ki so ključne za ocenjevanje kakovosti zraka.

Merilna mesta so razporejena po vsej Evropi in uporabljajo standardizirane metode v skladu z evropskimi direktivami o spremljanju kakovosti zraka. Pred objavo so podatki validirani, da se odstranijo napake, logične nedoslednosti in manjkajoči zapisi. Obdelani podatki so na voljo prek uradne storitve EEA [11], vključno z možnostmi prenosa v formatih Parquet in CSV.

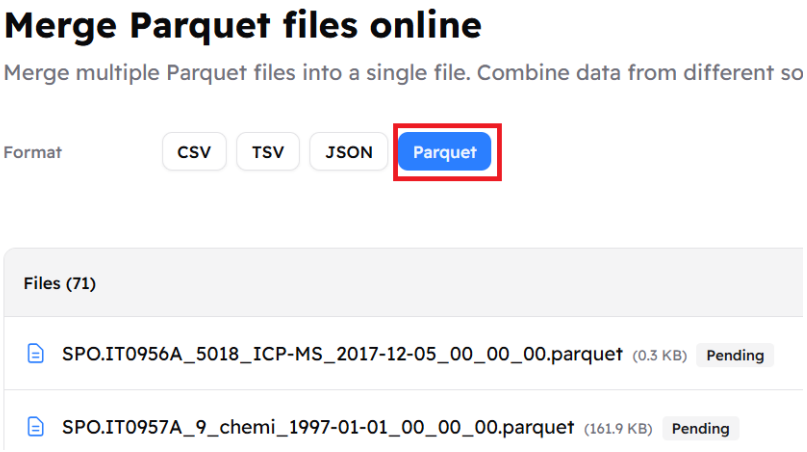
Po izbiri geografskega območja (slika 4.25), časovnega obdobja in želenega merilnega intervala sistem ustvari paket z izbranimi meritvami. Preneseni paket navadno vsebuje več datotek Parquet, po eno za vsako merilno postajo na izbranem območju. Takšna ureditev omogoča podrobno analizo, vendar je v analitičnih okoljih pogosto koristno vse zapise združiti v eno tabelo. To je mogoče storiti z razpoložljivimi orodji za združevanje datotek Parquet, kot je Agents for Data [12], ki uporabnikom omogoča, da z vlečenjem združijo več datotek v enoten podatkovni niz (slika 4.26).

Poudariti je treba tudi, da je nujno razumeti strukturo prenesenih podatkov, njihove atribute in pomen ter objavljene povezave, ki omogočajo preučevanje metapodatkov, podatkovnih slovarjev, besednjakov in tehnične dokumentacije objavljenih vmesnikov API.

.



Slika 4.25 Prenos podatkov EEA

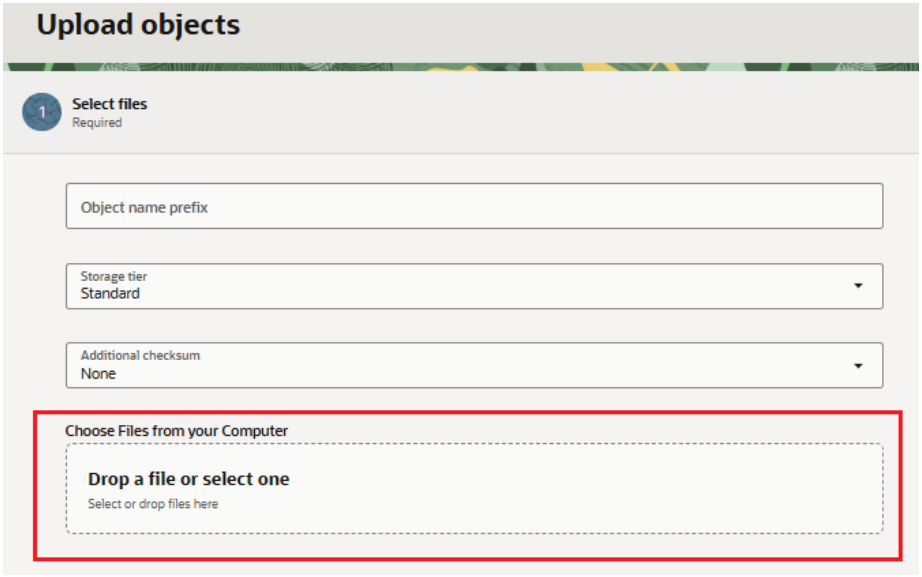


Slika 4.26 Združevanje več datotek Parquet na platformi Agents for Data

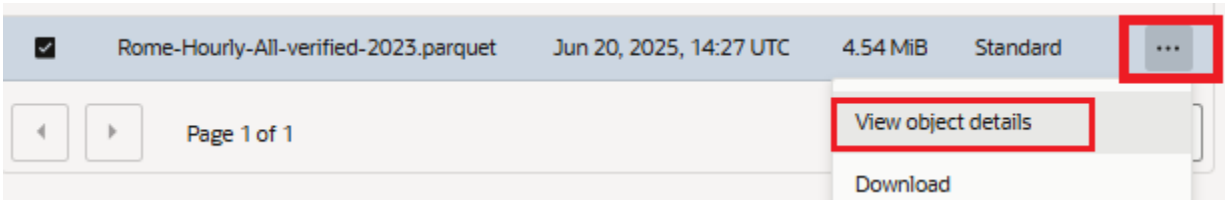
Shranjevanje meritev EEA v Oracle Object Storage

Po prenosu podatkov EEA jih je treba shraniti v predhodno konfiguriran vsebnik v storitvi Oracle Object Storage. V vsebniku izberite Objects, nato kliknite Upload Objects in naložite datoteko meritev v formatu Parquet (slika 4.27).

Vsakemu naloženemu objektu je dodeljen lasten URL, ki si ga je mogoče ogledati v lastnostih objekta (glejte sliko 4.28). Ta URL se pozneje uporabi pri uvozu podatkov v podatkovno skladišče.



Slika 4.27 Nalaganje datoteke EEA v vsebnik



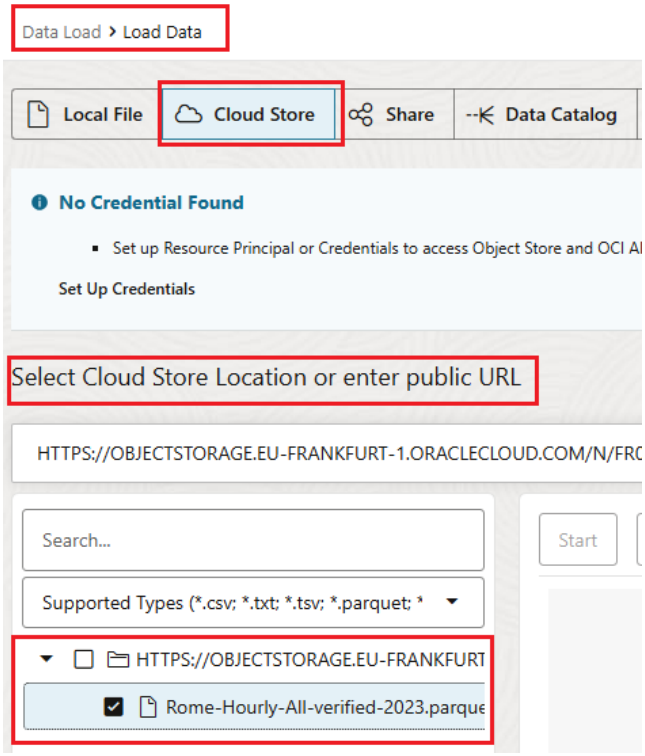
Slika 4.28: Ustvarjanje URL-ja objekta

Uvoz meritev v podatkovno skladišče

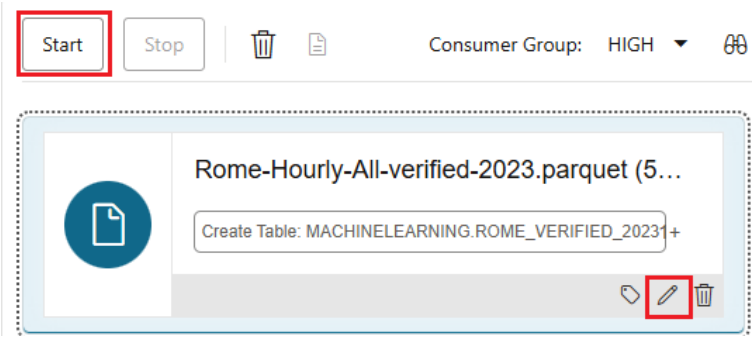
V naslednjem koraku v okolju Autonomous Database odprite vmesnik Data Load in izberite zavihek Cloud Store. V polje za vir podatkov vnesite predhodno kopirani URL objekta Parquet, shranjenega v vsebniku (slika 4.29).

Na levi plošči je prikazana struktura razpoložljivih podatkov, ki jih lahko izberete in preprosto povlečete na delovno površino za uvoz. Postopek nalaganja zaženete s klikom gumba Start (slika 4.30).

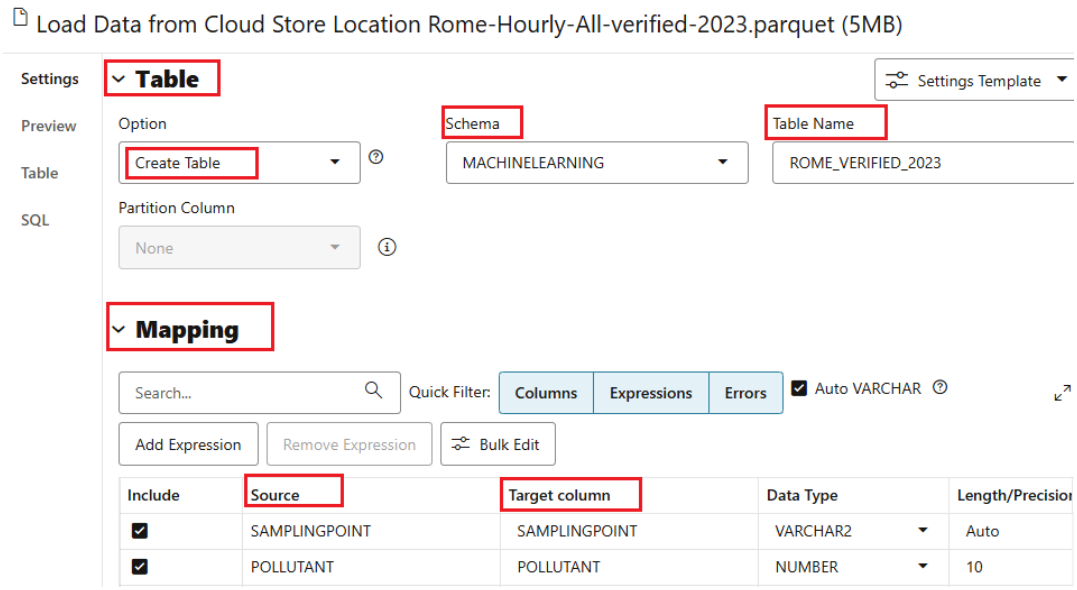
Pred uvozom je priporočljivo odpreti povezavo Settings, da pregledate strukturo datoteke in nastavite ključne možnosti nalaganja: ustvarjanje nove tabele ali posodobitev obstoječe, preslikavo stolpcev in izbiro uporabniške sheme (slika 4.31).



Slika 4.29 Priprava na nalaganje merilnih podatkov



Slika 4.30 Priprava na nalaganje merilnih podatkov



Slika 4.31: Prilagajanje parametrov prenosa

Primeri začetne analize podatkov

Spodaj so primeri osnovnih poizvedb, s katerimi si lahko ogledate vsebino tabele in razumete razpoložljive atribute. Ta začetni pregled pomaga razjasniti strukturo podatkov, razpon meritev ter način označevanja onesnaževal in merilnih postaj.

Vsa onesnaževala, navedena v tabeli meritev, si lahko ogledate z naslednjo poizvedbo SQL:

```
-------------------------------------------------------------------------------
-- 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;
```

Nato je koristno sestaviti seznam vseh merilnih postaj:

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

Določiti je mogoče tudi časovni razpon meritev za vsa poročana onesnaževala:

```
-------------------------------------------------------------------------------
-- 3. Determine the measurement range for each pollutant
```

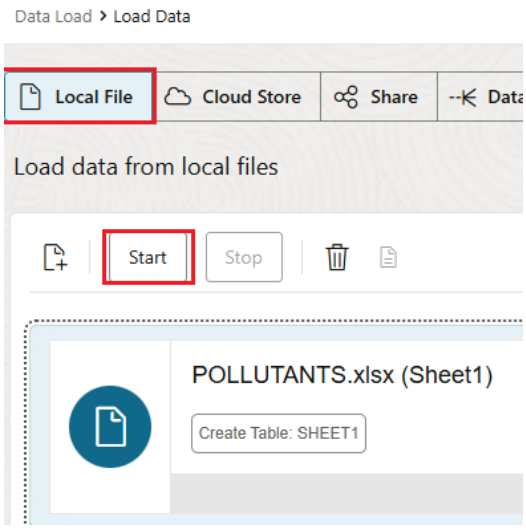
```
--    Shows the first and last measurement timestamp per pollutant.
-------------------------------------------------------------------------------
SELECT
    POLLUTANT,
    MIN(END_MEASUREMENT) AS FIRST_MEASUREMENT,
    MAX(END_MEASUREMENT) AS LAST_MEASUREMENT
FROM ROME_VERIFIED_2023
GROUP BY POLLUTANT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

Podobno lahko prikažemo tudi časovni razpon meritev za posamezno merilno postajo.

```
-------------------------------------------------------------------------------
-- 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;
```

Povezovanje meritev z opisi onesnaževal

Žal iz tabele meritev kakovosti zraka (ROME_VERIFIED_2023) ni mogoče ugotoviti, katera onesnaževala so vključena, saj so navedene le njihove številčne kode. Težavo rešimo tako, da to tabelo povežemo z drugo tabelo, ki vsebuje podrobne informacije o onesnaževalih zraka in njihove oznake. Vir podatkov, potreben za izdelavo dodatne tabele, je na voljo v slovarju EEA [13]. Seznam onesnaževal je mogoče preprosto kopirati v Excelovo datoteko in ga nato z okoljem Data Load naložiti v podatkovno skladišče. V tem primeru se podatki uvozijo v tabelo POLLUTANTS, kot je prikazano na sliki 4.32.



Slika 4.32 Uvoz kod onesnaževal zraka

Zdaj lahko meritve in vrste onesnaževal povežemo v eni poizvedbi 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;
```

Poleg tega lahko ocenimo letno povprečno koncentracijo NO_2 na vseh merilnih postajah.

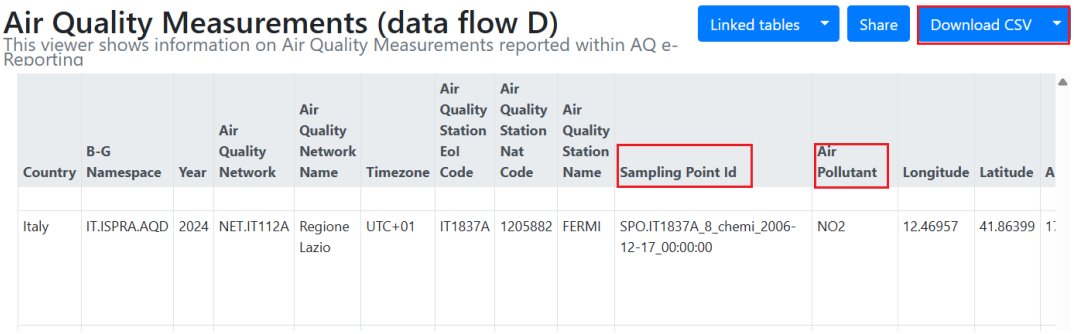
```
-------------------------------------------------------------------------------
-- Annual NO2 averages (pollutant = 8), excluding negative values.
-- Joined with POLLUTANTS for pollutant name, symbol and unit.
-------------------------------------------------------------------------------
SELECT 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;
```

Izvedba prejšnje poizvedbe pokaže, da merilna postaja z oznako IT/SPO.IT1837A_8_chemi_2006-12-17_00:00:00 beleži povprečno letno koncentracijo NO_2 47,074 µg/m³, kar je zdravju škodljivo (zgornja meja je 40,00 µg/m³).

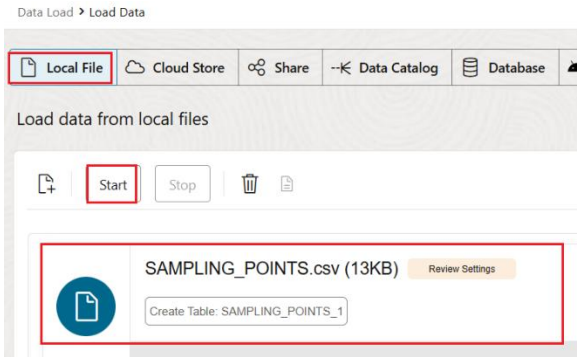
Povezovanje meritev z lokacijami merilnih postaj

Ko potrdimo, da posamezne merilne postaje beležijo preseganja NO_2, je nujno pridobiti dodatne informacije o njihovih lokacijah. Tabela ROME_VERIFIED_2023 vsebuje le identifikatorje merilnih postaj, zato njihovih natančnih lokacij ni mogoče določiti neposredno. To težavo odpravimo tako, da v podatkovno skladišče dodamo drugo tabelo, ki hrani opisne podatke o merilnih postajah, vključno z njihovimi imeni, geografskimi koordinatami in vrstami postaj.

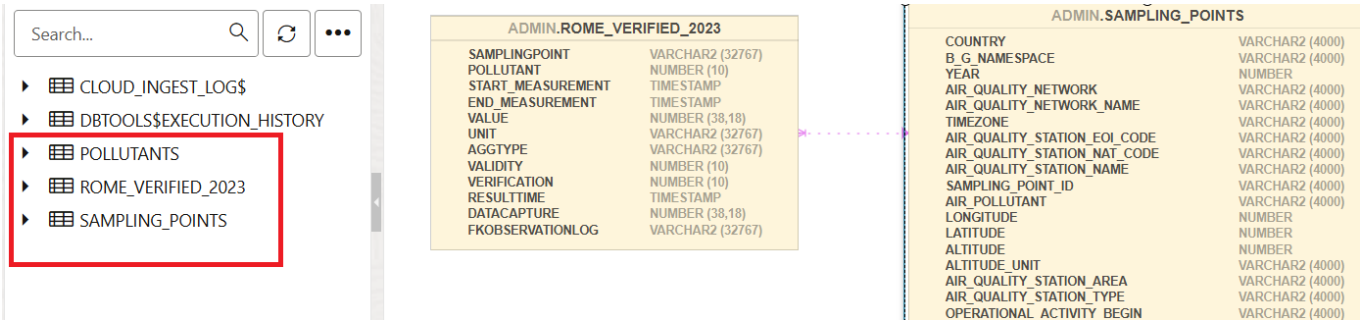
Vir takih podatkov je zgoraj omenjeni portal EEA [10], ki zagotavlja celoten seznam evropskih merilnih postaj. Seznam prenesemo kot datoteko CSV tako, da kliknemo povezavo Metadata in nato izberemo Download CSV (glejte sliko 4.29). Preneseno datoteko CSV nato z delovno površino Data Load naložimo nazaj v podatkovno skladišče (glejte sliko 4.30). S tem postopkom ustvarimo novo tabelo, v tem primeru poimenovano SAMPLING_POINTS, ki vsebuje vse ustrezne informacije o merilnih mestih (glejte sliko 4.31).



Slika 4.29 Prenos podatkov o merilnih postajah



Slika 4.30 Nalaganje podatkov o merilnih postajah



Slika 4.31 Tabele o onesnaženosti zraka

Po analizi tabel, ustvarjenih z uvozom podatkov, ugotovimo, da je mogoče tabelo ROME_VERIFIED_2023 povezati s tabelo POLLUTANTS prek stolpca POLLUTANT v tabeli ROME_VERIFIED_2023 in stolpca POLLUTANT_ID v tabeli POLLUTANTS.

Poleg tega je mogoče tabelo ROME_VERIFIED_2023 povezati s tabelo SAMPLING_POINTS prek stolpca SAMPLINGPOINT v tabeli ROME_VERIFIED_2023 in stolpca SAMPLING_POINT_ID v tabeli SAMPLING_POINTS.

Vendar pri poskusu povezovanja nastane težava, saj se vrednosti v stolpcu SAMPLINGPOINT tabele ROME_VERIFIED_2023 ne ujemajo neposredno z vrednostmi v tabeli SAMPLING_POINTS. V tabeli ROME_VERIFIED_2023 je na primer lahko vrednost IT/SPO.IT1837A_8_chemi_2006-12-17_00:00:00, medtem ko je ista lokacija v tabeli SAMPLING_POINTS zapisana le kot SPO.IT1837A_8_chemi_2006-12-17_00:00:00.

Zapis v tabeli ROME_VERIFIED_2023 zato vsebuje predpono države (IT/), ki je v tabeli SAMPLING_POINTS ni. Ker SQL JOIN zahteva natančno ujemanje vrednosti, teh tabel ni mogoče povezati neposredno.

Težavo odpravimo z uporabo funkcije SUBSTR (SAMPLINGPOINT, 4) v stavku JOIN, ki iz vrednosti v tabeli ROME_VERIFIED_2023 odstrani prve tri znake (IT/). Vrednosti se nato ujemajo, kar omogoča pravilno povezavo stolpcev SAMPLINGPOINT in SAMPLING_POINT_ID ter združitev vseh treh tabel v eni poizvedbi.

```
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;
```

Dokončamo lahko tudi agregatno poizvedbo, ki poveže vse tri tabele in izračuna letna povprečja onesnaženosti z NO_2 (oznaka 8). Ta poizvedba hkrati prikaže podatke o onesnaženosti in lokacijske atribute merilnih postaj, s čimer olajša razlago rezultatov meritev.

```
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;
```

Rezultati prejšnje poizvedbe so prikazani v tabeli 1, iz katere je razvidno, da izmerjena vrednost na merilni postaji Fermi v Rimu, na trgu Piazza Enrico Fermi, presega letno mejno vrednost povprečne koncentracije NO_2, ki znaša 40 µg/m³ (mikrogramov na kubični meter). Koordinate te postaje (41.86399° N, 12.46957° E) si je mogoče ogledati v storitvah Google Maps in Street View (slika 4.32).

Tabela 1: Pregled letnih povprečnih koncentracij NO_2 po merilnih postajah v Rimu (2023)

ONESNAŽEVALO	DRŽAVA	OBČINA	MERILNA POSTAJA KAKOVOSTI ZRAKA	OBMOČJE	VRSTA MERILNE POSTAJE KAKOVOSTI ZRAKA	ZEMLJEPISNA ŠIRINA	ZEMLJEPISNA DOLŽINA	LETO	VREDNOST
Dušik (zrak)	Italija	CORSO FRANCIA - Roma (RM)	C.SO FRANC	mestno	prometna	41,94745	12,46956	2023	35,61
Dušik (zrak)	Italija	LARGO MAGNA GRECIA - Roma (RM)	L.GO MAGNA GRECIA	mestno	prometna	41,88307	12,50895	2023	33,96
Dušik (zrak)	Italija	LARGO PERESTRELLO - Roma (RM)	L.GO PERESTRELLO	mestno	ozadje	41,88602	12,54162	2023	26,22
Dušik (zrak)	Italija	LGO BELOLLI - Roma (RM)	CINECITTA	mestno	ozadje	41,85772	12,56865	2023	25,77
Dušik (zrak)	**Italija**	**P.ZZA E. FERMI, FRONTE CIVICO16 - Roma (RM)**	**FERMS**	**mestno**	**prometna**	**41,86399**	**12,46957**	**2023**	**47,07**
Dušik (zrak)	Italija	Piazza Cairoli adiacente via dei Giubbonari - Roma (RM)	ARENUL	mestno	ozadje	41,89397	12,47539	2023	33,00
Dušik (zrak)	Italija	VIA AURELIA KM 20 CASTEL DI GUIDO - Roma (RM)	CASTEL DI GUIDÓ	podeželsko	ozadje	41,88945	12,26633	2023	8,17
kaže, da je letna mejna vrednost povprečne koncentracije NO_2 40 µg/m³ (mikrogramov na kubični meter) na merilni postaji Fermi, ki se nahaja v	Italija	VIA DELLA MELORIA, FRONTE CIVICO 27 - Roma (RM)	CIPRO	mestno	ozadje	41,90636	12,44759	2023	30,04
Dušik (zrak)	Italija	VIA MONTE CANDA, ADIACENTE CIVICO 12 - Roma (RM)	BUFALOTTA	mestno	ozadje	41,94765	12,53367	2023	29,94
Dušik (zrak)	Italija	VIA PAOLO EMILIO PAVOLINI 69/71 - Roma (R)	MALAGROT	predmestno	ozadje	41,87478	12,34559	2023	16,90
Dušik (zrak)	Italija	VIA SALARIA - Roma (RM)	VILLA ADA	mestno	ozadje	41,93289	12,50695	2023	19,76
Dušik (zrak)	Italija	VIA TENUTA DEL CAVALIERE 102 - Roma (RM)	TENUTA DEL CAVALIER	predmestno	ozadje	41,92947	12,65848	2023	23,89
Dušik (zrak)	Italija	VIA TIBURTINA, ADIACENTE CIVICO 654 - Roma (RM)	TIBURNIA	mestno	prometna	41,91027	12,54885	2023	35,50



Slika 4.32 Pogled Google Street View na merilno postajo Fermi

Praktične vaje

- Povežite se z Oracle Cloud Console.
- Ustvarite nov primerek Autonomous Database, v katerem boste vzpostavili podatkovno skladišče.
- Ustvarite vsebnik v storitvi Oracle Object Storage.
- Povežite se s portalom EEA ter preglejte dokumentacijo in opise podatkovnih modelov (povezavi Metadata in Vocabulary).
- Prenesite datoteko Parquet s trenutnimi meritvami onesnaženosti zraka za svojo državo in jo shranite v svoj podatkovni vsebnik.
- V podatkovno skladišče naložite meritve iz vsebnika ter vnaprej pripravljeni datoteki POLLUTANTS (Excel) in SAMPLING_POINTS (CSV).
- Analizirajte meritve delcev PM2.5. Kateri so viri tega onesnaženja in kako vpliva na zdravje ljudi?
- Določite lokacije, na katerih so bile izmerjene najvišje ravni onesnaženosti z delci PM2.5.
- Primerjajte izmerjene vrednosti z mejnimi vrednostmi.

Preverjanje znanja

Katere so ključne lastnosti in prednosti podatkovnih skladišč?

A. Omogočajo analizo velikih naborov zgodovinskih podatkov
B. Optimizirana so za hitre transakcije v realnem času
C. Poenotijo in povežejo podatke iz več virov
D. Uporabljajo se namesto izvajanja poročevalskih opravil v transakcijskih podatkovnih zbirkah

Če želite meteorološke podatke pripraviti v podatkovnem skladišču in uvesti preprosto zvezdno shemo, kateri najmanjši nabor tabel potrebujete? (En pravilen odgovor.)

A. Ena tabela meritev in ena tabela lokacij meritev
B. Ena tabela meritev, tabela merilnih senzorjev in tabela upravljavcev sistema
C. Samo ena tabela meritev, saj so vsi podatki že v njej
D. Vse navedeno

Katera lastnost najbolje opisuje tako imenovane »surove« (raw) tabele v okoljskem podatkovnem modelu?

A. Predstavljajo izvorne, neobdelane podatke, prevzete iz zunanjih sistemov, brez nadaljnjih pretvorb
B. Vsebujejo začasne podatke, prenesene iz zunanjih sistemov
C. Hranijo podatke iz zunanjih sistemov ter so zasnovane za hiter dostop in analizo
D. Optimizirane so za izvajanje agregatnih poizvedb in izračunavanje statističnih podatkov

Kateri koraki najbolje opisujejo postopek modeliranja podatkovnega skladišča v okviru predstavljene študije primera okoljskih podatkov?

A. Prenos podatkov iz zunanjih virov
B. Pregled strukture in razumevanje podatkov (metapodatki, besedišče), opredelitev ključnih atributov in čiščenje podatkov
C. Povezovanje tabel
D. Vse navedeno

Katere so osnovne funkcije vmesnika Data Load pri delu s podatkovnim skladiščem Oracle Autonomous Database?

A. Samodejno ustvarjanje poizvedb SQL za analizo podatkov
B. Uvoz podatkov iz vsebnika Object Storage
C. Upravljanje uporabniških računov in varnostnih nastavitev
D. Optimizacija analitičnih poizvedb SQL
E. Uvoz lokalnih datotek

Viri / dodatno branje

[1] »Autonomous AI Database«, Oracle. [Na spletu]. Dostopno na: https://www.oracle.com/autonomous-database/ (dostopano 9. decembra 2025).

[2] »Oracle Cloud Infrastructure (OCI)«, Oracle. [Na spletu]. Dostopno na: https://www.oracle.com/cloud/ (dostopano 9. decembra 2025).

[3] »Amazon Redshift«, AWS. [Na spletu]. Dostopno na: https://aws.amazon.com/redshift/ (dostopano 9. decembra 2025).

[4] »BigQuery | AI data platform | Lakehouse | EDW«, Google Cloud. [Na spletu]. Dostopno na: https://cloud.google.com/bigquery (dostopano 9. decembra 2025).

[5] »Azure Synapse Analytics«, Microsoft Azure. [Na spletu]. Dostopno na: https://azure.microsoft.com/en-us/products/synapse-analytics (dostopano 9. decembra 2025).

[6] »Apache Hive«, Apache Software Foundation. [Na spletu]. Dostopno na: https://hive.apache.org/ (dostopano 9. decembra 2025).

[7] »Solutions Cloud Program«, Oracle Academy. [Na spletu]. Dostopno na: https://academy.oracle.com/en/solutions-cloud-program.html (dostopano 9. decembra 2025).

[8] »Oracle Cloud Free Tier«, Oracle. [Na spletu]. Dostopno na: https://www.oracle.com/cloud/free (dostopano 9. decembra 2025).

[9] »Oracle SQL Developer – Downloads«. Dostopno na: https://www.oracle.com/database/sqldeveloper/technologies/download (dostopano 9. decembra 2025).

[10] European Environment Agency. Dostopno na: https://www.eea.europa.eu/en (dostopano 9. decembra 2025).

[11] European Environment Agency, Data Download Service. Dostopno na: https://eeadmz1-downloads-webapp.azurewebsites.net (dostopano 9. decembra 2025).

[12] Agents for Data, »Parquet Merge Tool«. Dostopno na: https://www.agentsfordata.com/parquet/merge (dostopano 9. decembra 2025).

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