Skladištenje okolišnih podataka s primjenama

Ishodi učenja

Nakon završetka ove nastavne jedinice studenti će moći:

- Objasniti pojam i ulogu skladišta podataka s primjenama na okolišne podatke.
- Postaviti skladište podataka i upravljati s njegovim osnovnim funkcijama.
- Preuzimati, pretraživati i učitavati okolišne podatke iz dokumentiranih vanskih izvora.
- Primijeniti SQL upite na izrađenom skladištu okolišnih podataka.

Uvod

Okolišni podaci nastaju kao rezultat različitih mjerenja, pri čemu svaki izvor obuhvaća samo jedan aspekt složenog okolišnog sustava. Neki skupovi podataka odnose se na fizikalna svojstva zraka, drugi na meteorološke uvjete, treći na stanje ekosustava ili intenzitet ljudskih aktivnosti. Iako svaki od njih pruža vrijedne informacije, ni jedan pojedinačno ne može ponuditi cjelovit uvid u procese koji utječu na okoliš. Zbog toga analitički rad nad okolišnim podacima gotovo uvijek počinje njihovim povezivanjem i usklađivanjem u jedinstvenu sliku stanja okoliša.

Pri tome se uočavaju dva osnovna izazova: pronalaženje relevantnih podataka, koji su često razasuti među različitim institucijama, te njihova pohrana u oblik prikladan za naknadnu analizu. Podaci mogu dolaziti u različitim formatima, vremenskim razmacima i stupnjevima obrade, pa ih je potrebno objediniti u sustav koji omogućuje pouzdano spajanje, agregiranje i usporedbu. Upravo zato uloga skladišta podataka (engl. *data warehouse*) postaje ključna. Ono omogućuje prikupljanje podataka iz više izvora i njihovo pretvaranje u strukturirani analitički model namijenjen složenim upitima i interpretaciji.

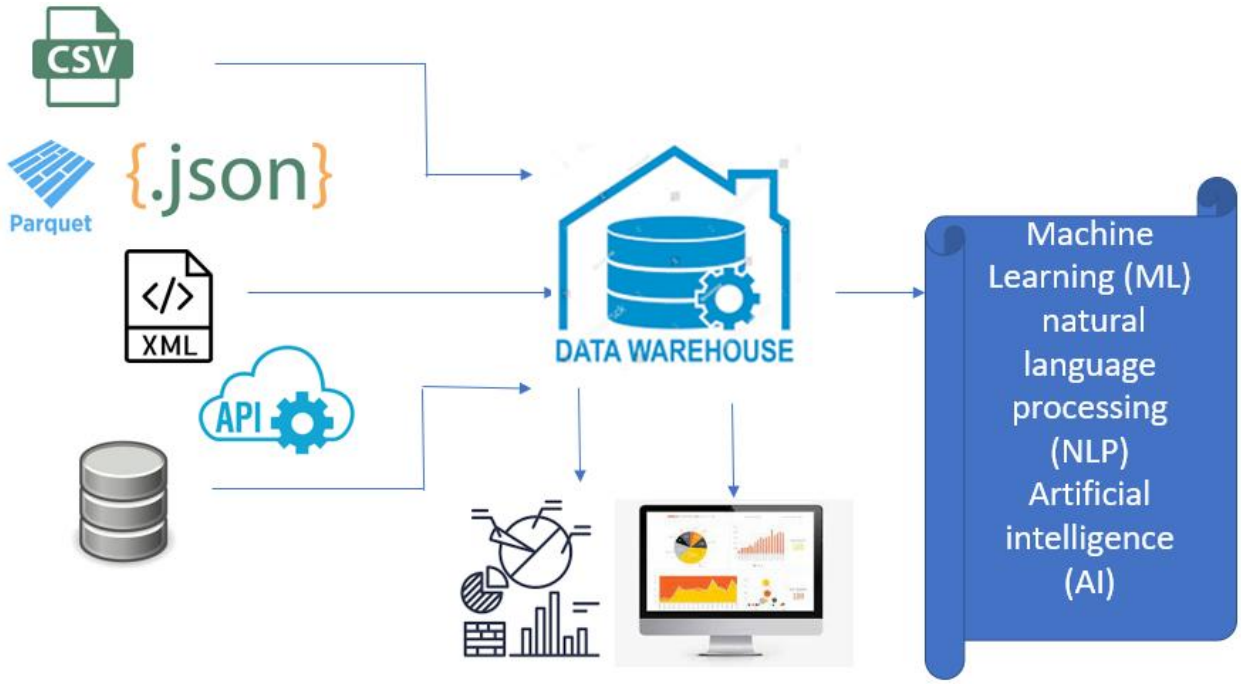
Skladište podataka razlikuje se od standardnih transakcijskih baza podataka po svojoj svrsi i načinu uporabe. Dok su transakcijski sustavi optimizirani za brze i česte zapise, skladište podataka usmjereno je na dugoročno čuvanje i analizu većih povijesnih skupova podataka. Naglasak je na analitičkim upitima, istraživanju trendova kroz vrijeme i integraciji heterogenih podataka koji zajedno omogućuju razumijevanje složenih pojava i njihove međuvisnosti.

U području okolišne znanosti, takav pristup je posebno važan jer se promjene u okolišu rijetko mogu objasniti izoliranim mjerenjima. Kvaliteta zraka, primjerice, rezultat je kombinacije meteoroloških uvjeta, emisijskih izvora, geografskih značajki i sezonskih varijacija. Tek kada se

ovi podaci promatraju zajedno i kroz dulje razdoblje, moguće je pouzdano procijeniti razvoj stanja okoliša, učinkovitost poduzetih mjera i potencijalne rizike.

Na slici 4.1. prikazana je pojednostavljena arhitektura skladišta podataka koje djeluje kao središnje mjesto kroz koje prolaze podaci različitih formata: od tradicionalnih CSV datoteka i polustrukturiranih JSON/XML zapisa do parquet formata i podataka preuzetih iz API-ja ili baza drugih sustava. Svi ti podaci ulaze u jedinstveni sustav u kojem se organiziraju, povezuju i pripremaju za analitičku uporabu. Time skladište podataka postaje pouzdana infrastruktura koju koriste istraživači, institucije i donositelji odluka, čime se omogućuje kontinuirano nadograđivanje i razvoj politika temeljenih na dokazima.

Nakon što se podaci učitaju u skladište, otvaraju se mogućnosti za različite oblike analize, i to od osnovnih statističkih prikaza i vizualizacija, preko utvrđivanja međusobne povezanosti među varijablama, do primjene naprednih metoda strojnog učenja i umjetne inteligencije. U području okoliša, to primjerice može omogućiti predviđanje epizoda zagađenja, procjenu utjecaja meteoroloških uvjeta na kvalitetu zraka te analizu dugoročnih trendova koji ukazuju na promjene u okolišnim procesima i sl.



Slika 4.1 Pojednostavljeni prikaz arhitekture skladišta podataka

Tehnološki kontekst

Danas je moguće koristiti vrlo širok raspon tehnologija za izgradnju skladišta podataka, koje se međusobno razlikuju prema skalabilnosti, zahtjevnosti upravljanja, dostupnim funkcionalnostima i ukupnim troškovima implementacije. Odabir odgovarajuće tehnologije ovisi o specifičnim potrebama korisnika, dostupnim resursima, vrsti podataka i planiranim analitičkim postupcima. Na jednom kraju spektra nalaze se lokalne instalacije tradicionalnih relacijskih baza podataka koje nude potpunu kontrolu nad sustavom, ali zahtijevaju značajnu tehničku podršku i redovito održavanje. Na drugom su kraju komercijalne platforme u oblaku, koje osiguravaju visoku razinu automatizacije, izvrsne performanse i mogućnost obrade vrlo velikih količina podataka, uz različite modele naplate i razine operativne složenosti.

Za potrebe ovog izdanja odabrana je Oracle tehnologija autonomne baze podataka (engl. *Oracle Autonomous Database*) [1] jer omogućuje izgradnju skladišta podataka u okruženju koje je stabilno, trajno dostupno i u najvećoj mjeri automatizirano. Kao dio Oracleova oblaka [2] (engl. *Oracle Cloud*) dostupna je i u okviru trajno besplatnog resursa (engl. *Always Free*), što početnicima olakšava rad bez dodatnih infrastrukturnih troškova. Ugrađeni mehanizmi automatskog upravljanja sigurnošću, konfiguracijom i nadogradnjama uklanjaju tehničke prepreke koje mogu predstavljati izazov korisnicima s manje iskustva te omogućuju da se pozornost usmjeri na analitički rad i interpretaciju podataka.

Naravno, uporaba ove platforme ne isključuje mogućnost primjene drugih rješenja. Principi prikazani u ovom poglavlju primjenjivi su i na druge tehnologije, pa zainteresirani čitatelji, ovisno o vlastitim potrebama i resursima, mogu razmotriti i alternativna rješenja poput Amazon Redshift [3], Google BigQuery [4], Microsoft Azure Synapse Analytics [5] ili sustave otvorenog koda kao što je npr. Apache Hive [6]. Svaka od ovih tehnologija nudi specifičan pristup skladištenju podataka, različite modele skaliranja i svoje prednosti, ali temeljna logika integracije podataka i njihove analitičke obrade ostaje nepromijenjena.

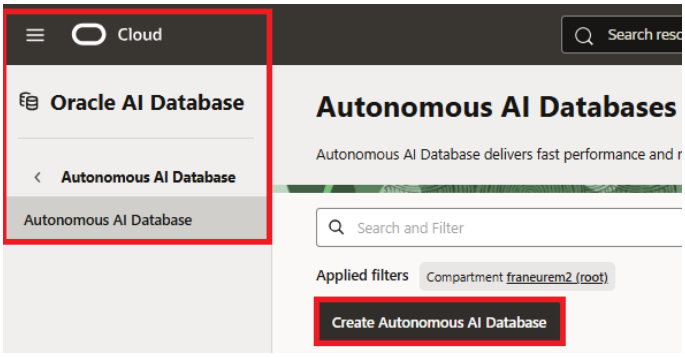
Kreiranje i konfiguriranje skladišta podataka

Postupak instalacije *Oracle Autonomous Database* skladišta podataka započinje prijavom na Oracle Cloud korisnički račun. Besplatan račun moguće je otvoriti putem *Oracle Academy* programa [7] ili prema uputama koje su dostupne na [8]. Nakon uspješne prijave, prvi korak jest odabir izbornika *Oracle AI Database*, a zatim opcije *Autonomous AI Database*, kao što je prikazano na slici 4.2. Time se otvara pregled svih postojećih baza podataka u korisničkom računu.

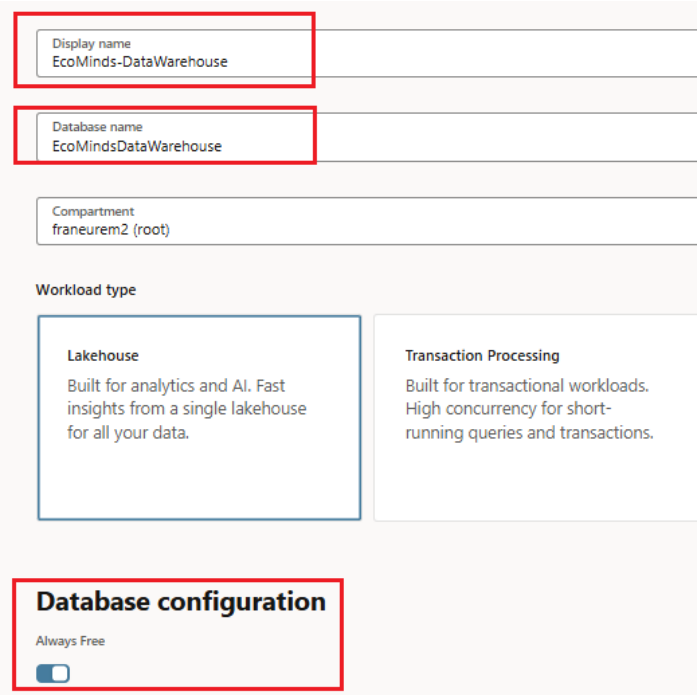
Za stvaranje nove baze podataka aktivira se poveznica *Create Autonomous Database*. U slijedećem obrascu unose se osnovni podaci o instanci, poput naziva baze podataka, korisničkog računa u oblaku na kojem će biti instalirana, te se odabire radno okruženje instance. Za potrebe izrade skladišta podataka preporučuje se odabrati tzv. *Lakehouse* okruženje, koje uključuje dodatne alate za administraciju baze, izvršavanje SQL upita i primjenu algoritama strojnog učenja nad pohranjenim skupovima podataka.

U sljedećem koraku potrebno je postaviti trajno besplatno korištenje i odabrati verziju baze podataka (najčešće je unaprijed ponuđena najčešće korištena verzija uz mogućnost nadogradnje na najnoviju). Nakon toga definiraju se administratorske vjerodajnice, odnosno lozinka za administratora sustava (ADMIN korisničko ime), koji u početnoj fazi ima puni pristup sustavu te služi za stvaranje novih korisnika, učitavanje podataka i osnovnu konfiguraciju baze. Lozinka mora biti u skladu s propisanim sigurnosnim pravilima. Posljednji korak odnosi se na odabir načina mrežnog pristupa instanci.

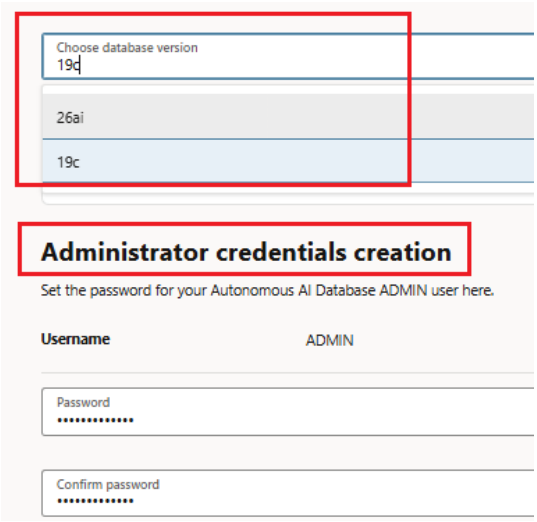
U trajno besplatnom okruženju najčešće se koristi opcija sigurnog pristupa bazi putem interneta s bilo kojeg uređaja (engl. *Secure access from everywhere*), pri čemu se upotrebljava digitalni novčanik (engl. *wallet*) i autentifikacijski mehanizmi ugrađeni u autonomnoj bazi podataka. Takav pristup znatno pojednostavljuje rad u obrazovnom okruženju, jer ne zahtijeva korištenje određene mreže niti dodatne infrastrukturne komponente, poput VPN tunela ili ograničavanja pristupa na pojedine IP adrese.



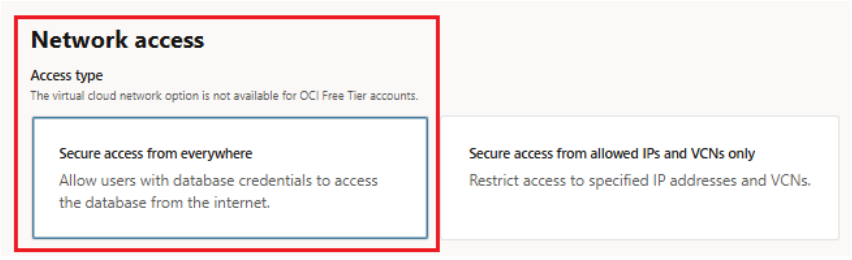
Slika 4.2 Pristup Oracle Autonomous Database platformi



Slika 4.3 Odabir imena i radnog okruženja skladišta podataka



Slika 4.4 Odabir verzije i završno konfiguriranje skladišta podataka



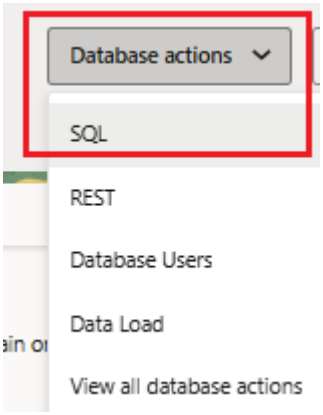
Slika 4.5 Odabir mrežnog pristupa

Autentikacija i razvojne okoline

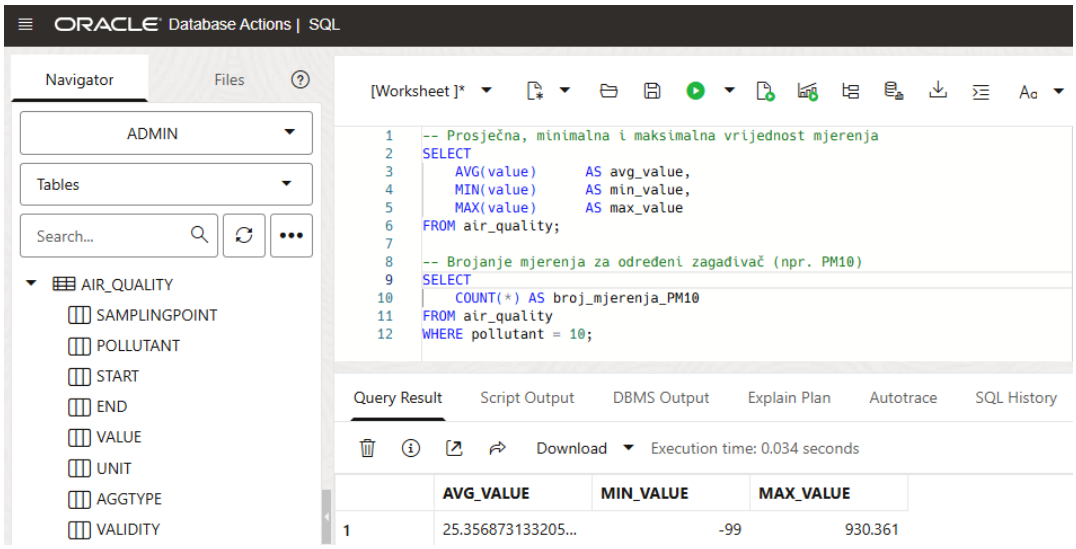
Sve analize okolišnih podataka u skladištu podataka temelje se na korištenju SQL jezika, kojim se podaci dohvaćaju, filtriraju, agregiraju i povezuju u cjelovite analitičke prikaze. Kako bi se takvi upiti mogli izvoditi, istraživači moraju imati pristup skladištu podataka i pouzdanu razvojnu okolinu.

Postoje dva osnovna načina povezivanja s Oracle autonomnom bazom podataka. Prvi, ujedno i najjednostavniji način jest korištenje *Oracle Cloud* web sučelja. Tijekom prijave aktivira se dvofaktorska autentikacija (engl. *Two-Factor Authentication- 2FA*), čime se osigurava pristup svim potrebnim resursima. Nakon uspješne prijave bira se opcija *Database Actions*, a zatim *SQL*, čime se otvara integrirana web razvojna okolina za izvršavanje SQL naredbi (slike 4.6 i 4.7). Prednost ovog pristupa jest to što ne zahtijeva nikakve dodatne instalacije, a budući je dostupan s bilo kojeg uređaja i preglednika, posebno je pogodan za početak rada.

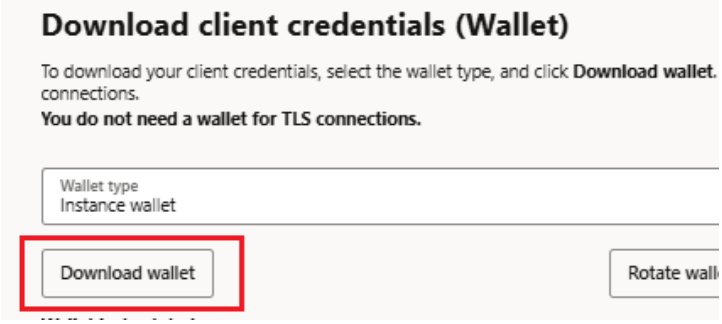
Drugi način pristupa jest korištenje klijentske desktop aplikacije *Oracle SQL Developer*, dostupne za besplatno preuzimanje na [9]. *SQL Developer* omogućuje naprednije upravljanje skriptama, rad na složenijim projektima, pregledavanje strukture baze i izvođenje administrativnih zadataka. Za povezivanje putem ove aplikacije potrebno je preuzeti digitalni novčanik (engl. *digital wallet*) iz *Oracle Cloud* konzole, odabirom opcije *Database connection* (slika 4.8). Paket sadrži certifikate i konfiguracijske datoteke potrebne za uspostavu sigurne veze prema skladištu podataka pomoću mTLS protokola koji osigurava obostranu autentikaciju klijenta i poslužitelja. Nakon učitavanja digitalnog novčanika u SQL Developer (slika 4.9) moguće je uspostaviti vezu i raditi sa skladištem podataka u standardnom desktop radnom okruženju (slika 4.10).



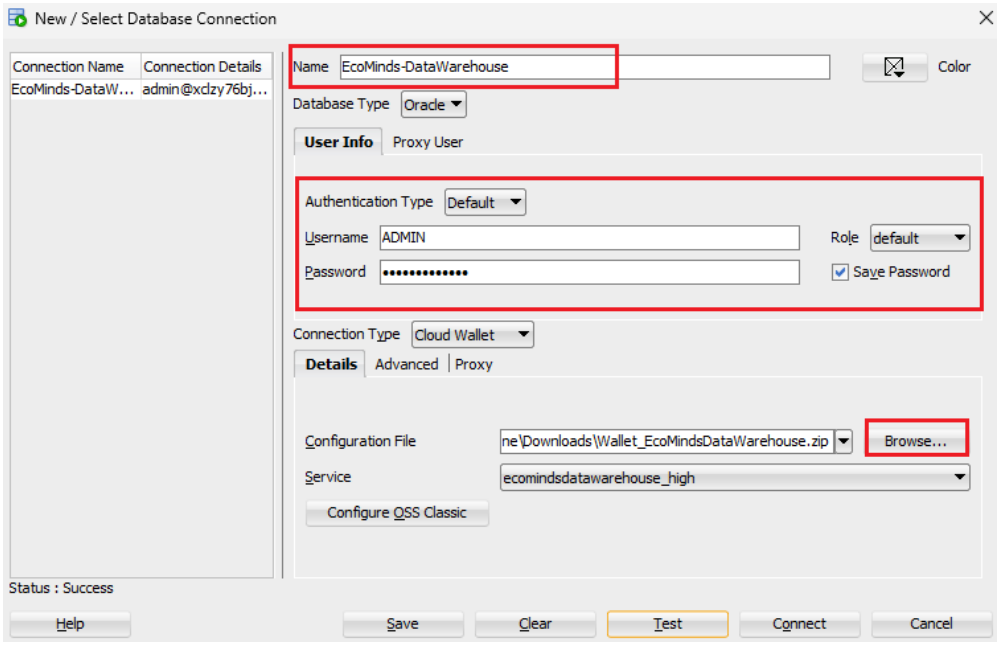
Slika 4.6 Pokretanje SQL web konzole



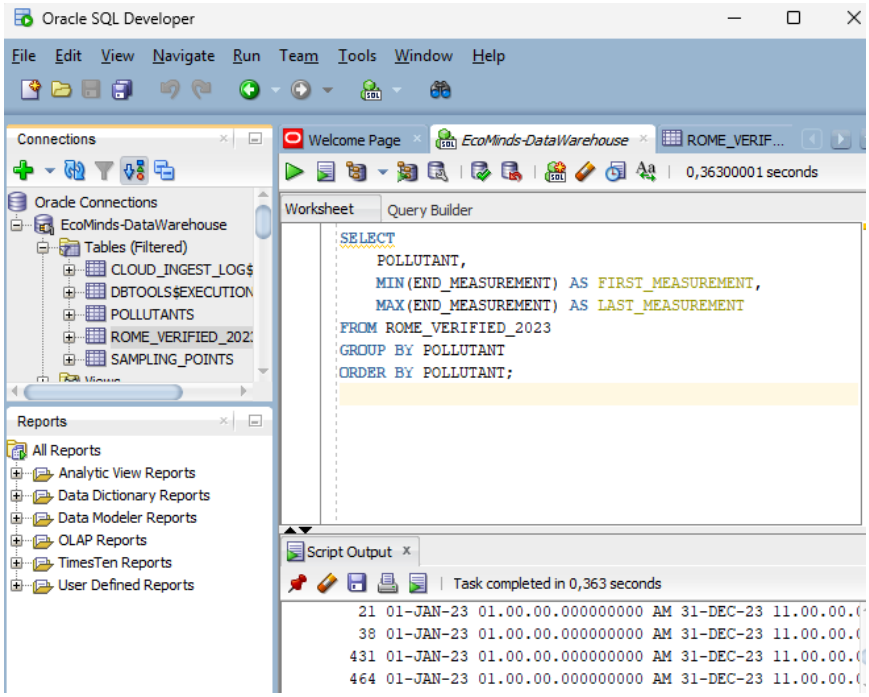
Slika 4.7 Prikaz rada na web konzoli



Slika 4.8 Preuzimanje digitalnog novčanika



Slika 4.9 Podešavanje veze na skladište podataka iz SQL Developer aplikacije



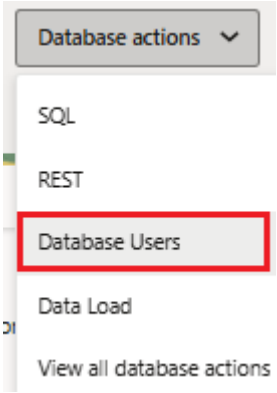
Slika 4.10 Prikaz rada u SQL Developer aplikaciji

Administriranje korisničkih prava

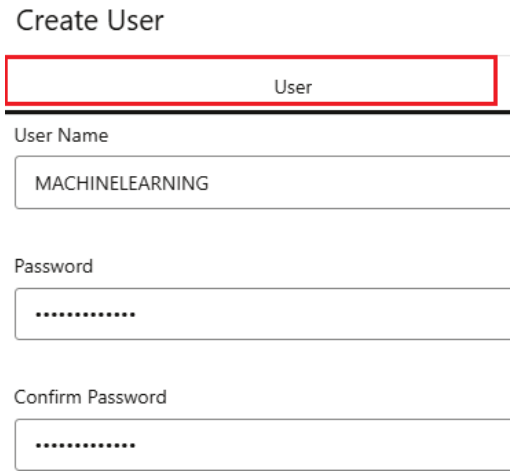
Nakon što je baza podataka kreirana, prvi dostupni korisnik u sustavu jest ADMIN. U edukacijskim okruženjima, osobito tijekom uvodnih vježbi, privremeno je moguće raditi s ADMIN računom kako bi se pojednostavilo učitavanje podataka i izvršavanje osnovnih SQL upita. Iako je takav pristup praktičan za početno upoznavanje sustava, ne smatra se prihvatljivim u profesionalnoj primjeni.

U produkcijskim okruženjima ADMIN se koristi isključivo za administrativne zadaće, poput stvaranja korisnika, dodjele privilegija i konfiguracije baze podataka. Budući da ADMIN posjeduje sve ovlasti, njegova uporaba za svakodnevni rad predstavlja sigurnosni rizik. Stoga se za korisnike koji provode analitičke ili razvojne zadatke kreiraju posebni računi ili uloge (roles) s ograničenim pravima pristupa, u skladu s načelom najmanjih privilegija.

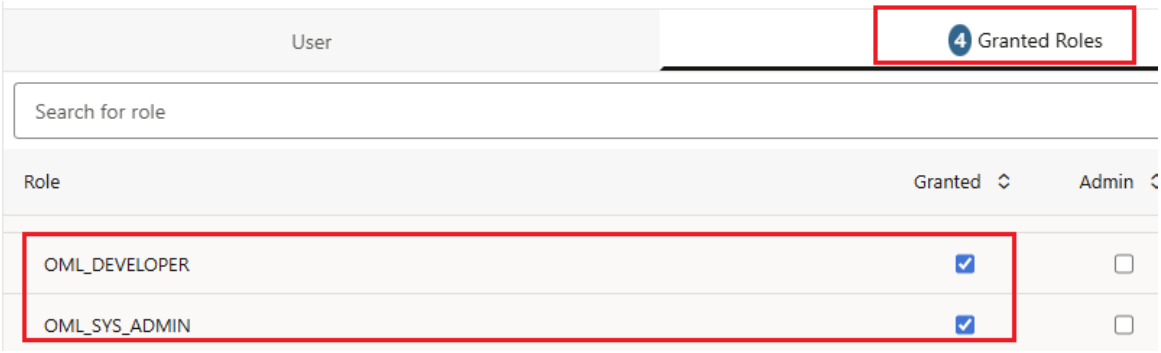
Važno je naglasiti da ADMIN nema pristup svim specijaliziranim alatima dostupnima u okruženju Autonomous Database. Pojedine funkcionalnosti, poput Oracle Machine Learninga (OML), graf-analitike ili određenih web-komponenti, zahtijevaju korisničke račune s posebno dodijeljenim ovlastima. Zbog toga se za rad s naprednim alatima uobičajeno stvaraju zasebni korisnici s jasno definiranom razinom pristupa, kao što je prikazano na slikama 4.11. - 4.13.



Slika 4.11 Odabir admistriranja korisnika



Slika 4.12 Unos novog korisnika



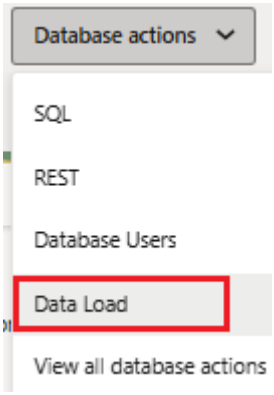
Slika 4.13 Definiranje korisničkih prava

Priprema i učitavanje podataka

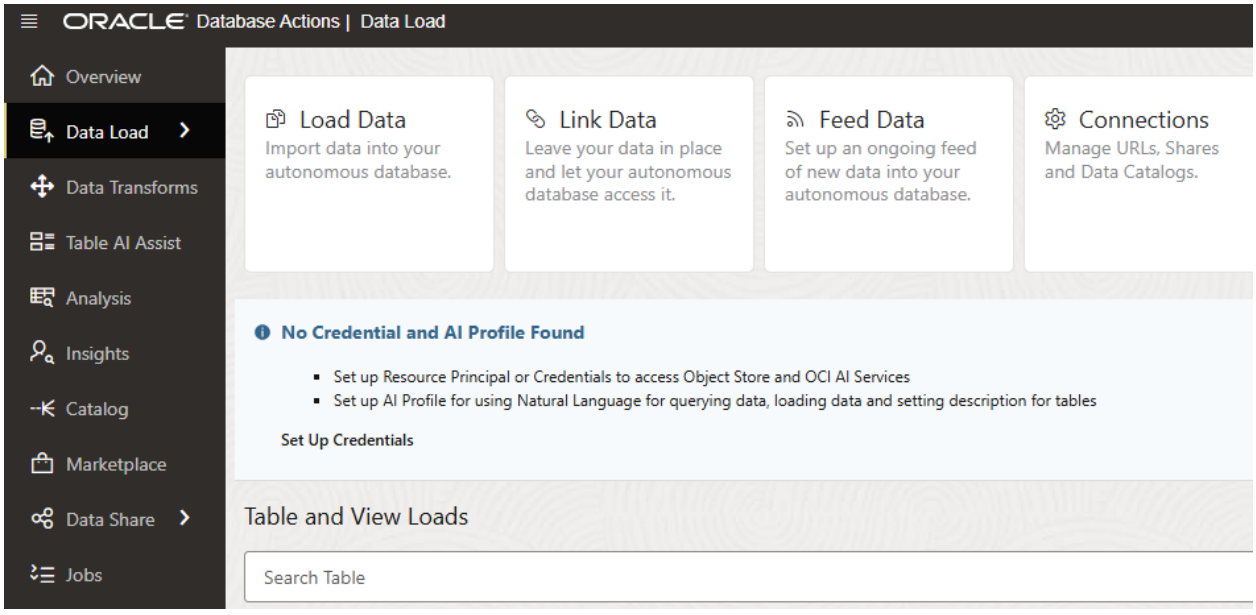
Kvalitetna priprema podataka ključan je preduvjet uspješnog rada skladišta podataka. Nakon što je skladište podataka s instancom autonomne baze spremno za korištenje,svi podaci se

učitavaju putem jedinstvenog Data Load sučelja koje se pokreće iz izbornika Database Actions (slike 4.14 i 4.15).

Najjednostavniji način uvoza u skladište podataka predstavlja učitavanje lokalnih datoteka koje se prenose jednostavnim povlačenjem (engl. *drag & drop*) na radnu površinu *Data Load* radne okoline, nakon čega sustav automatski prepoznaje strukturu, predlaže nazive stupaca, tipove podataka i nudi mogućnost stvaranja nove tablice. Ovaj način rada pogodan je za manje skupove podataka u CSV, JSON ili sličnim formatima i ne zahtijeva dodatne alate.



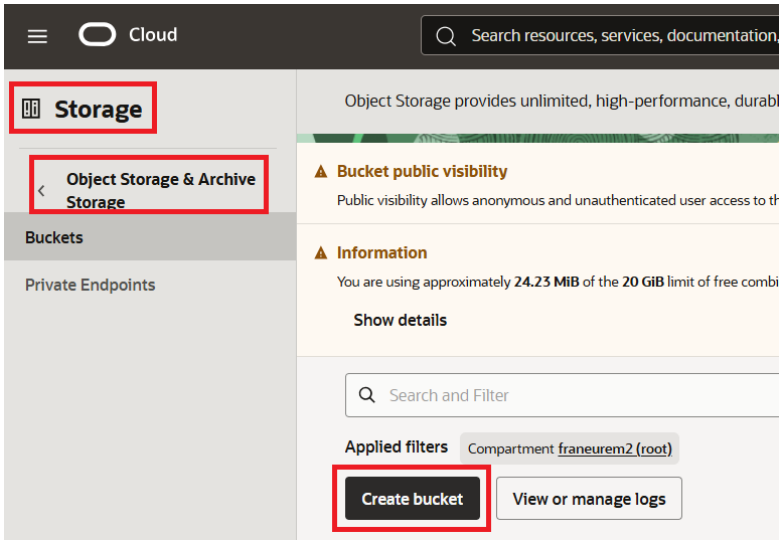
Slika 4.14 Pokretanje Data Load radne okoline



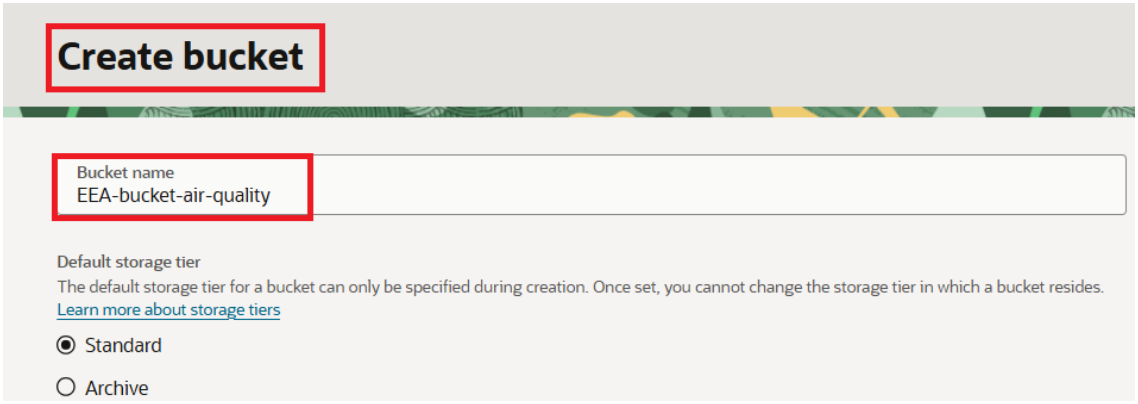
Slika 4.15 Prikaz Data Load radne okoline

Za veće datoteke, osobito u parquet formatu, preporučuje se njihovo prethodno učitavanje na *Oracle Object Storage* servis unutar Oracle oblaka. U navedenom repozitoriju podaci se pohranjuju u spremnike (engl. *buckets*), a u trajno besplatnom načinu korištenja osigurano je do 20 GB radnog prostora, što je dovoljno za brojne analitičke primjene.

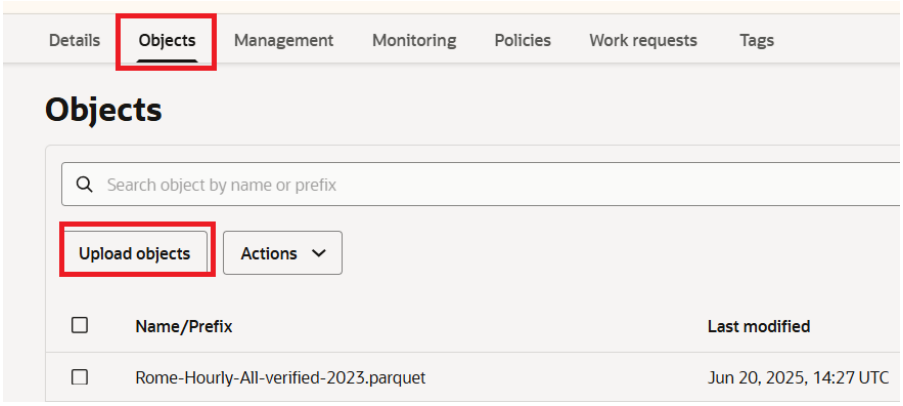
Spremnik se jednostavno izrađuje putem izbornika *Storage* - *Object Storage*, gdje je potrebno zadati naziv spremnika i potvrditi zadane postavke (slike 4.16 i 4.17). Nakon toga moguće je učitati pripremljene datoteke koje će se kasnije preuzeti u skladište podataka (slika 4.18).



Slika 4.16 Izrada novog Oracle Object Storage spremnika

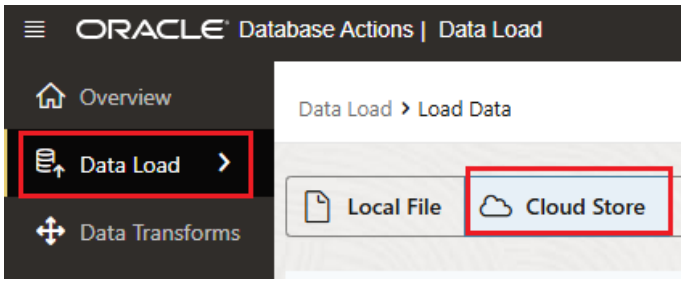


Slika 4.17 Izrada novog Oracle Object Storage spremnika



Slika 4.18 Učitavanje datoteka u spremnik

Kada su datoteke pohranjene u spremniku, njihov se uvoz u bazu ponovo izvodi iz istog *Data Load* sučelja. Umjesto lokalne datoteke, odabire se izvor *Cloud Store* (slika 4.19), a zatim se unosi URL objekta iz spremnika. U nastavku sustav automatski mapira podatke i kreira pripadajuću tablicu.



Slika 4.19 Učitavanje podataka iz Object Storagea putem Data Load modula

Modeliranje podataka

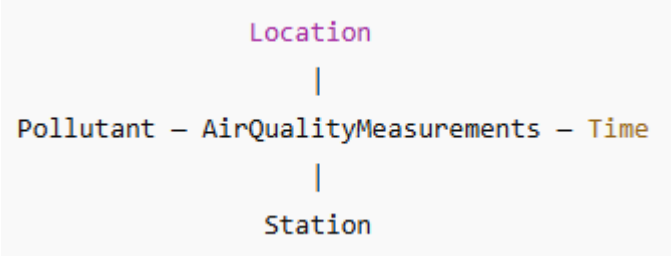
Nakon učitavanja podataka, sljedeći je korak njihovo modeliranje s ciljem jednostavnije buduće analize. Pri tome se posebna pažnja pridaje preglednoj i razumljivoj strukturi samih tablica u bazi podataka kako bi se ostvarilo optimalno izvođenje SQL upita.

U početnoj fazi tablice se najčešće stvaraju izravno iz učitanih datoteka. Pri tome se svaki skup podataka (primjerice mjerenja kvalitete zraka ili meteorološki zapisi) organizira kao zasebna tablica, a svaki redak se odnosi na pojedinačno mjerenje. Takve tablice obično se nazivaju sirovim tablicama (engl. *raw*) jer preslikavaju izvorni oblik podataka bez dodatnih transformacija. Iako je nad njima moguće provoditi osnovne analize, njihova se struktura može dodatno uređivati radi lakšeg i pouzdanijeg izvođenja upita.

Za složenije analize korisno je unaprijed uskladiti nazive stupaca, formate datuma i identifikatore lokacija ili mjernih postaja. Dosljedna struktura olakšava spajanje tablica i smanjuje rizik od pogrešaka pri interpretaciji rezultata. U određenim situacijama može se primijeniti i osnovna denormalizacija, primjerice dodavanjem informacija o mjernoj postaji iz pomoćne tablice izravno u tablicu mjerenja, čime se pojednostavljuju upiti koji bi inače zahtijevali česta spajanja tablica.

Jedan od najčešćih i konceptualno najjednostavnijih načina organizacije podataka jest tzv. zvjezdasta shema (engl. *star schema*) koje je prikazana na slici 4.20. U njoj se jedna središnja tablica (primjerice tablica mjerenja) koristi kao glavni izvor podataka, dok se opisni podaci o lokacijama mjerenja, mjernim postajama ili drugim kategorijama nalaze u manjim tablicama. Te

se tablice, poznate kao dimenzijske tablice, prema potrebi spajaju sa središnjom tablicom i omogućuju jednostavnije i preglednije analitičke upite.

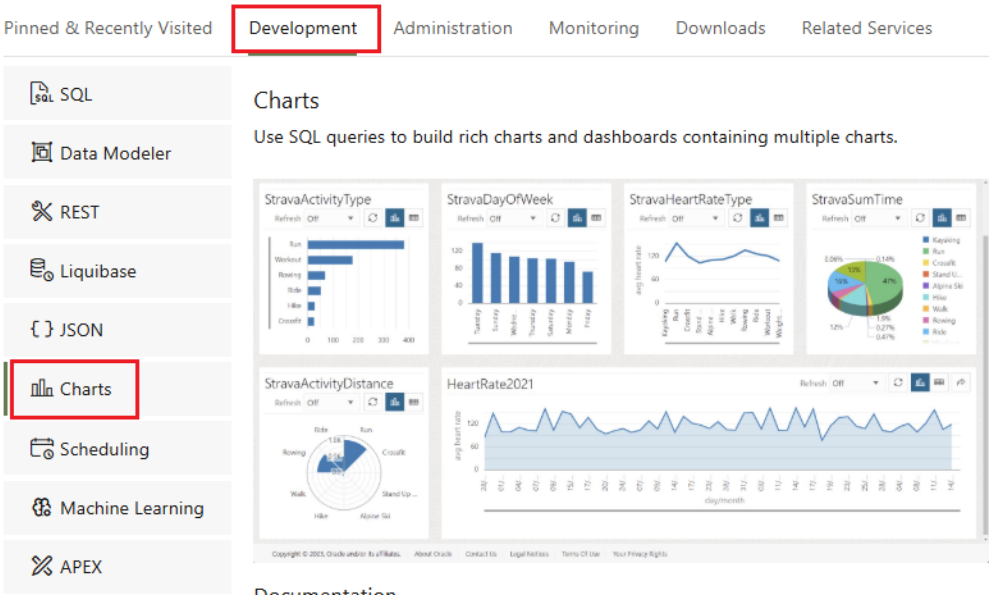


Slika 4.20 Jednostavna zvjezdasta shema

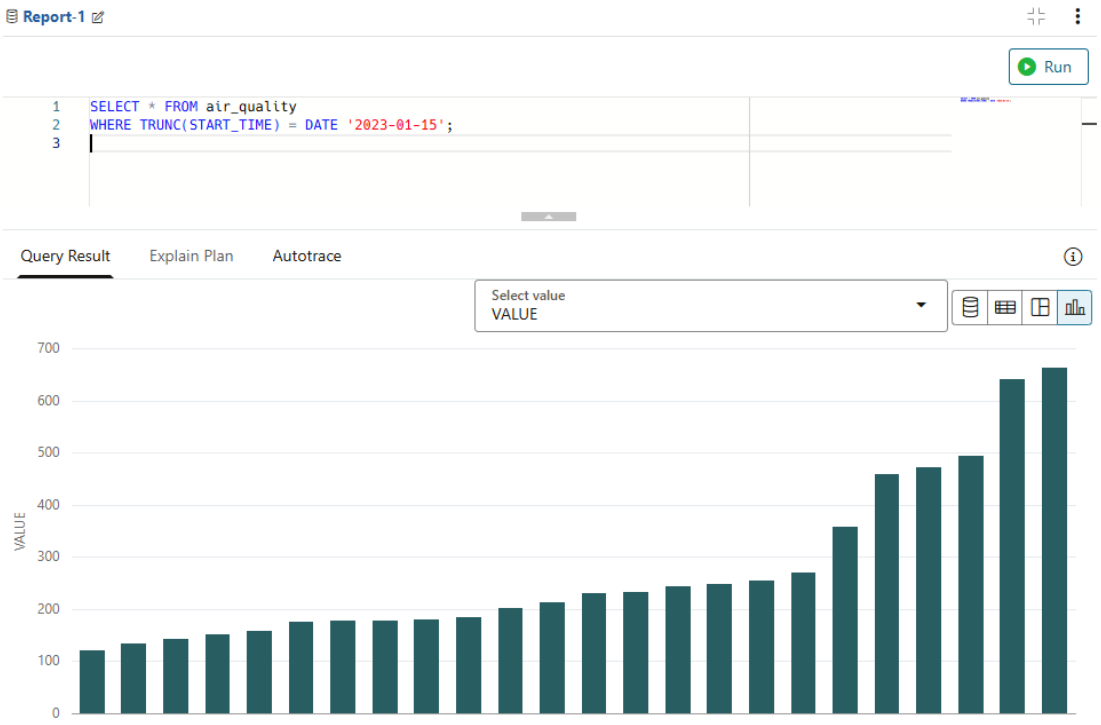
Vizualizacija podataka

Nakon izvođenja SQL upita, rezultati se mogu prikazati u obliku grafova korištenjem ugrađenog modula *Charts*. Do modula se dolazi putem izbornika *View all database actions*, zatim odabirom kartice *Data Studio*, unutar koje se nalazi opcija *Charts* kao na slici 4.21. Ovaj alat omogućuje izradu jednostavnih vizualizacija izravno nad rezultatima SQL upita, bez potrebe za vanjskim aplikacijama ili dodatnim instalacijama. Time se omogućuje brza i pregledna interpretacija podataka, osobito pri analizama trendova, usporedbama među skupovima podataka ili uočavanju odstupanja.

Tako primjerice možemo zamisliti tablicu koja sadrži satna mjerenja koncentracije određenog polutanta u zraku tijekom duljeg razdoblja. U tom slučaju često postoji potreba da se izdvoje samo dnevna mjerenja. U prikazanom primjeru korišten je stupčasti grafikon: na osi X nalaze se vremena mjerenja tijekom odabranog dana, dok su na osi Y prikazane njihove pripadajuće vrijednosti (slika 4.22).



Slika 4.21 Pokretanje Charts modula



Slika 4.22 Primjer jednostavne vizualizacije podataka

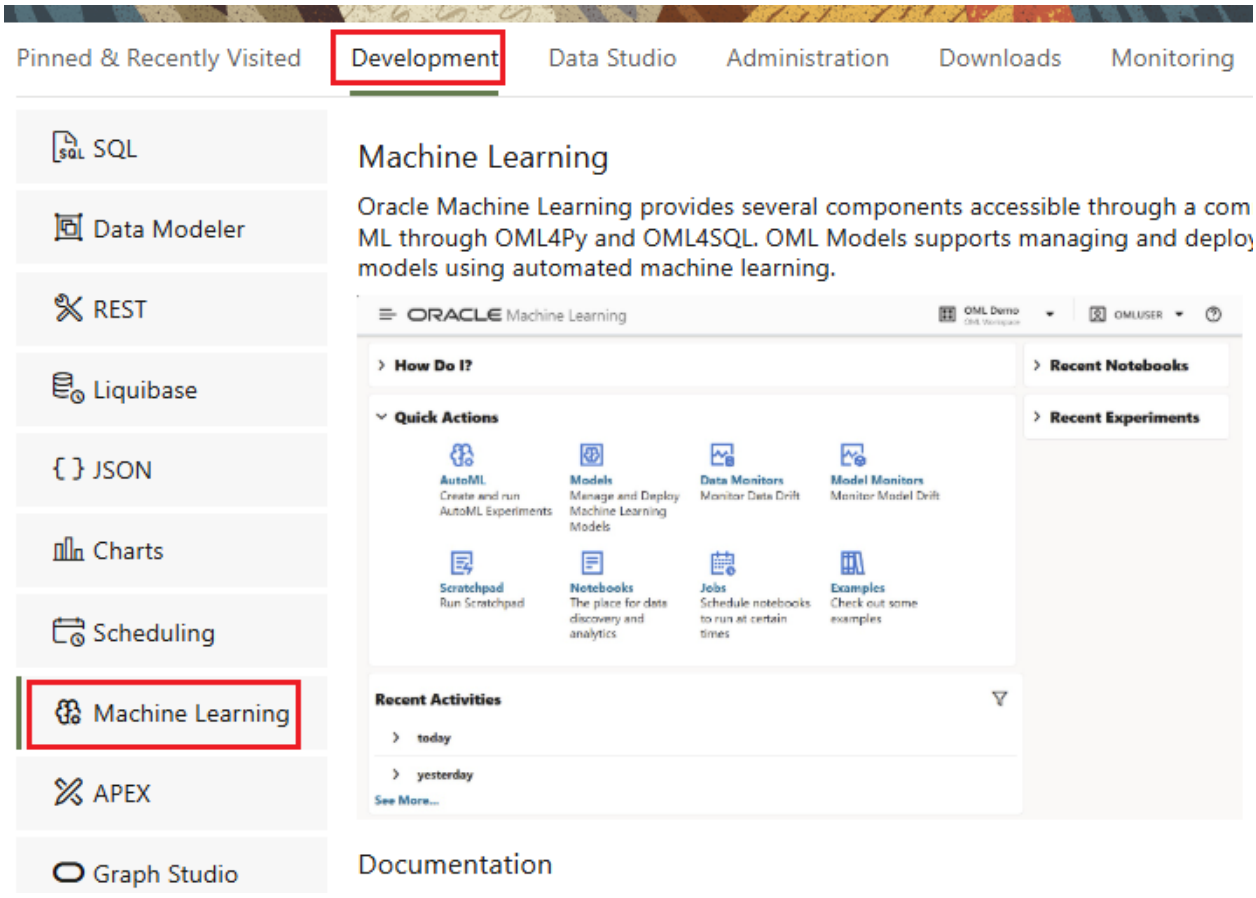
Strojno učenje u bazi podataka

Skladište podataka s Oracle autonomnom bazom podataka uključuje ugrađeni sustav za strojno učenje pod nazivom Oracle Machine Learning (OML), koji omogućuje treniranje i primjenu modela izravno nad podacima pohranjenima u bazi podataka. Za razliku od pristupa u kojem je

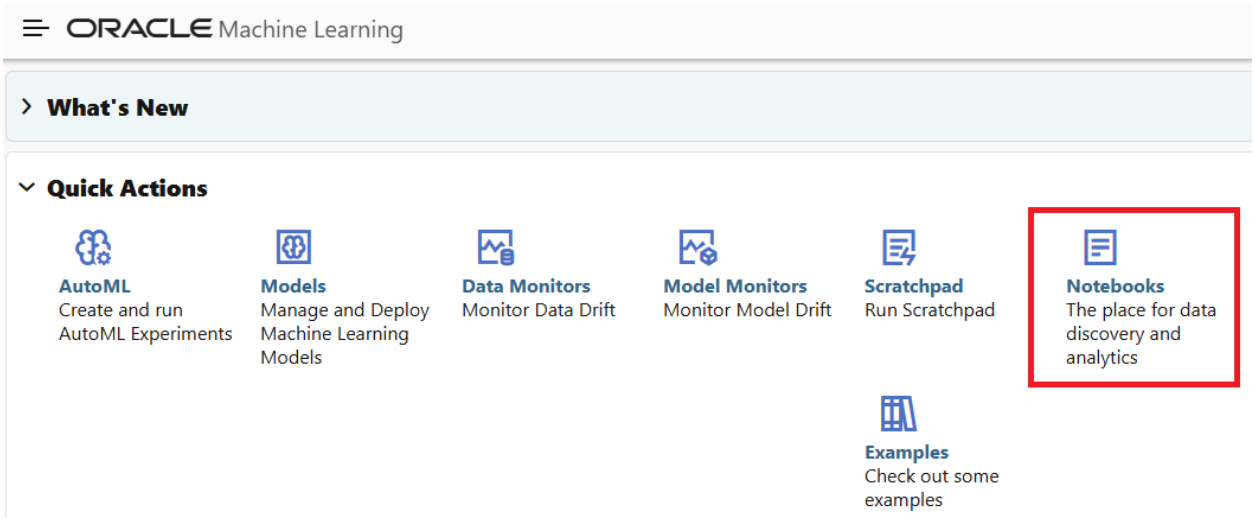
potrebno izvoziti podatke u vanjske alate, OML omogućuje da se priprema podataka, izrada modela i evaluacija provedu unutar baze, bez dodatne infrastrukture i bez premještanja velikih skupova podataka.

Pristup OML okruženju ostvaruje se putem izbornika *Database Actions - View all database actions*, nakon čega se odabire kartica *Development* i izbornik *Machine Learning* (slika 4.23). Time se otvara OML radno okruženje, prikazano na slici 4.24.

Najjednostavniji način rada u OML-u jest korištenje *OML Notebooks* radne okoline (slika 4.24), koje omogućuje pisanje i izvršavanje SQL i Python skripti izravno u web pregledniku. Dostupne su funkcionalnosti za klasifikaciju, regresiju, klasteriranje, otkrivanje anomalija i automatizirano treniranje modela.



Slika 4.23 Pokretanje OML radne okoline



Slika 4.24 Prikaz OML radne okoline

Studija slučaja: Analiza zagađenja zraka

U ovoj studiji slučaja prikazuje se rad sa stvarnim podacima o kvaliteti zraka kakvi se u praksi koriste u europskim okolišnim agencijama, istraživačkim ustanovama i analitičkim sustavima. Riječ je o satnim mjerenjima dostupnima putem Europske agencije za okoliš (EEA).

Cilj primjera jest prikazati cjelovit postupak rada s okolišnim skupovima podataka. U početnom dijelu opisuje se način preuzimanja podataka i njihove pohrane u skladište podataka, dok se u nastavku prikazuje proces analize podataka na stvarnom problemu iz područja zaštite okoliša, primjerice pri praćenju koncentracija dušikova dioksida (NO_2) ili lebdećih čestica $PM_{2.5}$.

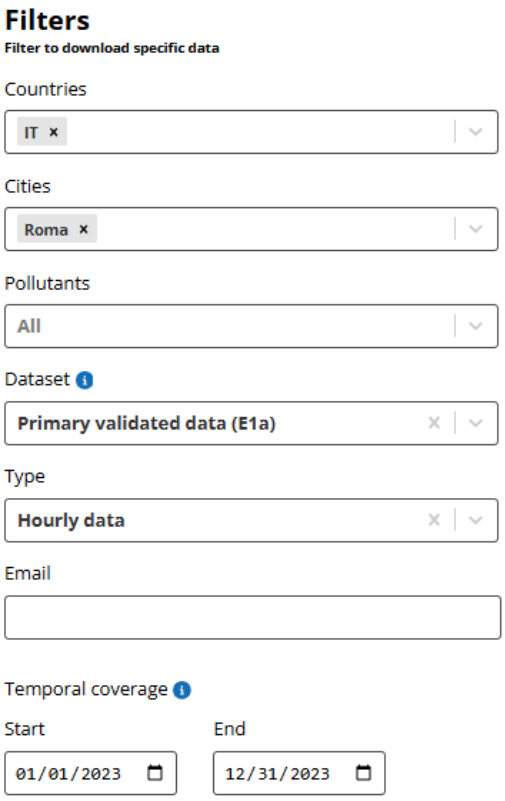
Europski izvori podataka o kvaliteti zraka

Podaci o kvaliteti zraka mogu se preuzeti iz sustava Europske agencije za okoliš (European Environment Agency – EEA) [10], koja prikuplja, validira i objavljuje mjerenja iz nacionalnih mreža za praćenje okoliša. Sustav EEA pruža pristup detaljnim satnim, dnevnim i godišnjim zapisima koncentracija različitih polutanata, uključujući dušikov dioksid (NO_2), ozon (O_3), lebdeće čestice $PM_{2.5}$ i PM_{10}, kao i druge tvari relevantne za procjenu kvalitete zraka.

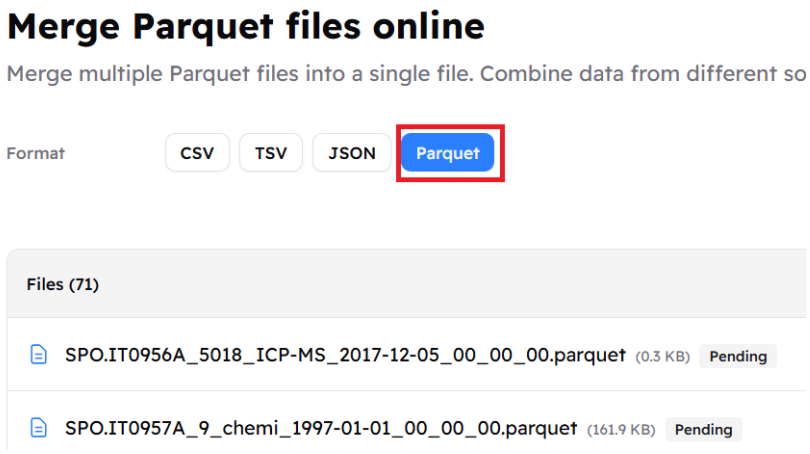
Mjerna mjesta raspoređena su diljem Europe i koriste standardizirane metode prema europskim direktivama o praćenju kvalitete zraka. Prije objave podaci prolaze provjeru kojom se uklanjaju pogreške, nelogične vrijednosti i nedostajući zapisi. Tako pripremljeni podaci dostupni su putem službenih EEA servisa [11], uključujući mogućnost preuzimanja u parquet i CSV formatu.

Nakon odabira geografskog područja (slika 4.25), vremenskog razdoblja i željenog intervala mjerenja, sustav generira paket s odabranim mjerenjima. Preuzeta datoteka obično sadrži više parquet zapisa, po jedan za svaku mjernu stanicu na odabranom području. Takav raspored omogućuje detaljnu analizu, no u analitičkim scenarijima često je korisno objediniti sve zapise u jedinstvenu tablicu. To se može učiniti pomoću dostupnih alata za spajanje parquet datoteka, poput servisa *Agents for Data* [12], u kojem se više datoteka jednostavnim povlačenjem objedinjuje u jedan skup podataka (slika 4.26).

Također je važno naglasiti potrebu poznavanja strukture preuzetih podataka, njihovih atributa i značenja, što omogućuju objavljene poveznice za proučavanje metapodataka (engl. *metadata*), rječnika podataka (engl. *vocabulary*) i tehničke dokumentacije za objavljena API sučelja.



Slika 4.25 Preuzimanje EEA podataka

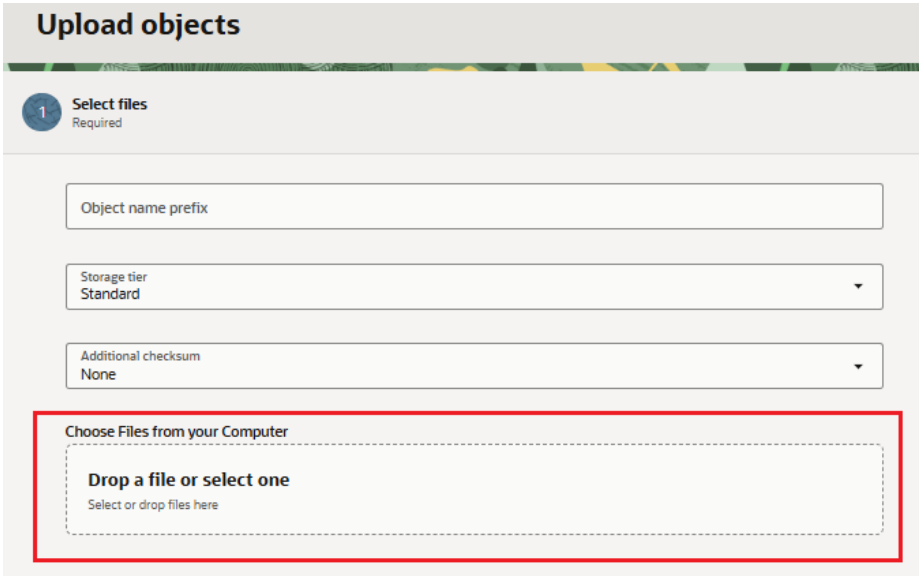


Slika 4.26 Spajanje više parquet datoteka na Agents for Data platformi

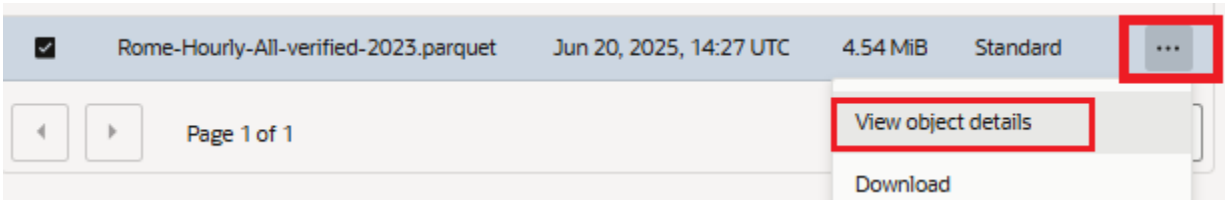
Pohrana EEA mjerenja na Oracle Object Storage

Nakon preuzimanja podataka sa servisa EEA, sljedeći je korak njihova pohrana u prethodno pripremljeni spremnik na *Oracle Object Storage* servisu. Unutar spremnika odabire se kartica *Objects*, zatim naredba *Upload objects*, nakon čega se učita parquet datoteka s mjerenjima (slika 4.27).

Svaki učitani objekt dobiva vlastitu URL oznaku, kojoj je moguće pristupiti putem pregleda svojstava objekta (slika 4.28). Ta se oznaka koristi pri kasnijem uvozu podataka u skladište podataka.



Slika 4.27 Učitavanje EEA datoteke u spremnik



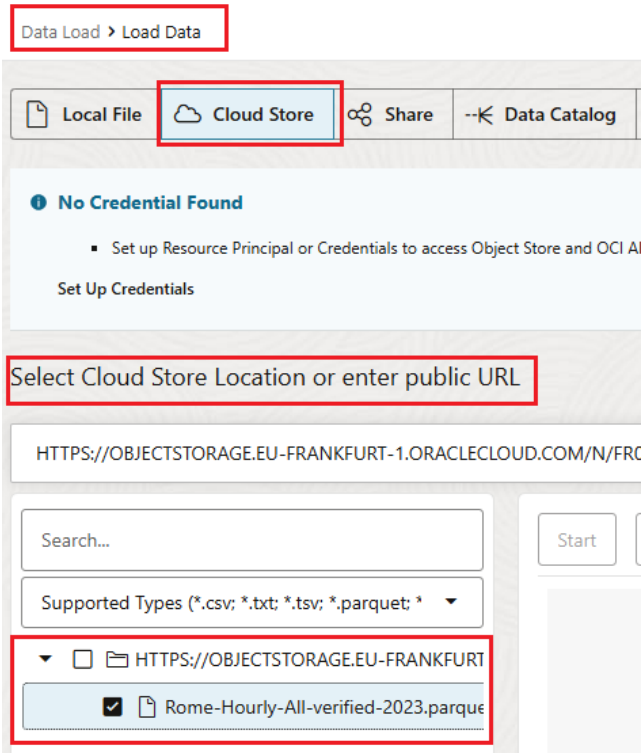
Slika 4.28 Dohvat URL oznake objekta

Uvoz mjerenja u skladište podataka

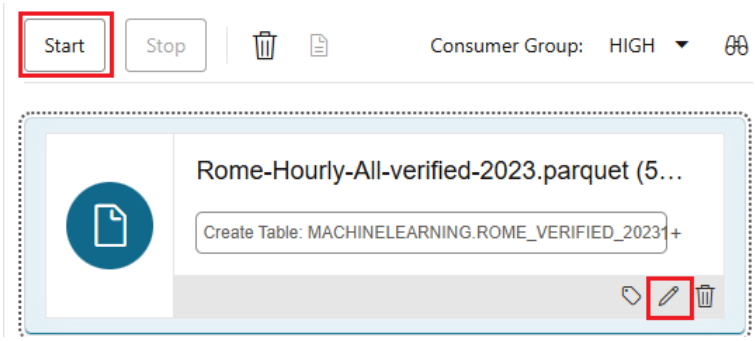
U sljedećem koraku otvara se sučelje Data Load u Autonomous Database okruženju te se odabire kartica Cloud Store. U polje za izvor podataka upisuje se prethodno kopiran URL oznaka parquet objekta pohranjenog u spremniku (slika 4.29).

U lijevom panelu prikazuje se struktura dostupnih podataka, koje je moguće odabrati i jednostavno povući na radnu površinu radi pripreme uvoza. Proces učitavanja započinje klikom na gumb *Start* (slika 4.30).

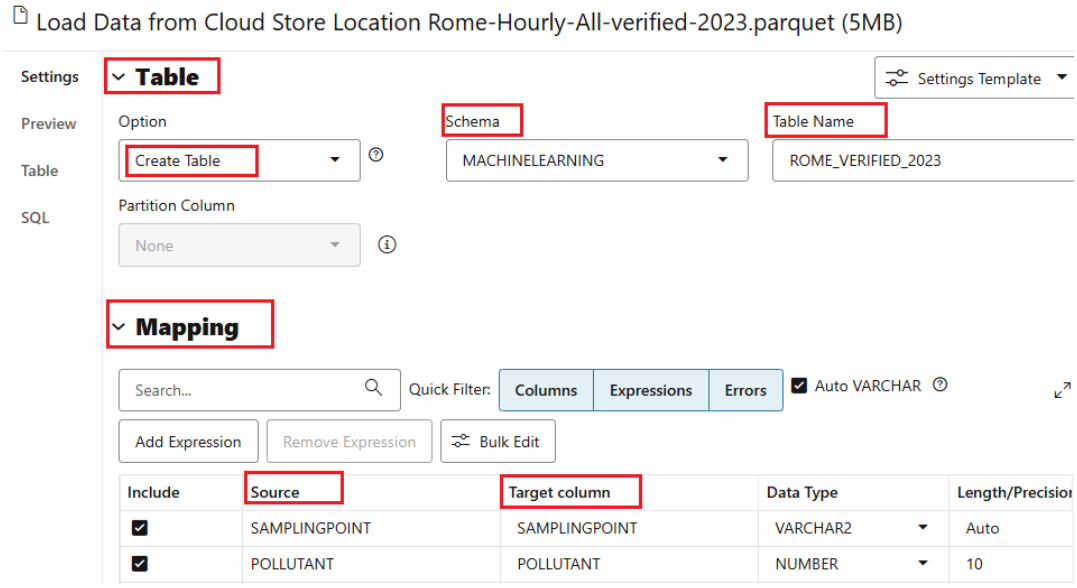
Prije samog uvoza preporučljivo je otvoriti poveznicu *Settings*, gdje se može pregledati struktura datoteke i podesiti ključni parametri prijenosa: stvaranje nove tablice ili nadopunjavanje postojeće, mapiranje stupaca te odabir korisničke sheme (slika 4.31).



Slika 4.29 Priprema za učitavanje mjerenja



Slika 4.30 Priprema za učitavanje mjerenja



Slika 4.31 Podešavanje parametara prijenosa mjerenja u skladište podataka

Primjeri početne analize učitanih podataka

U nastavku su navedeni primjeri osnovnih upita koji omogućuju pregled sadržaja tablice i stjecanje uvida u dostupne atribute. Takav početni pregled koristan je za razumijevanje strukture podataka, raspona mjerenja i načina na koji su polutanti i mjerne stanice identificirani.

Pregled svih zagađenja koja se pojavljuju u tablici mjerenja je moguće prikazati slijedećim SQL upitom:

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

Zatim je korisno izdvojiti popis svih mjernih stanica:

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

Također je moguće utvrditi raspon izmjerenih vrijednosti za sva zabilježena zagađenja:

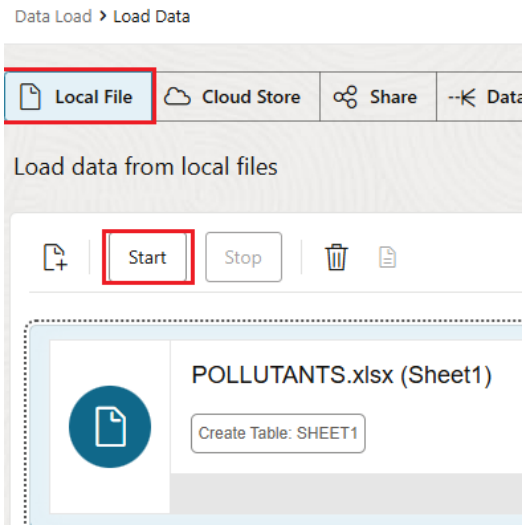
```
-------------------------------------------------------------------------------
-- 3. Determine the measurement range for each pollutant
--    Shows the first and last measurement timestamp per pollutant.
-------------------------------------------------------------------------------
SELECT
    POLLUTANT,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

Slično kao kod prethodnog upita, moguće je prikazati raspone izmjerenih zagađenja grupirane po mjernim stanicama:

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

Povezivanje mjerenja s opisima zagađenja

Nažalost, iz tablice s mjerenjima kvalitete zraka (ROME_VERIFIED_2023) nije moguće zaključiti o kojim je zagađujućim tvarima riječ, jer se u njoj navode samo njihove brojčane oznake. Problem se rješava povezivanjem ove tablice s dodatnom tablicom koja sadrži detaljan popis onečišćenja zraka i njihove oznake. Takav izvor podataka, potreban za izradu te druge tablice, dostupan je u EEA rječniku na [13], odakle se popis onečišćenja može jednostavno kopirati u excel datoteku i potom učitati u skladište podataka pomoću *Data Load* okoline. U ovom primjeru podaci se učitavaju u tablicu POLLUTANTS, kao što je prikazano na slici 4.32.



Slika 4.32 Uvoz oznaka onečišćenja zraka

Sada je moguće povezati mjerenja i vrste onečišćenja u jednom SQL upitu:

```
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;
```

Dalje se može istražiti kolike je godišnji prosjek NO_2 zagađenja za sve mjerne stanice.

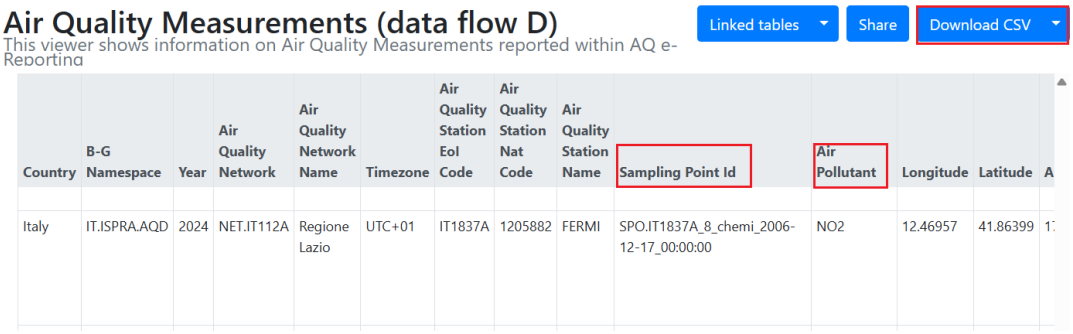
```
-------------------------------------------------------------------------------
-- Annual NO2 averages (pollutant = 8), excluding negative values.
-- Joined with POLLUTANTS for pollutant name, symbol and unit.
-------------------------------------------------------------------------------
SELECT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

Izvršavanjem prethodnog upita se dobiva da mjerna stanica s oznakom IT/SPO.IT1837A_8_chemi_2006-12-17_00:00:00 mjeri godišnji prosjek NO_2 od 47.074 ug.m-3 koji je opasan za zdravlje (gornja granica je 47.074 ug.m-3).

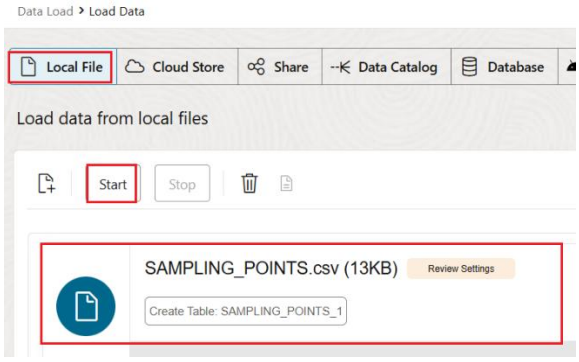
Povezivanje mjerenja s lokacijama mjernih stanica

Nakon što je utvrđeno da pojedine mjerne stanice bilježe prekoračenja dopuštenih razina NO_2, potrebno je saznati više o njihovim lokacijama. U tablici ROME_VERIFIED_2023 nalaze se samo oznake mjernih stanica, pa se iz nje ne može izravno zaključiti gdje je npr. njezina točna lokacija. Ovaj se nedostatak opet rješava uvođenjem dodatne tablice u skladište podataka koja sadrži opisne podatke o mjernim postajama, uključujući njihove nazive, geografske koordinate i tip postaje.

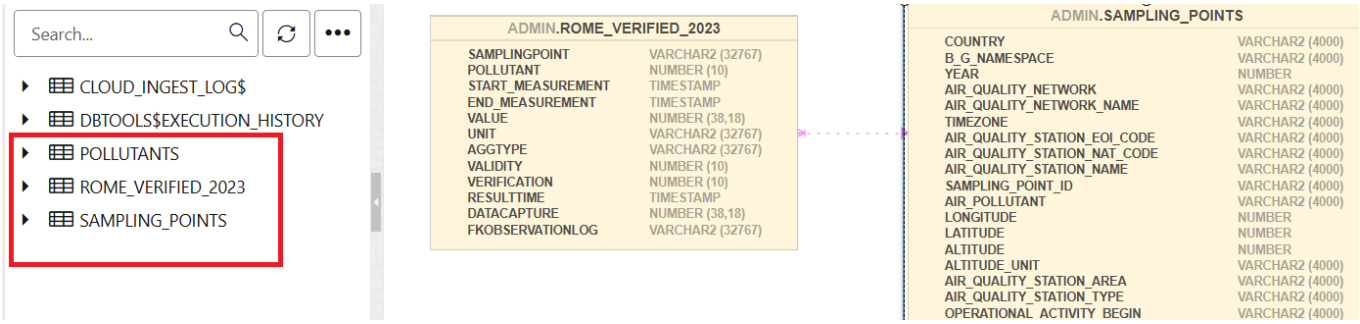
Izvor takvih podataka dostupan je na prethodno spomenutom EEA portalu [10], gdje se nalazi cjelovit popis europskih mjernih stanica. Popis se preuzima u CSV formatu odabirom poveznice *Metadata*, a zatim *Download CSV* (slika 4.29). Preuzeta CSV datoteka se opet učitava u skladište podataka pomoću *Data Load* radne okoline (slika 4.30). Time nastaje nova tablica, u ovom primjeru nazvana SAMPLING_POINTS, koja sadrži sve relevantne informacije o mjernim stanicama (slika 4.31).



Slika 4.29 Preuzimanje podataka o mjernim stanicama



Slika 4.30 Učitavanje podataka o mjernim stanicama



Slika 4.31 Pregled tablica u skladištu podataka

Nakon analize tablica koje su nastale uvozom podataka, utvrđuje se da se tablica ROME_VERIFIED_2023 može povezati s tablicom POLLUTANTS preko stupca POLLUTANT (u ROME_VERIFIED_2023) i stupca POLLUTANT_ID (u POLLUTANTS).

Također, tablica ROME_VERIFIED_2023 može se povezati i s tablicom SAMPLING_POINTS preko stupca SAMPLINGPOINT (u ROME_VERIFIED_2023) i stupca SAMPLING_POINT_ID (u SAMPLING_POINTS).

Međutim, pri pokušaju povezivanja javlja se problem jer se vrijednosti u stupcu SAMPLINGPOINT u tablici ROME_VERIFIED_2023 ne podudaraju izravno s vrijednostima u tablici SAMPLING_POINTS. Naime, u tablici ROME_VERIFIED_2023 vrijednost može biti IT/SPO.IT1837A_8_chemi_2006-12-17_00:00:00 dok istovremeno u tablici SAMPLING_POINTS ista lokacija postoji samo u obliku SPO.IT1837A_8_chemi_2006-12-17_00:00:00

Dakle, zapis u ROME_VERIFIED_2023 sadrži prefiks države (IT/), koji nije prisutan u tablici SAMPLING_POINTS. Budući da SQL JOIN zahtijeva potpuno podudaranje vrijednosti, ove se tablice ne mogu povezati izravno.

Problem se rješava tako da se u JOIN uvjetu primjenjuje funkcija SUBSTR(SAMPLINGPOINT, 4), kojom se uklanjaju prva tri znaka (IT/) iz vrijednosti u tablici ROME_VERIFIED_2023. Nakon toga, vrijednosti se podudaraju i moguće je ispravno povezati stupce SAMPLINGPOINT i SAMPLING_POINT_ID, čime se omogućuje spajanje sve tri tablice u jednom upitu.

```
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;
```

Završno se može postaviti i agregacijski upit koji povezuje sve tri tablice te izračunava godišnje prosjeke zagađenja za NO_2 (oznaka 8). Takav upit istovremeno prikazuje podatke o zagađenju i lokacijske atribute mjernih stanica, što omogućuje kvalitetniju interpretaciju rezultata mjerenja.

```
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 prethodnog upita su prikazani u Tablici 1 iz koje vidljivo prekoračenje NO_2 granične godišnje vrijednosti od 40 µg/m³ (mikrograma po kubnom metru) na mjernoj stanici Fermi koja se nalazi u RImu na Piazza Enrico Fermi. Dobivene koordinate za ovu mjernu stanicu (41.86399° N i 12.46957° E) moguće je pregledati na Google Mapsu te detaljnije prikazati putem Street View opcije (slika 4.32).

Tablica 1 Pregled godišnjih prosječnih koncentracija NO_2 po mjernim stanicama u Rimu (2023.)

POLLUTANT	COUNTRY	MUNICIPALITY	AIR QUALITY STATION NAME	AREA	AIR QUALITY STATION TYPE	LATITUDE	LONGITUDE	YEAR	VALUE
Nitrogen dioxide (air)	Italy	CORSO FRANCIA - Roma (RM)	C.SO FRANCIA	urban	traffic	41,94745	12,46956	2023	35,61
Nitrogen dioxide (air)	Italy	LARGO MAGNA GRECIA - Roma (RM)	L.GO MAGNA GRECIA	urban	traffic	41,88307	12,50895	2023	33,96
Nitrogen dioxide (air)	Italy	LARGO PERESTRELLO - Roma (RM)	L.GO PERESTRELLO	urban	background	41,88602	12,54162	2023	26,22
Nitrogen dioxide (air)	Italy	LGO. BELOLLI - Roma (RM)	CINECITTA	urban	background	41,85772	12,56865	2023	25,77
Nitrogen dioxide (air)	**Italy**	**P.ZZA E. FERMI, FRONTE CIVICO16 - Roma (RM)**	**FERMI**	**urban**	**traffic**	**41,86399**	**12,46957**	**2023**	**47,07**
Nitrogen dioxide (air)	Italy	Piazza Cairoli adiacente via dei Giubbonari - Roma (RM)	ARENULA	urban	background	41,89397	12,47539	2023	33,00
Nitrogen dioxide (air)	Italy	VIA AURELIA KM 20 CASTEL DI GUIDO - Roma (RM)	CASTEL DI GUIDO	rural	background	41,88945	12,26633	2023	8,17
Nitrogen dioxide (air)	Italy	VIA DELLA MELORIA, FRONTE CIVICO 27 - Roma (RM)	CIPRO	urban	background	41,90636	12,44759	2023	30,04
Nitrogen dioxide (air)	Italy	VIA MONTE CANDA, ADIACENTE CIVICO 12 - Roma (RM)	BUFALOTTA	urban	background	41,94765	12,53367	2023	29,94
Nitrogen dioxide (air)	Italy	VIA PAOLO EMILIO PAVOLINI 69/71 - Roma (RM)	MALAGROTTA	suburban	background	41,87478	12,34559	2023	16,90
Nitrogen dioxide (air)	Italy	VIA SALARIA - Roma (RM)	VILLA ADA	urban	background	41,93289	12,50695	2023	19,76
Nitrogen dioxide (air)	Italy	VIA TENUTA DEL CAVALIERE 102 - Roma (RM)	TENUTA DEL CAVALIERE	suburban	background	41,92947	12,65848	2023	23,89
Nitrogen dioxide (air)	Italy	VIA TIBURTINA, ADIACENTE CIVICO 654 - Roma (RM)	TIBURTINA	urban	traffic	41,91027	12,54885	2023	35,50



Slika 4.32 Google Street View na lokaciji mjerne stanice Fermi

Zadaci za vježbu

- Povežite se na Oracle Cloud konzolu.
- Otvorite novu instancu autonomne baze podataka u kojoj ćete realizirati skladište podataka.
- Stvorite spremnik u Oracle Object Storageu.
- Povežite se na EEA sjedište i pregledajte dokumentaciju i opise podatkovnih modela podataka (poveznice Medatdata, Vocabulary).
- Preuzmite parquet datoteku s aktualnim mjerenja o zagađenju zraka za svoju državu i postavite je u vaš spremnik podataka.
- Učitajte podatke u skladište podataka (mjerenja iz spremnika, prethodno pripremljene datoteke s popisom zagađenja (excel dateka POLLUTANTS, csv datoteka SAMPLING_POINTS)

- Analizirajte mjerenja za PM2.5 čestice. Koji su izvori ovog zagađenja i kako utječe na zdravlje ljudi?
- Identificirajte lokacije na kojima su izmjerene najviše razine PM2 5 zagađenja.
- Usporedite izmjerene vrijednosti s graničnim vrijednostima.

Pitanja za provjeru znanja

Koja su ključna svojstva i koristi skladišta podataka?

A. Omogućuje analizu velikih povijesnih skupova podataka
B. Optimizirano je za brze transakcije u stvarnom vremenu
C. Ujednačava i povezuje podatke iz više izvora
D. Služi kao zamjena za poslove izvještavanja na transakcijskim bazama podataka

Ako želite u skladištu podataka pripremiti meteorološke podatke kako biste postavili jednostavnu zvjezdastu shemu, koji bi minimalni skup tablica bio potreban (jedan točan odgovor)?

A. Jedna tablica mjerenja i jedna tablica lokacija mjerenja
B. Jedna tablica mjerenja, tablica mjernih senzora i tablica operatera sustava
C. Samo jedna tablica mjerenja jer se sve informacije već nalaze u njoj
D. Sve od prethodno navedenog

Koje svojstvo najbolje opisuje tzv. „sirove“ (raw) tablice u modelu okolišnih podataka?

A. Predstavljaju izvorne, neobrađene podatke preuzete iz vanjskih sustava, bez dodatnih transformacija
B. Sadrže privremene podatke koji su preuzeti iz vanjskih sustava
C. Pohranjuju podatke iz vanjskih sustava koji su oblikovani za brzi pristup i analizu
D. Optimirane su izvršavanje agregacijskih upita i izračun statistika

Koji koraci najbolje opisuju postupak modeliranja skladišta podataka u kontekstu prikazane studije slučaja s okolišnim podacima?

A. Preuzimanje podataka iz vanjskih izvora
B. Pregled strukture i razumijevanje podataka (metadata, vocabulary), identifikacija ključnih atributa i čišćenje podataka

C. Povezivanje tablica
D. Sve navedeno

Koje su osnovne funkcije Data Load sučelja tijekom rada s Oracle Automous Database skladištem podataka?

A. Automatsko generiranje SQL upita za analizu podataka
B. Uvoz podataka iz Object Storage spremnika
C. Upravljanje korisničkim računima i sigurnosnim postavkama.
D. Optimizacija analitičkih SQL upita.
E. Uvoz lokalnih datoteka

Dodatni izvori

[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”, 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 (EIONET), “Air Quality Pollutant Vocabulary”, Available: https://dd.eionet.europa.eu/vocabulary/aq/pollutant (accessed Dec. 9, 2025).