Obrada i optimizacija velikih skupova podataka

Ishodi učenja

Nakon završetka ove nastavne cjeline studenti će moći:

- Objasniti pojam velikih skupova podataka (Big Data) i njihova obilježja (količinu, brzinu nastajanja i raznolikost).
- Prepoznati glavne izvore velikih skupova podataka o okolišu (satelite, senzore i otvorene repozitorije).
- Razumjeti glavne izazove pohrane, čišćenja i obrade velikih skupova podataka.
- Primijeniti odgovarajuće alate za obradu velikih skupova podataka (Google BigQuery, Spark, Hadoop i dr.).
- Optimizirati podatkovne upite i postupke obrade radi boljih performansi i nižih troškova.
- Vizualizirati velike skupove podataka na nadzornim pločama i tumačiti rezultate u kontekstu stanja okoliša.
- Osmisliti i provesti skalabilan postupak obrade, analize i vizualizacije podataka o okolišu.

Uvod

Znanosti o okolišu ušle su u razdoblje u kojem količina podataka, njihova složenost i brzina nastajanja znatno nadmašuju mogućnosti za koje su razvijene tradicionalne metode upravljanja podacima i njihove analize. Izraz veliki skupovi podataka (Big Data) odnosi se na skupove toliko opsežne, promjenjive ili raznolike da je za izdvajanje korisnih informacija potrebno koristiti napredne računalne alate, distribuirane sustave i nove analitičke pristupe. U znanostima o okolišu nije riječ samo o veličini podataka, nego i o neprekidnim tokovima podataka, različitim formatima te izraženoj prostornoj i vremenskoj dimenziji koje treba zajednički analizirati.

Veliki skupovi podataka najčešće se opisuju s nekoliko ključnih obilježja, poznatih kao svojstva „V“:

- Količina (Volume) – goleme količine podataka, od terabajta do petabajta, koje svakodnevno nastaju satelitskim promatranjima, u senzorskim mrežama i simulacijama.
- Brzina (Velocity) – velika brzina nastajanja podataka, primjerice na meteorološkim postajama, mjernim postajama kakvoće zraka i oceanografskim plutačama koje mjere u stvarnom vremenu.

- Raznolikost (Variety) – različite vrste podataka, uključujući numerička mjerenja, slike, videozapise, vremenske nizove, tekstualna izvješća i slojeve geoprostornih podataka.
- Vjerodostojnost (Veracity) – različita kvaliteta, nesigurnost i pouzdanost podataka, što je osobito važno pri mjerenjima u okolišu.
- Vrijednost (Value) – mogućnost izdvajanja primjenjivih spoznaja za klimatsko modeliranje, zaštitu ekosustava, sprječavanje katastrofa i oblikovanje javnih politika.

U znanostima o okolišu veliki skupovi podataka imaju ključnu ulogu u razumijevanju složenih procesa na Zemlji, otkrivanju dugoročnih trendova te donošenju odluka o održivosti i zaštiti okoliša utemeljenih na dokazima.

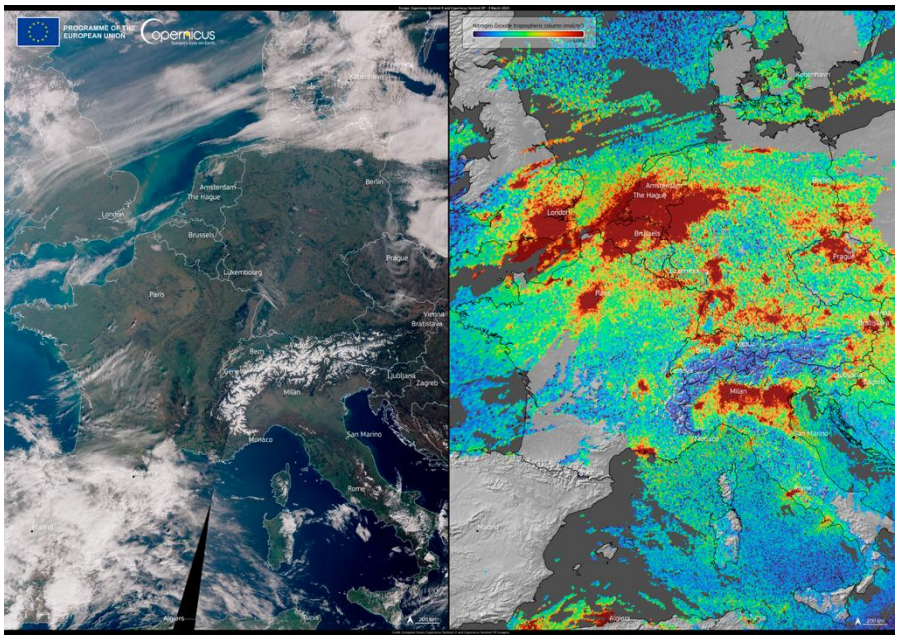
Izvori podataka o okolišu

Veliki skupovi podataka o okolišu potječu iz brojnih izvora. Svaki izvor daje poseban uvid u prirodne procese i procese na koje utječe čovjek.

Satelitski podaci jedan su od najvećih izvora. Sateliti za promatranje Zemlje neprekidno prikupljaju informacije o uporabi zemljišta, temperaturi površine mora, sastavu atmosfere, stanju vegetacije, ledenom pokrovu i dinamici oblaka. Ti podaci u pravilu imaju visoku prostornu razlučivost, višespektralni su i obuhvaćaju cijeli svijet, zbog čega su iznimno vrijedni, ali i računalno zahtjevni.

Senzori in situ (mjerenja in situ) provode se instrumentima postavljenima neposredno na mjernom mjestu, u dodiru s pojavom koja se proučava) osiguravaju neposredna mjerenja i podatke visoke vremenske učestalosti. Među njima su meteorološke postaje koje mjere temperaturu, vlagu, tlak i oborine; oceanografske plutače koje prate salinitet, struje i koncentraciju klorofila; senzori kakvoće zraka koji mjere onečišćujuće tvari poput PM2,5, NO_2 i ozona; te riječni i zemljišni senzori koji bilježe vodostaj, vlažnost i sadržaj hranjivih tvari. Takvi senzori često rade neprekidno i stvaraju tokove podataka koje treba obrađivati gotovo u stvarnom vremenu.

Globalne otvorene baze podataka dodatno obogaćuju istraživanja okoliša. Meteorološke i svemirske agencije te međunarodni istraživački konzorciji objavljuju otvoreno dostupne skupove podataka o klimi, bioraznolikosti, oceanografiji i atmosferi. Istraživači tako mogu povezati lokalna mjerenja s opažanjima globalnih razmjera te provoditi usporedne i dugoročne studije. Svi ti izvori zajedno čine složen i međusobno povezan podatkovni ekosustav velikih skupova podataka u znanostima o okolišu.



Slika 6.1. Satelitski prikaz programa Copernicus Sentinel-5P od 3. ožujka 2025. s istaknutim koncentracijama dušikova dioksida (NO_2).

Životni ciklus velikih skupova podataka

Upravljanje velikim skupovima podataka o okolišu obuhvaća strukturiran proces, poznat kao životni ciklus velikih skupova podataka (slika 6.2), koji se sastoji od nekoliko povezanih faza.

Prikupljanje podataka: podaci se preuzimaju sa satelita i senzora, terenskim mjerenjima, iz simulacija i baza podataka. Pritom treba voditi računa o učestalosti uzorkovanja, prostornoj razlučivosti i formatima kako bi podaci bili međusobno usklađeni i znanstveno relevantni.

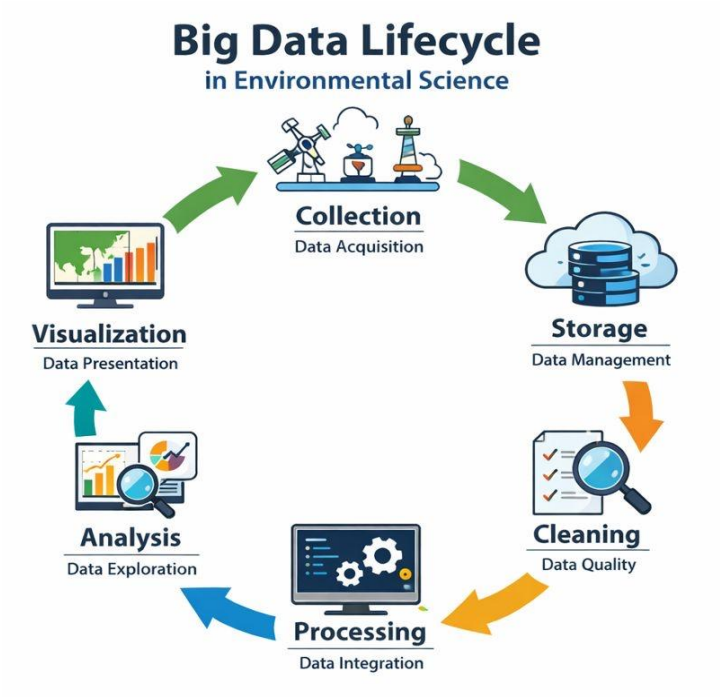
Pohrana podataka: zbog veličine i složenosti skupova podataka o okolišu često se koriste distribuirani sustavi, platforme u oblaku i specijalizirane baze za vremenske nizove i geoprostorne podatke. Učinkovita pohrana nužna je za skalabilnost i dugoročnu dostupnost.

Čišćenje i pretprocesiranje podataka: podaci o okolišu često sadržavaju nedostajuće vrijednosti, šum, odstupajuće vrijednosti ili neusklađena mjerenja. Čišćenje obuhvaća otkrivanje pogrešaka, popunjavanje praznina, ujednačavanje mjernih jedinica te vremensko i prostorno usklađivanje. Ova je faza ključna jer loša kvaliteta podataka može dovesti do pogrešnih zaključaka.

Obrada podataka: sirovi podaci pretvaraju se u strukturirani oblik prikladan za analizu. Postupci mogu uključivati agregiranje, filtriranje, ponovno uzorkovanje i integraciju više skupova podataka. Kod velikih skupova obrada se često paralelizira kako bi se velike količine podataka učinkovito obradile.

Analiza podataka: analitičke metode obuhvaćaju statističku analizu, otkrivanje trendova, strojno učenje i prediktivno modeliranje. U znanostima o okolišu služe za prepoznavanje obrazaca, procjenu rizika i simulaciju budućih scenarija.

Vizualizacija podataka: složeni skupovi podataka prikazuju se preglednim kartama, grafikonima, nadzornim pločama i animacijama. Učinkovita vizualizacija nužna je za prenošenje rezultata znanstvenicima, donositeljima odluka i javnosti, osobito kada je riječ o prostornim i vremenskim pojavama.



Slika 6.2. Životni ciklus velikih skupova podataka

Glavni izazovi pri radu s velikim skupovima podataka o okolišu

Unatoč znatnom potencijalu, rad s velikim skupovima podataka o okolišu donosi brojne izazove. Posebno je važna heterogenost: skupovi podataka razlikuju se po formatu, razlučivosti, mjerilu i nesigurnosti. Povezivanje satelitskih snimaka sa senzorskim mjerenjima i povijesnim zapisima zahtijeva pažljivo usklađivanje.

Brzina nastajanja podataka izazov je kada podaci neprekidno pristižu i moraju se obrađivati u stvarnom vremenu, primjerice tijekom ekstremnih vremenskih pojava ili pri upozorenjima o onečišćenju. I sama količina podataka može nadmašiti mogućnosti tradicionalnih računalnih sustava. Dugoročni klimatski zapisi i satelitske snimke visoke razlučivosti zahtijevaju skalabilnu infrastrukturu i učinkovite algoritme.

Kvaliteta i nesigurnost podataka osobito su važne u primjenama povezanima s okolišem. Kvarovi senzora, pogreške umjeravanja i nedostajuća mjerenja česte su pojave, pa nesigurnost treba izričito uzeti u obzir kako bi znanstveni zaključci bili pouzdani.

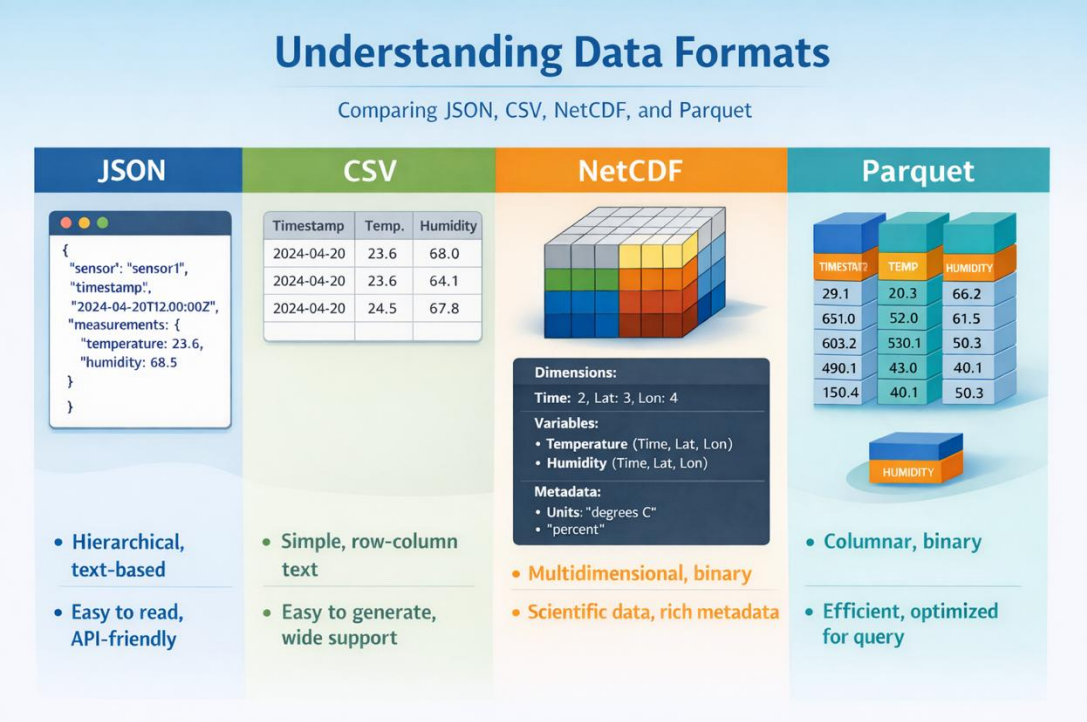
Formati i načini pohrane velikih skupova podataka

Jedan od najjednostavnijih i najraširenijih formata je CSV (Comma-Separated Values). CSV datoteke pohranjuju podatke kao običan tekst: svaki redak predstavlja jedan zapis, a stupci su odvojeni zarezima. Zbog jednostavnosti lako ih je izraditi, pregledavati i razmjenjivati među platformama. Često se koriste za tablične podatke, poput meteoroloških opažanja, mjerenja onečišćenja i statističkih sažetaka. Međutim, CSV nema ugrađenu podršku za metapodatke, nije prikladan za hijerarhijske podatke i postaje neučinkovit pri vrlo velikim skupovima.

JSON (JavaScript Object Notation) polustrukturirani je format za prikaz hijerarhijskih i ugniježđenih podataka. Posebno je prikladan za razmjenu podataka putem web API-ja, IoT uređaja i usluga u oblaku. U sustavima praćenja okoliša senzorske mreže često šalju podatke u JSON formatu, pri čemu se mjerenja, vremenske oznake i identifikatori uređaja povezuju u strukturirane objekte. JSON je čitljiv i fleksibilan, ali je zbog opširnog tekstualnog zapisa manje učinkovit od binarnih formata pri obradi golemih skupova podataka.

Radi poboljšanja performansi i skalabilnosti suvremeni sustavi za velike skupove podataka često koriste stupčane binarne formate, poput Parqueta. Apache Parquet optimiziran je za analitičke zadatke i često se koristi u distribuiranim sustavima poput Apache Sparka te u podatkovnim skladištima u oblaku. Pohranom podataka po stupcima, umjesto po redcima, Parquet ubrzava izvođenje upita, poboljšava kompresiju i smanjuje troškove pohrane. Posebno je prikladan za velike povijesne skupove, primjerice klimatske zapise i dugoročna mjerenja okoliša.

U znanstvenim područjima, osobito klimatologiji i oceanografiji, široko se primjenjuje NetCDF (Network Common Data Form). Namijenjen je pohrani višedimenzionalnih podataka, primjerice temperature, saliniteta ili atmosferskog tlaka kroz prostor i vrijeme. Za razliku od općih formata, NetCDF sadržava bogate metapodatke o varijablama, mjernim jedinicama i koordinatnim sustavima, zbog čega je vrlo prikladan za ponovljiva znanstvena istraživanja.



Slika 6.3. Usporedba formata podataka JSON, CSV, NetCDF i Parquet

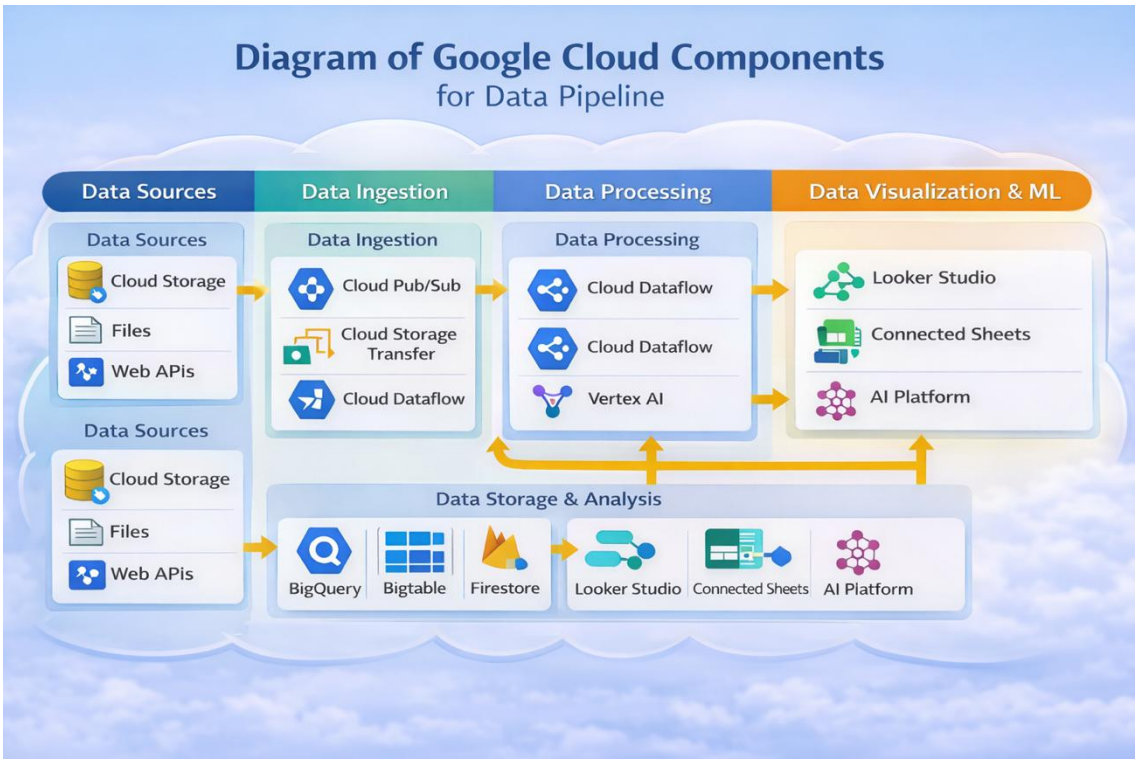
Osim o formatima (slika 6.3), sustavi za velike skupove podataka ovise i o načinima pohrane. Tradicionalna lokalna pohrana, primjerice datoteke na osobnim računalima ili laboratorijskim poslužiteljima, jednostavna je za upravljanje, ali ima ograničenu skalabilnost i pouzdanost. S rastom količine i složenosti podataka više nije dostatna za suradničku analizu velikih skupova.

Obrada velikih skupova podataka zahtijeva usklađene programske ekosustave koji upravljaju pohranom, računanjem i raspodjelom resursa na više računala. Jedan od najranijih i najutjecajnijih sustava bio je Apache Hadoop, koji je uveo distribuiranu pohranu (HDFS) zajedno s distribuiranim računanjem prema programskom modelu MapReduce. Hadoop je omogućio obradu skupova podataka koji su prije bili preveliki za jedno računalo.

Iako je Hadoop postavio temelje, noviji programski okviri poput Apache Sparka postali su široko prihvaćeni zahvaljujući boljim performansama i fleksibilnosti. Spark podržava računanje u radnoj memoriji, što znatno ubrzava obradu. Omogućuje i programska sučelja visoke razine za analizu podataka, strojno učenje i obradu grafova. U analitici okoliša često se koristi za obradu velikih senzorskih skupova ili arhiva satelitskih snimaka.

Posljednjih godina platforme u oblaku preobrazile su područje velikih skupova podataka nudeći upravljane usluge koje uklanjaju potrebu za ručnim održavanjem infrastrukture. Google Cloud Platform (GCP) – slika 6.4, Amazon Web Services (AWS) i Microsoft Azure nude integrirane ekosustave pohrane, obrade, analitike i vizualizacije. Tako se istraživači mogu usredotočiti na analizu, a ne na administriranje sustava.

Sustavi u oblaku posebno su privlačni interdisciplinarnim područjima poput znanosti o okolišu, u kojima je količina podataka velika, a računalni se zahtjevi mijenjaju tijekom vremena. Omogućuju elastično skaliranje: resursi se dinamički povećavaju ili smanjuju u skladu s opterećenjem, čime se optimiziraju performanse i troškovi.



Slika 6.4. Komponente platforme Google Cloud

Pretprocesiranje i čišćenje podataka velikog opsega

Sirovi podaci rijetko su odmah prikladni za analizu. Veliki skupovi podataka često sadržavaju nedostajuće vrijednosti, nedosljednosti, duplikate i pogreške mjerenja. Zato su pretprocesiranje i čišćenje ključne faze, posebno pri radu s podacima o okolišu ili tokovima podataka IoT uređaja.

OpenRefine široko je korišten alat otvorenog koda za interaktivno čišćenje i transformaciju podataka. Omogućuje pregled skupova podataka, otkrivanje anomalija, ujednačavanje

vrijednosti i ispravljanje pogrešaka bez poznavanja programiranja. Posebno je koristan pri pripremi podataka preuzetih iz javnih repozitorija ili terenskih istraživanja.

Trifacta Wrangler, u izvorniku naveden i kao Google Cloud Dataprep, proširuje takve mogućnosti na skupove podataka obrađene u oblaku. Pruža vizualno sučelje temeljeno na pravilima, u kojem se definiraju koraci transformacije koje je moguće izvršavati u velikom opsegu. To je korisno pri pripremi velikih skupova pohranjenih na uslugama poput Google Cloud Storagea ili BigQueryja.

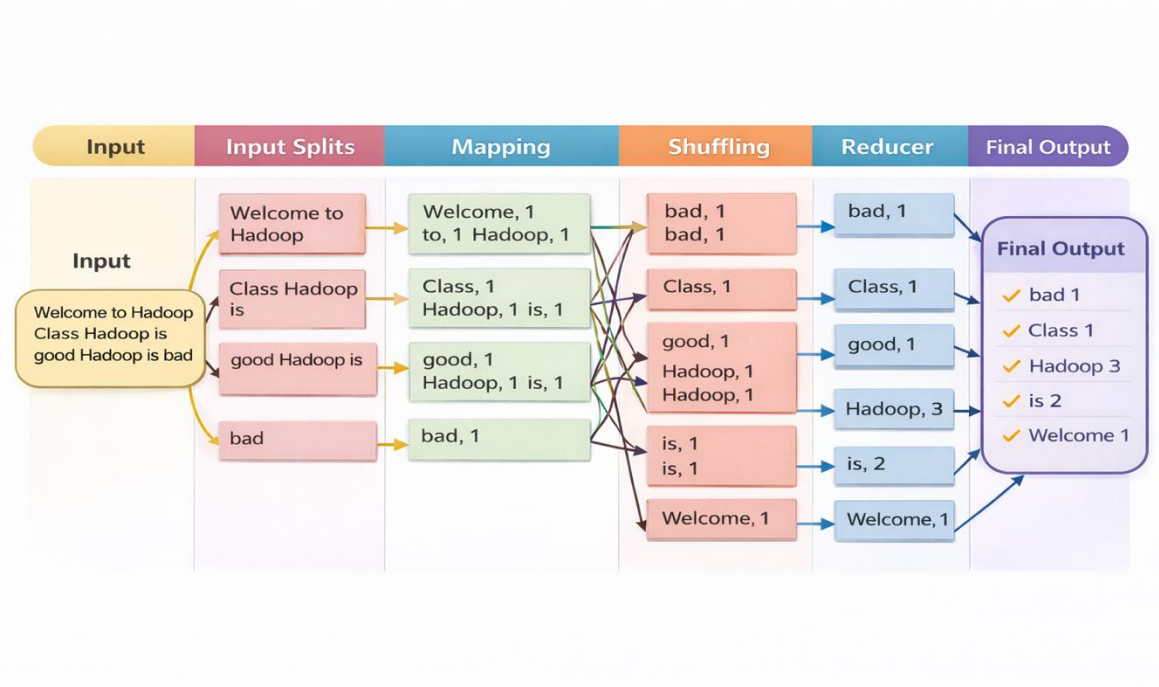
Korisnicima s programerskim iskustvom biblioteka Pandas za Python pruža snažne mogućnosti rada s podacima. Često se koristi u bilježnicama Jupyter i okruženju Google Colab za čišćenje, filtriranje i transformaciju podataka prije obrade većeg opsega. Iako je Pandas najprikladniji za srednje velike skupove, ima važnu ulogu u istraživačkoj analizi podataka i izradi prototipova.

Načela distribuirane i paralelne obrade

Temelj analitike velikih skupova podataka čine distribuirana i paralelna obrada. Umjesto uzastopne obrade na jednom računalu, sustav dijeli podatke u manje cjeline koje se istodobno obrađuju na više čvorova.

MapReduce (slika 6.5) model je obrade vrlo velikih skupova podataka podjelom posla na mnogo manjih zadataka koji se mogu izvoditi paralelno. Postao je popularan jer omogućuje obradu podataka koji ne stanu na jedno računalo, bez potrebe za jednim iznimno snažnim računalom. Umjesto toga, koristi se više uobičajenih računala ili instanci u oblaku, među kojima se podaci raspodjeljuju. To je osobito korisno kada su podaci raspoređeni u više datoteka ili poslužitelja, primjerice zapisi događaja, senzorska mjerenja, satelitski segmenti ili dugoročni arhivi okoliša.

MapReduce najlakše je razumjeti prema ulogama dviju funkcija. Funkcija Map pretvara sirove ulazne podatke u međurezultate koje je poslije lako objediniti. Ona ne završava cijeli zadatak, nego priprema strukturirane gradivne elemente, odnosno parove ključ–vrijednost. Funkcija Reduce zatim spaja međurezultate s istim ključem u konačan rezultat. Na ovaj se način mogu izraziti brojni važni zadaci: prebrojavanje, zbrajanje, grupiranje, računanje srednjih vrijednosti, pronalaženje minimuma i maksimuma te izrada statistika po lokaciji.



Slika 6.5. Postupak MapReduce

Faza Shuffle, odnosno preraspodjela, ključna je jer osigurava da se svi zapisi s istim ključem nađu zajedno. Često je i jedna od najskupljih faza jer zahtijeva prijenos podataka mrežom između računala. Ako se velik skup preslika na ključeve poput „station_id“ ili „city“, sustav mora sve zapise pojedine postaje usmjeriti istom izvršitelju Reduce. Budući da je mrežni prijenos sporiji od pristupa radnoj memoriji, dobar dizajn sustava MapReduce nastoji smanjiti međupodatke i ravnomjerno rasporediti ključeve kako se pojedini izvršitelj ne bi preopteretio.

Važna prednost modela MapReduce otpornost je na kvarove. U velikim distribuiranim sustavima povremeni kvar računala uobičajen je. Programska okruženja MapReduce, poput Hadoopa, omogućuju da se zadatak neuspjelog radnog čvora automatski dodijeli drugome i obrada nastavi. To je važno u stvarnim primjenama u okolišu, gdje obrada golemih količina podataka može trajati satima, a pouzdanost postaje jednako važna kao brzina.

U analitici okoliša i kakvoće zraka načelo MapReduce primjenjuje se u mnogim zadacima, čak i kada korisnik izravno ne piše odgovarajući kod. Zamislimo višegodišnja mjerenja PM2,5 s brojnih mjernih postaja. Faza Map za svako mjerenje može proizvesti par (station_id, pm25_value), a faza Reduce izračunati prosjek po postaji, dnevni maksimum ili broj dana u kojima je prekoračen zadani prag.

Mnogi suvremeni sustavi, poput Sparka, BigQueryja i postupaka u Dataflowu, više ne zahtijevaju ručnu implementaciju modela MapReduce, ali je taj model i dalje koristan za

razumijevanje obrade. Objašnjava zašto alati za velike skupove podataka naglašavaju grupiranje, agregiranje i distribuirano spajanje: podaci se najprije razvrstavaju u smislene skupine, zatim se povezani zapisi preraspodjeljuju i na kraju svode na sažetke koje ljudi mogu tumačiti.

Indeksiranje, optimizacija upita i upravljanje potrošnjom energije

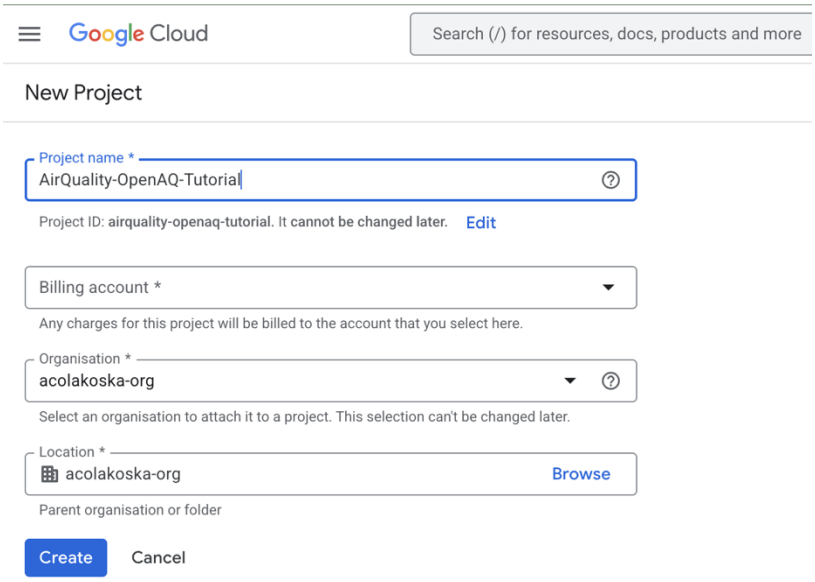
Kako skupovi podataka rastu, učinkovit pristup postaje sve važniji. Indeksiranje ubrzava izvođenje upita jer sustavu omogućuje brzo pronalaženje relevantnih zapisa bez pregledavanja cijelog skupa. U podatkovnim skladištima u oblaku, poput Google BigQueryja, strategije indeksiranja i particioniranja mogu znatno smanjiti trajanje i cijenu izvođenja upita.

Optimizacija upita podrazumijeva odabir učinkovitih planova izvršavanja analitičkih upita. Suvremeni sustavi automatski smanjuju prijenos podataka, iskorištavaju paralelizam i koriste predmemoriranje. Takve su optimizacije posebno važne u oblaku, gdje se računalni resursi naplaćuju prema korištenju.

Uz performanse, sve je važnije upravljanje električnom energijom i energetskom učinkovitošću. Veliki podatkovni centri troše mnogo energije, što potiče razvoj energetski učinkovite opreme i algoritama raspoređivanja zadataka. Pružatelji usluga u oblaku sve više ulažu u obnovljive izvore energije i raspodjelu resursa prilagođenu potrošnji, nastojeći takva rješenja učiniti održivijima od tradicionalne lokalne infrastrukture.

Primjer primjene: obrada i analiza podataka o kakvoći zraka u oblaku uz Googleove alate

U ovom primjeru kao glavni izvor koristi se skup podataka OpenAQ dostupan putem Google Cloud Marketplacea i BigQueryja. Sadržava globalna opažanja kakvoće zraka (primjerice PM2,5, PM10, NO_2 i O_3), prikupljena od različitih pružatelja podataka i s tisuća mjernih lokacija. Pojedini zapisi obično sadržavaju prostorno-vremenske metapodatke (državu, grad, mjernu postaju i često geografske koordinate), mjerenu onečišćujuću tvar ili parametar, vrijednost s mjernom jedinicom i vremensku oznaku. Zbog toga je skup prikladan za analitiku okoliša velikog opsega, uključujući analizu trendova vremenskih nizova, otkrivanje područja povišenog onečišćenja i izvještavanje na nadzornim pločama u alatima poput BigQueryja i Looker Studija. Šire gledano, OpenAQ prikuplja, objedinjuje, usklađuje i objavljuje interoperabilne povijesne podatke i podatke o kakvoći zraka u stvarnom vremenu, čime olakšava njihovu uporabu i usporedbu.



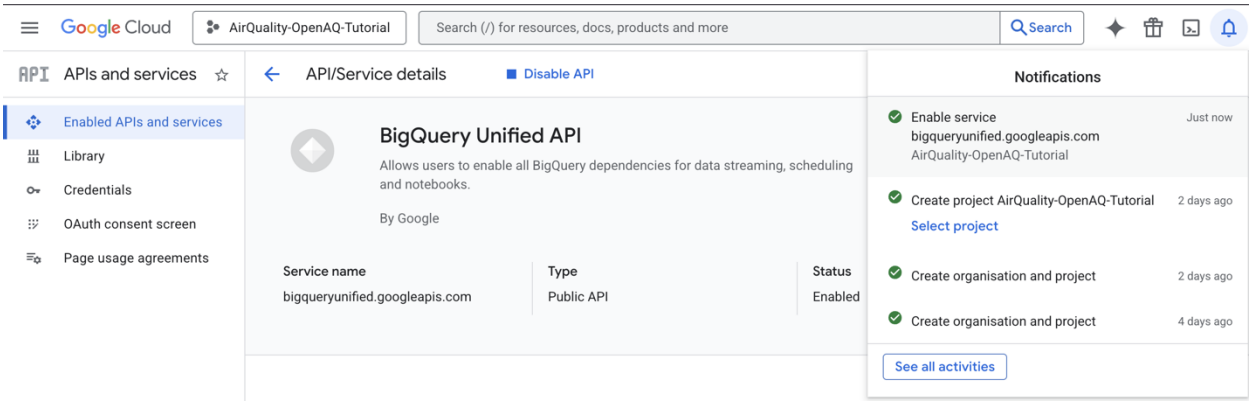
Slika 6.6. Izrada projekta u Google Cloudu

Najprije treba izraditi poseban projekt u Google Cloudu (slika 6.6) kako bi svi resursi za vježbu s podacima o kakvoći zraka bili organizirani i odvojeni od ostalih aktivnosti. U sučelju Google Cloud Console odaberite izbornik projekata na gornjoj navigacijskoj traci te mogućnost New Project. Upišite prepoznatljiv naziv, primjerice AirQuality-OpenAQ-Tutorial, radi lakšeg pronalaska pri promjeni projekta, upravljanju pravima pristupa ili pregledu troškova. Nakon

upisa naziva (i odabira organizacije ili lokacije, ako se to zatraži) kliknite Create. Time nastaje radni prostor u kojem ćete upravljati kasnijim BigQuery skupovima podataka, upitima i nadzornim pločama.

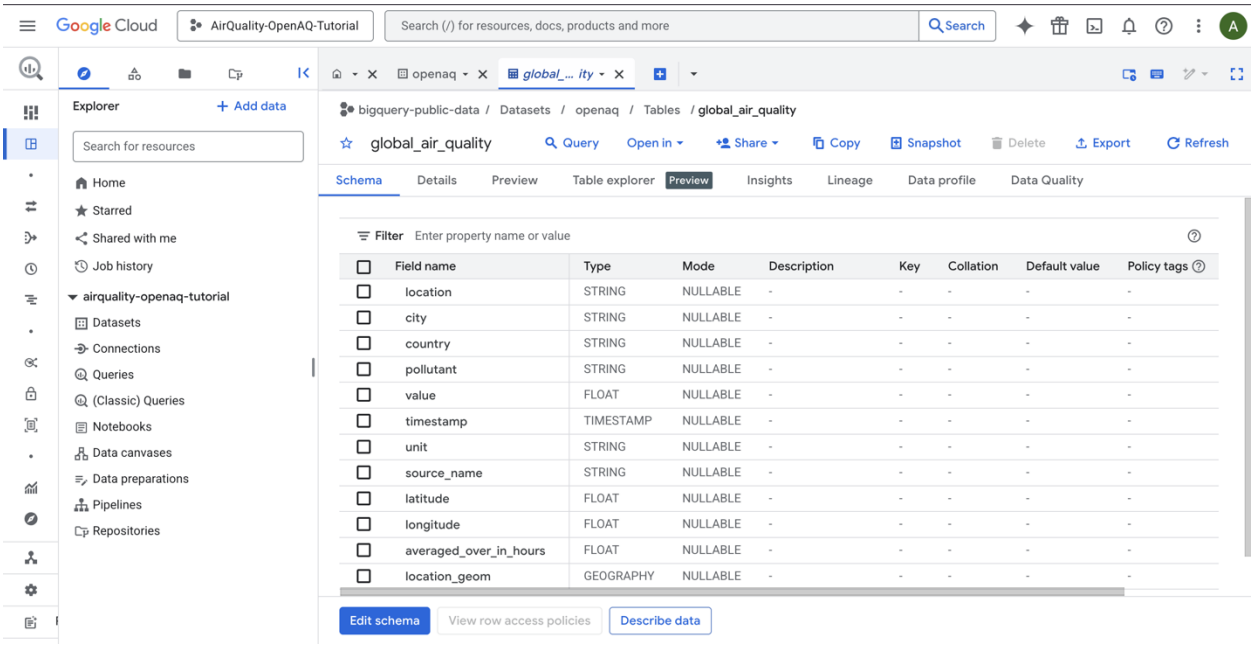
Zatim u Google Cloud Consoleu otvorite BigQuery. U traku za pretraživanje upišite „BigQuery“ i otvorite odgovarajuću uslugu. Provjerite naziv aktivnog projekta u gornjoj navigacijskoj traci kako biste bili sigurni da će se novi skupovi, upiti i spremljeni pogledi povezati s ispravnim projektom.

Provjerite je li omogućen BigQuery API (slika 6.7). Obično se aktivira automatski pri prvom otvaranju usluge, ali ako se pojavi pogreška pristupa ili API-ja, uključite ga ručno. Otvorite APIs & Services → Enabled APIs & services i provjerite nalazi li se na popisu BigQuery API. Ako nije aktivan, odaberite ga i kliknite Enable kako bi projekt mogao izvršavati upite i izrađivati poglede.



Slika 6.7. Omogućavanje API-ja BigQuery Unified

Nakon toga dodajte skup podataka OpenAQ. Najbrže je otvoriti Google Cloud Marketplace: kliknite Add Data, odaberite Public datasets, zatim pretražite „OpenAQ“, otvorite odgovarajući rezultat i odaberite View dataset. Skup će postati dostupan u trenutačnom projektu Google Cloud, bez potrebe za fizičkim prijenosom podataka ili dodatnim troškom njihove pohrane (slika 6.8).



Slika 6.8. Pregled skupa podataka u Google BigQueryju

U središnjem dijelu odabran je prikaz Schema, koji navodi polja (stupce) tablice global_air_quality. Prikaz sadrži zaglavlja Field name, Type i Mode te atribute kakvoće zraka, poput location, city, country, pollutant, value i timestamp. Tu su i kontekstualna polja unit i source_name te geografski podaci latitude, longitude i prostorno polje location_geom tipa GEOGRAPHY. To pokazuje da skup, uz uobičajene upite nad vremenskim nizovima, podržava kartografski prikaz i prostornu analizu.

Za ponovljivu analizu podataka OpenAQ i pripremu nadzornih ploča izradit ćemo nekoliko upita u BigQueryju te ih spremiti kao poglede (views). Prvi upit objedinjuje dnevna mjerenja PM2,5 u vremenski niz, pogodan za praćenje dugoročnih promjena i trendova. Drugi pronalazi dane s najvećim prosječnim koncentracijama PM2,5 i tako ističe epizode izraženog onečišćenja. Treći računa prosječne koncentracije po mjernim lokacijama i stvara usporedni poredak područja s trajno povišenim vrijednostima. Četvrti izdvaja najnovije mjerenje za svaku lokaciju s valjanim koordinatama, čime nastaje sloj za prikaz na karti. Budući da je dostupni skup u BigQueryju povijesna snimka čija je najnovija vremenska oznaka iz 2022., upiti koriste razdoblja određena u odnosu na posljednji zapis, a ne na današnji datum. Tako rezultati ostaju usporedivi i kada nema novih podataka. Ovi pogledi predstavljaju analitičku osnovu koju je moguće povezati s Looker Studiom radi interaktivnog prikaza i transparentnog izvještavanja (slika 6.9).

1) Dnevni trend PM2,5 tijekom posljednjih 90 dana u odnosu na max_ts

```
WITH t AS (
```

```
 SELECT MAX(timestamp) AS max_ts
 FROM `bigquery-public-data.openaq.global_air_quality`
)
SELECT
 DATE(g.timestamp) AS day,
 AVG(g.value) AS avg_pm25
FROM `bigquery-public-data.openaq.global_air_quality` g
CROSS JOIN t
WHERE g.pollutant = 'pm25'
 AND g.timestamp >= TIMESTAMP_SUB(t.max_ts, INTERVAL 90 DAY)
GROUP BY day
ORDER BY day;
```

Spremite kao pogled: Run → SAVE → Save view → skup podataka openaq_dashboard → naziv v_pm25_daily_trend

2) Deset dana s najvećim PM2,5 tijekom posljednje godine u odnosu na max_ts

```
WITH t AS (
 SELECT MAX(timestamp) AS max_ts
 FROM `bigquery-public-data.openaq.global_air_quality`
)
SELECT
 DATE(g.timestamp) AS day,
 AVG(g.value) AS avg_pm25,
 COUNT(*) AS measurements
FROM `bigquery-public-data.openaq.global_air_quality` g
CROSS JOIN t
WHERE g.pollutant = 'pm25'
 AND g.timestamp >= TIMESTAMP_SUB(t.max_ts, INTERVAL 365 DAY)
GROUP BY day
ORDER BY avg_pm25 DESC
LIMIT 10;
```

Spremite kao pogled: v_pm25_worst_days

3) Petnaest lokacija s najvišim prosječnim PM2,5 (posljednjih 90 dana u odnosu na max_ts)

```
WITH t AS (
 SELECT MAX(timestamp) AS max_ts
```

```
 FROM `bigquery-public-data.openaq.global_air_quality`
)
SELECT
 COALESCE(g.country, 'Unknown') AS country,
 COALESCE(g.city, 'Unknown') AS city,
 g.location,
 AVG(g.value) AS avg_pm25,
 COUNT(*) AS measurements
FROM `bigquery-public-data.openaq.global_air_quality` g
CROSS JOIN t
WHERE g.pollutant = 'pm25'
 AND g.timestamp >= TIMESTAMP_SUB(t.max_ts, INTERVAL 90 DAY)
GROUP BY country, city, location
HAVING measurements >= 50
ORDER BY avg_pm25 DESC
LIMIT 15;
```

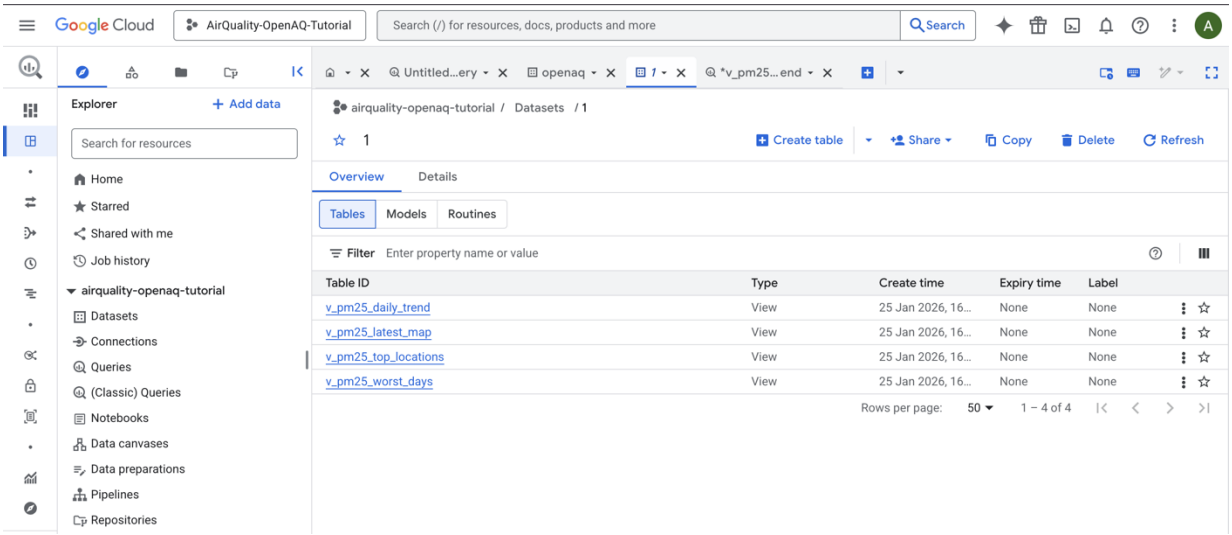
Spremite kao pogled: v_pm25_top_locations

4) Najnovije mjerenje PM2,5 po lokaciji (posljednjih 14 dana u odnosu na max_ts)

```
WITH t AS (
 SELECT MAX(timestamp) AS max_ts
 FROM `bigquery-public-data.openaq.global_air_quality`
),
latest AS (
  SELECT
    g.location,
  g.city,
  g.country,
  g.latitude,
  g.longitude,
  g.value,
  g.unit,
  g.timestamp,
  ROW_NUMBER() OVER (PARTITION BY g.location ORDER BY g.timestamp DESC) AS rn
 FROM `bigquery-public-data.openaq.global_air_quality` g
```

```
 CROSS JOIN t
  WHERE g.pollutant = 'pm25'
  AND g.latitude IS NOT NULL
  AND g.longitude IS NOT NULL
  AND g.timestamp >= TIMESTAMP_SUB(t.max_ts, INTERVAL 14 DAY)
)
SELECT
 location,
 city,
 country,
 latitude,
 longitude,
 value AS latest_pm25,
 unit,
 timestamp AS latest_time
FROM latest
WHERE rn = 1;
```

Rezultirajući pogledi prikazani su na slici 6.9.

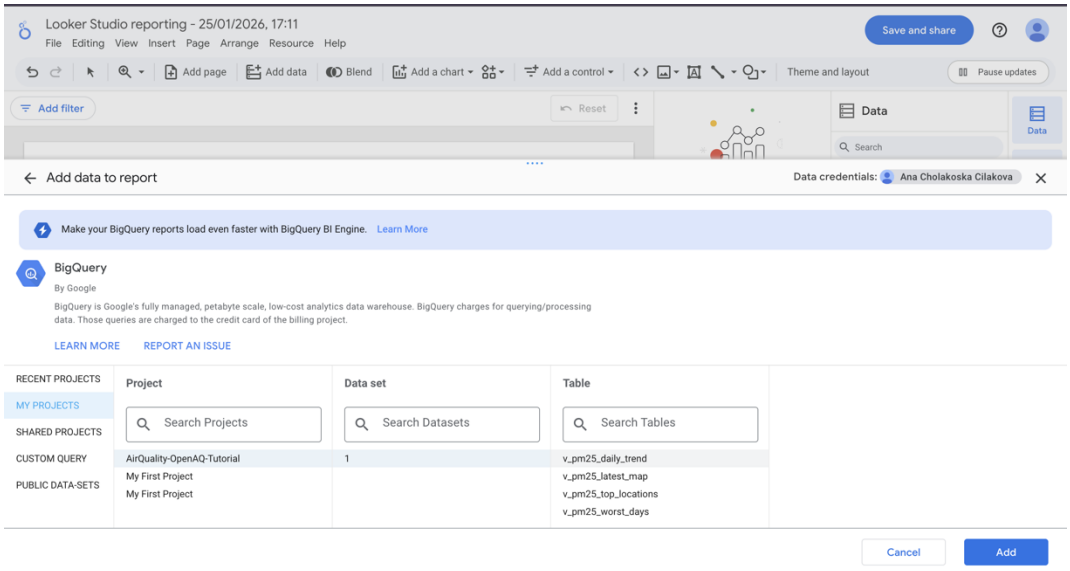


Slika 6.9. Četiri pogleda – rezultati SQL upita

U nastavku ćemo u Looker Studiju izraditi interaktivnu nadzornu ploču kakvoće zraka koja se izravno povezuje s četiri prethodno izrađena pogleda u BigQueryju, temeljena na podacima OpenAQ. Svaki je pogled zasebna analitička komponenta: dnevni trend PM2,5, ekstremni dani, usporedba lokacija i kartografski sloj najnovijih mjerenja. Odvajanjem pripreme podataka od

vizualizacije povećavaju se preglednost i ponovljivost postupka. Dobiva se cjelovita nadzorna ploča za istraživačku analizu grafikonima, tablicama i kartama, prikladna i za izvještavanje.

Otvorite Looker Studio i odaberite Create → Report. Po potrebi se prijavite Google računom i dodijelite prava za pristup resursima Google Clouda. Nakon otvaranja praznog izvješća kliknite Add data te odaberite BigQuery kao poveznik. Zatim izaberite odgovarajući Project (npr. AirQuality-OpenAQ-Tutorial), skup podataka openaq_dashboard i pogled v_pm25_daily_trend. Kliknite Add, a zatim po potrebi Add to report kako biste dodali prvi izvor podataka za nadzornu ploču (slika 6.10).



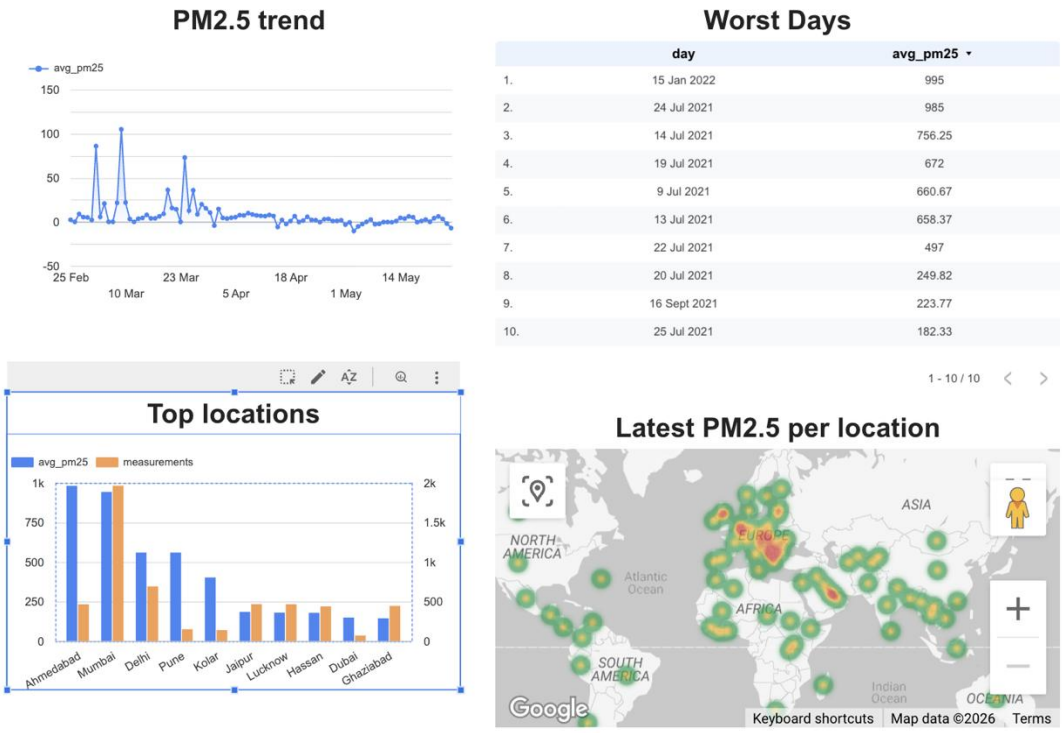
Slika 6.10. Dodavanje podataka u izvješće

Nakon dodavanja pogleda v_pm25_daily_trend izradite grafikon vremenskog niza (Time Series) za prikaz promjena onečišćenja tijekom vremena. Umetnite grafikon u izvješće te polje datuma day postavite kao glavnu dimenziju, a prosječni PM2,5 (avg_pm25) kao mjeru. Tako dobivate pregled dnevnih oscilacija i dugoročnijih kretanja koncentracija čestica, uz mogućnost prepoznavanja trajnih porasta, padova ili povremenih vršnih vrijednosti povezanih s meteorološkim uvjetima ili emisijama.

Pomoću pogleda v_pm25_top_locations umetnite stupčasti grafikon (Bar Chart) za usporedbu prosječnih koncentracija PM2,5 među mjernim lokacijama. Kao dimenziju odaberite location (ili city ako želite grupiranje po gradu), a kao mjeru avg_pm25. Poredajte stupce silazno kako bi lokacije s najvišim koncentracijama bile na vrhu. Polje measurements može pomoći pri procjeni pouzdanosti procjene jer prikazuje broj mjerenja.

Za prikaz najizraženijih epizoda onečišćenja dodajte tablicu (Table) temeljenu na pogledu v_pm25_worst_days. Postavite day kao dimenziju, a avg_pm25 i, gdje je dostupno, measurements kao mjere. Sortirajte prema avg_pm25 silazno kako bi najonečišćeniji dani bili prvi. Tablica omogućuje precizno izvještavanje i usporedbu vršnih vrijednosti te sažet prikaz razdoblja povećanog rizika za okoliš.

Za prostorno tumačenje izradite kartu Google Maps na temelju pogleda v_pm25_latest_map. Kao geografska polja postavite latitude i longitude, a latest_pm25 kao glavnu mjeru koja određuje intenzitet prikaza, primjerice veličinu oznake. Dopunite informacije pri prelasku pokazivačem atributima location, city i latest_time kako bi se pojedine mjerne postaje mogle pregledavati zajedno s pripadajućim vrijednostima. Svi izrađeni prikazi prikazani su na slici 6.11.



Slika 6.11. Dovršena nadzorna ploča

Po želji možete optimizirati četvrti upit kako biste smanjili računalno opterećenje i troškove, a zadržali isti analitički cilj: dohvatiti posljednje mjerenje PM2,5 za svaku postaju. Najprije se zajednički tablični izraz i naknadni filtar WHERE rn = 1 zamjenjuju klauzulom QUALIFY, koja omogućuje filtriranje rezultata prozorske funkcije unutar istog upita. Time se upit pojednostavnjuje i često ubrzava. Zatim se provjerom timestamp IS NOT NULL izostavljaju

nevaljani zapisi, a vremenski prozor od 14 dana određuje se prema najvećoj dostupnoj oznaci MAX(timestamp), što je važno za povijesne skupove koji se više ne ažuriraju. Odabirom samo polja potrebnih za kartu i izvještavanje smanjuje se količina obrađenih podataka i poboljšavaju performanse. Potpuni SQL upit prikazan je u nastavku:

```
SELECT
 location,
 city,
 country,
 latitude,
 longitude,
 value AS latest_pm25,
 unit,
 timestamp AS latest_time
FROM `bigquery-public-data.openaq.global_air_quality`
WHERE pollutant = 'pm25'
 AND timestamp IS NOT NULL
 AND latitude IS NOT NULL
 AND longitude IS NOT NULL
 -- Use dataset's latest timestamp window if your snapshot is old:
 AND timestamp >= TIMESTAMP_SUB(
  (SELECT MAX(timestamp) FROM `bigquery-public-data.openaq.global_air_quality` WHERE timestamp IS NOT NULL),
  INTERVAL 14 DAY
 )
QUALIFY ROW_NUMBER() OVER (PARTITION BY location ORDER BY timestamp DESC) = 1;
```

Praktična vježba

Mini-projekt: Odaberite skup podataka o okolišu (primjerice o kakvoći zraka, moru ili bioraznolikosti). Očistite i obradite podatke, izvedite upite i vizualizirajte rezultate koristeći alate Google Clouda. Postupak i rezultate predstavite kratkom prezentacijom u trajanju od 5 do 10 minuta.

Provjera znanja

1. Koja mogućnost točno opisuje tri obilježja velikih skupova podataka (3V)?

A) Volume = točnost podataka; Velocity = veličina podataka; Variety = brzina podataka

B) Volume = količina podataka; Velocity = brzina nastajanja/obrade; Variety = različite vrste i izvori podataka

C) Volume = broj korisnika; Velocity = brzina interneta; Variety = broj poslužitelja

D) Volume = trošak pohrane; Velocity = trošak upita; Variety = broj nadzornih ploča

2. Koje je uparivanje najtočnije?

A) CSV = nestrukturirani; JSON = strukturirani; slike = polustrukturirani podaci

B) CSV = strukturirani; JSON = polustrukturirani; slike/videozapisi = nestrukturirani podaci

C) CSV = polustrukturirani; JSON = nestrukturirani; slike = strukturirani podaci

D) CSV = nestrukturirani; JSON = nestrukturirani; slike = strukturirani podaci

3. Potrebna vam je brza analiza velikih dnevnika događaja (stupci poput timestamp, device_id, PM2.5 i temperature) pohranjenih u objektnoj pohrani u oblaku. Što je najbolji izbor?

A) Parquet/ORC (stupčani format)

B) Obični tekst (.txt)

C) PNG slike

D) XML s duboko ugniježđenim oznakama

4. Koji skup ispravno povezuje uslugu i njezinu namjenu?

A) BigQuery = objektna pohrana; Cloud Storage = SQL analitika; Dataflow = vizualizacija

B) Cloud Storage = objektna pohrana; BigQuery = analitičko podatkovno skladište bez upravljanja poslužiteljima; Dataflow = distribuirana obrada podataka (ETL)

C) BigQuery = red poruka; Pub/Sub = nadzorne ploče; Looker = unos podataka

D) Dataflow = pohrana datoteka; Cloud Storage = učenje ML modela; BigQuery = upravljanje identitetima

5. Koja radnja obično smanjuje trošak i trajanje upita nad vremenskim nizovima u podatkovnom skladištu u oblaku, primjerice BigQueryju?

A) Koristiti SELECT * kako nijedan stupac ne bi bio izostavljen

B) Particionirati prema datumu/vremenu i grupirati prema čestim ključevima filtriranja (npr. station_id/device_id)

C) Pohraniti sve u jedan veliki JSON stupac radi pojednostavnjenja sheme

D) Isključiti predmemoriranje i ponavljati upite kako bi se sustav „zagrijao“

Literatura i dodatni izvori

[1] BigQuery public datasets (overview + how to access). Dostupno na: https://docs.cloud.google.com/bigquery/public-data, Pristupljeno: 15. siječnja 2026.https://docs.cloud.google.com/bigquery/public-data

[2] BigQuery: Introduction to optimizing query performance. Dostupno na: https://docs.cloud.google.com/bigquery/docs/best-practices-performance-overview, Pristupljeno: 15. siječnja 2026.https://docs.cloud.google.com/bigquery/docs/best-practices-performance-overview

[3] BigQuery: Optimize query computation (best practices). Dostupno na: https://docs.cloud.google.com/bigquery/docs/best-practices-performance-compute, Pristupljeno: 15. siječnja 2026.https://docs.cloud.google.com/bigquery/docs/best-practices-performance-compute

[4] Looker Studio: Connect to Google BigQuery. Dostupno na: https://docs.cloud.google.com/looker/docs/studio/connect-to-google-bigquery, Pristupljeno: 15. siječnja 2026.https://docs.cloud.google.com/looker/docs/studio/connect-to-google-bigquery

[5] Looker Studio Documentation. Dostupno na: https://docs.cloud.google.com/looker/docs/studio, Pristupljeno: 15. siječnja 2026.https://docs.cloud.google.com/looker/docs/studio

[6] Looker Studio quick start guide . Dostupno na: https://docs.cloud.google.com/looker/docs/studio/quick-start-guide, Pristupljeno: 15. siječnja 2026.https://docs.cloud.google.com/looker/docs/studio/quick-start-guide

[7] Cloud Billing: Budgets and budget alerts. Dostupno na: https://docs.cloud.google.com/billing/docs/how-to/budgets, Pristupljeno: 15. siječnja 2026.https://docs.cloud.google.com/billing/docs/how-to/budgets

[8] Cloud Billing documentation hub. Dostupno na: https://docs.cloud.google.com/billing/docs, Pristupljeno: 15. siječnja 2026.https://docs.cloud.google.com/billing/docs

[9] BigQuery sharing (formerly Analytics Hub): Introduction. Dostupno na: https://docs.cloud.google.com/bigquery/docs/analytics-hub-introduction, Pristupljeno: 15. siječnja 2026.https://docs.cloud.google.com/bigquery/docs/analytics-hub-introduction

[10] BigQuery Explained (architecture overview, Google Cloud Blog). Dostupno na: https://cloud.google.com/blog/products/data-analytics/new-blog-series-bigquery-explained-overview, Pristupljeno: 15. siječnja 2026.https://cloud.google.com/blog/products/data-analytics/new-blog-series-bigquery-explained-overview