Clinical Trials Intelligence Platform

End-to-End Serverless Data Engineering Pipeline on Google Cloud Platform

Index

1. Introduction

1.1 Project Background

Clinical trial data is distributed across multiple international registries and regulatory platforms. These sources differ in format, structure, and completeness, making centralized analytics difficult.

1.2 Business Problem

Manual ingestion and cleaning of registry data was:

- Slow
- Error-prone
- Non-scalable
- Not idempotent
- Lacking monitoring and automation

1.3 Project Objective

Design and implement a fully automated, idempotent, serverless, monitored, scalable data pipeline on Google Cloud Platform that integrates multi-source clinical trial data into a unified analytics warehouse.

---

2. Data Sources

2.1 ClinicalTrials.gov (API Subset)

- Accessed via REST API
- JSON format
- Condition-filtered extraction

2.2 WHO ICTRP

- Aggregated global registry dataset
- Structured downloadable format

2.3 EU Clinical Trials Register (EUCTR)

- Structured CSV/TXT datasets
- Drug trials in EU/EEA

2.4 ISRCTN Registry

- API / CSV export
- Global intervention studies

2.5 EMA Clinical Data

- Regulatory study reports
- Structured data extracts

---

3. Data Characteristics & Handling Strategy

3.1 Multi-Format Data

Sources provided:

- JSON (API responses)
- CSV
- Excel
- TXT

All formats were programmatically parsed and converted into Pandas DataFrames before normalization.

3.2 Inconsistent Schemas

Observed differences:

- Different column names for same field

- Nested vs flat structures
- Status value inconsistencies

Solution:

- Canonical schema mapping
- Defensive extraction using `.get()`
- Standardized field names before warehouse load

3.3 Missing Identifiers

Some records lacked:

- Secondary IDs
- Consistent registry identifiers

Solution:

- Used primary registry ID
- Validation warnings logged
- MERGE-based idempotent load

3.4 Free-Text Conditions

Medical conditions were:

- Non-standardized
- Multi-value strings
- No ontology mapping

Decision:
Stored raw values.
Deferred ontology harmonization to future enrichment phase.

3.5 Encoding Issues

Observed:

- Non-ASCII characters
- Special characters in sponsor names

Solution:
Explicit UTF-8 handling during ingestion.

3.6 Sponsor Inconsistencies

Example:

- “Pfizer”
- “Pfizer Inc.”
- “Pfizer Ltd”

Decision:
Preserved raw sponsor text.
Entity resolution deferred.

---

4. System Architecture (GCP)

4.1 Architecture Overview

Serverless stack:

- IAM
- BigQuery
- Docker
- Artifact Registry
- Cloud Run (Jobs)
- Cloud Scheduler
- Cloud Logging
- Log-based Alerts

4.2 IAM & Security Configuration

Configured service accounts with:

- BigQuery Data Editor
- BigQuery Job User
- Cloud Run Invoker
- Logging Writer

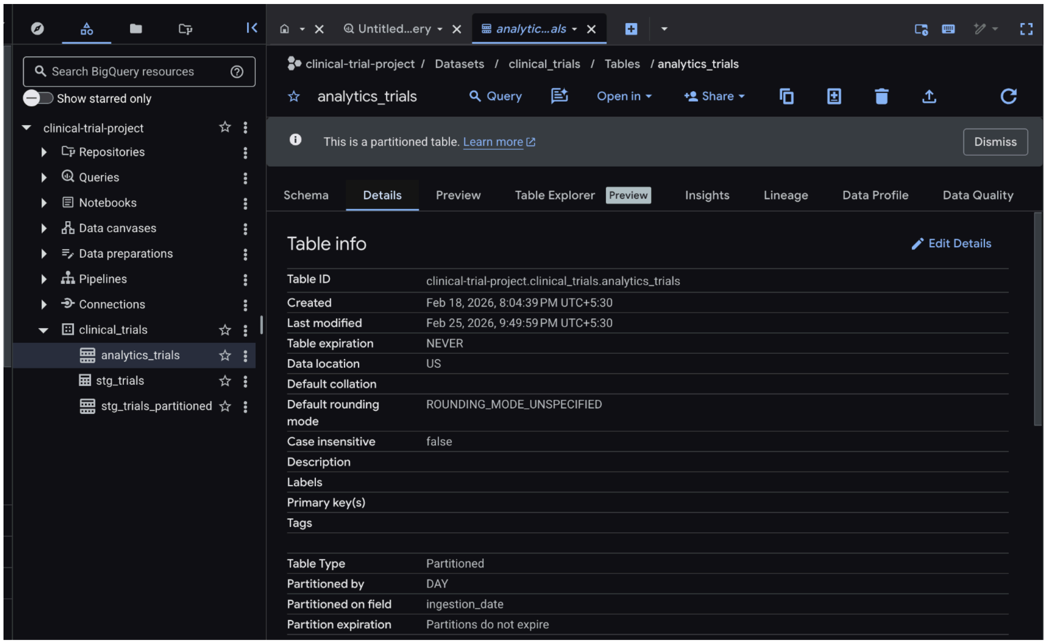
Resolved 403 permission errors during setup.

4.3 BigQuery Data Warehouse

Dataset: `clinical_trials`

Tables:

- `stg_trials`
- `Analytics_trials`



This is a partitioned table. Learn more

Table info

Table ID	clinical-trial-project.clinical_trials.analytics_trials
Created	Feb 18, 2026, 8:04:39 PM UTC+5:30
Last modified	Feb 25, 2026, 9:49:59 PM UTC+5:30
Table expiration	NEVER
Data location	US
Default collation	
Default rounding mode	ROUNDING_MODE_UNSPECIFIED
Case insensitive	false
Description	
Labels	
Primary key(s)	
Tags	
Table Type	Partitioned
Partitioned by	DAY
Partitioned on field	ingestion_date
Partition expiration	Partitions do not expire

Features:

- Partitioned by ingestion_date
- Idempotent MERGE
- Columnar storage optimization

4.4 Docker Containerization

Pipeline containerized using:

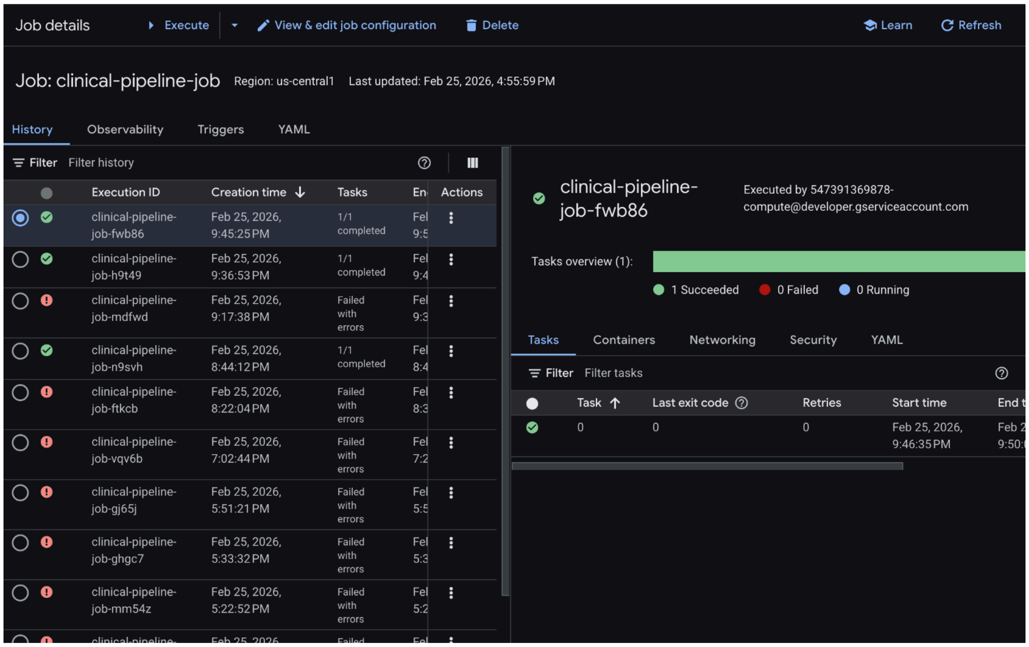
```
FROM python:3.10
WORKDIR /app
COPY requirements.txt .
RUN pip install --no-cache-dir -r requirements.txt
COPY . .
CMD ["python", "main.py"]
```

4.5 Artifact Registry

Used to store Docker images for Cloud Run execution.

4.6 Cloud Run (Compute Layer)

- Executes ingestion pipeline
- Serverless container execution
- Auto-scaling
- Timeout configured
- Service account attached



Job: clinical-pipeline-job Region: us-central1 Last updated: Feb 25, 2026, 4:55:59 PM

		Execution ID	Creation time ↓	Tasks	En	Actions
◉	✓	clinical-pipeline-job-fwb86	Feb 25, 2026, 9:45:25 PM	1/1 completed	Fel 9:5	⋮
○	✓	clinical-pipeline-job-h9t49	Feb 25, 2026, 9:36:53 PM	1/1 completed	Fel 9:4	⋮
○	!	clinical-pipeline-job-mdfwd	Feb 25, 2026, 9:17:38 PM	Failed with errors	Fel 9:3	⋮
○	✓	clinical-pipeline-job-n9svh	Feb 25, 2026, 8:44:12 PM	1/1 completed	Fel 8:4	⋮
○	!	clinical-pipeline-job-ftkcb	Feb 25, 2026, 8:22:04 PM	Failed with errors	Fel 8:3	⋮
○	!	clinical-pipeline-job-vqv6b	Feb 25, 2026, 7:02:44 PM	Failed with errors	Fel 7:2	⋮
○	!	clinical-pipeline-job-gj65j	Feb 25, 2026, 5:51:21 PM	Failed with errors	Fel 5:5	⋮
○	!	clinical-pipeline-job-ghgc7	Feb 25, 2026, 5:33:32 PM	Failed with errors	Fel 5:3	⋮
○	!	clinical-pipeline-job-mm54z	Feb 25, 2026, 5:22:52 PM	Failed with errors	Fel 5:2	⋮
○	!	clinical-pipeline-	Feb 25, 2026	Failed	Fel	⋮

✓ clinical-pipeline-job-fwb86

Executed by 547391369878-compute@developer.gserviceaccount.com

Tasks overview (1):

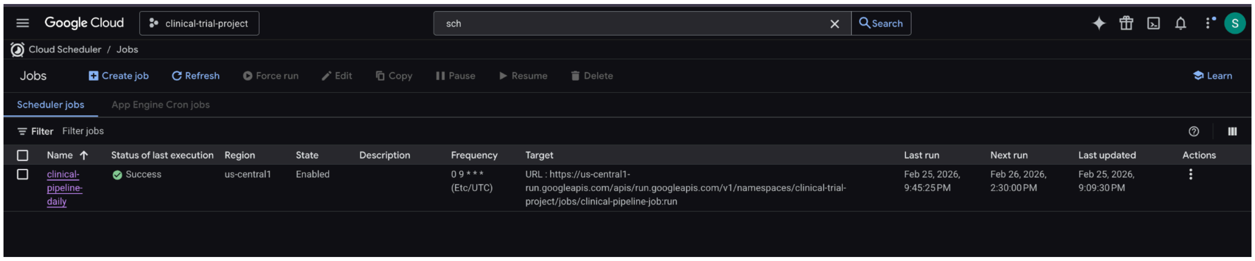
● 1 Succeeded ● 0 Failed ● 0 Running

●	Task ↑	Last exit code ⓘ	Retries	Start time	End t
✓	0	0	0	Feb 25, 2026, 9:46:35 PM	Feb 2 9:50:

4.7 Cloud Scheduler (Automation Layer)

- Daily trigger
- Cron: `0 9 * * *`
- HTTP trigger to Cloud Run Job

Timezone to be in UTC, 9am UTC is 1430 IST



□	Name ↑	Status of last execution	Region	State	Description	Frequency	Target	Last run	Next run	Last updated	Actions
□	clinical-pipeline-daily	✓ Success	us-central1	Enabled		0 9 * * * (Etc/UTC)	URL : https://us-central1-run.googleapis.com/apis/run.googleapis.com/v1/namespaces/clinical-trial-project/jobs/clinical-pipeline-job:run	Feb 25, 2026, 9:45:25 PM	Feb 26, 2026, 2:30:00 PM	Feb 25, 2026, 9:09:30 PM	⋮

4.8 Cloud Logging

Logs captured:

- Pipeline start/end
- Record counts
- Validation warnings
- Errors

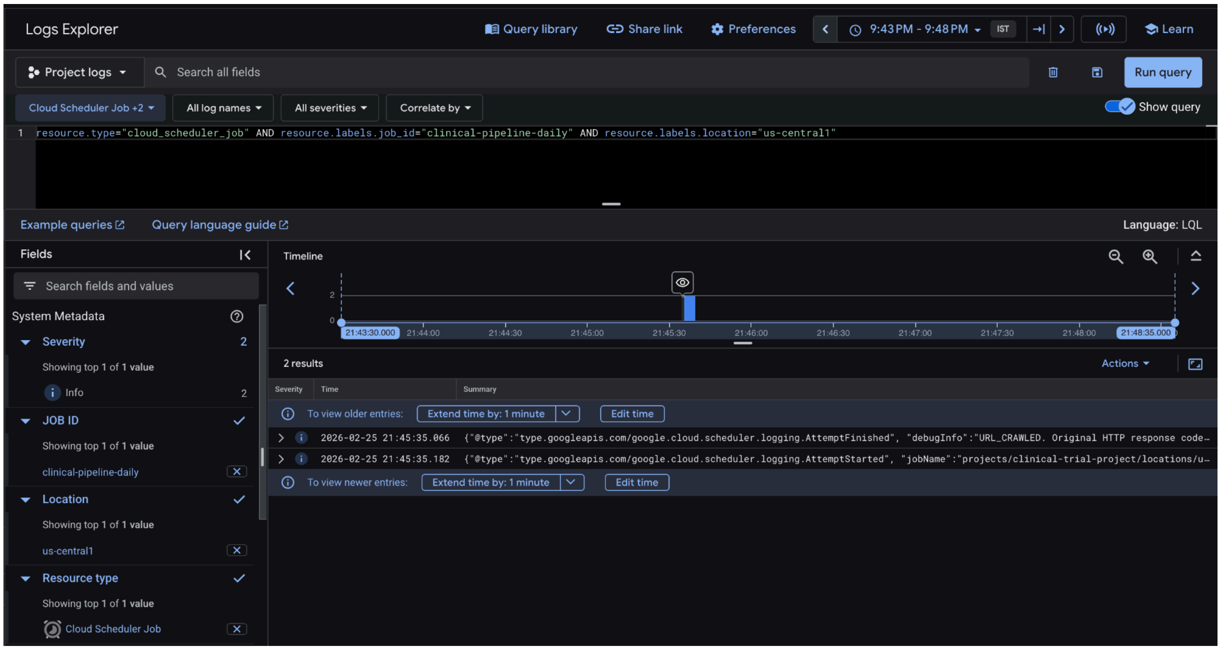
4.9 Logging-Based Alerts

Monitoring policy created:

Query:

resource.type="cloud_run_revision"
severity>=ERROR

Triggers notification if job fails.



```
resource.type="cloud_scheduler_job" AND resource.labels.job_id="clinical-pipeline-daily" AND resource.labels.location="us-central1"
```

System Metadata

- Severity 2 — Showing top 1 of 1 value — Info 2
- JOB ID — Showing top 1 of 1 value — clinical-pipeline-daily
- Location — Showing top 1 of 1 value — us-central1
- Resource type — Showing top 1 of 1 value — Cloud Scheduler Job

```
Severity  Time                     Summary
To view older entries:
          2026-02-25 21:45:35.066  {"@type":"type.googleapis.com/google.cloud.scheduler.logging.AttemptFinished", "debugInfo":"URL_CRAWLED. Original HTTP response code…
          2026-02-25 21:45:35.182  {"@type":"type.googleapis.com/google.cloud.scheduler.logging.AttemptStarted", "jobName":"projects/clinical-trial-project/locations/u…
To view newer entries:
```

5. Data Pipeline Design

5.1 Ingestion Layer

- API calls
- File parsing
- Pagination handling

5.2 Cleaning & Normalization

- Canonical schema mapping
- Type casting
- Null handling

5.3 Validation Layer

Implemented in `validation.py`:

- Empty dataset detection
- Null ID warnings
- Record count checks

5.4 Idempotent Loading Strategy

BigQuery MERGE:

```
MERGE target T
USING staging S
ON T.registry_id = S.registry_id
WHEN MATCHED THEN UPDATE SET *
WHEN NOT MATCHED THEN INSERT ROW
```

Ensures safe reruns.

5.5 Partitioning Strategy

Partitioned by ingestion_date for:

- Cost reduction
- Faster scans
- Efficient incremental loads

---

6. Codebase Structure

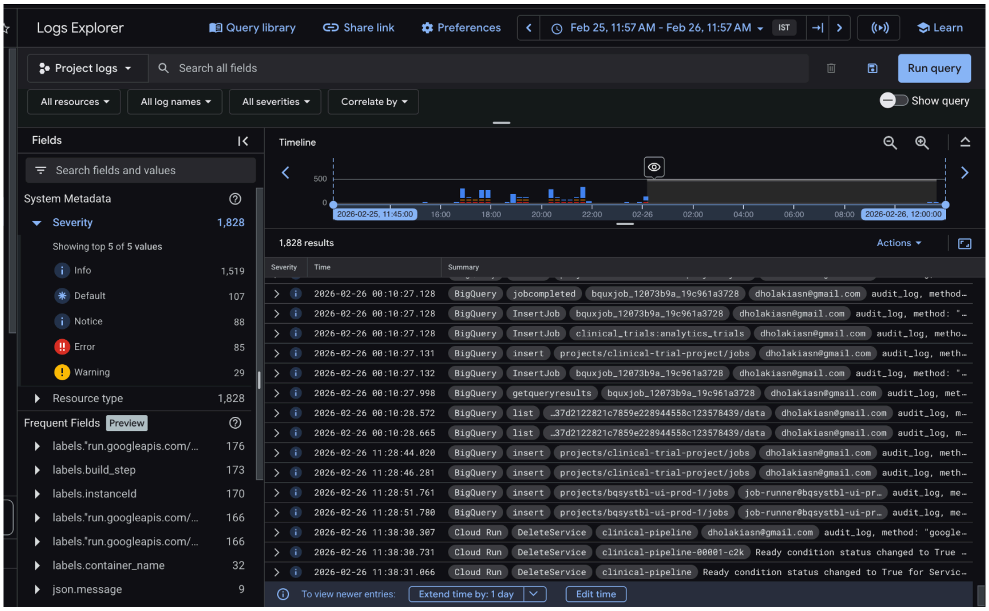
```
clinical-trial-pipeline/
|
├── main.py
├── validation.py
├── requirements.txt
├── Dockerfile
└── README.md
```

---

7. Monitoring & Observability

7.1 Log Capture

Cloud Run logs streamed to Cloud Logging.



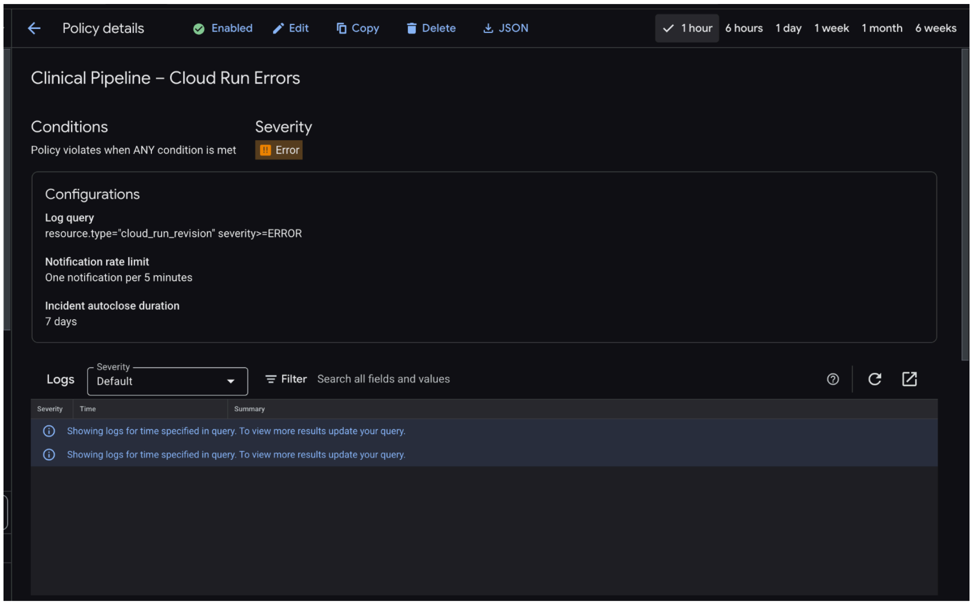
System Metadata	
Severity	1,828
Showing top 5 of 5 values	
Info	1,519
Default	107
Notice	88
Error	85
Warning	29
Resource type	1,828
Frequent Fields Preview	
labels."run.googleapis.com/...	176
labels.build_step	173
labels.instanceId	170
labels."run.googleapis.com/...	166
labels."run.googleapis.com/...	166
labels.container_name	32
json.message	9

1,828 results

```
Severity  Time                     Summary
i  2026-02-26 00:10:27.128  BigQuery  jobcompleted  bquxjob_12073b9a_19c961a3728  dholakiasn@gmail.com  audit_log, method…
i  2026-02-26 00:10:27.128  BigQuery  InsertJob  bquxjob_12073b9a_19c961a3728  dholakiasn@gmail.com  audit_log, method: "…
i  2026-02-26 00:10:27.128  BigQuery  InsertJob  clinical_trials:analytics_trials  dholakiasn@gmail.com  audit_log, metho…
i  2026-02-26 00:10:27.131  BigQuery  insert  projects/clinical-trial-project/jobs  dholakiasn@gmail.com  audit_log, meth…
i  2026-02-26 00:10:27.132  BigQuery  InsertJob  bquxjob_12073b9a_19c961a3728  dholakiasn@gmail.com  audit_log, method: "…
i  2026-02-26 00:10:27.998  BigQuery  getqueryresults  bquxjob_12073b9a_19c961a3728  dholakiasn@gmail.com  audit_log, met…
i  2026-02-26 00:10:28.572  BigQuery  list  …37d2122821c7859e228944558c123578439/data  dholakiasn@gmail.com  audit_log, m…
i  2026-02-26 00:10:28.665  BigQuery  list  …37d2122821c7859e228944558c123578439/data  dholakiasn@gmail.com  audit_log, m…
i  2026-02-26 11:28:44.020  BigQuery  insert  projects/clinical-trial-project/jobs  dholakiasn@gmail.com  audit_log, meth…
i  2026-02-26 11:28:46.281  BigQuery  insert  projects/clinical-trial-project/jobs  dholakiasn@gmail.com  audit_log, meth…
i  2026-02-26 11:28:51.761  BigQuery  insert  projects/bqsystbl-ui-prod-1/jobs  job-runner@bqsystbl-ui-pr…  audit_log, me…
i  2026-02-26 11:28:51.780  BigQuery  insert  projects/bqsystbl-ui-prod-1/jobs  job-runner@bqsystbl-ui-pr…  audit_log, me…
i  2026-02-26 11:38:30.307  Cloud Run  DeleteService  clinical-pipeline  dholakiasn@gmail.com  audit_log, method: "google…
i  2026-02-26 11:38:30.731  Cloud Run  DeleteService  clinical-pipeline-00001-c2k  Ready condition status changed to True …
i  2026-02-26 11:38:31.066  Cloud Run  DeleteService  clinical-pipeline  Ready condition status changed to True for Servic…
```

7.2 Error Detection

Severity-based logging implemented.



Clinical Pipeline – Cloud Run Errors

Conditions

Policy violates when ANY condition is met

Severity

Error

Configurations

Log query
resource.type="cloud_run_revision" severity>=ERROR

Notification rate limit
One notification per 5 minutes

Incident autoclose duration
7 days

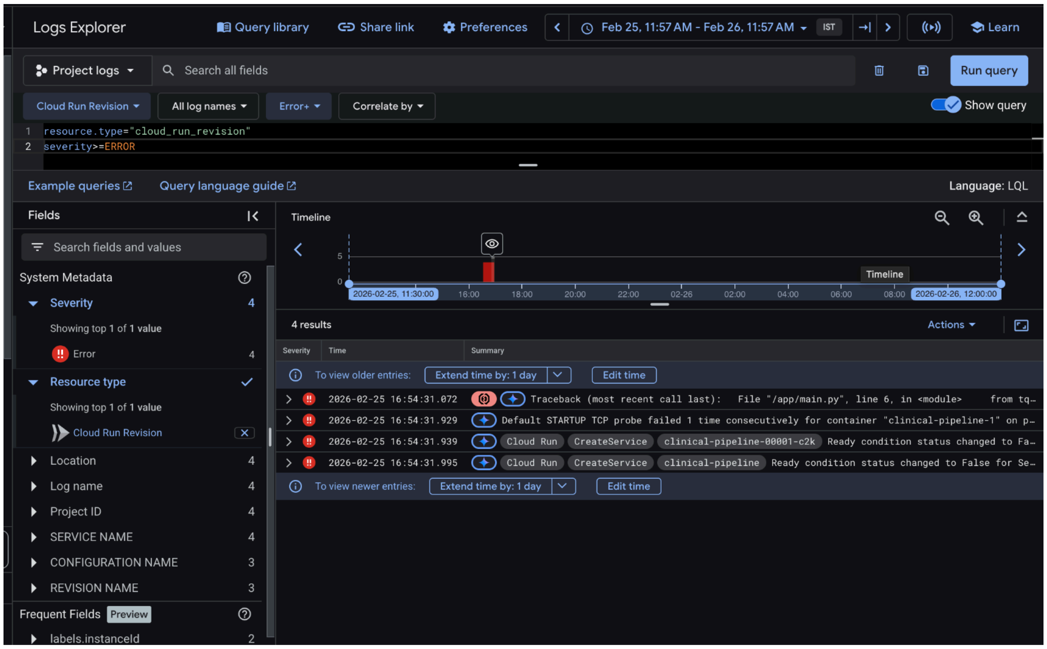
Severity: Default

Showing logs for time specified in query. To view more results update your query.

Showing logs for time specified in query. To view more results update your query.

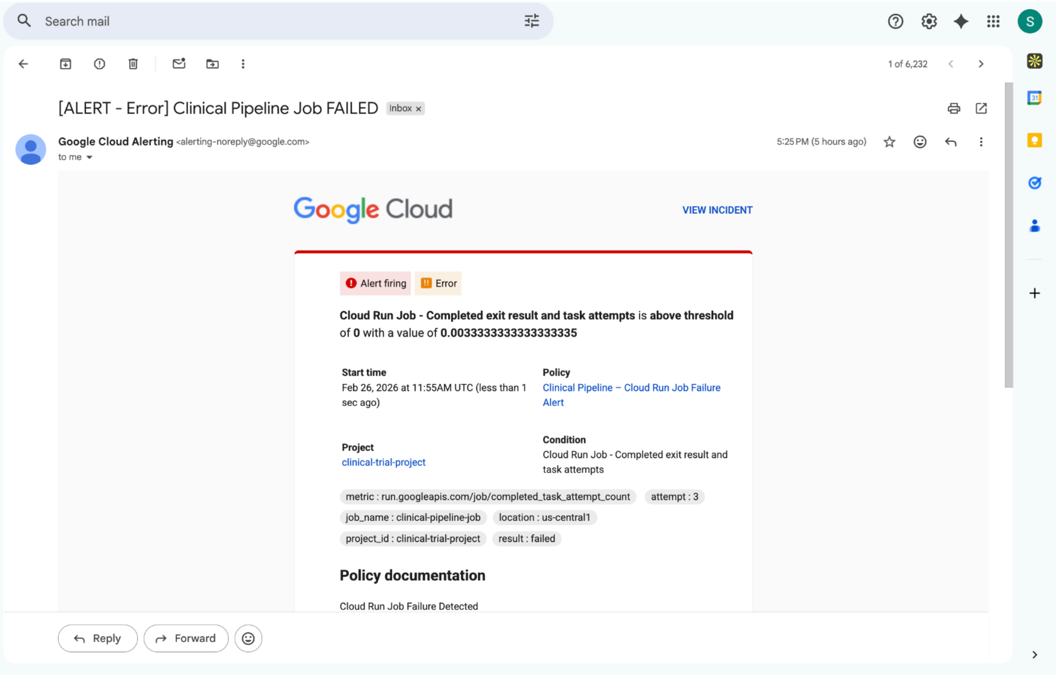
7.3 Alert Policy

Log-based alert triggers on ERROR.



```
resource.type="cloud_run_revision"
severity>=ERROR
```

```
2026-02-25 16:54:31.072  Traceback (most recent call last):   File "/app/main.py", line 6, in <module>     from tq…
2026-02-25 16:54:31.929  Default STARTUP TCP probe failed 1 time consecutively for container "clinical-pipeline-1" on p…
2026-02-25 16:54:31.939  Cloud Run  CreateService  clinical-pipeline-00001-c2k  Ready condition status changed to Fa…
2026-02-25 16:54:31.995  Cloud Run  CreateService  clinical-pipeline  Ready condition status changed to False for Se…
```



[ALERT - Error] Clinical Pipeline Job FAILED

Google Cloud Alerting <alerting-noreply@google.com>
to me

Alert firing | Error

Cloud Run Job - Completed exit result and task attempts is **above threshold** of **0** with a value of **0.0033333333333333335**

Start time
Feb 26, 2026 at 11:55AM UTC (less than 1 sec ago)

Policy
Clinical Pipeline – Cloud Run Job Failure Alert

Project
clinical-trial-project

Condition
Cloud Run Job - Completed exit result and task attempts

metric : run.googleapis.com/job/completed_task_attempt_count
attempt : 3
job_name : clinical-pipeline-job
location : us-central1
project_id : clinical-trial-project
result : failed

Policy documentation

Cloud Run Job Failure Detected

8. Analytics & Insight Demonstration

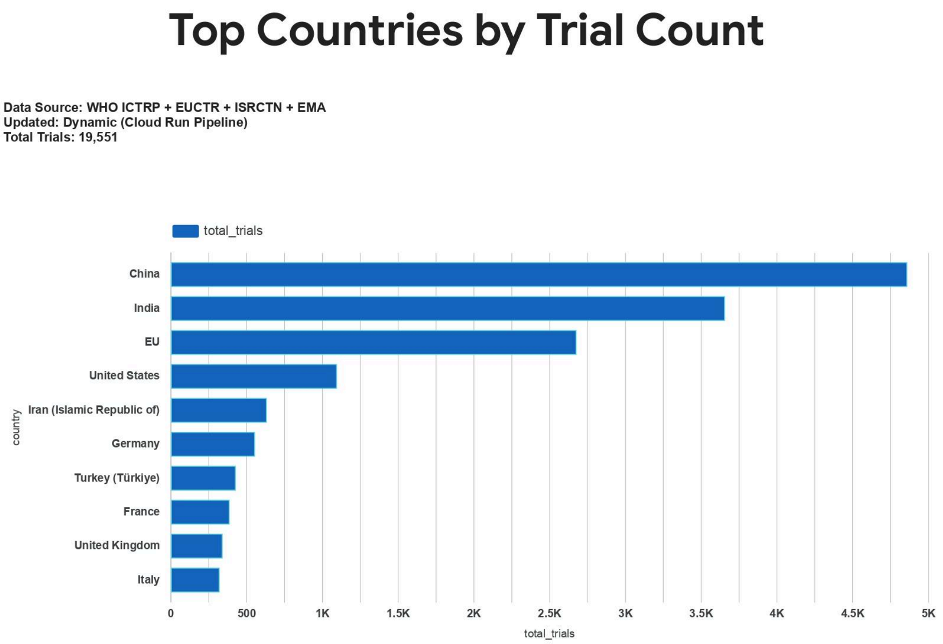
8.1 Sample BigQuery Queries

- Trials by year
- Trials by status
- Top sponsors
- Condition distribution

8.2 Looker Studio Dashboards

Visualizations:

- Top countries- https://lookerstudio.google.com/reporting/e3ebfdf2-9e34-48f9-8b06-088484ae1464

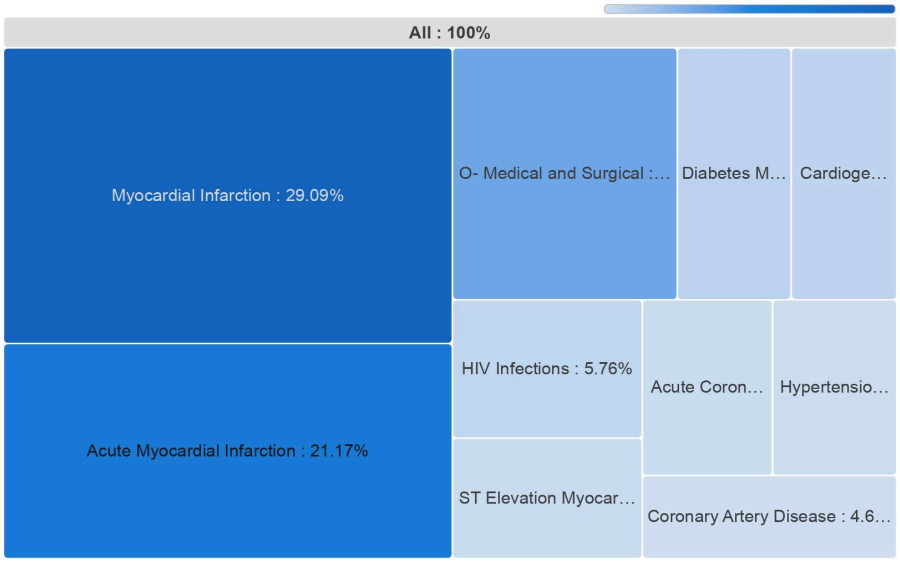


Top Countries by Trial Count

Data Source: WHO ICTRP + EUCTR + ISRCTN + EMA
Updated: Dynamic (Cloud Run Pipeline)
Total Trials: 19,551

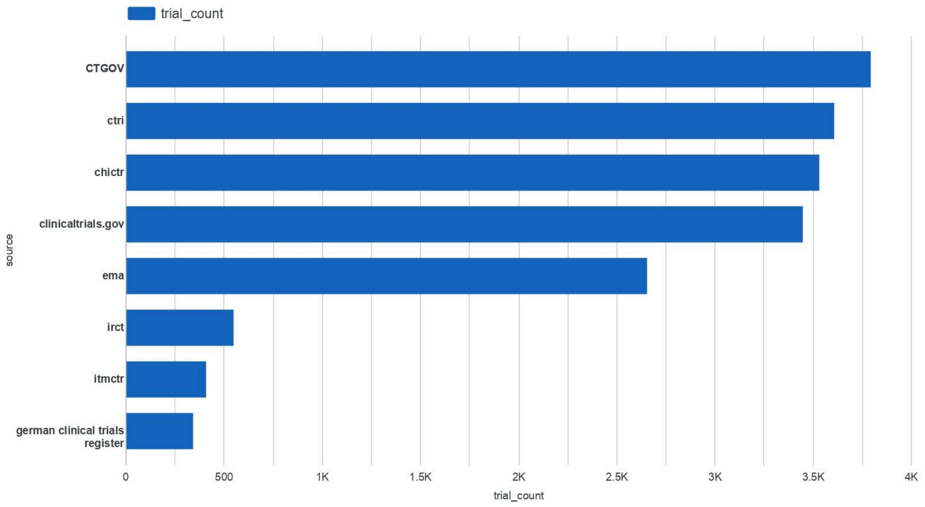
- Top conditions- https://lookerstudio.google.com/reporting/cd51f9f3-8f28-4db9-a3ae-05846badaee5

Top 10 Clinical Trial Conditions



- Source distribution - https://lookerstudio.google.com/reporting/0c5b3228-b4f5-492d-b2d3-ea6c02a18471

Clinical Trial Distribution by Registry Source



“It shows the distribution of integrated clinical trials by registry source. We can see the majority of the dataset comes from CTGov, CTRI, and ChiCTR, indicating strong representation from US, Indian, and Chinese registries.

9. Challenges Faced & Resolutions

9.1 Idempotency

Resolved via MERGE.

9.2 Scheduler Timezone Issues

Resolved via explicit timezone configuration.

9.3 IAM Permission Errors

Resolved via role assignment.

9.4 Docker Dependency Failures

Resolved via updated requirements.txt.

9.5 Cloud Run Startup Failures

Resolved via correct container CMD & configuration.

9.6 API Pagination

Resolved via nextPageToken looping.

---

10. Cost Optimization Strategy

- Fully serverless
- No idle compute
- Pay-per-execution model
- Partitioned warehouse
- No persistent VMs

---

11. Conclusion

This project successfully delivers:

- Multi-source ingestion
- Schema normalization
- Idempotent warehouse loading
- Automated scheduling
- Monitoring & alerting
- Scalable, serverless architecture
- Analytical dashboards

It transforms heterogeneous clinical trial data into a centralized, analytics-ready platform using modern cloud-native engineering principles.