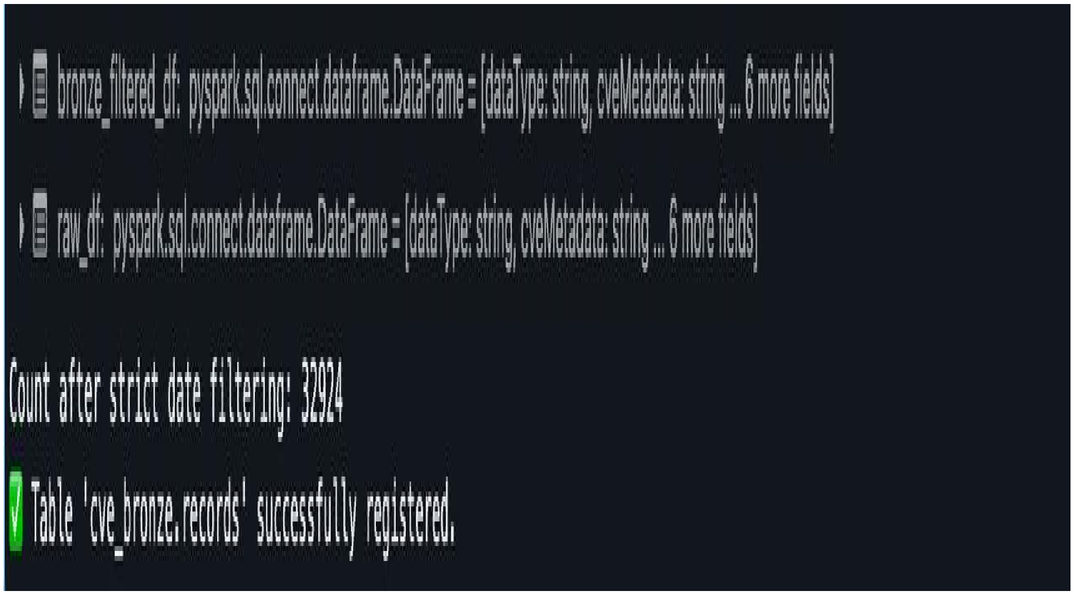
CVE Lakehouse Assignment – Final Report

1. Introduction

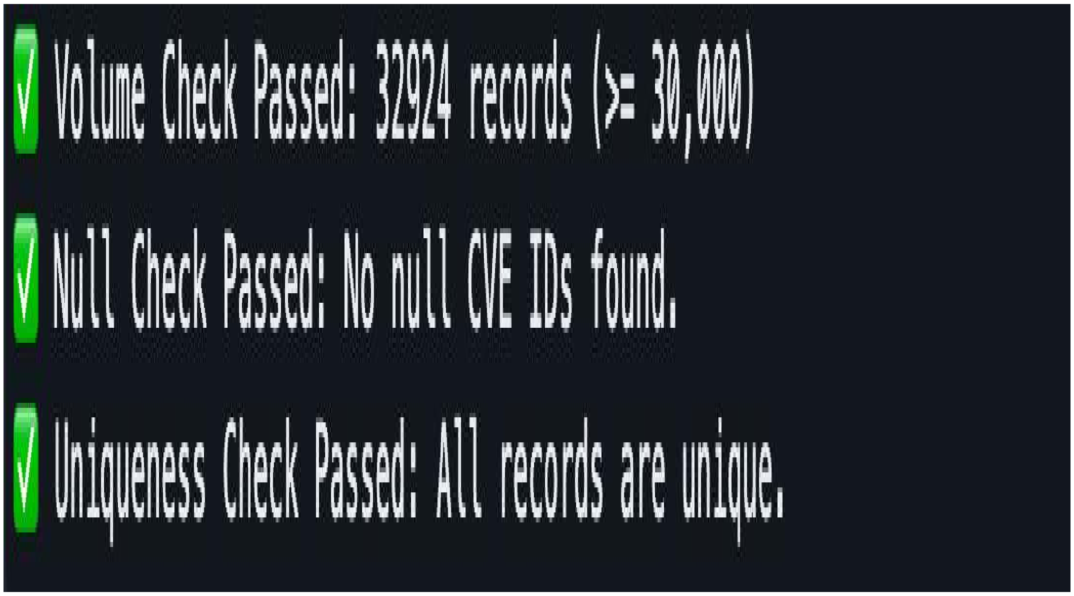
This report documents the end-to-end implementation of a CVE (Common Vulnerabilities and Exposures) Lakehouse analytics pipeline using Databricks Community Edition. The project transforms raw JSON vulnerability records from the CVE cvelistV5 repository into structured Bronze and Silver Delta tables, followed by Gold-level SQL analytics generating insights about vendors, severity distribution, and vulnerability disclosure timelines.

2. Bronze Layer – Raw Ingestion

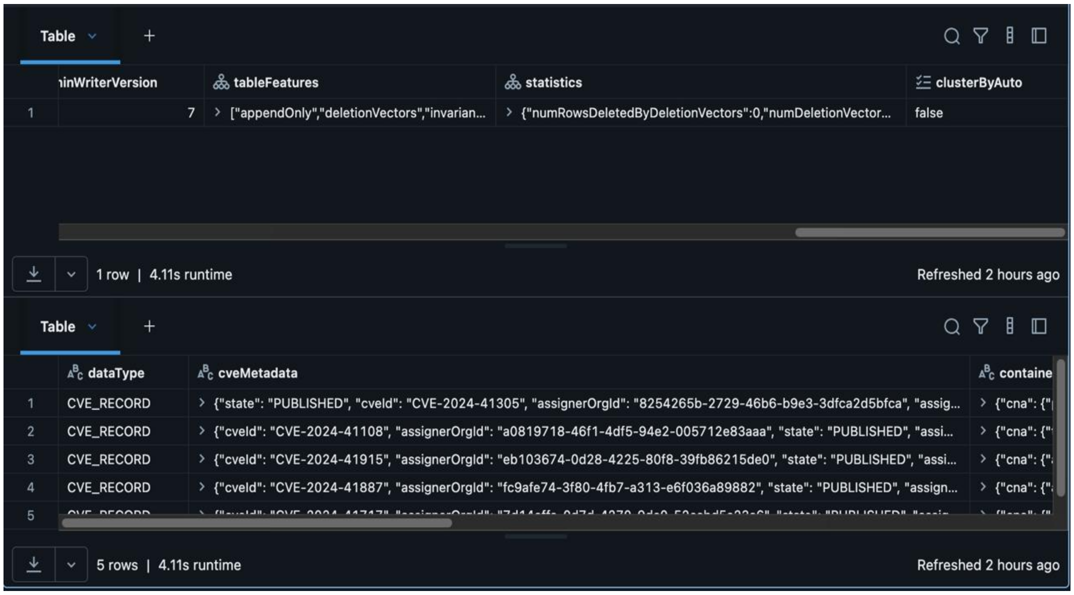
The Bronze layer loads raw JSON CVE records after downloading and extracting the cvelistV5 repository. In Databricks CE, ingestion was performed using */databricks/driver* storage to avoid DBFS restrictions. After extraction, recursive JSON loading produced a raw DataFrame, filtered to records published in 2024. Data quality checks validated record volume, null IDs, and uniqueness.



```
▸ ▤ bronze_filtered_df: pyspark.sql.connect.dataframe.DataFrame = [dataType: string, cveMetadata: string ... 6 more fields]

▸ ▤ raw_df: pyspark.sql.connect.dataframe.DataFrame = [dataType: string, cveMetadata: string ... 6 more fields]

Count after strict date filtering: 32924
✅ Table 'cve_bronze.records' successfully registered.
```



```
✅ Volume Check Passed: 32924 records (>= 30,000)

✅ Null Check Passed: No null CVE IDs found.

✅ Uniqueness Check Passed: All records are unique.
```



	ıinWriterVersion	tableFeatures	statistics	clusterByAuto
1	7	["appendOnly","deletionVectors","invarian...	{"numRowsDeletedByDeletionVectors":0,"numDeletionVector...	false

1 row | 4.11s runtime

Refreshed 2 hours ago

	dataType	cveMetadata	containe
1	CVE_RECORD	{"state": "PUBLISHED", "cveId": "CVE-2024-41305", "assignerOrgId": "8254265b-2729-46b6-b9e3-3dfca2d5bfca", "assig...	{"cna": {"
2	CVE_RECORD	{"cveId": "CVE-2024-41108", "assignerOrgId": "a0819718-46f1-4df5-94e2-005712e83aaa", "state": "PUBLISHED", "assi...	{"cna": {"
3	CVE_RECORD	{"cveId": "CVE-2024-41915", "assignerOrgId": "eb103674-0d28-4225-80f8-39fb86215de0", "state": "PUBLISHED", "assi...	{"cna": {"
4	CVE_RECORD	{"cveId": "CVE-2024-41887", "assignerOrgId": "fc9afe74-3f80-4fb7-a313-e6f036a89882", "state": "PUBLISHED", "assign...	{"cna": {"
5	CVE_RECORD	[illegible]	[illegible]

5 rows | 4.11s runtime

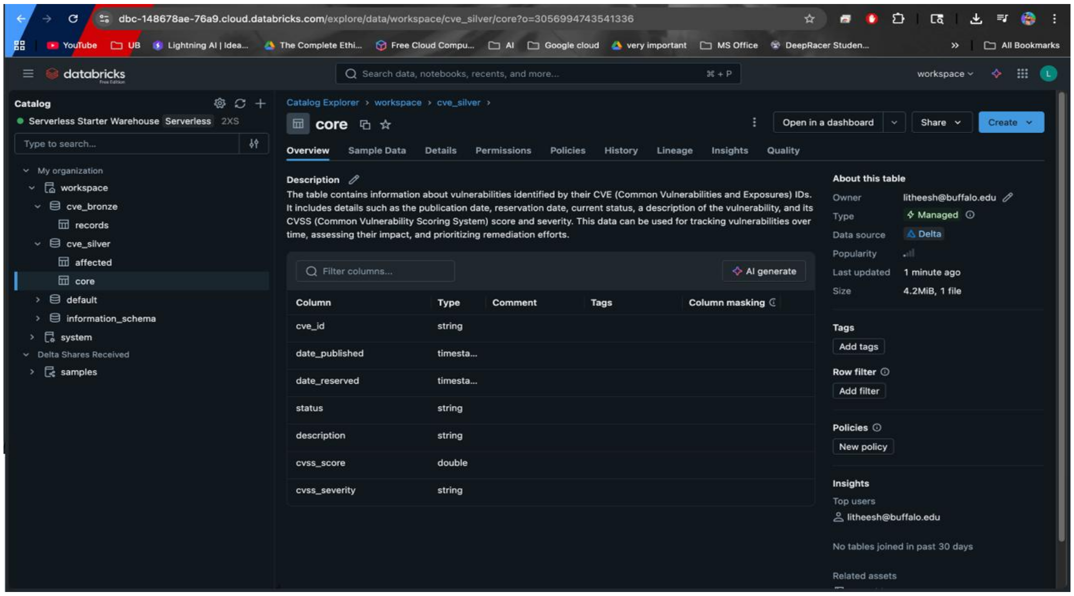
Refreshed 2 hours ago

3. Silver Layer – Normalized Tables

The Silver layer transforms the Bronze dataset into two analytical tables:

- **core** – Flattened metadata including CVE ID, publication dates, CVSS score, severity, and description.
- **affected** – Exploded vendor–product associations for each CVE.

These normalized tables enable vendor-level analytics and severity scoring.



core

Description

The table contains information about vulnerabilities identified by their CVE (Common Vulnerabilities and Exposures) IDs. It includes details such as the publication date, reservation date, current status, a description of the vulnerability, and its CVSS (Common Vulnerability Scoring System) score and severity. This data can be used for tracking vulnerabilities over time, assessing their impact, and prioritizing remediation efforts.

Column	Type	Comment	Tags	Column masking
cve_id	string			
date_published	timesta...			
date_reserved	timesta...			
status	string			
description	string			
cvss_score	double			
cvss_severity	string			

About this table

Owner	litheesh@buffalo.edu
Type	Managed
Data source	Delta
Popularity	
Last updated	1 minute ago
Size	4.2MiB, 1 file

Tags

Row filter

Policies

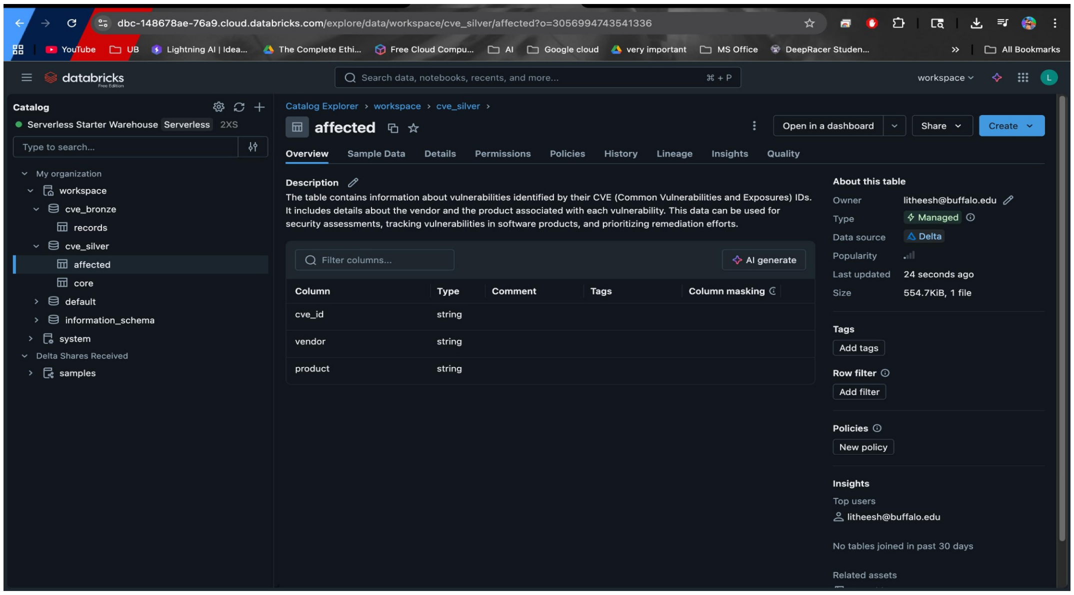
Insights

Top users

litheesh@buffalo.edu

No tables joined in past 30 days

Related assets



affected

Description

The table contains information about vulnerabilities identified by their CVE (Common Vulnerabilities and Exposures) IDs. It includes details about the vendor and the product associated with each vulnerability. This data can be used for security assessments, tracking vulnerabilities in software products, and prioritizing remediation efforts.

Column	Type	Comment	Tags	Column masking
cve_id	string			
vendor	string			
product	string			

About this table

Owner	litheesh@buffalo.edu
Type	Managed
Data source	Delta
Popularity	
Last updated	24 seconds ago
Size	554.7KiB, 1 file

Tags

Row filter

Policies

Insights

Top users

litheesh@buffalo.edu

No tables joined in past 30 days

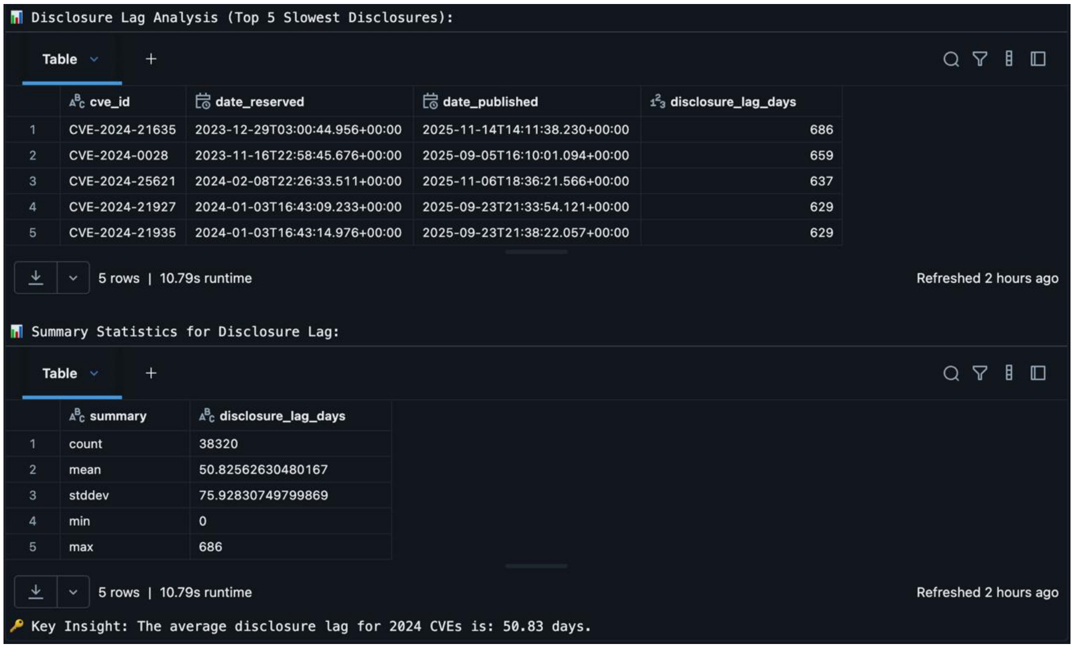
Related assets

4. Gold Layer – SQL Analytics (EDA)

SQL was applied on Silver tables to generate insights including:

- Top affected vendors
- Severity distribution
- Disclosure lag analysis

These insights help identify high-risk vendors, vulnerability severity proportions, and the efficiency of disclosure.



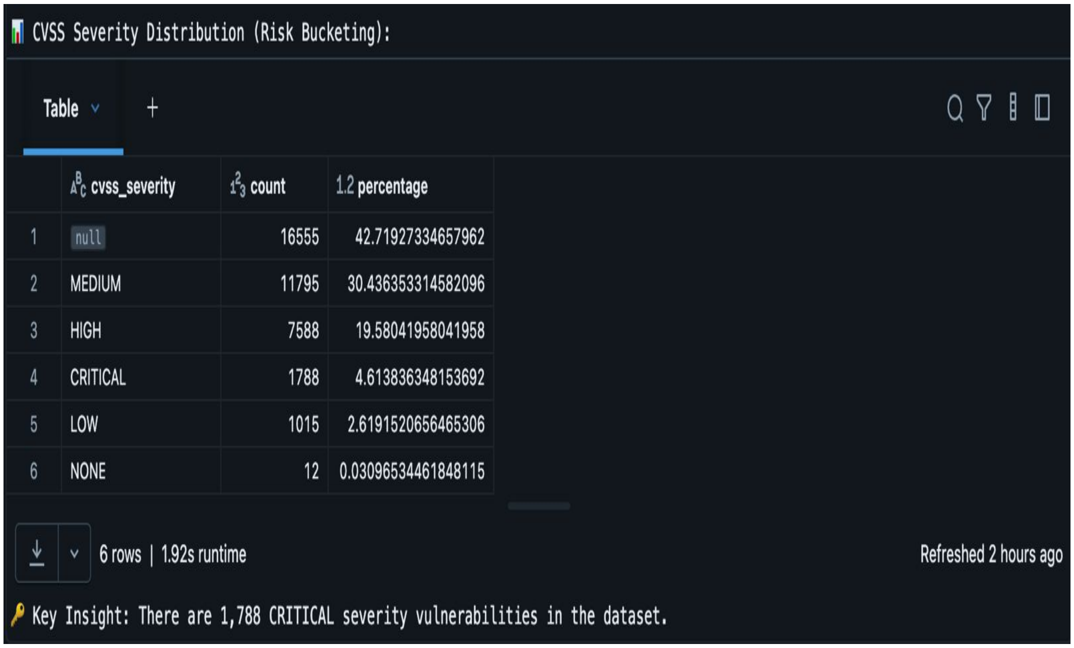
📊 Disclosure Lag Analysis (Top 5 Slowest Disclosures):

	cve_id	date_reserved	date_published	disclosure_lag_days
1	CVE-2024-21635	2023-12-29T03:00:44.956+00:00	2025-11-14T14:11:38.230+00:00	686
2	CVE-2024-0028	2023-11-16T22:58:45.676+00:00	2025-09-05T16:10:01.094+00:00	659
3	CVE-2024-25621	2024-02-08T22:26:33.511+00:00	2025-11-06T18:36:21.566+00:00	637
4	CVE-2024-21927	2024-01-03T16:43:09.233+00:00	2025-09-23T21:33:54.121+00:00	629
5	CVE-2024-21935	2024-01-03T16:43:14.976+00:00	2025-09-23T21:38:22.057+00:00	629

📊 Summary Statistics for Disclosure Lag:

	summary	disclosure_lag_days
1	count	38320
2	mean	50.82562630480167
3	stddev	75.92830749799869
4	min	0
5	max	686

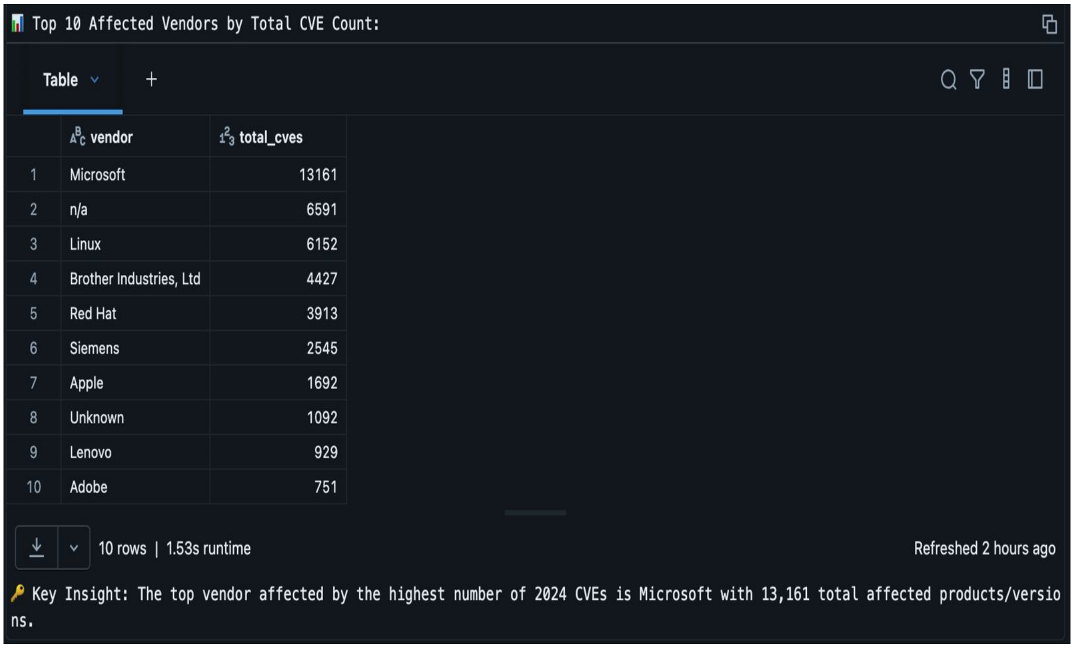
🔑 Key Insight: The average disclosure lag for 2024 CVEs is: 50.83 days.



📊 CVSS Severity Distribution (Risk Bucketing):

	cvss_severity	count	percentage
1	null	16555	42.71927334657962
2	MEDIUM	11795	30.436353314582096
3	HIGH	7588	19.58041958041958
4	CRITICAL	1788	4.613836348153692
5	LOW	1015	2.6191520656465306
6	NONE	12	0.03096534461848115

🔑 Key Insight: There are 1,788 CRITICAL severity vulnerabilities in the dataset.



📊 Top 10 Affected Vendors by Total CVE Count:

	vendor	total_cves
1	Microsoft	13161
2	n/a	6591
3	Linux	6152
4	Brother Industries, Ltd	4427
5	Red Hat	3913
6	Siemens	2545
7	Apple	1692
8	Unknown	1092
9	Lenovo	929
10	Adobe	751

10 rows | 1.53s runtime

Refreshed 2 hours ago

```
🔑 Key Insight: The top vendor affected by the highest number of 2024 CVEs is Microsoft with 13,161 total affected products/versio
ns.
```

5. Conclusion

This assignment demonstrates a complete, production-inspired Lakehouse pipeline using Databricks CE. Despite storage restrictions, the workflow successfully implemented:

- Bronze ingestion using driver storage
- Silver normalization (core + affected)
- Gold-level analytics producing actionable cybersecurity insights

The structured pipeline is scalable, reproducible, and aligns with modern Delta Lake best practices.