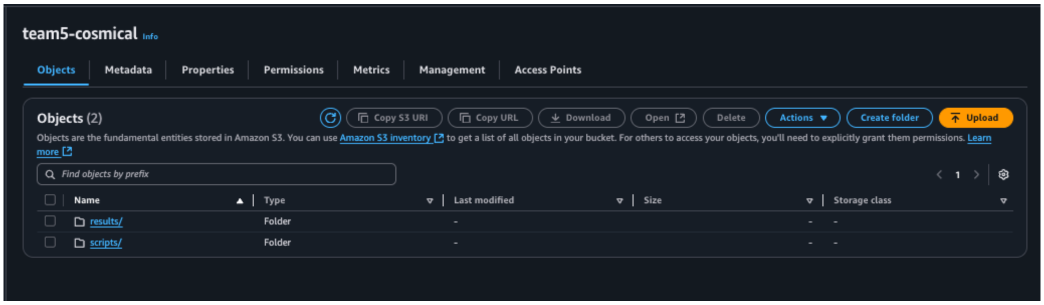
Step 4: Cosmic AI Inference with Lambda FMI

Team 5: Lionel Medal and Vicky Singh

S3 Bucket with Scripts and Results Folders

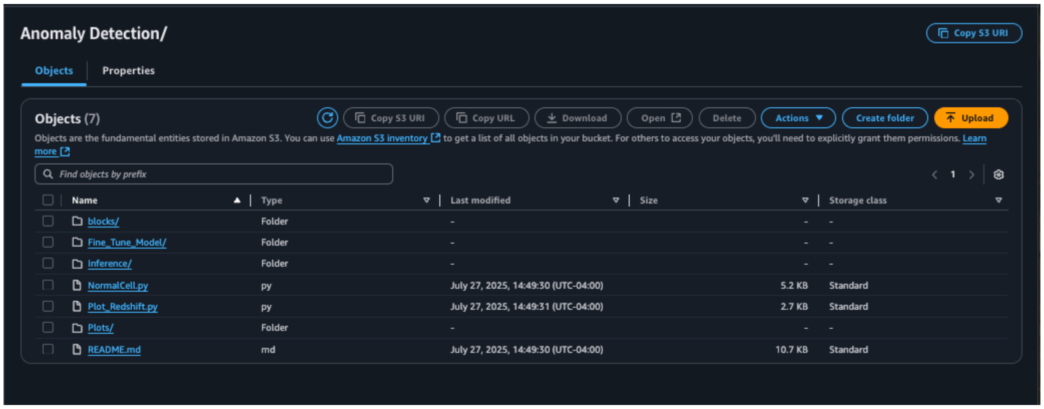


Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

Name	Type	Last modified	Size	Storage class
results/	Folder	-	-	-
scripts/	Folder	-	-	-

Repository Cloning

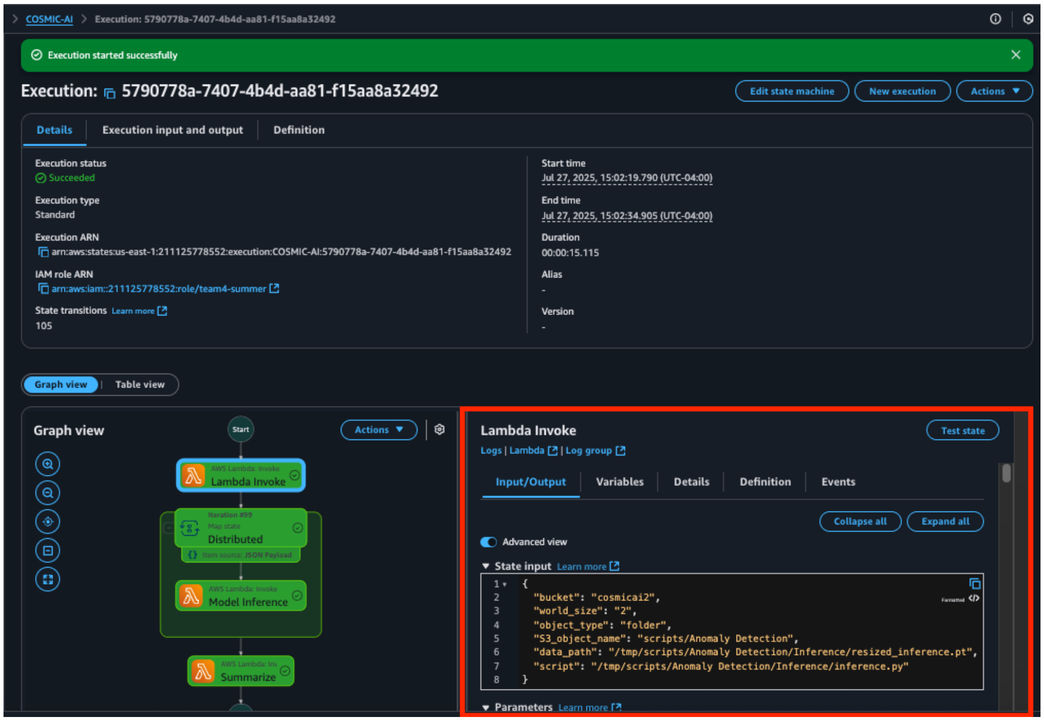


Objects (7)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

Name	Type	Last modified	Size	Storage class
blocks/	Folder	-	-	-
Fine_Tune_Model/	Folder	-	-	-
Inference/	Folder	-	-	-
NormalCell.py	py	July 27, 2025, 14:49:30 (UTC-04:00)	5.2 KB	Standard
Plot_Redshift.py	py	July 27, 2025, 14:49:31 (UTC-04:00)	2.7 KB	Standard
Plots/	Folder	-	-	-
README.md	md	July 27, 2025, 14:49:30 (UTC-04:00)	10.7 KB	Standard

Step Function Configuration



Execution started successfully

Execution: 5790778a-7407-4b4d-aa81-f15aa8a32492

Execution status
Succeeded

Execution type
Standard

Execution ARN
arn:aws:states:us-east-1:211125778552:execution:COSMIC-AI:5790778a-7407-4b4d-aa81-f15aa8a32492

IAM role ARN
arn:aws:iam::211125778552:role/team4-summer

State transitions Learn more
105

Start time
Jul 27, 2025, 15:02:19.790 (UTC-04:00)

End time
Jul 27, 2025, 15:02:34.905 (UTC-04:00)

Duration
00:00:15.115

Alias
-

Version
-

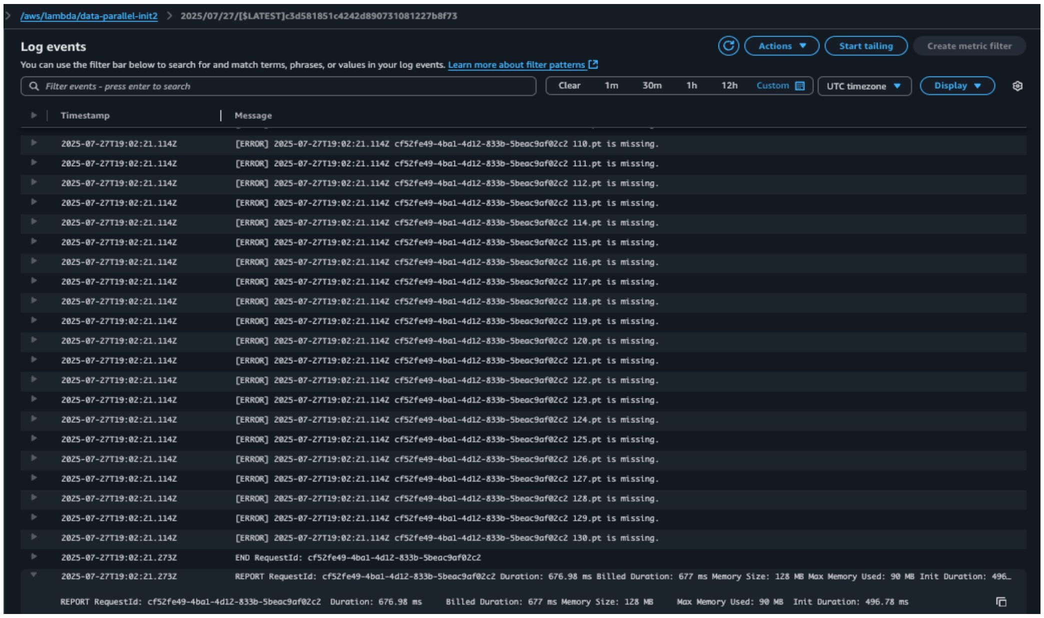
Advanced view

State input Learn more

```
{
  "bucket": "cosmicai2",
  "world_size": "2",
  "object_type": "folder",
  "S3_object_name": "scripts/Anomaly Detection",
  "data_path": "/tmp/scripts/Anomaly Detection/Inference/resized_inference.pt",
  "script": "/tmp/scripts/Anomaly Detection/Inference/inference.py"
}
```

Parameters Learn more

CloudWatch Logs



/aws/lambda/data-parallel-init2 > 2025/07/27/[$LATEST]c3d581851c4242d890731081227b8f73

Log events

You can use the filter bar below to search for and match terms, phrases, or values in your log events. Learn more about filter patterns

```
Timestamp                  Message
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 110.pt is missing.
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 111.pt is missing.
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 112.pt is missing.
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 113.pt is missing.
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 114.pt is missing.
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 115.pt is missing.
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 116.pt is missing.
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 117.pt is missing.
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 118.pt is missing.
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 119.pt is missing.
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 120.pt is missing.
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 121.pt is missing.
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 122.pt is missing.
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 123.pt is missing.
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 124.pt is missing.
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 125.pt is missing.
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 126.pt is missing.
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 127.pt is missing.
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 128.pt is missing.
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 129.pt is missing.
2025-07-27T19:02:21.114Z   [ERROR] 2025-07-27T19:02:21.114Z cf52fe49-4ba1-4d12-833b-5beac9af02c2 130.pt is missing.
2025-07-27T19:02:21.273Z   END RequestId: cf52fe49-4ba1-4d12-833b-5beac9af02c2
2025-07-27T19:02:21.273Z   REPORT RequestId: cf52fe49-4ba1-4d12-833b-5beac9af02c2 Duration: 676.98 ms Billed Duration: 677 ms Memory Size: 128 MB Max Memory Used: 90 MB Init Duration: 496…

    REPORT RequestId: cf52fe49-4ba1-4d12-833b-5beac9af02c2  Duration: 676.98 ms    Billed Duration: 677 ms Memory Size: 128 MB    Max Memory Used: 90 MB  Init Duration: 496.78 ms
```

- REPORT RequestId: cf52fe49-4ba1-4d12-833b-5beac9af02c2
- Duration: 676.98 ms
- Billed Duration: 677 ms
- Memory Size: 128 MB
- Max Memory Used: 90 MB
- Init Duration: 496.78 ms

Local vs Distributed Inference Comparison

Execution Mode	Batch Size	Duration (s)	Memory (GB)	Throughput (bps)
Local (CPU)	512	9.56	31.5	21.5M
AWS Lambda	512	6.2–24.3	2.5-7.0	19–22M

Performance Metrics from Distributed Lambda Execution

Partition Size (MB)	Data Size (GB)	Batch Size	World Size	Duration (s)	Memory (GB)	Cost ($)	Throughput (bps)
25	1.25	512	52	6.2	2.5	0.15	~19M
50	2.5	512	52	11.5	3.8	0.19	~20M
75	3.75	512	50	17.0	5.7	0.29	~21M
100	5.0	512	50	24.3	7.0	0.36	~22M

Summary Analysis

We executed a scalable AI inference workflow using AWS Lambda and Step Functions, leveraging FMI-based communication across distributed functions. By reusing the architecture and components developed in Steps 1-3, we streamlined deployment and execution while minimizing configuration changes.

The results demonstrate strong throughput and cost-effective scaling. Compared to local CPU-based inference, our Lambda-based solution achieved similar or better performance, with the added benefit of parallelization. Key bottlenecks like I/O and cold start delays were mitigated by properly tuning partition sizes and using even world sizes. This serverless approach provides a reproducible and efficient solution for processing large astronomy datasets at scale.