aws

COMMUNITY DAY

PUNE

ANNUAL EDITION

Saturday, 3rd August 2024

aws COMMUNITY DAY
PUNE
ANNUAL EDITION 2024

aws USER GROUP PUNE

From 24 Hours to 4: How Serverless Transformed Data Ingestion Pipeline

Mandar Gokhale Avinash Dalvi

“In this talk, we are not going to learn a solution, but we are going to learn a thought process or approach.”

aws
COMMUNITY DAY
PUNE
ANNUAL EDITION 2024

What’s on the table today ?

- ❑ The Problem
- ❑ Understanding the Complexity
- ❑ Why Serverless?
- ❑ Thought Process & Implementation
- ❑ Measuring Success
- ❑ Lessons Learned
- ❑ Real-World Examples
- ❑ Q&A

aws
COMMUNITY DAY
PUNE
ANNUAL EDITION 2024

aws
USER GROUP
PUNE

“Every problem has solution; it may sometime just need another perspective.”

- Katherine Russell

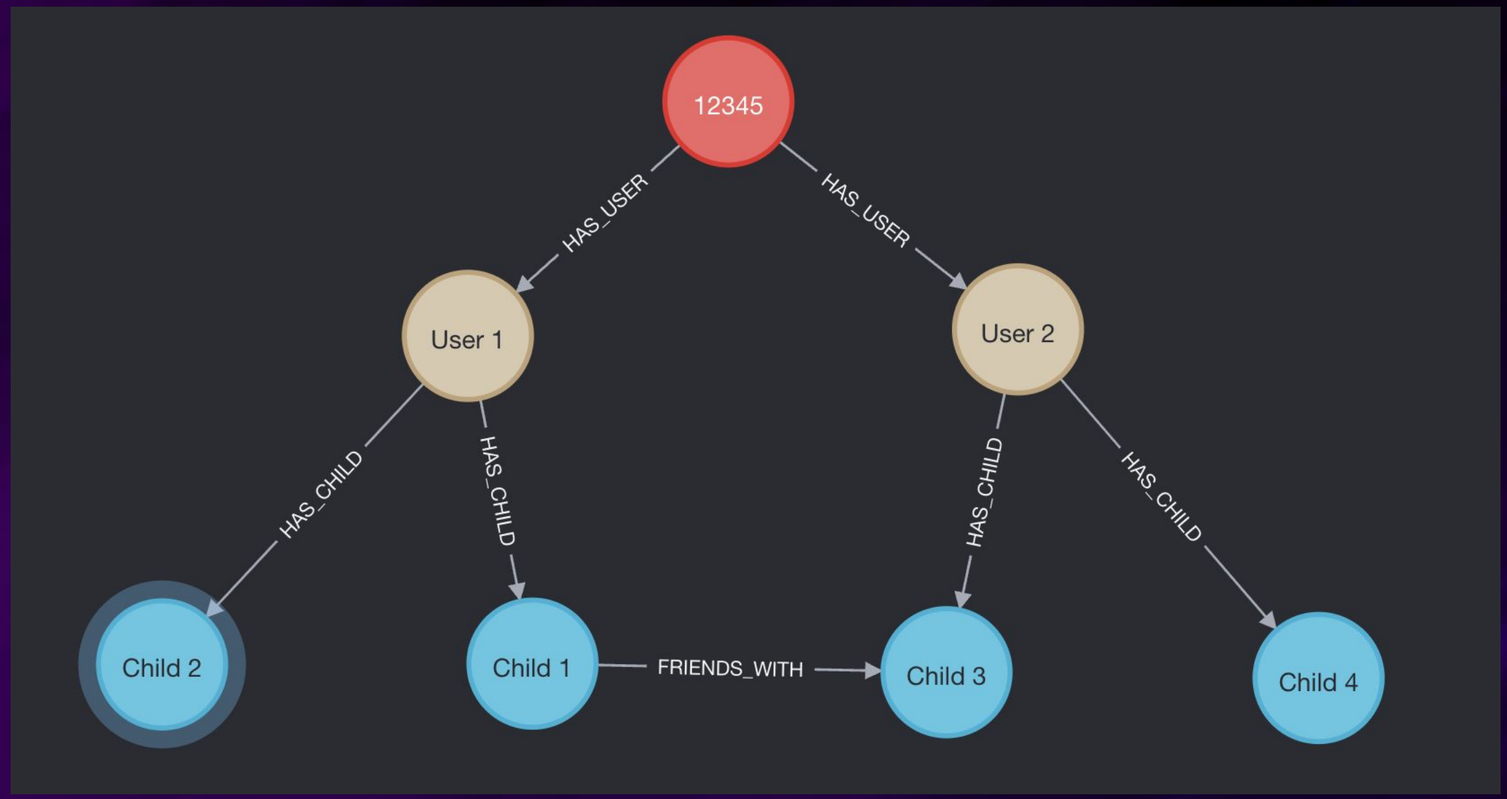
Source Data

- CMS managing business critical data
- Provides XML data exports of size ~1GB
- Contains ~200K entries
- Each entry contains large number of deeply nested, cross referencing nodes

aws COMMUNITY DAY
PUNE
ANNUAL EDITION 2024

aws USER GROUP PUNE

Data Visualization

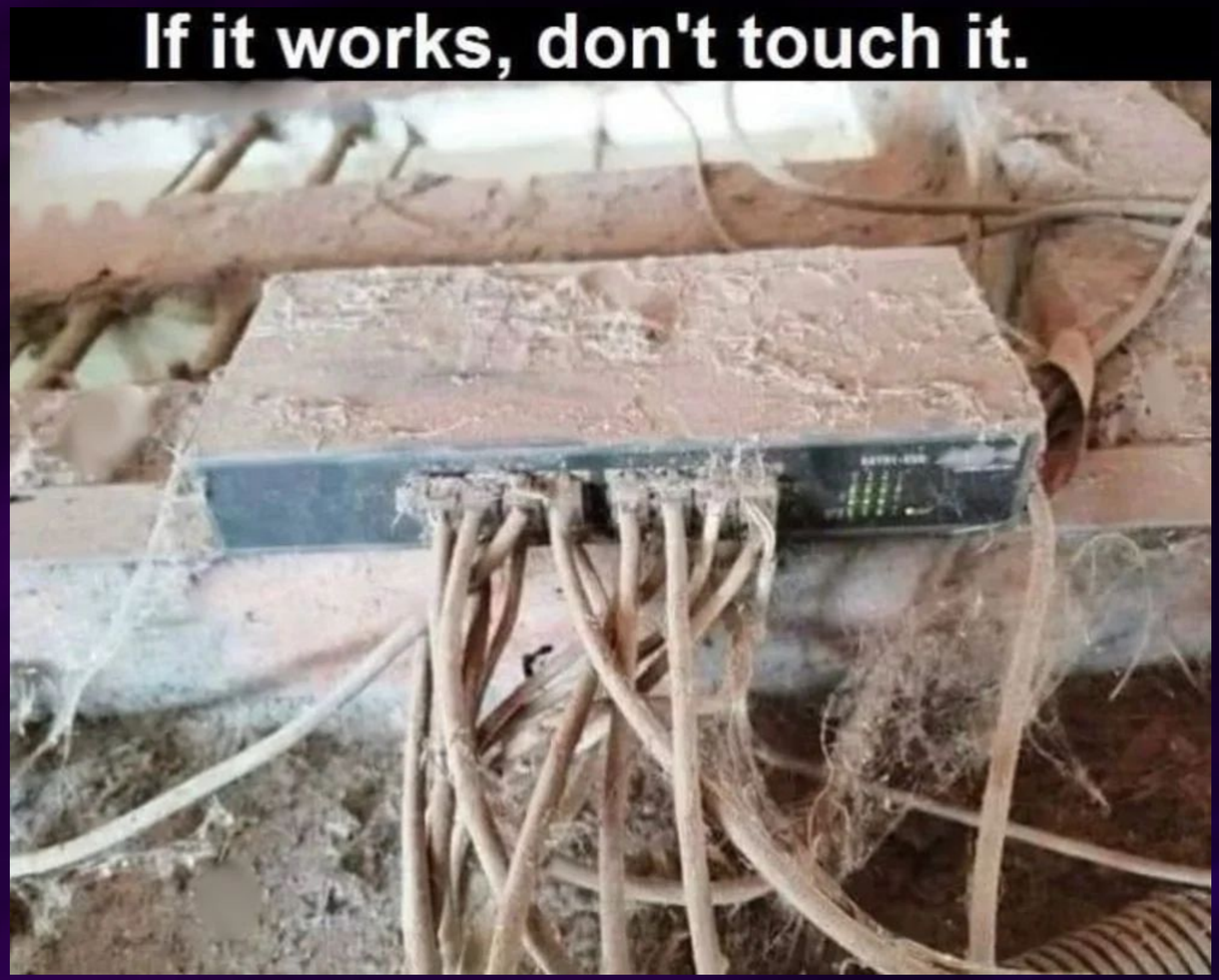


Existing Data Ingestion Pipeline

- NodeJS based app
- Running on single kubernetes pod
- Using Redis graph database for intermediate results
- Using elasticsearch for final storage

aws
COMMUNITY DAY
PUNE
ANNUAL EDITION 2024

aws
USER GROUP
PUNE



Pain Points

- By design pipeline couldn’’t be horizontally scaled to speed up ingestion
- Taking ~24 hrs to finish ingestion
- No observability
- No logic built in to retry in the event of failure
- Uncertainty or delay in content delivery
- Cost

aws COMMUNITY DAY
PUNE
ANNUAL EDITION 2024

aws USER GROUP PUNE

Objectives for Redesign

Accelerate Data Ingestion

Enhance Observability

Optimize Cost Efficiency

Facilitate easy ingestion retries

Simplify Developer Operations

aws
COMMUNITY DAY
PUNE
ANNUAL EDITION 2024

aws
USER GROUP
PUNE

Understanding the Complexity

Know your data

- ❑ Format
- ❑ Structure
- ❑ Volume
- ❑ Quality
- ❑ Source
- ❑ Lifecycle

aws COMMUNITY DAY PUNE ANNUAL EDITION 2024

aws USER GROUP PUNE

Scalability Issues

- Resource Allocation
- Load Balancing
- Maintenance Overhead
- Cost Management

aws
COMMUNITY DAY
PUNE
ANNUAL EDITION 2024

aws
USER GROUP
PUNE

Why Serverless ?

aws COMMUNITY DAY
PUNE
ANNUAL EDITION 2024

aws USER GROUP PUNE

Ordering Pizzas

- What toppings will be - RAM, Storage etc
- What base will be - Which programming language
- Small/Medium/Large - code size
- Pizza delivery time whether immediate or schedule - execution time

aws COMMUNITY DAY
PUNE
ANNUAL EDITION 2024

aws USER GROUP PUNE

Why Serverless ?

- Scalability
- Cost Efficiency
- Reduced Operational Complexity
- Faster Time to Market
- Flexibility and Innovation
- Handling Asynchronous Tasks
- Improved Reliability and Security

aws COMMUNITY DAY
PUNE
ANNUAL EDITION 2024

aws USER GROUP PUNE

Thought Process & Implementation

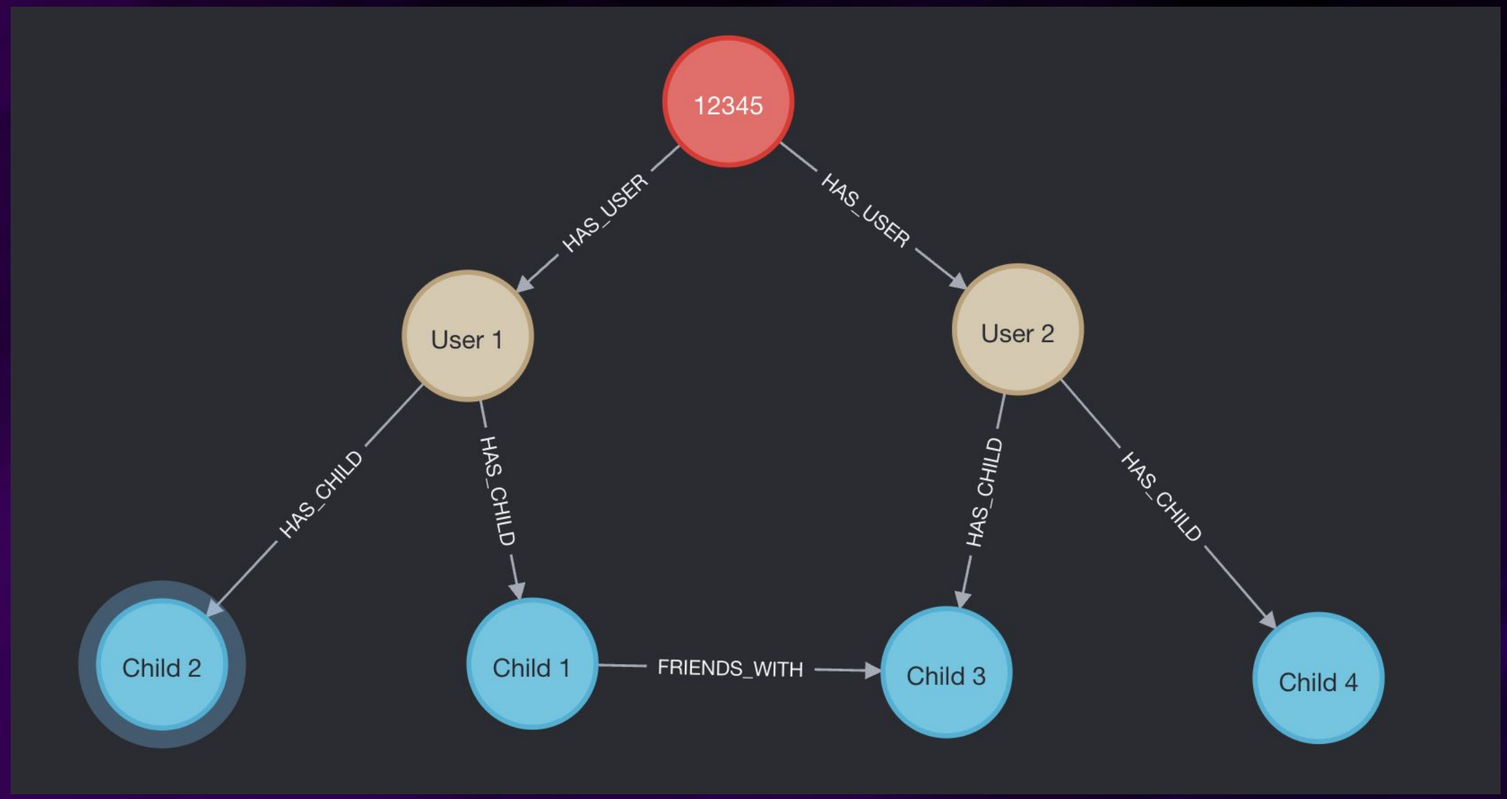
Our approach

- Lambda for compute & parallel processing
- S3 to store input xml, intermediate xml files and artifacts
- SQS for queueing records & to perform asynchronous processing
- Cloudwatch for logging & monitoring
- Step function for orchestration of lambda functions
- Neo4j for storing the data
- Elasticsearch to store data only needed for search operations

aws COMMUNITY DAY
PUNE
ANNUAL EDITION 2024

aws USER GROUP PUNE

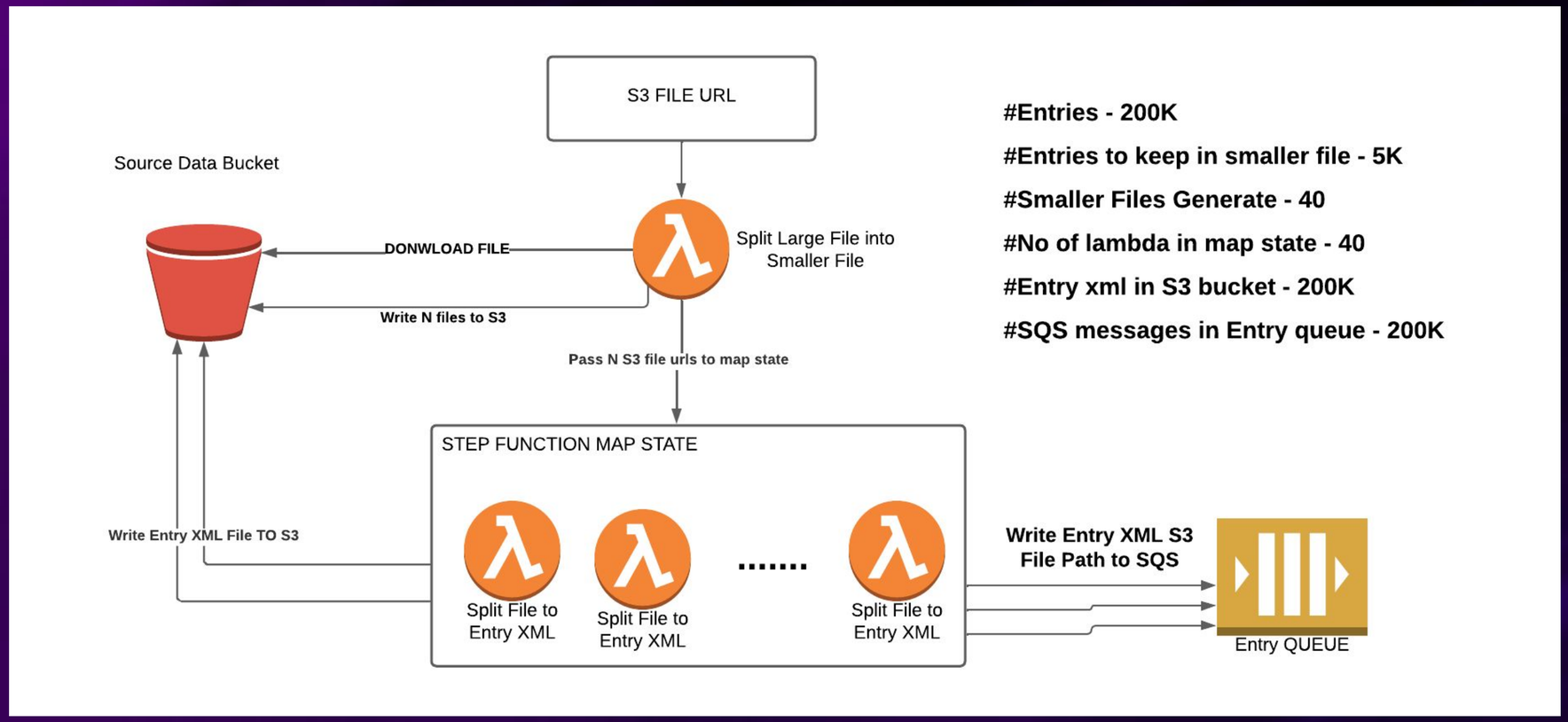
Data Visualization



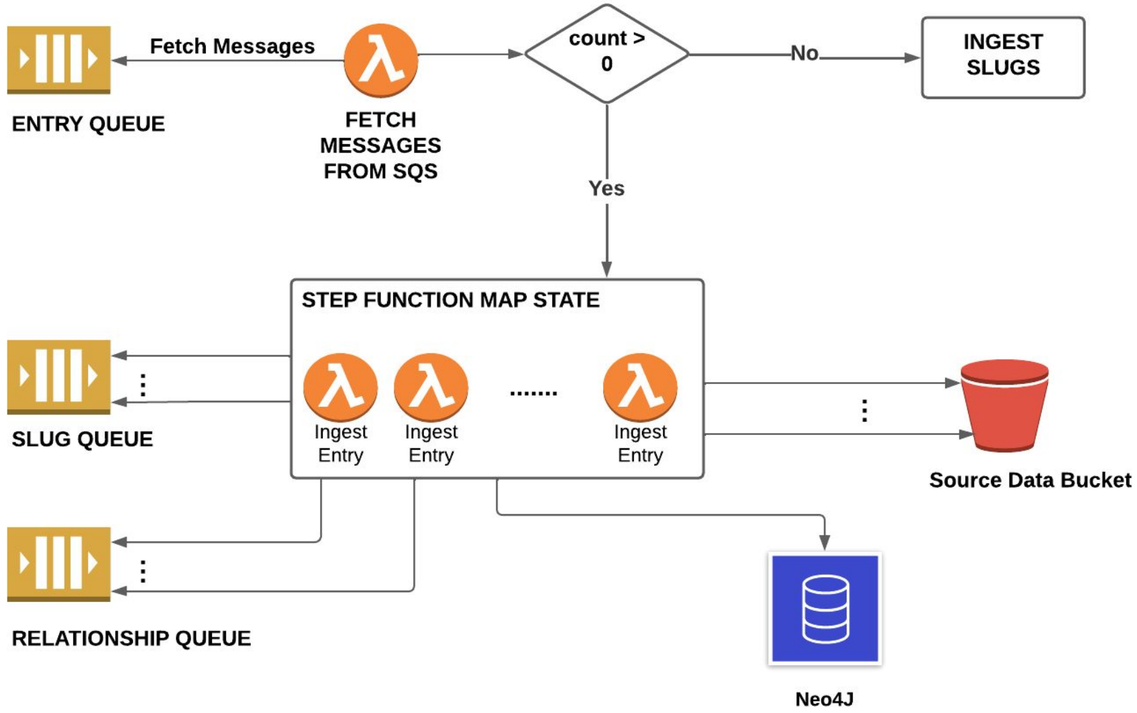
Implementation Steps

- ❑ Split source xml to individual entry xml files
- ❑ Ingest entries
- ❑ Ingest slugs
- ❑ Ingest relationships
- ❑ Generate artifacts

Split Source File



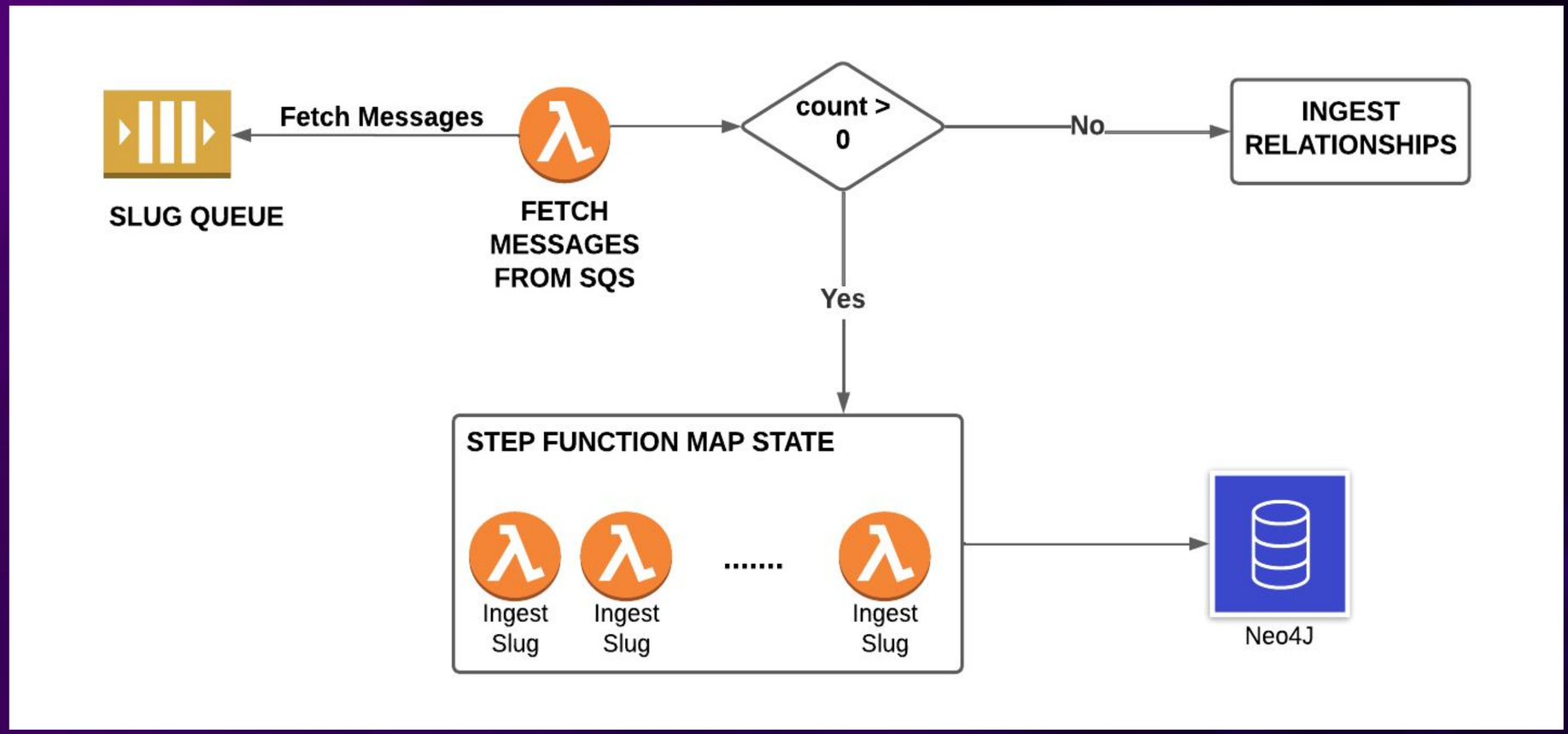
Ingest Entries



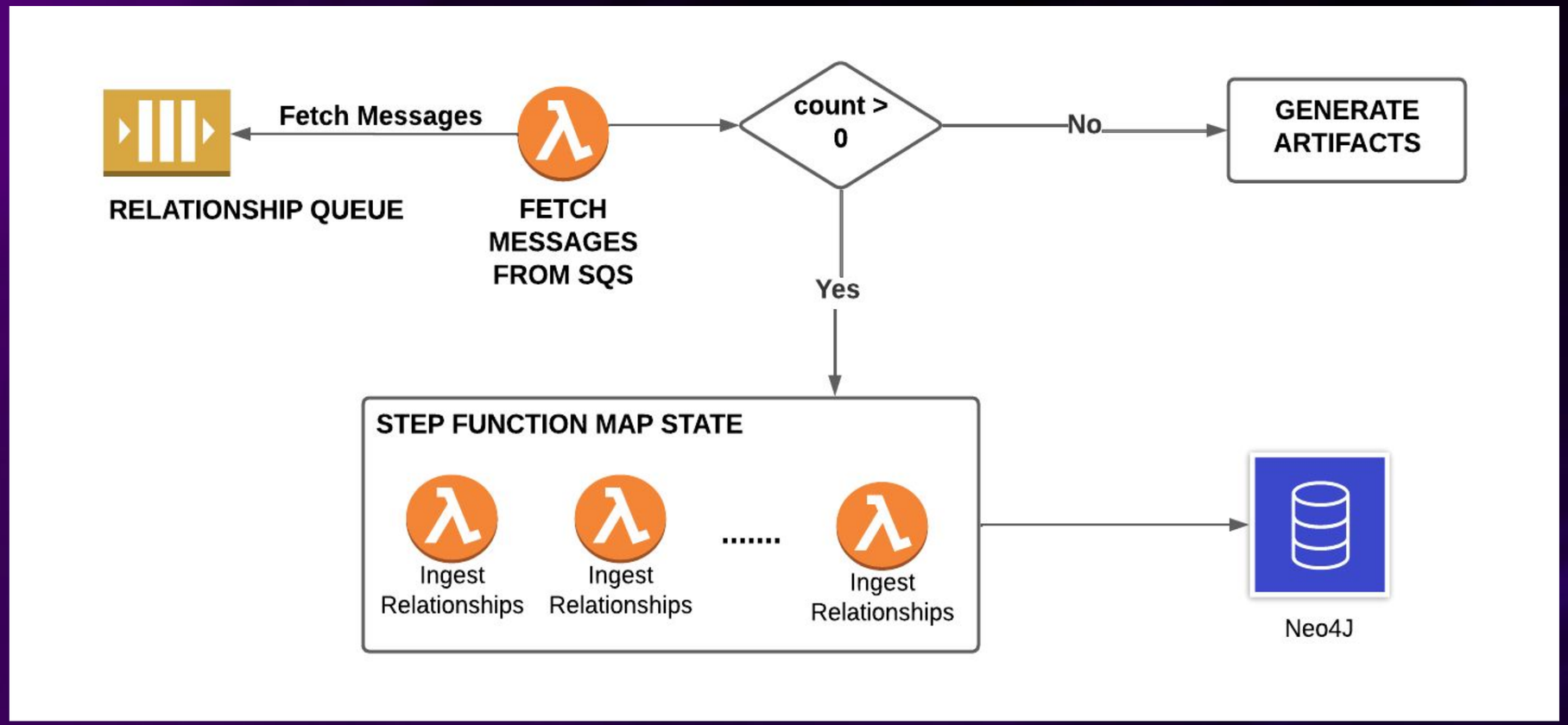
Approach 1 -
1 Receive Message call = max 10 messages
Run 10 lambda in parallel
Ingest 10 entries in one run
200K/ 10 = 20000 loops needed

Approach 2 -
1 Receive Message call = max 10 messages
Fetch 10 * 10 = 100 Messages from SQS
Run 100 lambda in parallel
Ingest 100 entries in one run
200k/100 = 2000 loops needed

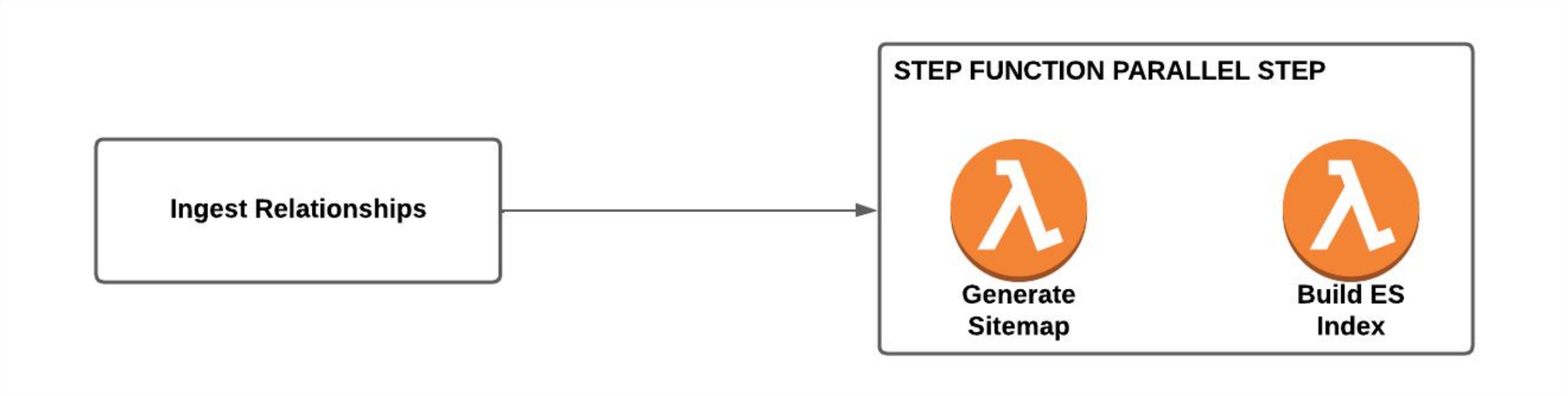
Ingest Slugs



Ingest Relationships



Generate Artifacts



Measuring Success

- Ingestion time reduced from ~24hrs to ~4hrs
- Use of step function provided visual workflow
- Enhanced debugging due to in built observability
- Step function provided ability redrive / retry the ingestion
- Self serving, improved Quality of life (QoL)
- Reduced cost

aws COMMUNITY DAY
PUNE
ANNUAL EDITION 2024

aws USER GROUP PUNE

Lessons Learned

Listen: Know Your Limits

- SQS message size - 256KB
- Step function data transfer limit between states - 256KB
- DB connection pool limit
- Step function maximum events limit in single run - 25000

aws COMMUNITY DAY
PUNE
ANNUAL EDITION 2024

aws USER GROUP PUNE

Learn: Use Batch Processing

- ❏ Fit more records to process in single SQS message
- ❏ Write data to DB in batches
- ❏ Drop your indexes while performing large data writes

aws COMMUNITY DAY PUNE ANNUAL EDITION 2024

aws USER GROUP PUNE

Improve

- Use child workflows to avoid reaching step function events limit of 25000
- Store pointers to the data in SQS instead of actual data

aws
COMMUNITY DAY
PUNE
ANNUAL EDITION 2024

aws
USER GROUP
PUNE

Real World Applications

- Financial report processing
- Customer data processing
- Data migration
- Log processing

aws COMMUNITY DAY
PUNE
ANNUAL EDITION 2024

aws USER GROUP PUNE

aws
COMMUNITY DAY
PUNE
ANNUAL EDITION 2024

aws
USER GROUP
PUNE

Q/A

aws COMMUNITY DAY
PUNE
ANNUAL EDITION 2024

aws USER GROUP PUNE

Connect with us

Mandar Gokhale

https://www.linkedin.com/in/mandar-gokhale-31125733/

@mandarg13

Avinash Dalvi

www.internetkatta.com

avinashdalvi.com

@AvinashDalvi_

@LearnWIthAvinashDalvi