

Analyzing WARC on Serverless Computing

Yinlin Chen

Digital Library Architect & Assistant Professor

ylchen@vt.edu

Virginia Tech Libraries

Web Archiving Conference 2021
15 - 16 Jun 2021

Agenda

- Challenges
- Serverless
- Architecture design
- Experiment setup
- Results
- Conclusion
- Future work

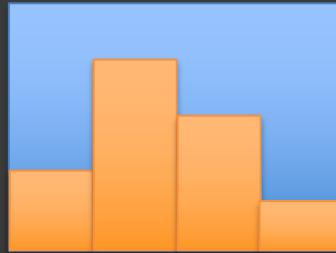
Challenges

- Limited in-house resource to maintain a cluster 24/7
 - Configure servers, monitor tools, logs, etc.
 - Software version upgrade, security patches, etc.
 - Hardware replacement
- Scalability is limited by the compute resource
- Time to process a set of datasets is fixed
- Moving the entire setup to the cloud would be expensive

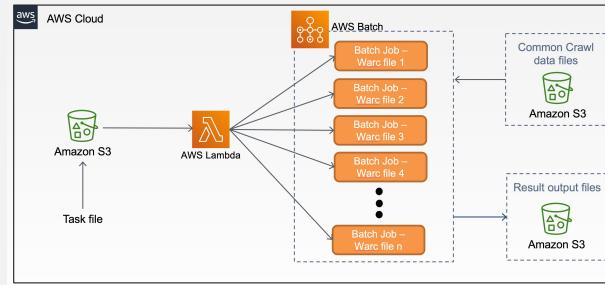
Before



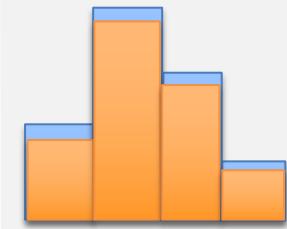
Resource Usage



Now



Resource Usage



- Cluster 24/7
- Handle scalability under a fixed resource
- Maintenance burden

- Ready to receive requests 24/7
- Handle scalability automatically with more flexible resources
- Use the servers, no need to maintain them
- Resource usage is closer to actual usage

Principles of Serverless Design



Design push-based, event-driven pipelines



Write single-purpose, stateless functions



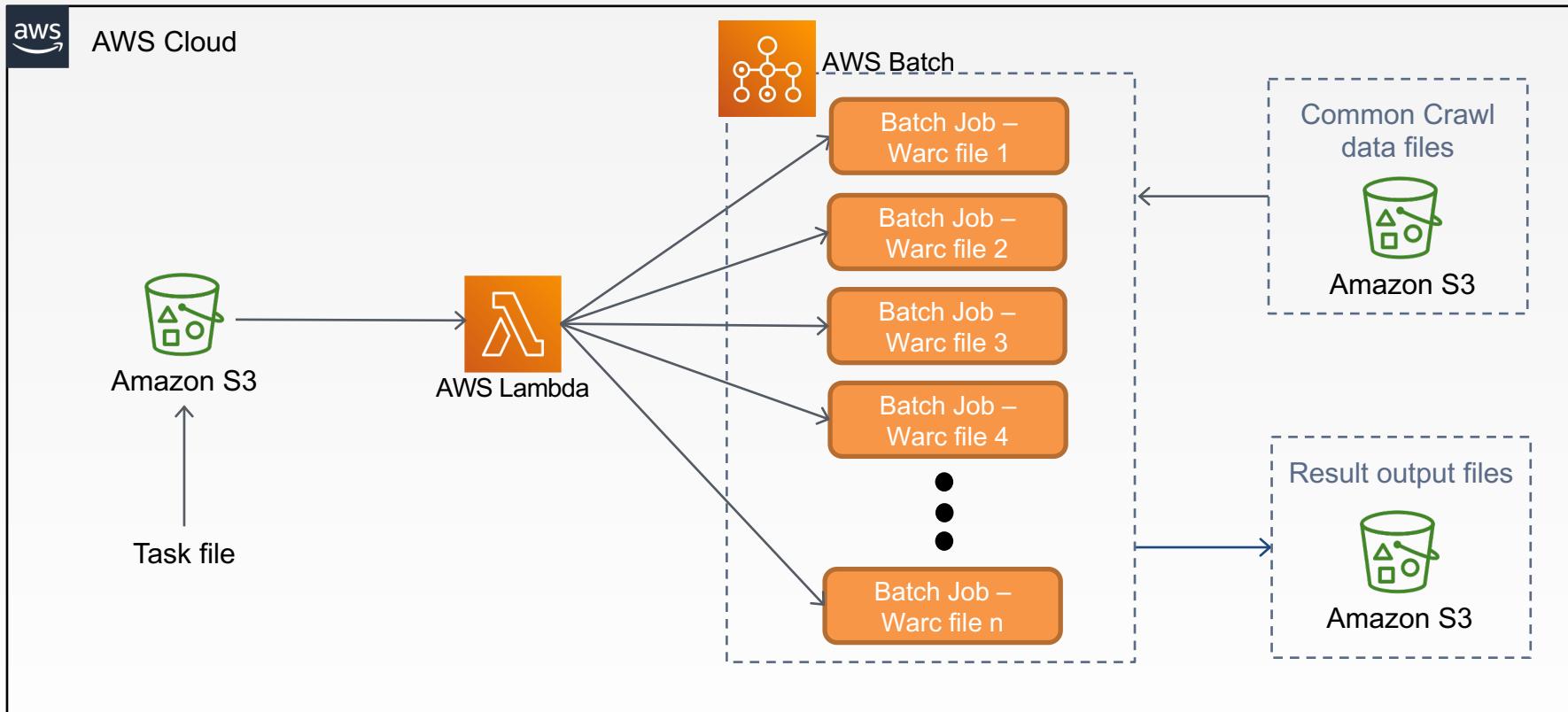
Execute functions on demand and use the exact resources that are needed



Three Pillars of Observability: Logs, Metrics, & Traces

Event-driven Pipeline Workflow

- Task files upload to a S3 bucket
 - S3 URL location of Common Crawl data
- Trigger a Lambda function
- Lambda function reads content in a task file and submits a batch job
- Each batch job runs a container and processes the program workflow
- Results upload to a S3 bucket

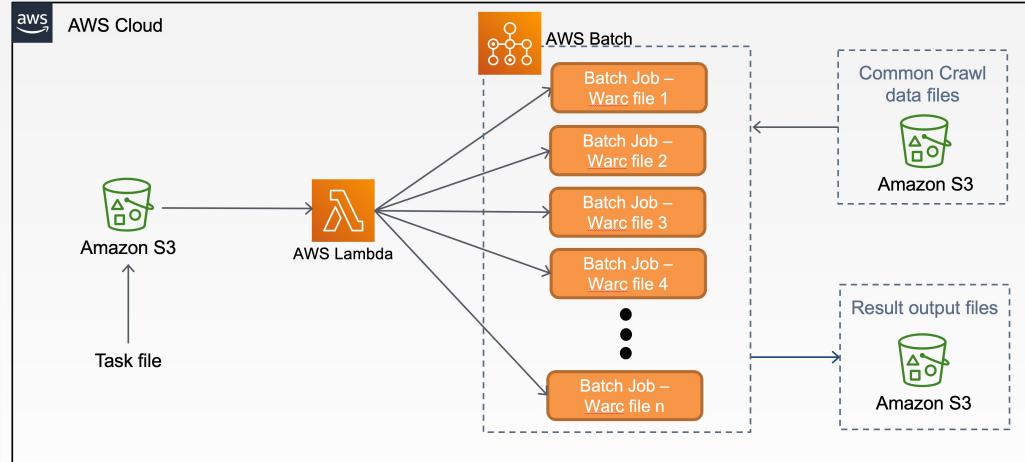


Program Workflow

- Get a file from Common Crawl
- Load a file into ArchiveSpark
- Create CDX
- Extract page with links
- Save all the derivative data
- Record execution times
- Upload results to a S3 bucket

Infrastructure as Code (DevOps)

```
ComputeEnvironment:
  Type: AWS::Batch::ComputeEnvironment
  Properties:
    Type: MANAGED
  ComputeResources:
    Type: SPOT
    MinvCpus: 0
    DesiredvCpus: 0
    MaxvCpus: 256
    InstanceTypes:
      - optimal
    Subnets:
      - Ref: Subnet
    SecurityGroupIds:
      - Ref: SecurityGroup
    ImageId: ami-00487452f317c9be1
    InstanceRole:
      | Ref: IamInstanceProfile
    BidPercentage: 40
    SpotIamFleetRole:
      | Ref: SpotIamFleetRole
    Tags: {"name" : "IIPCSpot"}
  ServiceRole:
    | Ref: BatchServiceRole
BatchProcessS3Bucket:
  Type: AWS::S3::Bucket
  DependsOn: BatchProcessBucketPermission
  Properties:
    BucketName:
      | !Sub '${S3BucketName}-${AWS::AccountId}'
  NotificationConfiguration:
    LambdaConfigurations:
      - Event: 's3:ObjectCreated:*'
```



One click deploy

```
CloudFormation stack changeset
Operation LogicalResourceId ResourceType
+ Modify ConvertFileFunction AWS::Lambda::Function
+ Modify UploadBucket AWS::S3::Bucket

Changeset created successfully. arn:aws:cloudformation:us-east-1:989117335741:changeSet:samcli-deploy1614752401/f3165ad7-122e-4644-bd52-f4b4c8b1c8b1

2021-03-03 01:20:17 - Waiting for stack create/update to complete
CloudFormation events from changeset

ResourceStatus ResourceType LogicalResourceId
UPDATE_IN_PROGRESS AWS::Lambda::Function ConvertFileFunction
UPDATE_COMPLETE AWS::Lambda::Function testhello
*UPDATE_COMPLETE_CLEANUP_IN_PROGRESS AWS::CloudFormation::Stack testhello
UPDATE_COMPLETE AWS::CloudFormation::Stack testhello
```

Experiment Setup

- Run 10, 20, 40, 80, 100, 1000 # of tasks
- Each task processes one Common Crawl file
- Run tasks using
 - On-demand instance
 - Spot instance (up to 40% off the on-demand instance price)
- Record execution time, logs, and cost

Monitor Cluster Status

New ECS Experience
[Tell us what you think](#)

Amazon ECS

Clusters

Task Definitions

Account Settings

Amazon EKS

Clusters

Amazon ECR

Repositories

AWS Marketplace

Discover software

Subscriptions 

Clusters

An Amazon ECS cluster is a regional grouping of one or more container instances on which you can run task requests. Each account receives a default cluster the first time you use the Amazon ECS service. Clusters may contain more than one Amazon EC2 instance type.

For more information, see the [ECS documentation](#).

[Create Cluster](#)

[Get Started](#)

View

 list

 card

[view all](#)



[ComputeEnvironment-0b4be640e8b4f2e_Batch_f3822058-b8a1-39ab-bd8f-ea2b19846e1e >](#)

CloudWatch monitoring

 Default Monitoring

FARGATE

0

Services

0

Running tasks

0

Pending tasks

EC2

0

Services

8

Running tasks

0

Pending tasks

0.32%

CPUUtilization

0.05%

MemoryUtilization

2

Container instances



Managed Compute Environment Automatically

New EC2 Experience Tell us what you think X

EC2 Dashboard New

Events

Tags

Limits

Instances

Instances New

Instance Types

Launch Templates

Spot Requests

Savings Plans

Reserved Instances New

Dedicated Hosts

Scheduled Instances

Capacity Reservations

Instances (10) Info

Filter instances

Instance state: running X Clear filters

	Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Public IPv4 DNS	Public IPv4 ...
<input type="checkbox"/>	T4g	i-020791c8eac31552d	Running	t4g.micro	2/2 checks ...	No alarms +	us-east-1e	ec2-3-84-192-33.com...	3.84.192.33
<input type="checkbox"/>	-	i-0d7d0dedbe4c3ecd	Running	c4.8xlarge	2/2 checks ...	No alarms +	us-east-1e	ec2-34-201-37-17.co...	34.201.37.17
<input type="checkbox"/>	-	i-0b52802169d10465d	Running	c4.8xlarge	2/2 checks ...	No alarms +	us-east-1e	ec2-100-24-18-20.co...	100.24.18.20
<input type="checkbox"/>	-	i-099c0a2915493b22f	Running	c4.8xlarge	2/2 checks ...	No alarms +	us-east-1e	ec2-54-152-43-249.co...	54.152.43.249
<input type="checkbox"/>	-	i-00b78addf23f0d056	Running	c4.xlarge	2/2 checks ...	No alarms +	us-east-1e	ec2-3-84-53-51.comp...	3.84.53.51
<input type="checkbox"/>	-	i-0df7b472422bc3973	Running	c4.2xlarge	2/2 checks ...	No alarms +	us-east-1e	ec2-18-212-60-155.co...	18.212.60.155
<input type="checkbox"/>	-	i-0218294f2764a5b90	Running	c4.4xlarge	2/2 checks ...	No alarms +	us-east-1e	ec2-35-174-172-163.c...	35.174.172.163
<input type="checkbox"/>	-	i-0b8b4a5b0ef0786ce	Running	m4.10xlarge	2/2 checks ...	No alarms +	us-east-1e	ec2-52-202-113-143.c...	52.202.113.143
<input type="checkbox"/>	-	i-099a905538aad1d77	Running	m4.10xlarge	2/2 checks ...	No alarms +	us-east-1e	ec2-54-242-135-140.c...	54.242.135.140
<input type="checkbox"/>	-	i-0be62bc5942c589ae	Running	m4.10xlarge	2/2 checks ...	No alarms +	us-east-1e	ec2-3-95-203-16.com...	3.95.203.16

Monitor Lambda Status

AWS Lambda ×

Updated console (preview) Tell us what you think

Dashboard

- Applications
- Functions

Additional resources

- Code signing configurations
- Layers

Related AWS resources

- Step Functions state machines

The new Lambda console experience becomes permanent on 2021-03-30
You can switch between the new and old experiences by using the toggle on the left sidebar. On 2021-03-30 the toggle will be removed and the new experience will be the only one available.

Create function

Resources for US East (N. Virginia)

Lambda function(s)	Code storage	Full account concurrency	Unreserved account concurrency
43	206.2 MB (0% of 75.0 GB)	1000	1000

Account-level metrics
The charts below show metrics across all your Lambda functions in this AWS Region.

Add to dashboard 1h 3h 12h 1d 3d 1w custom ▾

Error count and success rate (%)	Throttles	Invocations
<p>Count: 1 No unit: 100 0.5 0 Errors Success rate (%)</p>	<p>Count: 1 0.5 0 Throttles</p>	<p>Count: 417 236 54 invocations</p>

Duration	ConcurrentExecutions	UnreservedConcurrentExecutions
<p>Milliseconds: 305 187 69.9 Duration Minimum Duration Average Duration Maximum</p>	<p>Count: 2 1 0 ConcurrentExecutions Maximum</p>	<p>Count: 2 1 0 UnreservedConcurrentExecutions Maximum</p>

Monitor Job Progress

AWS Batch X AWS Batch > Dashboard Last updated: 12:12:23 PM. Auto-refreshes every 60 seconds. 

New Batch experience Dashboard Jobs Job definitions Job queues Compute environments Wizard

Dashboard

Jobs overview

RUNNABLE	RUNNING	SUCCEEDED	FAILED
936	64	260	1

Job queue overview

Job queue	SUBMITTED	PENDING	RUNNABLE	STARTING	RUNNING	SUCCEEDED	FAILED
WARCJobQueue	0	0	936	0	64	260	1

Compute environment overview

Name	Type	Provisioning model	Instance types	Status	State	Minimum vCPUs	Desired vCPUs	Maximum vCPUs
ComputeEnvironment-64dd36f67ae0cd5	MANAGED	EC2	optimal	VALID	ENABLED	0	256	256

Batch Job Logs

AWS Batch		WARJobQueue											
Dashboard	Jobs	Filter resources by property or value										Status: SUCCEEDED	< 1 2 >
		Name	ID	Job type	Array size	Created at	Started at	Stopped at	Total run time	Status			
		iipc-4f9251519ba54ff12ba312160f5c5258f	8dfc946b-2cb0-4cd5-a4a9-368927908e2d	single	--	Mar 22 2021 19:47:07	Mar 22 2021 20:07:29	Mar 22 2021 20:28:49	00:21:19	SUCCEEDED			
		iipc-35fc365316b419eb4db703b79d5b	a78ad6f-e340-4caf-87e0-0dd7e9b3abe3	single	--	Mar 22 2021 19:47:06	Mar 22 2021 20:07:29	Mar 22 2021 20:28:27	00:20:57	SUCCEEDED			
		iipc-abe17239dfe4f2aa6397d77538525a	b6ed92d-647e-431c-8e2b-bbe6e19d28ba	single	--	Mar 22 2021 19:47:05	Mar 22 2021 20:07:28	Mar 22 2021 20:30:12	00:22:43	SUCCEEDED			
		iipc-2f4458d451c4492b52845b5390ef9312	44e439b-78a1-495b-b0ef-5288bdf12b9	single	--	Mar 22 2021 19:47:05	Mar 22 2021 20:06:55	Mar 22 2021 20:26:11	00:19:16	SUCCEEDED			
		iipc-fa6117da5bb448988344186788eb2b99	46a350ef-ff5a-4d98-9135-ce5716cd522	single	--	Mar 22 2021 19:47:04	Mar 22 2021 20:06:56	Mar 22 2021 20:27:32	00:20:35	SUCCEEDED			
		iipc-73befa5fa7ae4103ae4905843a20d511	8760e947-6d78-4a12-ad73-9cbx576ac1a0	single	--	Mar 22 2021 19:47:03	Mar 22 2021 20:06:55	Mar 22 2021 20:26:35	00:19:39	SUCCEEDED			
		iipc-08c5d7c63e694891a3721b613f01304	ec6bf8f51-a612-4b18-a0c-d72c05e829	single	--	Mar 22 2021 19:47:03	Mar 22 2021 20:06:56	Mar 22 2021 20:27:39	00:20:42	SUCCEEDED			
		iipc-e9f324a4834743899fb24a8ed4c2854	62be63d8-38e4-4781-9f52-a277ab2c2ce5	single	--	Mar 22 2021 19:47:02	Mar 22 2021 20:06:56	Mar 22 2021 20:27:33	00:20:36	SUCCEEDED			
		iipc-f5b765f831741109b96821ba739a1b	a68bec9-e496-428e-ac05-8c8bd0a8587e	single	--	Mar 22 2021 19:47:01	Mar 22 2021 20:06:25	Mar 22 2021 20:26:29	00:20:04	SUCCEEDED			
		iipc-534294dfa8d14f048f41dc118c14603	0b96df23-5b0a-4cfe-a87c-94075b23a373	single	--	Mar 22 2021 19:47:01	Mar 22 2021 20:06:25	Mar 22 2021 20:26:06	00:19:41	SUCCEEDED			
		iipc-541d71a5434644f29b6456a6c18c35c9	6486ceda-1ead-4371-ac38-8431ed0e57a	single	--	Mar 22 2021 19:47:00	Mar 22 2021 19:59:04	Mar 22 2021 20:25:25	00:26:20	SUCCEEDED			
		iipc-4e410616af04fdcaab7701603d63689ce	f689aaec-ec24-449c-a05b-96a0c0f3692	single	--	Mar 22 2021 19:46:59	Mar 22 2021 20:06:25	Mar 22 2021 20:26:37	00:20:12	SUCCEEDED			
		iipc-9cae3f6cc096449c858f1716c2-7d59ca	91e9580f-f6e6-4fe2-9c3c-249953a1ff35	single	--	Mar 22 2021 19:46:58	Mar 22 2021 19:58:31	Mar 22 2021 20:24:56	00:26:24	SUCCEEDED			
		iipc-760c010f196f4f6b4d426aea575dc257a	5a894d1c-8051-45ac-893d-7106684c85e6	single	--	Mar 22 2021 19:46:58	Mar 22 2021 19:58:32	Mar 22 2021 20:24:21	00:25:48	SUCCEEDED			
		iipc-a67a9618313a411c94a61dc08254e5a1	20995ade-d8ac-4b81-a89d-ecf2294a1c112	single	--	Mar 22 2021 19:46:57	Mar 22 2021 19:58:31	Mar 22 2021 20:26:01	00:27:30	SUCCEEDED			
		iipc-d9b5939dee08a4de792223505f691ead	f963d61f-9429-43ae-a2bc-ae6b9668f15a	single	--	Mar 22 2021 19:46:56	Mar 22 2021 19:59:04	Mar 22 2021 20:28:16	00:29:11	SUCCEEDED			
		iipc-7879134280f4a42ac18cef91089980	b7e4ad2-05e1-409b-97a1-db7d1bf86568	single	--	Mar 22 2021 19:46:56	Mar 22 2021 19:58:33	Mar 22 2021 20:24:48	00:26:14	SUCCEEDED			
		iipc-a71c101e044cb0c906379504495fb5d	8a9f9b92-79b4-4a48-91e0-094734ccbd0e	single	--	Mar 22 2021 19:46:55	Mar 22 2021 19:59:04	Mar 22 2021 20:25:52	00:26:47	SUCCEEDED			
		iipc-cd22za50d9a74fb9489abe8ff7b8e74	fa69b7e70-6c48-4070-834e-eabd5ddba5b1	single	--	Mar 22 2021 19:46:54	Mar 22 2021 19:58:32	Mar 22 2021 20:24:52	00:26:20	SUCCEEDED			
		iipc-2696fd30fa894a239601b806ccb3364	e0e53324-c41d-4124-aed-885f192d57d	single	--	Mar 22 2021 19:46:54	Mar 22 2021 19:58:31	Mar 22 2021 20:07:37	00:09:05	SUCCEEDED			
		iipc-e420a016b3f94d43b0da4bb097b2648	f61db6ea-6d6d-4fe6-9f7b-abe3b91f1ae7	single	--	Mar 22 2021 19:46:53	Mar 22 2021 19:58:31	Mar 22 2021 20:07:37	00:09:05	SUCCEEDED			
		iipc-01841d62727e4c48aae8853a0544a57c	d74b2aa2-6054-4168-8a2a-25460439b0c4	single	--	Mar 22 2021 19:46:52	Mar 22 2021 19:58:31	Mar 22 2021 20:24:21	00:25:49	SUCCEEDED			
		iipc-e731df62ad3f4d97a4ab2d9c28c89ad	6b53d0b1-e9b5-4493-b843-cf433745af5d	single	--	Mar 22 2021 19:46:51	Mar 22 2021 19:58:32	Mar 22 2021 20:24:33	00:26:00	SUCCEEDED			
		iipc-41d5097305394b949636c741f20880d	00bc0a16-8d4d-4c6f-ba17-6d813666c8632	single	--	Mar 22 2021 19:46:51	Mar 22 2021 19:58:32	Mar 22 2021 20:25:42	00:27:10	SUCCEEDED			
		iipc-942a18be3a040b54a933893841a7a38	8ef06fb-e6c0-4ccb-ac8d-f206ea1dc751	single	--	Mar 22 2021 19:46:50	Mar 22 2021 19:58:31	Mar 22 2021 20:24:48	00:26:16	SUCCEEDED			
		iipc-ac7a646db43417b9f0258123505544	4fd277d1-9b9d-4f4c-abfd-9a6eb714e8	single	--	Mar 22 2021 19:46:49	Mar 22 2021 19:58:32	Mar 22 2021 20:07:37	00:09:05	SUCCEEDED			
		iipc-4195be6b5ac04eb9b0f5040a3ef5c44	6c45e53c-905c-4466-9604-eb03d5e98873	single	--	Mar 22 2021 19:46:49	Mar 22 2021 19:58:32	Mar 22 2021 20:30:00	00:31:28	SUCCEEDED			

Batch Job Logs

CloudWatch ×

- ▶ Dashboards
- ▶ Alarms In alarm 0
- ▶ Insufficient data 0
- ▶ OK 0
- ▶ Billing
- ▶ Logs Log groups
 - ▶ Insights
- ▶ Metrics New
- ▶ Events New
 - ▶ Rules
 - ▶ Event Buses
- ▶ ServiceLens New
 - ▶ Service Map
 - ▶ Traces
- ▶ Container Insights New
 - ▶ Resources
 - ▶ Performance monitoring
- ▶ Lambda Insights New
 - ▶ Performance monitoring
- ▶ Synthetics

- ▶ 2021-03-30T19:48:08.462-04:00 duration = 753.365644969
- ▶ 2021-03-30T19:48:08.548-04:00 summary: String = CDX:106.264873529|PagesWithLinks:753.365644969
- ▶ 2021-03-30T19:48:08.639-04:00 pw: java.io.PrintWriter = java.io.PrintWriter@4d5a6b60
- ▶ 2021-03-30T19:48:08.853-04:00 21/03/30 23:48:08 [WARN] o.a.t.k.p.v.s.KernelOutputStream - Suppressing empty output: ''
- ▶ 2021-03-30T19:48:08.877-04:00 Unauthorized system.exit detected!
- ▶ 2021-03-30T19:48:08.877-04:00 21/03/30 23:48:08 [INFO] o.a.t.Mains\$anon\$1 - Shutting down kernel
- ▶ 2021-03-30T19:48:08.877-04:00 21/03/30 23:48:08 [INFO] o.a.t.Mains\$anon\$1 - Shutting down interpreters
- ▶ 2021-03-30T19:48:08.878-04:00 21/03/30 23:48:08 [INFO] o.a.t.k.i.s.ScalaInterpreter - Shutting down interpreter
- ▶ 2021-03-30T19:48:08.880-04:00 21/03/30 23:48:08 [INFO] o.a.t.Mains\$anon\$1 - Shutting down actor system
- ▶ 2021-03-30T19:48:08.930-04:00 21/03/30 23:48:08 [INFO] o.a.t.Mains\$anon\$1 - Shutting down interpreters
- ▶ 2021-03-30T19:48:08.930-04:00 21/03/30 23:48:08 [INFO] o.a.t.k.i.s.ScalaInterpreter - Shutting down interpreter
- ▶ 2021-03-30T19:48:09.091-04:00 [NbConvertApp] Writing 1823 bytes to CC-MAIN-20200524210325-20200525000325-00020_output.ipynb
- ▶ 2021-03-30T19:48:09.510-04:00 Completed 1.8 KiB/134.6 MiB (54.1 KiB/s) with 10 file(s) remainingupload: ../results/CC-MAIN-20200524210325-20200525000325-00020/CC-MAIN-20200524210325-20200525000325-00020_output...
- ▶ 2021-03-30T19:48:09.530-04:00 Completed 1.8 KiB/134.6 MiB (54.1 KiB/s) with 9 file(s) remainingCompleted 1.8 KiB/134.6 MiB (34.8 KiB/s) with 9 file(s) remainingupload: ../results/CC-MAIN-20200524210325-20200525000325-00020/cdx.gz/_complete to s3://warcresults/CC-MAIN-202005...
- ▶ 2021-03-30T19:48:09.536-04:00 Completed 1.8 KiB/134.6 MiB (34.8 KiB/s) with 8 file(s) remainingupload: ../results/CC-MAIN-20200524210325-20200525000325-00020/cdx.gz/_complete to s3://warcresults/CC-MAIN-202005...
- ▶ 2021-03-30T19:48:09.538-04:00 Completed 1.8 KiB/134.6 MiB (34.8 KiB/s) with 7 file(s) remainingCompleted 1.8 KiB/134.6 MiB (30.4 KiB/s) with 7 file(s) remainingupload: ../results/CC-MAIN-20200524210325-202005...
- ▶ 2021-03-30T19:48:09.540-04:00 Completed 1.8 KiB/134.6 MiB (30.4 KiB/s) with 6 file(s) remainingCompleted 1.8 KiB/134.6 MiB (29.5 KiB/s) with 6 file(s) remainingupload: ../results/CC-MAIN-20200524210325-202005...
- ▶ 2021-03-30T19:48:09.541-04:00 Completed 1.8 KiB/134.6 MiB (29.5 KiB/s) with 5 file(s) remainingupload: ../results/CC-MAIN-20200524210325-20200525000325-00020/pages-with-links.gz/_complete to s3://warcresults/C...
- ▶ 2021-03-30T19:48:09.572-04:00 Completed 1.8 KiB/134.6 MiB (29.5 KiB/s) with 4 file(s) remainingCompleted 257.8 KiB/134.6 MiB (3.5 MiB/s) with 4 file(s) remainingCompleted 513.8 KiB/134.6 MiB (6.5 MiB/s) with 4...
- ▶ 2021-03-30T19:48:09.630-04:00 Completed 1.0 MiB/134.6 MiB (11.0 MiB/s) with 3 file(s) remainingCompleted 1.3 MiB/134.6 MiB (9.7 MiB/s) with 3 file(s) remainingCompleted 1.5 MiB/134.6 MiB (11.5 MiB/s) with 3 f...
- ▶ 2021-03-30T19:48:09.979-04:00 Completed 3.5 MiB/134.6 MiB (23.4 MiB/s) with 2 file(s) remainingCompleted 3.8 MiB/134.6 MiB (24.6 MiB/s) with 2 file(s) remainingCompleted 4.0 MiB/134.6 MiB (25.7 MiB/s) with 2 f...
- ▶ 2021-03-30T19:48:14.224-04:00 Completed 49.4 MiB/134.6 MiB (98.7 MiB/s) with 1 file(s) remainingCompleted 49.7 MiB/134.6 MiB (98.3 MiB/s) with 1 file(s) remainingCompleted 49.9 MiB/134.6 MiB (98.7 MiB/s) with ...
- ▶ 2021-03-30T19:48:14.279-04:00 Done.
- ▶ 2021-03-30T19:48:14.279-04:00 Task Done.

Query Log Results

AWS Services ▾

CloudWatch X

CloudWatch > CloudWatch Logs > Logs Insights

Select log group(s)

/aws/batch/job Clear

```
1 fields @timestamp, @message
2 | parse @message "summary: String = *" as summary
3 | filter summary like "PagesWithLinks"
4 | sort @timestamp desc
```

Run query Save History

Logs Visualization

Showing 998 of 998 records matched ⓘ
20,284,936 records (2.2 GB) scanned in 11.0s @ 1,839,069 records/s (200.4 MB/s)

Export results ▾ Add to dashboard

Hide histogram

@timestamp @message summary

#	@timestamp	@message	summary
▶ 1	2021-03-25T01:39:17...	summary: String = CDX:125.669154332 PagesWithLinks:897.522650...	CDX:125.669154332 PagesWithLinks:897.522650781
▶ 2	2021-03-25T01:37:48...	summary: String = CDX:125.425643731 PagesWithLinks:874.46451...	CDX:125.425643731 PagesWithLinks:874.464511474
▶ 3	2021-03-25T01:37:21...	summary: String = CDX:125.911768471 PagesWithLinks:904.810172...	CDX:125.911768471 PagesWithLinks:904.81017273
▶ 4	2021-03-25T01:36:22...	summary: String = CDX:126.699601735 PagesWithLinks:887.192927...	CDX:126.699601735 PagesWithLinks:887.192927855
▶ 5	2021-03-25T01:36:21...	summary: String = CDX:124.922903359 PagesWithLinks:878.758452...	CDX:124.922903359 PagesWithLinks:878.758452552
▶ 6	2021-03-25T01:36:08...	summary: String = CDX:107.011999682 PagesWithLinks:737.620613...	CDX:107.011999682 PagesWithLinks:737.620613719
▶ 7	2021-03-25T01:35:17...	summary: String = CDX:102.57925977 PagesWithLinks:726.04894947	CDX:102.57925977 PagesWithLinks:726.04894947
▶ 8	2021-03-25T01:35:10...	summary: String = CDX:107.412546462 PagesWithLinks:752.619504...	CDX:107.412546462 PagesWithLinks:752.619504649
▶ 9	2021-03-25T01:34:59...	summary: String = CDX:108.659221843 PagesWithLinks:744.074575...	CDX:108.659221843 PagesWithLinks:744.074575421
▶ 10	2021-03-25T01:34:52...	summary: String = CDX:127.018653886 PagesWithLinks:884.161975...	CDX:127.018653886 PagesWithLinks:884.161975368
▶ 11	2021-03-25T01:34:18...	summary: String = CDX:125.825652285 PagesWithLinks:900.629645...	CDX:125.825652285 PagesWithLinks:900.629645172
▶ 12	2021-03-25T01:34:14...	summary: String = CDX:108.297303066 PagesWithLinks:740.564994...	CDX:108.297303066 PagesWithLinks:740.564994068

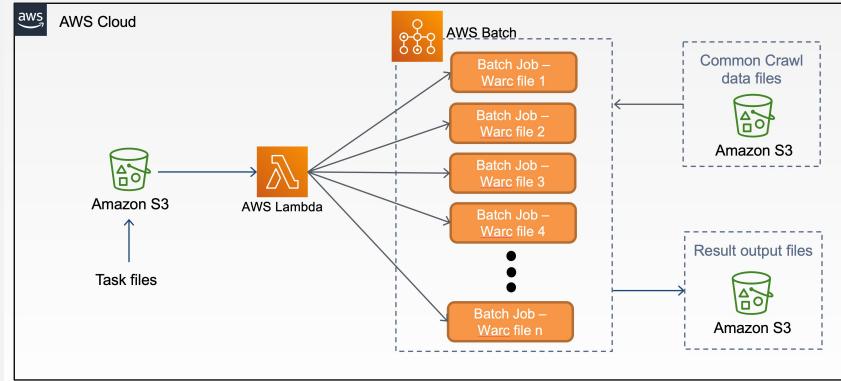
Results

- Avg. execution time for processing a WARC file is 23.3 mins
- Execution times for processing 10, 20, 40, 80, 100 # of WARC files are under 30 mins using on-demand instance
 - 1000 tasks took about 6 hours
- Each job (1G file) cost:
 - On-demand instance: \$0.066
 - Spot instance: \$0.022
- Success rate:
 - On-demand instance: 99%
 - Spot instance: depends (need to run more experiments)

Cost: Before



Cost: Now



- Add one server, plus the cost
- Plan server capacity each year
 - Overestimation
 - Underestimation

- Pay what we use
 - Use it well and pay less
 - Pay 0 when services are idle
- Some (invisible) costs vanish
 - Server maintenance cost
 - Labor, time, electric, etc.

Conclusion

- This serverless architecture design can process almost any number of tasks
- Use only the resources that are needed
- Eliminate the need to manage underlying servers
- Event-driven pipeline automates the entire workflow
- A suite of services for three pillars of observability: logs, metrics, and traces
- On-demand instance is more stable than spot instance
 - Need more experiments to verify
 - Spot instance cost is cheaper

Future Work

- Extend WARC processing program for more complex tasks
- Batch resource (compute environment) tuning
 - CPU and Memory
 - Container
- Spot instance usage tuning
- Apply to other use cases using this serverless pipeline

Q & A

Thank You!

This research work was supported by
AWS Educate program