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

VIRGINIA TECH

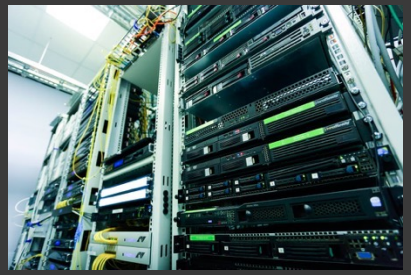
Agenda

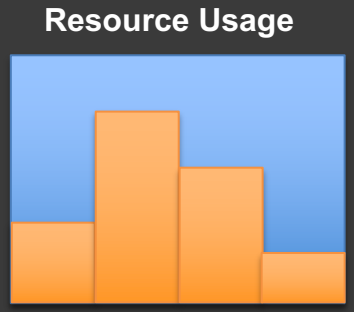
- Challenges
- Serverless
- Architecture design
- Experiment setup
- Results
- Conclusion
- Future work

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

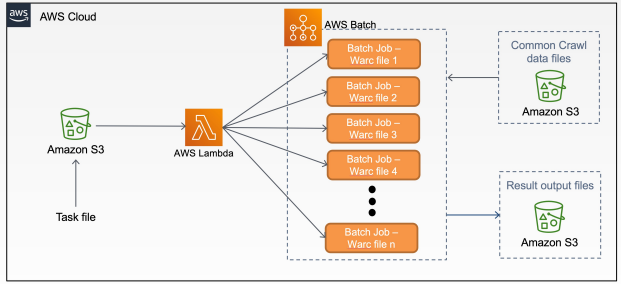
Before

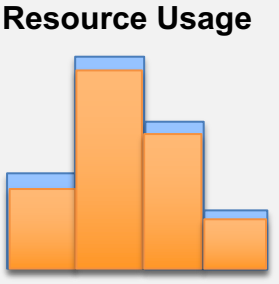




- Cluster 24/7
- Handle scalability under a fixed resource
- Maintenance burden

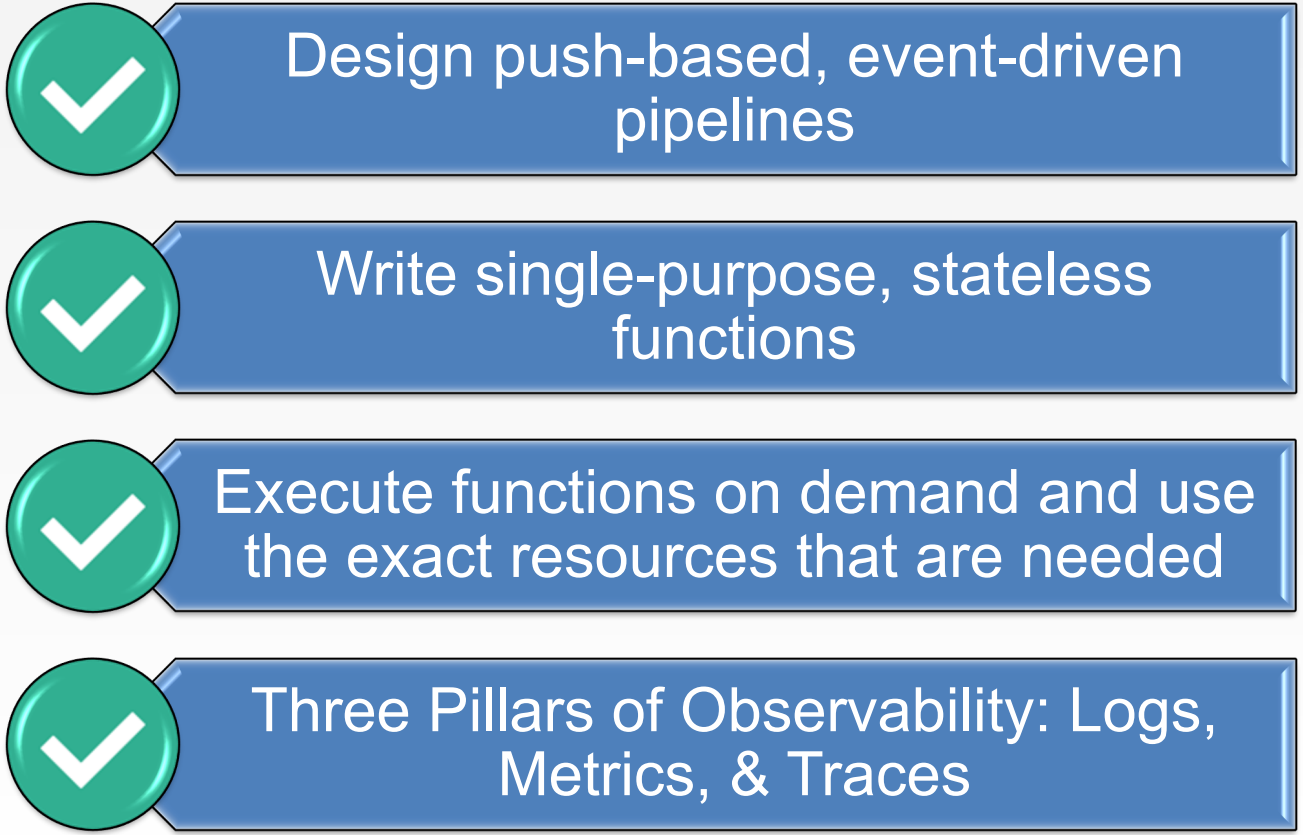
Now





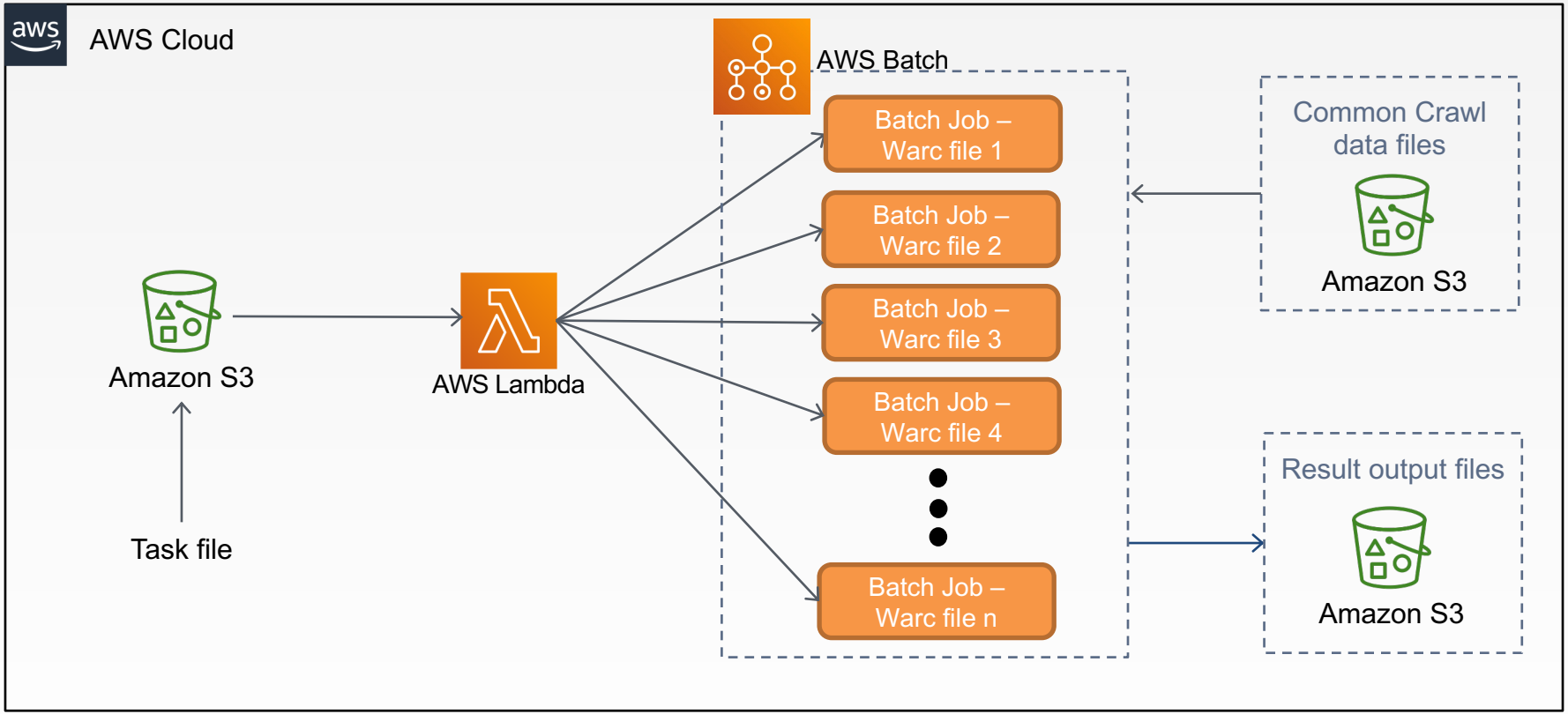
- Ready to receive requests 24/7
- Handle scalability automatically with more flexible resources
- Use the servers, no need to maintain them
- Resource usage is closer to actual usage

Principles of Serverless Design



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

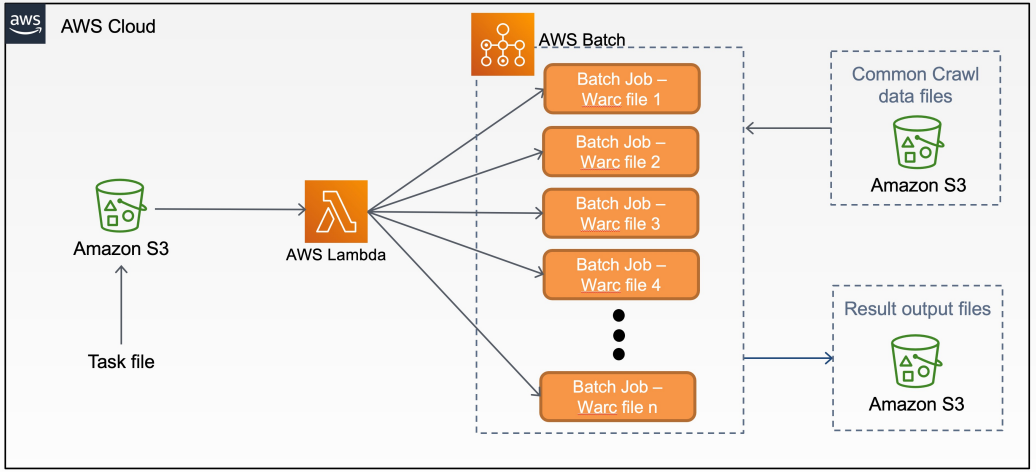

AWS Cloud
AWS Batch
Batch Job – Warc file 1
Batch Job – Warc file 2
Batch Job – Warc file 3
Batch Job – Warc file 4
Batch Job – Warc file n
Common Crawl data files
Amazon S3
Amazon S3
AWS Lambda
Amazon S3
Task file
Result output files
Amazon S3

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:*'
```



Launch Stack

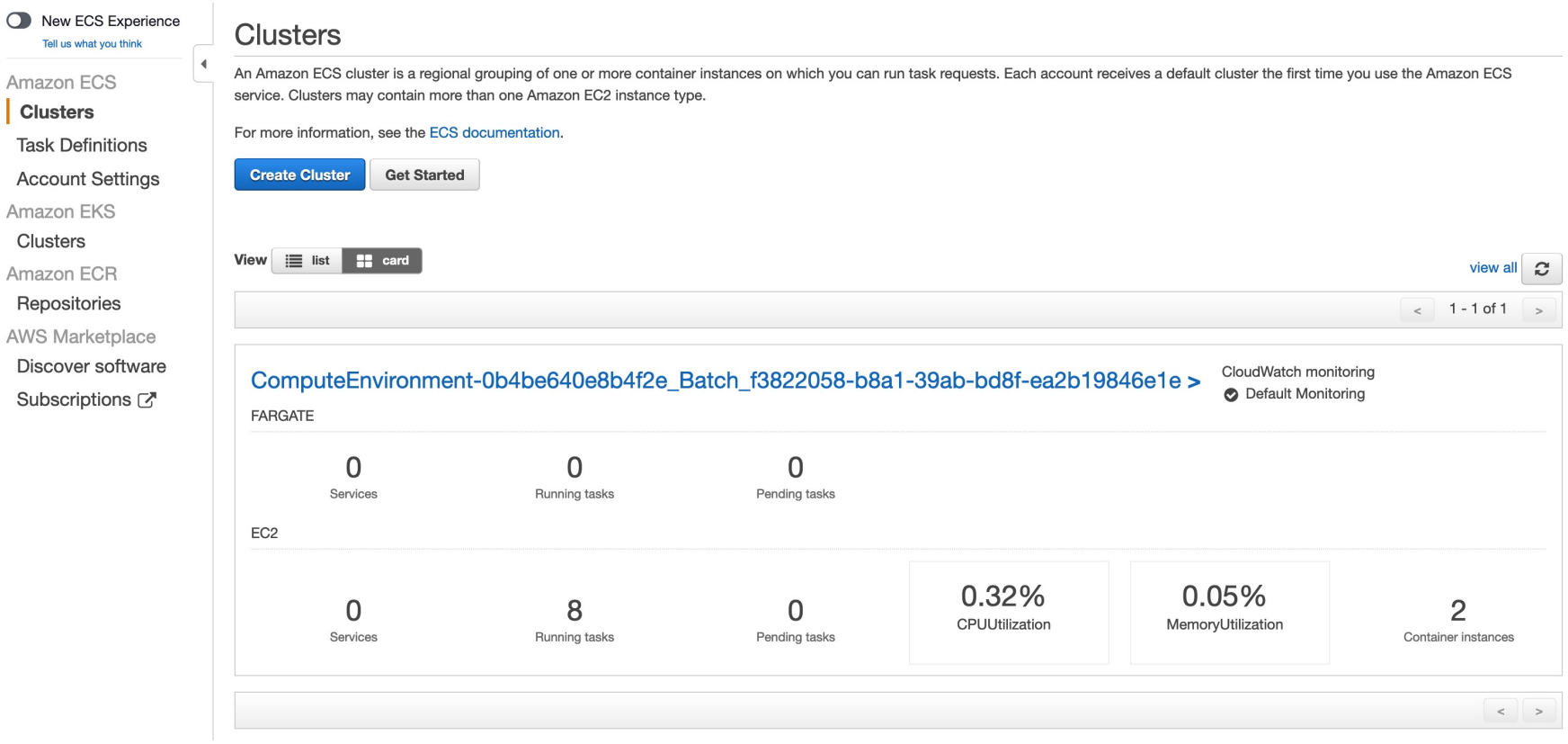
One click deploy

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

Changeset created successfully. arn:aws:cloudformation:us-east-1:909117335741: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       ConvertFileFunction
^[UPDATE_COMPLETE_CLEANUP_IN_PROGRESS  AWS::CloudFormation::Stack  testhello
UPDATE_COMPLETE               AWS::CloudFormation::Stack  testhello
```

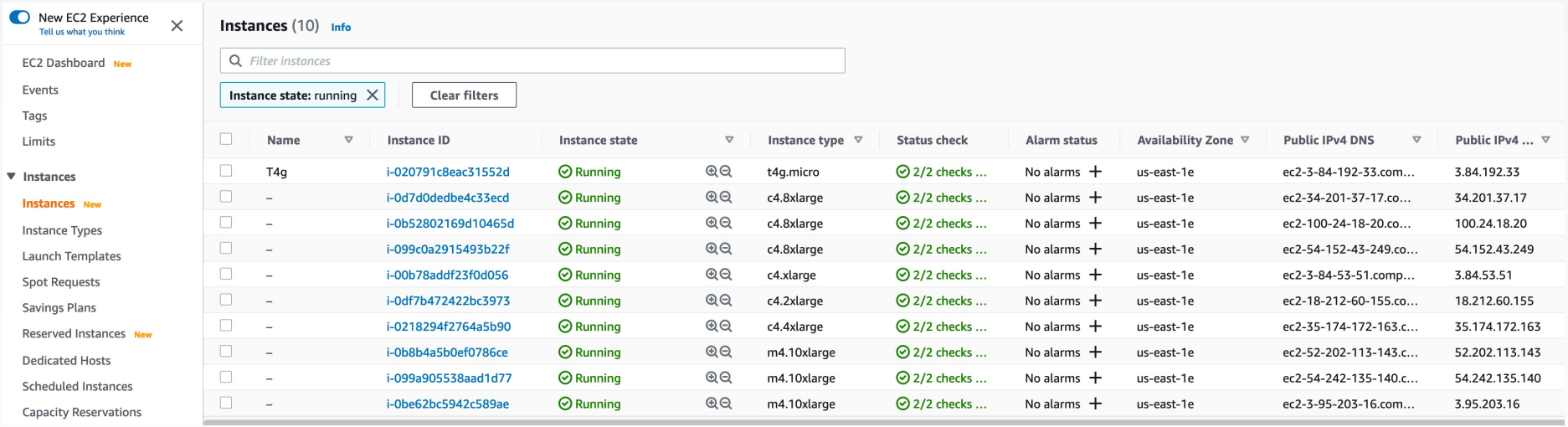
Experiment Setup

- 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

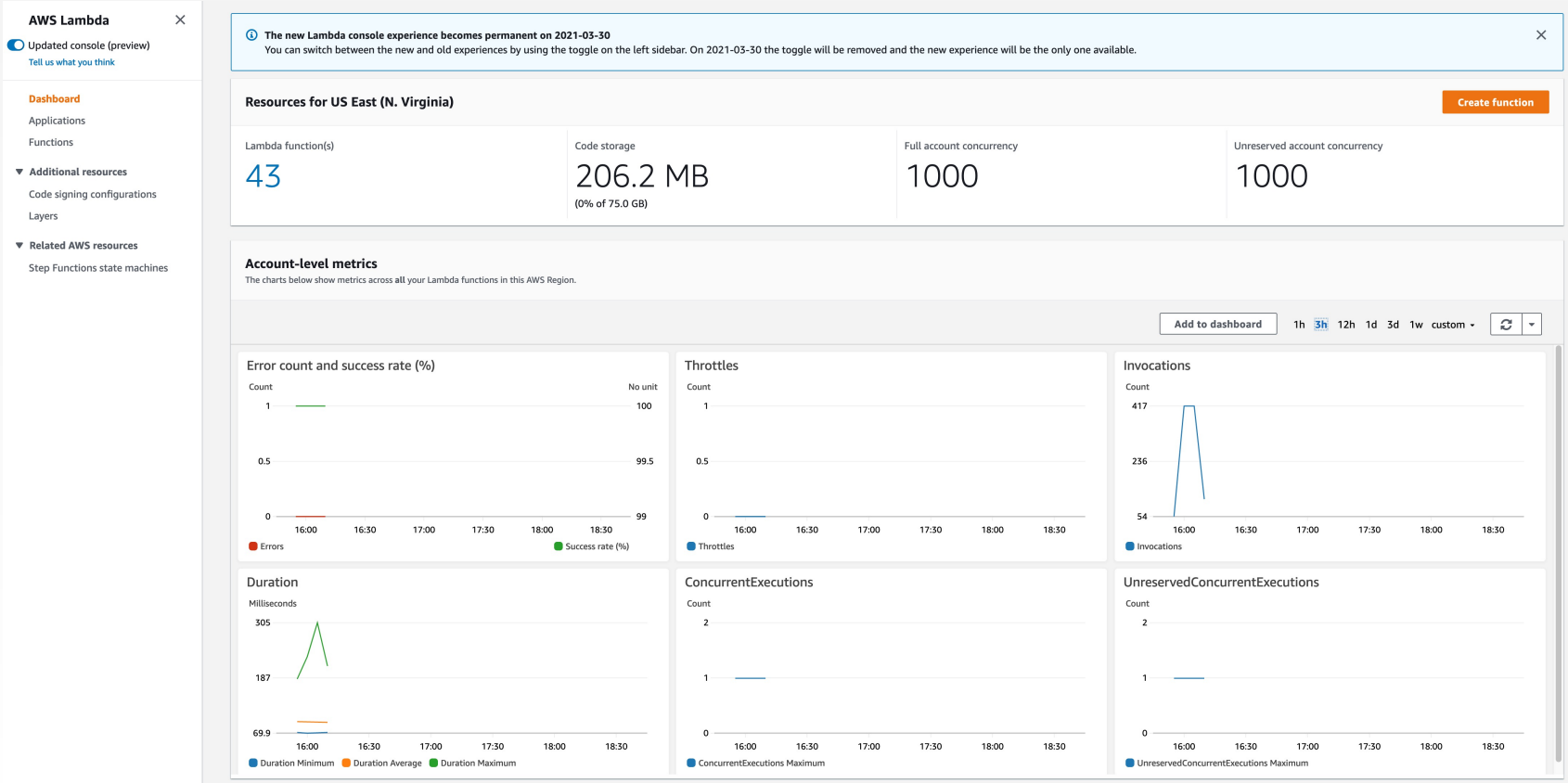


Managed Compute Environment Automatically



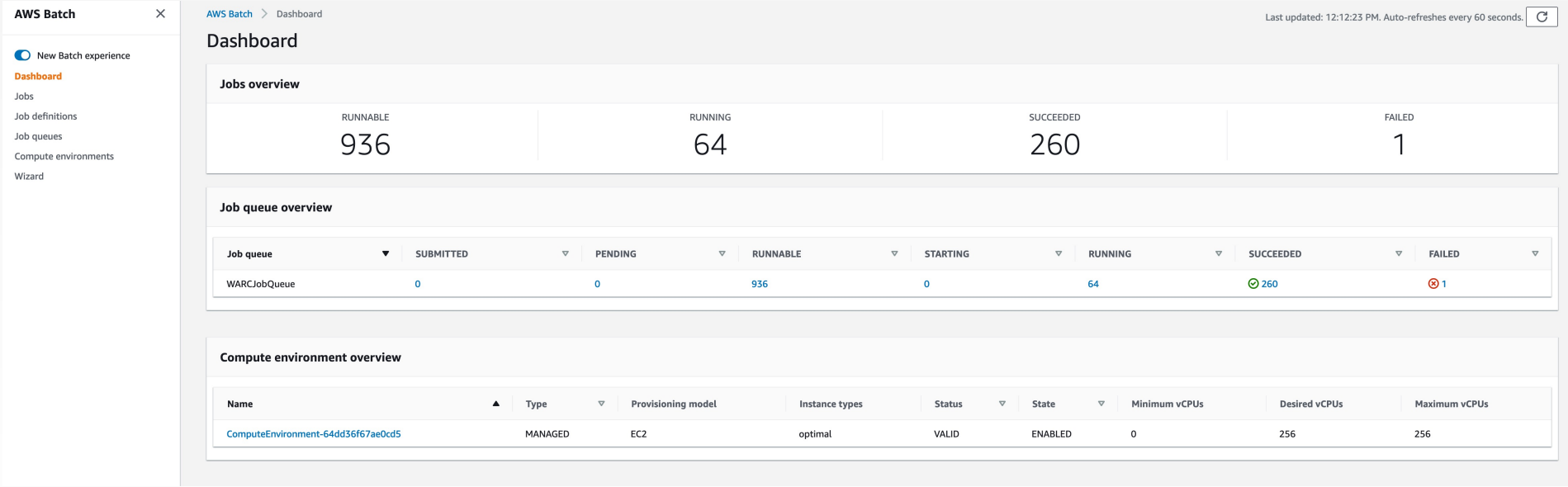
VIRGINIA TECH

Monitor Lambda Status



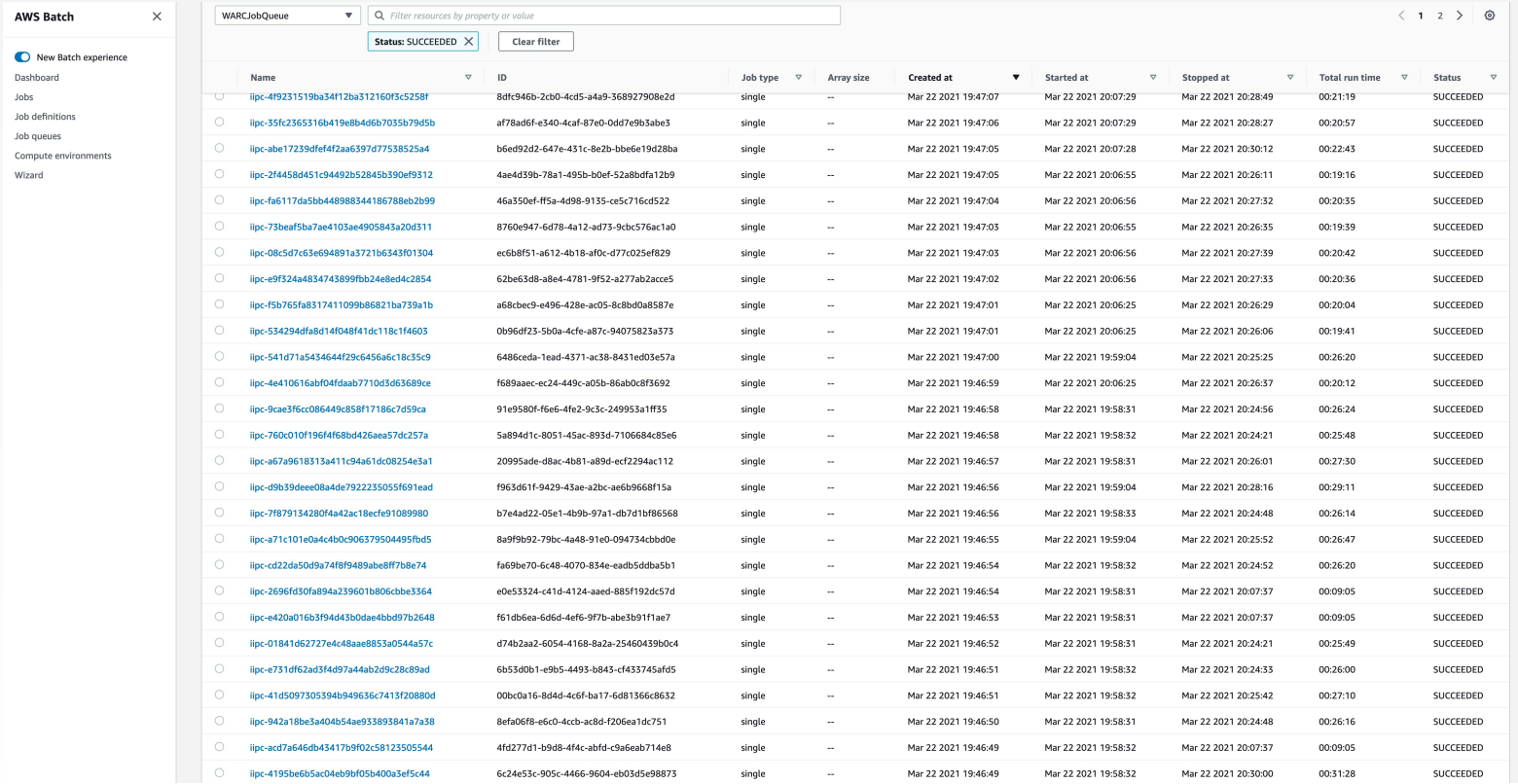
VIRGINIA TECH

Monitor Job Progress



VIRGINIA TECH

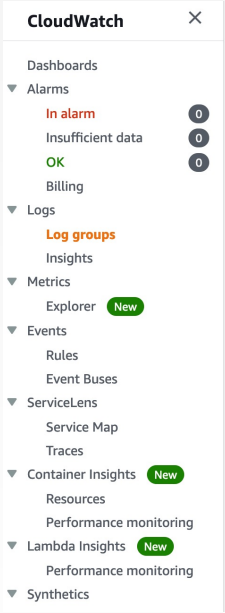
Batch Job Logs



Name	ID	Job type	Array size	Created at	Started at	Stopped at	Total run time	Status
iipc-4f9231519ba34f12ba312160f3c5258f	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-35fc2365316b419e8b4d6b7035b79d5b	af78ad6f-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-abe17239dfef4f2aa6397d77538525a4	b6ed92d2-647e-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-2f4458d451c94492b52845b390ef9312	4ae4d39b-78a1-495b-b0ef-52a8bdfa12b9	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-ce5c716cd522	single	--	Mar 22 2021 19:47:04	Mar 22 2021 20:06:56	Mar 22 2021 20:27:32	00:20:35	SUCCEEDED
iipc-73beaf5ba7ae4103ae4905843a20d311	8760e947-6d78-4a12-ad73-9cbc576ac1a0	single	--	Mar 22 2021 19:47:03	Mar 22 2021 20:06:55	Mar 22 2021 20:26:35	00:19:39	SUCCEEDED
iipc-08c5d7c63e694891a3721b6343f01304	ec6b8f51-a612-4b18-af0c-d77c025ef829	single	--	Mar 22 2021 19:47:03	Mar 22 2021 20:06:56	Mar 22 2021 20:27:39	00:20:42	SUCCEEDED
iipc-e9f324a48347438 99fbb24e8ed4c2854	62be63d8-a8e4-4781-9f52-a277ab2acce5	single	--	Mar 22 2021 19:47:02	Mar 22 2021 20:06:56	Mar 22 2021 20:27:33	00:20:36	SUCCEEDED
iipc-f5b765fa8317411099b86821ba739a1b	a68cbec9-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-534294dfa8d14f048f41dc118c1f4603	0b96df23-5b0a-4cfe-a87c-94075823a373	single	--	Mar 22 2021 19:47:01	Mar 22 2021 20:06:25	Mar 22 2021 20:26:06	00:19:41	SUCCEEDED
iipc-541d71a5434644f29c6456a6c18c35c9	6486ceda-1ead-4371-ac38-8431ed03e57a	single	--	Mar 22 2021 19:47:00	Mar 22 2021 19:59:04	Mar 22 2021 20:25:25	00:26:20	SUCCEEDED
iipc-4e410616abf04fdaab7710d3d63689ce	f689aaec-ec24-449c-a05b-86ab0c8f3692	single	--	Mar 22 2021 19:46:59	Mar 22 2021 20:06:25	Mar 22 2021 20:26:37	00:20:12	SUCCEEDED
iipc-9cae3f6cc086449c858f17186c7d59ca	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-760c010f196f4f68bd426aea57dc257a	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-a67a9618313a411c94a61dc08254e3a1	20995ade-d8ac-4b81-a89d-ecf2294ac112	single	--	Mar 22 2021 19:46:57	Mar 22 2021 19:58:31	Mar 22 2021 20:26:01	00:27:30	SUCCEEDED
iipc-d9b39deee08a4de7922235055f691ead	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-7f879134280f4a42ac18ecfe91089980	b7e4ad22-05e1-4b9b-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-a71c101e0a4c4b0c906379504495fbd5	8a9f9b92-79bc-4a48-91e0-094734cbbd0e	single	--	Mar 22 2021 19:46:55	Mar 22 2021 19:59:04	Mar 22 2021 20:25:52	00:26:47	SUCCEEDED
iipc-cd22da50d9a74f8f9489abe8ff7b8e74	fa69be70-6c48-4070-834e-eadb5ddba5b1	single	--	Mar 22 2021 19:46:54	Mar 22 2021 19:58:32	Mar 22 2021 20:24:52	00:26:20	SUCCEEDED
iipc-2696fd30fa894a239601b806cbbe3364	e0e53324-c41d-4124-aaed-885f192dc57d	single	--	Mar 22 2021 19:46:54	Mar 22 2021 19:58:31	Mar 22 2021 20:07:37	00:09:05	SUCCEEDED
iipc-e420a016b3f94d43b0dae4bbd97b2648	f61db6ea-6d6d-4ef6-9f7b-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-e731df62ad3f4d97a44ab2d9c28c89ad	6b53d0b1-e9b5-4493-b843-cf433745afd5	single	--	Mar 22 2021 19:46:51	Mar 22 2021 19:58:32	Mar 22 2021 20:24:33	00:26:00	SUCCEEDED
iipc-41d5097305394b949636c7413f20880d	00bc0a16-8d4d-4c6f-ba17-6d81366c8632	single	--	Mar 22 2021 19:46:51	Mar 22 2021 19:58:32	Mar 22 2021 20:25:42	00:27:10	SUCCEEDED
iipc-942a18be3a404b54ae933893841a7a38	8efa06f8-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-acd7a646db43417b9f02c58123505544	4fd277d1-b9d8-4f4c-abfd-c9a6eab714e8	single	--	Mar 22 2021 19:46:49	Mar 22 2021 19:58:32	Mar 22 2021 20:07:37	00:09:05	SUCCEEDED
iipc-4195be6b5ac04eb9bf05b400a3ef5c44	6c24e53c-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

VIRGINIA TECH

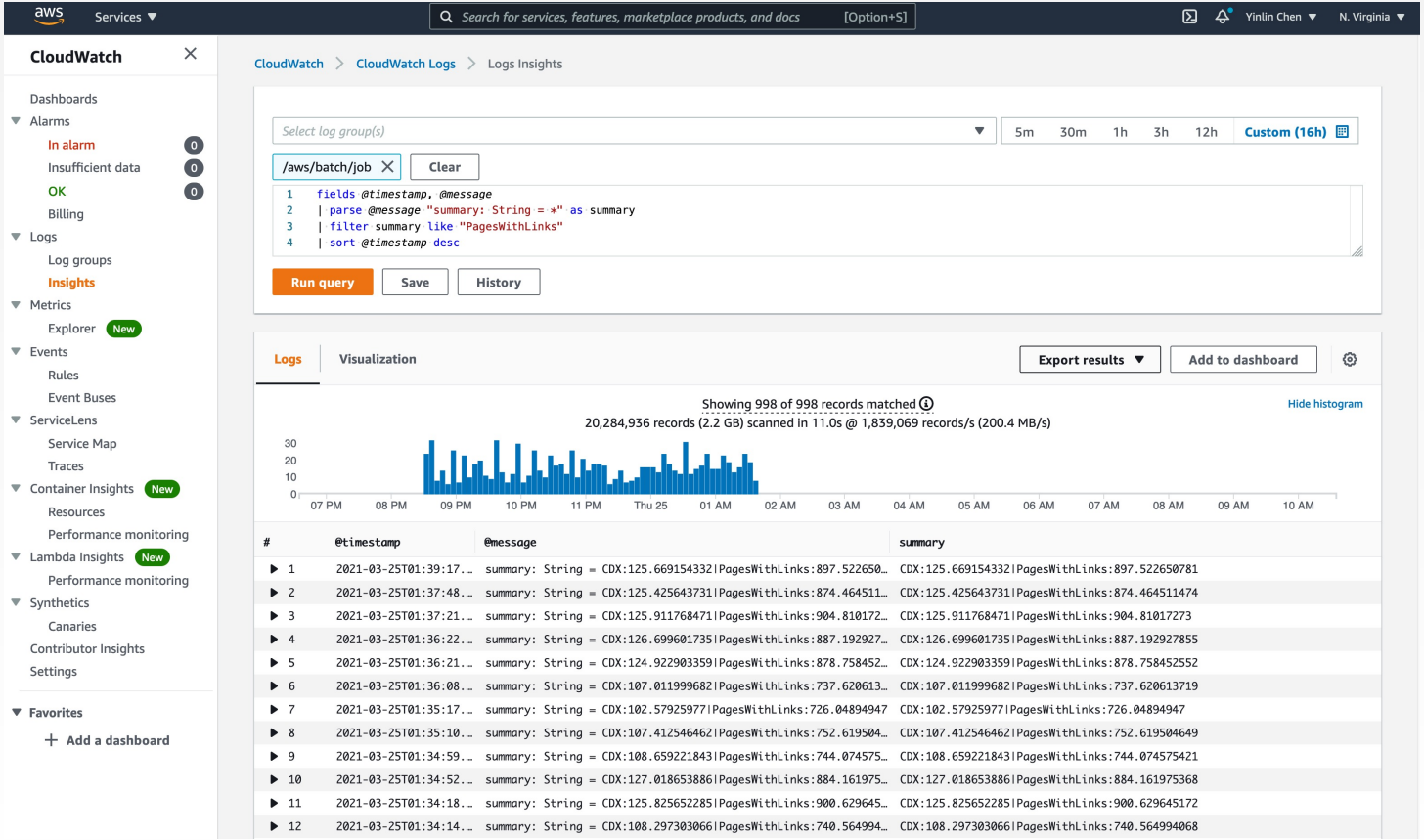
Batch Job Logs



2021-03-30T19:48:08.462-04:00	duration: Double = 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.Main$$anon$1 - Shutting down kernel
2021-03-30T19:48:08.877-04:00	21/03/30 23:48:08 [INFO] o.a.t.Main$$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.Main$$anon$1 - Shutting down actor system
2021-03-30T19:48:08.930-04:00	21/03/30 23:48:08 [INFO] o.a.t.Main$$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
2021-03-30T19:48:08.930-04:00	21/03/30 23:48:08 [INFO] o.a.t.Main$$anon$1 - Shutting down actor system
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-20200524210325-2020052…
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-20200524210325-2020052…
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-20200524210325-2020052…
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) remaining Completed 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.

VIRGINIA TECH

Query Log Results



#	@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:874.464511…	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

VIRGINIA TECH

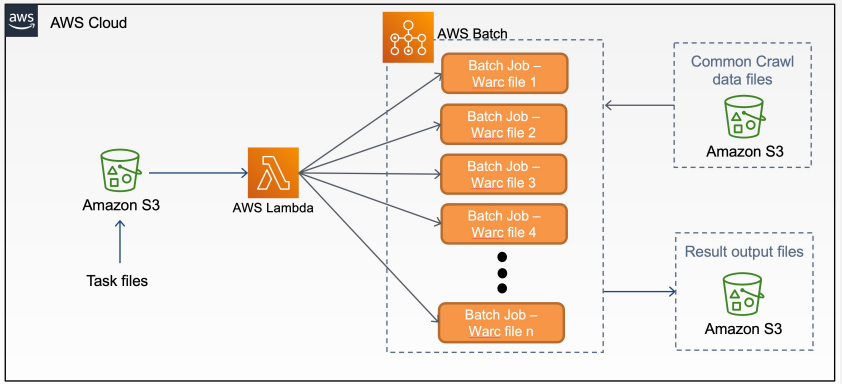
Results

- 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

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

Cost: Now



- 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
- GitHub: https://github.com/yinlinchen/warc-analytics-wac2021

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