



AWS Document Retrieval

Team Members:

Daniel Kim

Fadi Durah

Xavier Pleimling

Brandon Le

Matthew Hwang

CS4624, Multimedia, Hypertext, and Information Access

Dr. Edward A. Fox

Virginia Polytechnic Institute and State University

Blacksburg, VA 24061

May 3, 2020



Outline

- Project Overview
- Project Design
- Data Pipeline
- Front End Application & Database
- Future Goals
- Lessons Learned
- Acknowledgements
- References



Project Overview

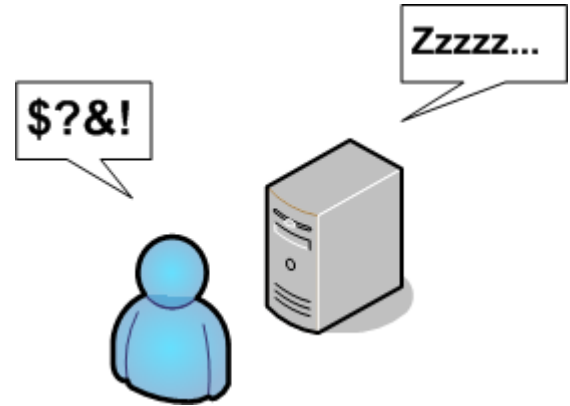


Project Overview - Desired Deliverable

A fully operational version of the CS 5604 search engine on Amazon Web Services (AWS). In short, moving the functional search engine to AWS.

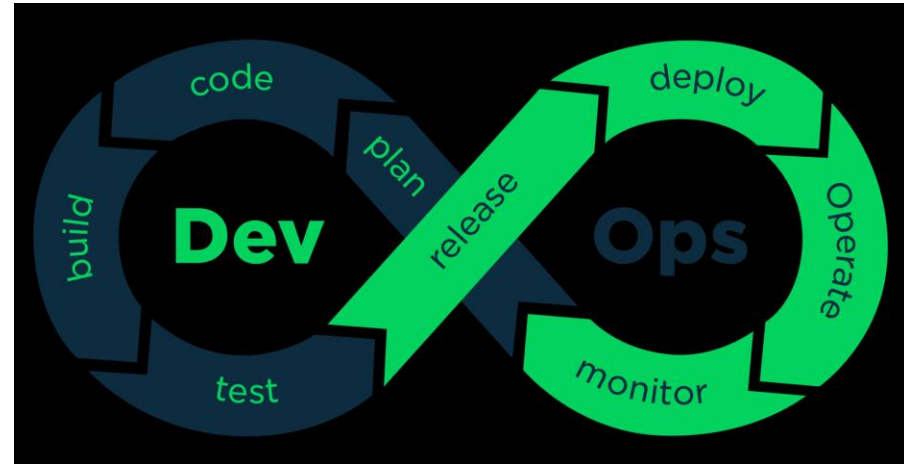
Project Overview - Need

- A reliance on the Virginia Tech network could mean downtime when the servers or a node is down
- AWS is a more stable service
- An experiment to investigate other infrastructures



Project Overview - Nature

- Less need of “developing”, original project was functional
- Different aspect of CS projects, more of DevOps nature
- Building infrastructure for existing code





Project Design

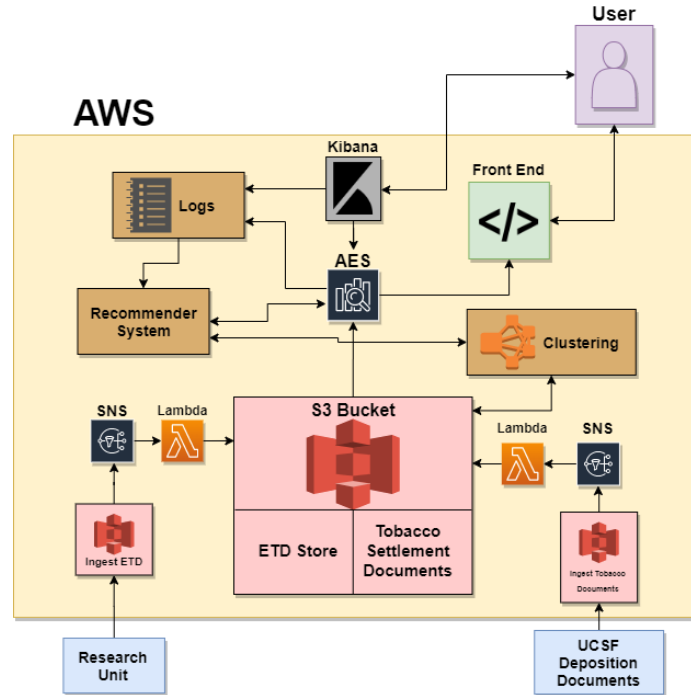


Main Concerns

- Setting up common storage
- Ingesting data into search engine (ElasticSearch)
- Port front-end application and deploy into AWS ECS
- Set up streaming platform for new data
- Reach Goal: CICD for front end application or any other containers

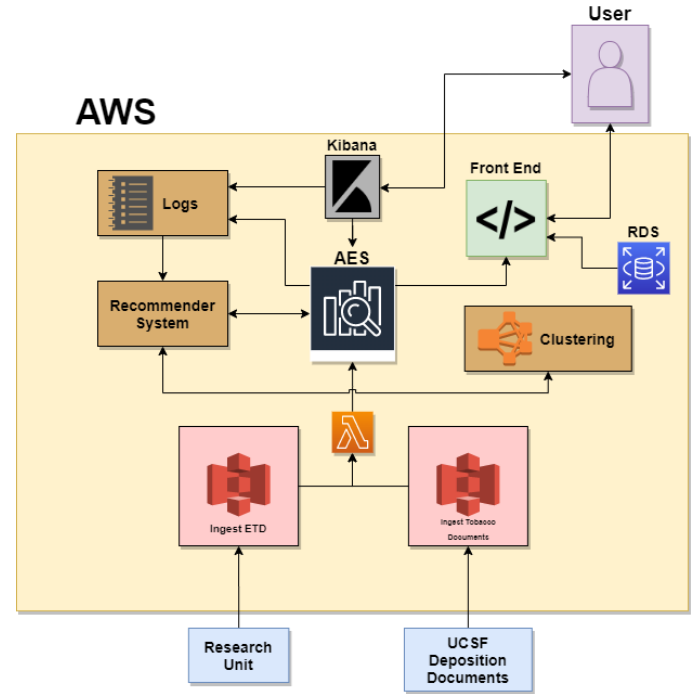
Initial Design

- Data comes in from external source
- Fed into large common storage (AWS S3)
- ElasticSearch ingests data from S3
- Kibana interact with ES, logs, and recommender
- Front-End Application



Updated Design

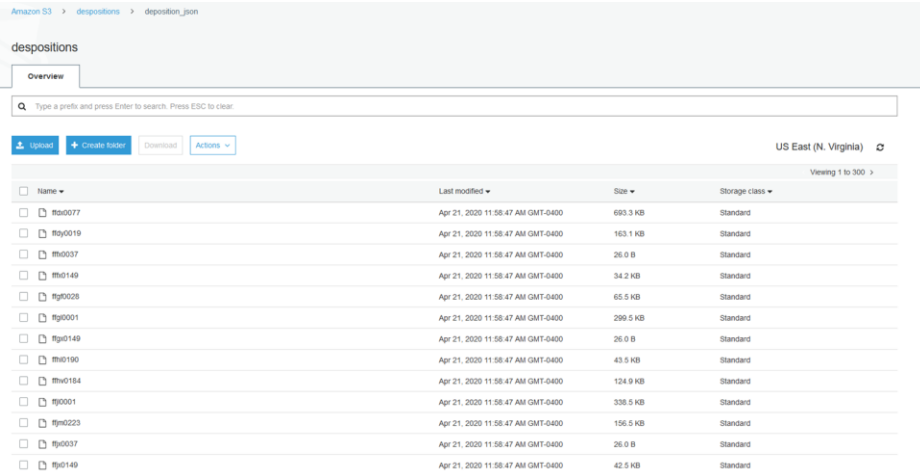
- No centralized S3 bucket
- Lambda now used to stream new data into ElasticSearch
- SNS not needed
- Included RDS for login verification



Data Pipeline

Common Storage

- Used AWS S3 buckets to contain our data
- One for ETDs and one for Tobacco Documents
- Configured to allow access from other AWS services



Amazon S3 > depositions > deposition_jon

depositions

Overview

🔍 Type a prefix and press Enter to search. Press ESC to clear.

📁 Upload 📄 Create folder 📄 Download ⚙️ Actions

US East (N. Virginia) 🌐

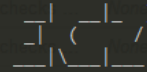
Viewing 1 to 300 >

<input type="checkbox"/> Name	Last modified	Size	Storage class
<input type="checkbox"/> rfd0077	Apr 21, 2020 11:58:47 AM GMT-0400	693.3 KB	Standard
<input type="checkbox"/> rfd0019	Apr 21, 2020 11:58:47 AM GMT-0400	163.1 KB	Standard
<input type="checkbox"/> rfd0037	Apr 21, 2020 11:58:47 AM GMT-0400	26.0 B	Standard
<input type="checkbox"/> rfd0149	Apr 21, 2020 11:58:47 AM GMT-0400	34.2 KB	Standard
<input type="checkbox"/> rfd0028	Apr 21, 2020 11:58:47 AM GMT-0400	65.5 KB	Standard
<input type="checkbox"/> rfd0001	Apr 21, 2020 11:58:47 AM GMT-0400	299.5 KB	Standard
<input type="checkbox"/> rfd0149	Apr 21, 2020 11:58:47 AM GMT-0400	26.0 B	Standard
<input type="checkbox"/> rfd0190	Apr 21, 2020 11:58:47 AM GMT-0400	43.5 KB	Standard
<input type="checkbox"/> rfd0184	Apr 21, 2020 11:58:47 AM GMT-0400	124.9 KB	Standard
<input type="checkbox"/> rfd0001	Apr 21, 2020 11:58:47 AM GMT-0400	338.5 KB	Standard
<input type="checkbox"/> rfd0223	Apr 21, 2020 11:58:47 AM GMT-0400	156.5 KB	Standard
<input type="checkbox"/> rfd0037	Apr 21, 2020 11:58:47 AM GMT-0400	26.0 B	Standard
<input type="checkbox"/> rfd0149	Apr 21, 2020 11:58:47 AM GMT-0400	42.5 KB	Standard

External File System

- Created two AWS EC2 instances to act as external servers
- Can ssh with a private .pem key
- Each one has a bucket mounted on it
- Able to list, read, and write to the buckets

```
Last login: Tue Apr 21 23:20:44 2020 from c-98-249-4-214.hsd1.va.comcast.net
Last login: Tue Apr 21 23:20:44 2020 from c-98-249-4-214.hsd1.va.comcast.net
```



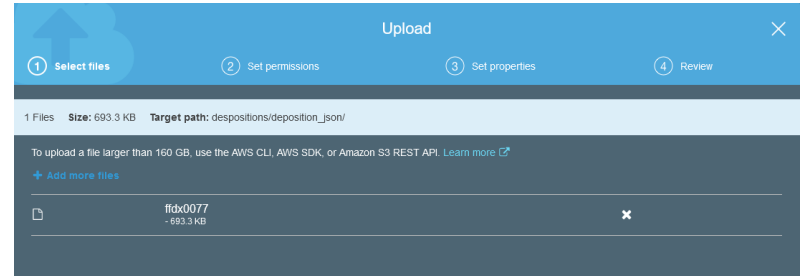
Amazon Linux 2 AMI

```
https://aws.amazon.com/amazon-linux-2/
[ec2-user@ip-10-0-1-106 ~]$ cd /buckets/deposition_json/
```

```
root@ip-10-0-1-106 deposition_json# ls
Ffdx0077  Fnlm0226  Fyfx0217  gmh10001  gxwg0021  hljx0149  httpj0184  jzjlp0061
Ffdy0019  Fnlkx0149  Fyjl0226  gmh10191  gxxd0001  hlk10001  httpp0183  jzjlv0190
FFfx0037  Fnl10001  Fy110001  gmhy0219  gx110001  hlkp0061  hrtwg0021  jzjpp0183
FFfx0149  Fnlv0103  Fy1v0103  gmkd0100  gx00037  hlkw0181  hrtwx0149  jzjpp0184
FFfg0028  Fnlw0100  Fy1w0100  gmkd0224  gycg0219  hlkx0149  hrtwx0225  jzjxd0001
FFgx0001  Fnlw0221  Fyxm0132  gmkl0190  gycv0082  hlml0001  hrtxd0001  jzjxf0028
FFfx0149  Fnm10001  Fyng0184  gmkm0226  gyfg0225  hlrw0087  hrtxm0020  jzxl10001
FFfh0190  Fnp00183  Fypp0042  gmkn0226  gyfp0183  hlpd0100  hrtxw0181  jzxp0018
FFhv0184  Fnw0021  Fyvw0082  gmkp0018  gyfx0149  hlpm0191  hrtxx0037  jzxx0037
FFj10001  Fnw0149  Fywx0149  gmks0149  gygf0028  hlvb0187  htyy0083  jzzy0083
FFjmm0223  Fnxp0180  Fyxf0028  gm110001  gyg10001  hlwg0021  hxdb0184  jzzy0184
FFjx0037  Fnxw0100  Fyxx0037  gm1p0018  gyhf0028  hlwx0149  hxdy0019  jkby0019
FFjy0149  Fny0097  Fyvg0225  gm1p0183  gyhh0225  hlxx0191  hxfg0021  jkc01014
FFlw0221  Fnyg0225  Fyyp0183  gm1w0221  gyhj0223  hlxl0001  hxg10001  jkfg0021
FFm10190  Fpb10191  Fzby0019  gmpr0227  gyhy0219  hlxm0008  hxg10091  jkfg0028
FFmm0056  Fpc10154  Fzdy0019  gmwx0149  gyjz0191  hlxp0018  hxhb0089  jkhf0028
FFmp0018  Fpcc0225  Fzfj0223  gmxx0191  gyj10001  hllyk0135  hxh10001  jkhg0225
FFmx0149  Fpd10011  Fzfl0001  gm110001  gyjp0180  hmb10056  hxh10191  jkhk0083
FFfn0219  Fpdv0184  Fzfx0149  gmxp0180  gykd0224  hmbw0082  hxhv0184  jkh10001
FFpf0100  Fpdy0019  Fzgd0001  gmxx0100  gyl10190  hmbv0019  hxj10001  jkhy0042
FFpy0034  Fpfg0023  Fzg10001  gnc10154  gykx0149  hmcv0184  hxj0180  jkj10001
FFvn0178  Fpfg0225  Fzhd0001  gncv0082  gy110001  hmcx0225  hxj0037  jkjpp0018
FFvy0189  Fpfx0149  Fzfh0028  gndn0226  gy1v0103  hmdg0145  hxj0149  jkjx0149
FFwd0086  Fpg10001  Fzhh0225  gnfx0149  gy1w0100  hmdy0019  hxkx0149  jkl10001
FFwg0021  Fphb0089  Fzhj0223  gngf0028  gy1w0221  hmfg0021  hlxb0019  jkpp0061
FFfx0028  Fphj0223  Fzh10001  gnhf0028  gyl10001  hmfv0191  hlj10015  jkpw0181
FFxh0225  Fph10001  Fzhp0018  gnhh0225  gylk0135  hmfx0149  hlp10061  jkxx0149
FFxj0223  Fphy0219  Fzhx0037  gngh0223  gypp0183  hmgb0089  hxmn0226  jkz10001
FFxn0191  Fpj0035  Fzhy0219  gn110001  gyvj0223  hmfg0028  hxmp0018  jkm10001
FFxx0037  Fpj0180  Fzj10001  gnhy0219  gyvw0082  hmhd0001  hxn10001  jkmp0018
ggch0065  Fpkh0228  Fzj10011  gnk10001  gywx0149  hmhg0225  hxxd0001  jkn10001
ggc10154  Fpk10001  Fzjpp018  gnkm0226  gyl10001  hmh10001  hxxf0028  jkpk0135
ggcv0184  Fpk10190  Fzk10001  gnkx0149  gyxx0037  hmhw0181  hxxx0037  jkvf0228
ggdy0019  Fplkx0149  Fzk10190  gnkx0217  gzby0019  hmhy0219  hxyh0225  jkgw0021
ggfx0149  Fplkx0217  Fzk10223  gn110001  gzc01000  hmjpp018  hycy0019  jkxd0001
gggf0028  Fplj0191  Fzkkp0018  gn1v0103  gzcg0145  hmkd0100  hydb0184  jkx10001
```

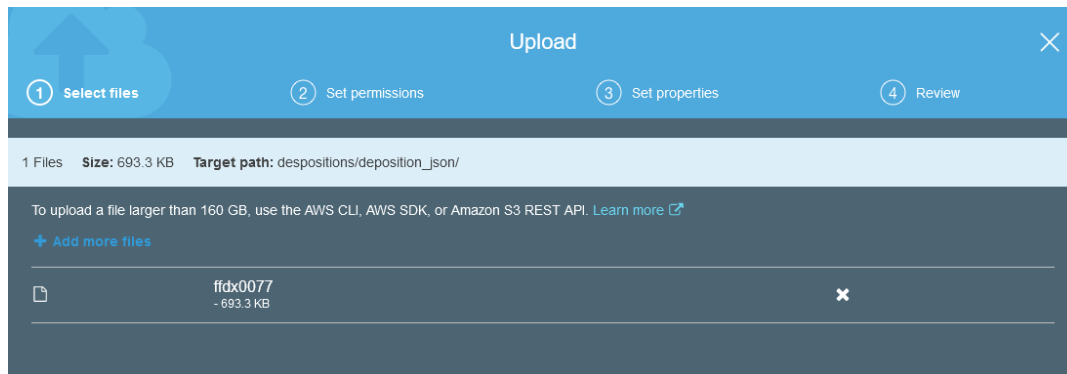
Data Streaming

- Whenever new data is placed in the common storage (S3), it will be automatically processed into AWS Elasticsearch
- AWS Lambda script is triggered whenever new data is put into the buckets



▼ etd_metadata

Count	1
Size in bytes	3.77 KiB
Query total	8
Mappings	



▼ etd_metadata

Count 1
Size in bytes 3.77 KiB
Query total 8
Mappings



▼ etd_metadata

Count 1
Size in bytes 3.77 KiB
Query total 32
Mappings

System Monitoring

- Using AWS Cloudwatch allows easy viewing of function metrics and logs
- Viewing what failed or succeeded with logs

CloudWatch > Log Groups

Create Metric Filter Actions

Filter: Log Group Name Prefix

Log Groups	Insights	Expire Events After
<input type="radio"/> /aws/ec2/containerinsights/abc/performance	Explore	1 day
<input type="radio"/> /aws/lambda/stream-s3-std	Explore	Never Expire
<input type="radio"/> /ecs/aatask	Explore	Never Expire
<input type="radio"/> /ecs/first-run-task-definition	Explore	Never Expire
<input type="radio"/> /ecs/front-end	Explore	Never Expire
<input type="radio"/> /ecs/summary-task	Explore	Never Expire

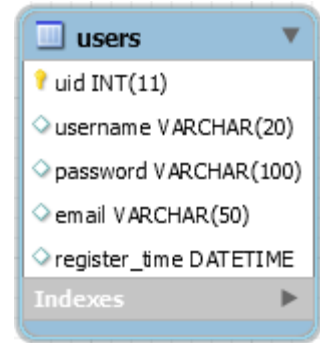
Filter events 2020-04-22 (20:50:00) - 2020-04-22 (23:50:00)

Time (UTC +00:00)	Message
2020-04-22	No older events found at the moment. Retry
23:43:04	START RequestId: 86f319a9-6aaa-4611-984f-01db5e290702 Version: SLATEST
23:43:05	"NameType" object has no attribute "group": AttributeError Traceback (most recent call last): File "/var/task/reader.py", line 40, in handler ip = (ip_pattern.search(line) group(1) AttributeError: 'NoneType' object has no attribute 'gr
23:43:05	END RequestId: 86f319a9-6aaa-4611-984f-01db5e290702
23:43:05	REPORT RequestId: 86f319a9-6aaa-4611-984f-01db5e290702 Duration: 346.29 ms Billed Duration: 400 ms Memory Size: 128 MB Max Memory Used: 76 MB Init Duration: 599.68 ms
23:44:12	START RequestId: 86f319a9-6aaa-4611-984f-01db5e290702 Version: SLATEST
23:44:12	"NameType" object has no attribute "group": AttributeError Traceback (most recent call last): File "/var/task/reader.py", line 40, in handler ip = (ip_pattern.search(line) group(1) AttributeError: 'NoneType' object has no attribute 'gr
23:44:12	END RequestId: 86f319a9-6aaa-4611-984f-01db5e290702
23:44:12	REPORT RequestId: 86f319a9-6aaa-4611-984f-01db5e290702 Duration: 250.21 ms Billed Duration: 300 ms Memory Size: 128 MB Max Memory Used: 76 MB
23:46:00	START RequestId: 86f319a9-6aaa-4611-984f-01db5e290702 Version: SLATEST
23:46:01	"NameType" object has no attribute "group": AttributeError Traceback (most recent call last): File "/var/task/reader.py", line 40, in handler ip = (ip_pattern.search(line) group(1) AttributeError: 'NoneType' object has no attribute 'gr
23:46:01	END RequestId: 86f319a9-6aaa-4611-984f-01db5e290702
23:46:01	REPORT RequestId: 86f319a9-6aaa-4611-984f-01db5e290702 Duration: 241.86 ms Billed Duration: 300 ms Memory Size: 128 MB Max Memory Used: 77 MB

Front End Application & Database

RDS Instance

- User login database
- Table schema from FEK group
- Previous login information was not migrated
- Front-end sign up form populates new entries



RDS Connection and Accessibility



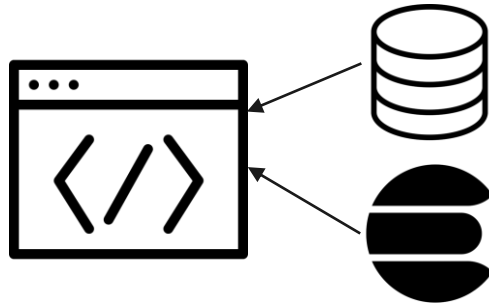
- Publicly accessible RDS instance
- Connection test with MySQL Workbench 8.0
- Workbench helps with schema configuration
- Learned to connect to RDS endpoint and set up MySQL environment

	uid	username	password	email	register_time
▶	1	1234	\$5\$rounds=535000\$zscTNVylZT7rFPZ\$IVJ44Bf...	123456	2020-04-15 21:06:41
	2	username	\$5\$rounds=535000\$9J9pnAgUuTKxDTFs\$7EGx...	email@email	2020-04-15 21:15:31
	3	username1	\$5\$rounds=535000\$KygQ9lvLGjH3q7Bj\$kiWwB...	email1@email	2020-04-16 00:33:59
	4	username2	\$5\$rounds=535000\$XyBNDGf88DsaZP\$WZSp...	asdfasdf	2020-04-16 00:46:57
	5	asdfasdf	\$5\$rounds=535000\$q9aZGwQtirhJy/mO\$NFp/...	asdfasdfasdf	2020-04-16 00:50:54
	6	asdfasdfasdf	\$5\$rounds=535000\$2Cid5JuGsXnezKHg\$40IBF...	asdfasdfasdfasdf	2020-04-16 00:55:44
	7	asdfasdfasdfasdf	\$5\$rounds=535000\$7IEEBi5NHgtrO6nK\$7GwT...	asdfasdfasdfasdf	2020-04-16 01:05:45
	8	asdfasdfasdfasdf	\$5\$rounds=535000\$A5UPrXpPnPIz4fsH\$4jmdF...	asdfasdfasdfasdf	2020-04-16 01:10:25
*	NULL	NULL	NULL	NULL	NULL

Connected Database and ElasticSearch

- Learned that connections established through settings file
- Created settings.yaml file
- Inserted Database and ElasticSearch endpoints

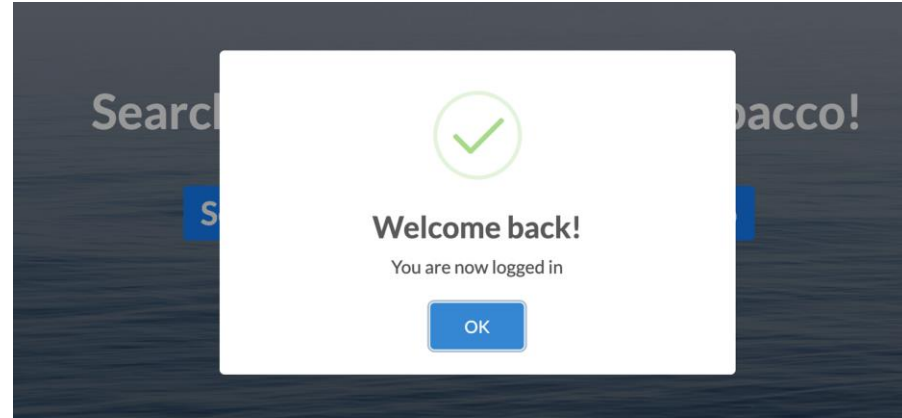
```
settings.yaml
1 default:
2   mysql:
3     host: [REDACTED]
4     port: 3306
5     user: admin
6     password: [REDACTED]
7     database: verification
8   elasticsearch: https://search-dlrl-elasticsearch-rmd75zwsfd7rmi7doj52t
9   baseuri: http://0.0.0.0:3001
```





Established Login Functionality

- Initially only create account worked
- Learned that original original code required missing field
- Modified code to work with new database





ElasticSearch UI

- Learned that search UI is separate application
- Displayed in main app through Iframe
- Had to build separately before able to be displayed

Search for ETD

date-issued

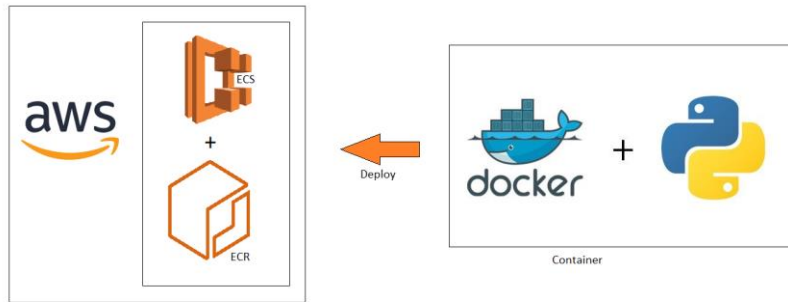
-



Future Goals

Front End Container

- Upload container image to ECR
- Deploy through ECS





Lessons Learned



Lessons Learned

- Preventing permissions issues; ensuring all project resources are available to use
- Process of setting up infrastructure
- Understanding how the original project worked

Acknowledgements



Acknowledgements

Rahul Agarwal (Our Client)

- Worked on original project in CS 5604
- Part of Integration and Implementation (INT) Team of the original project

INT and FEK Groups

- CS 5604 Class Students
- Implemented original project

Bill Ingram

- Director of IT at VT Library
- ETD documents



Acknowledgements Cont.

ETD Project Grant IMLS LG-37-19-0078-19

- Project funding

Dr. David Townsend

- Tobacco documents

Dr. Edward Fox

- Thank you for the continual guidance

Carlos Augusto

- Thank you for the advice on how to approach presenting and illustrating our project



References



References

- R. Agarwal, H. Albahar, E. Roth, M. Sen, and L. Yu, "Integration and Implementation (INT) CS5604 Fall 2019," *VTechWorks*, 11-Dec-2019. [Online]. Available: <http://hdl.handle.net/10919/96488>. [Accessed: 04-Feb-2020].
- ETDs: Virginia Tech Electronic Theses and Dissertations. *VTechWorks*. [Online]. Available: <https://hdl.handle.net/10919/5534>. [Accessed: 14-Feb-2020].
- K. K. Kaushal, R. Kulkarni, A. Sumant, C. Wang, C. Yuan, and L. Yuan, "Collection Management of Electronic Theses and Dissertations (CME) CS5604 Fall 2019," *VTechWorks*, 23-Dec-2019. [Online]. Available: <https://hdl.handle.net/10919/96484>. [Accessed: 03-May-2020].
- Y. Li, S. Chekuri, T. Hu, S. A. Kumar, and N. Gill, "ElasticSearch (ELS) CS5604 Fall 2019," *VTechWorks*, 12-Dec-2019. [Online]. Available: <http://hdl.handle.net/10919/96310>. [Accessed: 10-Mar-2020].
- R. S. Mansur, P. Mandke, J. Gong, S. M. Bharadwaj, A. S. Juvekar, and S. Chougule, "Text Analytics and Machine Learning (TML) CS5604 Fall 2019," *VTechWorks*, 29-Dec-2019. [Online]. Available: <http://hdl.handle.net/10919/96226>. [Accessed: 11-Feb-2020].
- S. Muhundan, A. Bendelac, Y. Zhao, A. Svetovidov, D. Biswas, and A. Marin Thomas, "Collection Management Tobacco Settlement Documents (CMT) CS5604 Fall 2019," *VTechWorks*, 11-Dec-2019. [Online]. Available: <https://hdl.handle.net/10919/96437>. [Accessed: 03-May-2020].
- E. Powell, H. Liu, R. Huang, Y. Sun, and C. Xu, "Front-End Kibana (FEK) CS5604 Fall 2019," *VTechWorks*, 13-Jan-2020. [Online]. Available: <http://hdl.handle.net/10919/96418>. [Accessed: 04-Feb-2020].
- Truth Tobacco Industry Documents Library. [Online]. Available: <https://www.industrydocuments.ucsf.edu/tobacco/>. [Accessed: 14-Feb-2020].



Questions?