Architecting A Cloud-native Data Analysis Application for ETDs

Yinlin Chen & Edward A. Fox

{ylchen, fox}@vt.edu

Virginia Tech, Blacksburg, VA 24061, USA

26 Sept. Presentation at ETD 2018, Taiwan

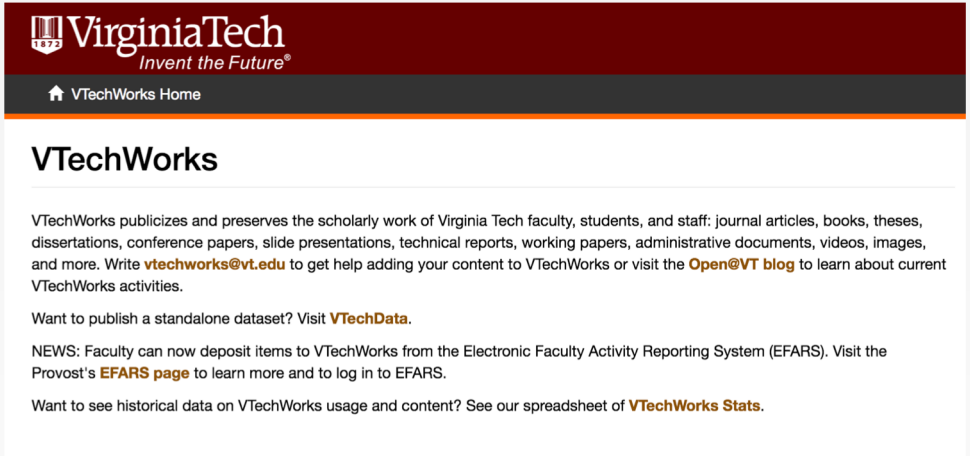
VIRGINIA TECH

Agenda

- ***Introduction***
- Monolithic vs. Cloud-native applications
- Cloud-native approach
- Data analytics architecture
- Future work

VTechWorks

- Research Documents
  - 60,000+ scholarly works
  - Inc. 30,000+ ETDs
- Using DSpace
  - Open digital repository
  - Open source project
  - Monolithic architecture



DSPACE

DURASPACE™

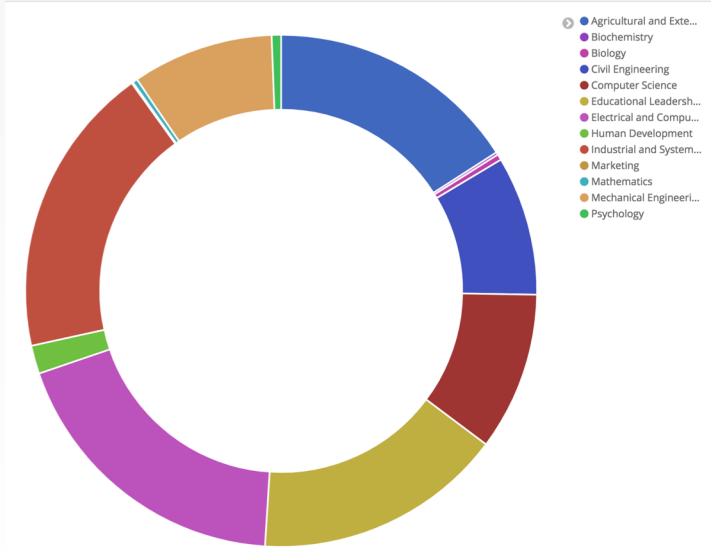
Problems

- Need for advanced analytics functionalities
  - Customizable
  - Visualization
  - Usable and flexible user interface (UI)
- Extending existing codebase is a difficult task.
  - Customization that meets community needs is rare.
  - Version update will lose or break customized functionalities.

From Research Dataset to Visualization

id	collection	dc.contributc	dc.contributc	dc.contributc	dc.contributc	dc.contributc	dc.contributc	dc.contributc	dc.contributc	dc.contributc	dc.contributc	dc.co
50440	10919/11041	||10919/717	Larochelle, Catherine			Alwang, Jeffrey R.						Amacher, Gregory
43029	10919/9291	||10919/7175	Ketene, Alperen Nurullah			Agah, Masoud						Behkam, Bahareh|
42401	10919/9291	||10919/7175	Smith, Ryan Christopher			Geller, E. Scott						Spencer, Edward F
43494	10919/9291	||10919/7175	Zareian-Jahromi, Mohammad Amin			Agah, Masoud						Raman, Sanjay| |M
43064	10919/9291	||10919/7175	Schuler, Matthew Michael			FitzPatrick, William J.						Klagge, James C.| |
89430	10919/9291	||10919/7175	Kohler, Rachel Elizabeth			Luther, Kurt						North, Christopher
89034	10919/9291	||10919/7175	Dellinger, Elizabeth Aalseth			Gardner, Thomas M						Swenson, Karen| |
88738	10919/9291	||10919/7175	Williams, Rebecca Jean			Jones, Kathleen W						Shumsky, Neil Larr
11228	10919/11041		Lee, Dong-Ho			Lee, Fred C.						Chen, Dan Y.| |Bor
11229	10919/11041		Kim, Byung-ki			Stutzman, Warren L.						Sweeney, Dennis G
11230	10919/11041		Kriz, Kerri-Lynn Murphy			Madison-Colmore, Octavia D.						Messier, Louis| |H
11231	10919/11041		Lipkovich, Ilya A			Smith, Eric P.			Ye, Keying			Foutz, Robert| |Bir
11232	10919/11041		Lorca, Tatiana Andrea			Pierson, Merle D.						Eifert, Joseph D.| |
11441	10919/11041		Crozier, James Brooks			Stromberg, Erik L.						
11442	10919/11041		Cooper, Jamie S.			Cooper, Robin K. Panneton						Lickliter, Robert E.
11443	10919/11041		Clark, Paul Alexander			Oyama, Shigeo Ted						Vandsburger, Uri|
11444	10919/11041		Hafsteinsson, Leifur Geir			Donovan, John J.						Carlson, Kevin D.| |
11445	10919/11041		Rowland, Amy Lee			Stratton, Richard K.						Poole, Jon R.| |Kro
11235	10919/11041		Prater, Mary Renee			Holladay, Steven D.						Wong, Eric A.| |Be
11447	10919/11041		Zhao, Xiaopeng			Nayfeh, Ali H.						Dankowicz, Harry J
11448	10919/11041		Rosario, Astrid Christa			Riffle, Judy S.						Long, Timothy E.| |
11238	10919/11041		Torrence, Vera D.			Krill, Cecelia W.						Parson, Stephen R.
11450	10919/11041		Niu, Sanjun			Saraf, Ravi F.						Oyama, Shigeo Tet
11240	10919/11041		Parrett, Matthew Barton						Stegeman, Mark| |Eckel, Catherine C.			Haller, Hans H.| |M
11242	10919/11041		Lahouar, Samer						Al-Qadi, Imadeddin L.| |Brown, Gary S.			de Wolf, David A.|
11243	10919/11041		Bradley, Kevin Michael			Hauenstein, Neil M. A.						Finney, Jack W.| |D
11244	10919/11041		Seock, Yoo-Kyoung			Norton, Marjorie J. T.						Littlefield, James E
11453	10919/11041		Robinson, Tammy Renee'			Giddings, Valerie L.						Bailey, Carol A.| |S
11454	10919/11041		Clevenger, Jennifer Lynn			Giddings, Valerie L.						Kincade, Doris H.| |
11455	10919/11041		Gough, Christopher Michael			Seiler, John R.						Fox, Thomas R.| |F
11456	10919/11041		Ahmed, Farzana			Larson, Timothy J.						Popham, David L.|
11457	10919/11041		Lee, Yuri			Kincade, Doris H.						Giddings, Valerie L
11458	10919/11041		Liang, Hongping			Hilu, Khidir W.						Opell, Brent D.| |P
11459	10919/11041		Liu, Sixin						Griffey, Carl A.| |Maroof, M. A. Saghai			Wilkinson, Carol A.
11460	10919/11041		Li, Yuxin			Huang, Alex Q.						Nelson, Douglas J.
11461	10919/11041		van Aardt, Jan Andreas Nicholaas			Wynne, Randolph H.						Oderwald, Richard
11462	10919/11041		Renneckar, Scott Harold						Zink-Sharp, Audrey G.| |Glasser, Wolfg			Ducker, William A.
11463	10919/11041		Perry, Elizabeth A.			Salehi-Isfahani, Djavad						Mills, Bradford F.|

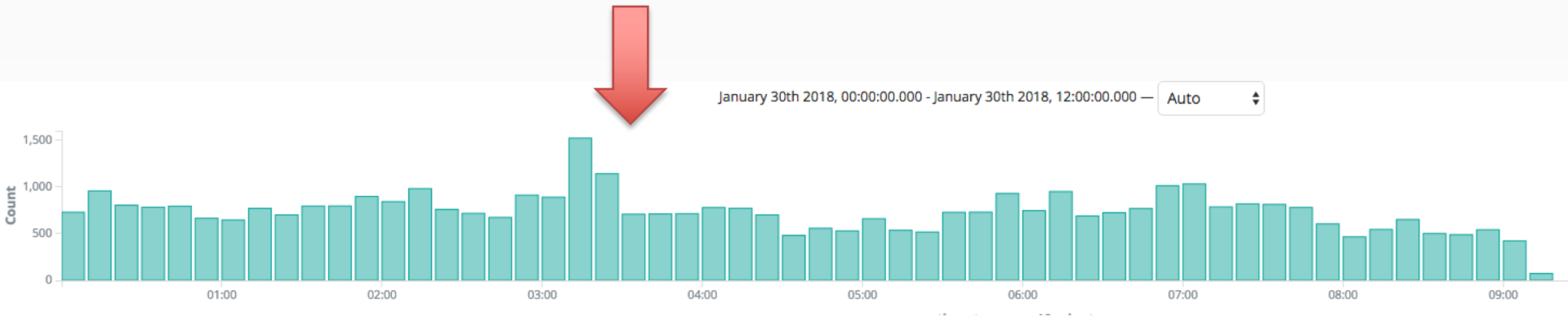
Categorized by Department



VIRGINIA TECH

From Log Data to Visualization


```
"2018-01-30T08:13:13.200Z","192.168.2.6","prod_log: 202.248.84.158 vtechworks.lib.vt.edu [30/Jan/2018:03:13:15 -0500] GET /bitstream/handle/10919/33268/WGraf_Thesis_2005.pdf?sequence=1 HTTP/1.1   Mozilla/5.0 (Windows NT 6.3; WOW64; Trident/7.0; rv:11.0) like Gecko"
"2018-01-30T07:54:58.955Z","192.168.2.6","prod_log: 104.220.158.2 vtechworks.lib.vt.edu [30/Jan/2018:02:55:01 -0500] GET /bitstream/handle/10919/10361/TX715.C543_1849.pdf?sequence=1&isAllowed=y HTTP/1.1   Mozilla/5.0 (Windows NT 6.1; WOW64; rv:55.0) Gecko/20100101 Firefox/55.0"
"2018-01-30T07:54:59.018Z","192.168.2.6","prod_log: 40.77.167.27 vtechworks.lib.vt.edu [30/Jan/2018:02:55:01 -0500] GET /bitstream/handle/10919/76427/LD5655.V855_1984.D465.pdf.jpg?sequence=3&isAllowed=y HTTP/1.1   Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)"
"2018-01-30T07:55:12.635Z","192.168.2.6","prod_log: 157.55.39.137 vtechworks.lib.vt.edu [30/Jan/2018:02:55:14 -0500] GET /browse?type=author&value=Yang%2C+S HTTP/1.1   Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)"
"2018-01-30T08:13:13.810Z","192.168.2.6","prod_log: 202.248.84.158 vtechworks.lib.vt.edu [30/Jan/2018:03:13:15 -0500] GET /bitstream/handle/10919/33268/WGraf_Thesis_2005.pdf?sequence=1 HTTP/1.1   Mozilla/5.0 (Windows NT 6.3; WOW64; Trident/7.0; rv:11.0) like Gecko"
"2018-01-30T08:13:14.127Z","192.168.2.6","prod_log: 202.248.84.158 vtechworks.lib.vt.edu [30/Jan/2018:03:13:16 -0500] GET /bitstream/handle/10919/33268/WGraf_Thesis_2005.pdf?sequence=1 HTTP/1.1   Mozilla/5.0 (Windows NT 6.3; WOW64; Trident/7.0; rv:11.0) like Gecko"
"2018-01-30T08:13:06.173Z","192.168.2.6","prod_log: 99.229.202.186 vtechworks.lib.vt.edu [30/Jan/2018:03:13:08 -0500] GET /handle/10919/5531/search-filter?field=author&filter_0=%5B1980%20TO%201989%5D&filter_relational_operator_0=equals&filtertype_0=dateIssued&starts_with=j HTTP/1.1   Mozilla/5.0 (compatible; TinEye-bot/1.31; +http://www.tineye.com/crawler.html)"
"2018-01-30T07:56:54.832Z","192.168.2.6","prod_log: 157.36.157.247 vtechworks.lib.vt.edu [30/Jan/2018:02:56:56 -0500] GET /wp-login.php HTTP/1.1   Mozilla/5.0 (Windows NT 6.1; WOW64; rv:40.0) Gecko/20100101 Firefox/40.1"
"2018-01-30T18:54:50.951Z","192.168.2.6","prod_log: 207.46.13.132 vtechworks.lib.vt.edu [30/Jan/2018:13:54:53 -0500] GET /handle/10919/24211/discover?filtertype_0=author&filtertype_1=subject&filtertype_2=subject&filter_relational_operator_1=equals&filtertype_3=subject&filter_relational_operator_0=equals&filter_2=galaxies%3A+Seyfert&filter_relational_operator_3=equals&filter_1=galaxies%3A+active&filter_relational_operator_2=equals&filter_0=Costantini%2C+E.&filter_3=quasars%3A+absorption+lines&filtertype=dateIssued&filter_relational_operator=equals&filter=2011 HTTP/1.1   Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)"
```

VIRGINIA TECH

Agenda

- Introduction
- ***Monolithic vs. Cloud-native applications***
- Cloud-native approach
- Data analytics architecture
- Future work

Monolithic Architecture

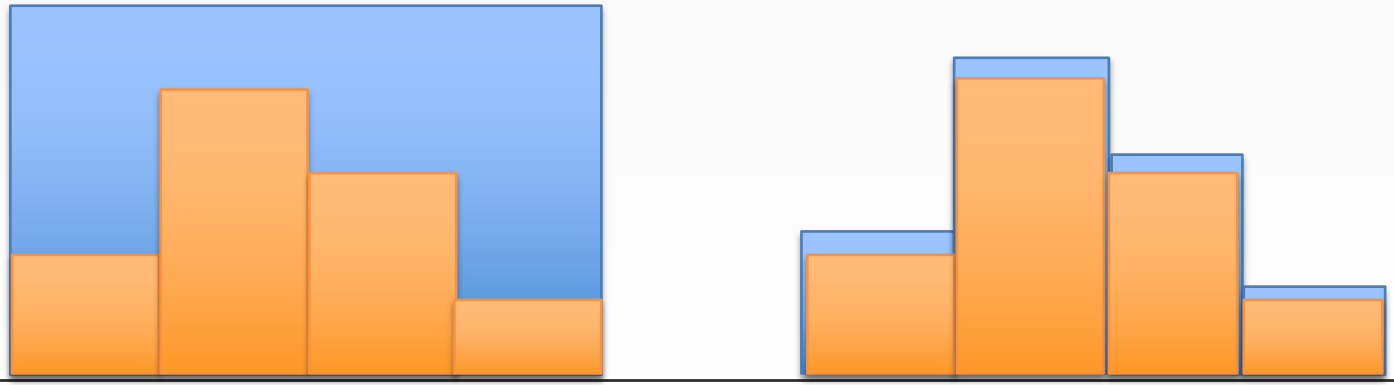
- Develop and deploy as a single unit
- Often requires human intervention
- Long-term commitment to a technology stack or even version
- Hard to scale development
- Difficult to scale the application

Why Towards Cloud Native

- Limited resources:
  - Developers, DevOps, Infrastructure, Time
- Reduce need to build everything from scratch
- Use services that can help deliver the project
- Facilitate the development process
- Provide better services: fault-tolerant, auto-scale, update/rollback without downtime, etc.
- Optimize resource utilization

Resource Usage Optimization and Automation

- Consume only the required resources for the applications
- Scale up and down automatically
- Service and function oriented, not server oriented
- Utilize cloud services to help understand applications (CloudWatch, Auto Scaling, Trusted Advisor, etc.)



Agenda

- Introduction
- Monolithic vs. Cloud-native applications
- ***Cloud-native approach***
- Data analytics architecture
- Future work

What is Cloud Native?

It is not just putting applications in the cloud.

It is about building applications in the cloud that utilize the **advantages** provided by the cloud

AS MUCH AS POSSIBLE.

Cloud Native

- Cloud Native Computing Foundation (CNCF)
  - An open source software foundation dedicated to making cloud native computing universal and sustainable
- Microservices oriented
- Containerized
- Dynamically orchestrated

Microservices oriented

Microservice

- Small software piece
- Messaging enabled – communicate with messages
- Decentralized
  - Autonomously developed
  - Independently deployable
  - Can change independently of each service
  - Scale individually by load
- Built and released with automated processes
- More complex architecture

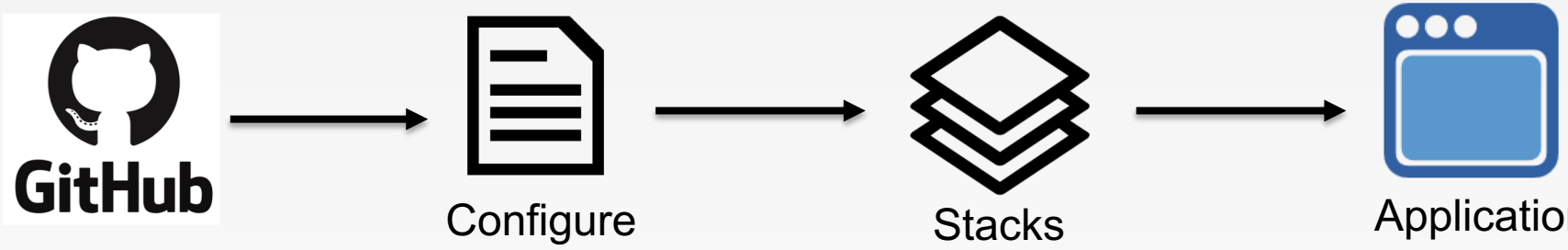
Serverless

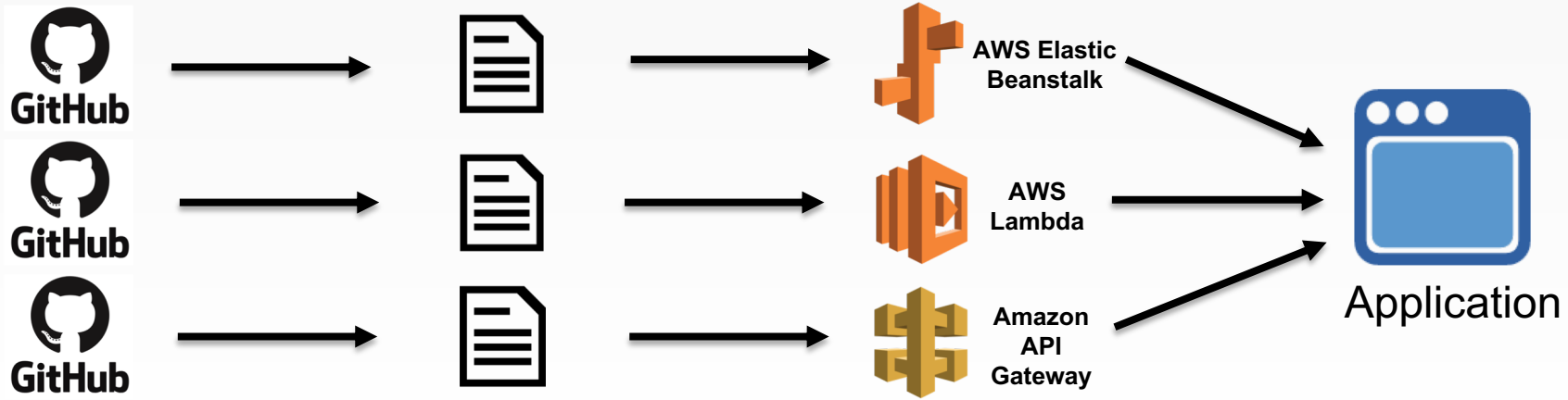
Does not mean “There are no servers at all”.

Does mean “Use fully managed services”.

Focus on application development,
not server maintenance

Parallel Development and Deployment





Containerized

Containerization

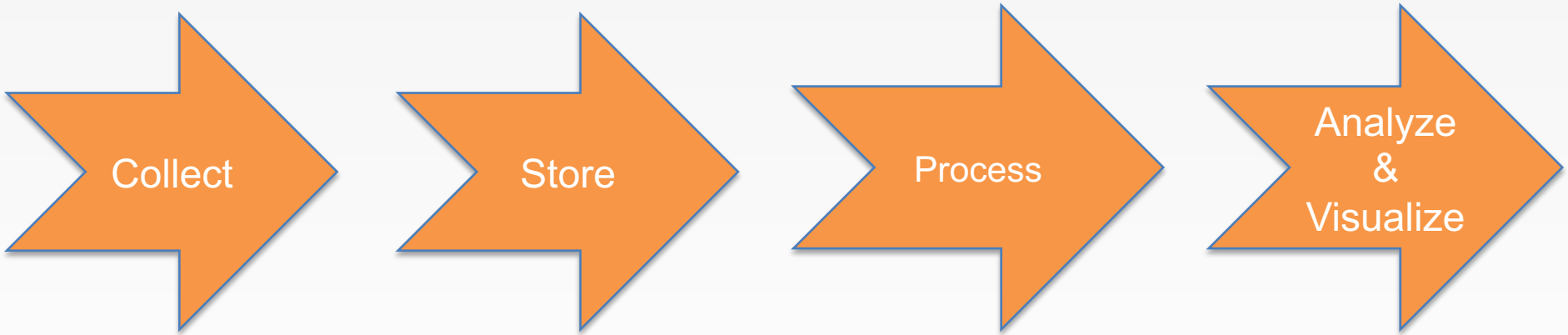
- BaaS (Backend as a Service), CaaS (Container as a Service), and FaaS (Function as a Service)
- Best practice is each service in its own isolated environment, e.g., Docker container.
- But a container can run multiple services, or an entire application.
- Everything at Google runs in a container
  - 4 Billion containers per week in 2018

Dynamically orchestrated

Orchestration

- Infrastructure as Code
- Automatic deployment and operation
- Optimize resource utilization dynamically

Data Analytics Pipeline



Design Pattern and Best Practice

- The Twelve-Factor App (http://12factor.net)

Codebase	Dependencies	Config
Backing services	Build, release, run	Processes
Port binding	Concurrency	Disposability
Dev/prod parity	Logs	Admin processes

Design Strategies

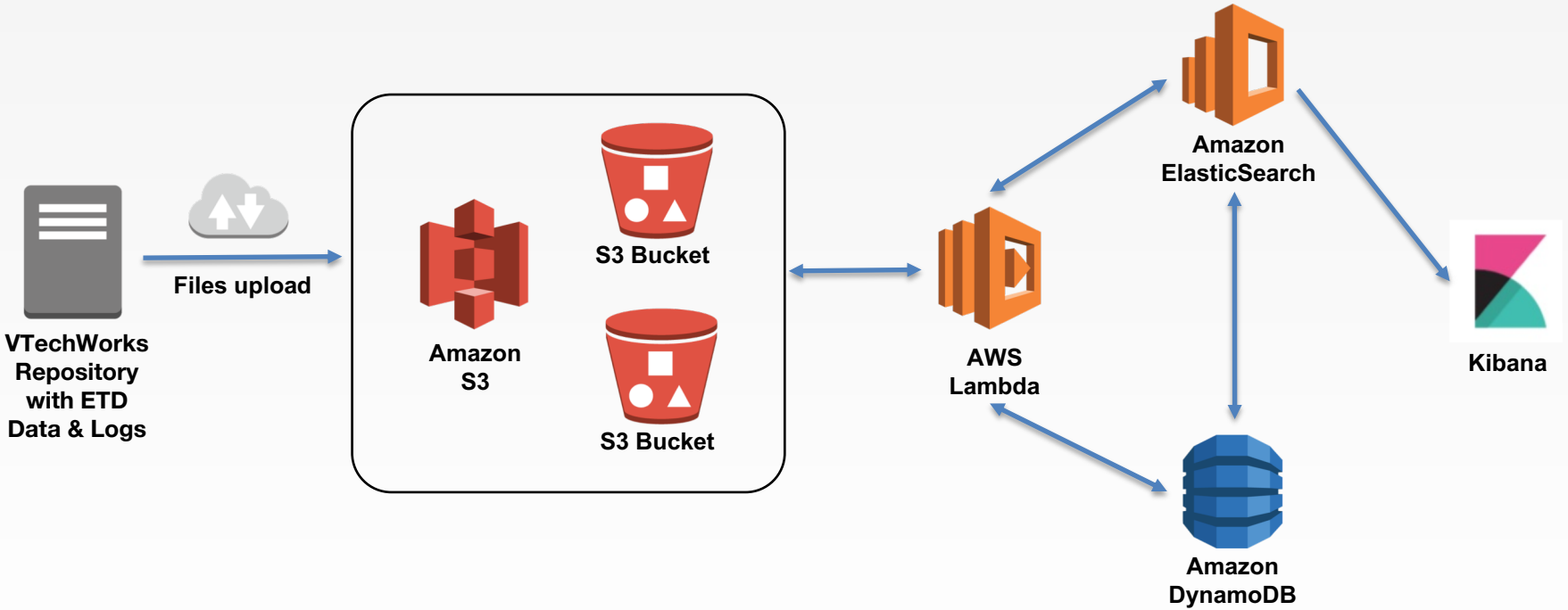
- Microservices
- Containers
- Orchestration

- Serverless (managed service)
- Scalability
- Automation
- Optimization (resource usage, cost, etc.)

Agenda

- Introduction
- Monolithic vs. Cloud-native applications
- Cloud-native approach
- ***Data analytics architecture***
- Future work

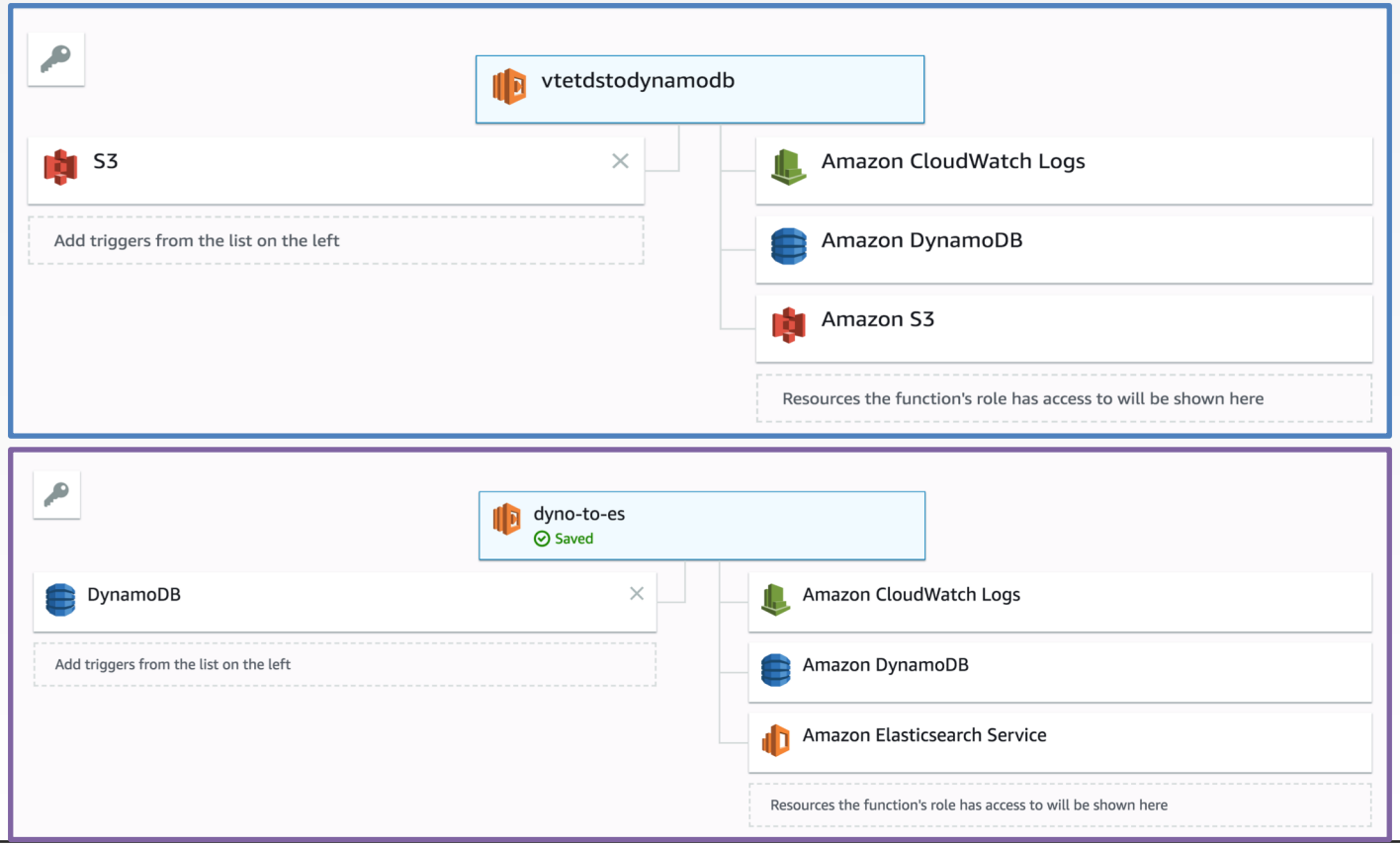
Application Architecture Overview



AWS Services

- AWS S3: Object storage in the cloud
- AWS Lambda: Serverless compute platform for stateless code execution in response to events
- Amazon Dynamodb: Fully managed NoSQL database service
- Amazon ElasticSearch: Search engine based on Lucene
- Kibana: An open source data visualization plugin for ElasticSearch

Microservice – Using AWS Lambda



Metadata Transformation Using AWS Lambda

```
9 43029,10919/9291||10919/71751,,,,,"Ketene, Alperen Nurullah",,,"Agah, Masoud",,,,,,"Behkam, Bahareh||Schmelz,
  Eva M.",,Mechanical Engineering,,,,5/31/11,,,5/6/11,5/31/12,,5/18/11,,,,"According to the American Cancer
  Society, Cancer is the second most common cause of death in the United States, only exceeded by heart
  disease. Over the past decade, deciphering the complex structure of individual cells and understanding the
  symptoms of cancer disease has been a highly emphasized research area. The exact cause of Cancer and the
  genetic heterogeneity that determines the severity of the disease and its response to treatment has been a
  great challenge. Researchers from the engineering discipline have increasingly made use of recent
  technological innovations, namely the Atomic Force Microscope (AFM), to better understand cell physics and
  provide a means for cell biomechanical profiling.
```

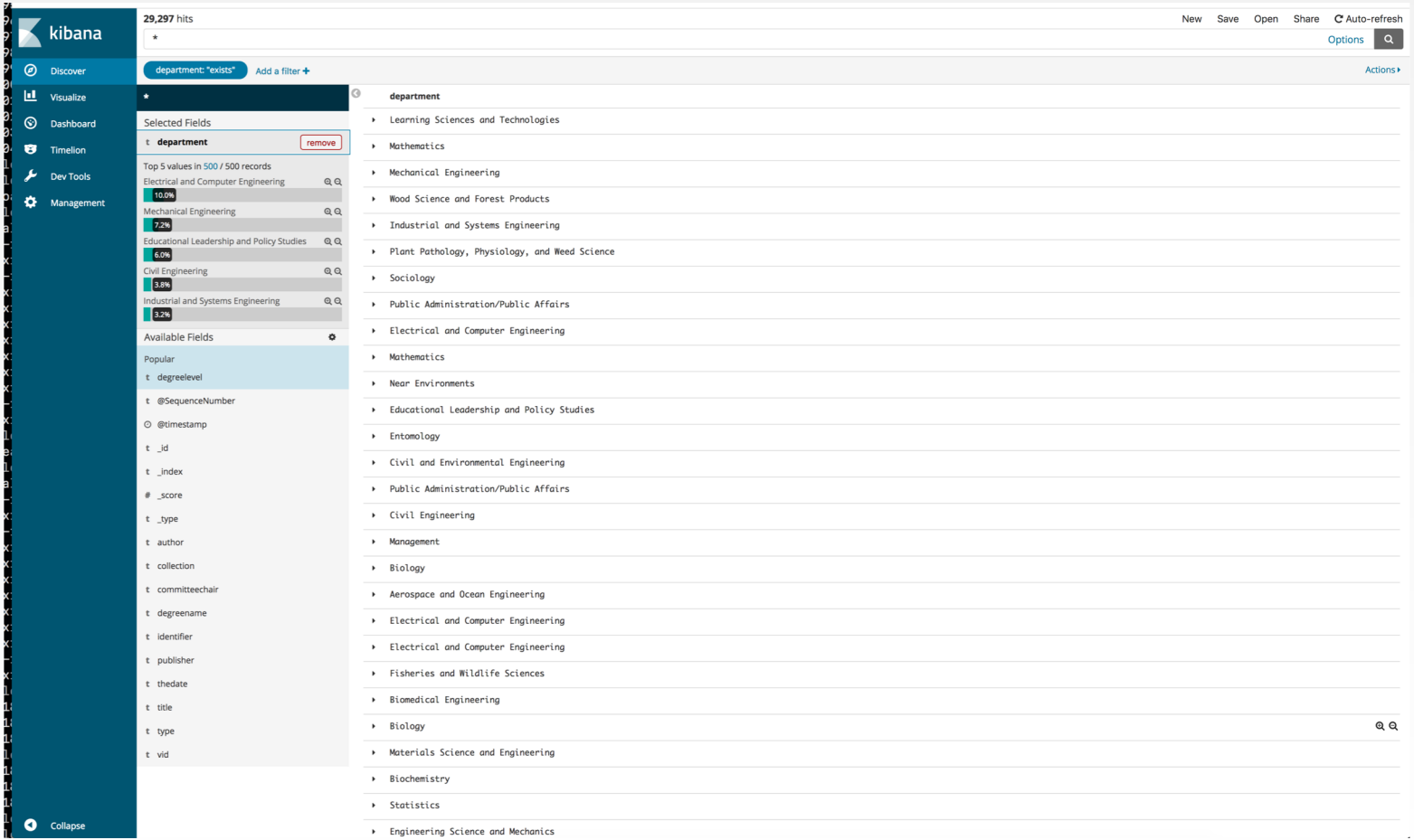

```
{
  "author": "\"Alperen Nurullah Ketene\"",
  "collection": "10919/9291||10919/71751",
  "committeechair": "\"Masoud Agah\"",
  "degreelevel": "masters",
  "degreename": "Master of Science",
  "department": "Mechanical Engineering",
  "identifier": "etd-05182011-152552",
  "publisher": "Virginia Tech",
  "thedate": "5/6/11",
  "title": "The AFM Study of Ovarian Cell Structural Mechanics in the Progression of Cancer",
  "type": "Thesis",
  "vid": "43029"
}
```

Log Data Transformation Using AWS Lambda

```
"2018-01-30T08:13:13.200Z","192.168.2.6","prod_log: 202.248.84.158 vtechworks.lib.vt.edu
[30/Jan/2018:03:13:15 -0500] GET
/bitstream/handle/10919/33268/WGraf_Thesis_2005.pdf?sequence=1 HTTP/1.1   Mozilla/5.0
(Windows NT 6.3; WOW64; Trident/7.0; rv:11.0) like Gecko"
```


```
{
    "website": "vtechworks.lib.vt.edu",
    "client_IP": "202.248.84.158",
    "http_version": "HTTP/1.1",
    "http_method": "GET",
    "timestamp": "2018-01-30T08:13:13.200Z",
    "web_client": "Mozilla/5.0",
    "request_uri": "/bitstream/handle/10919/33268/WGraf_Thesis_2005.pdf?sequence=1"
}
```

Visualization – Kibana



VIRGINIA TECH

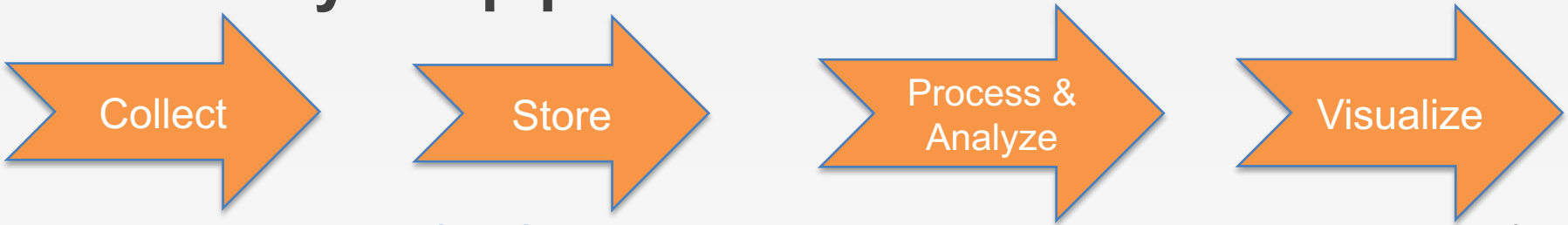
Results

- Facilitate application development
  - Parallel development and deployment
  - Switch services or techniques more flexibly
- Decouple the data analytics pipeline
  - More complex architecture and testing setup
  - More things to learn in order to select right tools
- Delegate maintenance tasks to cloud providers

Agenda

- Introduction
- Monolithic vs. Cloud-native applications
- Cloud-native approach
- Data analytics architecture
- ***Future work***

Data analytics pipeline in AWS



Real-time
Amazon Kinesis Firehose

Data Import
Amazon Import/Export Snowball

Message Queuing
Amazon SQS

Web/app Servers
Amazon EC2

Object Store
Amazon S3
Amazon Glacier

Real-time
Amazon Kinesis Streams

RDBMS
Amazon RDS

NoSQL
DynamoDB

Search
Amazon CloudSearch

IoT
Amazon IoT

Hadoop Ecosystem
Amazon EMR

Real-time
AWS Lambda
Amazon Kinesis Analytics

Data Warehousing
Amazon Redshift

Machine Learning
Amazon Machine Learning

Elastic Search Analytics
Amazon ElasticSearch

Process & Move Data
AWS Data Pipeline

Business Intelligence & Data Visualization
Amazon QuickSight

Elastic Search Analytics
Amazon ElasticSearch

Blue: current
Green: alternate
Black: other possible

Cloud Native Data Analysis Platform in AWS

IAM

AWS Organizations

Security & Identity

AWS Lambda

Amazon DynamoDB

Amazon ElasticSearch

Compute & Database

Amazon S3

Amazon Glacier

Storage

Kibana

*Amazon Athena

*Amazon Machine Learning

Analytics

AWS CLI

AWS CloudFormation

AWS CloudTrail

AWS Trusted Advisor

Amazon CloudWatch

AWS Config

Management

Other Cloud Platforms

- Amazon Web Services (AWS) - done
- Google Cloud Platform (GCP) – easily done
- Microsoft Azure, etc. – also possible

AWS	GCP	Azure
Elastic Compute Cloud	Compute Engine	Virtual Machines
Elastic Beanstalk	Google App Engine	Cloud Services
EC2 Container Service Kubernetes (EKS)	Kubernetes Engine	Container Service (AKS)
Lambda	Cloud Functions	Functions
Simple Storage Services	Cloud Storage	Storage
Virtual Private Cloud	Virtual Private Cloud	Virtual Network

Q & A

Supported by
Virginia Tech Libraries and
AWS Cloud Credits for Research program

Thank You!