Lessons from the trenches

Improving response by being “data wrangling” amateurs in AWS

Swetha Balla, BSides Budapest, May 2021

Agenda

- Challenge with IR in AWS
- Logs
- Data wrangling
- Metrics
- Visualisation
- Questions

How many “things” do you need to improve incident response in AWS?

Just one! “Data wrangling”. (with a lot of caveats!)

Lot of caveats (logs, damn’ed logs!)

- ~~The security team want me to enable ALL logs!~~ Have I enabled key logs?
- ~~My SIEM costs are high!~~ How do I store these logs?
- ~~I need to transfer logs off-platform to create dashboards!~~ How can I visualise logs?

Amazon CloudTrail

VPC

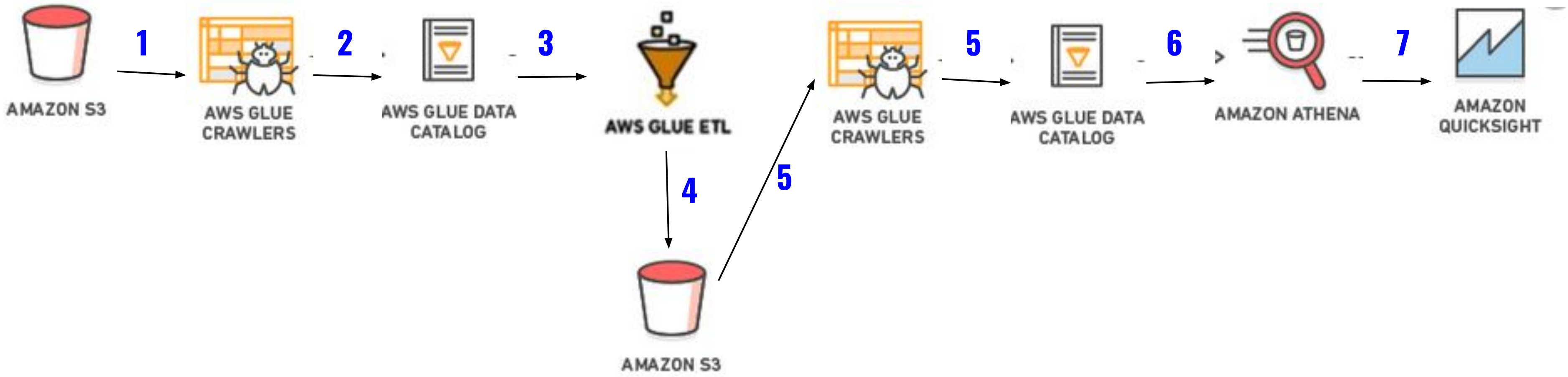
amazon S3

ELB

Amazon CloudWatch

AWS Config

Data wrangling



1. AWS Glue “Crawler” reads the CloudTrail logs from S3 (.json.gz)
2. Metadata for CloudTrail JSON logs added to AWS Glue “Tables”. (**Note**: It’s important to change the Data Type for the fields “*requestparameters*” and “*responseelements*” to string. By default, Glue sets them to struct)
3. AWS Glue ETL “jobs” convert JSON files to Parquet.
4. Converted parquet data stored in S3.
5. AWS Glue “Crawler” reads the CloudTrail logs from S3 (.json.gz).
6. Query parquet data in Athena
7. Visualise logs in Quicksight

Pre-requisites

- IAM role with the right permissions.
    - “AwsGlueServiceRole” policy
    - Other perms are also required - e.g. access to the S3 bucket with logs
- CloudTrail logs are available in S3 (.json.gz).
- Patience, in case something fails.

Demo

Example metrics

Query	json (unpartitioned)	parquet

API Errors
SELECT eventTime, eventSource, eventName, errorCode, errorMessage, responseElements, awsRegion, userIdentity.arn,
sourceIPAddress, userAgent
FROM <<*table*>>
WHERE errorCode IN ('Client.UnauthorizedOperation' , 'Client.InvalidPermission.NotFound ' , 'Client.OperationNotPermitted' ,'AccessDenied')
ORDER BY eventTime DESC limit 25
	Run time: 17.25 seconds
	Data scanned: 1.53 GB
		Run time: 5.37 seconds
		Data scanned: 989.51 MB

Activity from malicious IP
SELECT eventTime, eventSource, eventName, awsRegion, userIdentity.arn, sourceIPAddress, userAgent
FROM <<*table*>>
WHERE sourceIPAddress = '5.205.62.253'
ORDER BY eventTime DESC limit 25
	Run time: 19.32 seconds
	Data scanned: 1.53 GB
		Run time: 2.98 seconds
		Data scanned: 9.05 MB

EC2 Instance enumerating S3
SELECT useridentity.principalid, eventsource, eventname, count(*) AS total
FROM <<*table*>>
WHERE useridentity.principalid LIKE '%:i-%' AND eventsource = 's3.amazonaws.com' AND eventname = 'ListBuckets'
GROUP BY useridentity.principalid,eventsource,eventname
ORDER BY total DESC limit 25
	Run time: 13.23 seconds
	Data scanned: 1.53 GB
		Run time: 1.19 seconds
		Data scanned: 38.87 MB

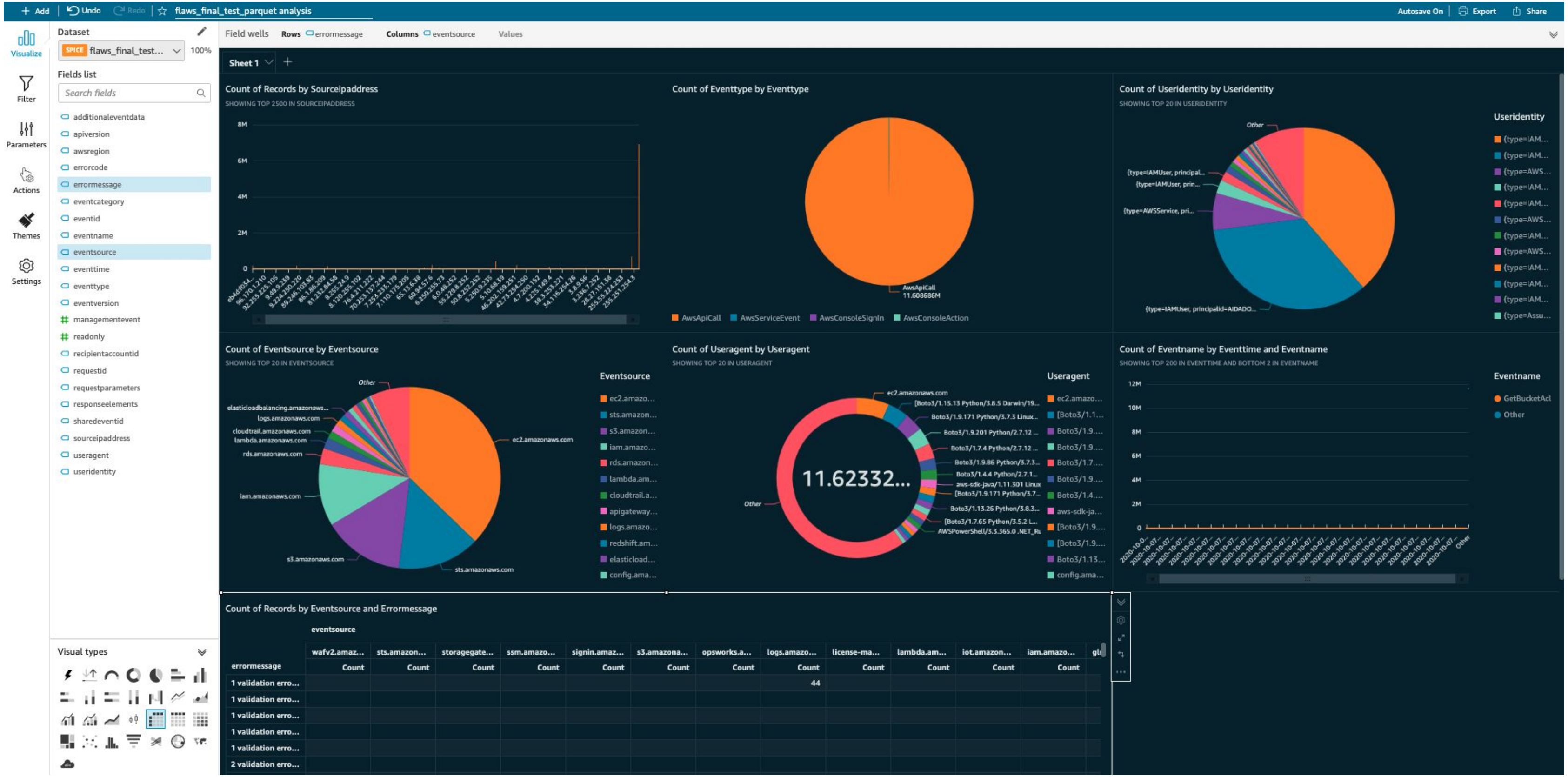
~ 74% less data scanned

~ 77% quicker

Parquet vs. JSON (unpartitioned) - query performance improvement

(admittedly, small data set!)

Example dashboard



Field wells Rows errormessage Columns eventsource Values

Count of Records by Sourceipaddress
SHOWING TOP 2500 IN SOURCEIPADDRESS

Count of Eventtype by Eventtype

Count of Useridentity by Useridentity
SHOWING TOP 20 IN USERIDENTITY

Count of Eventsource by Eventsource
SHOWING TOP 20 IN EVENTSOURCE

Count of Useragent by Useragent
SHOWING TOP 20 IN USERAGENT

Count of Eventname by Eventtime and Eventname
SHOWING TOP 200 IN EVENTTIME AND BOTTOM 2 IN EVENTNAME

Count of Records by Eventsource and Errormessage

	eventsource											
	wafv2.amaz...	sts.amazon...	storagegate...	ssm.amazo...	signin.amaz...	s3.amazona...	opsworks.a...	logs.amazo...	license-ma...	lambda.am...	iot.amazon...	iam.amazo...
errormessage	Count	Count	Count	Count	Count	Count	Count	Count	Count	Count	Count	Count
1 validation erro...								44				
1 validation erro...												
1 validation erro...												
1 validation erro...												
1 validation erro...												
2 validation erro...												

Take-aways

- ~~The security team want me to enable ALL logs!~~ Have I enabled key logs?
- ~~My SIEM costs are high!~~ How do I store these logs?
- ~~I need to transfer logs off-platform to create dashboards!~~ How can I visualise logs?

References

- Logging in the cloud: https://ponderthebits.com/wp-content/uploads/2020/02/Logging-in-the-Cloud-From-Zero-to-Incident-Response-Hero-Public.pdf
- Dataset: https://summitroute.com/blog/2020/10/09/public_dataset_of_cloudtrail_logs_from_flaws_cloud/ & http://summitroute.com/downloads/flaws_cloudtrail_logs.tar
- Example queries for AWS: https://github.com/easttimor/aws-incident-response
- AWS Glue:
    - https://aws.amazon.com/glue/
    - https://docs.aws.amazon.com/athena/latest/ug/glue-best-practices.html
    - https://aws.amazon.com/blogs/database/how-to-extract-transform-and-load-data-for-analytic-processing-using-aws-glue-part-2/
- AWS Athena: https://aws.amazon.com/athena/
- AWS Quicksight: https://aws.amazon.com/quicksight/
- Columnar data storage: https://docs.aws.amazon.com/athena/latest/ug/columnar-storage.html