Guppy

A Scalable Security Data Lake

Faisal Alghamdi, Senior Security Engineer, Saudi Aramco

Hafiz Farooq, Senior Security Architect, Saudi Aramco

Project Name – Guppy

Security Data Lake

Guppy Fish can stay in fresh water lakes adaptively for longer durations

Security Data Lake will also retain security big data for **extended durations**, before **correlation** & **analysis**

This project is named as **Guppy**, a tropical fish that resides in fresh water lakes

Security Data Lake
Code Name
Guppy

Presentation Agenda

- Guppy - Security Data Lake Architecture
- Why Security Data needs a Data Lake nowadays?
- Guppy Security Features
- Take Aways

This project is named as **Guppy**, a tropical fish that resides in fresh water lakes

Security Data Lake
Code Name
Guppy

Project Name
Guppy

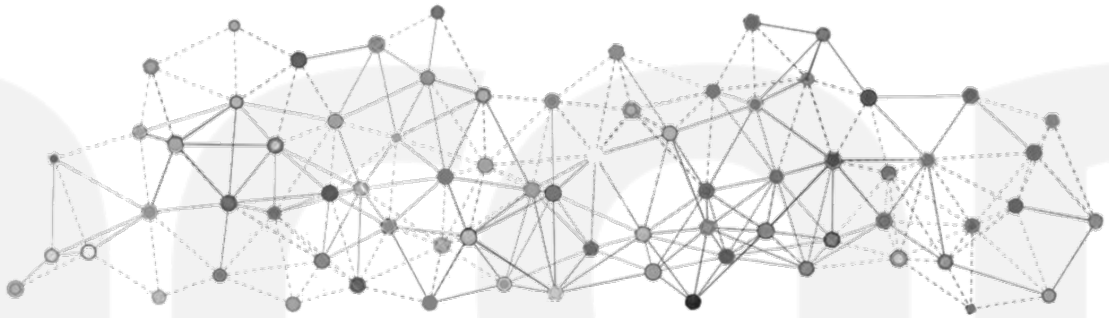
1 Why Security Data Needs a Data Lake?

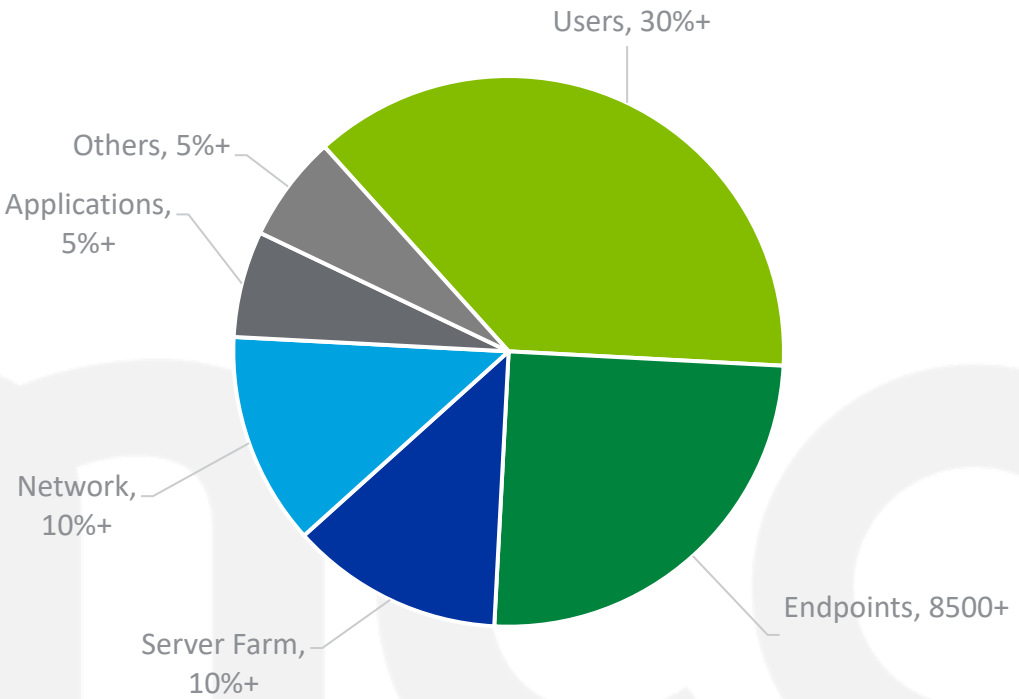
Best Practices & Guidelines

01/11/2023

Why Security Data Needs a Data Lake?

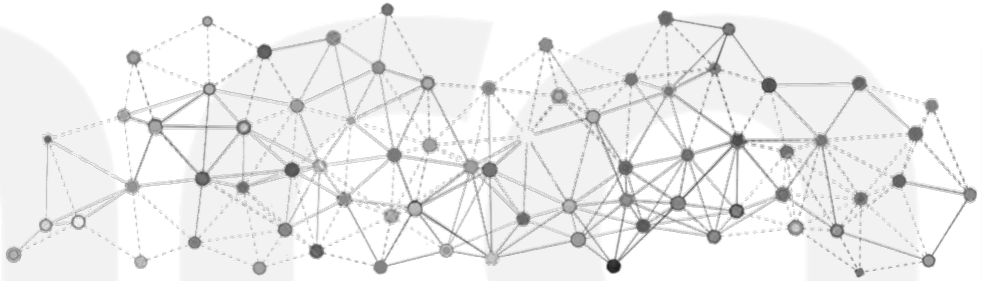
- Any Large Scale SOC currently monitors minimum **10,000+ IT Assets**
- Security Data volume in all larger enterprises is increasing exponentially, currently around **TeraBytes per day**
- Managing such a growing Big Data is a challenge

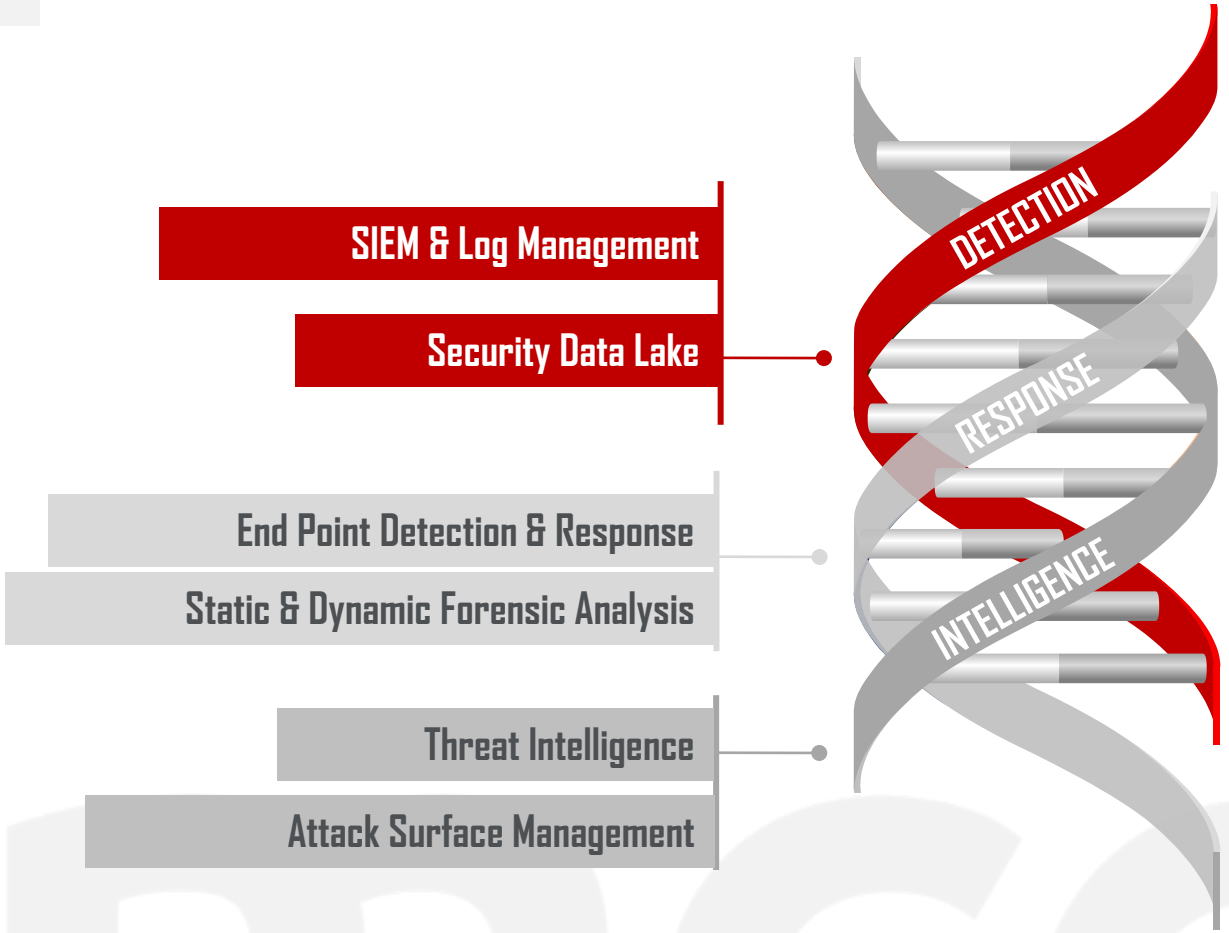




Generic SOC Eco System

- 3 Key SOC Requirements
  Detection, Response & Intelligence
- Detection via SIEM is not enough
  Data Lake is the Future
- Data Engineers are a must for SOC
  They are soon replacing SIEM engineers





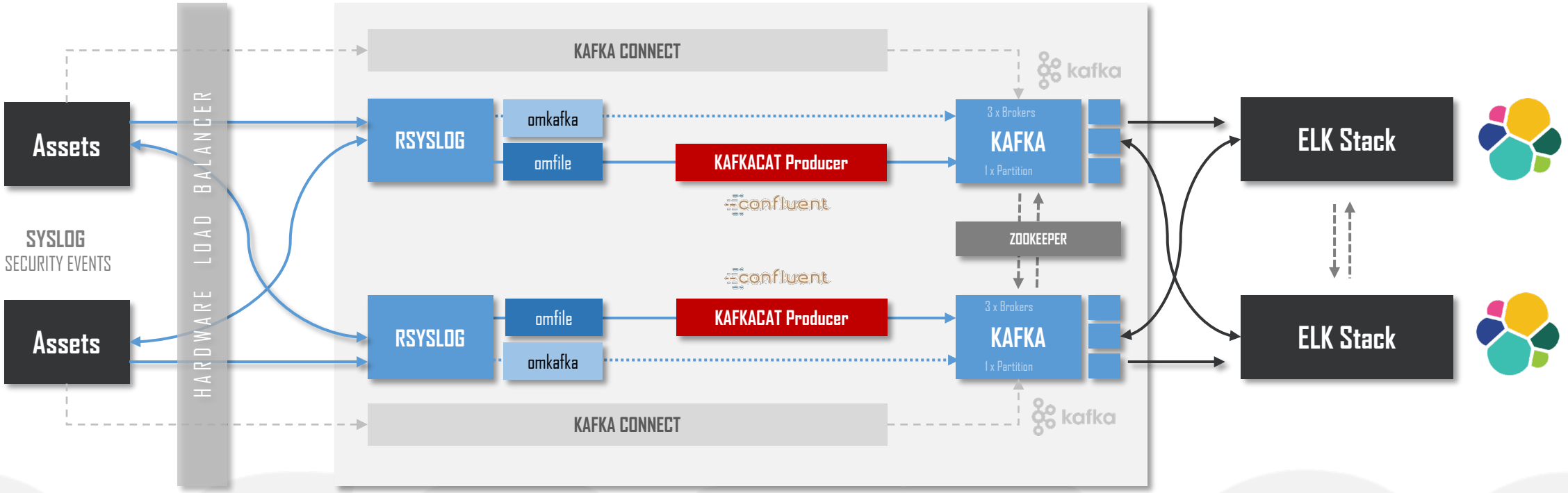
Project Name
Guppy

2 Security Data Lake Architecture

Project Code Name: **Guppy**

Streaming Ingestion

Project Name
Guppy



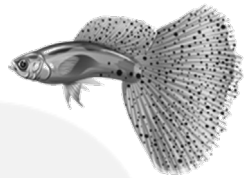
Log Ingestion Middleware

RSYSLOG
Fast | In-Memory Queues | QoS | TLS | Kafka Compatible

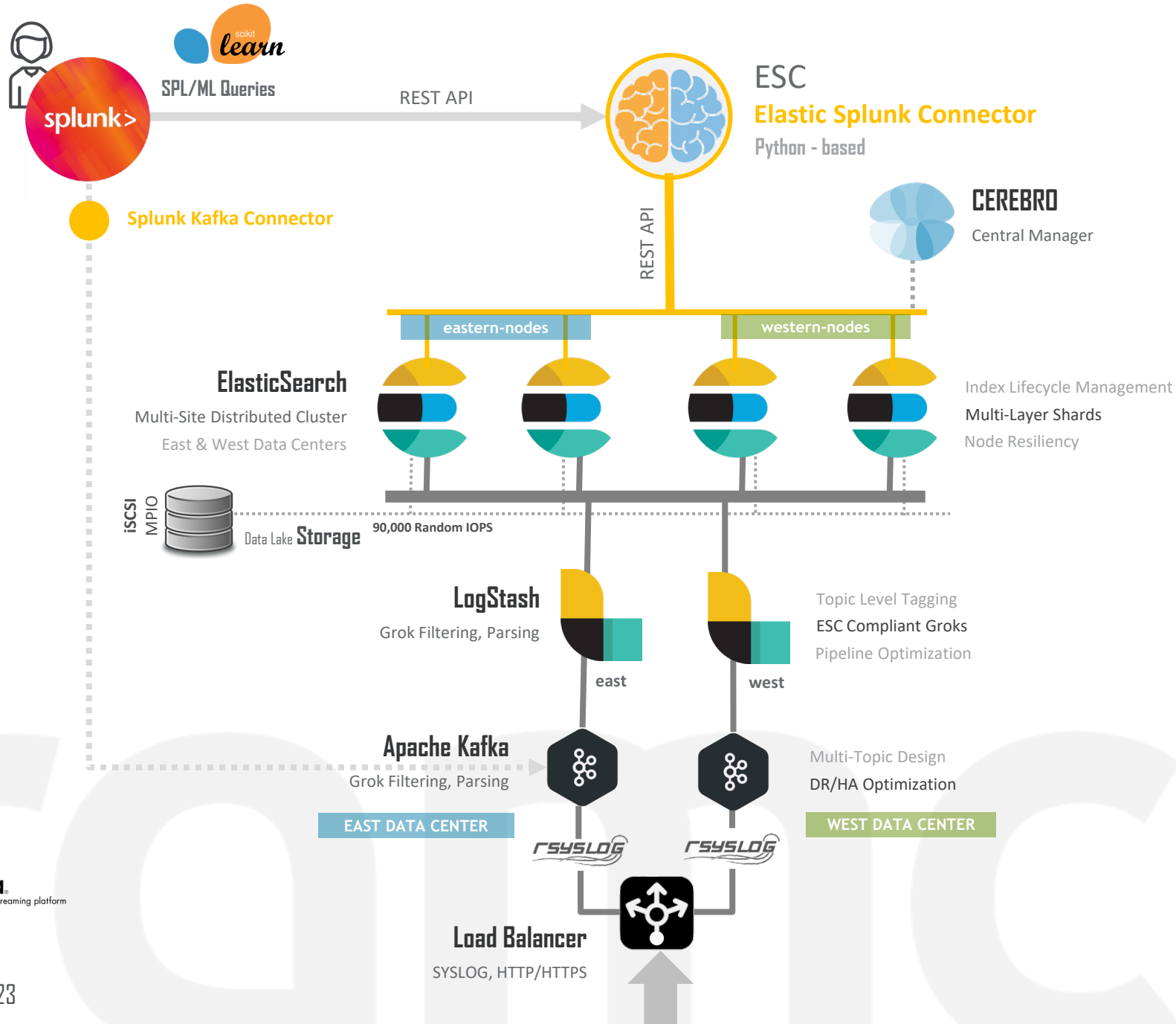
KAFKACAT
KAKFA Debugger | Producer | Consumer | Query Topics

ZOOKEEPER
Central Coordinator | Synchronization | Cluster Management

Guppy
Data Lake Architecture



GUPPY
A Scalable
Security Data Lake



Analytics
Processing
Streaming Ingestion

Project Name
Guppy

2 Guppy Security Features

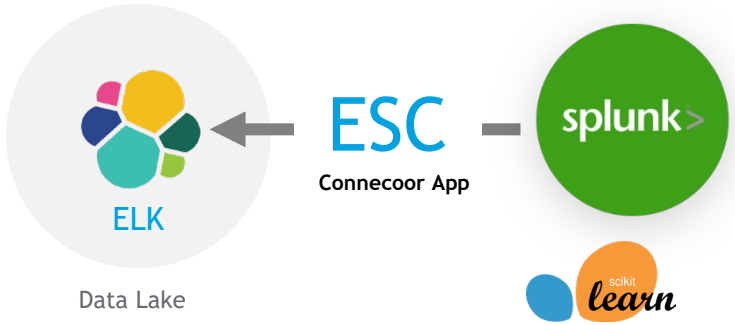
Project Code Name: **Guppy**

01/11/2023

Elastic - Splunk Connector ESC

Project Name
Guppy

- A custom connector, that queries **ElasticSearch** Data using **Splunk GUI**
- Much **easier** to search for SOC analyst
- No **Licensing** Overheads on Splunk
- Splunk **Machine Learning** capabilities can be applied on Data Lake events



Splunk ESC Query

```
| esc `ELK` action=search index=smtp.2022.12
latest=now earliest=now-1m query="message:*email*"
include_raw=true include_es=true
| stats count by message
| vader textfield=message full_output=true
```

Machine Learning Capabilities

Project Name
Guppy

Prediction | Numeric
Linear Regression
LASSO
Ridge, Kernel Ridge
ElasticNet
DecisionTree Regressor
RandomreForenst Regressor
Stochastic Gradient Descent

Clustering | Numerical
KMeans / Xmeans
DBSCAN
BIRCH
Spectral Clustering

Prediction | Categorical
Logistic Regression
Support Vector Machines
Bernoulli Naïve Bayes
Guassian Naïve Bayes
SGDClassifier
DecisionTree Classifier
RadomForenst Classifier

Feature Extraction
FieldSelector
PCA
Kernel PCA
TFIDF

Outliers | Categorical
OneClass SVM
AnomalyDetection (command)

Outliers | Numerical
OneClass SVM
Median, Mean, P25, P75

Preprocessing
StandardScalaer
AnalyzeFields (command)

Forecast | Time Series
ARIMA
Kalman Filter

Access to over 300
Open-Source Algorithms

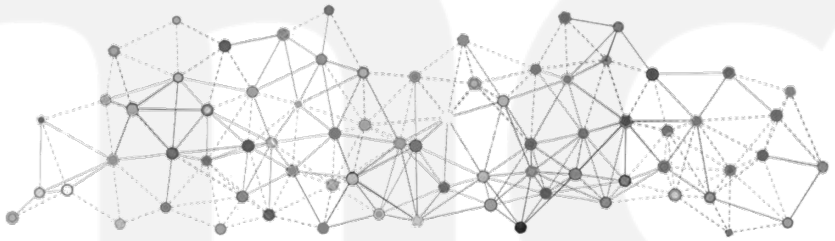
Elastic Common Schema (ECS)

- **Elastic Common Schema (ECS)** defines a standard naming convention for data ingested into Elasticsearch, helping you correlate data from diverse vendors and technologies

github.com/elastic/ecs

Without ECS	With ECS
src: 10.42.42.42 client_ip: 10.42.42.42 src_addr: 10.42.42.42	source.ip: 10.42.42.42

NORMALIZATION



Event Query Language (EQL)

- **Event Query Language (EQL)** for threat hunting and real-time detection with a simple syntax that helps SOC Analysts for writing complex queries
- Supports functions like join, unique, filter, head, sort, tail

```
network
| filter total_out_bytes > 100000000
| sort total_out_bytes
| tail 5
```

- Used for development of threat cases / correlation alerts

```
GET /my-data-stream/_eql/search
{ "query": """ library where process.name == "regsvr32.exe"
   and dll.name == "scrobj.dll" """ }
```

Project Name
Guppy

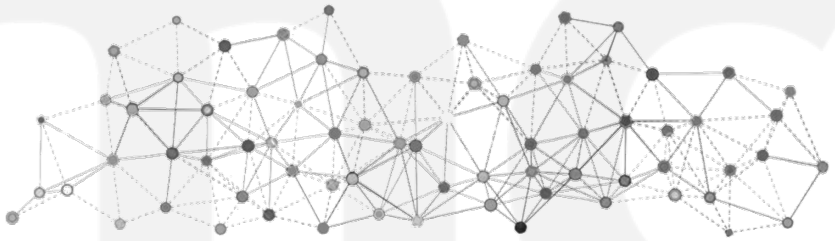
2 Take Aways

Project Code Name: **Guppy**

01/11/2023

Take Aways

- SOC require Security Data Lakes for **better visibility**
- Data Lake may be built, using **Open Source products**
- Data Lake should be utilized to keep **nice-to-have** security data
  i.e. Cloud API, Application Debug, OS System Logs
- Data Retention can be achieved through **low-cost storage**
- It will enable **Machine** & **Deep Learning** analysis of security data



Project Name – Guppy

Security Data Lake

Questions & Answers

Using Data to Defend

This project is named as **Guppy**, a tropical fish that resides in fresh water lakes

Security Data Lake

Code Name

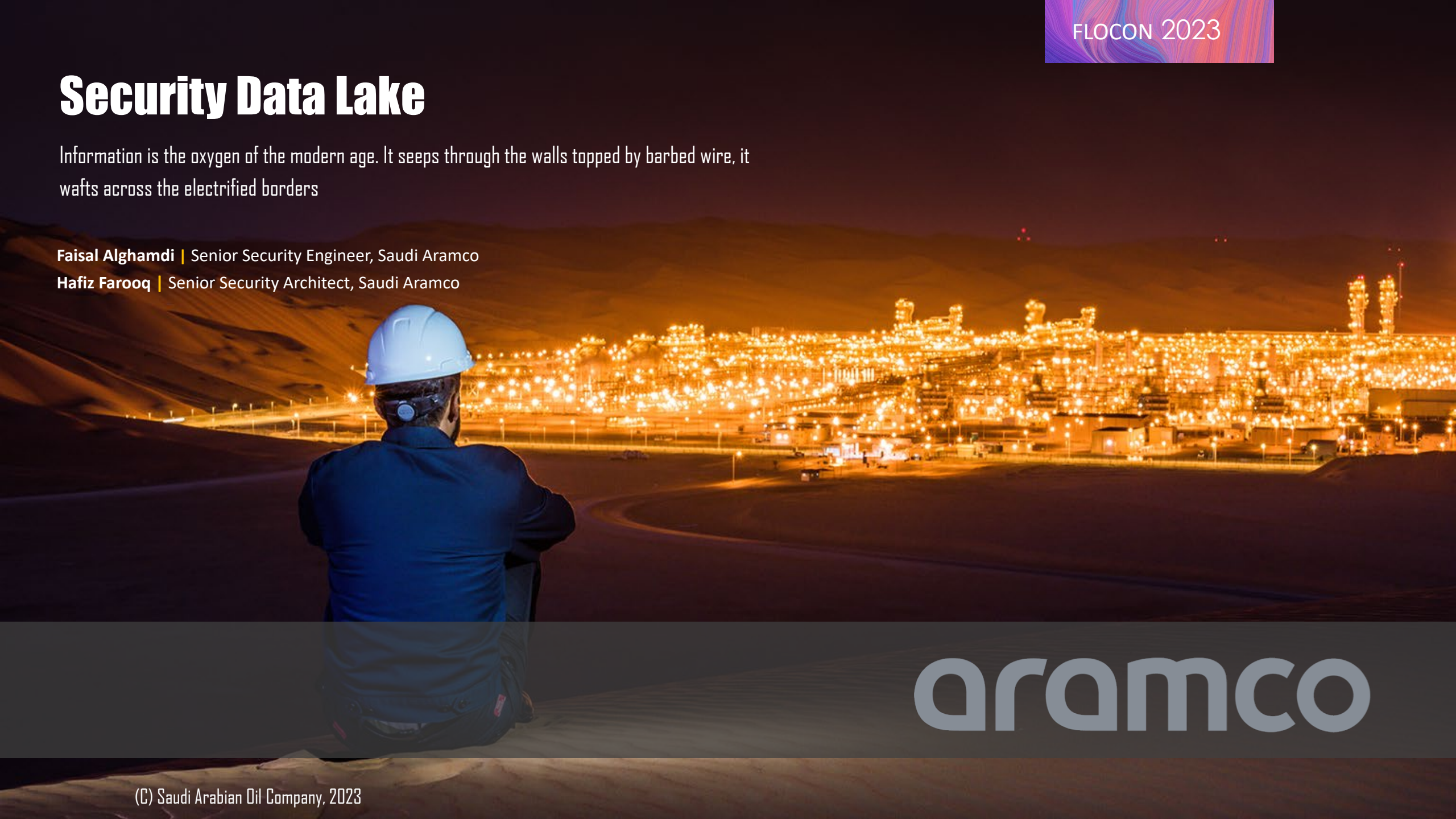
Guppy

FLOCON 2023

Security Data Lake

Information is the oxygen of the modern age. It seeps through the walls topped by barbed wire, it wafts across the electrified borders

Faisal Alghamdi | Senior Security Engineer, Saudi Aramco
Hafiz Farooq | Senior Security Architect, Saudi Aramco



aramco