EFFICIENTLY STANDING UP A CLOUD-BASED CYBERSECURITY DATA LAKE WITH MINIMAL RESOURCING

FLOCON 2023

ROSALIE BAKKEN, PHD; MATTHEW SPITZER, PHD; AND MS. JENNIFER MARR

PROJECT BACKGROUND

- Gaps in cybersecurity tools exist – no single tool provides 100% coverage of all data and events for complete awareness
- Options to help fill the gaps
  - Additional vendor solution(s)
    - Vendor solution functionality takes time to understand, compare, purchase, and deploy
    - Cost will vary, but involves initial purchase price plus ongoing support
    - Vendor tools are never fully customizable to an organization's unique situation
  - New in-house custom solution
    - Allows complete control over the solution, including future functionality
    - Expensive solution – costs, time to implement, ongoing resourcing

CUSTOM TOOL OPTION – TYPICAL SCENARIO

- Large projects must be proposed, funded, and managed
- Lengthy build process
- Laborious maintenance required
- Storage and compute power needed (processing and analytics)
  - On-premise components
  - Cloud-based components
- What specific use cases must the tool support?
- Does the delivered value of the tool exceed the level of investment?

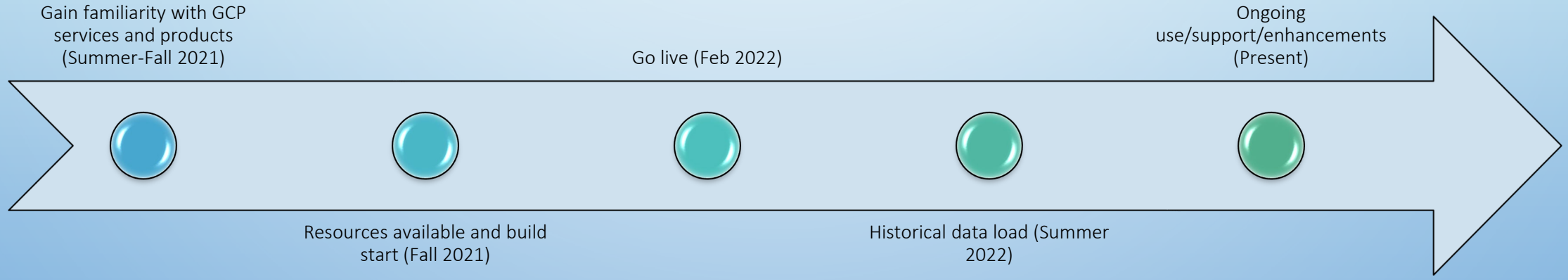
APPROACH TAKEN

- Indiscriminately filling the lake with all available security data *would not* achieve the goal of effective usability
- Focus on specific goals for use of the lake
  - Identify the highest priority use cases
  - Evaluate only those data sources required for these use cases, including data quality
  - Create an information model that clearly identifies each data field and how it relates to other fields from the other sources
  - Plan to store data in a way that will make it most searchable for the use cases
  - Create reliable, parameterized queries that can be re-used across multiple use cases

TIMELINE

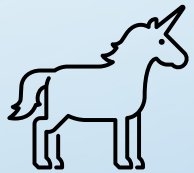
Build to go-live took 7 months and includes:

- Multiple disparate data sources (Trillions of rows/petabytes of data)
- Historical data load complete (~750,000 files)
- Integration and searchability (Clean data poised for analytics)
- Comprehensive monitoring in place (Alerts, gaps, signal loss)



TEAM COMPOSITION

- Small fusion team of 5 people
- Variety of backgrounds and a committed willingness to rise to the challenge as a team
  - Cybersecurity experience/certifications
  - Network architecture and operations
  - Software and script development
  - Database administration and operations
  - Statistical analysis of large datasets
  - Expertise across multiple operating systems and platforms
  - Source system familiarity
  - Technical writing and documentation
  - Artificial intelligence and machine learning



CHALLENGES

- Many challenges occurred during the build of the data lake
  - Paradigm challenges
    - Understanding the nature of working in the cloud – new expectations
    - Handling the unexpected within this context – understanding the limitations in control
    - Accomplishing our specific goals within a much larger organizational implementation as part of an over-arching strategy
  - Architectural challenges
    - On premise – consistently and robustly obtaining data from disparate sources with various owners
    - Cloud
      - Tenancy management for designing/consumption of cloud services
      - Initial data migration design
      - High Availability/Disaster Recovery
  - Data volume challenges
    - Transfer – managing data velocity and scheduling
    - Processing – quota management
    - Schema – partitioning strategy and signal loss monitoring

CLOUD MIGRATION EXPECTATIONS AND RISKS

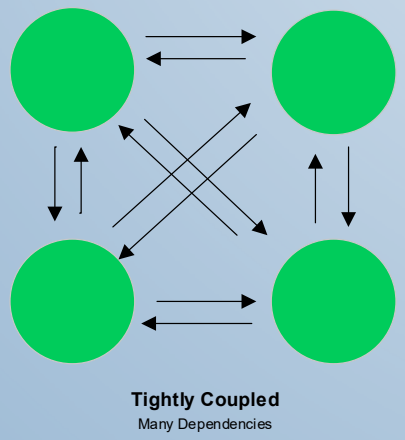
- Paradigm shift for technical, business, and operational activities
- Continual changes, consumers made aware through release notes

- Typical on-premise environment
- Changes are carefully made with multiple business units aware before, during, and after the change

- Focus on approach and clearly document needs from the environment
- Architect the project in a robust and resilient way, with ability to investigate issues caused by changes
  - Dig into changes to determine impacts
- Service provided by cloud provider is “different” - easier in some ways, but much more challenging in other ways

ROLE BINDINGS

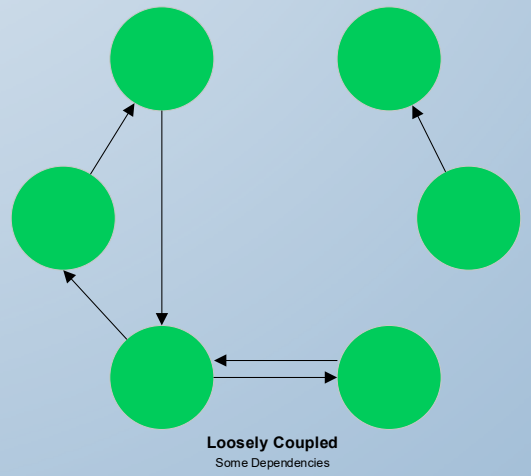
- Environment includes both custom and native roles
  - Workflows require roles not available in custom bindings
  - Increases complexity in support/maintenance efforts
- Roles included in updates to cloud services are not readily available in corresponding custom roles for use in our projects (high overhead)
  - Unable to validate new bindings to discern impacts
- Changes to custom bindings require coordination among teams sharing the same role definitions



Tightly Coupled
Many Dependencies

<u>Custom roles</u>

- Requires ongoing maintenance to keep in step with cloud provider changes



Loosely Coupled
Some Dependencies

<u>Mix of native and custom roles</u>

- Allows for change without extensive coordination

LESSONS LEARNED – BUILDING THE LAKE

- Pre-set guard rails for clarity
  - Technical and process security controls defined and in place
  - Staff were assigned focus areas for subject matter expertise
- The value of documentation
  - Decision log tool
  - Operational runbook
- Diligence in adhering to processes – reduce chaos to increase efficiency
  - Code repository management
  - Deployment processes
  - Issues tracking process

QUESTIONS?