Building a Modern, PCI-DSS Compliant Data Processing Platform

Eoin O'Flanagan - Lead Data Engineer, NewDay
Darragh McConville - Solution Architect, Kainos

Thursday 2nd May 2019
NewDay

- ~ 20yr Old Credit Card Issuer
- ~ 6m Customers
- Own-branded & Co-branded Products
- Digital Transformation Programme in mid-flight
Existing Data Technology Landscape

- **Challenges**
  - Scalability & Performance
  - Extensibility
  - Fine-grained data segregation
  - Data Science limited
  - No streaming support
Solution

- **User Features**
  - Modelled & Un-modelled Data
  - Self-service Reporting
  - Self-service Data Preparation
  - Exploratory Data Analysis
  - Built for Data Science

- **Technical Features**
  - PCI Compliant
  - Data Sensitivity Segregation
  - Metadata-driven
  - Stream & Batch Ready
  - Proportionate Cost & Chargeback

- **Guiding Principles**
  - Many self-contained systems over monolith ([http://scs-architecture.org/](http://scs-architecture.org/))
  - Open source over open source APIs over proprietary software
  - Serverless (FaaS, DBaaS) over CaaS over IaaS
Data Ingestion
Ingestion

- **Motivation**
  - Make all received datasets immediately accessible
  - Initial tranches of 100GB/day (EBCDIC, CSV, XML)
  - File sizes from 5MB – 30GB
  - PCI Protection for Personal Account Numbers (PANs)
Ingestion
Ingestion

- **Features**
  - Entirely event-driven and ephemeral
  - File-specific cost appropriation
  - Configuration-based throughput
  - Reversible PAN Tokenisation

- **Retrospective**
  - + Step Functions
  - + DynamoDB
  - - Lambda functions instead of Akka Actors
Data Streaming
Streaming

- **Motivation**
  - Provide a single ingress/egress interface for streaming data
  - Perform ETL on streaming data
Streaming

- **Features**
  - PII Data Filtering
  - PAN Scanner
  - Central store for Avro Schemas

- **Retrospective**
  - + Exactly once processing performance
  - - Use the Connectors
  - - Streams Logging & Instrumentation
Data Lake
Data Lake

- **Motivation**
  - Persist an immutable history of all ingested and derived data
  - Manage metadata in a central, searchable catalogue
  - Transform, verify and publish datasets
Data Lake

- **Features**
  - Schema-driven ETL
  - S3 as our source of truth
  - Glue Data Catalogue & Collibra for metadata

- **Retrospective**
  - + Large Parquet files not small Avro files
  - - AWS Glue ETL developer experience
  - - ETL micro-apps not ETL monoliths
  - - Data Pipeline Orchestration
Metadata-Driven
Data Presentation
Data Presentation

- **Motivation**
  - Provide query-optimized datasets
  - Provide authorised datasets to authorised consumers
  - Publish to heterogeneous targets (DB, S3, SFTP)

- **Features**
  - Account-isolated PCI datasets
  - Resource-isolated PII datasets
  - Presentation-layer Joins
Data Presentation

- **Features**
  - Separate storage and compute
  - S3 buckets as the source of truth
  - Departmental transient clusters

- **Retrospective**
  - + Spectrum Performance
  - - Redshift-Postgres DB Link Performance
  - - AD Integration
  - - Metadata Governance in AWS Glue
End-user Tooling
End-User Tooling

**Motivation**

- Offer the latest in data analysis and data science technologies
- Provide secure, audited access to data
- Support the broadest range of users from novice to expert
- Prefer Python as a query and ETL language
End-User Tooling

- **Features**
  - AD-driven resource assignment
  - Easy extensibility
  - Seamless scalability
  - User customisation

- **Retrospective**
  - + JupyterLab extensibility
  - - You get what you pay for (e.g. Superset)
  - - Security and usability will never be friends
  - - Excel integration is difficult without Microsoft
Key Takeaways

- **Batch your batches, stream your streams**
  - From Kappa to Lambda Architecture

- **Design for change as a first principle**
  - Wrong sooner rather than later

- **Build orthogonally**
  - Push data to users and build technical features

- **Bypass the network for bulk loads**
  - Snowballing 300TB was simple

- **Centralise data access management**
  - Unifying data access and security is challenging
Looking Ahead

- Unification of data preparation and access
- More fine-grained User Access
- Enabling Data-as-a-service
- Support an even wider range of tooling
We're living in a new era of constant sabotage, misinformation, and fear, in which everyone is a target, and you're often the collateral damage in a growing conflict among states. From crippling infrastructure to sowing discord and doubt, cyber is now the weapon of choice for democracies, dictators, and terrorists.

David Sanger explains how the rise of cyberweapons has transformed geopolitics like nothing since the invention of the atomic bomb. Moving from the White House Situation Room to the dens of Chinese, Russian, North Korean, and Iranian hackers to the boardrooms of Silicon Valley, David reveals a world coming face-to-face with the perils of technological revolution—a conflict that the United States helped start when it began using cyberweapons against Iranian nuclear plants and North Korean missile launches. But now we find ourselves in a conflict we're uncertain how to control, as our adversaries exploit vulnerabilities in our hyperconnected nation and we struggle to figure out how to deter these complex, short-of-war attacks.

David Sanger
The New York Times

David E. Sanger is the national security correspondent for the New York Times as well as a national security and political contributor for CNN and a frequent guest on CBS This Morning, Face the Nation, and many PBS shows.