From Data Discovery to Data Insight: How to Leverage Automation and Machine Learning

A conversation with

Nimrod Vax
Chief Product Officer
BigID

Peggy Tsai
VP, Data Solutions
BigID
Moderated by **Mike Meriton**
Co-Founder & COO, EDM Council

- Joined EDM Council full-time 2015 to lead Industry Engagement
- EDM Council Co-Founder & First Chairman (2005-2007)
- Former CEO GoldenSource (2002-2015)
- Former Executive for D&B Software and Oracle
- FinTech Innovation Lab – Executive Mentor (2011 – Present)
Speaker Profiles

Nimrod Vax
Co-Founder & Chief Product Officer, BigID

Nimrod has over 15 years of experience in enterprise security software and identity management serving in various capacities ranging from engineering, product and business management.

Peggy Tsai
VP of Data Solutions, BigID

Peggy has over 18 years of practitioner experience in data management, stewardship and governance in the financial services industry.
Agenda

1. Data Management Challenges
2. Data Discovery as a core foundation activity
3. How Data Discovery accelerates your data maturity level and Use Cases
4. Demo of Data Discovery in action
5. Q&A
The Stick versus the Carrot
Data Management Lifecycle

1. Scope of Project
2. Define Data Elements
3. Create Data Quality Rules
4. Collect Information
5. Access Business/Risk Impact
6. Monitor results
7. Repeat with different scope

The Result:
Slow to scale and manual to maintain
Data Knowledge is a Foundation for Data Use

People
- Data Stewards and Compliance Officers

Process
- Data Governance
  - Policy and Workflows

Technology
- Critical/Sensitive Data Discovery
  - Find, Inventory and Map
What Proceeds Good Data Lifecycle Management?

Before cataloging and classification...Know Your Data (not just your metadata)!
Data Growth Challenges

- More data in more places
- Difficult to identify critical data
- Difficulty to know where all that data is (dark data)
- Rise in redundant, duplicate, and similar data
- Silo’d data
Critical Data is Everywhere

No longer confined to your relational databases

Files
Unstructured File Shares
Structured Databases
Big Data & NoSQL
Applications
Messaging
Data Pipelines
Cloud / SaaS
Definitions of Sensitive Data are Evolving

- Personal data
- Sensitive data
- Protected data
- Regulated data
- Inferred data
- Related data
Classification & Categorization

The Traditional Approach

Pattern Matching
Regular Expressions
String Sequences
Dictionaries
Manual Tagging
Category Labels
Machine Learning & AI

Next Generation Classification for Today’s Data Reality

- ML-based Content Classification
- Neural Network based Classifiers
- Enrichment
- Cluster Analysis
- Correlation
- Intelligent Classification
Unified Data Inventory

- Automatically inventory and map all personal and sensitive data
- Single source of truth
- PI, PII, Sensitive data, metadata
Operationalizing Data Data Discovery

How Data Discovery Fits into DCAM - Data Assessment Maturity Models

**Catalog Physical Data**
- Capture metadata on data elements
- Establish Catalog aligned to Data Domains

**Locate, Document, and Inventory Repositories**
- Expand beyond databases to unstructured sources
- Logical to physical mappings

**Profile & Analyze Data**
- Profiled data scores embedded into discovery
- Continuous monitoring

**Govern Adherence to Contractual Terms**
- 3rd party data alignment and monitoring
- Enable policy enforcement
How Data Discovery Accelerates Data Management Program Maturity

Top 3 Use Cases:

1) Information Lifecycle Management - meet retention schedules
2) Cloud Migration Projects - manage data duplications
3) Security Access - identifying open access systems
Data Minimization

- **Minimize & Consolidate**
  Remove redundant data for IT savings & cloud migration

- **Reduce Risk**
  Backup or delete unnecessary data reducing risk exposure

- **Identify Originals**
  Identify original versions vs similar and duplicates
Cloud Migration

- **Inventory**
  Know what you have, where it is, and what's getting migrated

- **Clean up your data**
  Minimize duplicate and redundant data, & fix data quality issues

- **Reduce Risk**
  Align policy and tagging to consistently monitor on-prem and in the cloud
Highlight Vulnerable Data

- **Find open access**
  See exposed sensitive data

- **Find excess permissions**
  Identify over permissioned employee access

- **Reduce Risk**
  Reduce risk on sensitive data
Map Data to Process

- **Map to Business Processes**
  Document the systems relevant to business processes

- **Workflow Review**
  Assign owners to review and verify the data that flows between systems

- **Track data activity**
  Simplify and maintain processing activity and data usage including external data sharing
{ Business Use Case on Finding and Managing your Critical Data }
Data Integrations, Interoperability, Enhancements

- Alation technologies
- Collibra
- MANTA
- SAP

Broad Ecosystem of Strategic Partnerships

- SAP
- Microsoft
- IBM Security
- AWS
- Salesforce
- Immuta
- Collibra
- OneTrust
- asg technologies
- TrustArc
- ServiceNow
- IONIC
- RSA Archer
- CYBERARK
- Box
- MANTA
- databricks
- teradata
- Office 365
Where Data Discovery Fits into Your Data Architecture

External Sources

Data Registry - Acquisition
- Catalog
- Classify

Collection & Analysis
- Correlate
- Cluster

Provisioning
- Access Control

Search Engine
Identify and Pinpoint

Governance
Metadata / Data Quality / Data Access / Retention & Remediation
Questions?