ARMA METRO NYC CHAPTER

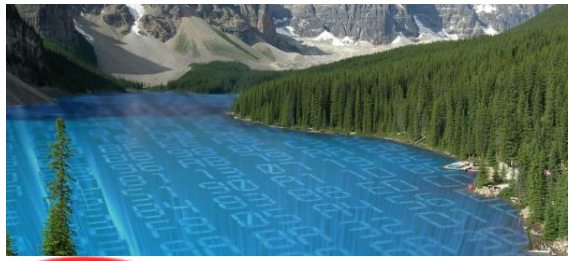
ARMA Northern New Jersey | ARMA International | Long Island Chapter | ARMA International Connecticut Chapter

BIG DATA AS AN ASSET: USING IG TO NAVIGATE THE DATA LAKE AND ITS SECURITY PITFALLS

Galina Datskovsky, Ph.D., CRM, CEO, Vaporstream
Ronald J. Hedges, J.D., Vaporstream Advisory Board

Learning Objectives

- The emerging world of a data lake
- Legal and operational risks of retaining data in a large repository such as a data lake
- Security Considerations
- Identify ways to leverage Big Data as an asset consistent with relevant Information Governance Principles



Northern New Jersey

Long Island Chapter

ARMA International ARMA Connecticut Chapter In the Heart of New England

ARMA METRO NYC CHAPTER

“Data Lake” Defined

- “A data lake is a large object-based storage repository that holds data in its native format until it is needed.”
  - searchaws.techtarget.com/**definition/data-lake** (last visited July 25, 2015)
- Gartner’s definition:

  “A data lake is a collection of storage instances of various data assets additional to the originating data sources. These assets are stored in a near-exact, or even exact, copy of the source format. The purpose of a data lake is to present an unrefined view of data to only the most highly skilled analysts, to help them explore their data refinement and analysis techniques independent of any of the system-of-record compromises that may exist in a traditional analytic data store (such as a data mart or data warehouse).”
  - http://www.gartner.com/it-glossary/data-lake (last visited July 31, 2015)

ARMA Northern New Jersey | ARMA International | Long Island Chapter | ARMA International ARMA Connecticut Chapter In the Heart of New England | ARMA METRO NYC CHAPTER

Further Definition

- While a hierarchical data warehouse stores data in files or folders, a data lake uses a flat architecture to store data. Each data element in a lake is assigned a unique identifier and tagged with a set of extended metadata tags
- When a business question arises, the data lake can be queried for relevant data, and that smaller set of data can then be analyzed to help answer the question
- The term data lake is often associated with Hadoop-oriented object
  - *storagehttp://searchaws.techtarget.com/definition/data-lake (last visited 7/27/15*

Beware Of The Data Lake?

- “While the marketing hype suggests audiences throughout an enterprise will leverage data lakes, this positioning assumes that all those audiences are highly skilled at data manipulation and analysis, as data lakes lack semantic consistency and governed metadata”
- “Gartner Says Beware of the Data Lake Fallacy” (Press Release) (July 28, 2014)
  http://www.gartner.com/newsroom/id/2809117 (last visited July 25, 2015)

The Data Lake In Practice

- Some questions related to a data lake:
  - Who decides what goes in?
  - What goes in?
  - How is “content” organized?
  - Who has access rights and how do you secure information and resulting objects?
  - Chain of custody?
  - How long is the data really useful?

Applying The Relevant Principles

- Integrity
  - Provenance
  - Chain of Custody
- Protection
  - Security (data may be quite disparate)
  - Privacy
  - Personally identifiable information (PII)
- Compliance
  - What are the applicable regulation for the raw data and the end product?
  - Discovery and Production

Northern New Jersey | Long Island Chapter | ARMA International Connecticut Chapter In the Heart of New England | ARMA METRO NYC CHAPTER

Applying The Relevant Principles

- Availability
  - Access Rights
  - Available Metadata
  - System longevity
- Transparency
  - Discovering the content and providing appropriate response
- Retention
  - Length of useful life
  - Data migration
- Disposition
  - When, what and how

BIG Data and BIG Security

- Data Origin
- Original Security
- Securing the combined blobs
- Maintaining access controls

Making The Data Lake Consistent With Good Governance

- Make sure you are in the conversation
- Know what data is going into the lake and point out policy considerations consistent with the principles just covered
- Don’t dwell on data retention – data will certainly overstay the original retention schedule
- Emphasize Data Security
- Focus on other principles that will allow the business easy access and ethical use

Making the data lake consistent with good governance (Continued)

- Don’t try to stop the inevitable
- Be a business enabler, therefore focus on anything other than disposition
- Manage the lake as a big blob
- Manage the mined results in accordance with data usage
- Notify Legal Counsel of new data location
- Discovery of results as well as raw data is an important consideration

Resources for "Data Lake" Slides

- "The Principles of the Business Data Lake" (Capgemini: undated) http://pivotal.io/big-data/white-paper/the-principles-of-the-business-data-lake (last visited July 27, 2015)
- Gartner IT Glossary http://www.gartner.com/it-glossary/data-lake (last visited July 31,2015)
- M. Fowler, "DataLake" (Feb. 5, 2015) (available at http://martinfowler.com/bliki/DataLake.html (last visited July 27, 2015)
- "Data Miming from A to Z: Better Insights, New Opportunities" (SAS: undated) (available at http://www.sas.com/en_us/whitepapers/data-mining-from-a-z-104937.html (last visited July 27, 2015)
- B. Stein & A. Morrison, "The Enterprise Data Lake: Better Integration and Deeper Analytics" (PWC: June 1, 2014) (available at http://www.pwc.com/en_US/us/technology-forecast/2014/cloud-computing/assets/pdf/pwc-technology-forecast-data-lakes.pdf) (last visited July 27, 2015)

QUESTIONS?

COMMENTS?

THANK YOU!

Contact Information

Galina Datskovsky, Ph.D., CRM, Vaporstream CEO

galina.datskovsky@vaporstream.com

Ronald J. Hedges, J.D., Principal of Ronald J. Hedges, LLC and Vaporstream Advisory Board

r_hedges@live.com