Data Lakes – The Key to a Scalable Data Architecture

May 24th, 2017

Ben Sharma | CEO
ben@zaloni.com
Enabling the data-powered enterprise

1. Industry-leading enterprise data lake management, governance and self-service platform
2. Expert data lake professional services (Design, Implementation, Workshops, Training)
3. Solution-based packaged offerings to simplify implementation and reduce business risk
Data lakes are central to the modern data architecture

- Store all types of data in its raw format
- Create Refined, Standardized, Trusted datasets for various use cases
- Store data for longer periods of time to enable historical analysis
- Query and access data using a variety of methods
- Manage streaming and batch data in a converged platform
- Provide shorter time-to-insight with proper management and governance
Data architecture modernization

Traditional

Sources → ETL → EDW → Data Discovery

Modern

Streaming → Data Lake → EDW → Data Science

Unstructured Data → Data Lake → Discovery Sandbox

Various Sources → Data Lake → Data Discovery
## Zaloni Big Data Maturity Model

<table>
<thead>
<tr>
<th>Stage:</th>
<th>Ignore</th>
<th>Store</th>
<th>Manage</th>
<th>Automate</th>
<th>Optimize</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Characteristics:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emphasis on structured data</td>
<td>Hadoop on premises or in the Cloud</td>
<td>Acquire useful data from across the enterprise</td>
<td>Self-Service Ingestion &amp; Provisioning</td>
<td>Self-improving data lake via machine learning algorithms</td>
<td></td>
</tr>
<tr>
<td>Limited ability to leverage data at scale</td>
<td>Limited visibility and usability of data</td>
<td>Improved visibility and understanding via managed Ingestion of data and metadata</td>
<td>360 View of Customer, Product, etc</td>
<td>True democratization of big data and analytics</td>
<td></td>
</tr>
<tr>
<td>Business emphasis on retrospective reporting and analysis</td>
<td>Limited corporate oversight &amp; governance</td>
<td>Ensure security and privacy of sensitive data</td>
<td>Enterprise Data Discovery</td>
<td>Intelligent data remediation and curation</td>
<td></td>
</tr>
<tr>
<td>Strong governance and security policies</td>
<td>Sandbox or Dev Environments</td>
<td>Operationalize data at scale</td>
<td>Operationalize analytical models into business fabric</td>
<td>Recommended Data Security, and Governance policies</td>
<td></td>
</tr>
<tr>
<td><strong>Value Realized:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Business Impact:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slow to accommodate business changes</td>
<td>Ad-hoc and exploratory insights for individual use cases</td>
<td>Scalable production data lake for new and improved business insights</td>
<td>Enables immediate data impact on business operations</td>
<td>Lights out business operations optimized for business success</td>
<td></td>
</tr>
<tr>
<td><strong>Descriptor:</strong></td>
<td>Data Warehouse</td>
<td>Data Swamp</td>
<td>Managed Data Lake</td>
<td>Responsive Data Lake</td>
<td>Self-Organizing Data Lake</td>
</tr>
<tr>
<td><strong>Stage Today:</strong></td>
<td>66% of market</td>
<td>22% of market</td>
<td>10% of market</td>
<td>2% of market</td>
<td>0% of market</td>
</tr>
</tbody>
</table>
Data Lake Reference Architecture

**Transient Landing Zone**
- Temporary store of source data
- Consumers are IT, Data Stewards
- Implemented in highly regulation industries

**Raw Zone**
- Original source data ready for consumption
- Consumers are ETL developers, data stewards, some data scientists
- Single source of truth with history

**Trusted Zone**
- Standardized on corporate governance/ quality policies
- Consumers are anyone with appropriate role-based access
- Single version of truth

**Refined Zone**
- Data required for LOB specific views - transformed from existing certified data
- Consumers are anyone with appropriate role-based access

**Sandbox**
- Enables ad-hoc, exploratory analytics, experimentation
- Consumers are anyone with appropriate role-based access

**Data Lake**
Data Lake 360°: Zaloni’s integrated platform for data lakes

- Leverage the full power of a scale-out architecture with an actionable, scalable data lake
  
  1. Enable the lake

- Improve data visibility, reliability and quality to reduce time-to-insight

  2. Govern the data

- Safeguard sensitive data and enable regulatory compliance

- Foster a data-driven business through self-service data discovery and preparation

  3. Engage the business
Data Governance

1. Based on a foundation of metadata management
2. Lightweight and distributed
3. Hybrid – top down and bottom up approach

Gartner Data and Analytics 2017
Metadata Management

- Central to a well-managed data lake – provides visibility, reliability and enables data governance
- Capture and manage operational, technical and business metadata
- Reduced time to insight for analytics
- Types of metadata:
  - Where it resides, and how was it ingested
  - What it means, and how it should be interpreted
  - What governance policies apply to it
  - What it's worth, and how its value can be expressed
  - Who it's accessed and consumed by
  - Which business processes downstream consume it
Metadata Exchange Framework

1. Metadata sharing is critical for an integrated approach
2. Federated approach for metadata collection
3. Two way metadata exchange between the Data Lake and other Enterprise repositories
Managed Ingestion

- Ability to ingest vast amounts of data
- Ability to handle a wide variety of formats (streaming, files, custom) and sources
- Build in repeatability via automation to pick up incoming data and apply pre-defined processing
Data Lineage

• See how data moves and how it is consumed in the data lake
• Safeguard data and reduce risk, always knowing where data has come from, where it is, and how it is being used
Data Quality

- Rules based data validation
- Integration with the managed data pipeline
- Stats and metrics for reporting and actions
- Automation, Remediation, Notifications

Future:
- ML based classification
Data Security and Privacy

- Secure infrastructure for data in motion and data at rest
- Role based access control for Metadata and the data
- Mask or tokenize data before published in the lake for consumption
- Audit, access logs, alerts and notifications
Data Lifecycle Management

1. Hot -> Warm -> Cold on an entity level based on policies/SLAs

2. Provide data management features to automate scheduling and orchestration of data movement between heterogeneous storage environments

3. Across on-premise and cloud environments
Data Catalog

• See what data is available across your enterprise
• Contribute valuable business information to improve search and usage
• Use a shopping cart experience to create sandbox for ad-hoc and exploratory analytics
Self-service Data Preparation

- Blend data in the lake without a costly IT project
- Perform interactive data-driven transformations
Considerations for data lake in the cloud

- How do you create a cloud agnostic data lake platform?
- How do you deploy a cost-effective compute layer?
  - Elastic compute layer
  - Batch and near real-time
- How do you optimize storage?
  - Support polyglot persistence
  - Data Lifecycle Management
- How do you optimize network connectivity between Ground to Cloud?
- How do you meet enterprise security requirements?
Building your blueprint

1. Questions

Business Drivers AND Business Questions:

- e.g. Where is fraud occurring? How do I optimize inventory?

2. Inputs

Data + Use Cases + Platform

- Subject Areas Source System
- Capabilities, Process
- Ingest, Organize, Enrich, Explore

3. Outcomes

Roadmap + Managed Data Lake + Analytics Strategy
New Buyer’s Guide on Data Lake Management and Governance

- Zaloni and Industry analyst firm, Enterprise Strategy Group, collaborated on a guide to help you:

  1. Define evaluation criteria and compare common options
  2. Set up a successful proof of concept (PoC)
  3. Develop an implementation that is future-proofed

Download now at: resources.zaloni.com
Free eBooks to help you future-proof your data lake initiative

Download now at: resources.zaloni.com
DATA LAKE MANAGEMENT
AND GOVERNANCE PLATFORM

The Next Generation
DATA LAKE
Data Lake 360°

MICA
SELF-SERVICE DATA PLATFORM

BEDROCK
DATA LAKE MANAGEMENT
AND GOVERNANCE PLATFORM