Governed by your cloud-based enterprise data lake

Selwyn Collaco
Chief Data Officer
TMX

Ben Sharma
Founder and CEO, Zaloni

Strata Data Conference
TMX Data Strategy
Unlocking The Business Value Of DATA, An Enterprise Asset
“Our strategy is that we use DATA as an Enterprise (TMX) Asset to unlock Business Value and enable Growth”
Foundational Enablers for Transforming TMX Data Assets

1. Deliver Enhanced Business Value through Data
   - “Uplift” Strategy (Significant Business Value)
     - Enterprise focused use case driven approach
     - Data Synergies
       - Critical data sets across business pillars
       - Data commonality
       - Augmentation & enrichment with alternate data sets
     - Enablement of Monetization Strategies
     - Define business driven consumption patterns

2. Modernize Data Assets
   - Build data synergies across business units eliminating data silos
   - Structure historical data with significant business value residing in RAW format
   - Migrate away from unsupported platforms
   - Avoid data graveyard and associated high costs

3. Enable Advanced BI & Analytical Capabilities
   - Efficient use of data to create business value through:
     - Self serve capabilities (clients, business)
     - Common tooling (ETL, BI)
     - Data mining capabilities

4. Establish Enterprise Data & MDM Capabilities
   - Build out the Enterprise Data Platform (EDP)
   - Data lake capabilities, business glossary, data quality, data security
   - Implement Client Master Data Management

5. Establish Enterprise Data Governance
   - Define & Implement Data Governance:
     - Data governance council
     - Data ownership
     - Data classification
     - Data quality
Update On Our Journey

“Unlocking business value through our DATA”

Client C360
Holistic view of our customers through TMX
- Enable cross sell & up sell
- Enterprise-Wide Account management
- Enterprise Client Relationship view

Enterprise BI
Enterprise wide BI and Data Visualization
- Reduce time and effort for Reporting & Analysis
- Self-serve BI and Empowered business users
- Improved Quality and Better insights

Analytics as a Service
Use case driven search based analytics
- Ad-hoc analytical capability in milliseconds
- Real-time calculations
- White label for monetization opportunities

Enterprise Data Platform
Enterprise Data Lake – Enterprise Wide data repository
- Data Platform supporting Business Insights
- Technology enabling new Business Capabilities
  - Research & Product Development
  - Improved data fulfilment
  - Advanced Analytics
- Data ownership
- Data classification
- Data quality
Non-Invasive Data Governance Approach

In order to develop a Data Governance structure that is right-sized for TMX, the proposed Enterprise Data Governance structure has been developed based on key observations identified through stakeholder interviews, which provided insights into key opportunities and challenges, data domain structure and in-flight data initiatives.

Key Opportunities and Challenges Identified

- Lack of data ownership
- Data accessibility and security
- Duplicative Data
- Third Party Data Management
- Data Sharing Capabilities
- Manual Processes
- Data retention / access to historical data

An ongoing responsibility of the Data Governance bodies will be to prioritize and address data opportunities and challenges.

Data Domains & In-flight Initiatives

Data Domains:
- Trading – Deriv.
- Trading – Equities
- Trading – FI
- Listings
- Clearing – Deriv.
- Clearing – Equities
- Regulatory – Deriv.
- Finance

In-flight Initiatives:
- Marketing
- Customer
- Legal
- Human Resources
- Risk
- Tech. Operations
- External/Alternative

Client 360
Issuer Services
Excel.
Tableau
Advance Analytics Platform
Workday
Atlas

Owners for each domain have been identified and will serve as the foundation for the EDGC; an in-flight project will be selected for preliminary implementation.

Based on the identified key opportunities and challenges, as well as TMX's data domain structure, a two-tier structure is created.

1 Data domains and data domain owners are further detailed in the TMX Data Asset Catalogue.
There are number of Technology solutions available in market place for building the Enterprise Data Platform.

Technologies required for Data Platform falls in THREE categories.

Chart describe the categories and key options within those categories.
Transform your business with an integrated self-service data platform

September 12, 2018
Today’s data confusion
The key is an integrated self-service data platform

1. Manage data across on-prem and cloud environments
2. Metadata management to find useful data and enable governance
3. Automation of data pipelines for scale and efficiency
4. Improve time to insight with self-service for data consumers
No matter where you are on your data lake journey...

Modernize your platform
- Need more agility and flexibility
- Need advanced analytics to reduce time to insight
- Do not want the complexity of building and integrating the platform

Improve governance and adoption
- Early Hadoop adopters
- Not sure what data is in the swamp
- Need to right-size governance
- Data duplicated across ponds
- Enterprise adoption
- Need to demonstrate ROI

Self-service data platform
- Acquire useful data from across enterprise
- Improved visibility and understanding via metadata management
- Ensure security/privacy of sensitive data
- Scalable production data lake for new and improved business insights

Traditional Architecture
- Data Warehouse

Scale Out Architecture
- Data Swamp/Data Ponds

Governed Data Lake

Govern
Typical data lake reference architecture

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

- **Raw Zone**
  - Original source data ready for consumption
  - Consumers are 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

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

- **Sandbox**
  - Ad-hoc exploratory analysis
  - Data scientists, data consumers with proper privileges

Data Lake

- Log (or other unstructured data)
- Sensors (or other time series data)
- Social and shared data

Relational Data Stores (OLTP/ODS/DW)
Zaloni’s integrated self-service data platform (ZDP) accelerates time to insight

Enable
- Batch Ingestion
- Streaming Ingestion
- Metadata Capture
- Auto Discovery

Govern
- Data Quality
- Data Lineage
- Data Mastering
- Data Privacy/Security
- Data Enrichment
- Data Lifecycle Management

Engage
- Data Marketplace
- Self-Service Operations

Cloud, On-Premises, Hybrid
Infrastructure and cloud-platform agnostic

Provides the foundation for your data initiatives
Enable the Data Lake with ZDP

- Scale out ingestion of data
  - Batch and streaming ingestion
  - DB ingestion with support for full/incremental updates
  - Metadata capture and integration with Data Pipeline
  - Ingestion Wizard simplifies ingestion and creates reusable workflows

- Automated data inventory
  - Crawls the data store to catalog datasets
  - Automatically detects metadata in several cases

- Workflow and orchestration
  - Robust workflow management feature with scalable and extensible architecture

- Monitor the health of the data lake
  - Single unified log repository for all data lake operations
  - Dashboards for operational health of the data lake
Govern the Data Lake with ZDP

Metadata management
- Business, technical and operational metadata
- Captures lineage from ingestion to consumption
- Integration with Enterprise Metadata repositories

Security and access control
- Masking and Tokenization of sensitive data
- RBAC for data lake artifacts – Entities, DQ rules, ingestion with push down to underlying security framework
- Supports kerberized environment

Data quality
- Rules Engine for DQ
- DQ reporting and analysis
- Automation of DQ remediation process

Data enrichment
- Drag and drop enrichment of the data
- Support for data operations – JOIN, FILTER, UNION, etc.
- Out of the box format conversion, parsers with ability to extend with custom implementations

Data lifecycle management
- Policy based DLM
- Leverages HDFS storage tiers and supports S3 interface
Engage the Business with ZDP

Data marketplace
- Rich data catalog that aggregates business, technical and operational metadata
- Global search allows data consumers and producers to search across data lake zones, projects, workflows, data quality rules, transformations and other related content
- Annotate, Tag, Rate and create custom metadata
- Workspaces for collaboration

Data provisioning
- Shopping cart experience
- Sandbox provisioning into the data lake or RDBMS
AWS Data Lake Platform Components

**Consumption Layer**
- Tableau
- Kibana
- ML
- Quicksight

**Serving Layer**
- EMR
- ES
- Redshift
- RDS
- DynamoDB
- Spark
- Hive

**Processing Layer**
- EMR
- Athena
- Lambda
- Glue
- Quicksight

**Storage Layer**
- S3
- EFS
- Glacier
- Zaloni Data Platform
- Databases
- Files
- Streams
- SaaS APIs
- Ingestion
- Storage Gateway
- Kinesis
- Zaloni Data Platform

Security and Networking components are included but not shown in this architecture.
Solution Architecture for S3 Data Lake
Visit Booth #1115

The Data Imperative
Ben Sharma, CEO Zaloni
Thursday Keynotes, 10:20am – 3E

Complimentary copy of “Architecting Data Lakes” 2nd edition