The Data Reservoir: Architecture, Best Practices and Governance

Jo Ramos, Distinguished Engineer, Chief Architect for Information Integration & Governance
Agenda for Today

• How to become a data driven analytics organization

• Modern Data Architecture Considerations

• The Data Reservoir & Data Governance
### Three major shifts in our industry

<table>
<thead>
<tr>
<th>Data is becoming the world’s new natural resource</th>
<th>Social, mobile and access to data are changing how individuals are understood and engaged</th>
<th>The emergence of cloud is transforming IT and business processes into digital services</th>
</tr>
</thead>
<tbody>
<tr>
<td>500 million DVDs worth of data is generated daily</td>
<td>80% of individuals are willing to trade their information for a personalized offering</td>
<td>85% of new software is being built for cloud</td>
</tr>
<tr>
<td>1 trillion connected objects and devices by 2015</td>
<td>84% of millennials say social and user-generated content has an influence on what they buy</td>
<td>25% of the world's applications will be available in the cloud by 2016</td>
</tr>
<tr>
<td>80% of the world’s data is unstructured</td>
<td>5 minutes: response time users expect once they have contacted a company via social media</td>
<td>72% of developers say cloud-based services or APIs are central to the applications they are designing</td>
</tr>
</tbody>
</table>
Data Monetization Drivers

- Data volumes
- Data variety
- Real-time execution
- Cost of data storage
- BI & Analytics visibility
- Increasing value of data
- Under-utilized resource
Data driven analytics platform capabilities move from information to insights and from reporting to action.

**Real-Time Decision Management**: What action should I take? What is the Next Best Action?

**Prediction / Simulation**: What will happen? Uses data to forecast based on complex algorithms or rules; forecasting of customers’ buying behavior.

**Evaluation**: Why did it happen? Thoroughly analyzes data to support important decisions / understand root causes for unusual observations; evaluation of campaign responses to understand changes in customer behavior.

**Data Mining**: Why did it happen? Looks for patterns in the data to explain a not yet understood observation; understand why certain customers show a certain behavioral pattern.

**Monitoring**: What is happening now? Looks for trigger information in the data indicating need for action; monitoring live campaigns and making optimization decisions as necessary.

**Reporting**: What happened? Slices and dices data to create transparency on campaign performance and financial or quality outcomes.
Multi-structured Data Mashups provide the Greatest Enterprise Value

**Systems of Record**
Structured data from operational systems

20% of all data generated

**Systems of Insight**
Diverse data types that combine structured and unstructured data for business insight

**Systems of Engagement**
Data that “connects” companies with their customers, partners and employees

80% of all data generated

**Data Warehouses**
- Transaction data
- ERP Data
- Electronic Health Records
- Mainframe Data
- OLTP System Data

**Hadoop, Streams, Spark**
- Audio
- Documents
- Images
- RFID
- Emails
- Sensors
- Social Data
- Video
- Web Logs

**Transaction data**
- ERP Data
- Electronic Health Records
- Mainframe Data
- OLTP System Data

**Advanced Analytics**
- Context
- Accumulation
- Enterprise
- Integration

**Unstructured Data**
- Big Data
- Language based
- Qualitative
- Subjective
- Intuitive
- Mystery
- Exploratory
dynamic

**New Data Sources**
- Audio
- Documents
- Images
- RFID
- Emails
- Sensors
- Social Data
- Video
- Web Logs

**Traditional Sources**
- Structured Data
- Small Data
- Clearly formatted
- Quantitative
- Objective
- Logical
- Puzzle
- Repeatable
- linear
A growing data demand … and organizational tensions

**Data Scientists** seeking data for new analytics models.

**Marketer** seeking data for new campaigns.

**Fraud investigator** seeking data to understand the details of suspicious activity.
Modern Data Architectures: What to aim for

We need IT solutions that are like Legos

- Agile, flexible, adaptive, efficient
- Able to give fast responses to business needs
- Taking advantage of cloud, mobile, social, localization, Big Data

IT should not be like pouring cement

- Rigid, hardcoded business processes
- Monolithic
- Expensive to change
- Out of sync with business needs
Modern Data Architecture

What to aim for

- Simplification of the IT environment
- Eliminate redundancies
- Reduce cost
- Decouple systems
- Decouple transactions
- Fast, easy reuse
- Faster time-to-market
- Explore opportunities
- React to problems
- Generate insights from information in real time
- Use insights to improve customer experience, anticipate facts
Modern Data Architecture

Enables better Customer Engagement

**Social**
- Profiles
- Preferences
- Activities

**Sensors**
- Geolocation
- Movement
- Events
- Internet of Things

**Transactional**
- Customer info
- Transaction history
- Product rules

**Analytics**
- “Next Best Action”
- Expert systems
- Future trends
- Big Data

**Processes**
- Tasks & milestones
- Integration
- Monitoring & SLAs

**Smart Channels**
- Mobile & Web
- “SoLoMo”
- Context-enriched

**Context-Aware Experiences**
- Personalized services
- Real-time Intelligence
- “Moments of Truth”
- Opportunistic offers
Analytics Lifecycle

Search & Survey
(understand what data is available)

Analytics Discovery
(application development)

Deploy & Consume
(application & workflow)

Online Collaboration Workspace
The Data Reservoir

- Built to extract value from the data.
- Managed, Trusted and Governed
Big Data Lakes or Swamps?

• As we bring data together, are we creating a data swamp?
  - No one is sure of the origin or purity of data.
  - No one can find the data they need.
  - No one knows what data is present and if it is being adequately protected.

• How do we build trust in big data?
  - Need trust both to share and to consume data.
  - Need understanding of **quality**, **origin** and **ownership** of data.
  - Need **classification** of data to **govern** and **protect** it.
  - Need **timely**, **reliable** data feeds and **results**.
  - All built on **secure** and **reliable** infrastructure.
IBM’s Data Lake

IBM’s Data Lake = Efficient Management, Governance, Protection and Access.
Users supported by the Data Lake
The Data Lake Subsystems (Services)
Considerations for a well-managed and governed data lake

1. Multiple repositories organized based on source and usage; hosted on appropriate data platforms for workload.
2. Catalog of data, ownership, meaning and permitted usage.
3. No direct access to repositories.
4. Effective interchange of data and insight with other systems.
5. Curation of all data to define meaning and classifications.
6. Business-led information governance and management.
7. Active monitoring and management of data.
9. Access to raw data to develop new production analytics.
10. Moderated, view-based self-service access to data and analytics for line of business.
View from the user community - fraud

- Detect and prevent fraud
- Develop new fraud models
- Conform to regulations
- Investigate Fraud Case
Data repositories support multiple zones
Big data needs a variety of repositories for cost, access and performance reasons.
Data lake logical architecture
Differing user perspectives

Provision Sand Boxes.

Search for, locate and download data and related artifacts.

Define governance policies, rules and classifications. Monitor compliance.

Information Governance Catalogue

View lineage (business and technical) and perform impact analysis.

Add additional insight into data sources through automated analysis.

Develop data management models and implementations.

Curation of Metadata about Stores, Models, Definitions

Data Stores
Governance Rules
Defined for each classification for each situation

Sensitive information masked here

Personal information masked here
Integrated Metadata

- Data Lineage (Traceability)
  - Where does this data come from?
  - Why is this data incorrect?
  - Why is this data incomplete?
  - Can I trust this value?
- Impact Analysis
  - Where is this element used?
  - What happens if I change this?
- Optimization
  - Where is the redundancy?
  - How can I make this run more efficiently?
- Understanding
  - What does this mean?
  - How is this used?
- Control
  - Why is this parameter set to this value?
  - Who made this change?
  - I can change this to meet new business requirements
Secure access to the data lake’s data

- The data lake’s security is assured with this combination of business processes and technical mechanisms.
Building a data lake

- The **first step** in creating the lake is to establish the following:
  - The information **integration** and **governance** components,
  - The **staging areas** for integration,
  - The **catalog**,
  - The common data **standards**.

- The build out of the lake then proceeds iteratively based on the following processes:
  - Governance of a data lake subject area.
  - Managing an information source.
  - Managing an information view.
  - Enabling analytics.
  - Maintaining the data lake infrastructure.
Questions?
Thank You