Business Data Lakes

Venkat Gupta
BI & Analytics Practice Director
Venkat.Gupta@us.sogeti.com
LinkedIn: Venkat Gupta
Agenda

- Business data lakes
- Architecture of Modern Data Warehouse Environments
- Governance considerations for data lakes
Business Data Lakes
The Business Data Lake (BDL) represents a new approach to the creation of analytical insights for the business, from the acceleration of traditional enterprise reporting through to new analytics driven by data science.

It works with high volumes of all kinds of data (structured and unstructured), storing them at low cost and making insights rapidly available throughout the enterprise.

It can coexist with earlier investments, accelerating the evolution of the information landscape.
Concepts and Roles

**Business Use Case Owner**
- Dedicated business teams & governance

**Sandbox**
- Discovery Environment

**Usage 1**
- Selected data & insights

**Usage 2**
- Selected data & insights

**BDL Platform Owner**

**Data Owners**

**Contributors**
- Data Owners

**Business Use Case Contributors**
- Dedicated business teams & governance

**Business Compartments**
- Distill
- Keep the history
- Load everything
- Acquire

**Acquire BDL Platform**

**Owner**
- Business Use Case

---

**CONFIDENTIAL AND PROPRIETARY INFORMATION. © 2017 SOGETI USA LLC**
# Concepts of the BDL

<table>
<thead>
<tr>
<th>Data</th>
<th>Ingestion</th>
<th>Processing</th>
<th>Unified Data Management</th>
<th>Unified Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Data</td>
<td>• Batch Ingestion</td>
<td>• Lambda Architecture</td>
<td>• Master Data Management</td>
<td>• System Monitoring</td>
</tr>
<tr>
<td>• Metadata</td>
<td>• Real-Time Ingestion</td>
<td>• Batch Processing Workflow</td>
<td>• Reference Data Management</td>
<td>• System Management</td>
</tr>
<tr>
<td>• Event</td>
<td>• Micro-Batch Ingestion</td>
<td>• Analytics</td>
<td>• Audit &amp; Policy Management</td>
<td></td>
</tr>
<tr>
<td>• Stream</td>
<td>• Metadata generation</td>
<td>• Analytics Engine</td>
<td>• Privacy</td>
<td></td>
</tr>
<tr>
<td>• Insight</td>
<td></td>
<td>• Real-Time processing</td>
<td>• Information Security</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Business compartments</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Actions – Service Layer</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Existing IS landscape</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Unified Data Management**
- Master Data Management
- Reference Data Management
- Audit & Policy Management
- Privacy
- Information Security

**Unified Operations**
- System Monitoring
- System Management
Business Use Cases for BDL

**EDW Offload**
Accommodate massive data growth with existing EDW investments

**Data Discovery**
Unify Unstructured and Structured Data Access

**Big Data Apps**
Build analytic-led applications impacting top line revenue

**Data-Driven Ecosystem**
Serving multiple organizations (suppliers, partners and 3rd Party contributors) with as-a-service to the ecosystem members

**Data-Driven Enterprise**
Create new services and/or new revenue streams. New data and new analytics capabilities
Modern Data warehouse Environment (DWE) Reference Architecture
Modern Data Warehouse Environment

Reference Architecture

- **Traditional Data**
  - CRM, SFA, ERP...
  - Financials, billing, operations, call center, human resources...

- **New Data, Big Data**
  - Machine Data, IoT, Mobile Data, Web Data, Social Media...

- **CROSS-SYSTEM INTEGRATION**: Data Integration, Application Integration, Metadata, Federated Queries, Virtualization...

- **Many Ingestion Methods**
  - Ingestion
    - Data Landing and Staging
    - ETL/ELT Processing
    - Stream Capture
    - Complex Event Processing

- **Hadoop**
  - Data Lake
    - Massive Store of Raw Source Data (actual Data Lake)
    - Algorithmic Analytics
    - Departmental Data Domains (Mktg, Sales)

- **Core Warehouse**
  - Dimensions, cubes, subject areas, time series, metrics...
  - Data for reports, dashboards...
  - Set-Based Analytics (SQL, OLAP)

- **Diverse System Platforms**: Web, Client/Server, Clusters, Racks, Grids, Clouds, Hybrid Combinations...

- **Specialized DBMSs**
  - Based on clouds, appliances, columns, graph, sandboxes, and other specialized analytics
Information Governance considerations for organizations looking at data lakes
The biggest complaint in Data Governance...

The business won’t engage
The biggest complaint in Data Governance...

Want to know why?
The biggest complaint in Data Governance...

It's because they never cared about Data Governance.
The biggest complaint in Data Governance...

Because it didn’t actually govern the data in the way they wanted
Multiple internal views – consistently compromised
Multiple internal views - you already have a swamp
IT’s answer – just make the EDW ‘bigger’- one view for all
Why the single view & the golden record fails

Division 1
Sales
Finance
Supply chain
Marketing
R&D

Division 2
Sales
Finance
Supply chain
Marketing
R&D

Division 3
Sales
Finance
Supply chain
Marketing
R&D

Division 4
Sales
Finance
Supply chain
Marketing
R&D

Corporate KPIs

Now agree on everything
The biggest complaint in Data Governance...

But good news, Big Data is changing the game
Business Meta-Data – recognizing how people work

- Debt Recovery
- High Value Customers
- Marketing
- Prioritize

Explore the Lake

Data Lake
Business Meta-Data – recognizing how people work

Debt Recovery
High Value Customers

Marketing
Exclude
Purpose: Mktg Exclude

Business Meta-Data
Purpose: Debt Recovery

Data Lake

Explore the Lake
So what do we need?

**Govern where it matters**
- Focus on Meta-data, MDM and RDM
- Enforce only when sharing
- Treat Corporate as aggregation of Local.

**Encourage local requirements**
- Let the business decide what they need
- Build from the bottom
- Enable traceability to source
- Disposable data views.

**Distill on demand**
- Select only what you want
- Business friendly tooling
- Re-usable information maps
- Rapid change cycle.

**Store everything**
- Store everything ‘as is’
- Include structured and unstructured data
- Store it cheaply.
Why Data Lakes work

Supply Chain Optimization
- SAS

Fraud Detection
- MADLib

Next Best Action
- R

Marketing

One Place for Everything

Data Explosion
- Geographic Data

Archives

Operational systems
- Operations
- Case Mgmt
- Scheduling

CONFIDENTIAL AND PROPRIETARY INFORMATION. © 2017 SOGETI USA LLC
Keep governance small and focused – the MDM RADAR
The cross-reference is the most important thing an MDM project can deliver
Quality is a side-effect – not a goal of the first phase of your MDM project
Why the X-Ref matters

Corporate view

Customer

Orders

Invoices

Customer x-ref

Customer MDM

Information governance

Corporate standards

Master data and reference data

Local view

Customer

Invoices

Orders

BU1

BU2

BU3
The Quality mess

ETL

Reports

Reports

Reports

Reports

Internal systems

Market Data

Back Office

CRM

Trading

Trading

Trading
The current situation – add a new System
The current situation – Where does Data Quality go?

Fix at Source

- Market Data
- Back Office
- Trading
- Internal systems

ETL DQ

Reports

Patch

CRM
Back Office
Trading

Reports

Reports

Reports

Reports
Lets try another way – Fix at Source is still the ideal but...
Then lets go further.
Why the Business Data Lake succeeds

Now agree
where it counts
Help me ingestion – you’re my only hope

Sources

- Real-time ingestion
- Batch ingestion

Ingestion tier

- Real time
- Micro batch
- Mega batch

Unified operations tier

- System monitoring
- System management

Unified data management tier

- Data mgmt. services
- MDM
- RDM
- Audit and policy mgmt.

Workflow management

- Processing tier
- In-memory

Distillation tier

- MPP database
- Distillation tier

HDFS storage

- Unstructured and structured data

Insights tier

- SQL
- NoSQL

Action tier

- Real-time insights
- Interactive insights
- Batch insights
If your business isn’t trivial – you’ll end up with multiple lakes
Federation requires **THREE** things

**The cross-reference**
- If you can’t link data sets its impossible

**Rationalization**
- The lesson of Google & Amazon is less choice, not more

**Business Meta-Data**
- Technical Meta-Data is no-longer enough
Building a Lake and the challenge of federation

To do this you need to standardize technology
Business Meta-Data and the x-ref – the **two** things that must be shared.
The new philosophy

- Encourage local
- Govern only the common
- Store everything
- Treat global as a local view

Business Data Lake

It’s all about insight at the point of action
But remember

Your next challenge will be analytical collaboration **beyond** your company boundaries