The Future of Data Management
Managing Data Beyond Boundaries

Ron Agresta
Director of Product Management - Data Management Product Line
SAS
The Future of Data Management

Managing Data Beyond Boundaries

#analyticsx
1969
2006

THE POWER TO KNOW.

Enhance Relations
U.S. Bank uses SAS to foster significant change in retention and a better customer experience.

2016

SAS Global Forum 2016
Our premier event is almost here. Find out what's happening in Las Vegas - and how you can check in on the action.

SAS ranks 8th on Forbes' America's Best Employers list

SAS Health Analytics
SAS Global Forum
Executive Program
SAS Health Analytics Virtual Forum
### SAS Global Forum Abstract Term Comparison

<table>
<thead>
<tr>
<th>2006</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enterprise Guide</td>
<td>Enterprise Miner</td>
</tr>
<tr>
<td>ODS</td>
<td>Enterprise Guide</td>
</tr>
<tr>
<td>IntrNet</td>
<td>Analytics</td>
</tr>
<tr>
<td>Macros</td>
<td>Big Data</td>
</tr>
<tr>
<td>Excel</td>
<td>Hadoop</td>
</tr>
<tr>
<td>Model</td>
<td>DATA Step</td>
</tr>
<tr>
<td>ETL</td>
<td>DS2</td>
</tr>
<tr>
<td>Programmer</td>
<td>Forecasting</td>
</tr>
<tr>
<td>PROC SQL</td>
<td>Amazon</td>
</tr>
<tr>
<td></td>
<td>Grid</td>
</tr>
</tbody>
</table>
New Directions in Data Management
The “Problem” with Data Management Today

- Manually intensive
- Desktop-based tools
- ETL-based processes
- Disenfranchised business users
- Opaque data provisioning techniques
- Advanced analytics often the end, not the means
<table>
<thead>
<tr>
<th>Trends</th>
<th>Scalability</th>
<th>Hybrid Cloud Deployments</th>
<th>Big Data</th>
<th>Operational Integration</th>
<th>Self-Service</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Customers have an increasing number of data domains fueled by data from an increasing number of sources of many different types</td>
<td>Customers have a desire to seamlessly combine data and processes operating locally and in the Cloud</td>
<td>Customers want to take advantage of investments in and promise of new big data infrastructure</td>
<td>Customers see a proliferation of internal and external APIs and task-specific applications that beg for an integrated user experience</td>
<td>Customers want to empower their own technical and non-technical users with an “App-like” experience that minimizes IT involvement</td>
</tr>
</tbody>
</table>
Challenges Extracting Value from Data

- Prioritizing data to generate insights: 65%
- Integrating multiple sources of data: 62%
- Determining the right data sets to collect and analyze: 56%
- Managing data in a privacy-compliant environment: 35%

Source: Forbes Magazine, May 2016
Business Users
Non-technical users looking for an intuitive, app-like experience that decreases the time it takes to go from data acquisition to answers.

Coders, DBAs, IT
Technical experts and traditional data owners who may be reluctant to give up control of data cleansing, ingestion, and provisioning responsibilities.

Data Scientists
Domain experts with deep knowledge looking for an environment where it’s easy to onboard and transform data to foster rapid experimentation.
Data Management Activities

Managed Data
- ETL Developer or Coder
- Team/Department
- Repeated Use

Monitored Data
- IT Staff
- Cross-Team
- Measured Use

Governed Data
- Data Stewards
- Cross-Department
- Administered Use

Governance

Coding, Job Building, Tuning, & Scheduling

in-SAS · in-Memory · in-Database
Data Management in Action

Access → Profile → Clean → Transform

Load → Schedule → Monitor → Share
Fad or Fundamental Changes?

1968

1969

1970

1971
The New Normal

• Interactive data preparation
• Intelligent ingestion
• In-stream data management
• Cognitive data management
• Runtime auto-deployment and optimization
• Big data quality and governance
• Data management anywhere (on any data)
New Data Management in Action

- Discover
- Ingest
- Profile
- Categorize
- Analyze
- Suggest
- Interactively Transform
- Optimize
- Deploy
- Operationalize
- Monitor
Discover & Ingest

**Yesterday**
- Relational databases
- Text files
- ERP systems
- Excel files
- XML

**Today**
- Web services
- Twitter
- LinkedIn
- YouTube
- Data streams
- Cloud platforms
- JSON
- Hadoop | HDFS
- Spark
- Machine data
- Doc management systems
Profile & Categorize

Automatically generated profiling metrics let users quickly find problems in their data with little manual investigation.

Combined data profiling and advanced analytics can lead to semantic tagging where the meaning of data can be inferred. Tagged data then leads to any number of automated actions.
Analyze & Suggest

Software-delivered suggestions (Apple Siri, Google Now, Microsoft Cortana, Netflix, etc.) are now the norm. Users intuitively know that data available behind the scenes can greatly reduce manual investigation of problems, and once analyzed, can lead to insightful recommendations.

**These techniques:**
- Inferring data meaning by tags
- Comparing data to other data
- Analyzing usage patterns
- Evaluating data decay and fetching new data automatically
- Tracking data provenance

**Can lead to:**
- Automatic standardization
- Smart data parsing and merging
- Data invalidation alerts
- Real-time data ranking
- Security policy violation detection
- Advanced pattern discovery
Data Preparation

- data wrangling
- data shaping
- data munging
- data sharpening
- data stitching
- data blending
- data curation

Actions necessary to transform data into the appropriate structure and format for operational, reporting, business intelligence, or analytics use.
Interactive Transformation

Interactive data preparation lets users explore, manipulate and combine data. Transformation can occur in real-time. Suggestions can be reviewed and then confirmed, modified, or rejected. A set of instructions that can be saved and delivered to the appropriate execution environment is generated as a result of user-driven data transformations.

Clean, match, standardize, split, combine, filter, join, merge, hide, arrange, format, sort, append, augment, enrich...
Big (and Little) Data on Big Servers

Powerful commodity-based computing has changed some of the underlying assumptions about how to empower both business users and data experts working in the data preparation area. It is expected that complex transformations on millions of rows will be completed in seconds, not minutes or hours.
Optimize & Deploy

Jobs or code generated from data management activities can be:

- Analyzed to determine an appropriate runtime environment
- Optimized for the target runtime environment
- Automatically deployed and monitored with processing metrics routed back to machine learning algorithms for improved results

Code Generation, Optimization, Process Distribution, & Monitoring

in-SAS · in-Hadoop · in-Memory · in-Database · in-Cloud · in-Stream
Data Management Anywhere

Interest is growing for data management that spans deployment locations yet provides integrated movement, transformation, monitoring, and governance capabilities.

<table>
<thead>
<tr>
<th></th>
<th>Local Data</th>
<th>Local Processing</th>
<th>Remote Data</th>
<th>Remote Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>On-Premises Only</td>
<td></td>
<td>●</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>On-Premises + Cloud Data</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>On-Premises + Cloud Platform</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>Cloud Only</td>
<td></td>
<td></td>
<td>●</td>
<td></td>
</tr>
</tbody>
</table>

THE FUTURE
The Promise of the Cloud

- Opportunities for integration with new and always current data sources
- Storage of large data volumes can be cheaper, especially for less frequently accessed data
- A way to test-drive new technologies
- Elastic infrastructure

- Platform for distributed teams and distributed applications
- Rapid on-boarding of cloud data sources through packaged connectors
- Little to no IT involvement
- “Guaranteed” up-time
- Self-service enablement
- Cloud providers may offer "value added" services
A Few Words about Data Governance

This data free-for-all introduces new challenges for the enterprise that only a comprehensive data management environment with integrated data governance can address.
Data Management Activities

New Users
- Report builders
- Data scientists
- Data analysts
- Compliance staff

Managed Data
- ETL Developer or Coder
- Team/Department
  Repeated Use

Monitored Data
- IT Staff
  Measured Use

Governed Data
- Data Stewards
  Administered Use

Governance

Coding, Job Building, Tuning, & Scheduling
in-SAS · in-Memory · in-Database

New Processes
- Data discovery
- Different runtimes
- Alternate data locations
- Machine learning
# A New Data Management Paradigm

## Examples:

<table>
<thead>
<tr>
<th>Interactive data quality &amp; rule capture</th>
<th>Data quality rule deployment</th>
<th>Data quality process monitoring</th>
<th>Policy rules for data quality</th>
<th>Auto-profiling and data quality rule discovery</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ad Hoc Data</strong></td>
<td><strong>Managed Data</strong></td>
<td><strong>Monitored Data</strong></td>
<td><strong>Governed Data</strong></td>
<td><strong>Activated Data</strong></td>
</tr>
<tr>
<td>Data Scientist or Analyst</td>
<td>ETL Developer or Coder</td>
<td>IT Staff</td>
<td>Data Stewards</td>
<td>Automated</td>
</tr>
<tr>
<td>Individual Self-Service</td>
<td>Team/Department Repeated Use</td>
<td>Cross-Team Measured Use</td>
<td>Cross-Department Administered Use</td>
<td>Cross-Organization Always-On</td>
</tr>
</tbody>
</table>

## Code Generation, Optimization, Process Distribution, & Monitoring

- in-SAS
- in-Hadoop
- in-Memory
- in-Database
- in-Cloud
- in-Stream
Projects with collaboration built-in
**Import data on the fly**

In the SAS Data Management application, the Import Data page allows users to select and import data on the fly. The page features a list of available files, with a highlighted option for "NC_WakeCo_PublicSchools". The data table includes columns for ID, School Name, Street Address, City, and State. The table is populated with records for various schools, each with associated details such as school name, address, city, and state. The interface also includes options to apply settings to all files from the selected location and a button to import selected data.
Discovery, profiling, and categorization
Data preparation with deep data quality
Process and data monitoring combined
Comprehensive lineage and auditing
Impact analysis from data sources to targets
Data Management Evolves

• Self-service data management is here but it doesn’t replace the need for managed, monitored, and governed data

• Cognitive data management is on the rise where deep data analysis uncovers hidden patterns and relationships that enable suggestions and data process automation
Data Management Evolves

• Technologies such as Hadoop, data streaming, in-memory processing and hybrid Cloud deployments require a mix of “standard” and novel approaches.

• Democratizing data management is great as long as users are protected from actions that might lead to incorrect results or security risks.