A FRESH APPROACH TO DATA INTEGRATION FOR ANALYTICS

Lessons from the Field

Jordan Martz
Director, Technology Solutions
Attunity
Who is Attunity?

The leading platform for delivering data efficiently and in real-time to data lake, streaming and cloud architectures

#1 provider of change data capture (CDC)
Support most sources with best performance and least impact

#1 cloud database migration technology
Already moved over 57,000 databases to Cloud platforms like AWS

#1 in ease of use
Designed to accelerate deployments by data architects and DBAs instead of developers

Microsoft Azure
Amazon Web Services
Verizon
Cardinal Health
Zurich
Ford
ENABLE AND ACCELERATE DATA AVAILABILITY
Attunity’s Journey

- From component to platform
- From tactical to strategic
- From small to large deals

- 2004
  - CDC for ETL
  - Database Replication
    - Oracle, SQL, ...

- 2011
  - Universal Replication
    - Cloud, Hadoop, ...

- 2015
  - Modern Data Integration Platform
    - Pipeline automation, metadata, ...

- 2018
Your Data Journey
Your Critical Data is Here

- 45% market share RDBMS
- 80% of the corporate data on mainframes
- 64,000 unique companies globally
Your Technology Landscape is Shifting

CLOUD
DATA LAKES
STREAMING
ON PREM
DATA WAREHOUSE
BATCH
Your Choice of Many Modern Data Platforms

- cloudera
- nifi
- Google Cloud Platform
- Apache Spark
- Hortonworks
- confluent
- Microsoft Azure
- snowflake
- MAPR
- kafka
- Amazon Web Services

DATA LAKES
STREAMING
CLOUD
DATA WAREHOUSE
BATCH
ON PREM
## You Have Many Usage Patterns Too

<table>
<thead>
<tr>
<th>Type</th>
<th>Mode</th>
<th>Description</th>
<th>Reason</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>Stream</td>
<td>Observing real-time operational feed</td>
<td>Personalization, real-time scoring request</td>
<td>Click stream, tick stream, sensor outputs, M2M, gameplay metrics</td>
</tr>
<tr>
<td>Event</td>
<td>State</td>
<td>Long-lived, low velocity data attributes</td>
<td>Data required for application operation</td>
<td>Version, location, user profiles, point-of-presence data</td>
</tr>
<tr>
<td>Big Data Analytic Outputs</td>
<td>State</td>
<td>OLAP reporting</td>
<td>Scoring rubrics, user segmentation profiles</td>
<td>Scoring Models, seasonal usage, demographic trends</td>
</tr>
<tr>
<td>Event responses and alerts</td>
<td>Events</td>
<td>Alerts/notifications of exceptional events or sequences of events</td>
<td>Decisions and corrective actions</td>
<td>Authorizations, policy decisions, triggers, threshold alerts</td>
</tr>
</tbody>
</table>
Through 2018, 90% of deployed data lakes will be useless as they are overwhelmed with information assets captured for uncertain use cases.

Key Challenges

Data lakes store raw data and their business value is entirely determined by the skills of data lake users.

Many technologies used to implement data lakes are new and lack the necessary information capabilities that organizations normally take for granted.

Without data lineage within data lakes, data must be collected, assembled and refined by each user separately and independently to drive meaningful business insights.

Source: Metadata Is the Fish Finder in Data Lakes, 2017
Changing Data Integration

Lack of:
- SKILLS
- TOOLS
- TRUST

Requires:
- Automated Data Pipelines
- Integrated Platform Built for Data Lakes, Cloud, & Streaming
- Effective & Ubiquitous Metadata
## Your Analytics Requirements

<table>
<thead>
<tr>
<th>Modern Analytics</th>
<th>Requires</th>
<th>Ideas</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI/ML/IoT Analytics</td>
<td>Scale</td>
<td>Use data from more sources, 100s to 1000s, with minimal impact and development resources</td>
</tr>
<tr>
<td>Streaming Analytics</td>
<td>Real-time</td>
<td>Create real-time streams from database transactions</td>
</tr>
<tr>
<td>Cloud Analytics</td>
<td>Efficiency</td>
<td>Optimized data transfer over limited network bandwidth and across multiple data centers</td>
</tr>
<tr>
<td>Agile, Fast Deployment</td>
<td>Simplification</td>
<td>Enable “data people” (not developers) to deploy solutions and iterate fast</td>
</tr>
<tr>
<td>Diverse Analytics Platforms</td>
<td>Flexibility</td>
<td>Rapidly adapt and adopt new platforms</td>
</tr>
</tbody>
</table>
Pipeline Automation is Critical to Analytics Success
Landing Data

LAND
Hive
AWS S3
Azure ADLS

STREAM
CDC
BULK

Informatica
talend

SAP
RDBMS
DATA WAREHOUSE
FILES
MAINFRAME
Staging Data

LAND
- Hive
- AWS S3
- Azure ADLS

STAGE
- Hive
- S3
- ADLS

Informatica
Talend
Trifacta
Paxata
Provisioning Data

LAND
- Hive
- AWS S3
- Azure ADLS

STAGE
- Hive
- S3
- ADLS

PROVISION
- Hive
- Cloud DW
- EDW, RDB
- Kudu...

STREAM
CDC
BULK

ATTUNITY
PROCREATE
ATTUNITY
COMPOSE

SAP
RDBMS
DATA WAREHOUSE
FILES
MAINFRAME

UNSTRUCTURED DATA

IOT

TRIFACTA
Paxata

python
R
Sas
Qlik
Tableau
Managing the Process

MANAGEMENT AND METADATA SERVICES

- LAND: Hive, AWS S3, Azure ADLS
- STAGE: Hive, S3, ADLS
- PROVISION: Hive, Cloud DW, EDW, RDB, Kudu...

STREAM
CDC
BULK
Automating Your Data Lake Pipeline

ATTUNITY REPLICATE

Ingest
- CREATE
- UPDATE

Data ingested into data lake

ATTUNITY COMPOSE FOR HIVE

Staging
- STANDARDIZE
- FORMAT
- MERGE

Data continuously updated and merged into historic data store

Provision
- ENRICH
- UPDATE

Subsets created to meet analytic requirements

Consume
- ANALYZE
- PREPARE, CLEANSE & JOIN

Data continuously updated and merged into historic data store
Universal Data Ingest/Replication

Platforms you know

- Data Warehouses – Teradata, Netezza, Vertica
- RDBMS – Oracle, MSSQL, MySQL, Postgres...
- Virtual Machines – Windows, Linux...
- Hadoop – Cloudera, Hortonworks, Map-R
- Message Streams – Apache Kafka
- File Storage – Windows Shares, NAS devices...

- Redshift
- RDS
- EC2
- EMR
- Kinesis
- AWS S3
- SQL DW
- SQL DB
- Azure VM
- HDInsight
- Event Hubs
- Azure Blob
- Data Lake Store

Platforms you know (Microsoft Azure)

- Big Query
- Cloud SQL
- Compute
- Dataproc
- Pub/Sub
- Cloud Storage

Platforms you know (Google Cloud Platform)
Customer Success

PROBLEM

- Needed efficient, real-time data integration
- SAP ERP with Oracle DB to Hortonworks Hadoop
- Available tools insufficient

ATTUNITY SOLUTION

- Application level data integration
  - SAP directly to Hadoop
  - Decoded SAP pool and cluster tables
- Reduced labor and cost
  - Single data integration platform
  - Eliminated most manual coding

“Our financial analytics and modeling teams now have instant visibility across the business.”

Arvind Rajagopalan
DIRECTOR - GLOBAL TECHNOLOGY SERVICES

Fortune 20 Company
Demo: Automating a Data Pipeline

1. Hide underlying complexity
   Include all customizations
   Native app extension

2. Universal integration
   Simply powerful
   Change data streams

3. No manual coding
   Complete transactional integrity
   Both operational & historic data
Questions and Answers
Thank you