Synthesis

A New Webinar Series

WHAT ARE THE BUILDING BLOCKS OF A MODERN DATA PIPELINE?
What are the Building Blocks of a Modern Data Pipeline?
Destination: Purpose and End Point

Origin

Dataflow

Storage

Workflow

Processing

Monitoring

Destination

Stores:
- Staging
- Warehouse
- Data Mart
- MDM
- ODS
- Data Lake
- Sandbox

Applications:
- Reporting
- OLAP
- Scorecards
- Dashboards
- Exploration
- Analytics

Technology
What requirements for real time data?
What criteria for right time data?
For which data is latency okay? And how much latency?
Origin: Data Supply and Begin Point

**Sources:**
- Legacy
- Transaction
- Web
- 3rd Party
- Social Media
- Machine
- Geospatial

**Stores:**
- Staging
- Warehouse
- Data Mart
- MDM
- ODS
- Data Lake
- Sandbox

**Dataflow**

**Storage**

**Processing**

**Workflow**

**Monitoring**

**Technology**

- **Origin:**
  - Data Supply and Begin Point

- **Destination:**
Which data is event based & which is entity data?
Is event data stored or streamed?
How quickly must data be gathered from sources?
How frequently must data be gathered from sources?
Data Flow: Pipeline Boundaries

Origin

Dataflow
origin

Destination

dataflow
destination

one pipeline

- Orders
- A/R

Process:
- Extract
- Cleanse
- Load

Destination:
- Data Warehouse
- Data Mart
- A/R

two pipelines

- Orders
- A/R

Process:
- Extract
- Cleanse
- Load

Destination:
- Data Warehouse
- Data Mart
- A/R

Technology
Data Storage: Data at Rest

**Origin**

- Dataflow

**Storage**

- temporary files
- staging tables
- data lake
- data warehouse
- master data repository
- analytics sandbox

**Workflow**

**Processing**

**Monitoring**

**Destination**

**Technology**
Data Storage: Which is the Right Data Store?

Volume of data?
Structure & format?
Duration & retention?
Query frequency & volume?
Other users and uses?
Governance constraints?
Privacy & security?
Disaster recovery?

Origin

Dataflow

Storage
- temporary files
- staging tables
- data lake
- data warehouse
- master data repository
- analytics sandbox

Workflow

Processing

Technology

Destination

Dataflow scheduling execution failover distribution verification

Monitoring

Volume of data?
Structure & format?
Duration & retention?
Query frequency & volume?
Other users and uses?
Governance constraints?
Privacy & security?
Disaster recovery?
Processing: Adding Value and Creating Data Products

Origin

- Dataflow

Storage

- Staging
- Warehouse
- Data Mart
- MDM
- ODS
- Data Lake
- Sandbox

Applications:

- Reporting
- OLAP
- Scorecards
- Dashboards
- Exploration
- Analytics

Destination

- Stores:
  - Staging
  - Warehouse
  - Data Mart
  - MDM
  - ODS
  - Data Lake
  - Sandbox

- Applications:
  - Reporting
  - OLAP
  - Scorecards
  - Dashboards
  - Exploration
  - Analytics

Stores:

- Origin
  - Staging
  - Warehouse
  - Data Mart
  - MDM
  - ODS
  - Data Lake
  - Sandbox

Workflow

Processing

- Source:
  - extract
  - map
  - sample
- Transform:
  - transform
  - abstract
  - blend
- Destination:
  - load
  - reduce
  - format
  - publish

Workflow

- Scheduling
- Execution
- Failover
- Distribution
- Verification

Technology

Process:

- Extraction
- Transformation
- Loading

Monitoring

- Health Check
- Performance
- Logging
- Debugging
Processing: Stages of the Data Lifecycle

**Origin**
- Ingest
- Persist
- Transform
- Deliver

**Workflow**
- Ingest
- Persist
- Transform
- Deliver

**Technology**
- Map
- Reduce
- Load
- Transform
- Abstract
- Convert
- Sample
- Blend
- Format

**Storage**
- Database
- Temporary files
- Staging tables
- Data warehouse
- Data mart
- Operational data store
- Master data repository

**Processing**
- Extract
- Transform
- Load
- Filter
- Pull
- Merge
- Join
- Reduce
- Map
- Reduce
- Connect
- Abstract
- Publish
- Sample
- Blend
- Format

**Monitoring**
- Health check
- Performance
- Logging
- Debugging

**Destination**
- Reporting
- OLAP
- Scorecards
- Dashboards
- Exploration
- Analytics

**Sources**
- Legacy
- Transaction
- Web
- 3rd Party
- Social Media
- Machine
- Geospatial

**Stages**
- Stores: Staging, Warehouse, Data Mart, MDM, ODS, Data Lake, Sandbox
- Applications: Reporting, OLAP, Scorecards, Dashboards, Exploration, Analytics

**Other**
- Derive
- Append
- Aggregate
- Sort, sequence, & pivot
- Sample, select, filter, & mask
- Assemble & construct
- Standardize & conform
- Cleanse & quality assure
- De-duplicate
Workflow: Sequence and Dependencies

<table>
<thead>
<tr>
<th>Task Dependencies</th>
<th>Downstream Dependencies</th>
<th>Synchronous Dependencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requires successful completion of one or more preceding tasks</td>
<td>Tasks later in execution sequence wait for successful completion</td>
<td>Parallel execution of multiple tasks requires all tasks to finish successfully</td>
</tr>
<tr>
<td>Job Dependencies</td>
<td>Requires successful completion of one or more preceding jobs</td>
<td>Jobs later in execution sequence wait for successful completion</td>
</tr>
</tbody>
</table>

Workflow:
- scheduling
- execution
- distribution
- failover
- verification

Monitoring:
- health check
- performance
- logging
- debugging

Technology:
- Workflow
- Scheduling
- Execution
- Distribution
- Failover
- Verification
Monitoring: Pipeline Health

What to watch?  
Who is watching?  
Using what tools?  
What thresholds & limits?  
What actions & when?

Technology:

- Storage  
- Processing  
- Workflow  
- Monitoring  
- Dataflow

Origin

Destination

Source:
- Legacy  
- Transaction  
- Web  
- 3rd Party  
- Social Media  
- Machine  
- Geospatial

Stores:
- Staging  
- Warehouse  
- Data Mart  
- MDM  
- ODS  
- Data Lake  
- Sandbox

Applications:
- Reporting  
- OLAP  
- Scorecards  
- Dashboards  
- Exploration  
- Analytics

Destination:

Origin:

Storage:
- temporary files  
- staging tables  
- data warehouse  
- data mart  
- operational data store  
- master data repository

Processing:
- extract  
- transform  
- load  
- map  
- reduce  
- extract  
- load  
- transform  
- connect  
- abstract  
- publish  
- sample  
- blend  
- format

Workflow:
- scheduling  
- execution  
- failover  
- distribution  
- verification

Monitoring:
- health check  
- performance  
- logging  
- debugging  
- health check  
- debugging
Stores:
- Staging
- Warehouse
- Data Mart
- MDM
- ODS
- Data Lake
- Sandbox

Applications:
- Reporting
- OLAP
- Scorecards
- Dashboards
- Exploration
- Analytics

Destination:
- Legacy
- Transaction
- Web
- 3rd Party
- Social Media
- Machine
- Geospatial

Storage:
- temporary files
- staging tables
- data warehouse
- data mart
- operational data store
- master data repository

Processing:
- extract
- transform
- load
- map
- reduce
- extract
- load
- transform
- connect
- abstract
- publish
- sample
- blend
- format

Workflow:
- scheduling
- execution
- failover
- distribution
- verification

Technology: Hadoop, Databases, ETL, Automation, Virtualization, Analytics, Cataloging, Data Preparation

Dataflow:
- Origin
- Destination

Technology: Pipeline Tools
Design Summary: Scope and Complexity

Dataflow:
- Origin: Legacy, Transaction, Web, 3rd Party, Social Media, Machine, Geospatial
- Destination: Staging, Warehouse, Data Mart, MDM, ODS, Data Lake, Sandbox

Storage:
- Temporary files
- Staging tables
- Data lake
- Data warehouse
- Master data repository
- Analytics sandbox

Processing:
- Extract
- Transform
- Load
- Map
- Reduce
- Connect
- Abstract
- Publish
- Sample
- Blend
- Format

Workflow:
- Scheduling
- Execution
- Distribution
- Failover
- Verification

Monitoring:
- Performance
- Logging
- Health check
- Debugging

Technology: Hadoop, Databases, ETL, Automation, Virtualization, Analytics, Cataloging, Data Preparation...
Infoworks Overview

The Automated Software Platform for Agile Data Engineering

July 18, 2018
What Will The Data & Analytics World Look Like In 3-5 Years?
The Goal: Companies Want to Emulate - Google, Facebook and Amazon

- Winners and losers will be determined by the data & analytics agility of the company –
  - The ability to handle:
    - large number of analytic use cases
    - large amounts of data
    - a large number of users
    - rapid and frequent changes

- This requires companies to:
  - Automate data engineering
  - Have end-to-end functionality in a single place
  - Design once and deploy anywhere – on-premise or cloud
The Challenge: Data Engineering is “Death by 1000 Paper Cuts”

Data Ingestion
- Change Data capture
- Parallelization of data load
- Slowly changing dimensions
- Conversion of source types to big data types

Data Synchronization
- Data Merge
- Data Synch
- History table creation

Data Transformation
- Building initial load data pipelines
- Building CDC pipelines
- Building SCD pipelines
- Pipeline change management
- End to end lineage creation

Data Models
- Building semantic models
- Building OLAP cubes
- Building in-memory models

Data Governance
- Data access control
- Change management tracking
- Enabling compliance reporting

Performance Optimization
- Tuning of data load
- Tuning of data transformation
- Tuning of cube generation
- Tuning of in memory models

Production Orchestration
- Scaling jobs
- Migration from dev to production
- Operationalizing data science models
- Monitoring operational environment
- Identifying and restarting failed jobs
The Solution: An Agile Data Engineering Platform

- **Automation**
  - Code-free automation of data engineering from data source to point of consumption

- **Infrastructure Independence**
  - Portable between and across environments on premise or in the cloud

- **Platform Extensibility**
  - Supports customer or 3rd party applications
Infoworks Agile Data Engineering Platform

- End to end automation
- Portable across all data & compute platforms
- All components are API accessible
Automation: Allows You to Focus on Business Results

Infoworks Autonomous Data Engine

**DATA INGESTION & SYNC**
- Automatic Ingestion / CDC
- Automatic Data Type Conversion
- Auto Crawling
- Automatic Schema Change
- Automatic Merge

**DATA TRANSFORMATION**
- Automated Incremental Pipelines
- Automated Data Validation
- Automated Dependency Management
- Suggest New Data Connections

**HI-PERF MODELS**
- Automatically optimize data models
- Auto create OLAP cubes
- Automatically maintain time axis
- Automated metadata lineage to source

**PRODUCTION OPERATIONS**
- Automated Fault tolerance
- Restartability
- Monitor/ Debug

- Eliminates the need for specialized talent and consultants
- Enables new use cases to be launched 10x faster with fewer resources
Customer Case Studies

**Advanced Analytics**  
*(Fortune 10 Healthcare Co.)*  
Implemented a complex, machine learning, near-real-time, business process in 19 days

- Synchronized with Teradata every 10 mins
- **15 min** data-availability SLA
- Implemented by 2 engineers in 19 days from requirements to production

**Data Lake Creation**  
*(Fortune 100 Technology Co.)*  
Implemented Enterprise wide Data Lake involving 1500 data sources

- Synchronized data (CDC/Merge) from 1500 sources
- **Serving reference data** for all enterprise analytics
- Implemented by 2 engineers in < 2 months including a data shopping cart

**Self Service BI & Cloud Portability**

Built self-serve BI use case dashboards in 4 days and migrated from Azure to GCP in 1 day

- 7 data sources,
- 8 pipelines
- 8 optimized models
- 3 cubes
- 13 reports & dashboards
- Sub-second query response

<table>
<thead>
<tr>
<th>Without Infoworks</th>
<th>Infoworks</th>
</tr>
</thead>
<tbody>
<tr>
<td>~6 months</td>
<td>19 days</td>
</tr>
</tbody>
</table>

**9.5x Improvement**

<table>
<thead>
<tr>
<th>Without Infoworks</th>
<th>Infoworks</th>
</tr>
</thead>
<tbody>
<tr>
<td>~2 years</td>
<td>60 days</td>
</tr>
</tbody>
</table>

**10x Improvement**

<table>
<thead>
<tr>
<th>Without Infoworks</th>
<th>Infoworks</th>
</tr>
</thead>
<tbody>
<tr>
<td>~6 months</td>
<td>1 day</td>
</tr>
</tbody>
</table>

**180x Improvement**
Infoworks: **The Agile Data Engineering Platform**

- **End to end functionality in an integrated platform**
  - Full data, metadata & business logic in one place
    - Ingestion
    - Merge
    - Data transformation
    - Data models
    - Data acceleration
    - BI/Al/ML Integration
    - Data governance & lineage
    - Workload migration

- **End to end data engineering automation**
  - Schema evolution, incremental pipelines, type management, query routing, query optimization, model recommendation, dependency management, …

- **Infrastructure independence**
  - Portable across different environments—on-premise or cloud

- **Platform extensibility**
  - Third party apps, customer apps, API integration
THANK YOU!

Every new beginning comes from some other beginning’s end.
Q&A