DataOps 101:
A DataOps Framework For Enterprise Data Management
Agenda

Why DataOps?
A DataOps Framework
DataOps Success Stories
Getting Started With DataOps
Panel Discussion
Traditional companies have significant “legacy drag coefficient”

Manage data from their business systems more as “exhaust” than “asset” > “significant data debt”

**Problem**: Thousands of systems generating data every day that were built over decades to support business processes - idiosyncratic to that time/context.

Data is idiosyncratic to each system - creates fundamental **“data disconnect”** and **“data decay”**

**Result**: “Random Data Salad”
Data debt from constant change/entropy

**Consequences**:
1. Too much time spent on data prep vs. analysis / action.
2. High failure rate of BI / analytics projects
3. Game changing initiatives deemed ‘impossible’ and never start
Why now? 7 years ago: we need data scientists!
Today: we have data scientists! (and want to do cool AI stuff)
But what about cleaning up/preparing our data

The New York Times

For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights

What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

Data Scientist Survey by Figure Eight
Unique Moment in Time: Enterprise Data as an Asset

Now that we’ve established Data Science as critical component of enterprise:

It’s time for each enterprise in the Global 2000 to build their data engineering muscles to enable them to “compete on analytics” over the coming decades.

- Decades of treading data as operational exhaust
- Deeply Fragmented/Siloed Data Environments
- Inability to leverage new sources - esp external “AI Cart” before the “Data Horse”
- Significant “lift & shift” opportunity
- Potential for behavioral changes
- New infra good/secure enough

- Fear of disruption by “Data Natives”
- Low Hanging Analytical Fruit
- “Competing on Analytics”/Strategic Imperative
- Popularity of AI but really of Core Data Quality
- Inability of traditional tech to scale
- Lack of innovation from old vendors
- Maturing Big Data Tech (HDFS/Lakes)
- Democratization of analytics
- Rise of the CDO
- Generational Change in Enterprise Data Management
- Enterprise Migration to the Cloud
What is DataOps? = Modern Data Engineering Practice

**DataOps is an automated, process oriented methodology, used by analytic and data teams to improve the quality and reduce the cycle time of data analytics.**

DataOps is an automated, process-oriented methodology, used by analytic and data teams, to improve the quality and reduce the cycle time of data analytics. While DataOps began as a set of best practices, it has now matured to become a new and independent approach to data analytics. DataOps applies to the entire data lifecycle from data preparation to reporting, and recognizes the interconnected nature of the data analytics team and information technology operations. From a process and methodology perspective, DataOps applies Agile software development, DevOps and the statistical process control used in lean manufacturing, to data analytics.

In DataOps, development of new analytics is streamlined using Agile software development, an iterative project management methodology that replaces the traditional Waterfall sequential methodology. Studies show that software development projects complete significantly faster and with far fewer defects when Agile Development is used. The Agile methodology is particularly effective in environments where requirements are quickly evolving—a situation well known to data analytics professionals.

DevOps focuses on continuous delivery by leveraging on-demand IT resources and by automating test and deployment of analytics. This merging of software development and IT operations has improved velocity, quality, predictability and scale of software engineering and deployment. Borrowing methods from DevOps, DataOps seeks to bring these same improvements to data analytics.

Like lean manufacturing, DataOps utilizes statistical process control (SPC) to monitor and control the data analytics pipeline. With SPC in place, the data flowing through an operational system is constantly monitored and verified to be working. If an anomaly occurs, the data analytics team can be notified through an automated alert.

# This Is A Solved Problem - Look At Software

## Waterfall Development

<table>
<thead>
<tr>
<th>Top-Down</th>
<th>Architects drive the spec</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Priori Modeling</td>
<td>Front-loaded view of all components</td>
</tr>
<tr>
<td>Monolithic</td>
<td>Single application</td>
</tr>
<tr>
<td>Quality Assurance</td>
<td>Manual QA</td>
</tr>
<tr>
<td>Traditional SDLC</td>
<td>Dev/Test/Prod → Major/Minor release</td>
</tr>
</tbody>
</table>

## Agile Development

<table>
<thead>
<tr>
<th>Bottom-Up</th>
<th>Users drive the spec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learn From Use</td>
<td>Emergent feature set</td>
</tr>
<tr>
<td>Distributed</td>
<td>Loosely-coupled, scalable</td>
</tr>
<tr>
<td>Continuous Integration</td>
<td>Automated testing</td>
</tr>
<tr>
<td>Modern DevOps</td>
<td>☄ Continuous delivery</td>
</tr>
</tbody>
</table>

Just as DevOps drove rapid delivery of high-quality, scalable software applications, DataOps is the path forward for data
A DataOps Framework: Technology, Organization, Process
DataOps Framework Components

**Technology**
- **Architecture** - selection of tools which comprise data supply chain
- **Infrastructure** - selection of platform to support architecture

**Organization**
- **Roles** - division of labor across mixed-skill teams
- **Structure** - working model for projects across technical and business teams

**Process**
- **Agile** - incremental delivery model
# Technology - Architectural Principles

<table>
<thead>
<tr>
<th>Internal Tabular Data</th>
<th>Technology, Organization, Process</th>
<th>Consumers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale Out/Distributed</td>
<td>Cloud First</td>
<td>Citizens</td>
</tr>
<tr>
<td>Collaborative (Humans at the Core)</td>
<td>Highly Automated - automate whenever possible</td>
<td>Analysts</td>
</tr>
<tr>
<td>Open/Best of Breed (not one platform/vendor)</td>
<td>Service Oriented (clear endpoints for data)</td>
<td>Data Scientists</td>
</tr>
<tr>
<td>Continuous (assume data will change)</td>
<td>Both aggregated AND federated storage</td>
<td>Developers</td>
</tr>
<tr>
<td>Lineage/Provenance is essential</td>
<td>Both batch AND Streaming</td>
<td></td>
</tr>
</tbody>
</table>

**Sources**

- Internal Tabular Data
- External Tabular Data
Infrastructure - Key Components

Sources

Internal Tabular Data

External Tabular Data

Technology, Organization, Process

Management

Compute

Storage

Infrastructure

Search

Consumers

Citizens

Analysts

Data Scientists

Developers
Organization - Roles

Sources
- Internal Tabular Data
- External Tabular Data

Technology, Organization, Process
- CIO
- CDO
- Business Owners and Other CxOs
  - Source Owner
  - DBA
  - IT Professional
  - Data Engineer
  - Curator
  - Steward

Consumers
- Citizens
- Analysts
- Data Scientists
- Developers
<table>
<thead>
<tr>
<th>Role</th>
<th>Goals</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Citizen</strong></td>
<td>Use data to make business decisions</td>
<td>Viz, CRM, Excel, PowerPoint, Word, Web Search</td>
</tr>
<tr>
<td><strong>Analyst</strong></td>
<td>Deliver insights to the business, typically through dashboards and reports</td>
<td>Viz, Excel, SSDP, Web Search</td>
</tr>
<tr>
<td><strong>Scientist</strong></td>
<td>Deliver insights to the business, typically through models and algorithms</td>
<td>R, Python, SAS, SSDP</td>
</tr>
<tr>
<td><strong>Developer</strong></td>
<td>Build applications which leverage corporate data</td>
<td>Python, Java, JS, SQL, REST</td>
</tr>
<tr>
<td><strong>Engineer</strong></td>
<td>Deliver and manage data pipelines</td>
<td>ETL, SQL</td>
</tr>
<tr>
<td><strong>Curator</strong></td>
<td>Ensure consumers have the data they need, in the form they need it</td>
<td>MDM, Catalog</td>
</tr>
<tr>
<td><strong>Steward</strong></td>
<td>Create policies and drive governance</td>
<td>MDM, Catalog, Governance</td>
</tr>
<tr>
<td><strong>Source Owner</strong></td>
<td>Define and manage purpose, processes (data creation, consumption) &amp; users (i.e., access) of the data source</td>
<td>EDW, SQL, ERWin, LDAP, SAP</td>
</tr>
</tbody>
</table>
Organization - Structure

**Advisory Model**
*Bootstraps projects with best of breed tools and approach, but does not complete them*

**Advantages**
- Centralized technical knowledge
- Minimal resourcing - experts, not implementers
- Flexibility - options to deviate from standard tools

**Disadvantages**
- Resource burden in on each project / department - both in development and ongoing maintenance
- Limited feedback - does the advice get better after each project?

**Shared Services Model**
*Full-service development of data applications, in collaboration with business*

**Advantages**
- Centralized technical knowledge
- Centralized resourcing - one-stop shop
- Accretive experience

**Disadvantages**
- Bandwidth contention - how to prioritize competing projects?

Appropriate model will fluctuate with scale of DataOps project work
Process - The Wrong Way

Sources
- Internal Tabular Data
- External Tabular Data

Technology, Organization, Process
- Labor-intensive
- Monolithic
- IT driven

Consumers
- Citizens
- Analysts
- Data Scientists
- Developers

Remaining Work
- Modeling
- Testing
- Rules

Time

Delivery
Process - The Right Way

Sources

- Internal Tabular Data
- External Tabular Data

Technology, Organization, Process

- Automated
- Incremental
- Collaborative

Consumers

- Citizens
- Analysts
- Data Scientists
- Developers
DataOps Success Stories
Know Your Customer (KYC)

Goals
● Understanding over 400,000 institutional customers and their business affiliations
● Complying with strict regulations to combat illegal activity
● Quickly and reliably link transaction activity to relevant institutions

Organization
● Project coordinated by Customer Data Services group
● Impacts and depends upon several major groups (e.g. onboarding, risk mgmt.)
● Multiple high-level stakeholders, including direct CDO involvement

Challenges
● Regulatory deadlines -- MASSIVE potential fines
● Extremely high visibility
● Immediate and constant demand for progress and results
Where we started:

Datasets
- 30+ total datasets
- Mix of external and internal data
- Customers, accounts, transactions
- Many datasets were completely unfamiliar to the team

Small Working Team
- 2 active contributors from Customer Data Services
- 6 consulting subject matter experts (SMEs)
- Me! (+ massive support from Tamr team)
Case Study - Financial Institution

Sound Familiar?

“This software can do <impossible thing> right?”

“We need results ASAP!”

“We are 100% certain this dataset is clean and curated”

“We’ll test along the way”

“Can’t we just do that ourselves?”

“We can always throw more resources at it”
Case Study - Financial Institution

Breaking down the customer vs. vendor relationship

- Everyone’s fortunes are tied up in the same success/failure
- Mistrust wastes a lot of time
- Get the expert on the “front foot” - proactive rather than reactive
- Enable a lot more creativity from all sides
- Better understanding and confidence in the outcomes
Case Study - Financial Institution

“This software can do *impossible thing* right?”
“We need results ASAP!”

**Examples**
- Magic Machine Learning
- Handling badly mangled data
- Provide answers that aren’t in the data

**Counteroffer**
“Let’s figure out what we CAN get now, but also push for improvements in the data”

**Outcomes**
- Much shorter time to value
- Set aside initial assumptions to focus on the real goals
Case Study - Financial Institution

“We are 100% certain this dataset is clean and curated”
“We’ll test along the way”

Examples
● Missing primary keys
● Malformed data
● Inconsistent meaning of attributes

Counteroffer
Clearly define blocking issues, show everyone the examples, and develop mitigations

Outcomes
● “Cleanliness” issues often reveal fundamental problems with the data or logic
● Push back all the way up the chain!
Case Study - Financial Institution

“Can’t we just do that ourselves?”
“We can always throw more resources at it”

Examples
● Scaling, data transformation, error handling

Counteroffer
“Sure, we CAN do that for now, but we really need to…”

Outcomes
● You are now a great partner/employee! Way better than those other guys!
● Strongly quashes unreasonable expectations while still doing everything possible to help
● Often leads to a more direct role in the long-term solution. You figured it out first!
Case Study - Financial Institution

Key Points

1. Holistic Approach
   ● No software is a magic bullet for all problems
   ● We’re exposing data quality, data engineering, and process issues
   ● Keeping bad/unnecessary data away as much as possible

2. Iteration, Flexibility, Speed
   ● Many problems were multiple years in the making
   ● Existing processes haven’t worked -- don’t rely on them!
   ● Offer something newer and more productive

3. Money Talks
   ● Consistently deliver value through the iterations of a project
   ● Good results can quickly put an end to arguments
   ● Maintain focus on the real goals and realistic solutions
Getting Started With DataOps
DataOps Framework Components

**Technology**
- **Architecture** - selection of tools which comprise data supply chain
- **Infrastructure** - selection of platform to support architecture

**Organization**
- **Roles** - division of labor across mixed-skill teams
- **Structure** - working model for projects across technical and business teams

**Process**
- **Agile** - incremental delivery model
Getting Started - Process

Agile is the key
- If not already there, choose a model that works (Scrum, SAFe)

Inventory the set of available projects
- Score on availability of data vs. value of solving a problem

Define high-value, data-rich project that will demand a complex solution
- Forcing function to ensure end-to-end functionality will be covered
Getting Started - Organization

Inventory current team
● Identify existing key roles - data engineers and their consumers
● Find best candidates for new roles - data curators and data stewards

Create cross-functional team
● Data consumers - will depend upon project
● Data Engineer(s)
● Data Curator
● Data Steward

Choose your operating model
● Start with Shared Services for first project

Ensure executive alignment
● CDO or equivalent
Getting Started - Technology

Identify path to a modern, modular service architecture
- Create blueprint for next generation data management platform
- Revisit cloud migration strategy

Inventory current tool set
- TCO / skill requirements / etc.
- Determine which should be replaced, and when this is viable

Decouple monolithic processes
- Wrap components in APIs, expose as services

Start building with new tech
- Choose subset of tools for proof of concepts to replace old tech
In Parting - What NOT to do

- Avoid **boil the ocean/“waterfall”** (projects measured in years/quarters)
  - Build rational long term infra while delivering real analytic value along the way

- **Single “Platform”**: Don’t overestimate what single piece of software can do
  - Focus on thoughtfully designed ecosystem of loosely coupled best of breed tools

- **Single Vendor**: Don’t overestimate what single vendor can do
  - Align vendors with APIs and expectations that they MUST work together

- **Don’t Underestimate effort** required to make FOSS work
  - Just because Google does it doesn’t mean you can do it

- Don’t **underestimate human/behavioral challenges with data**
  - Most often the reason that projects fail/stall are human/behavioral
Appendix - Technology Examples
Technology - Movement
ETL/ELT
Technology - Mastering/Quality
Technology - Publish: Citizens
Technology - Publish: Analysts
Technology - Publish: Data Scientists
Technology - Publish: Developers

## Technology - Feedback

### Open Issues (11)  Clear filters

<table>
<thead>
<tr>
<th>Status</th>
<th>Reporter</th>
<th>Assignee</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**57 Members**

<table>
<thead>
<tr>
<th>Name</th>
<th>Email</th>
<th>Joined</th>
<th>Reported</th>
<th>Resolved</th>
<th>Assigned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lisa Skinner</td>
<td><a href="mailto:lisa.skinner@tamr.com">lisa.skinner@tamr.com</a></td>
<td>March 13, 2019 at 10:10 pm</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mike Gurica</td>
<td></td>
<td></td>
<td>0</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>ihab.ijisa@</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Son hgo</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Edward Cox</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Dave Theba</td>
<td></td>
<td></td>
<td>0</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Kevin Burke</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>scott.demer</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Todd Vendt</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Michael Guv</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Report data issue to Tamr Steward**

- **Workspace:** Tamr
- **Title:** Give a descriptive title
- **Comment:** Type @ to mention someone
- **Assignee:** Assign...

**Send a reminder?**
- 0
- 0
- 0

**End a reminder?**
- 0
- 0
- 0

Confidential 44
Technology - Governance