Seven Steps to High-Velocity Data Analytics with DataOps

2017-03-16 • San Jose, CA
Agenda

DataKitchen Summary

Analytic Landscape

Seven Shocking Steps

Case Study
DataKitchen Executive Summary

• Who we serve:
  • IT & Analytic Teams

• What we do:
  • World’s first company focused on enabling DataOps

• How we do it:
  • DataOps Software Platform

• Why DataOps matters? It allows analytic teams to:
  • Deliver insight fast
  • With high quality
  • Using the tools they love
  • Reuse what they’ve created
  • Continuously show value to their business customers
The world changed February 2005

“I am competing with Amazon”

-- Director of Analytics
Brand team support
Big Pharma Company
Expectations are going higher
Your competitors are moving fast

WATERFALL ➡ AGILE ➡ DEVOPS RELEASE FREQUENCY

12 Months
3 Months
3 Weeks
1 Week
11 Seconds

Average Release Frequency Over Time


From the authors of The Visible Ops Handbook

The Phoenix Project
A Novel About IT, DevOps, and Helping Your Business Win
Gene Kim, Kevin Behr, and George Spafford

Copyright © 2017 by DataKitchen, Inc. All Rights Reserved.
Your customers expect high quality

HAND MADE ➞ MASS PRODUCTION ➞ LEAN MANUFACTURING
How do you deliver on high expectations on speed and quality?

The answer is **DataOps**

DataOps combines Agile development, DevOps and statistical process controls and applies them to data analytics.
Seven Steps to Implement DataOps

1. Add Data and Logic Tests
2. Use a Version Control System
3. Branch and Merge
4. Use Multiple Environments
5. Reuse & Containerize
6. Parameterize Your Processing
7. Use Simple Storage
1. Add Data And Logic Tests

Are data inputs free from issues?

Is your business logic still correct?

Are you outputs consistent?

Access: Python Code

Transform: SQL Code, ETL Code

Model: R Code

Visualize: Tableau Workbook XML

Report: Tableau Online

And Save Test Results!
## Example Tests

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Verifying the inputs to an analytics processing stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count Verification - Check that row counts are in the right range, ...</td>
<td></td>
</tr>
<tr>
<td>Conformity - US Zip5 codes are five digits, US phone numbers are 10 digits, ...</td>
<td></td>
</tr>
<tr>
<td>History - The number of prospects always increases, ...</td>
<td></td>
</tr>
<tr>
<td>Balance - Week over week, sales should not vary by more than 10%, ...</td>
<td></td>
</tr>
<tr>
<td>Temporal Consistency - Transaction dates are in the past, end dates are later than start dates, ...</td>
<td></td>
</tr>
<tr>
<td>Application Consistency - Body temperature is within a range around 98.6F/37C, ...</td>
<td></td>
</tr>
<tr>
<td>Field Validation - All required fields are present, correctly entered, ...</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Business Logic</th>
<th>Checking that the data matches business assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Customer Validation - Each customer should exist in a dimension table</td>
</tr>
<tr>
<td></td>
<td>Data Validation - 90 percent of data should match entries in a dimension table</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Output</th>
<th>Checking the result of an operation, for example, a cross-product join</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Completeness - Number of customer prospects should increase with time</td>
</tr>
<tr>
<td></td>
<td>Range Verification - Number of physicians in the US is less than 1.5 million</td>
</tr>
</tbody>
</table>
2 Use a Version Control System

Access: Python Code
Transform: SQL Code, ETL Code
Model: R Code
Visualize: Tableau Workbook XML
Report: Tableau Online

At the end of the day, Analytic work is all just code

Source Code Control
Branching & Merging enables people to safely work on their own tasks.
4 Use Multiple Environments

Your Analytic Work Requires Coordinating Tools And Hardware

Copyright © 2017 by DataKitchen, Inc. All Rights Reserved.
Provide an Analytic Environment for each branch

- Analysts need a controlled environment for their experiments
- Engineers need a place to develop outside of production
- Update Production only after all tests are run!
5. Reuse & Containerize

Reuse
1. Do not create one ‘monolith’ of code
2. Reuse the code and results

Containerize
1. Manage the environment for each component (e.g. Docker, AMI)
2. Practice Environment Version Control
Parameterize Your Processing

• Parameters and named sets of parameters will increase your velocity
• With parameters, you can vary
  • Inputs [you can make a time machine]
  • Outputs
  • Steps in the workflow
Use Simple Storage

- **Data Lake**
  - Keep copies of all your raw data in simple, cheap storage (s3, HFDS, file system)
  - Data Restore: Be able to back up and restore your databases easily

- **“My Own Database”: Data Marts On Demand**
  - Create parameterized variations of your process that allow you to assemble data for experimentation, development, and production
DataOps case study
Analytics Team supporting Sales and Marketing in a Pharmaceutical Company

• One Data Engineer
  • Supports 12 analysts
  • Makes 12 schema changes a week without breaking anything
• One Data Analyst supports hundreds of sales people
• Other Data Analysts make visualization changes and publish them the next day
DataKitchen Executive Summary

• **Who we serve:**
  • IT & Analytic Teams

• **What we do:**
  • World’s first company focused on enabling **DataOps**

• **How we do it:**
  • DataOps Software Platform

• **Why DataOps matters? It allows analytic teams to:**
  • Deliver insight fast
  • With high quality
  • Using the tools they love
  • Reuse what they’ve created
  • Continuously show value to their business customers