Now wait a minute?

You Probably Aren't Monitoring Enough

Casey Stella

January 29, 2018

Hi, I'm Casey Stella!

Monitoring: My Observations

- **Most** people monitor machines

Monitoring: My Observations

- **Most** people monitor machines
  - Capability exists in dev ops and infrastructure groups

Monitoring: My Observations

- **Most** people monitor machines
  - Capability exists in dev ops and infrastructure groups
- **Some** people monitor applications

Monitoring: My Observations

- **Most** people monitor machines
  - Capability exists in dev ops and infrastructure groups
- **Some** people monitor applications
  - Most application monitoring is focused on performance

Monitoring: My Observations

- **Most** people monitor machines
  - Capability exists in dev ops and infrastructure groups
- **Some** people monitor applications
  - Most application monitoring is focused on performance
  - You should be monitoring for correctness too

Monitoring: My Observations

- **Most** people monitor machines
  - Capability exists in dev ops and infrastructure groups
- **Some** people monitor applications
  - Most application monitoring is focused on performance
  - You should be monitoring for correctness too
- **A few** people monitor their data pipelines

Monitoring: My Observations

- **Most** people monitor machines
  - Capability exists in dev ops and infrastructure groups
- **Some** people monitor applications
  - Most application monitoring is focused on performance
  - You should be monitoring for correctness too
- **A few** people monitor their data pipelines
  - It's almost always "did it complete or not?"

Monitoring: My Observations

- **Most** people monitor machines
  - Capability exists in dev ops and infrastructure groups
- **Some** people monitor applications
  - Most application monitoring is focused on performance
  - You should be monitoring for correctness too
- **A few** people monitor their data pipelines
  - It's almost always "did it complete or not?"

Monitor Your Data

- Ingesting Data without ongoing monitoring is like deploying code without unit tests

Monitor Your Data

- Ingesting Data without ongoing monitoring is like deploying code without unit tests
  - You want to ensure regressions of the assumptions of your data don't happen

Monitor Your Data

- Ingesting Data without ongoing monitoring is like deploying code without unit tests
  - You want to ensure regressions of the assumptions of your data don't happen
  - Things like completeness and consistency are key

Monitor Your Data

- Ingesting Data without ongoing monitoring is like deploying code without unit tests
  - You want to ensure regressions of the assumptions of your data don't happen
  - Things like completeness and consistency are key
- Your downstream analytics should **define** your data monitoring

Monitor Your Data

- Ingesting Data without ongoing monitoring is like deploying code without unit tests
  - You want to ensure regressions of the assumptions of your data don't happen
  - Things like completeness and consistency are key
- Your downstream analytics should **define** your data monitoring
  - Check pre-conditions: Assumptions about data consistency (e.g. seasonality, completeness, distribution)

Monitor Your Data

- Ingesting Data without ongoing monitoring is like deploying code without unit tests
  - You want to ensure regressions of the assumptions of your data don't happen
  - Things like completeness and consistency are key
- Your downstream analytics should **define** your data monitoring
  - Check pre-conditions: Assumptions about data consistency (e.g. seasonality, completeness, distribution)
  - Check post-conditions: Assumptions about analytic consistency (e.g. distribution of output for machine learning)

Monitor Your Data

- Ingesting Data without ongoing monitoring is like deploying code without unit tests
  - You want to ensure regressions of the assumptions of your data don't happen
  - Things like completeness and consistency are key
- Your downstream analytics should **define** your data monitoring
  - Check pre-conditions: Assumptions about data consistency (e.g. seasonality, completeness, distribution)
  - Check post-conditions: Assumptions about analytic consistency (e.g. distribution of output for machine learning)
  - This should be a **shared responsibility** between data science and data engineering teams.

Key Takeaway

Any data ingestion plan should include a plan for ongoing monitoring or else you're courting disaster...or at least incorrect results.