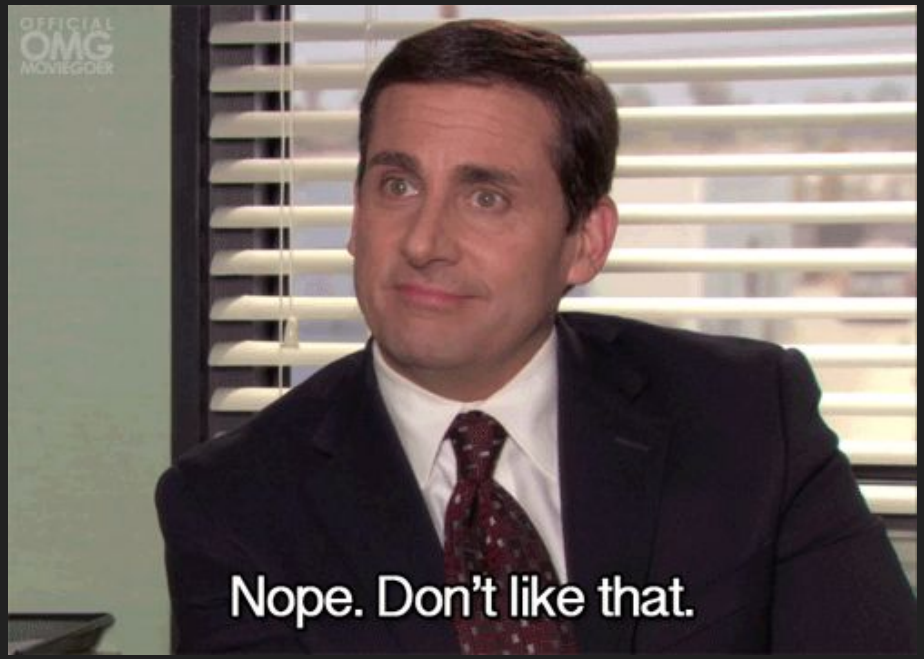
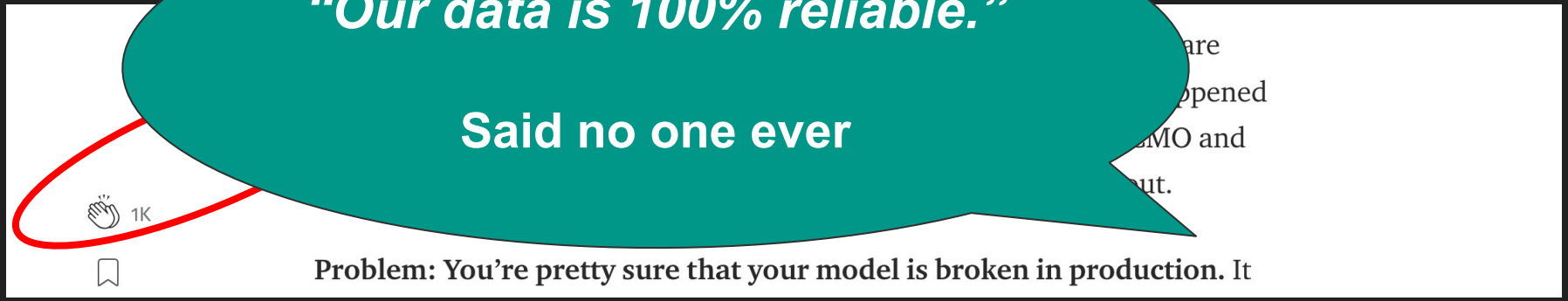
Introducing Data Downtime: From Firefighting to Winning

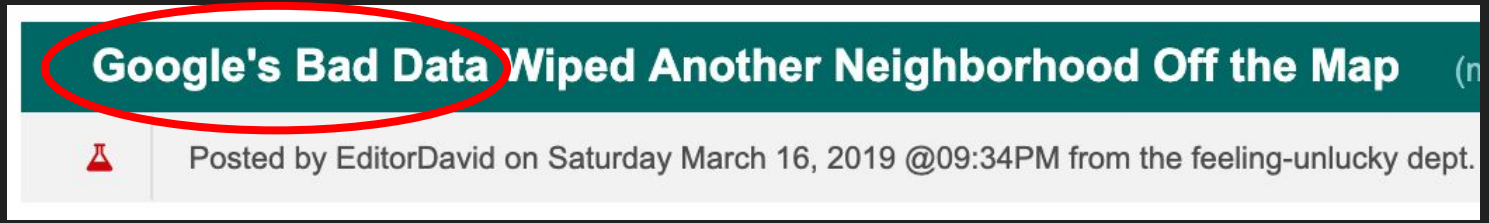
Barr Moses

When your CEO or customer says the data is wrong...

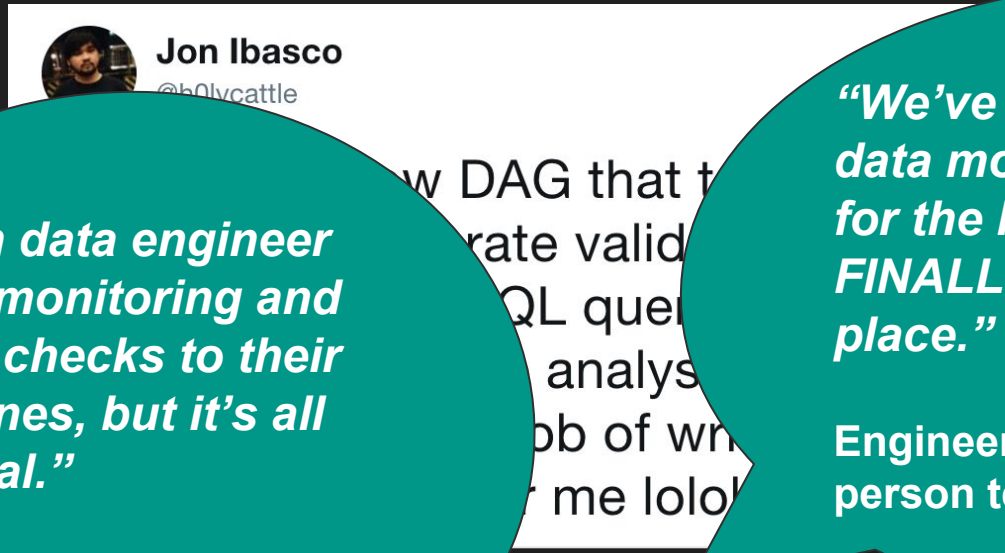


Unreliable data is more prevalent than we admit





Current solutions take heroic (or tedious) efforts



“Each data engineer adds monitoring and basic checks to their pipelines, but it’s all manual.”

Engineering Manager, 1500 person tech co

“We’ve been building out data monitoring solutions for the last year and FINALLY we’re in a good place.”

Engineering Manager, 300 person tech co

“Data downtime” refers to periods of time when your data is partial, erroneous, missing or otherwise inaccurate.

Signs you're experiencing data downtime


Your data team spends >10% of time on fire drills


Your company lost $$ because data was broken

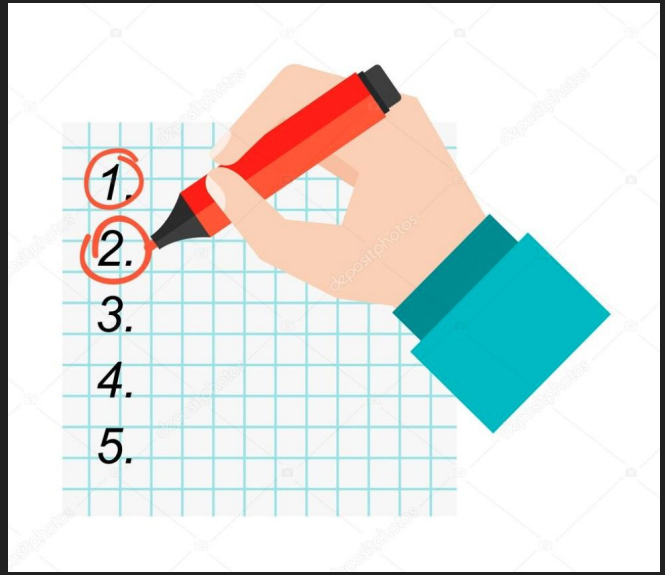

Critical analysis fails because missing data went unnoticed


Troubleshooting involves tedious step-by-step debugging

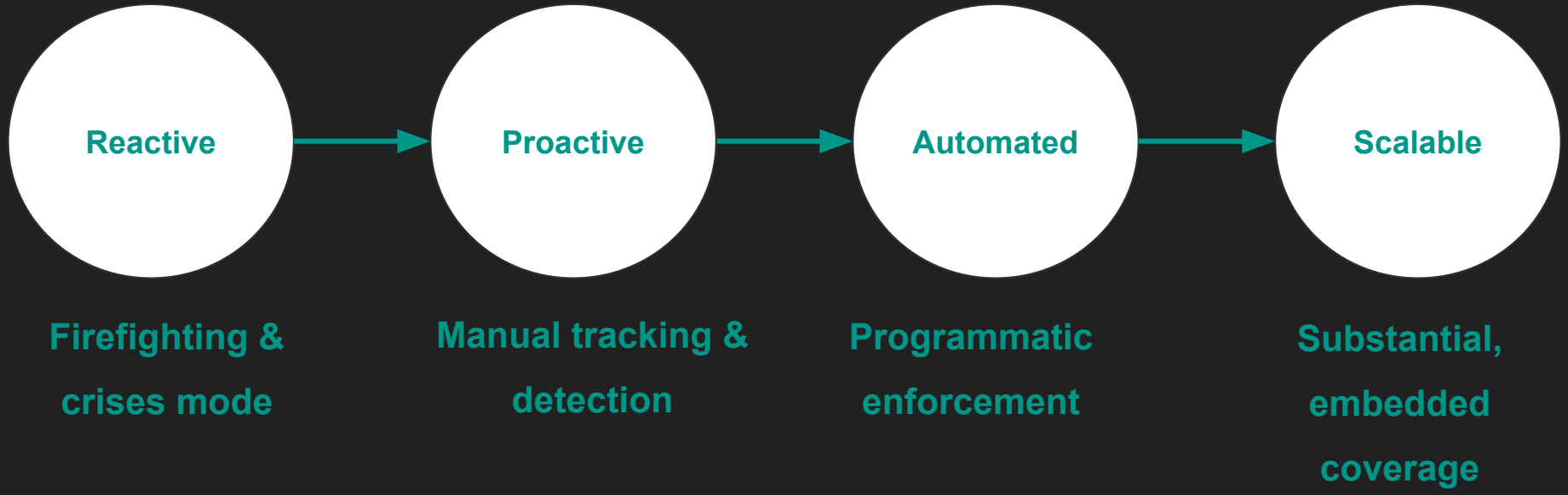
What can we do about it?

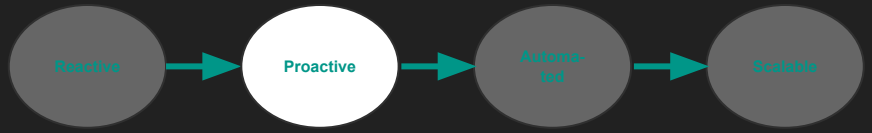
If you care about the data industry, make data downtime mitigation part of your culture

- Team SLAs
- Weekly meetings
- Post-mortems
- Pay data debt



World-class companies mitigate data downtime





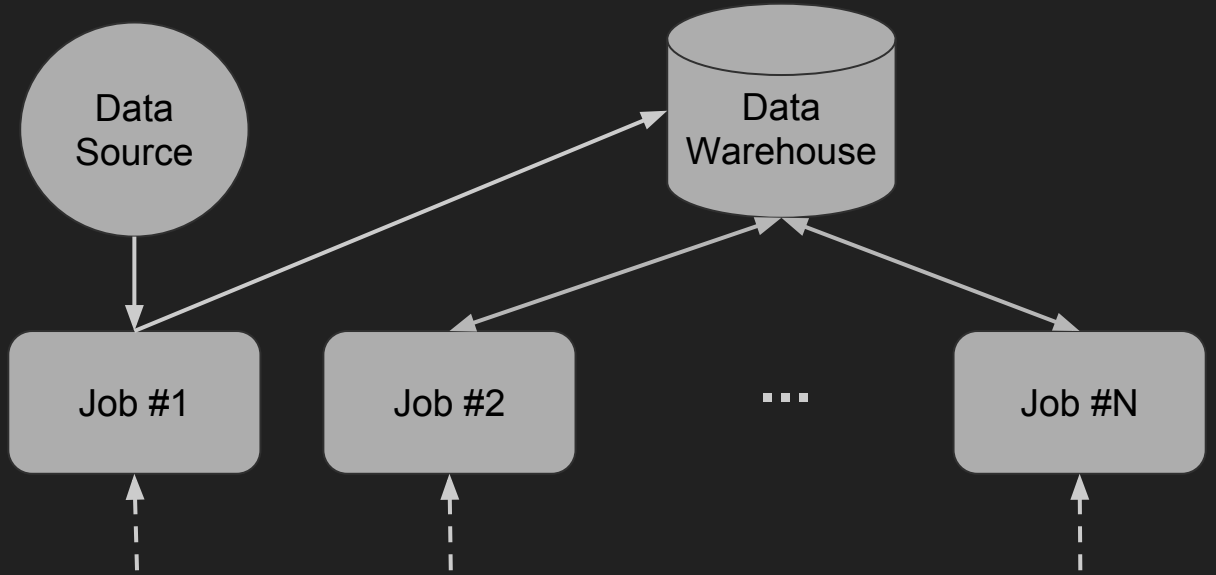
Proactive

Manual tracking & detection

Examples:

- Generate the data more frequently than you need it
- Manually write data QA queries and alert on it
- Track time stamps to ensure data isn’t stale
- Validate row counts in critical stages of pipeline

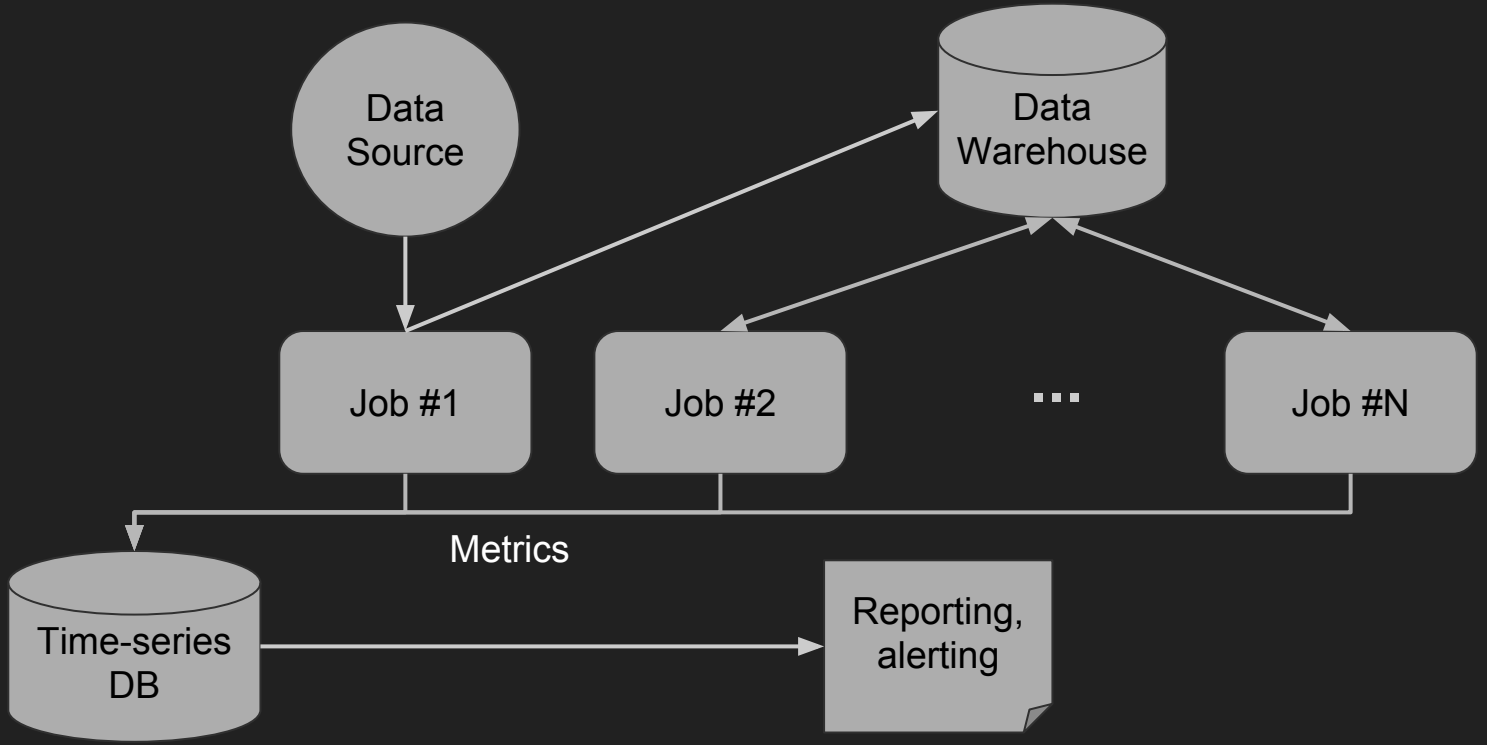
Example #1: Validate row counts in critical stages of pipeline



Validate number of rows for key tables (e.g. website visitors) is within reasonable range

Consider failing the job vs. generating a warning/alert

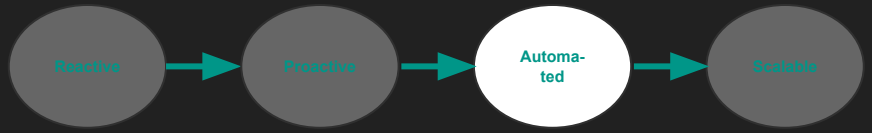
Example #1: Validate row counts in critical stages of pipeline



Example #1: Validate row counts in critical stages of pipeline

- **Particularly effective when**
  - Row count is somewhat predictable
  - Problems tend to result in substantial missing or duplicated data

- **Limitations**
  - Could require tuning and maintenance
  - Sensitive to seasonality
  - Doesn’t catch all issues



Automated

Programmatic enforcement

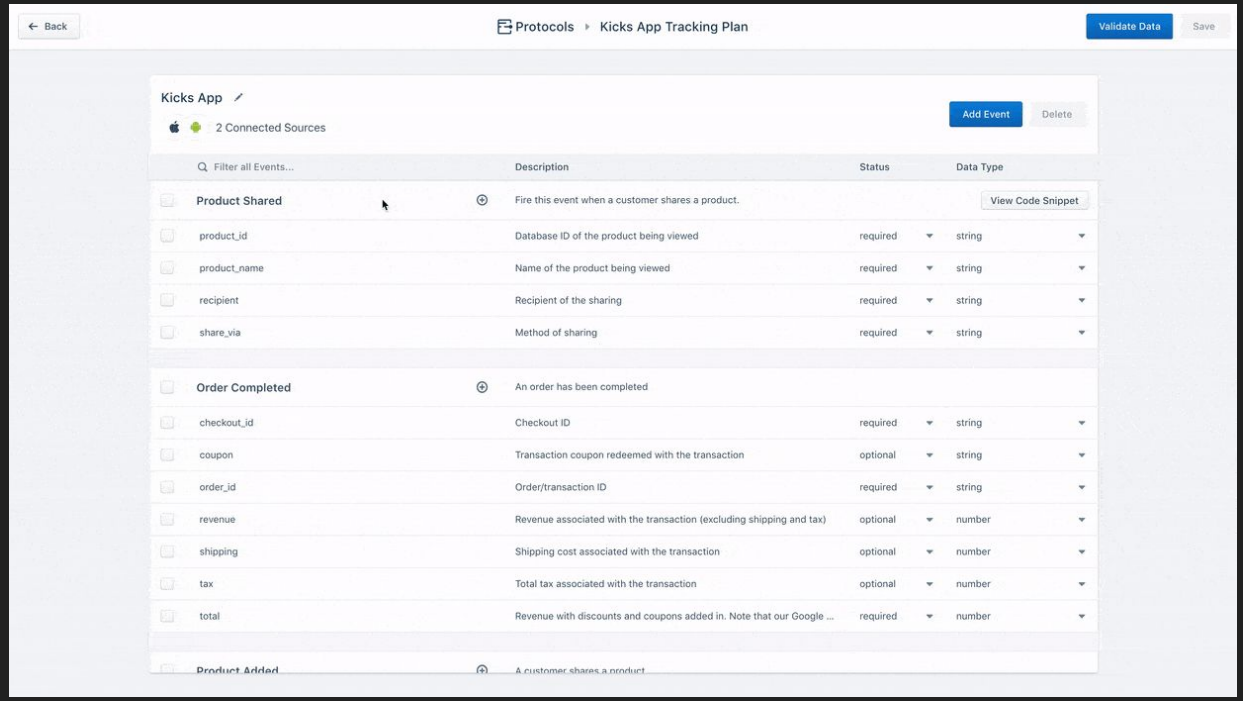
Examples:

- Automatically track metrics about dimensions & measures, compare to past periods
- Create a data health dashboard
- Document fields and tables
- Monitor/enforce schema and validity of upstream

Example #2: Monitor/enforce schema and validity of upstream

- **Why**
  - Data sources change (e.g. engineering pushes schema change)
  - Data team does not control and has no visibility
- **How**
  - Define expected schema
  - Validate data against schema before processing
  - Discard or alert on data that doesn’t match

Example #2: Monitor/enforce schema and validity of upstream



Example #2: Monitor/enforce schema and validity of upstream

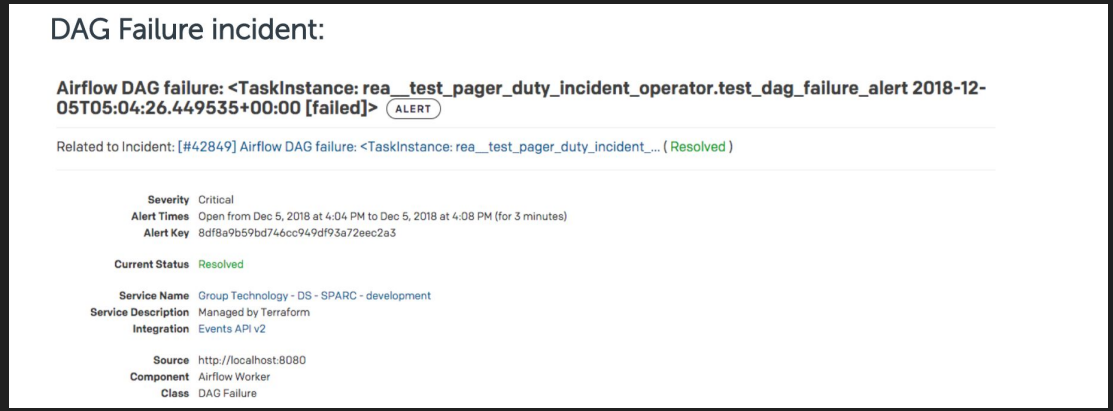
- Bonus: validate data, not just schema (e.g. great expectations open source)

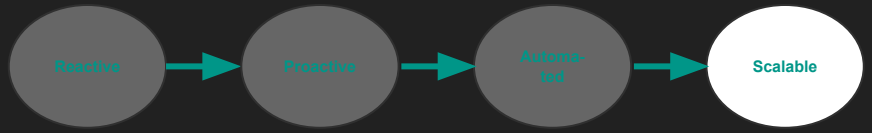
```
>> my_df.expect_column_values_to_be_between(
    "my_column",
    min_value=0,
    max_value=7,
    mostly=0.7
)
{
  "success": true
  "result": {
    "unexpected_percent": 0.2,
    "partial_unexpected_index_list": [
      8,
      9
    ],
    "partial_unexpected_list": [
      8,
      9
    ],
    "unexpected_percent_nonmissing": 0.2,
    "unexpected_count": 2
  }
}
```

Source: https://github.com/great-expectations/great_expectations

Example #2: Monitor/enforce schema and validity of upstream

- Bonus: automate ownership & page the right owner (e.g. pagerduty alerts)



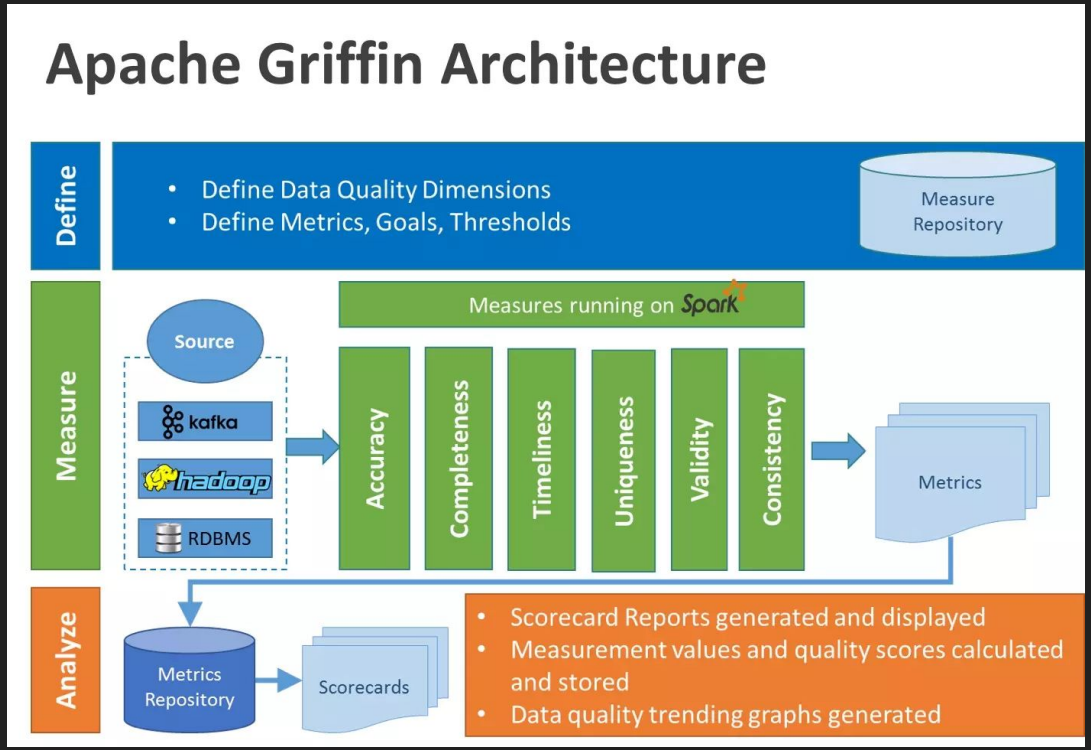


Substantial, embedded coverage

Examples:

- Validate measures across different tables, sources
- Track and annotate data issues and questions
- Create reusable components to calculate metrics
- Setup a staging environment that closely resembles production
- Embed automated data validation code across your pipeline

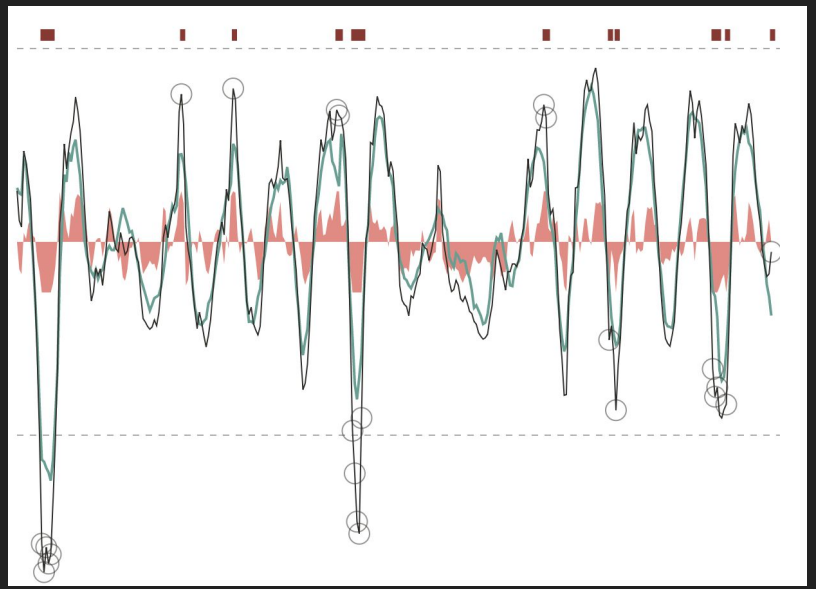
Example #3: Embed automated data validation code across your pipeline



Example #3: Embed automated data validation code across your pipeline

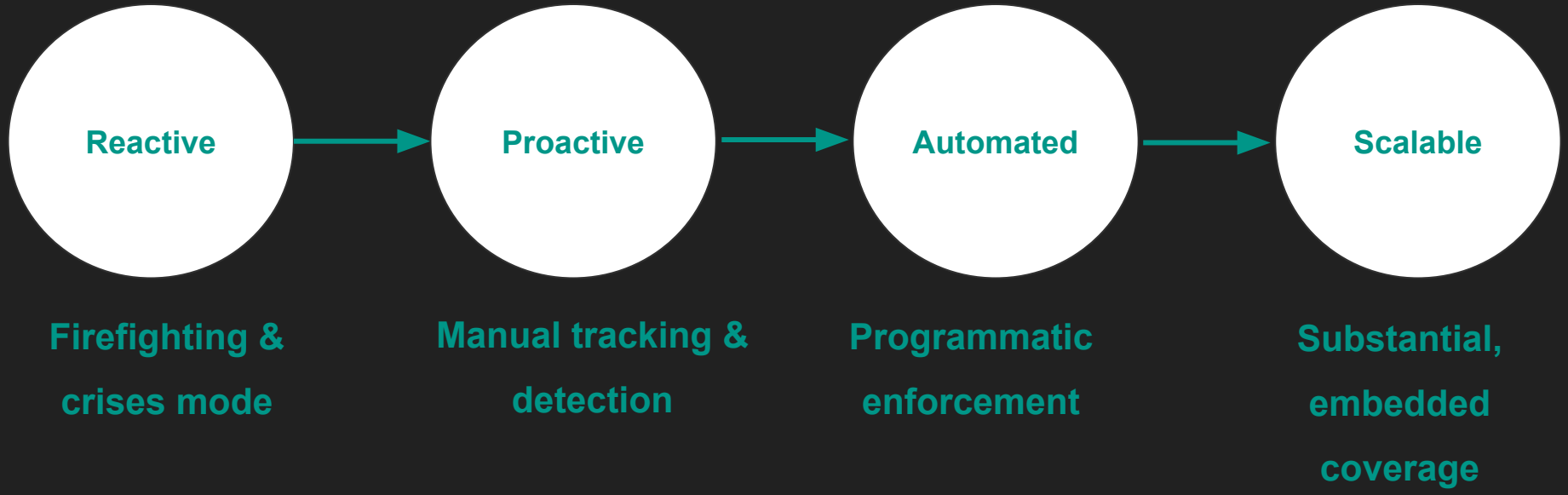
Case in Point: Netflix's RAD

- Need to validate 10000s of metrics, with seasonality
- RPCA algorithm to detect time series anomalies
- Pig wrapper allows pipeline engineers to validate key metrics
- Success with (a) transaction data per banking institution (b) sign up conversion by country and browser



Source: https://medium.com/netflix-techblog/rad-outlier-detection-on-big-data-d6b0494371cc

World-class companies mitigate data downtime



Your company will kickass if you figure this out

- Minimize fire drills
- Gain confidence
- Make better decisions
- Move faster

Join the data downtime movement.

Connect with me @barrmoses on Medium or LinkedIn