DataKitchen

DataKitchen Webinar: Actionable, Automated, Agile Data Quality Scorecards

February 19, 2025

Christopher Bergh
CEO, Head Chef
DataKitchen

DataKitchen

Before We Start

Today's slides and recording will be shared this week.

Put your questions in the chat window; we will answer questions at the end of the session.

We have targeted 45 minutes for our presentation, with questions at the end.

We may go long …

Agenda

- **Start DataOps With Data Quality**
- Challenges of Data Quality and Data Quality Leadership
- Announcing: Actionable, Automated, Agile Data Quality Scorecards
- Demonstration: Open Source DataOps Data Quality TestGen
- How To DataOps Your Data Quality
- Conclusion

#1 Problem In Data & Analytics: **Waste**

Wasted Time, Energy & Trust

60% of projects **fail**

– Gartner

79% have too many **errors**

– Eckerson

73% of data practitioners do not **trust** the their data *– IDC*

78% of data teams are **stressed** & want therapy

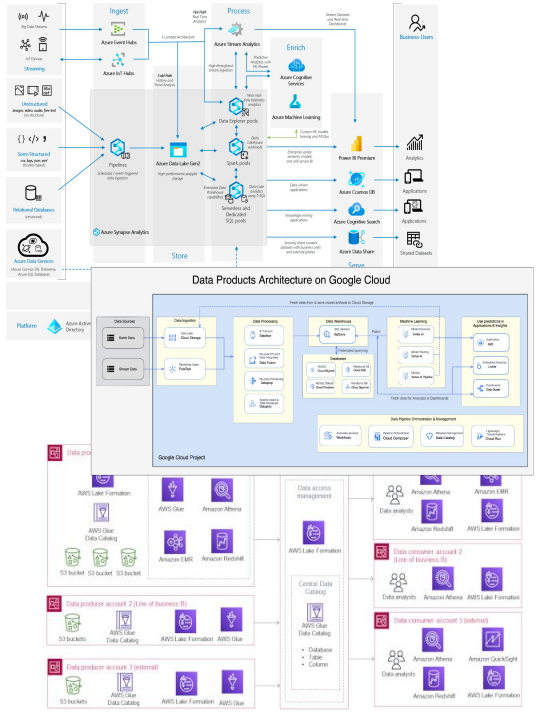
-DataKitchen

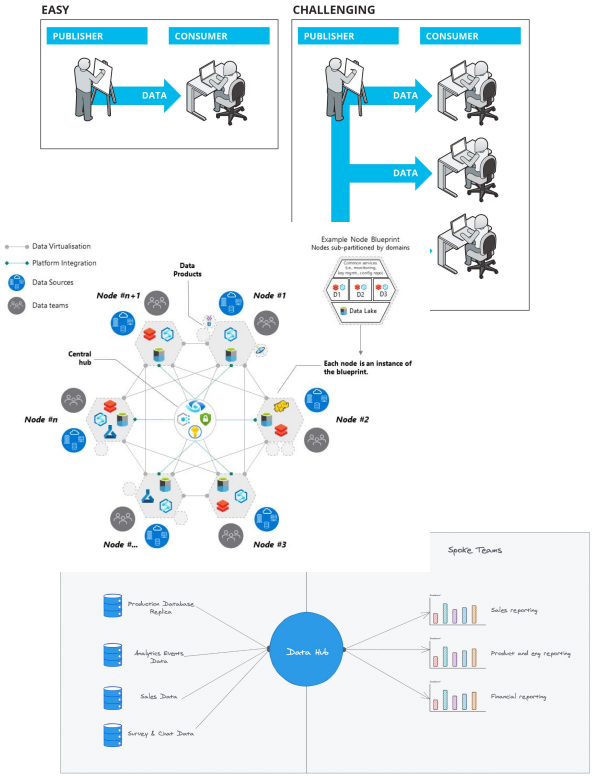
#2 Problem: **Little Boxes Everywhere**

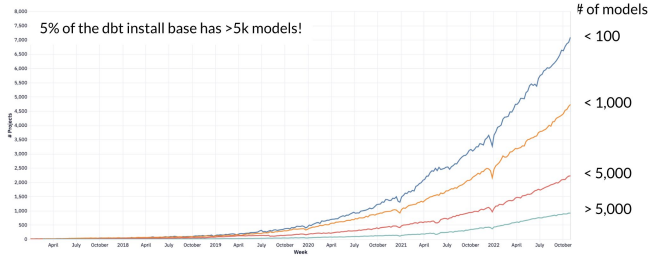
Data Stack Sprawl; Organizational Silos, Data Set Exploding

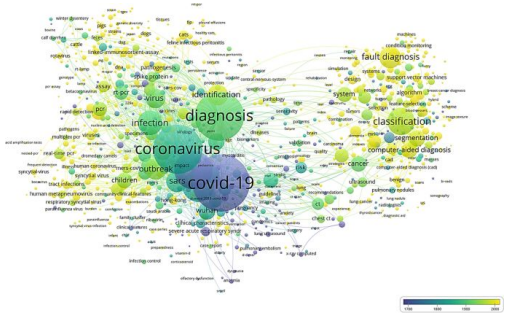
DataKitchen

Data Architecture Boxes	Team Organizational Boxes	Diverse Data Boxes









Why Waste, Failure, & Poor Results?

DataKitchen

Poor Data Quality ⇨

Complex & Fragile Toolchain ⇨

⇦ **Too Much To Do**

⇦ **Customer Visible Errors**

Why Waste, Failure, & Poor Results?

DataKitchen

Poor Data Quality ⇨

Complex & Fragile Toolchain ⇨

'Day 1' Focus:
Immediate Tasks & Building With Individual Tools

⇦ **Too Much To Do**

⇦ **Customer Visible Errors**

How To Stop Waste, Failure, & Poor Results?

DataKitchen

Day 2 Focus:

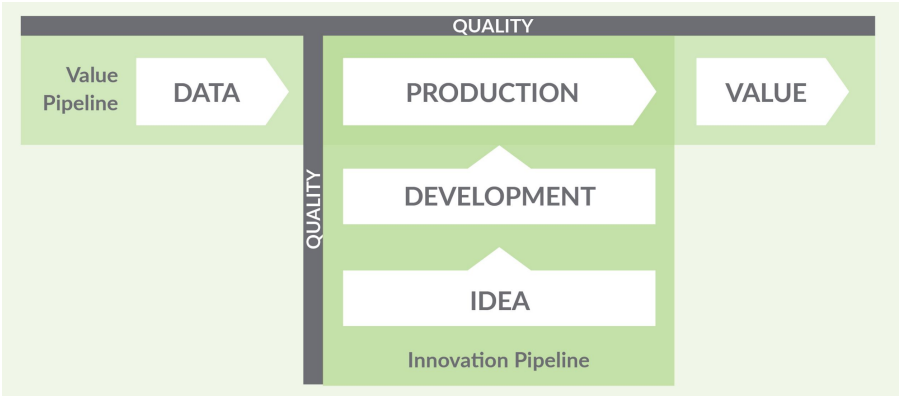
DataOps: Optimize The System Of People, Tools, Data, Work Process And Deliverables

The Answer Is DataOps

DataKitchen

*" teams guided by DataOps practices and tools will be **ten times more productive** than teams that do not use DataOps. " – **Gartner***

Increase Quality & Decrease Production Errors

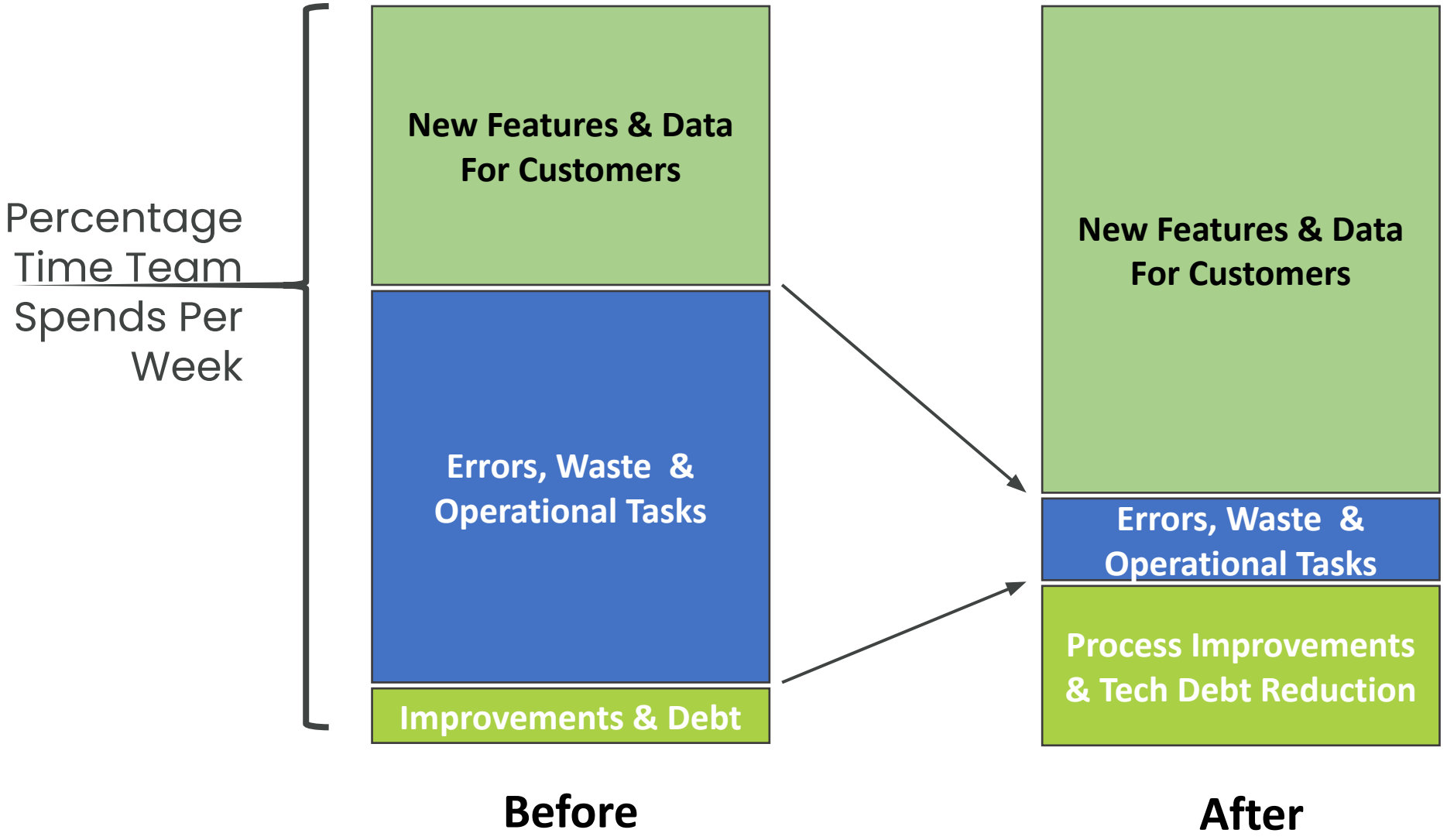


Increase Development Cycle Time

Result
Less Waste
10x Team Productivity Improvement
Improved Trust

DataOps Benefit: Time Well Spent

DataKitchen

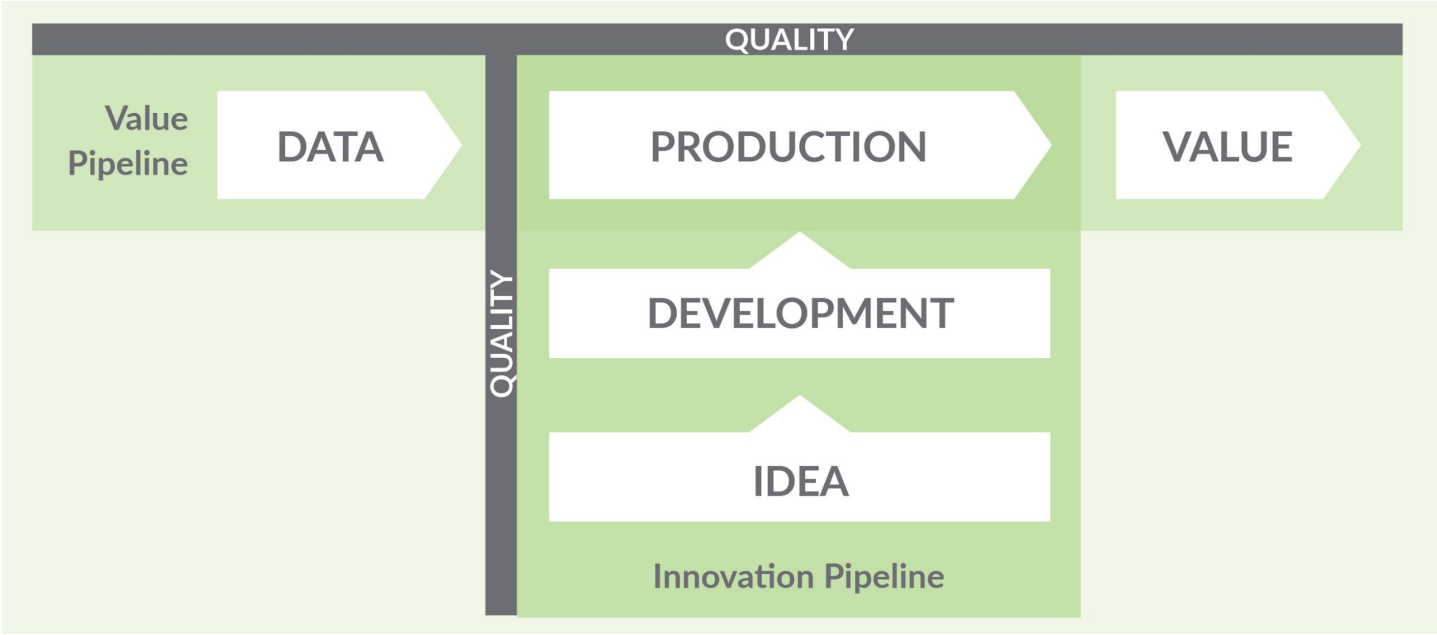


Where To Start DataOps

1. **Improve Data Quality**

2. **Stop Production Errors**

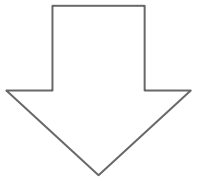
3. **Automate For Faster, Safer Deployment**



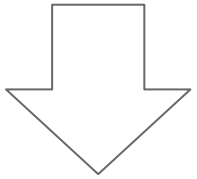
Where To Start DataOps

DataKitchen

Improve Data Quality



Stop Production Errors



Automate For Faster, Safer Deployment

Open Source Data Observability

DataOps Data Quality TestGen
Simple, Fast Data Quality Test Generation and Execution

Generative Data Quality

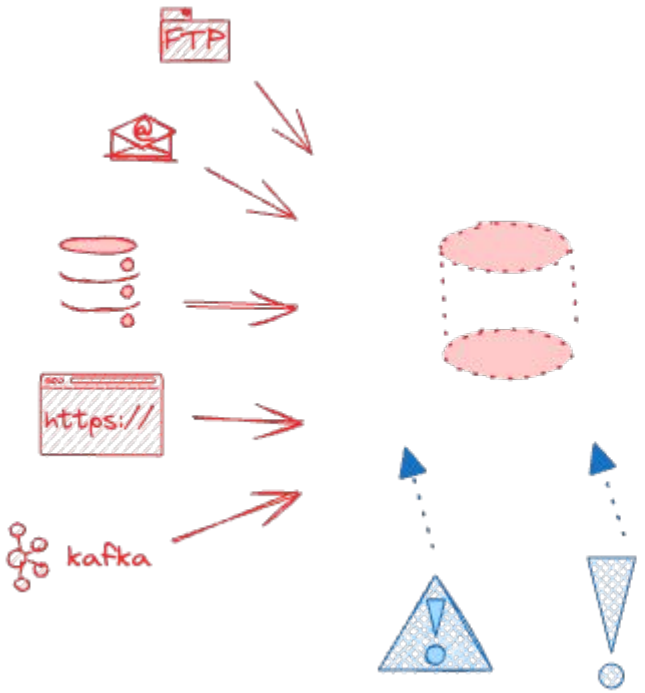
DataOps Observability
Data Journey Mission Control From Data Source To Customer Value

Anticipate, Track Production Errors Across the Whole Data Estate

DataOps Automation
Orchestrate, Manage And Test Your Complex Data Toolchain

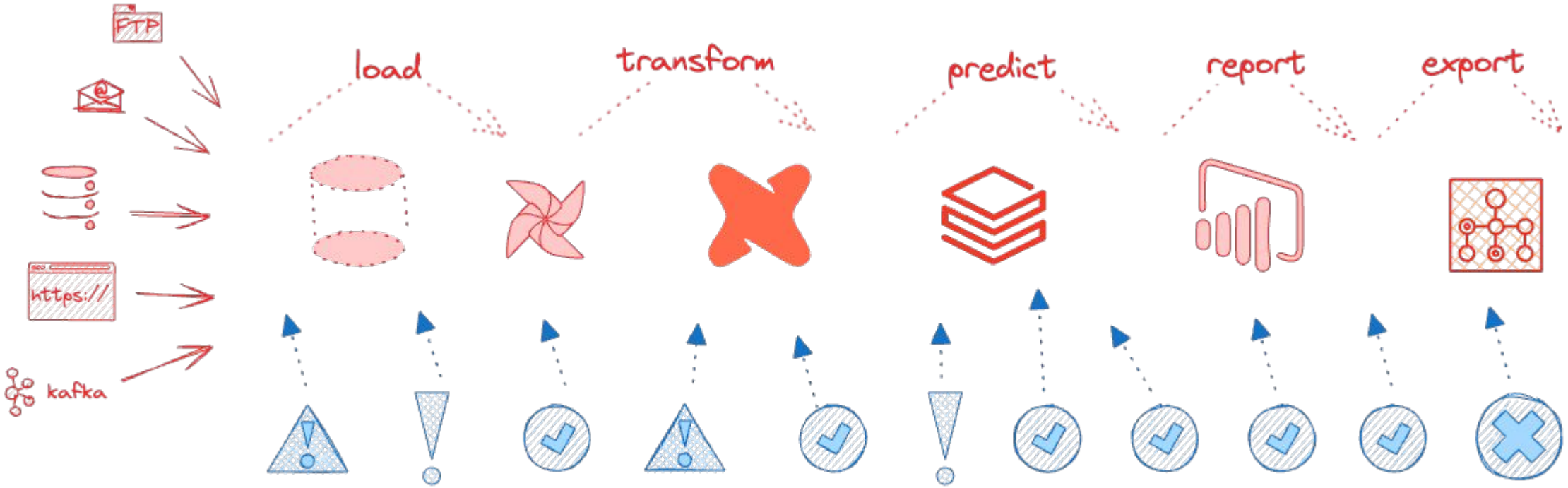
Reduce Data Team Cycle Time & Increase Productivity

Quality Problems Start In Data



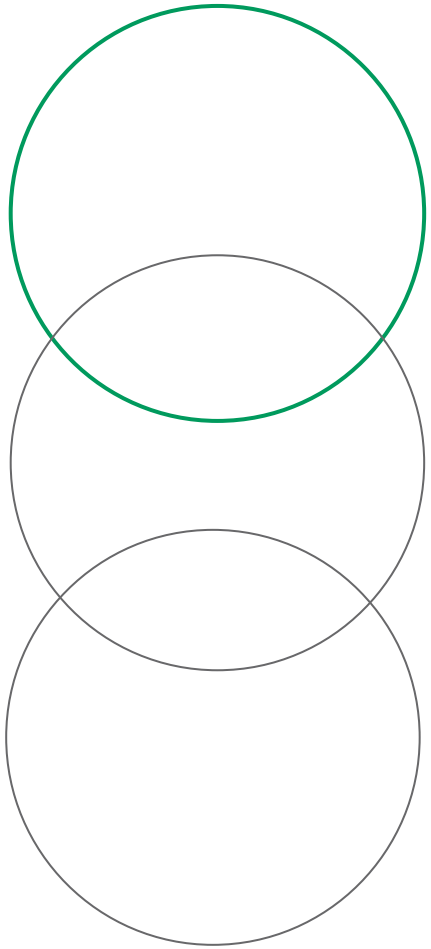
Quality Problems Are Not Just In Data

DataKitchen



Overlapping Roles In Quality

DataKitchen



Improve Data Quality

Data Quality Team

Stop Production Errors

Data Engineering Team

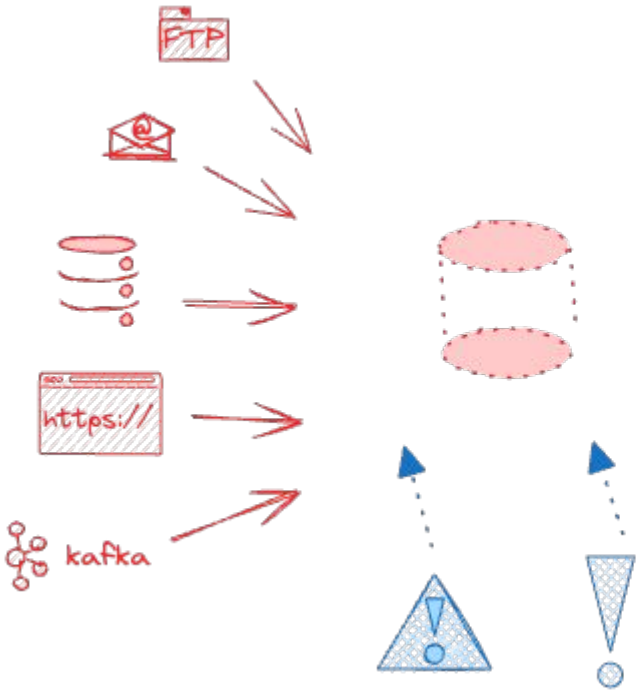
Automate For Faster, Safer Deployment

Entire Data Team

Focus Today Is On Data Quality Team

DataKitchen

Improve Data Quality



Data Quality Team

Agenda

- Start DataOps With Data Quality
- **Challenges of Data Quality and Data Quality Leadership**
- Announcing: Actionable, Automated, Agile Data Quality Scorecards
- Demonstration: Open Source DataOps Data Quality TestGen
- How To DataOps Your Data Quality
- Conclusion

Data Quality

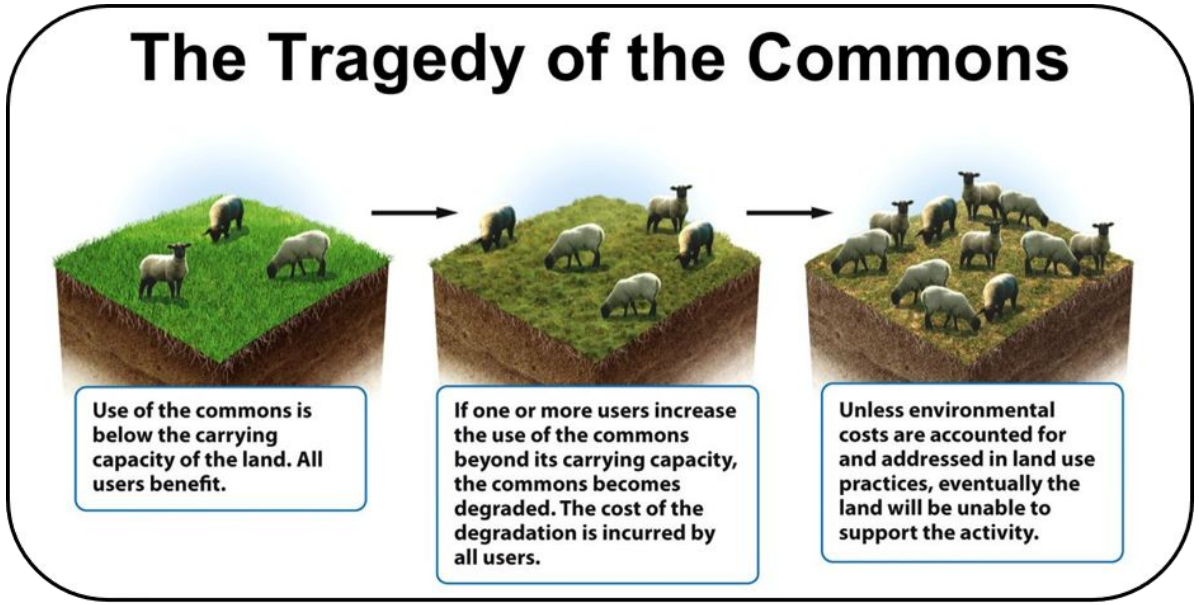
Data Quality is a comparison of the current state of your data vs. the desired state based on user expectations, usage requirements, and defined quality standards.

- When data quality is good, your data is **fit for its intended purpose** and conforms to the standards you've set.
- When the quality of your data is poor, it is out of compliance with established standards and unfit for use in operational and decision-making processes.

Data Quality remains a significant challenge for organizations.

- 57% of a 2024 dbt Labs survey respondents rated data quality as one of the three most challenging aspects of the data preparation process (up from 41%)
- 73% of data practitioners do not trust their data (IDC)
- Millions Lost In 2023 Due To Poor Data Quality, Potential For Billions To Be Lost With AI Without Intervention (Forrester)

Data Quality Is A ‘Tragedy Of The Commons’



Data Quality is an example of the Tragedy of the Commons.

How is data similar to similar to overgrazed land or polluted air?

- Shared Data Responsibility Leads to Neglect
- Short-Term Gains Over Long-Term Integrity
- Accumulation of Errors and Data Pollution
- Difficult to Reverse Once Degraded

Data Quality Leadership Is Frustrating

Data Quality Leaders are optimistic but frustrated.

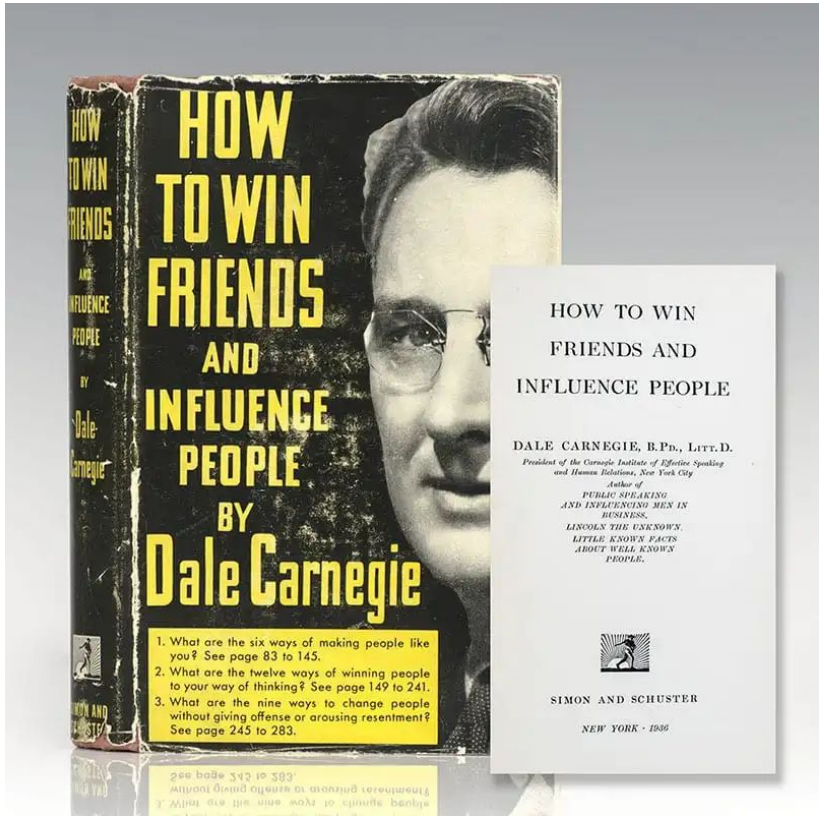
They find problems in data but need more direct power to change. '**Data Nags**'

They need to be focused to ensure they align with organizational goals.

Data Quality leaders have **influence but little power to cause change** in their organization.

This mismatch causes **frustration**!!

How to Win Friends and Influence People



Is a **1936 self-help book** written by Dale Carnegie.

Over **30 million copies** have been sold worldwide, making it one of the best-selling books of all time.

Why? It has proven principles to help people gain influence on others around them.

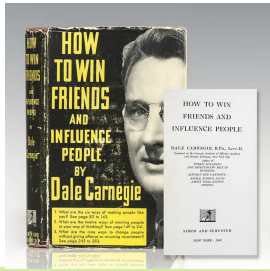
How does this help us as Data Quality Leaders?

Key Principles We Can Apply to Data Quality

Dale Carnegie Principles To Gain Influence And Reduce Frustration In Data Quality:

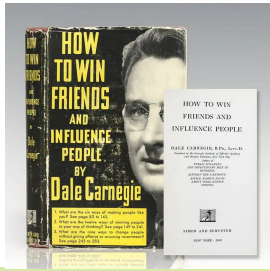
1. Don't Criticize, Condemn or Complain
8. Talk In Terms Of The Other Person's Interests
12. If You Are Wrong Admit It Quickly And Emphatically
18. Be Sympathetic With The Other Person's Ideas And Desires
19. Appeal To The Nobler Motives
29. Use Encouragement. Make The Fault Seem Easy To Correct
30. Make The Other Person Happy About Doing The Thing You Suggest

https://blog.synergyiq.com.au/blog-post/dale-carnegies-30-principles-to-win-friends-influence-people



How To Apply “Win … Influence People” to DQ

1. Don’t Criticize, Condemn or Complain
   Use Generative Data Quality To Automatically Identify Multiple Data Issues
8. Talk In Terms Of The Other Person’s Interests
   Build Data Quality Dashboards on Limited, Specific Data Items
12. If You Are Wrong Admit It Quickly And Emphatically
   Start Quickly, Small, and Work Iteratively: DataOps Data Quality
18. Be Sympathetic With The Other Person’s Ideas And Desires
   Give Them Specific Actionable Problems – They Are Busy!
19. Appeal To The Nobler Motives
   Make Multiple Data Quality Dashboards That Align To Organization Goals
29. Use Encouragement. Make The Fault Seem Easy To Correct
   Package Data Quality Issues So They Are Easy to Fix
30. Make The Other Person Happy About Doing The Thing You Suggest
   Measure And Show Data Quality Improvement Over Time

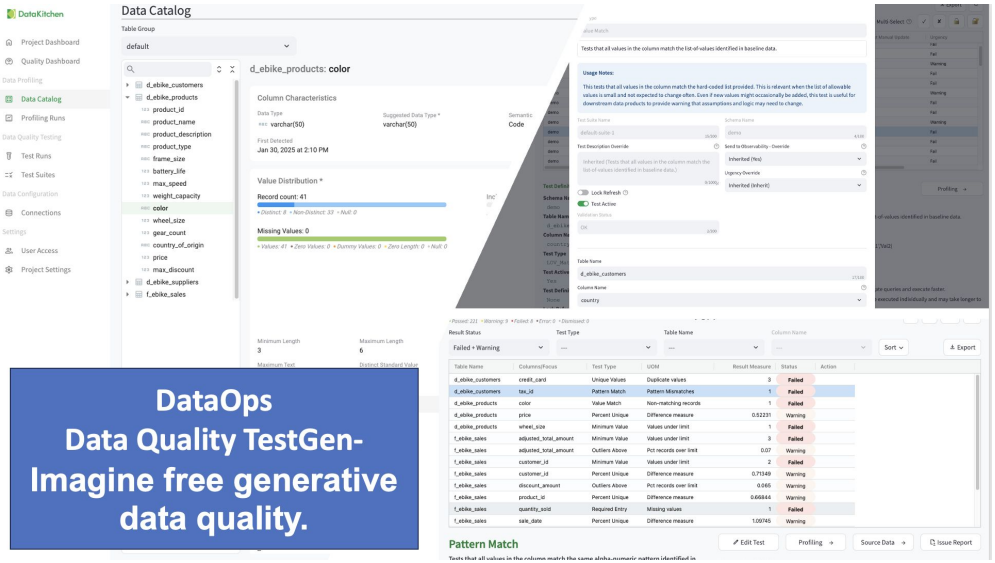


DataKitchen

Agenda

- Start DataOps With Data Quality
- Challenges of Data Quality and Data Quality Leadership
- **Announcing: Actionable, Automated, Agile Data Quality Scorecards**
- Demonstration: Open Source DataOps Data Quality TestGen
- How To DataOps Your Data Quality
- Conclusion

Dale Carnegie Data Quality & DataOps Data Quality TestGen



DataOps Data Quality TestGen:

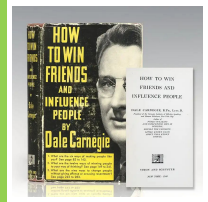
- Open Source, Full Featured, Data Quality Tool
- In Database Execution
- Full Featured (UI, AI, Rules) One User
- Enterprise Version starts at $100 per user per month

It Does Five Tasks:

1. Data Profiling
2. Dataset Screening And Hygiene Review
3. Algorithmic Generation of Data Quality Validation Tests
4. Ongoing Testing Of New Data Refreshes For Anomalies
5. ***Data Quality Scoring***

DataKitchen

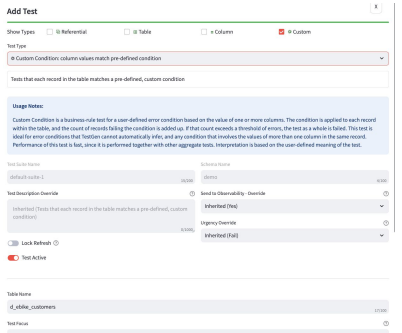
Dale Carnegie Data Quality Principle #1

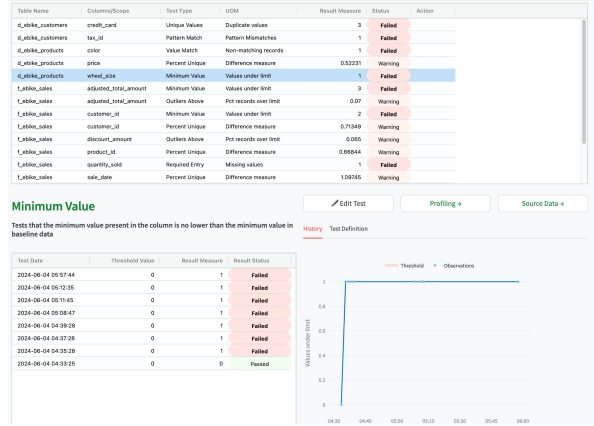


Don't Criticize, Condemn or Complain

Use Generative Data Quality To Automatically Identify Multiple Data Issues

DataOps Data Quality TestGen is a Generative Data Quality Application



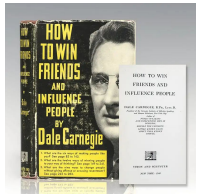


It Does Simple, Fast Data Quality Test Generation And Execution:

- **Fifty-One** Data Profiling Characteristics Collected
- **Twenty-Seven** Data Hygiene Detector Tests
- **Thirty-Two** Simple, Fast, Automatically Generated Data Quality Checks
- **Two** Custom Tests
- **Eight** Best Practice Data Validation Tests Configured With Fill-In-The-Blank Simplicity

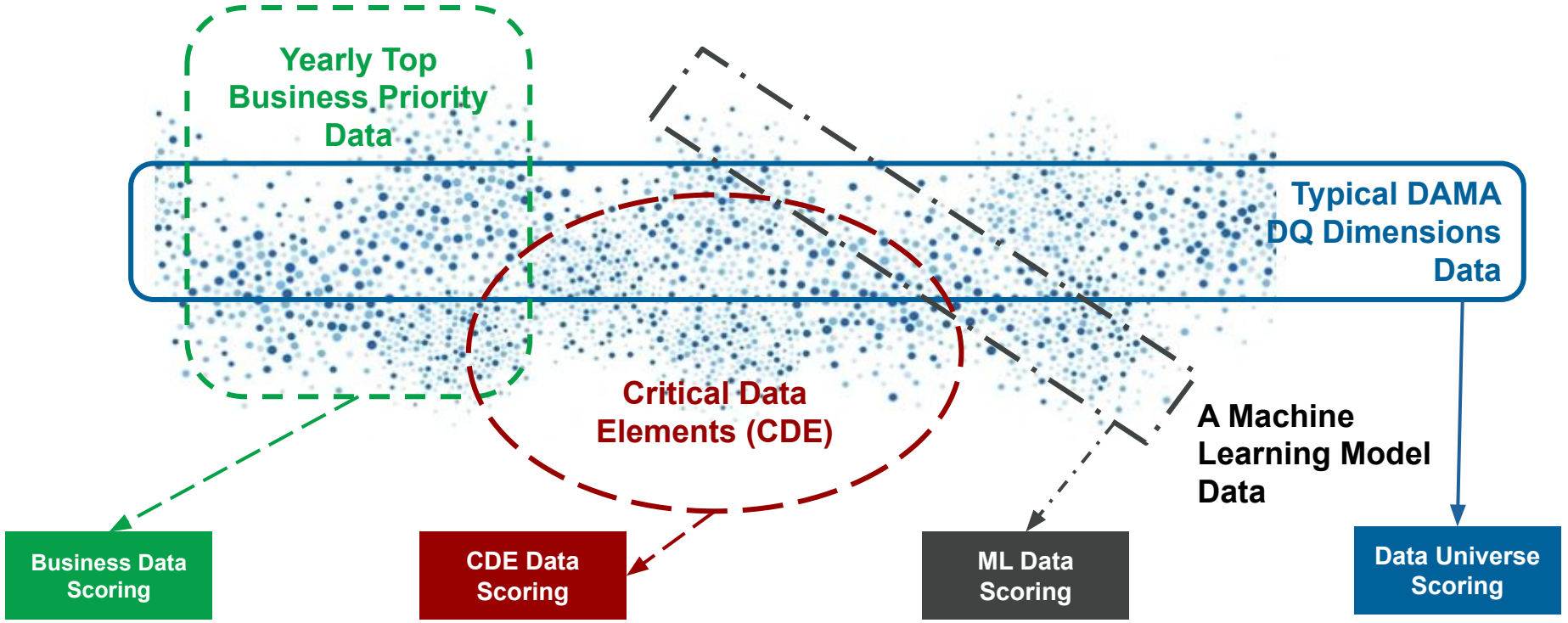
DataKitchen

Dale Carnegie Data Quality Principle #8



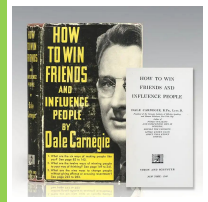
Talk In Terms Of The Other Person's Interests

Build Data Quality Dashboards on Limited, Specific Data Items



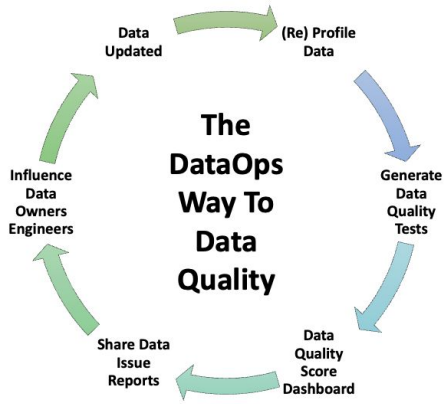
DataKitchen

Dale Carnegie Data Quality Principle #12



If You Are Wrong Admit It Quickly And Emphatically

Start Quickly, Small, and Work Iteratively: DataOps Data Quality

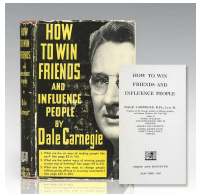


Agile/DataOps approach to Data Quality.

- Get a Data Quality process working right away, then iterate and improve.
- Generate 80% Data Quality Rules Automatically
- Start measuring/evaluating Data Quality before standards are established
- Then, use the resulting measurements to establish standards (where there are none) and improve them over time
- Cycle Quickly and maximize your learning as a Data Quality leader.

DataKitchen

Dale Carnegie Data Quality Principle #19



Appeal To The Nobler Motives

Make Multiple Data Quality Dashboards That Align To Organization Goals

Customer: VP of Sales

Use: Data for A Strategic Priority

Data Fixer: Data Engineer

Sales Data Quality Dashboard

Customer: CFO

Use: Data For Financial Compliance

Data Fixer: Accounting Data Person

CFO Data Quality Dashboard

Customer: ML Team

Use: Good Data for AI

Data Fixer: Data Scientist

ML Team Data Quality Dashboard

Customer: CDO

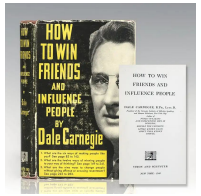
Use: Quarterly Review of DQ Team Progress

Data Fixer: IT Staff

CDO Quality Dashboard

DataKitchen

Dale Carnegie Data Quality Principle #18



Be Sympathetic With The Other Person's Ideas And Desires

Give Them Specific Actionable Problems – They Are Busy!

Customer: VP of Sales

Use: Data for A Strategic Priority

Data Fixer: Data Engineer

Sales Data Quality Dashboard

Actionable Test Results

Customer: CFO

Use: Data For Financial Compliance

Data Fixer: Accounting Data Person

CFO Data Quality Dashboard

Actionable Test Results

Customer: ML Team

Use: Good Data for AI

Data Fixer: Data Scientist

ML Team Data Quality Dashboard

Actionable Test Results

Customer: CDO

Use: Quarterly Review of DQ Team Progress

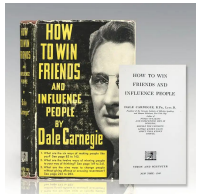
Data Fixer: IT Staff

CDO Quality Dashboard

Actionable Test Results

DataKitchen

Dale Carnegie Data Quality Principle #29



Use Encouragement. Make The Fault Seem Easy To Correct

Package Data Quality Issues So They Are Easy to Fix

DataOps Data Quality TestGen — DataKitchen

TestGen Hygiene Issue Report

Hygiene Issue	**Quoted Values Found in Column Values:** Column values were found within quotes. This likely contradicts user expectations and could be a sign of broader ingestion or processing errors.		Likely
Detail	**Cases Found: 11**		
Database/Schema	demo	**Profiling Date**	Feb 17, 12:57 PM EST
Table	d_ebike_products	**Table Group**	default
Column	product_name	**Disposition**	No Decision
Column Type	varchar(255)		View on TestGen >

Suggested Action

Review your source data, ingestion process, and any processing steps that update this column.

Sample Data

Product Name	Count
'Aero Ace'	1
'Epic Endeavor'	1
'Meteor Motion'	1
'Nebula Nomad'	1
'Nimbus Navigator'	1
'Odyssey Origin'	1

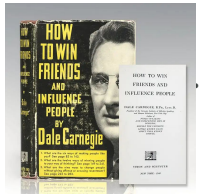
Product Name	Count
'Primal Pursuit'	1
'Skyward Sprinter'	1
'Tempest Twirl'	1
'Titan Trailblazer'	1
'Wanderer'	1

SQL Query

```
SELECT DISTINCT "product_name", COUNT(*) AS count FROM demo.d_ebike_products WHERE (CASE WHEN "product_name"
ILIKE '"%"' OR "product_name" ILIKE '''%''' THEN 1 ELSE 0 END) = 1 GROUP BY "product_name" ORDER BY
"product_name";
```

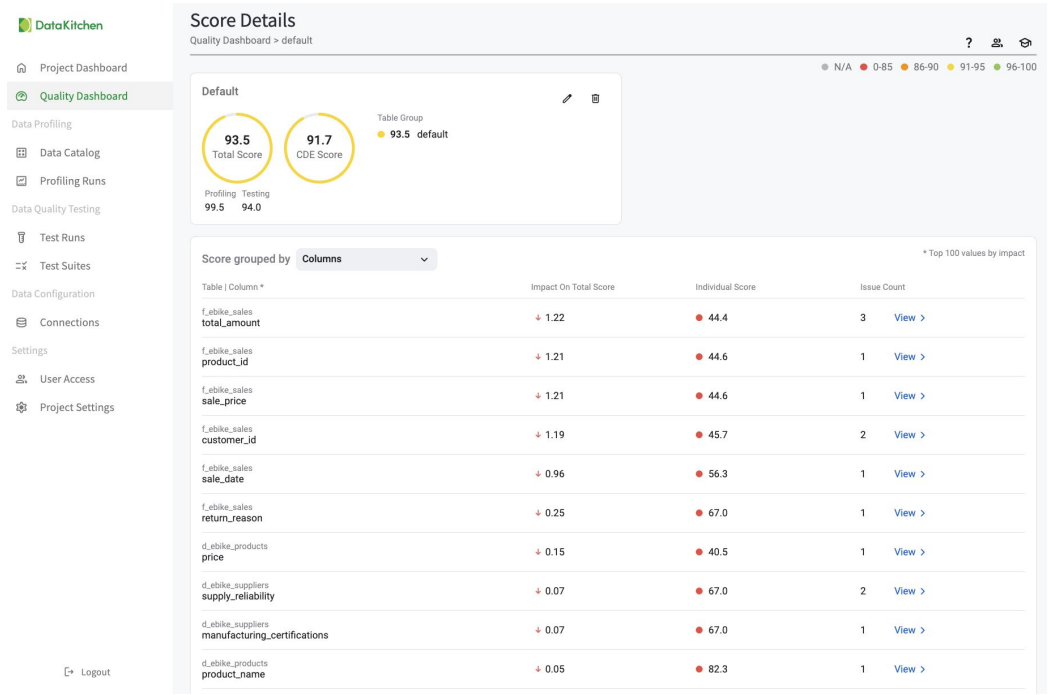
DataKitchen

Dale Carnegie Data Quality Principle #30



. Make The Other Person Happy About Doing The Thing You Suggest

Measure And Show Data Quality Improvement Over Time



As you get new data, your DQ scores will change. Revist the data quality tests you use to establish those scores:

- constantly refining your DQ standards
- get fewer false positives
- continue to learn about the data
- learn from real-world data.

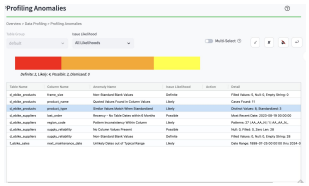
Watch the chart of DQ scores over time—improving the data and the evaluation criteria/standards affects both!

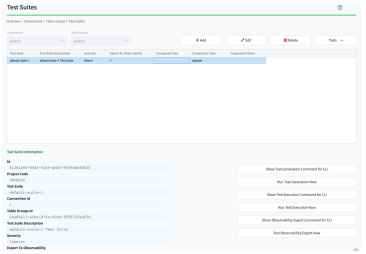
DataKitchen

Agenda

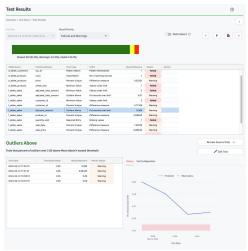
- Start DataOps With Data Quality
- Challenges of Data Quality and Data Quality Leadership
- Announcing: Actionable, Automated, Agile Data Quality Scorecards
- **Demonstration: Open Source DataOps Data Quality TestGen**
- How To DataOps Your Data Quality
- Conclusion

Open Source DataOps TestGen Workflow





Data Owners,
Data Engineers



Profile Tables

Screen for Data Hygiene “Gotcha”

Generate Data Quality (DQ) Tests

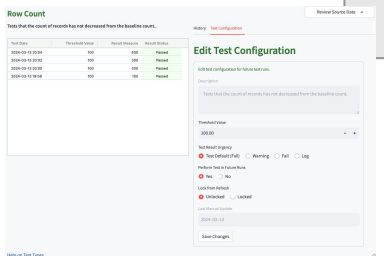
Share DQ Issues Details

Source Data Updates

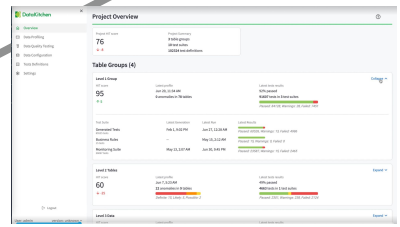
Review and Refine DQ Testing

Data Quality Person

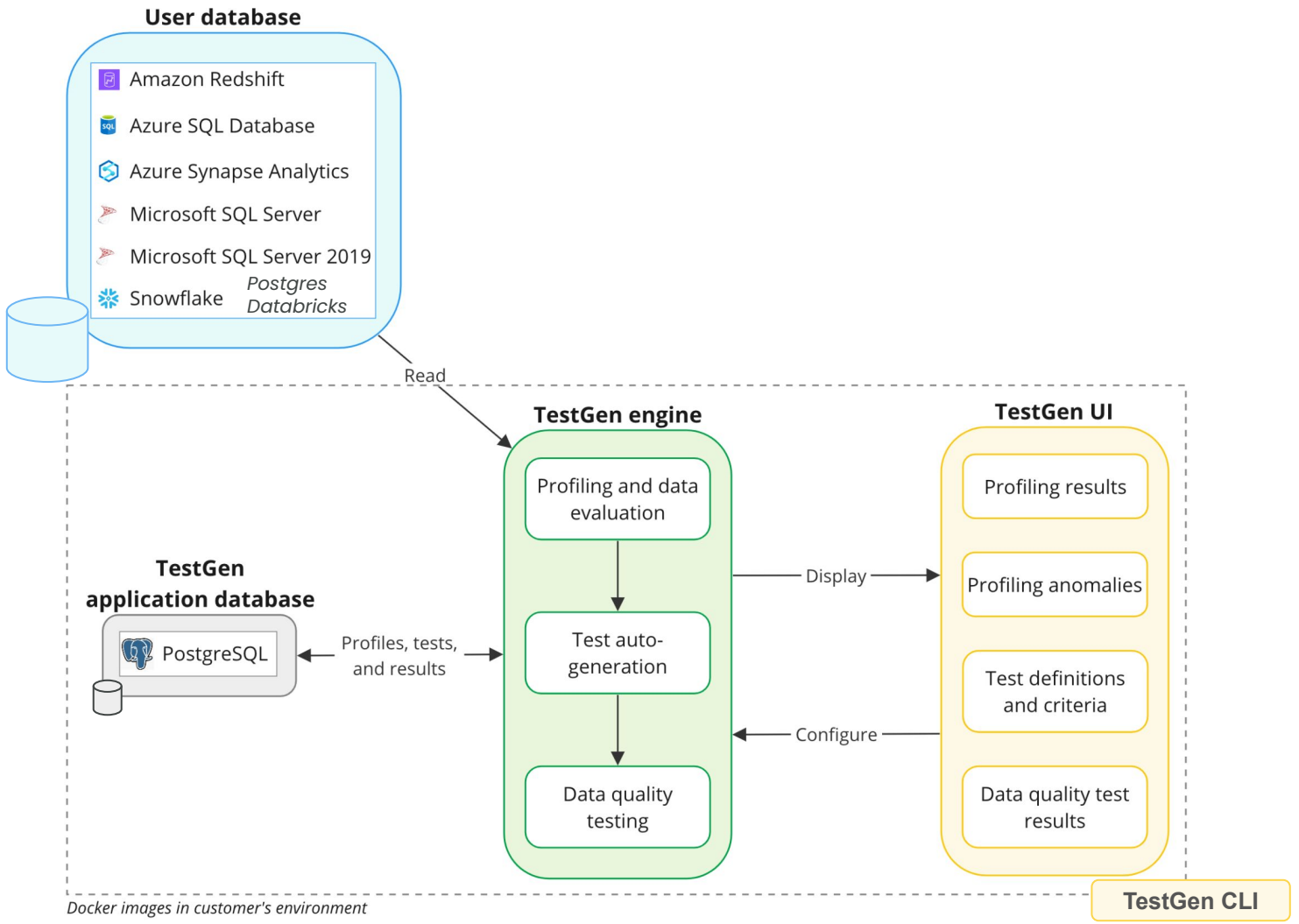
Execute DQ Tests on Schedule



Generate Data Quality Scores



DataOps DQ TestGen Technical Architecture

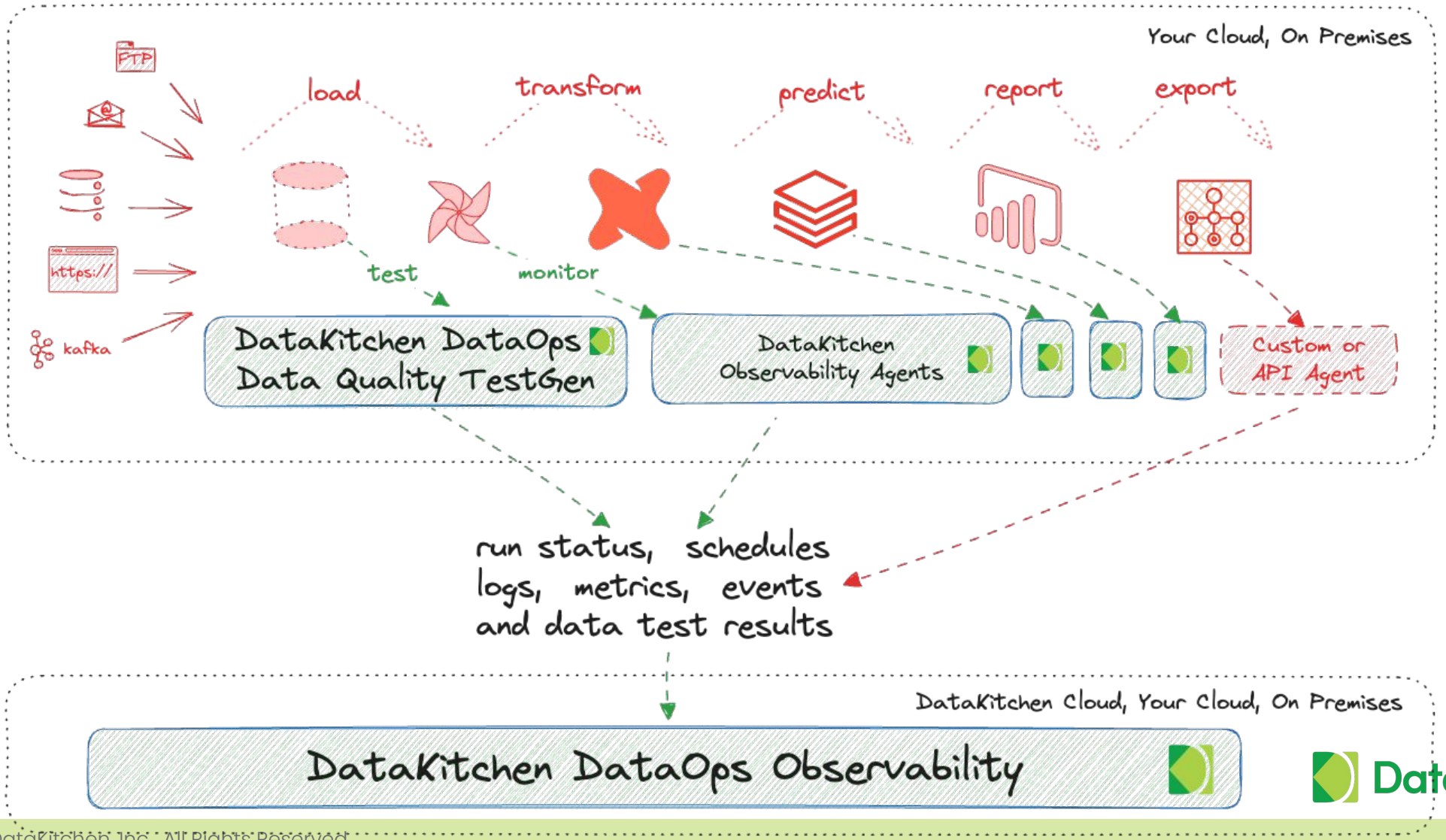


DataKitchen

DataOps Data Quality TestGen Architecture

DataKitchen

DataOps Observability Architecture



Agenda

- Start DataOps With Data Quality
- Challenges of Data Quality and Data Quality Leadership
- Announcing: Actionable, Automated, Agile Data Quality Scorecards
- Demonstration: Open Source DataOps Data Quality TestGen
- **How To DataOps Your Data Quality**
- Conclusion

DataOps Perspective on Data Quality

If you boil it all down, Data Quality Leaders have two major challenges:

How do I deal with all this data at scale?
How do I influence change to improve Data Quality?

Best method to solve those challenges

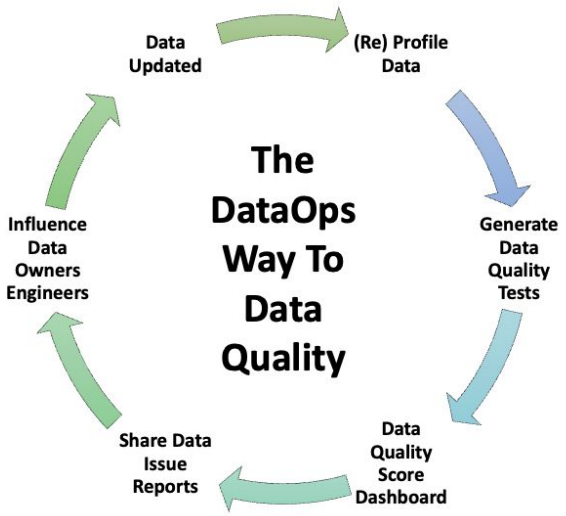
Get Going Quickly
Focus on Specific Customer Needs
Iterate, Improve, Influence

Agile/DataOps Data Quality At Scale

Advantages of this approach:

- Don’t boil the ocean, focus on specific data items related to your customers goals
- More iteration means better standards, more action to achieve them
- Create quality checks – automatically – don't spend months coding
- Make trade-offs based on knowledge, not hope
- Tap expertise when it counts
  - Leverage time of Subject Matter Experts and skilled engineers
- It is faster, easier to implement on real data, learn as you iterate
- Maximize your influence with customer focus!

DataOps Process: Iterative Data Quality At Scale



- **(Re) Profile Data:**
  - Establish a baseline by measuring the current state of the data with data profiling
  - Get foundation for tracking progress and identifying areas for improvement, repeatedly
- **Generate Data Quality Tests:**
  - Generative Data Quality tests can streamline this process, making it easier to catch problems before they become more significant without coding
- **Data Quality Score Dashboard:**
  - Scoring provide a quantifiable way to assess the health of the data and communicate the results.
  - Scores should be focused on the data elements that matter and reflect top organization needs
- **Share Data Quality Issues Report**
  - Give actionable, specific changes to those who can do the technical work.
- **Influence Data Owners and Engineer**:
  - Once data quality issues have been identified, the next step is to drive the necessary changes within the organization.
  - Influence data owners, system administrators, or other stakeholders to address the identified issues.

Open Source Goal: Data Quality Super Powers

Give a single person, with limited time, a tool to effect meaningful Data Quality change.

- No cost, run on their laptop.
- Learn and improve over time. No large ramp-up
- Automation and AI to help with testing
- Make it easy to get impact.
- Fit to your needs
- Measure your success

= INFLUENCE SUPERPOWERS

Read More ‘Data Quality The DataOps Way’

DataKitchen

Data Quality The DataOps Way

Why a DataOps Approach to Data Quality: Data Quality Challenges

Data quality is more critical than ever in the age of AI and organizations ‘competing with data.’ As teams handle increasingly vast amounts of data, the demand for accuracy, consistency, and reliability in data sources has risen in parallel. Despite advancements in data management technologies, a significant gap persists in ensuring data quality. According to a 2024 survey by dbt Labs, 57% of respondents rated data quality as one of the top three most challenging aspects of the data preparation process—a sharp increase from 41% in prior years. This statistic highlights an important reality: data quality is not merely an operational problem but a strategic imperative.

Moreover, IDC reports that 73% of data practitioners do not fully trust their data, signaling deep-seated issues with the reliability of organizational data. This mistrust directly impacts decision-making processes across departments. Forrester's research adds another layer of urgency, estimating that millions were lost in 2023 due to poor data quality, with the potential for these losses to escalate into billions as artificial intelligence (AI) becomes more integral to business operations. Without a concerted intervention, the symbiotic relationship between AI and data quality could become a global liability for organizations.

The severity and scale of these problems indicate that the traditional methods of addressing data quality must be revised. Many organizations have approached data quality issues through failed initiatives—avoiding fixing problems or on long, slow, get-little-done projects. However, data quality challenges are not static; they are dynamic and continuous, simultaneously impacting multiple aspects of an organization. Addressing these data quality challenges effectively requires more than avoidance or slow bureaucracy. It calls for an ongoing, adaptive methodology that can evolve with the data itself—a methodology found in DataOps.

Why a DataOps Approach to Data Quality: Leadership Challenges

While identifying and understanding data quality issues is crucial, data quality leaders face a different, more nuanced challenge: driving change at scale across their organizations. In many instances, the people responsible for the data quality may differ from those responsible for fixing it. Data quality problems often span multiple departments, crossing into different business units and functional areas. This creates an environment where effecting widespread improvement becomes incredibly complex, as it requires cooperation from a diverse range of stakeholders, many of whom might not see the immediate benefit of investing in data quality improvements.

One of the core issues faced by data quality leaders is that source system data is often deemed "good enough" for the operational systems where it originates. However, this data frequently needs to be revised

The DataOps approach to data quality offers a transformative path.

It empowers individuals to act swiftly, enables continuous improvement, and fosters collaboration across organizational silos.

With AI-driven insights and rapid iteration, DataOps tackles data quality issues at scale

Read More:
https://info.datakitchen.io/white-paper-data-quality-the-dataops-way

Agenda

- Start DataOps With Data Quality
- Challenges of Data Quality and Data Quality Leadership
- Announcing: Actionable, Automated, Agile Data Quality Scorecards
- Demonstration: Open Source DataOps Data Quality TestGen
- How To DataOps Your Data Quality
- **Conclusion**

Learn More About DataOps & Data Observability

Install Open Source TestGen
https://info.datakitchen.io/install-dataops-data-quality-testgen-today

Install Open Source DataOps Observability
https://docs.datakitchen.io/articles/#!open-source-data-observability/install-data-observability-products-open-source

Sign The DataOps Manifesto
http://dataopsmanifesto.org

Free DataOps Cookbook
https://datakitchen.io/the-dataops-cookbook/

Free DataOps Certification
https://info.datakitchen.io/training-certification-dataops-fundamentals

Free Data Quality & Observability Certification
https://info.datakitchen.io/data-observability-and-data-quality-testing-certification

Summary

1. **Improve Data Quality**

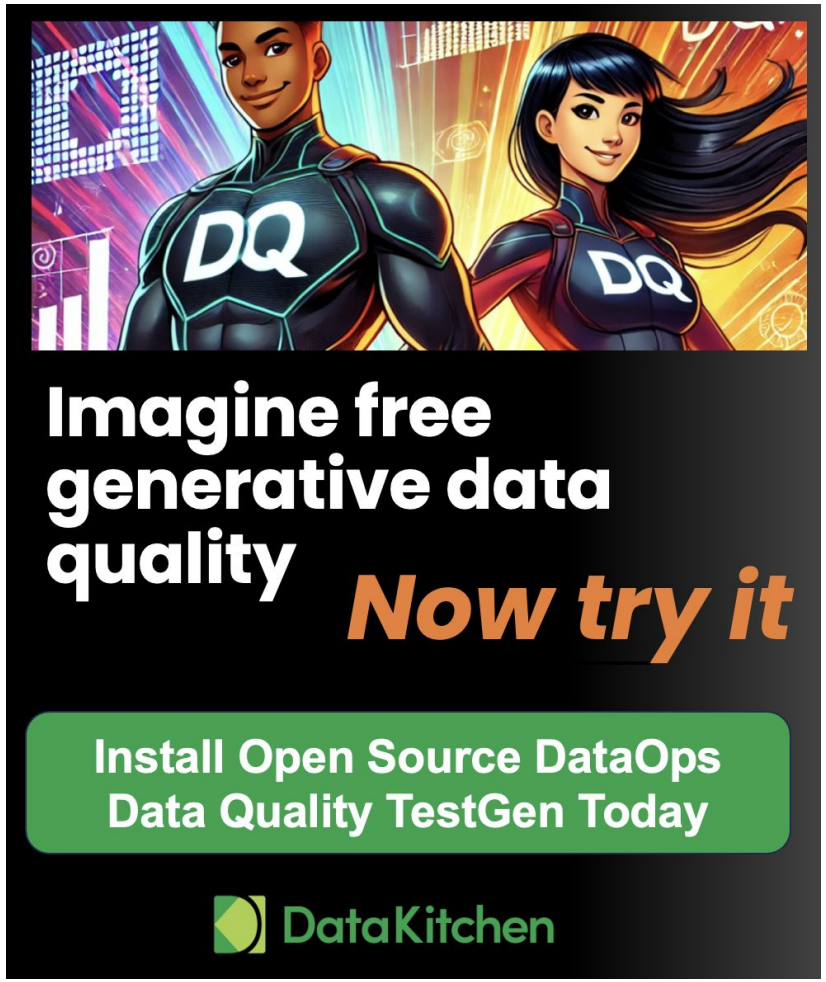
DataOps Data Quality TestGen

2. **Stop Production Errors**

DataOps Observability

3. **Enjoy your extra time**

Imagine …



Imagine a tool you can point at any dataset, that will learn from your data, screen for common data quality issues, then automatically generate and perform powerful tests, analyzing and scoring your data to pinpoint issues before they snowball.

Imagine an open-source tool that's free to download, but also minimal cost in time and effort to actually use. It assesses your data, deploys production testing, monitors progress, and helps you build a constituency within your company for lasting change.

Start Using it Today – https://info.datakitchen.io/install-dataops-data-quality-testgen-today