How Data Governance Enables Scalable Data Science

May 6, 2021
Our Mission is to democratize data skills for everyone
Trusted by over 1,600 data-driven companies

DataCamp is transforming the way businesses prepare their employees for the future of work.
Aaren Stubberfield

- Previously Manager of Data Governance and AI at a multinational food supplier
- 10+ years in data centric roles
- **Author of two Datacamp courses**

Data Scientist
Microsoft

linkedin.com/in/aaren-stubberfield
Motivation
Data-driven organizations have been a staple of the technology industry.
Data-driven organizations have been a staple of the technology industry
Data-driven organizations have been a staple of the technology industry. Amazon (1994), Netflix (1997), Google (1998), Facebook (2004), Airbnb (2008), Uber (2009), and Stripe (2010) are examples of companies that have transformed the industry by leveraging data-driven strategies.

Now every organization needs to become data-driven, including Financial Institutions, Healthcare Institutions, and Government Institutions.
The challenge is democratizing data
The challenge is democratizing data

- Empowered People
- Decisions are based on data
- Data is accessible
The challenge is democratizing data

Empowered People

Decisions are based on data

Data is accessible

Everyone knows how to access and work with the data they need to do their job
The importance of data democratization — by the numbers

- Organizations with data democratization are **2x** more likely to exceed business goals.
- Of organizations with high democratization of data “significantly exceeded” yearly business goals.
- High-performing organizations are **65%** more likely to provide data access to frontline employees whenever needed.


The north star — the Airbnb example

Source — Democratizing Data at Airbnb
The components of data democratization
The components of data democratization

**Tools**
The type of data tools data practitioners and experts use across all skill levels, and the availability of frameworks that democratize working with data.

**Organization**
Organizational models that enable data talent to scale data science throughout the organization.

**Processes**
Data processes to scale collaboration and efficiency, and ensure alignment with business objectives.

**People**
Forging a data culture where everyone understands the value of data, and ensuring that everyone has the skills to work with data to do their best work.

**Infrastructure**
Ensuring data is collected, discoverable, reliable, understood, compliant, and actionable throughout the organization.

Webinar: Path to Data Fluency
The components of data democratization

**Tools**
The type of data tools data practitioners and experts use across all skill levels, and the availability of frameworks that democratize working with data.

**Organization**
Organizational models that enable data talent to scale data science throughout the organization.

**Processes**
Data processes to scale collaboration and efficiency, and ensure alignment with business objectives.

**People**
Forging a data culture where everyone understands the value of data, and ensuring that everyone has the skills to work with data to do their best work.

**Infrastructure**
Ensuring data is collected, discoverable, reliable, understood, compliant, and actionable throughout the organization.

*Webinar: Path to Data Fluency*
Data governance & data quality are crucial for democratizing data science because they enable organization-wide trust in data.
Data governance
What Is Data Governance?

“Data Governance is defined as the exercise of authority and control (planning, monitoring, and enforcement) over the management of data assets.”

The Global Data Management Community
What Is Data Governance?

Planning

Monitoring

Enforcement
What Is Data Governance?

Planning

Creating rules for data sets

- What is the format of the “date” fields?
- How many characters are used in the “name” field?
- What encoding will be used for the string fields?

Monitoring

Enforcement
What Is Data Governance?

Planning

Creating rules for data sets

○ What is the format of the “date” fields?
○ How many characters are used in the “name” field?
○ What encoding will be used for the string fields?

Monitoring

Measuring compliance to rules

○ What is the percentage of missing values in a field over time?
What Is Data Governance?

Planning

Creating rules for data sets
- What is the format of the “date” fields?
- How many characters are used in the “name” field?
- What encoding will be used for the string fields?

Monitoring

Measuring compliance to rules
- What is the percentage of missing values in a field over time?

Enforcement

Remediation by urgency
- Records that need to be corrected right away
- Record that can be corrected later
- Suggest changes to records
What Is Data Governance?

Planning

Creating rules for data sets

- What is the format of the “date” fields?
- How many characters are used in the “name” field?
- What encoding will be used for the string fields?

```sql
CREATE TABLE `Customers` (  `CID` INT NOT NULL AUTO_INCREMENT,  `CompanyName` VARCHAR(255) NOT NULL,  `DateFounded` Date,  `CompanyIndustry` VARCHAR(255), PRIMARY KEY (`CID`)) ;
```

<table>
<thead>
<tr>
<th>CID</th>
<th>CompanyName</th>
<th>DateFounded</th>
<th>CompanyIndustry</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Broadview Holdings</td>
<td>1943-07-01</td>
<td>Technology</td>
</tr>
<tr>
<td>2</td>
<td>Epic Inc.</td>
<td>1921-01-03</td>
<td>Marketing</td>
</tr>
<tr>
<td>3</td>
<td>Damen Studios</td>
<td>1921-01-03</td>
<td>Unknown</td>
</tr>
</tbody>
</table>
What Is Data Governance?

Monitoring

Measuring compliance to rules

- What is the percentage of missing values in a field over time?

<table>
<thead>
<tr>
<th>CID</th>
<th>CompanyName</th>
<th>DateFounded</th>
<th>CompanyIndustry</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Broadview Holdings</td>
<td>1943-07-01</td>
<td>Technology</td>
</tr>
<tr>
<td>2</td>
<td>Epic Inc.</td>
<td></td>
<td>Marketing</td>
</tr>
<tr>
<td>3</td>
<td>Damen Studios</td>
<td>1921-01-03</td>
<td>Unknown</td>
</tr>
</tbody>
</table>
What Is Data Governance?

Enforcement

Remediation by urgency

- Records that need to be corrected right away
- Record that can be corrected later
- Suggest changes to records

<table>
<thead>
<tr>
<th>CID</th>
<th>CompanyName</th>
<th>DateFounded</th>
<th>CompanyIndustry</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Broadview Holdings</td>
<td>1943-07-01</td>
<td>Technology</td>
</tr>
<tr>
<td>2</td>
<td>Epic Inc.</td>
<td>1921-01-03</td>
<td>Marketing</td>
</tr>
<tr>
<td>3</td>
<td>Damen Studios</td>
<td></td>
<td>Unknown</td>
</tr>
</tbody>
</table>
What Is Data Governance?

Data Policies & Procedures

Data Stewards & Stakeholders

Datasets

Technology
What is Data Governance — Data Stewards

- Accountable over data quality
- Subject Matter Expert on the data
- High data literacy with the ability to communicate effectively
What is Data Governance — Data Stewards

- Accountable over data quality
- Subject Matter Expert on the data
- High data literacy with the ability to communicate effectively

Can influence data quality throughout the organization
What is Data Governance — Technology

Data Quality Management
- IBM
- Talend
- Informatica

Data Discovery Tools
- Amundsen
- Databook
- Dataportal

Data Governance Platforms
- Collibra
- /true.dat
- PRIVAGERA
The Benefits of Data Governance — Reducing Risk

Data security and limiting access

Adhering to Privacy
The Benefits of Data Governance — Scalable Data Science

**Data Quality & Accessibility**
- Data is no longer siloed
- Common data glossary and verbiage
- Increased confidence in data quality

**Impact**
- Consistent data-driven decision making
- Improved business performance

**Improved Efficiency at Scale**
- Clear rules for data processes
- Centralized data management mechanisms
- Increased data modularity

**Impact**
- IT and Business Units are more agile
- Reduced costs
Operationalizing data governance
What Is Data Governance?

- Data Policies & Procedures
- Datasets
- Technology
- Data Stewards & Stakeholders
Challenges In Data Governance
Challenges In Data Governance — Not an organizational priority

42%
Of data and analytics leaders do not assess, measure, or monitor data governance

60%
Of organizations do not plan to implement processes or tools to publish data provenance or lineage

Source — Gartner

Source — O'Reilly: The State of data quality in 2020
Challenges In Data Governance — Data Acquisition Challenges
Challenges In Data Governance — Inadequate Data Steward Representation

- Low organizational data literacy
- Lack of alignment and diversity in representation
- Lack of executive sponsorship
Challenges In Data Governance — Data Governance Ownership

Information Technology

Business Units
Best Practices for Data Governance
Best Practices — Start with the Right People

People set the agenda for scalable data governance
Best Practices — Start with the Right People

People set the agenda for scalable data governance

“To begin building the big picture, start with the people, then build your processes, and finally incorporate your technology. Without the right people, it’s difficult to build the successful processes needed for the technical implementation of data governance.”

Tableau
Best practices for implementing data governance in your organization
Best Practices — Data Governance is Marathon not a Sprint

- Be iterative — start in one domain and expand to another
- Account for new data over time
Best Practices — Track your Metrics
Best Practices — Track your Metrics

Value

Effectiveness

Sustainability
Best Practices — Track your Metrics

Value
- Time to insights by data practitioners
- Increased speed in customer onboarding

Effectiveness
- # of business users trained on a new policy
- % completion of data glossary
- Time to market for new data products

Sustainability
- Use of data management tools
- Compliance percentage of data sets over time
Best Practices — Communicate Early and Often

I HAVE AN ANNOUNCEMENT
Best Practices — Communicate Early and Often

Kübler-Ross Change Model Curve Template

Source — Matt Sevenoaks Coaching
Best Practices — Invest in Data Governance Tools

- Data Quality Management:
  - IBM
  - Talend
  - Informatica

- Data Discovery Tools:
  - Amundsen
  - Databook
  - Dataportal

- Data Governance Platforms:
  - Collibra
  - /truedat
  - PRIVAGERA
Getting started
Getting Started — Build the Business Case

- Identify pain points of low data quality
- Determine ROI for Data Governance Endeavors
- Garner Executive Support for Data Governance
Getting Started — Perform a Readiness Assessment

Do people have the necessary buy-in or data literacy?

What is the state of data quality today? Are there any rules set in place?

What is the tools and technology landscape?

Source — Informatica
Getting Started — Avoid Starting From Scratch

Source — DAMA-DMBOK 2nd Edition

Source — EDMCouncil
Wrap up and Q&A
Data governance & data quality are crucial for democratizing data science because they enable organization-wide trust in data.
High quality data enables data driven decision making at scale

Data Governance is a framework and process for scaling data quality in the organization

Data Governance is inextricably linked to a data-driven culture
Data Governance is a Team Effort
Q&A
Upcoming webinars in May

Panel: Learning and Development for the Data Driven Age

In this panel, Kerrie O'Sullivan & Sheil Naik from the Bloomberg Global Data team, and Marcus Robertson from NatWest Group, will discuss the importance of L&D for data-driven organizations, best practices designing digital-first learning experiences, and how they go about sustaining a learning culture in their organizations.

📅 Wednesday, May 12, 11:00 am ET

Building Data Cultures

In this webinar, Sudaman Thoppam Mohanchandralal, Chief Data Officer at Allianz Benelux, will deep dive into the ins and outs of building sustainable data cultures.

📅 Friday, May 28, 9:00 am ET

Register here
Thank you