# Book Summary: Driving Data Quality with Data Contracts

Hello, my name is Cynthia, and welcome to this review of “Driving Data Quality with Data Contracts” by Andrew Jones!

For the last two weeks, I’ve been deep into learning about Data Contracts and I’m excited to share a summary of this book, hoping to make these concepts clear for starters, spur the interest of the curious, and give a snapshot for those who haven’t got around to reading the whole book.

In this book I’ve discovered how data contracts can solve common problems many companies have with their data. Even with a lot of money spent, getting good quality data that actually helps businesses is a big challenge for everyone. Interestingly, while each company’s data may be different, the challenges are universally similar!

The solution data contracts provide consists of two aspects. First, they set up a contract-backed architecture which makes it easier to create and use good quality data through self-served, autonomous tooling. Second, they facilitate a shift in data culture, emphasizing data explicitly generated to meet use cases, fostering collaboration between data generators and consumers, and prioritizing data quality over quantity.

Both elements are integral to cultivating a genuinely data-driven organization, leveraging quality data to derive real business value. This book is a valuable resource for experienced data engineers, leaders, architects, or practitioners contemplating their data architecture and seeking innovative approaches to maximizing data value.

This book is packed with ten chapters, each addressing different facets of data contracts and their implementation. It begins with a historical overview of data platforms, revealing how architectural stagnation has hindered data quality and value delivery. And then it dives into the world of data contracts, exploring what they are and how they relate to data mesh.

Middle chapters are all about getting data generators and consumers to work closely, diving into data governance, and breaking down what a data contract actually includes. The book then transitions into practical aspects, showcasing a contract-driven data architecture, providing a sample implementation, and discussing how to integrate data contracts into a organization.

Finally, the book provides a practical look at working with data contracts daily, including designing, monitoring, enforcement, and publishing patterns for data generators. Every chapter is a step closer to understanding and using data contracts to make our data better and more valuable.

Again, this review is here to give you a quick peek and hopefully get you interested in learning more about data contracts.

Data is vast and varied, making it a challenge to get everyone on the same page. A one-size-fits-all explanation just doesn’t cut it for people from different backgrounds. So, I’m aiming to share stories and insights that connect with different folks because I believe that’s the way to make a real impact and understanding.

If you’re keen for more details, I totally recommend grabbing the book and, of course, you can always get in touch with me or John Thomas, a Staff Engineer of D&A Platforms and an expert reviewer of this book.

### Definition

Let us first familirized ourselves with some important terminologies.

A **Data Contract** is a formal agreement between data providers and data consumers about the format, quality, and other properties of the data being exchanged. But more on that later. Here are some key concepts and terms related to Data Contract:

1. **Schema**: Defines the structure of the data, including the fields, data types, and relationships. It acts as a blueprint for the data to be exchanged and helps in validating the data against the defined structure.
2. **Metadata**: Data about the data. It includes information like data lineage, timestamps, data quality metrics, and other details that describe the characteristics and context of the data.
3. **Data Quality**: A measure of the condition or caliber of data, which considers aspects such as accuracy, completeness, reliability, relevance, and timeliness. Ensuring high data quality is pivotal for making informed decisions.
4. **Data Lineage**: Visualization of the flow and transformation of data as it moves through the various stages of a system or process. It helps in understanding the origins, movements, and calculations applied to the data.
5. **Data Provider**: The entity or system that produces or supplies data. Data providers are responsible for ensuring that the data meets the agreed-upon standards and specifications outlined in the data contract.
6. **Data Generator/Producer/Provider**: These terms are often used interchangeably to describe the entity, system, or process that creates, supplies, or makes data available for use. They are responsible for maintaining the quality, accuracy, and security of the data according to the agreed-upon standards in the data contract.
7. **Data Consumer**: This entity, application, or individual utilizes the data provided by the data generator/producer/provider. Data consumers use the data for various purposes such as analysis, reporting, or to feed into other systems or processes. They rely on the data contract to understand the format, quality, and characteristics of the data they are consuming.
8. **Data Consumer**: The entity or system that receives or uses data. Data consumers rely on the data provided to perform analyses, generate insights, or make informed decisions.
9. **Service Level Agreement (SLA)**: A commitment between the data provider and the data consumer on the level of service, including data availability, timeliness, and quality.
10. **Data Governance**: The practice of managing and organizing data to ensure data quality, security, and compliance with policies and regulations. It involves defining roles, responsibilities, and processes related to data management.
11. **Versioning**: The practice of keeping multiple versions of data to track changes and updates over time. It helps in managing and controlling modifications to the data.
12. **Data Validation**: The process of checking and ensuring that the data meets the predefined standards and specifications before it is shared or used.

### Five-Minute Summary:

#### For a General Audience (Simple Terminology):

**Imagine a Data Contract as a Promise:**

You and your friend decide to exchange letters. You agree on what information to share, how to write it, and when to send it. This agreement is like a data contract.

1. **What’s in the Letter?**
   * **Data Contract**: It details what information is shared. It’s like deciding what stories, news, and updates you’ll put in your letter.
2. **How is it Written?**
   * **Format and Structure**: Just like you’d write a letter with a greeting, body, and closing, data has a specific way it’s written and structured so that the other person can understand it.
3. **Quality Check:**
   * **Accuracy and Timeliness**: You want to make sure your letter has the correct information and reaches on time. A data contract ensures the data is accurate and shared promptly.
4. **Keeping Promises:**
   * **Trust**: When both sides keep to the agreement, it builds trust. This is the same with data; the contract ensures everyone is on the same page, building trust between the data sharer and the user.

#### For a Technical Audience:

**Data Contract as a Formal Agreement:**

A Data Contract is a formal agreement detailing the specification of data exchanged between a Data Provider and a Data Consumer.

1. **Schema and Metadata:**
   * **Specification**: The contract defines the schema and metadata of the data, specifying the format, type, and structure, ensuring that data is consistent and usable.
2. **Data Quality and Integrity:**
   * **Requirements**: It outlines the quality requirements, such as accuracy, completeness, and timeliness, maintaining the reliability and integrity of the data exchanged.
3. **Compliance and Governance:**
   * **Standards and Regulations**: The contract addresses compliance with data standards, regulations, and governance policies, mitigating risks associated with data misuse and breaches.
4. **Dispute Resolution:**
   * **Clarity and Accountability**: It provides a clear framework for resolving disputes and ensuring accountability, fostering trust and collaboration between data producers and consumers.

In conclusion, a Data Contract is essential for ensuring the smooth, reliable, and secure exchange of data between different parties, whether explained simply or technically.

## Deep Dives

## 1. **Technical Teams:**

*Includes Data Engineers, Data Scientists, Application Developers, and Quality Assurance Teams.*

### What is Data Contract for you? You can consider Data Contract to be an API. It’s the blueprint that defines how data is structured, validated, and interacted with, ensuring consistent data quality and format across various applications and services.

### Why do you need Data Contract? With Data Contract, you can directly create value. You now have the flexibility and autonomy to change how your code is implemented in the background with ease of mind, knowing you will not break the downstream. It gives you a set of tools to help automate your workflows and maintain data integrity. You can negotiate with your downstream users, understanding their needs without having the data engineering teams as a bottleneck.

### How will your workflow look like? Incorporating Data Contracts means more streamlined and automated workflows. The clear specifications and validation checks result in fewer errors and less time spent on troubleshooting and data cleaning. It facilitates better collaboration between different technical teams and ensures smoother integration of services.

**Key Takeaways for Technical Teams:** - Enhanced Flexibility and Autonomy - Streamlined and Automated Workflows - Improved Collaboration and Integration - Reduction in Errors and Troubleshooting

## 2. **Data Analysis and Business Intelligence Teams:**

*Includes Data Analysts and Business Intelligence Professionals.*

### What is Data Contract for you? For Data Analysts and BI Professionals, Data Contract is a guarantee of the quality and consistency of the data they analyze. It provides a clear understanding of the data structure, enabling accurate and insightful analysis.

### Why do you need Data Contract? Data Contracts ensure that the data you work with is reliable and of high quality. They allow for more accurate analyses and predictions, which in turn lead to better business decisions and strategies. The clarity and reliability provided by Data Contracts enable you to focus more on deriving insights and less on data cleaning and preparation.

### How will your workflow look like? With the adoption of Data Contracts, the workflow of analysts and BI professionals becomes more streamlined. There’s less time spent on cleaning and preparing data, and more focus on conducting in-depth analyses, generating insights, and contributing to the development of data-driven strategies.

**Key Takeaways for Data Analysis and Business Intelligence Teams:** - Guaranteed Data Quality and Consistency - More Accurate Analyses and Predictions - Streamlined Workflows and Focused Analysis - Enhanced Contribution to Data-Driven Strategies

## 3. **Management and Strategy Teams:**

*Includes IT Management and Leadership, Data Stewards, Data Governance Teams, and Product Managers.*

### What is Data Contract for you? For management and strategy teams, Data Contract is a mechanism for enforcing data quality and governance. It serves as a cornerstone for data strategy, ensuring consistency, reliability, and compliance of data across the organization.

### Why do you need Data Contract? Implementing Data Contracts aligns with strategic objectives by ensuring data reliability and compliance. It reduces risks associated with data breaches and misuse, and it fosters trust and collaboration among different organizational units. It allows for better management of data resources and aids in the realization of data-driven goals.

### How will your workflow look like? With Data Contracts, management and strategy workflows will be more organized and focused on ensuring compliance and strategic alignment. There will be a clear process for establishing, reviewing, and enforcing data contracts, contributing to the overall data governance and strategy of the organization.

**Key Takeaways for Management and Strategy Teams:** - Enforced Data Quality and Governance - Risk Mitigation and Compliance - Strategic Alignment and Collaboration - Organized and Focused Workflows

## 4. **Compliance and Operational Teams:**

*Includes Legal and Compliance Teams, and Supply Chain and Logistics Teams.*

### What is Data Contract for you? For Compliance and Operational Teams, a Data Contract is a binding agreement that specifies the standards and regulations data must adhere to. It is an essential tool for maintaining legal compliance and operational integrity.

### Why do you need Data Contract? Data Contracts are crucial for ensuring that the data used and shared complies with legal and regulatory standards, thereby reducing the risk of legal repercussions. They help in maintaining the quality of operational data, ensuring smooth and efficient business processes.

### How will your workflow look like? Workflows for compliance and operational teams will involve regular audits and checks to ensure adherence to Data Contracts. It will include collaboration with various departments to ensure that the standards and regulations outlined in the contracts are consistently met.

**Key Takeaways for Compliance and Operational Teams:** - Legal Compliance and Risk Reduction - Maintenance of Operational Data Quality - Regular Audits and Checks - Cross-Departmental Collaboration