Compliments of

Starburst

Data Products

for dummies®

A Wiley Brand

Maximize data's value, shareability, and usability

Embrace a self-service data product approach

See the future of data products with large language models

Vishal Singh,
Ryo Komatsuzaki, and
Andrew Mott, MBA

Starburst Special Edition

About Starburst

For data-driven companies, Starburst offers a full-featured data lake analytics platform, built on open source Trino. Our platform includes the capabilities needed to discover, organize, and consume data without the need for time-consuming and costly migrations. We believe the lake should be the center of gravity, but support accessing data outside the lake when needed. With Starburst, teams can access more complete data, lower the cost of infrastructure, use the tools best suited to their specific needs, and avoid vendor lock-in. Trusted by companies like Comcast, Grubhub, and Priceline, Starburst helps companies make better decisions faster on all their data.

Start building data products: starburst.io/platform/features/data-products/

Sign up for a free trial: starburst.io/platform/starburst-galaxy/start/

starburst.io

Data Products

Starburst Special Edition

by Vishal Singh, Ryo Komatsuzaki, and Andrew Mott, MBA

for dummies®
A Wiley Brand

Data Products For Dummies®, Starburst Special Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

ISBN 978-1-394-21837-0 (pbk); ISBN 978-1-394-21838-7 (ebk)

Publisher's Acknowledgments

Development Editor: Rachael Chilvers

Project Editor: Saikarthick Kumarasamy

Acquisitions Editor: Traci Martin

Editorial Manager: Rev Mengle

Business Development Representative: Matt Cox

IN THIS CHAPTER

- Applying product management principles to data
- Elevating the business through the data product approach

Chapter 1
Defining Data Products

Maximizing value from data is a persistent business challenge. Among the latest contributions to the field of data is the concept of *data mesh* — a decentralized, distributed approach to enterprise data management — which brings the concept of the data product into the mainstream. In this chapter, we lift the lid on the idea of a data product and what it can do to modernize your data analytics strategy.

So, What is a Data Product?

A quick search of the internet results in two related, but different, terms in this area:

- *Data as a product* is the concept of applying product management principles to data to enhance its use and value.
- *Data product* is a combination of a curated, reusable dataset, engineered to deliver trusted data for downstream data consumers, with associated metadata to make this data easy to find and consume. A data product is the result of applying *data as a product* thinking to data for a specific purpose.

REMEMBER

The contents of a data product can take on a number of forms, such as a schema of tables in a relational database, a piece of code that outputs data, a SQL view, an embedded recommendation engine, or a fraud detection model.

Looking at Data Products and Data Mesh

REMEMBER

The data mesh perspective provides a more precise definition of a data product. Within this framework, a data product is a self-contained entity composed of code responsible for data collection, transformation, metadata definition, and infrastructure required to run the code, with the output being a dataset that is valuable in and of itself. While not every organization is mature enough or has the immediate need to adopt a data mesh approach, the concept of treating data as a product remains essential.

TIP

To make data products as useful and frictionless as possible, you need to consider the importance of certain attributes, including discoverability, addressability, trustworthiness, self-describing interoperability, and security.

Data mesh also advocates for application development teams to take ownership of the entire data product life cycle from inception through development and management to retirement. For example, the sales operations department, which manages the sales data system, would be responsible for serving sales data to the rest of the organization as a data product. Having lines of business responsible for data products ultimately increases the agility of the overall organization, because it avoids the bottleneck of relying on a central data team to provision and manage data.

Embracing all aspects required to adopt the data mesh data management philosophy, such as implementing new technologies and making significant organizational and cultural changes, is a major undertaking. But even without data mesh, by adopting a mindset of treating data as a product, organizations will prioritize stakeholder needs and value delivery, ensuring that data products are tailored to meet specific requirements.

TIP

Organizations can shorten the path between data and business value through the data product approach, which prioritizes delivering value as defined by stakeholders and linking data products to key performance indicators. Data product producers continuously and proactively iterate and refine data products towards the needs of the data product consumers. This iterative process enables data products to meet stakeholder needs and deliver organizational value both at the outset and as the corporate environment evolves, either through external competition or internal organizational change.

IN THIS CHAPTER

- Designing for value and usability
- Scaling data products
- Seeing the future of data products and large language models

Chapter 2
Creating, Managing, and Optimizing Data Products

In this chapter, we focus on key areas for your successful data product program: design, usability, scalability, and technology. Dive in!

Designing Data Products for Value

When establishing a data product initiative, know that the more visible organizational win will come from the business value provided by your initial data products. As your program expands, actively think of a plan for scalability and consistency across the organization. Here, we share a set of data product requirements that augment value and transcend specific business requirements.

Discoverability

Design data products in such a way as to promote data discoverability, ensuring that users can find and access the data they need easily:

- **Metadata:** Data products that include information about the data help users to understand its contents, context and find more easily. This metadata might include information about the data source, the date it was collected, business relevance, and other details that aid in comprehension.
- **Search functionality:** Use keywords and descriptions that are commonly used across your organization that can be indexed and searched via tools of your choice. If your organization utilizes a business glossary, applying the terms will help ensure consistent discovery. Many data products enable users to find specific data sets or topics. Optimize this search functionality to provide relevant results quickly and make it easier for users to find the needed data.
- **Data categorization:** Tag data with standard categories and subcategories such as domains and subject matter that are familiar across your organization to enable quick filtration and drilling down to data of relevance. Organize data products into categories and subcategories, such as topic, data source, date, or other relevant factors, making it easier for users to browse the available data sets and find their needs.
- **User feedback:** Collect user feedback about their experiences using the product so that developers can continuously improve the discoverability of data products.

Self-service and usability

Design data products to support self-service data exploration and analysis, making it easier for nontechnical users to work with data and gain insights. Here's how to promote self-service data product consumption:

- **User-friendly interface:** Provide a well-designed self-service, user-friendly interface that makes it easy for nontechnical users to work with data products independently.
- **Data accessibility:** Design data products so that they are easy and simple for users to access reliably, regardless of immediate or long-term business needs.
- **Documentation:** Generate documentation and user guides on data products that provide clear instructions on using the data. This might include providing definitions of data fields, explanations of data sources and limitations, or examples of how the data can be used to provide value to the business.

- **Data consumption:** Add charts, code snippets, and examples of consumption and output that can help the consumer reduce time to utilization.
- **Collaboration and sharing:** Inspire a data products culture by enabling users to collaborate on analysis and share insights. This might include features such as commenting, sharing, and publishing.
- **Training and support:** Provide training and support for data products to help users get up to speed with the product and learn how to use it effectively. This might include online tutorials, help documentation, and user forums.

Scaling Data Products

Every data product initiative needs a technology platform where users can easily design, manage, and govern all data products across all domains.

Building on the foundation

Data products are the foundational data building blocks for data insights and can be used independently, or combined downstream to create further ordinal data products. Secondary data products are derived from existing data products and add value by combining or transforming data to provide new insights. Tertiary data products combine secondary data with other sources or external data to provide even deeper insights or create new data products. This can help organizations become more competitive, agile, and responsive to changing market conditions.

REMEMBER

Data products serve as building blocks for secondary and tertiary data products in several ways:

- **Standardization:** By standardizing data products, organizations can ensure that all secondary and tertiary data products built on top of them use the same definitions and formats. This makes combining data from different sources easier and ensures consistency across different products.
- **Interoperability:** Data products designed to be interoperable can be easily combined with other data products to create new insights or applications. For example, a data

product that provides weather data could be combined with a data product that provides traffic data to create a new application that helps users plan their routes based on weather and traffic conditions.

- **Programmatic Access (API):** Data products that offer Application Programming Interfaces (APIs) can be easily integrated with other data products to provide new insights or create new applications. For example, an organization could create a custom application that combines data from multiple data products to provide a unique customer experience.

Governing at scale

Platforms to manage data products help administrators by providing a common framework to centrally manage and automate various data governance processes, improve data quality, and enable security and compliance across the organization. Here are some ways in which these platforms assist data admins in governing data products at scale:

- **Metadata management:** This enables admins to maintain a comprehensive and accurate inventory of data assets, including data lineage, quality, and usage information, to ensure that data is appropriately classified, tagged, and tracked across different systems and applications.
- **Data quality management:** Automated data quality checks and alerts enable admins to identify and address data quality issues promptly to improve data accuracy and consistency across the organization.
- **Access and security management:** Admins can control access to data assets, set data access policies, and monitor data access and usage to ensure that data is accessed and used in compliance with data privacy regulations and security policies.
- **Data integration and automation:** Admins can automate data processing workflows and streamline data integration processes, reducing the time and effort required to integrate data from different sources and ensuring that data is consistent and up to date.
- **Analytics and reporting:** Data admins can monitor and track data governance processes and performance metrics. This identifies areas for improvement and optimizes data governance processes at scale.

Reducing the cost of ownership

Data products reduce the total cost of ownership (TCO) of data management by improving data efficiency, reducing manual work, and streamlining processes, which helps organizations to save time and money while improving overall business performance. Here are some ways in which data products can assist in reducing the TCO:

- **Automated data management:** Data products automate various data management processes, including data integration, data quality, and data governance, reducing the need for manual intervention and increasing efficiency. This reduces operational costs and frees up resources to focus on strategic initiatives.
- **Improved data quality:** Data products provide automated data quality checks and alerts, improving data accuracy and consistency across the organization. This reduces the time and effort required to validate and correct data, leading to cost savings.
- **Self-service analytics:** Data products enable users to access and analyze data independently without relying on IT or data teams. This can reduce the burden on IT and data teams and free up resources to focus on other initiatives.
- **Scalable infrastructure:** Data products can exploit a scalable infrastructure and can grow with the organization's needs, reducing the cost of maintaining and upgrading infrastructure while ensuring that the organization can handle growing volumes of data.
- **Data-driven decision-making:** Data products can provide insights and analytics that help drive data-driven decision-making across the organization, helping to reduce the cost of making wrong decisions and improving overall business performance.

THE FUTURE OF DATA PRODUCTS

With the rise and adoption of generative AI and large language models (LLM), organizations will reduce the barrier to entry for users who may not have technical expertise or familiarity with complex interfaces. By leveraging LLM's natural language processing capabilities and

(continued)

(continued)

user-centric design, organizations can provide a more intuitive and user-friendly experience around data products.

Here's how LLMs will empower organizations for wider adoption of data products:

- **Democratizing data access:** LLM's natural language processing capabilities make data products more accessible to a wider audience. Users can interact with data products, eliminating the need for deep data expertise. Additionally, users who find data products easy to use and understand are more likely to engage with them and promote their use.
- **Data exploration and discovery:** LLMs can assist users in exploring and discovering insights from trusted data products. Users can ask questions, request specific analyses, or seek guidance, and LLMs can provide relevant responses and recommendations. This interactive data exploration facilitated by the LLM encourages users to explore and adopt data products as they gain valuable insights and make data-driven decisions.
- **Personalization and recommendations:** LLM's personalized recommendations based on user preferences and historical behavior create a more tailored experience for users. By suggesting relevant data products, or insights, the LLM increases user engagement and satisfaction, making users more likely to explore and promote data products to others.
- **Automation and efficiency:** LLMs automate data analysis tasks, such as data preprocessing, anomaly detection, or predictive modeling, saving users time and effort, enabling them to focus on deriving insights and value. The increased efficiency and productivity gained through LLM's automation features can contribute to the popularity of data products.

While LLMs drive the adoption of data products, factors such as the usefulness of the data products, value proposition, effective communication strategies, and addressing data privacy concerns also play significant roles. LLMs enhance the accessibility, usability, and user experience of data products, thereby promoting their adoption.

IN THIS CHAPTER

- Meeting engineers, product managers, and analysts
- Fostering consumption with the data product platform

Chapter 3
People and Processes

Data products serve as the units of exchange between the data producers and data consumers. In this chapter, we explore the crucial people involved: data platform engineers, data product producers, and data consumers. Data products need to be built atop a data product creation and exchange platform managed by a data platform engineering team, which we also cover.

Building Your Data Teams

Here, we explain the brainy bods you need on your data teams.

Data platform engineer

The data platform engineer is responsible for building and maintaining the infrastructure for the overall data ecosystem, including the data product platform, ensuring that data storage and compute capabilities meet the needs of data management and consumption. Oftentimes, this role is also responsible for providing the capabilities necessary to establish centralized governance policies to prevent the "Wild, Wild West" of data. The data platform engineer enables both the data product program and teams responsible for data products to be successful.

Additionally, the data platform engineer provides enforceable guardrails for data teams producing data products to freely

operate within. These may include treatment of personally identifiable information (PII) data, API standards, and data product templates and design standards.

The data product producer: Data product manager and data engineer

TIP

While multiple roles are responsible for the successful creation and consumption of a data product, the two key roles responsible for the success of a data product are the data product manager and the data engineer. In some organizations, the data product manager may be responsible for managing the strategy for the full data product life cycle, and in a separate role, the data product owner may be responsible for the day-to-day management of the data product. For a hybrid approach, we recommend a single role that spans the full responsibility of strategic-to-tactical data product execution and maintenance.

Having an understanding of the business needs of the data product is critical to the success of the data product initiative in its nascent stages. The data product manager and the data engineer must have a strong collaborative spirit to ensure that critical business needs are understood and being met, including:

- **Collaboration:** The data product manager defines the requirements based on the business needs of the data consumers, while the data engineer provides guidance on the technical feasibility and implementation design.
- **Prioritization:** The data product manager can provide guidance on the business value of different data engineering initiatives, and the data engineer can provide input on the technical complexity and feasibility of each project. Multi-domain data products may require engagement with different domain subject matter experts based on their potential impact and the available resources.
- **Technical implementation:** The data engineer is responsible for the technical implementation of the data product, encompassing the pipeline and the output of the pipeline. Understandability of the data product falls to both the data engineer and the data product manager. The data product manager is responsible for proactively obtaining feedback from the intended audience to ensure that the final data product output provides business value.

- **Continuous process improvement:** The two roles need to work together to identify areas for improvement in data product processes.

The data consumers: Data analysts and data scientists

Data consumers are the customers of the data product program and are responsible for consuming and transforming data into business value. The output of this group includes, but isn't limited to, dashboards, data and machine-learning (ML) models, reports, and much more. Often, their output is used for critical decision making.

As such, it's important to ensure that data consumers aren't in a strict "consumption-only" model. Rather, providing feedback after consuming data products is critical in providing the data product producers with actionable insights for continuous improvement of the data product.

Putting it All Together: Data Product Platform

Since a core goal of data products is to realize business value from data, fostering consumption is key.

Addressing the need for a data catalog

A vast majority of businesses still contend with the problem of only a handful of highly tenured data staff having knowledge about the data that isn't codified and gated.

This puts an undue burden on them to support a large number of employees with tasks that generally fall outside of their original responsibility, diverting time away from critical tasks to maintain data pipelines, data quality, and data infrastructure. Further, this type of arrangement risks business continuity when these "living data catalogs" leave the organization.

A successful data product initiative places the data consumers as the customers of the data team's output, and focuses on centralized governance and infrastructure to provide a high degree of discoverability and accessibility to data products.

Having centralized policies with decentralized governance

It stands to reason that data products are often an output of multiple data sources abstracted via a centralized data infrastructure. While data lakes and warehouses may constitute a single source of truth, this is, more often than not, an amalgamation of analytical data from multiple data sources. However, when considering agility, the various domain data teams either need to be empowered to manage their own corner of a centralized data infrastructure and the satellite operational and analytical data sources, or they need to have the ability to manage via a centralized data governance layer.


TIP

For the ultimate agility, we recommend hosting a centralized data product governance platform on top of or within a centralized data governance function. This is a layer that sits atop all of an organization's data sources and provides centralized governance around the following:

- Metadata management
- Data discovery
- Data integration
- Data delivery
- Security and access control


TIP

Modern solutions have started to realize the ability for centralized governance across a decentralized data ecosystem. Built around federated query engines, they realize the ability to create innovative federated data products, while dramatically decreasing the need for data movement and eliminating the need for centralized data infrastructure. The need to repetitively implement data governance across each data source and compute resources in each cloud region in each cloud vendor, compounded by each tool in the data stack, is starting to become a story of the past.

Further, with the ability to connect to a decentralized ecosystem that doesn't require data movement, this opens up opportunities that have traditionally been off limits: data products federated across the operational and analytical data planes. The main limitations become reduced to the types of data sources that are being used and the availability of the connectors with your solution of choice.

Chapter 4
Ten Tips for Adopting Data Products

Take inspiration from the following ten tips for smoothly and successfully adopting data products:

- **Focus on business values:** Ensure that technology is an enabler in reducing the cognitive load of data product developers (DPDs) so that they're able to focus on data and its related business context.
- **Link data products to key performance indicators (KPIs):** Leading KPIs provide a quantifiable framework to evaluate the impact and effectiveness of data products, ensuring their ongoing relevance and value within the organization, and driving continuous improvement.
- **Focus on user metrics and total cost of ownership (TCO) of data products:** By doing so, data products enable organizations to optimize their data strategies and investments, provide valuable insights into user behavior, drive user adoption, and help streamline resource allocation. As a result, organizations can make data-driven decisions that maximize the value and impact of their data assets while controlling costs and ensuring compliance.
- **Ensure governance by business role and responsibility:** By focusing on business governance, organizations establish trust in their data assets, mitigate risks, and ensure compliance. They promote accountability, data integrity, and ethical data

practices. With robust governance in place, businesses can make confident decisions based on reliable data, fostering a culture of data-driven decision-making across the organization.

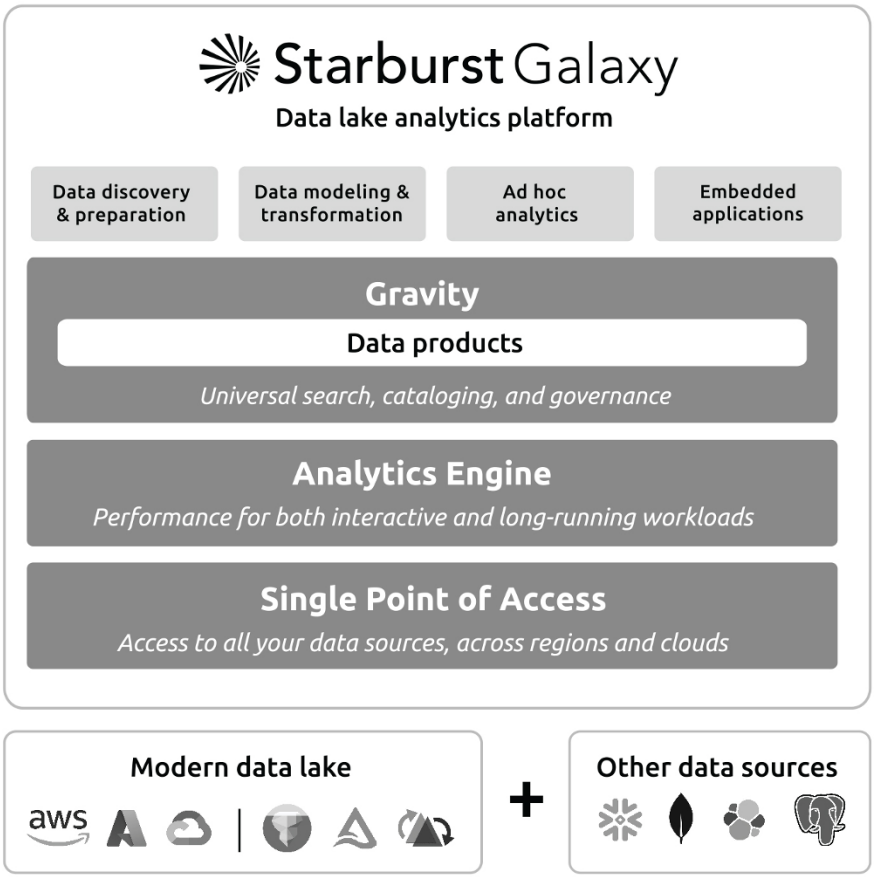
- **Design for use — with consumers in mind:** The design of data products for customers should be based on a deep understanding of their unique needs, preferences, and workflows, resulting in maximum value and fostering user adoption and engagement.
- **Reuse the same data products without making copies:** Data products that prioritize reuse without making multiple copies offer significant advantages in terms of cost savings, data consistency, and improved productivity. By leveraging technologies like data virtualization, data lakes, and data APIs, organizations can unlock the full potential of their data assets without the need for redundant copies.
- **Incentivize production and usage for data producers and data consumers — avoid shadow IT:** By aligning data products with user needs, providing adequate support and education, and ensuring the security and usability of authorized solutions, organizations can minimize the risks associated with shadow IT while maximizing the benefits of data-driven decision-making.
- **Invest in Data Product Owner/Manager Role:** The data product owner/manager is crucial in overseeing the entire life cycle of a data product. They act as the primary point of contact and advocate for optimizing the product, ensuring alignment between business objectives and product development. The data product owner collaborates with stakeholders to gather requirements, prioritize features, and make data-driven decisions.
- **Remember that iteration is key:** Through continuous feedback and iterative development around refinement, user-centric design, adaptation, testing, risk mitigation, continuous learning, and user adoption, data products can evolve, improve, and remain aligned with user needs and changing requirements, ultimately driving their effectiveness and value within the organization.
- **Invest in your organization's culture:** Adopting a data product mindset can build a culture of informed decision-making and continuous growth in your organization by promoting data-driven decision-making, increasing transparency, encouraging experimentation and innovation, facilitating collaboration, enabling self-service analytics, celebrating successes, and fostering continuous learning and improvement.

Starburst

Data products

Transform data collaboration and empower data consumers

- A single location to self-serve, discover, understand, and consume curated datasets
- Full visibility into data quality, lineage, and ownership
- Faster delivery times and improved self-service SLAs



Sign up for a free trial: starburst.io/platform/starburst-galaxy/start

Unlock the value in your data

Data has evolved from being mere information to a product that you need to oversee throughout its lifecycle. However, the challenges of trying to swiftly access that data, and easily extract meaning and value from it, remain.

Step forward *Data Products For Dummies,* your guide to unlocking the value in your data. Data products can inform decision-making, encourage experimentation and innovation, facilitate collaboration, and foster continuous learning and improvement. Read this book to discover how to maximize the value and impact of your data assets while controlling costs and ensuring compliance.

Inside…

- Understand data products
- Know the roles you need on your data teams
- Encourage data-driven decision-making
- Improve agility with centralized data governance
- Drive iterative development and refinement
- Look into the future of data products

Vishal Singh is Head of Product for Data Products and Data Governance at Starburst. **Ryo Komatsuzaki** is Senior Product Manager for Data Products at Starburst. **Andrew Mott, MBA,** is EMEA Head of Partner Solutions Architecture and a Data Mesh lead at Starburst.

ISBN: 978-1-394-21837-0
Not For Resale

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.