Gartner Data & Analytics Summit
29 – 30 May 2019 / São Paulo, Brazil

Upgrade From Self-Service to Enterprise Data Preparation for Analytics and Data Science Success

Rita Sallam

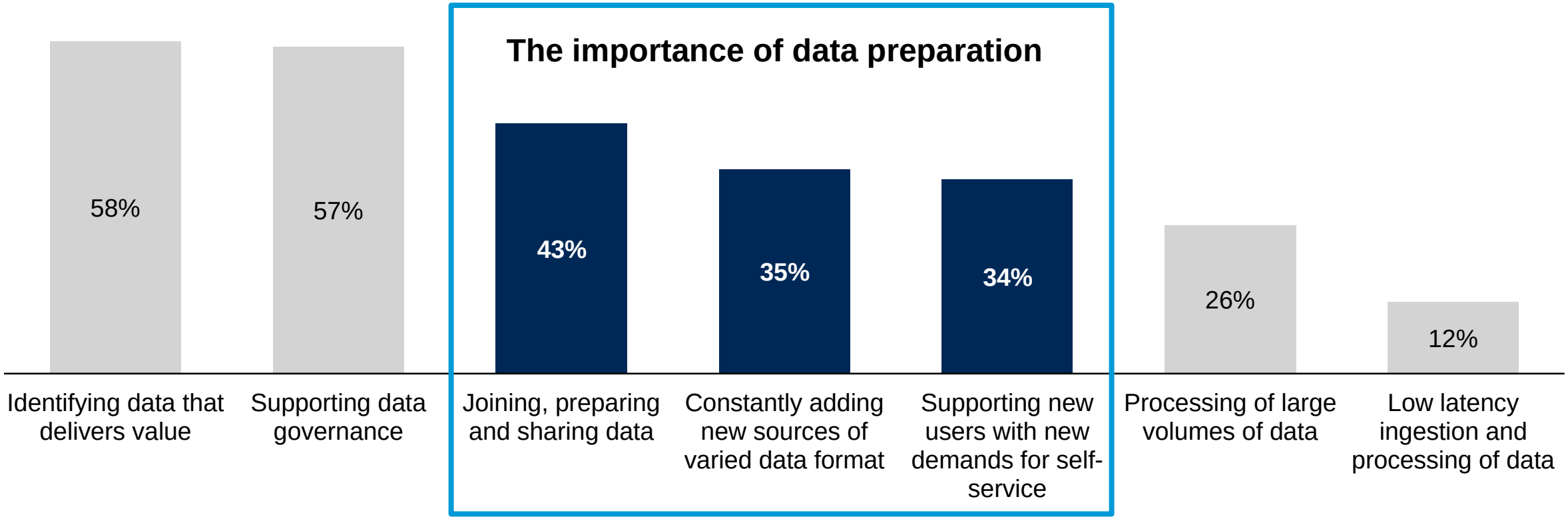
Gartner

Finding, Cleaning, Transforming and Sharing Data in a Trusted and Timely Manner Is the Biggest Challenge for Data and Analytics Teams

Gartner

Data Preparation — The Biggest Challenge for Data Management

Biggest challenges for a data management practice — Most are centered around data preparation!



Base: n = 113, Gartner Research Circle Members.
Q. New use cases drive the complexity of data management. What factors do you consider to be the most challenging to your data management practice? Please select up to three.

Gartner

Key Issues

1. What is data preparation and how has it evolved to support enterprise data integration, analytics/BI and data science needs?
2. What is the current state of the market and how will it evolve?
3. How can you best leverage data preparation capabilities, while at the same time ensuring governance and control?

Key Issues

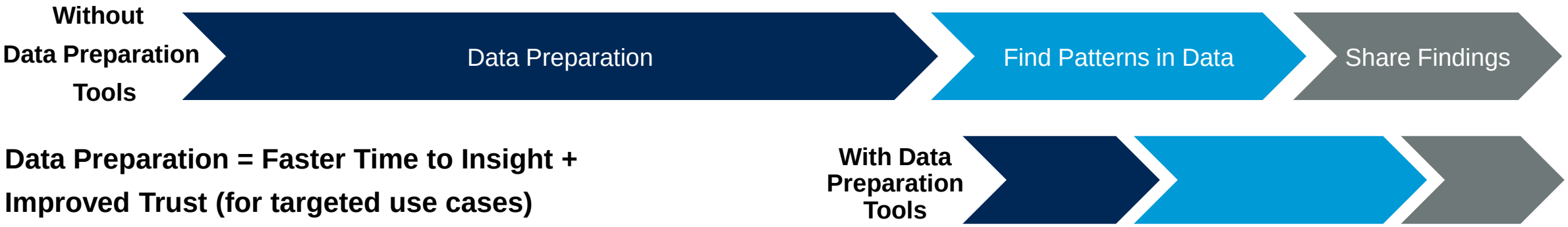
1. What is data preparation and how has it evolved to support enterprise data integration, analytics/BI and data science needs?
2. What is the current state of the market and how will it evolve?
3. How can you best leverage data preparation capabilities, while at the same time ensuring governance and control?

Gartner

What Is Data Preparation?

Organizations continue to report about 60%-80% time spent on finding, accessing, preparing and sharing data for further analysis

Data preparation reduces the time to insight for analytics and some operational use cases



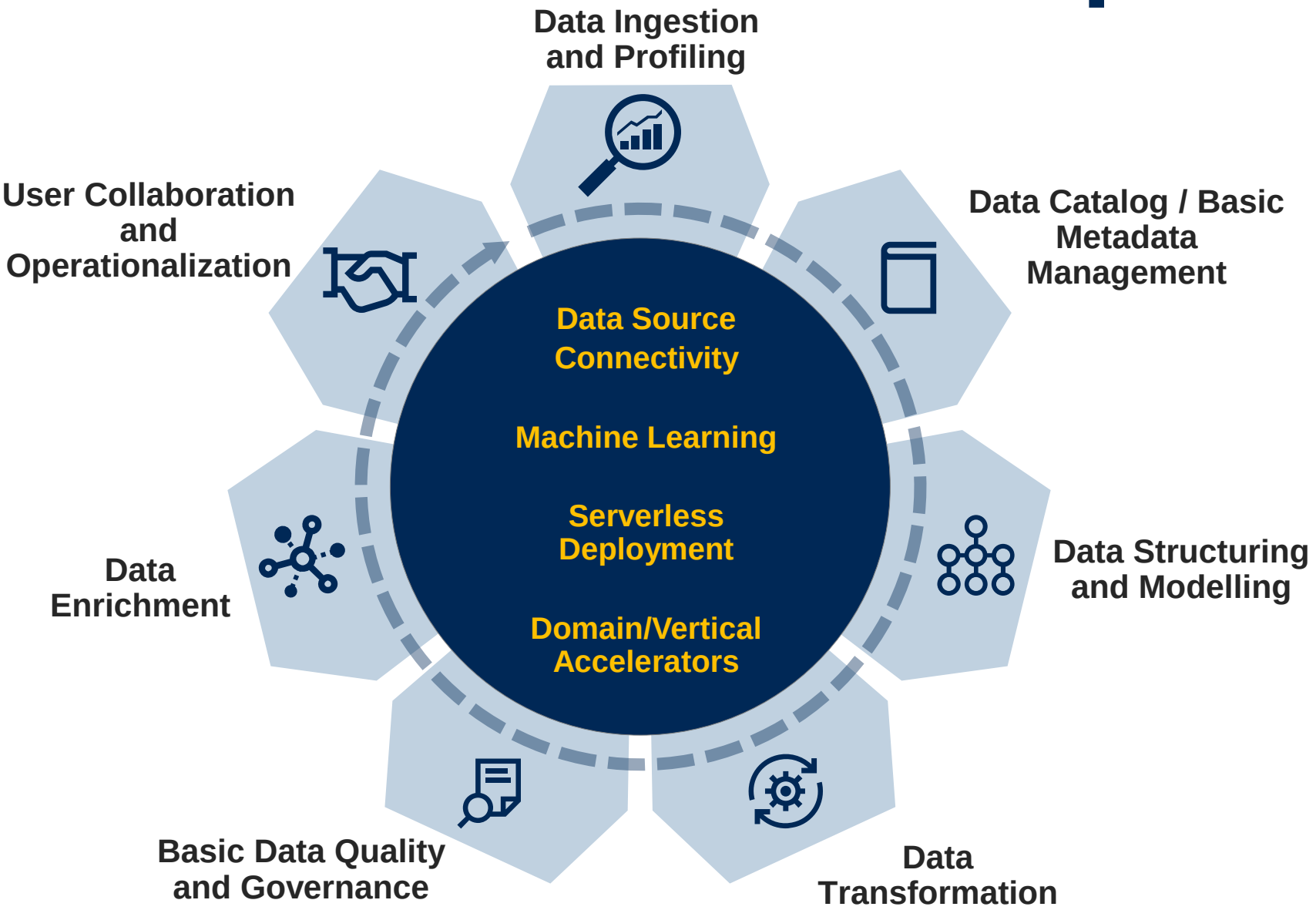
Data Preparation = Faster Time to Insight +

Improved Trust (for targeted use cases)

Data preparation is an iterative, agile process for *finding, combining, cleaning, transforming & sharing* raw data into curated datasets for self-service data integration, analytics/BI and data science use cases.

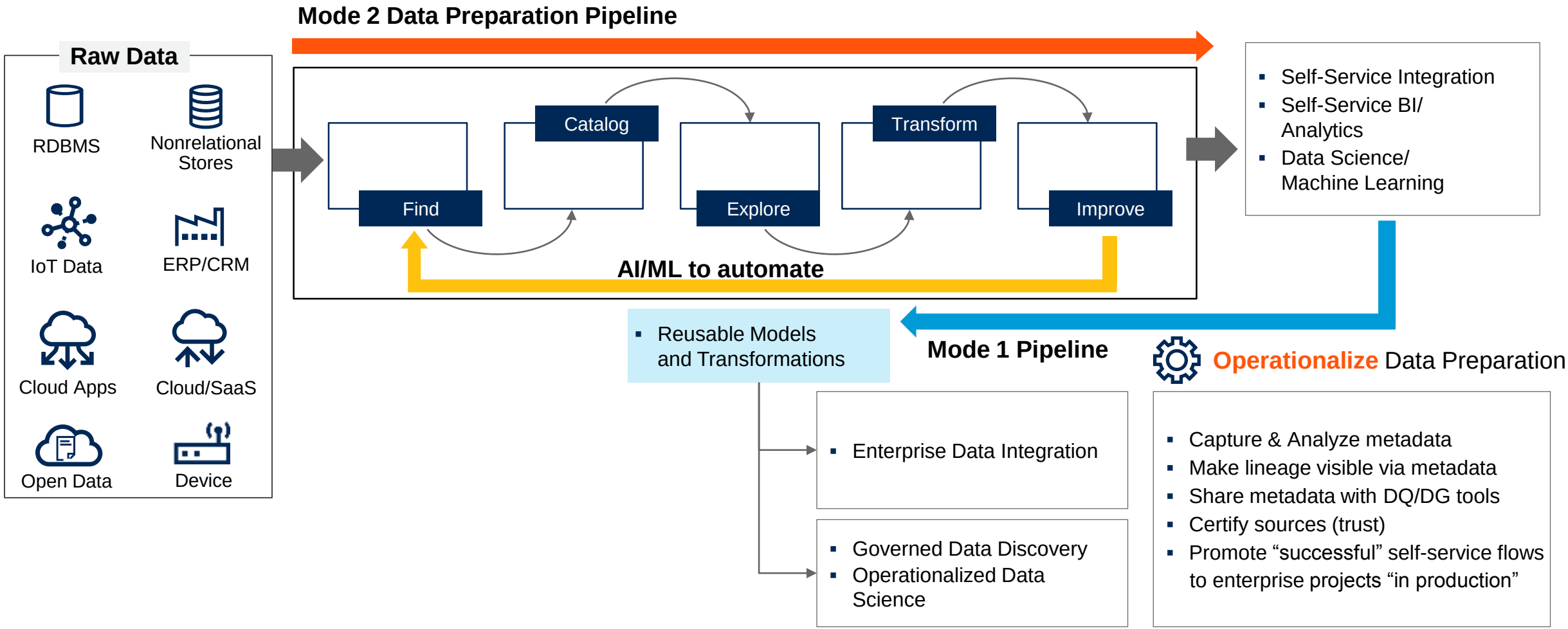
Gartner

Key Capabilities of a Modern Data Preparation Tool



Gartner

The Modern Enterprise Data Preparation Pipeline — Operationalization Is Key!

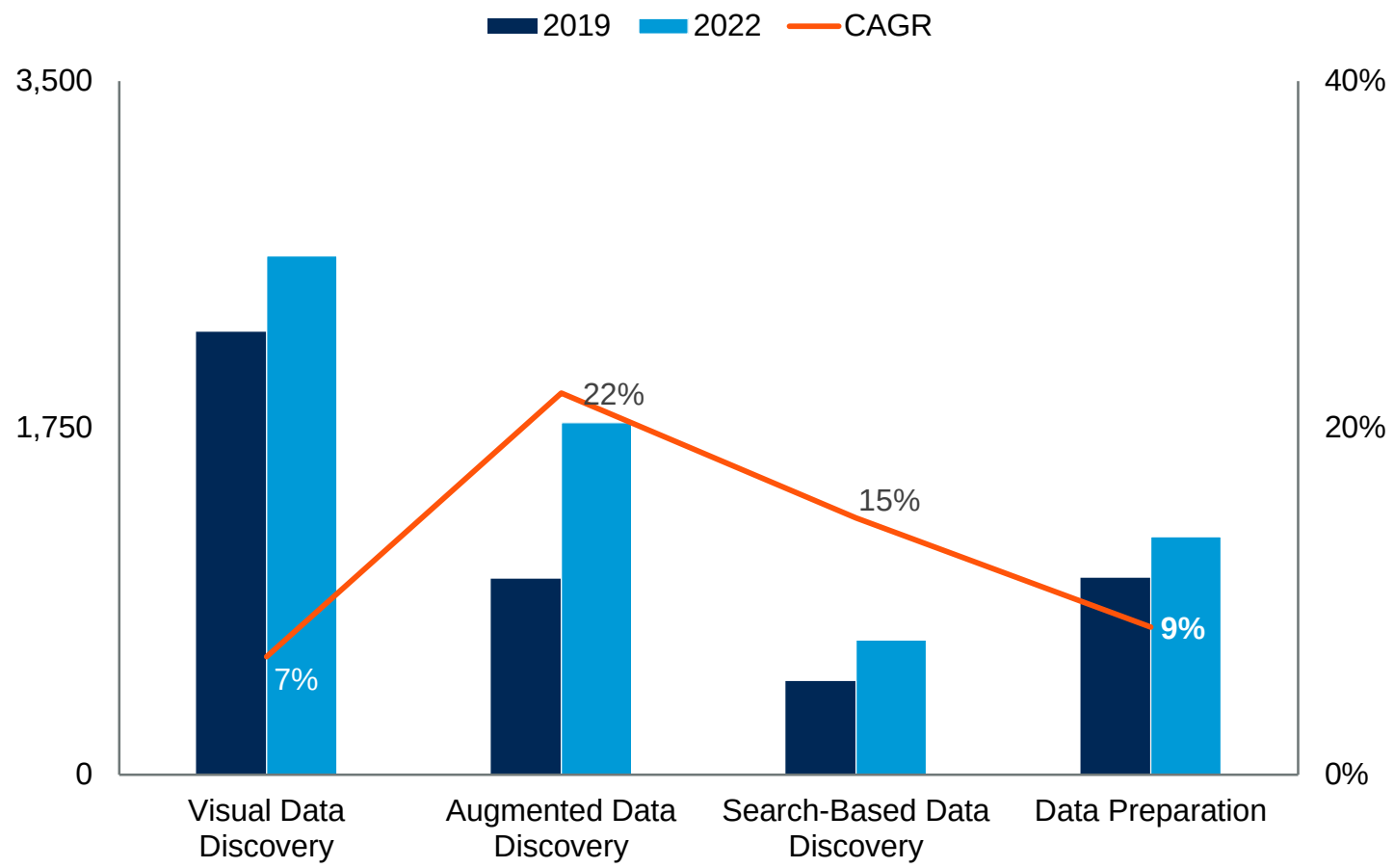


Key Issues

1. What is data preparation and how has it evolved to support enterprise data integration, analytics/BI and data science needs?
2. What is the current state of the market and how will it evolve?
3. How can you best leverage data preparation capabilities, while at the same time ensuring governance and control?

Gartner

The Market for Data Preparation Tools Continues to Grow!

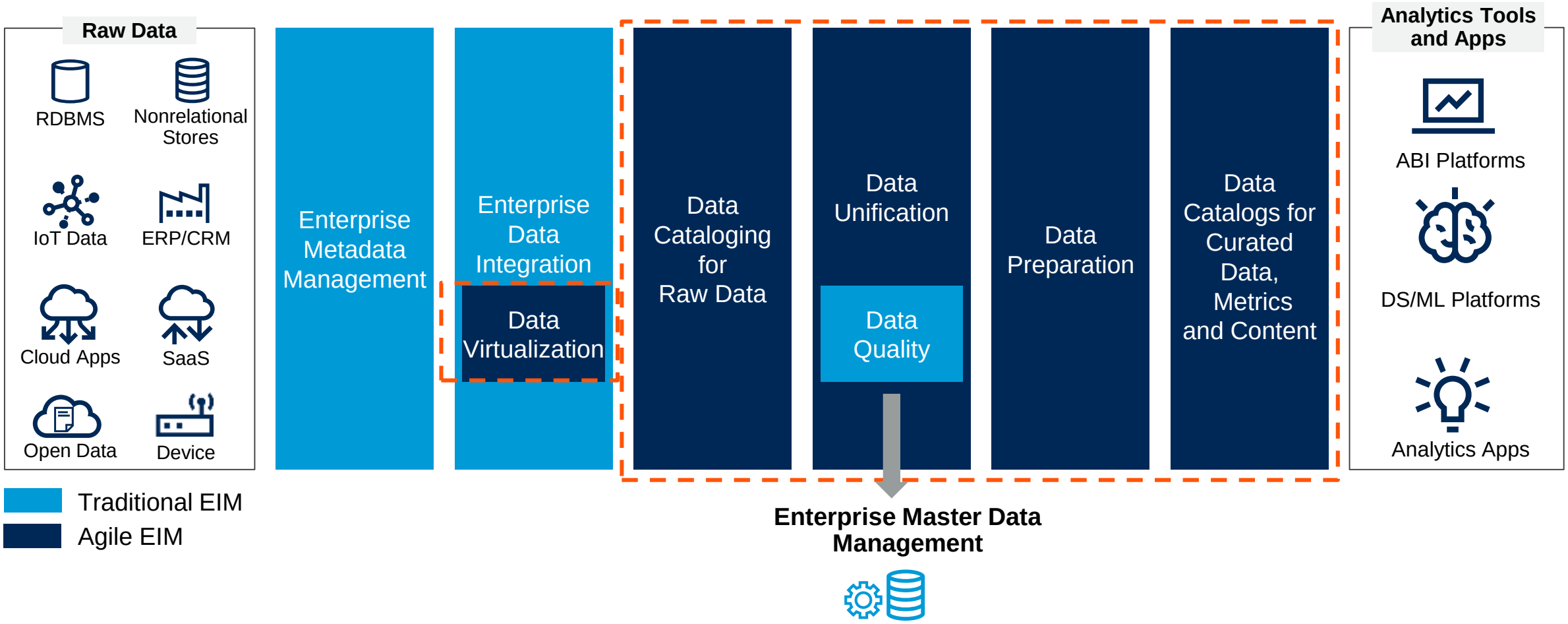


- The market for **data preparation** is currently estimated to be around $993 Million in software revenue
- It will continue to grow at a healthy 8.5% CAGR (through 2022), reaching an estimated value of **$1.2 billion by 2022**

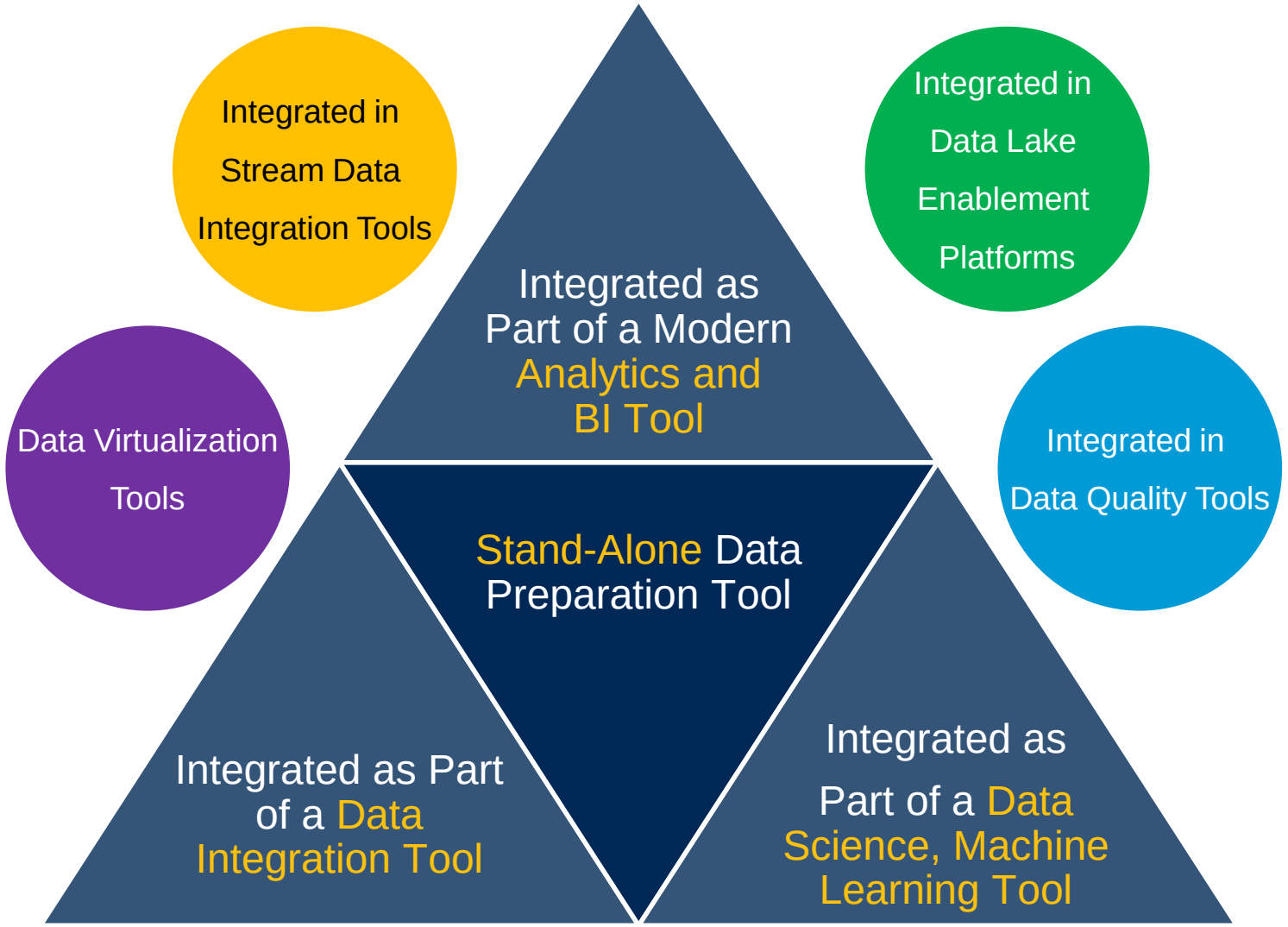
Source: "Forecast: Modern Business Intelligence Platforms by Selected Functionality, Worldwide, 2017-2022"

Gartner

The Data and Analytics Market Is Evolving — Point Solutions Will Likely Converge Into a Modern Consolidated Platform



Data Preparation Market — The Segments



- FIRST, **always evaluate your existing tools** in data management and analytics
- Then evaluate **existing and upcoming use cases**
- **Only then make gradual investments!**
- Invest in stand-alone data prep. tools for **general purpose and stack independent** data preparation

Gartner

	Stand-Alone Data Preparation Tool	Integrated as Part of a Data Integration/ Data Quality Tool	Integrated as Part of a Modern Analytics/ BI Tool	Integrated as Part of Data Science/ML Platform
Alteryx	Alteryx Designer		Alteryx Analytics Platform	Alteryx Analytics Platform
Cambridge Semantics	Anzo	Anzo	Anzo	Anzo
ClearStory Data	ClearStory Data		ClearStory Data	
Datameer	Datameer Enterprise			
Altair (Datawatch)	Datawatch Monarch, Datawatch Swarm			
Infogix (Lavastorm)	Data3Sixty Analyze	Data3Sixty	Data3Sixty	Data3Sixty
Lore IO	Lore Platform			
Paxata	Paxata Self-Service Data Preparation			
SAP	SAP Agile Data Preparation	SAP HANA EIM, SAP Data Hub	SAP Analytics Cloud	SAP Predictive Analytics
SAS	SAS Data Loader for Hadoop, SAS Data Preparation	SAS Data Loader for Hadoop	SAS Data Preparation	SAS Data Preparation
Talend	Talend Data Preparation	Talend Integration Cloud, Talend Data Fabric		
Tamr	Tamr Unify (Data Unification)			
Trifacta	Trifacta Wrangler, Google Cloud Dataprep			
Unifi Software	The Unifi Data Platform			
Yellowfin	Yellowfin Data Prep		Yellowfin Suite	
	***Representative list only — not an exhaustive collection of all data preparation tool vendors**			
Dataiku				Data Science Studio
Experian		Aperture Data Studio (Data Quality)		
IBM				Data Refinery
Informatica		Informatica Enterprise Data Preparation		
Microsoft			Power BI	
MicroStrategy			Microstrategy	
Oracle			Oracle Analytics Cloud	
Qlik (Podium Data)		Qlik Data Catalyst	QlikSense, QlikView	
Tableau			Tableau Desktop	
TIBCO			TIBCO Spotfire	

Machine Learning Is Significantly Impacting Data Preparation

Engagement

- **Collaboration support:** Automated tagging, profiling and annotation
- Automated assignment of **role-based user access**
- ML enabled-**crowdsourced curation, access & transformation**

Insight

- Automated **flagging** of risky datasets
- Automated **suggestions**
- AI powered **search and discovery**
- **Semantic relationships** using knowledge graphs

Automation

- Automated **metadata discovery** and extraction
- Automated **anomaly detection & reporting**
- **Self-healing** Optimization
- Detection & **Assignment of PII**
- **Automated repetitive transformations**
- **Recommendations** using **active metadata** analysis

Gartner

Strategic Planning Assumptions

- By 2020, data preparation tools will be used in more than 50% of new data integration efforts for analytics and data science.

- By 2020, organizations that offer users access to a curated catalog of internal and external prepared data will realize twice the business value from analytics investments than those that do not.

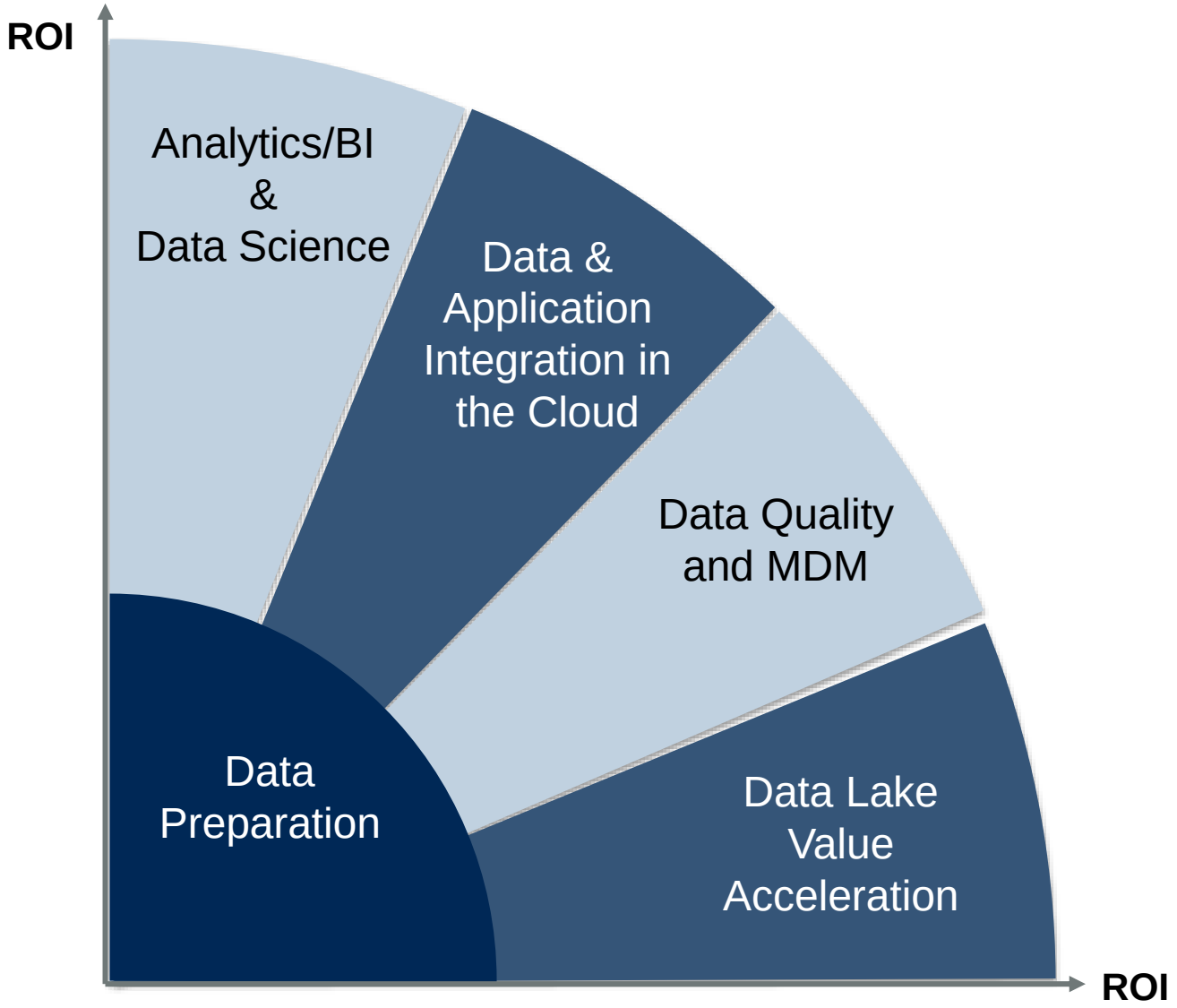
Gartner

Key Issues

1. What is data preparation and how has it evolved to support enterprise data integration, analytics/BI and data science needs?
2. What is the current state of the market and how will it evolve?
3. How can you best leverage data preparation capabilities, while at the same time ensuring governance and control?

Gartner

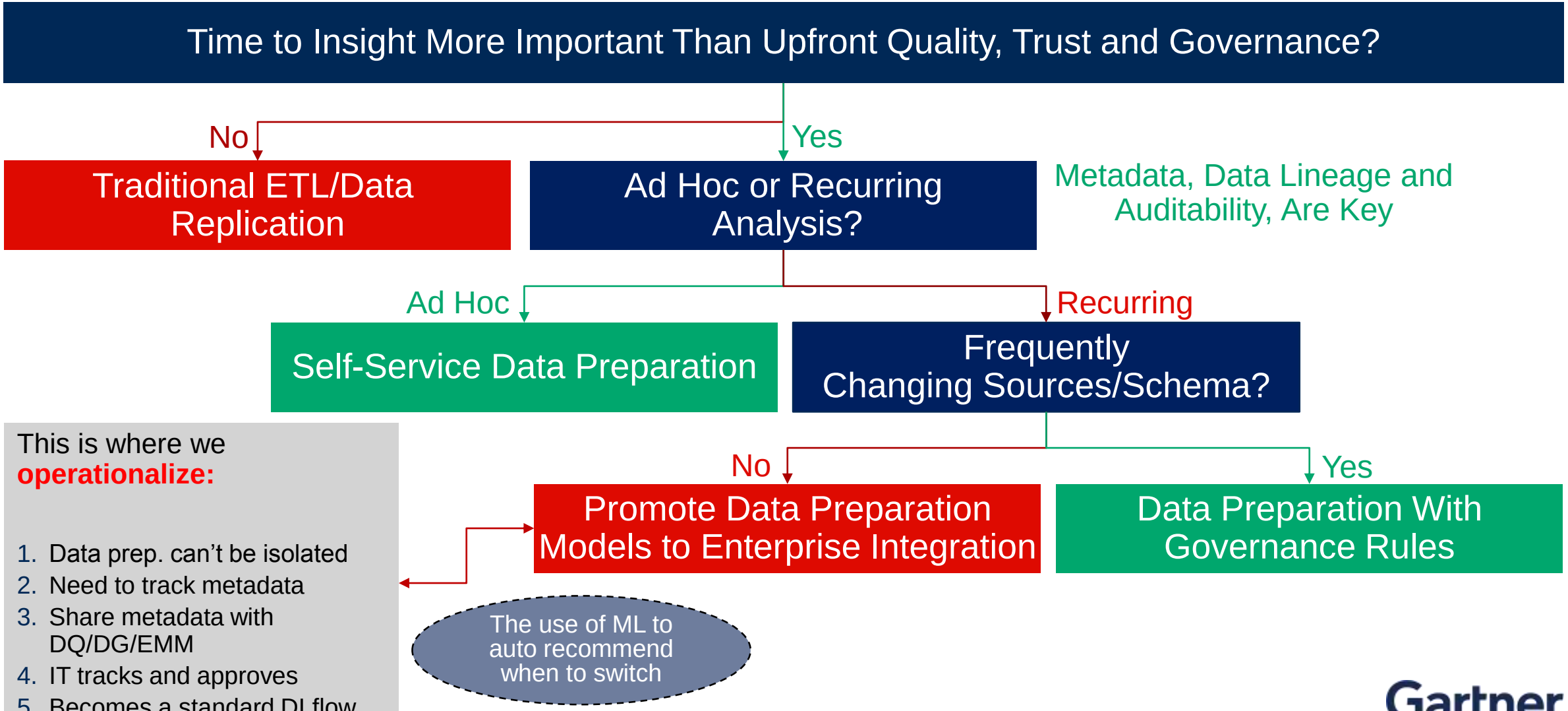
Data Preparation Use Cases



- Data Prep. creates an entry point to multiple use cases
- It matures to **Enterprise Use Cases** e.g.,
  - Governed Data Discovery Operationalized Data Science
  - Enterprise MDM
  - Operationalized Data Lakes etc.
  - Cloud Integration

Gartner

When to Use Data Preparation (Ensuring Consistency and Control)— Versus Using ETL?

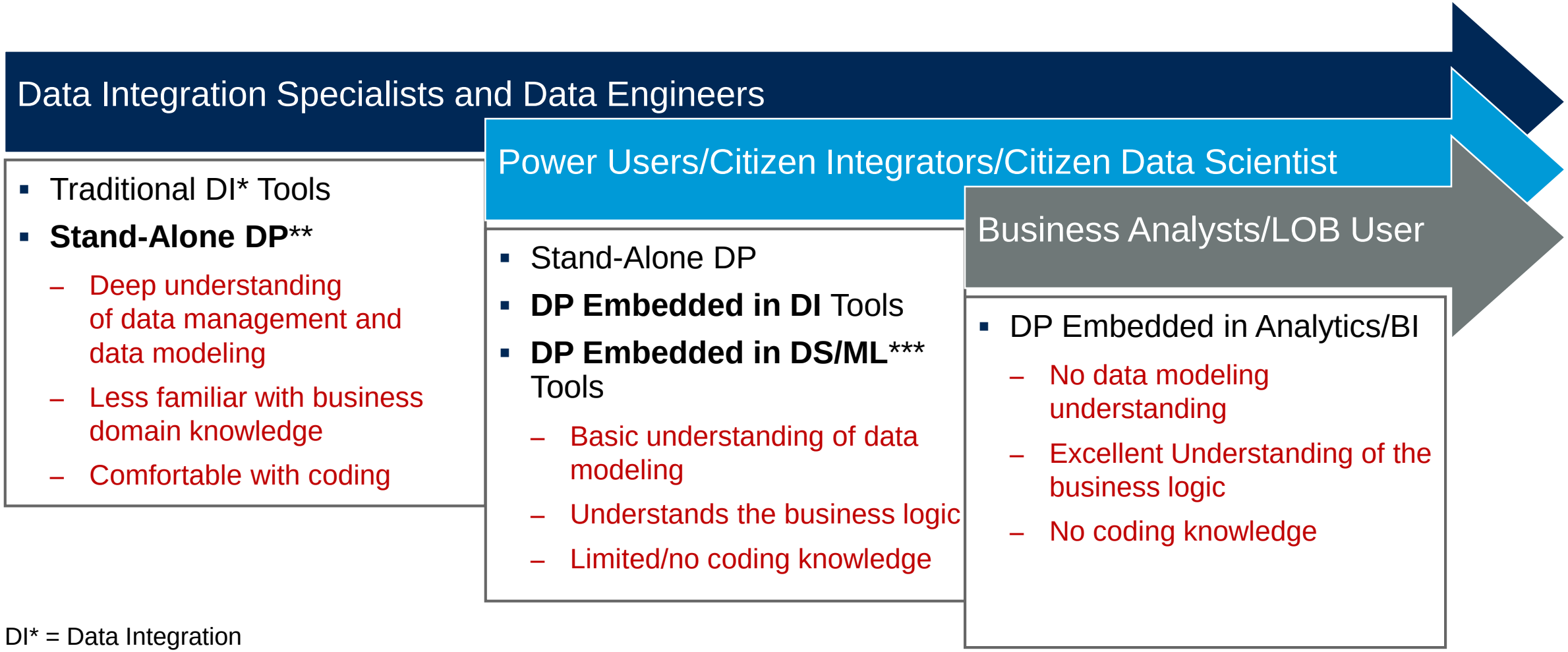


This is where we **operationalize:**

1. Data prep. can't be isolated
2. Need to track metadata
3. Share metadata with DQ/DG/EMM
4. IT tracks and approves
5. Becomes a standard DI flow

Gartner

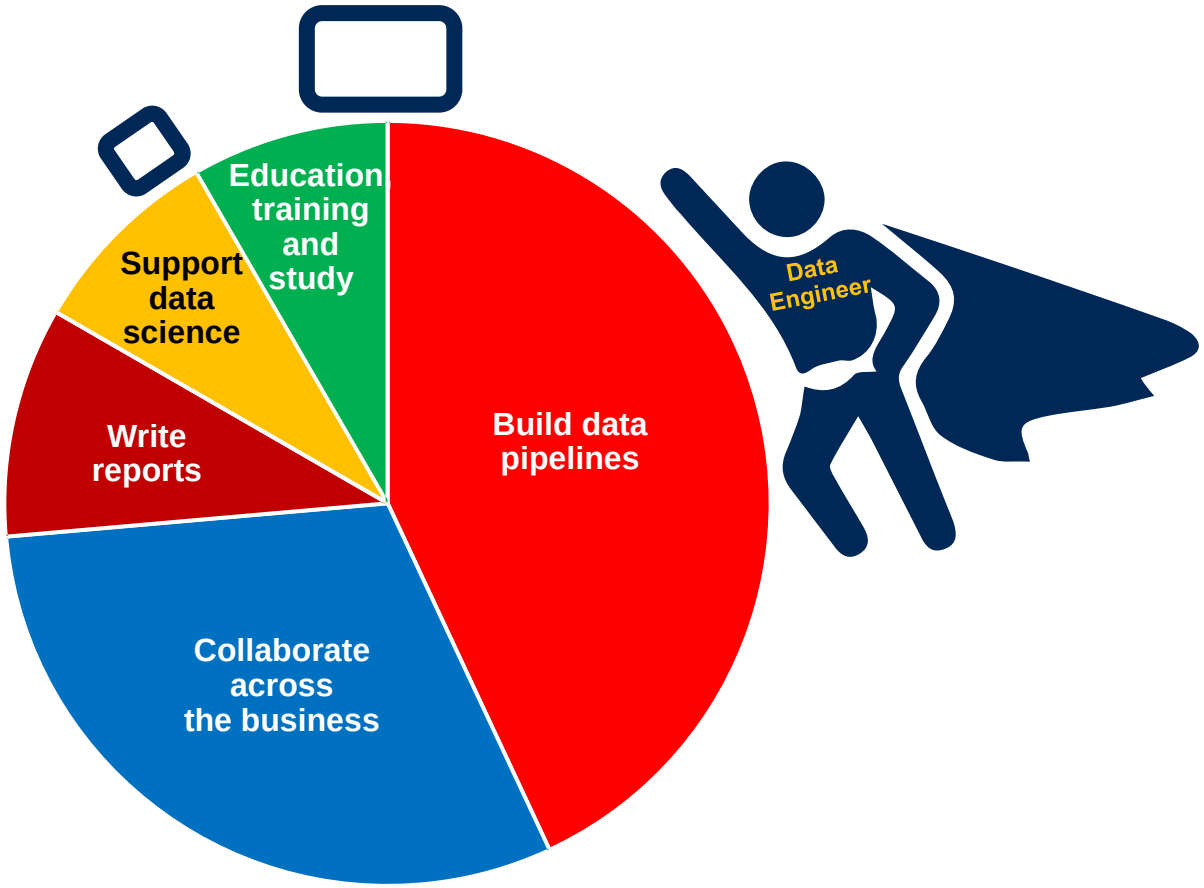
Positioning Different Types of Data Preparation Tools to the Right Persona



DI* = Data Integration
DP** = Data Preparation
DS/ML*** = Data Science & Machine Learning

Gartner

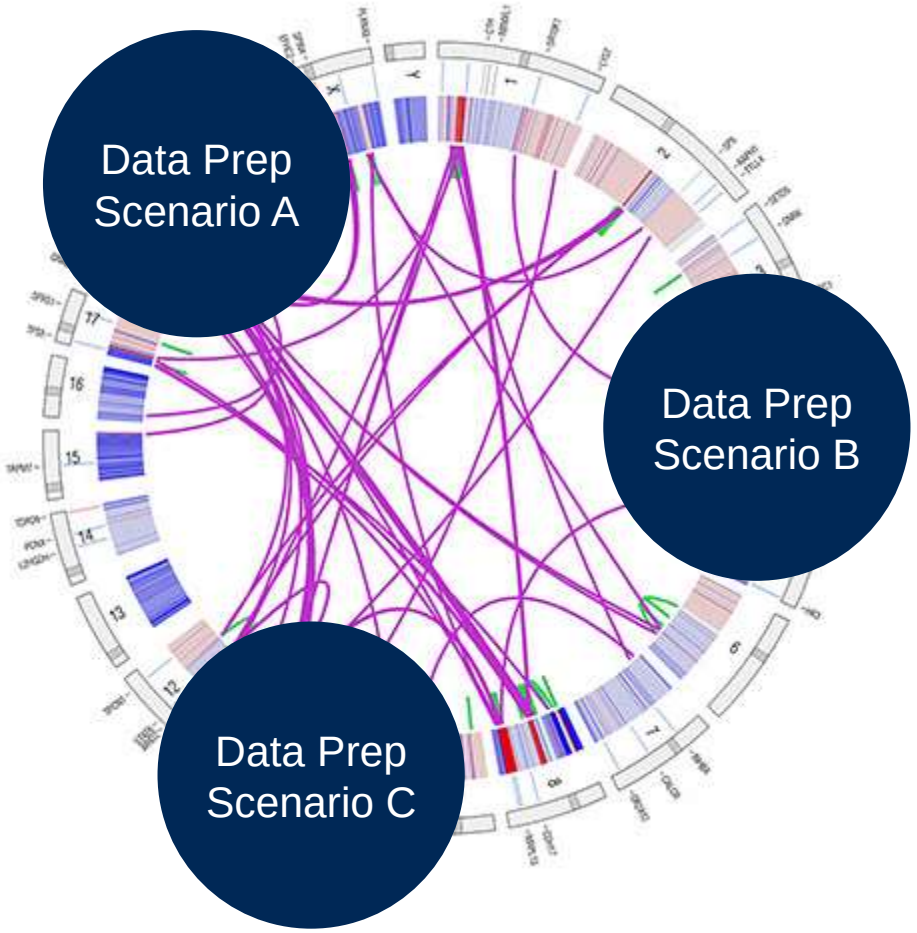
The Rise of the Data Engineer — Arm Them With Data Preparation for Efficiency



- **Data engineers** will be instrumental to your **data science** and **analytics success.**
- A significant portion of their time goes into **building data pipelines**. They are unable to help Analysts and Data Scientists
- As a result – Analysts & Data scientists end up doing data prep. which is not a productive use of their time.
- Engineers are expected to integrate new data sources quickly — the **proliferation of data sources, and volume is not helping!**
- **Armed with the right data preparation tool —** data engineers will be able to deliver integrated datasets for analytics/data science quicker.

"The role of **data engineer —** is emerging in IT, and is using Data Prep. **to speed the creation of curated, trusted data pipelines** for a range of **distributed analytics content authors**"

Metadata Management and Governance Are Vital!



- Discover and capture metadata, and synchronize models (with DQ, DG and modeling tools)
- Derive physical/logical models, discern and reconcile relationships
- **Track data lineage and impact**
- Share metadata with other tools (bidirectional) and synchronize with diverse instances
- Data Preparation DOES NOT replace the need for effective data quality and governance.

Without effective metadata management, data preparation becomes brittle, chaotic and siloed.

Gartner

Case Study: GE

Agile, Multidomain Entity Mastering & Prep Drives $500M+ In Value

Technical Challenge

- **Suppliers**: Build an integrated view from 75+ ERP systems and 2M supplier records
- **Parts:** 25M nonunique parts in purchasing systems across 8 BUs
- **M&A:** Integrate data from acquired entities with existing master views

Technical Outcome

- <6 months from pilot to globally deployed data pipeline; 2M records consolidated to 700k
- 25M reduced to 6.4M unique parts
- Data from three acquisition units integrated with GE's in <2 weeks

Business Challenge

- **Suppliers:** Get GE's best terms with a given supplier in every negotiation
- **Parts:** Optimize sourcing strategies to most cost-effective suppliers
- **M&A:** Increase the velocity of realizing synergies post-acquisition

Business Outcome

- **$80M savings** in year 1 — due to the mastered unified supplier view
- **$300M in annual savings** identified (0.5% reduction of direct spend)
- Supplier, purchasing and customer-base opportunities quickly identified

Source: Tamr

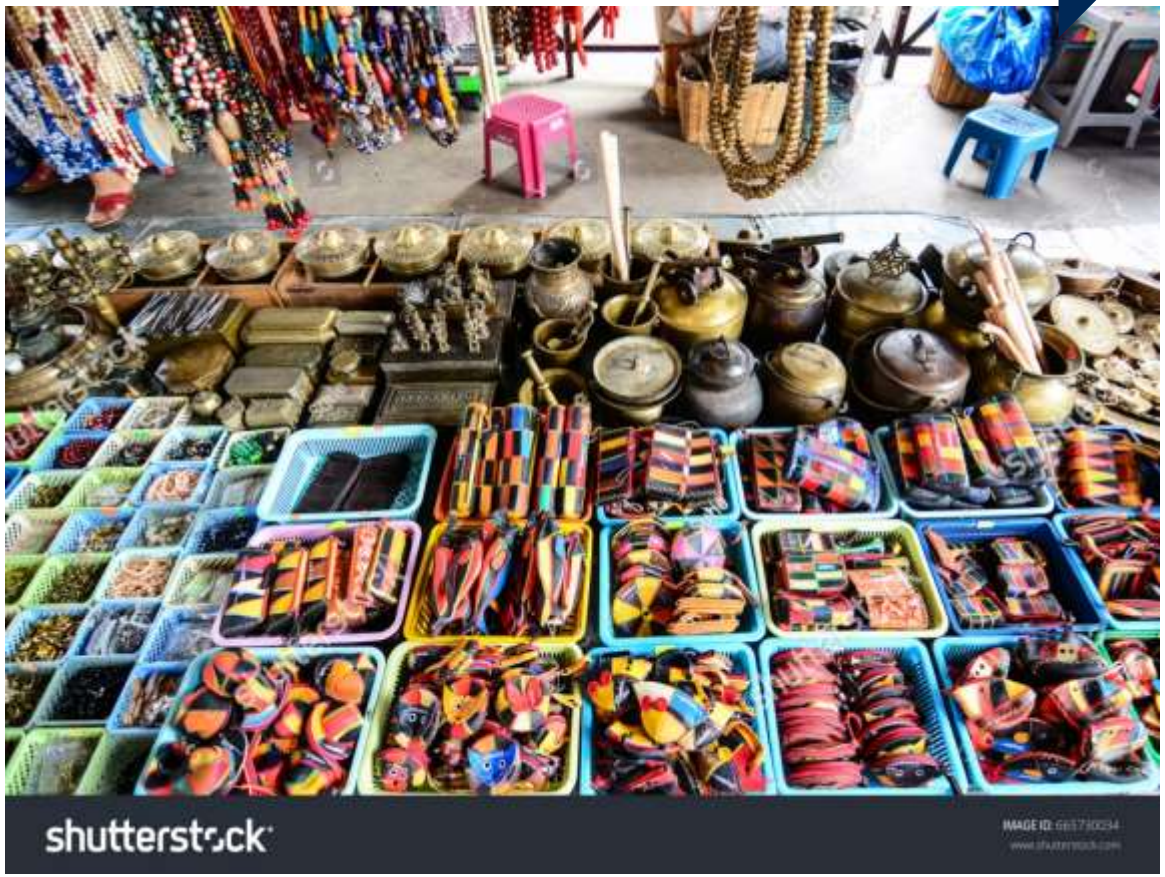
Case Study — Kaiser Permanente Improves Operational Efficiency

- Millions in **operational cost overruns** on **overestimation** of staffing & **medical supplies**
- Inability to analyze cost & **procedure data** across 38 hospitals, 628 facilities & 18k physicians
- **Existing spreadsheet-based** process unable to manage growing data diversity
- Kaiser **reduced time-to-analysis by 97%** with new data preparation process
- Uncovered **3x cost disparity** for the same procedure across different medical facilities

Source: Trifacta

Gartner

Data Preparation!



Gartner

Recommendations

- **Develop** a deployment strategy for data preparation to enhance user understanding of data, to increase agility in data integration.
- **Create** a formal process for vetting the self-service models, for operationalizing data preparation flows.
- **Recognize that**, while data preparation tools can be used for an increasing number of data integration use cases, they do not yet replace the need for enterprise data integration.
- **Investigate** your data preparation tool vendors' roadmap on their current/planned support for:
  - Data Science/ML Libraries
  - Improved Data Quality/Data Governance
  - Improved User Collaboration and Sharing
  - Embedded Data Cataloging Capability
- ✓ **Evaluate** and choose the most appropriate vendor offering by considering:
  - Deployment model
  - Pricing
  - Domain-/Industry-specific accelerators, KPIs or starter templates
  - Support for new data sources
  - Support for new end-user roles (and)
  - **Existing** vendor tools in which you have **already invested**

Gartner

Recommended Gartner Research

- **Market Guide for Data Preparation**
  Ehtisham Zaidi, Rita Sallam and Shubhangi Vashisth (G00315888)
- **Position Your Data Preparation Tools Effectively to Drive Product Growth**
  Sharat Menon, Terilyn Palanca and Ehtisham Zaidi (G00369469)
- **Toolkit: Self-Service Data Preparation**
  Rita Sallam, Paddy Forry, Ehtisham Zaidi and Shubhangi Vashisth (G00313410)
- **Toolkit: Job Description for the Role of a Data Engineer**
  Ehtisham Zaidi, Roxane Edjlali, Nick Heudecker and Mark Beyer (G00354447)
- **Data Catalogs Are the New Black in Data Management and Analytics**
  Ehtisham Zaidi, Guido De Simoni, Roxane Edjlali and Alan D. Duncan (G00338777)
- **Magic Quadrant for Business Intelligence and Analytics Platforms**
  Rita Sallam, Cindi Howson and Others (G00301340)

Gartner