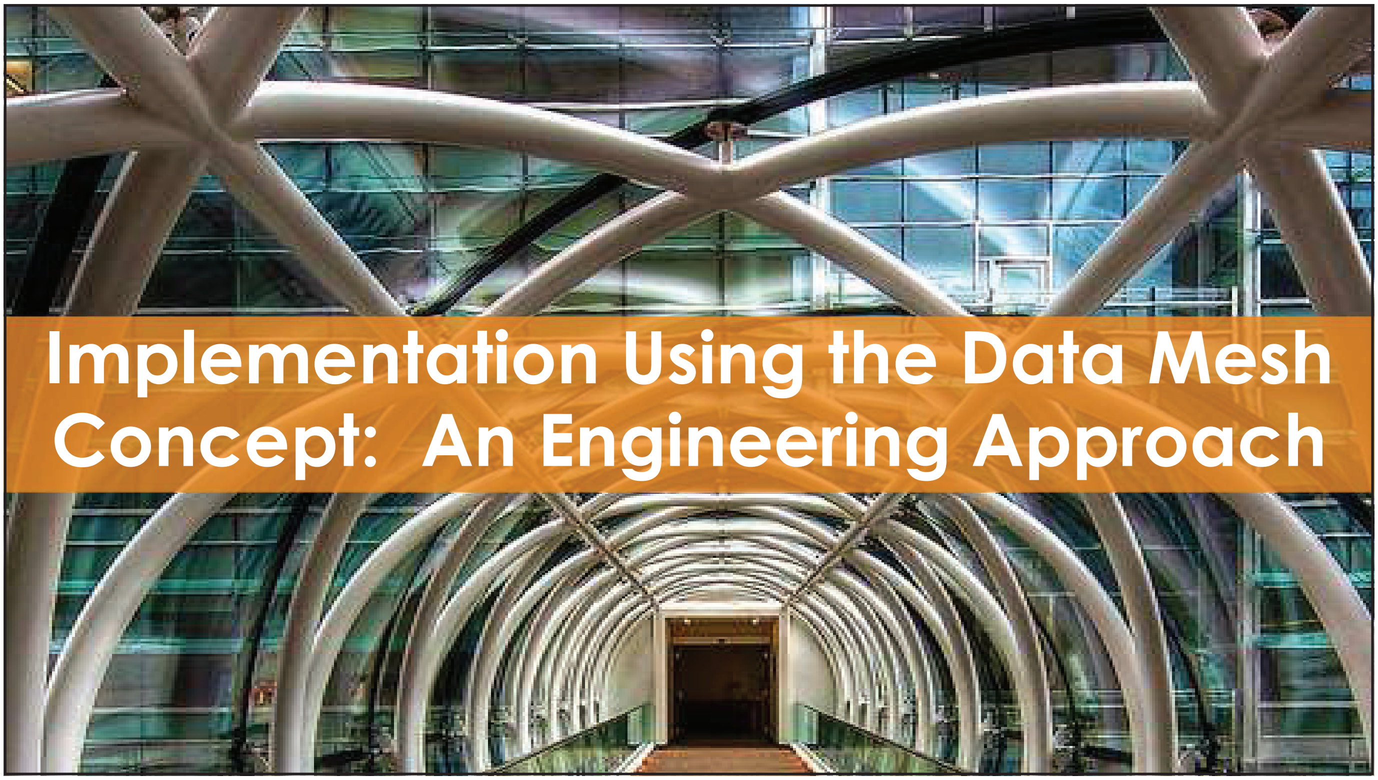
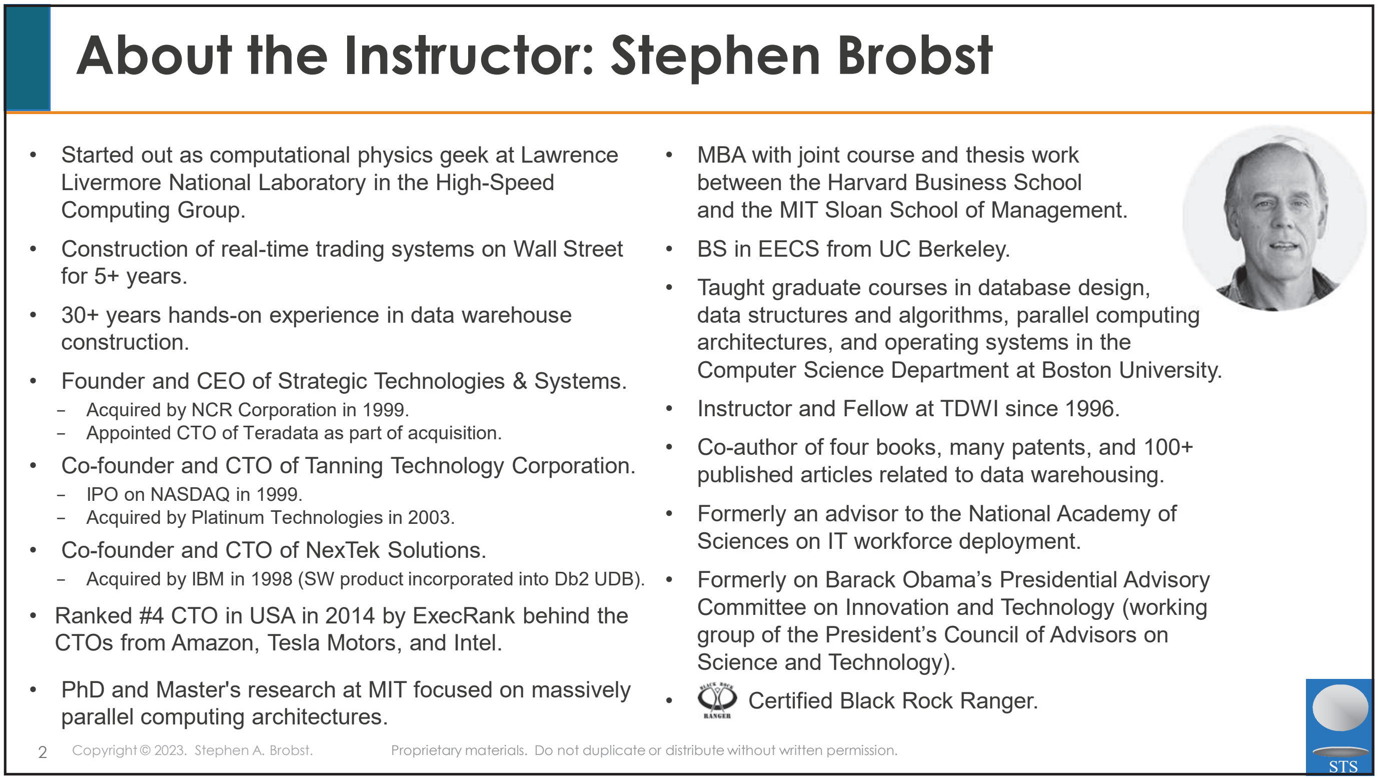


Implementation Using the Data Mesh Concept: An Engineering Approach

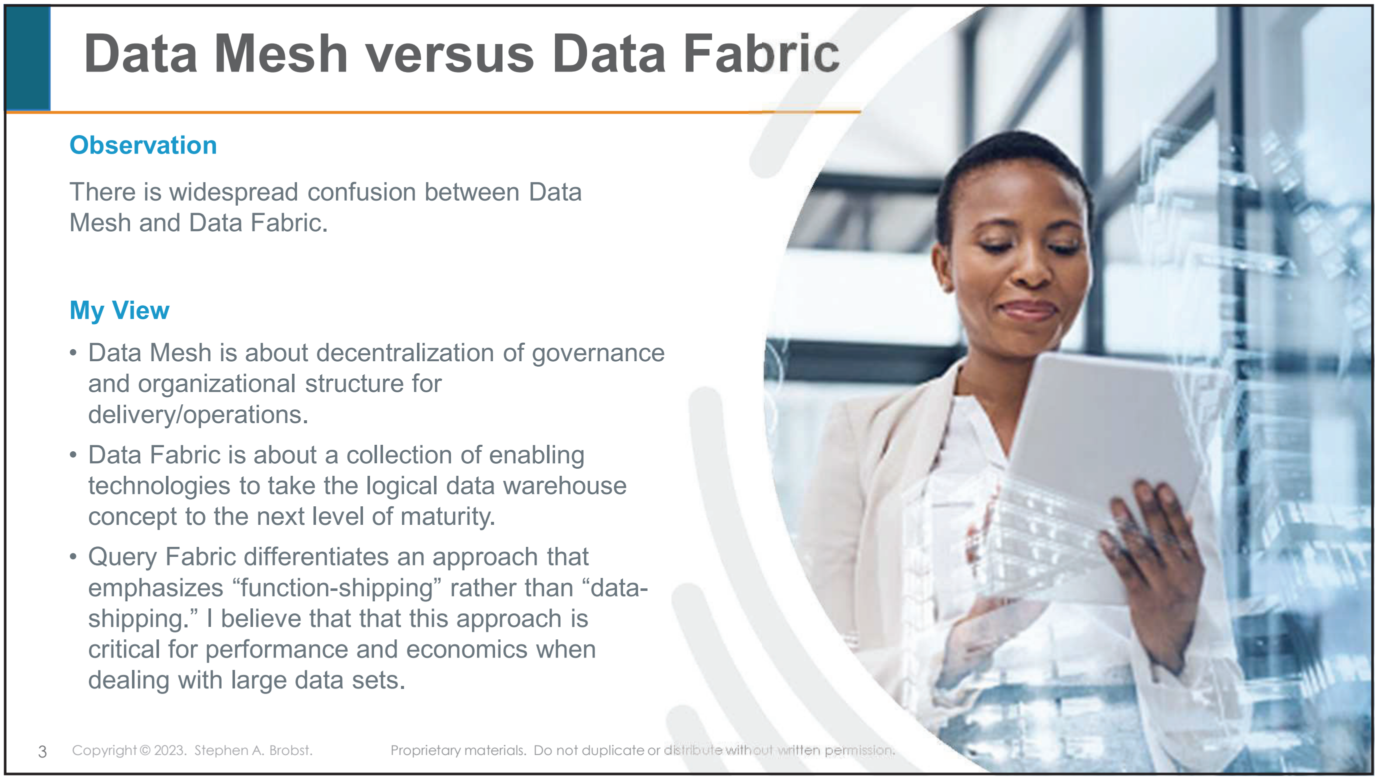
1



About the Instructor: Stephen Brobst

- Started out as computational physics geek at Lawrence Livermore National Laboratory in the High-Speed Computing Group.
- Construction of real-time trading systems on Wall Street for 5+ years.
- 30+ years hands-on experience in data warehouse construction.
- Founder and CEO of Strategic Technologies & Systems.
  - Acquired by NCR Corporation in 1999.
  - Appointed CTO of Teradata as part of acquisition.
- Co-founder and CTO of Tanning Technology Corporation.
  - IPO on NASDAQ in 1999.
  - Acquired by Platinum Technologies in 2003.
- Co-founder and CTO of NexTek Solutions.
  - Acquired by IBM in 1998 (SW product incorporated into Db2 UDB).
- Ranked #4 CTO in USA in 2014 by ExecRank behind the CTOs from Amazon, Tesla Motors, and Intel.
- PhD and Master's research at MIT focused on massively parallel computing architectures.
- MBA with joint course and thesis work between the Harvard Business School and the MIT Sloan School of Management.
- BS in EECS from UC Berkeley.
- Taught graduate courses in database design, data structures and algorithms, parallel computing architectures, and operating systems in the Computer Science Department at Boston University.
- Instructor and Fellow at TDWI since 1996.
- Co-author of four books, many patents, and 100+ published articles related to data warehousing.
- Formerly an advisor to the National Academy of Sciences on IT workforce deployment.
- Formerly on Barack Obama's Presidential Advisory Committee on Innovation and Technology (working group of the President's Council of Advisors on Science and Technology).
- Certified Black Rock Ranger.

2



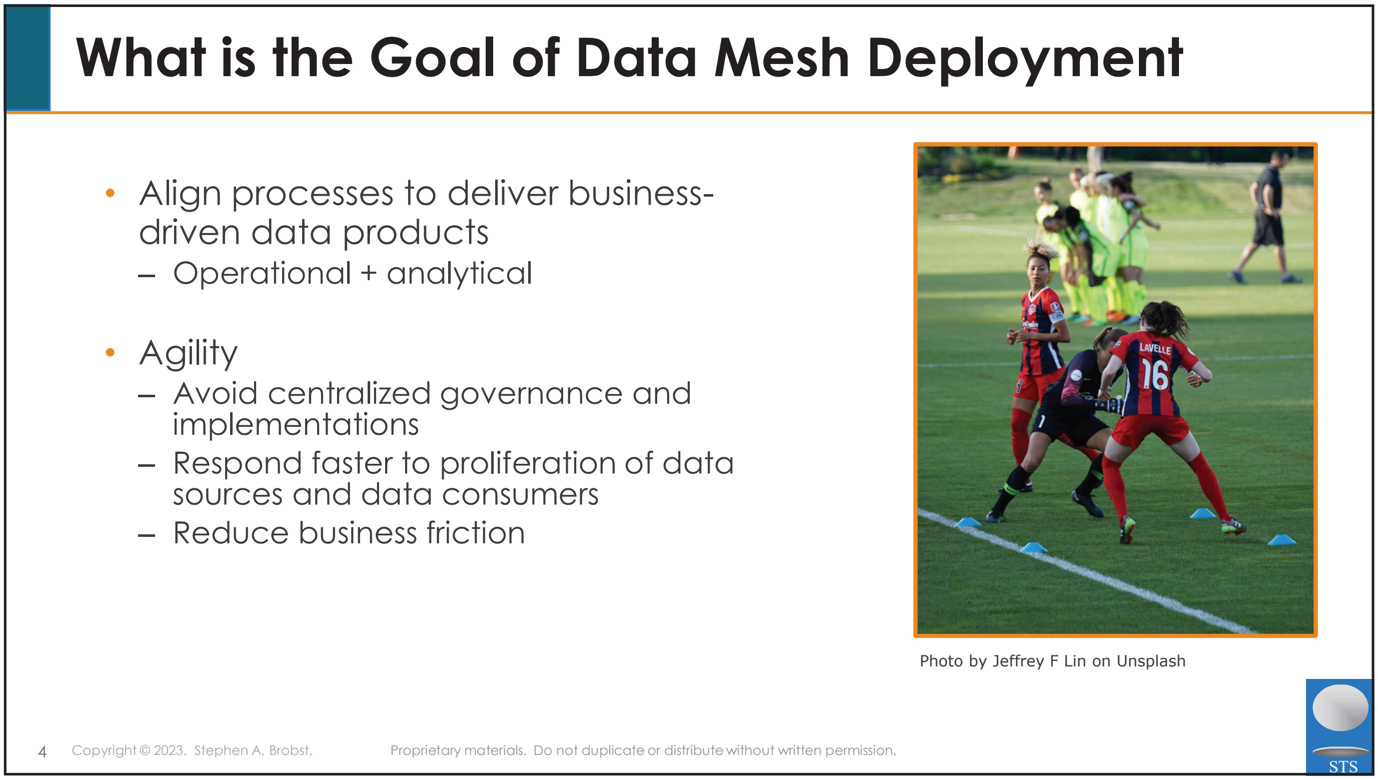
Data Mesh versus Data Fabric

Observation

There is widespread confusion between Data Mesh and Data Fabric.

My View

- Data Mesh is about decentralization of governance and organizational structure for delivery/operations.
- Data Fabric is about a collection of enabling technologies to take the logical data warehouse concept to the next level of maturity.
- Query Fabric differentiates an approach that emphasizes "function-shipping" rather than "data-shipping." I believe that that this approach is critical for performance and economics when dealing with large data sets.

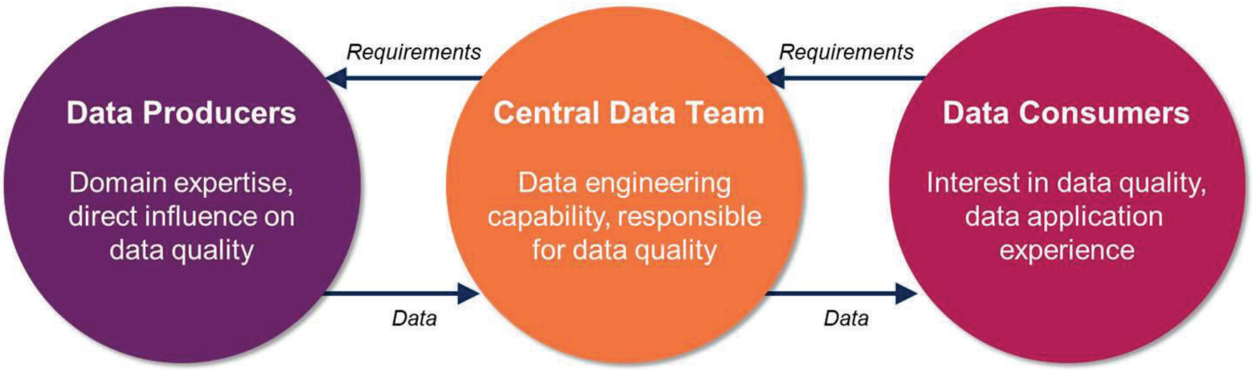


What is the Goal of Data Mesh Deployment

- Align processes to deliver business-driven data products
  - Operational + analytical
- Agility
  - Avoid centralized governance and implementations
  - Respond faster to proliferation of data sources and data consumers
  - Reduce business friction

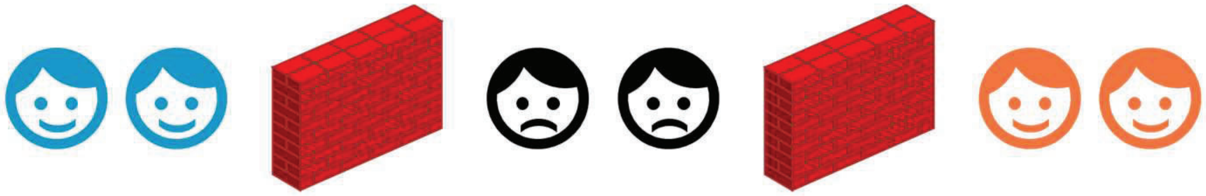
Photo by Jeffrey F Lin on Unsplash

Between a Rock and a Hard Place


Requirements
Requirements
Data Producers
Domain expertise, direct influence on data quality
Central Data Team
Data engineering capability, responsible for data quality
Data Consumers
Interest in data quality, data application experience
Data
Data

https://www.thoughtworks.com/insights/blog/data-mesh-its-not-about-tech-its-about-ownership-and-communication

One of These is Not Like the Others



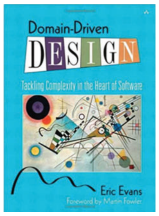
Domain-oriented **data producers**

Domain-agnostic **data and platform engineers**

Domain-oriented **data consumers**

Domain-Driven Design: Tacking Complexity in the Heart of Software

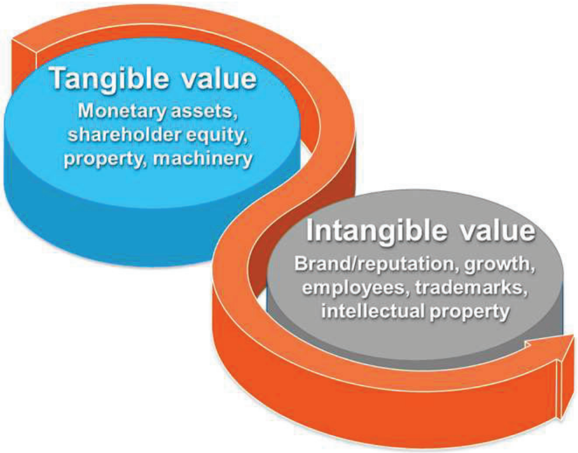
> **"…an approach to developing software for complex needs by deeply connecting the implementation to an evolving model of the core business concepts." Eric Evans, 2003**


Domain-Driven
DESIGN
Tackling Complexity in the Heart of Software
Eric Evans
Foreword by Martin Fowler

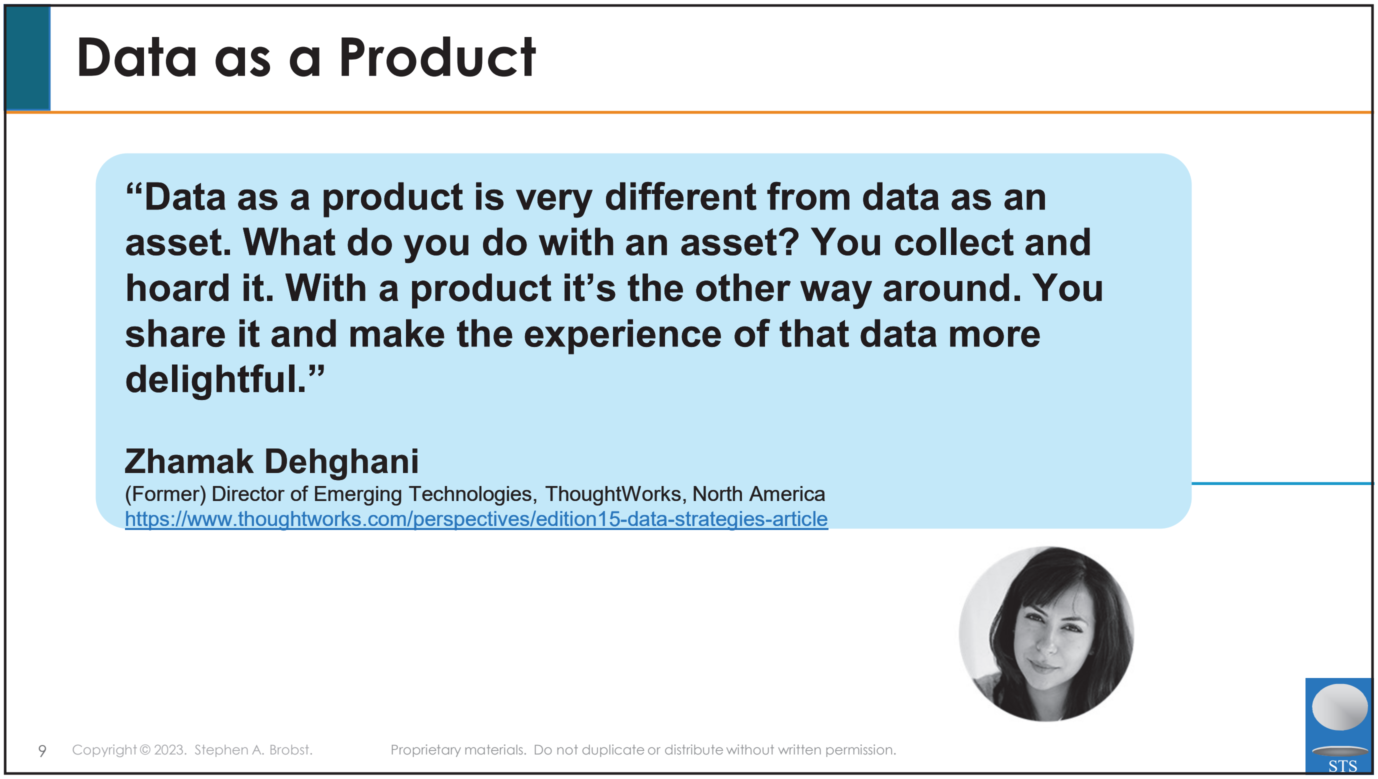
STS

Data Mesh: Business Value Driven

- 1st principle: business value
  - Goals drive teams & data architecture
- 2nd principle: business domains
  - Context within boundaries
  - Data & process ownership
- 3rd principle: distributed responsibility
  - Loosely coupled IT+LOB teams
  - Decentralized plans, decisions, actions
  - Inherent agility


Tangible value
Monetary assets, shareholder equity, property, machinery
Intangible value
Brand/reputation, growth, employees, trademarks, intellectual property

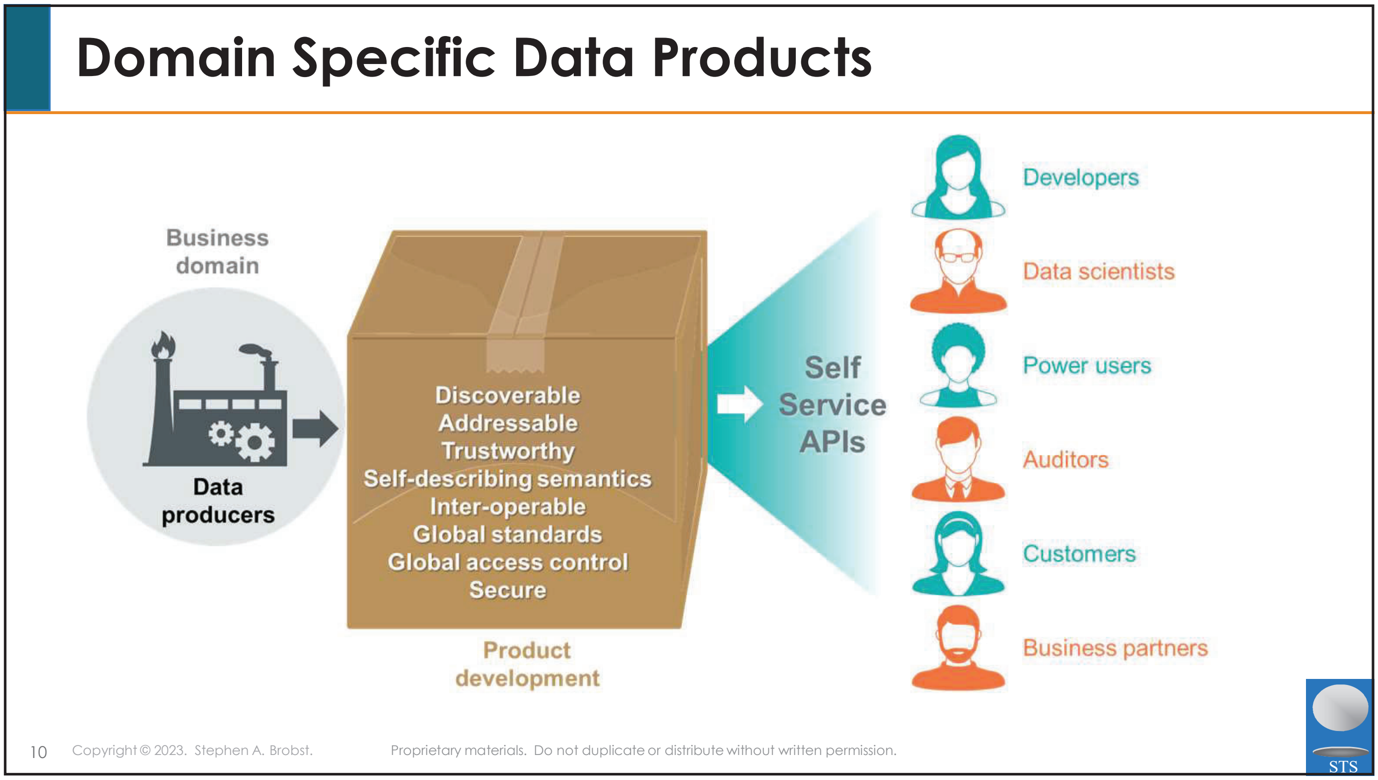
STS



Data as a Product

> "Data as a product is very different from data as an asset. What do you do with an asset? You collect and hoard it. With a product it's the other way around. You share it and make the experience of that data more delightful."
>
> **Zhamak Dehghani**
> (Former) Director of Emerging Technologies, ThoughtWorks, North America
> https://www.thoughtworks.com/perspectives/edition15-data-strategies-article

9



Domain Specific Data Products

10

What Data Mesh Gets Right

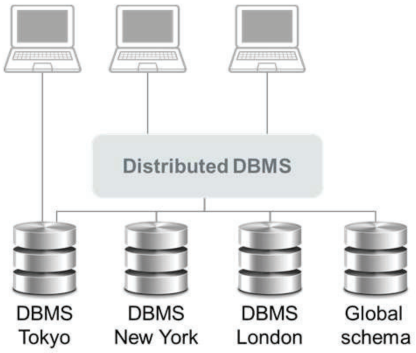
> **"Data Mesh attempts to strike a balance…it gives domain teams autonomy to have control of their local decision making, such as choosing the best data model for their data products…while uses…governance policies to impose a consistent experience across all data products; for example, standardizing on the data modeling language that all domains utilize."**
>
> **"Data Mesh places a domain-agnostic data platform team who empowers the domain teams with self-serve capabilities to manage the lifecycle of data products declaratively and consistently, to prevent team isolation and decrease cost of autonomy."**
>
> **"…fundamental assumptions that have remained unchallenged for the last few decades and must be closely evaluated:**
>
> - **Data must be centralized to be useful - managed by a centralized organization…**
> - **Data management architecture, technology and organization are monolithic.**
> - **The enabling technologies dictate the paradigm - architecture and organization."**
>
> Zhamak Dehghani, Data Mesh – Delivering Data Driven Value At Scale, O'Reilly, 6/22/2021

STS

Confusion with Data Mesh

- Decentralized teams do not necessarily imply distributed database deployment!
- Technology ≠ architecture
- Quiz: data mesh replaces
  - Data warehouses?
  - Data lakes?
  - Data marts?
  - ETL?
  - None of the above?

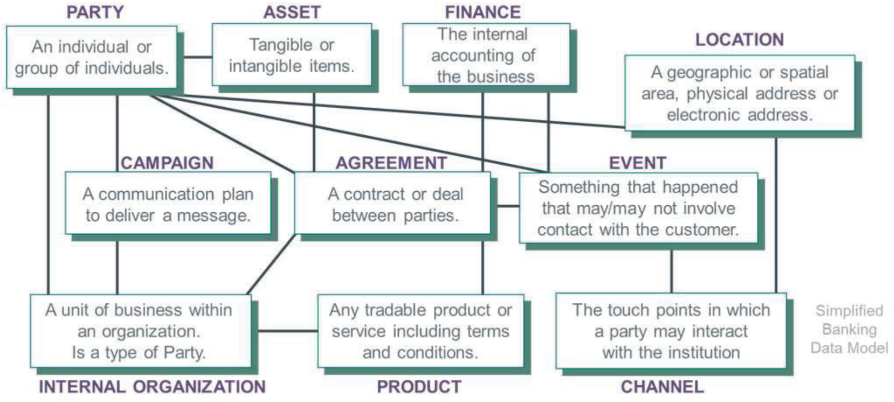
> **"Domain ownership distribution results in a *distributed data architecture*, where the data artifacts - datasets, code, metadata, and data policies - are maintained by their corresponding domains"**
> Zhamak Dehghani, Data Mesh – Delivering Data Driven Value At Scale, O'Reilly 6/22/21



STS

What is a Subject Area?

"Each subject area is a high-level classification of data representing a group of concepts pertaining to a major topic of interest to an organization. Subject areas can represent generic business concepts (customer, product, employee and finance), as well as be industry specific."
https://tdan.com/the-enterprise-data-model/5205



PARTY
An individual or group of individuals.

ASSET
Tangible or intangible items.

FINANCE
The internal accounting of the business

LOCATION
A geographic or spatial area, physical address or electronic address.

CAMPAIGN
A communication plan to deliver a message.

AGREEMENT
A contract or deal between parties.

EVENT
Something that happened that may/may not involve contact with the customer.

A unit of business within an organization. Is a type of Party.
INTERNAL ORGANIZATION

Any tradable product or service including terms and conditions.
PRODUCT

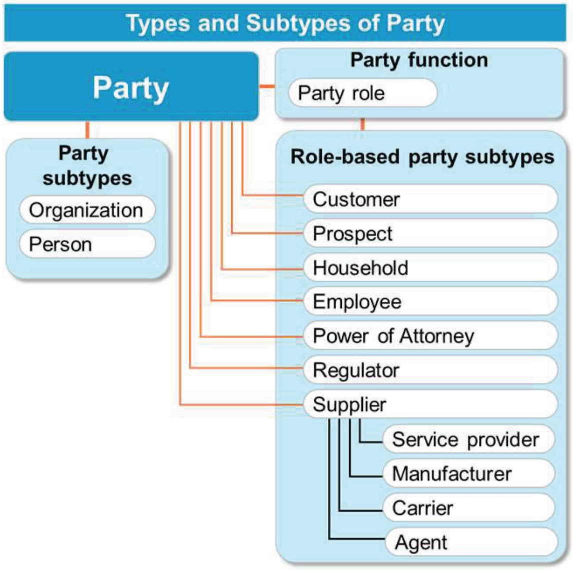
The touch points in which a party may interact with the institution
CHANNEL

Simplified Banking Data Model

13 Copyright © 2023. Stephen A. Brobst. Proprietary materials. Do not duplicate or distribute without written permission.

13

Subject Area: Types and Subtypes



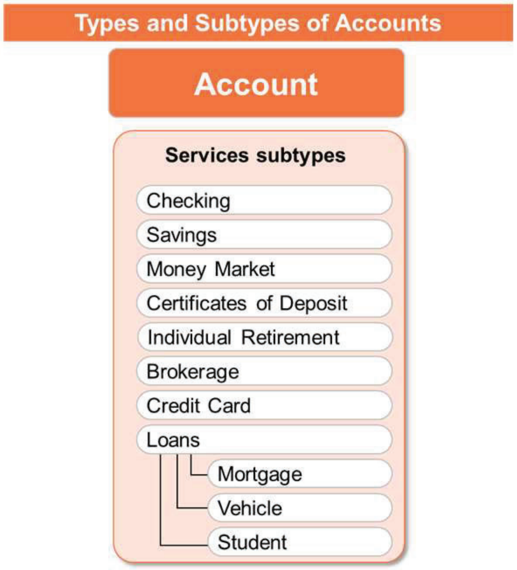
Types and Subtypes of Party

Party

Party function
- Party role

Party subtypes
- Organization
- Person

Role-based party subtypes
- Customer
- Prospect
- Household
- Employee
- Power of Attorney
- Regulator
- Supplier
  - Service provider
  - Manufacturer
  - Carrier
  - Agent



Types and Subtypes of Accounts

Account

Services subtypes
- Checking
- Savings
- Money Market
- Certificates of Deposit
- Individual Retirement
- Brokerage
- Credit Card
- Loans
  - Mortgage
  - Vehicle
  - Student

14 Copyright © 2023. Stephen A. Brobst. Proprietary materials. Do not duplicate or distribute without written permission.

14

What is a Business Process Domain?

Example banking domains:

- Credit Card
- Mortgage
- Direct Deposit (checking)
- Branch Operations
- Call Center Operations
- Finance & Accounting
- Treasury
- Marketing
- …

Domain definition:

- a field of action, thought, influence, etc.;
- a territory governed by a single ruler or government;
- a realm or range of personal knowledge, responsibility;

Dictionary.com

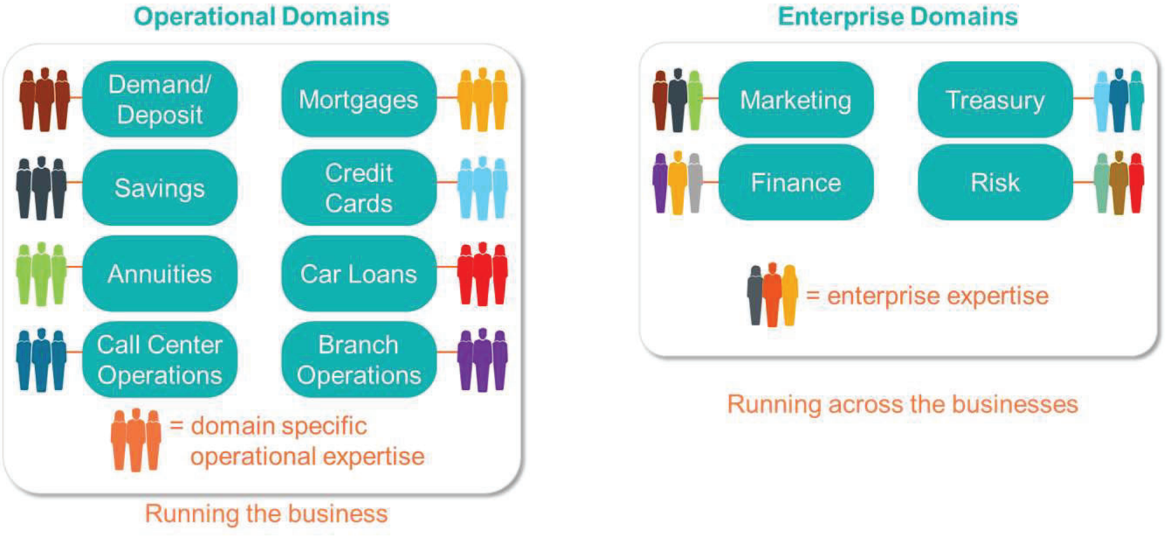
Domain definition:

A sphere of knowledge, influence, or activity. The subject area to which the user applies a program is the domain of the software.

Domain--Driven Design Reference, Eric Evans

Proprietary materials. Do not duplicate or distribute without written permission.

What is a Business Process Domain?



Operational Domains

Demand/ Deposit

Mortgages

Savings

Credit Cards

Annuities

Car Loans

Call Center Operations

Branch Operations

= domain specific operational expertise

Running the business

Enterprise Domains

Marketing

Treasury

Finance

Risk

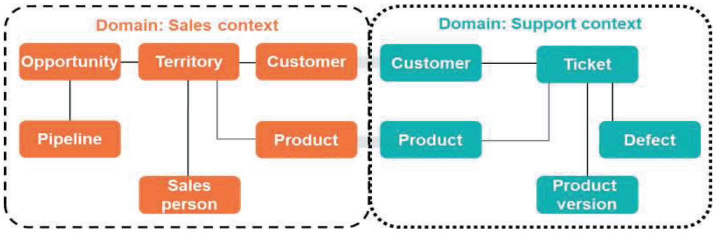
= enterprise expertise

Running across the businesses

Proprietary materials. Do not duplicate or distribute without written permission.

Bounded Context

- Domains have boundaries
  - Subject areas + workers + objectives
  - Analytical data & operational capabilities
- Entities in multiple domains
  - Different context → different attributes & treatments



Domain: Sales context

Opportunity

Territory

Customer

Pipeline

Product

Sales person

Domain: Support context

Customer

Ticket

Product

Defect

Product version

17

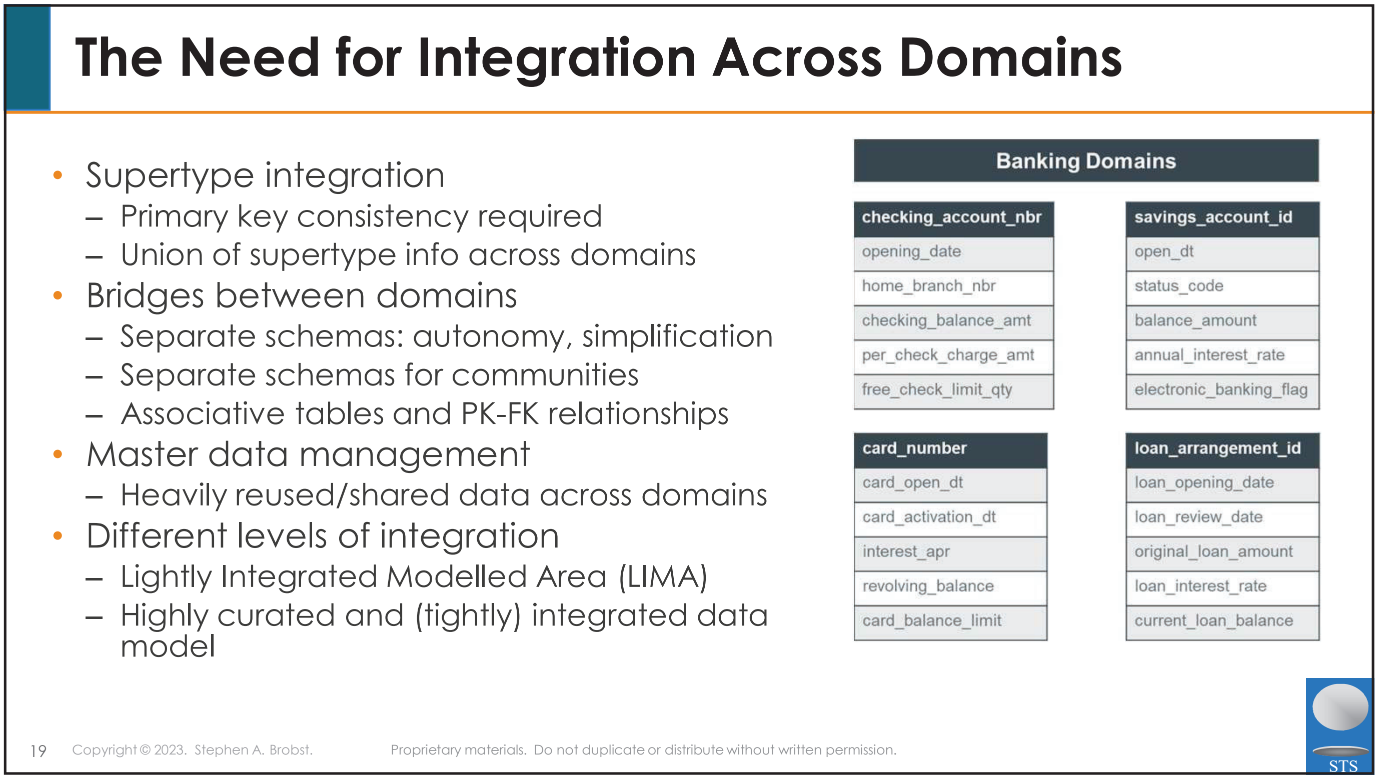
17

Data Sharing

> "As organizations are pulled into these ecosystems, sharing data becomes more important, and difficult, because they have to do it across trust boundaries. Even managing your own data is a challenge, and now you need solutions that go beyond the bounds of a particular organization."
>
> **Zhamak Dehghani**
> (Former) Director of Emerging Technologies, ThoughtWorks, North America
> https://www.thoughtworks.com/perspectives/edition15-data-strategies-article

18

18



The Need for Integration Across Domains

- Supertype integration
  - Primary key consistency required
  - Union of supertype info across domains
- Bridges between domains
  - Separate schemas: autonomy, simplification
  - Separate schemas for communities
  - Associative tables and PK-FK relationships
- Master data management
  - Heavily reused/shared data across domains
- Different levels of integration
  - Lightly Integrated Modelled Area (LIMA)
  - Highly curated and (tightly) integrated data model

Banking Domains

checking_account_nbr
opening_date
home_branch_nbr
checking_balance_amt
per_check_charge_amt
free_check_limit_qty

savings_account_id
open_dt
status_code
balance_amount
annual_interest_rate
electronic_banking_flag

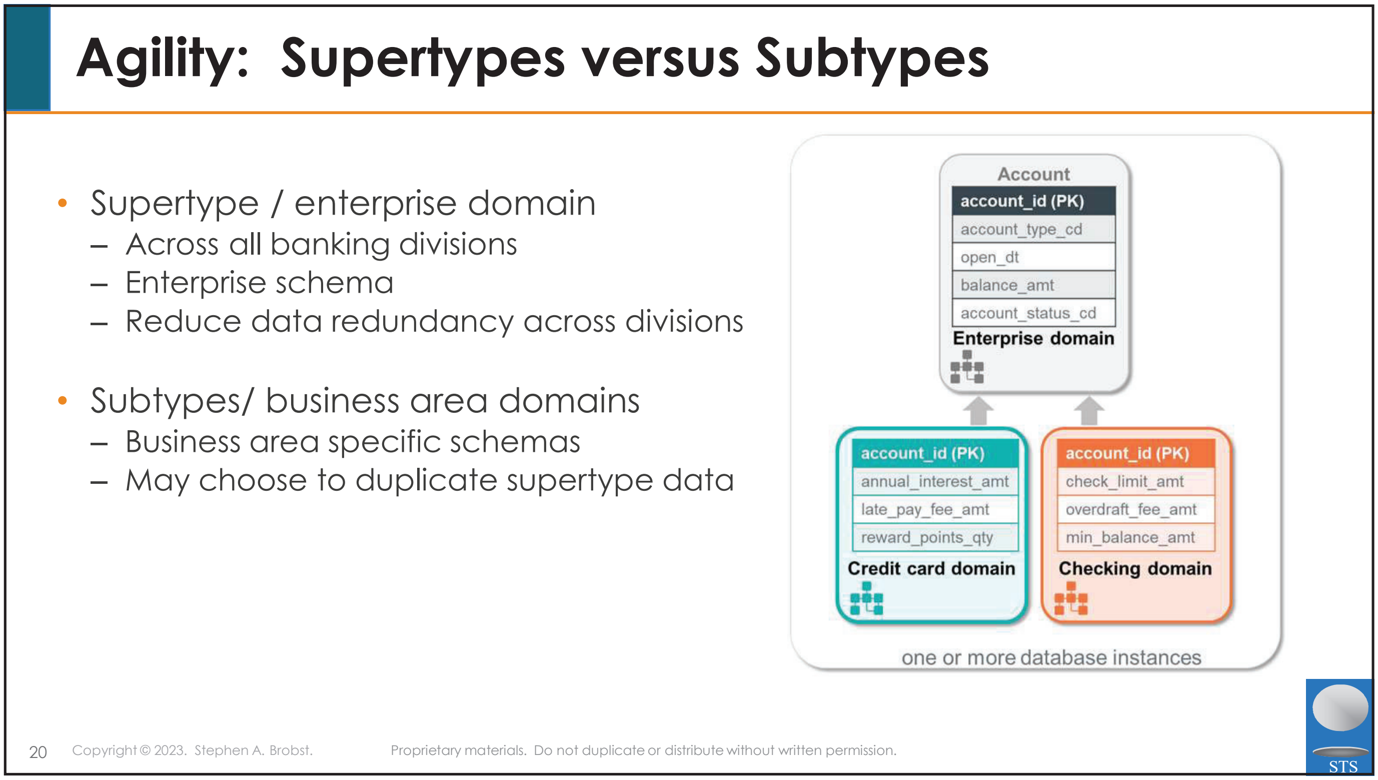
card_number
card_open_dt
card_activation_dt
interest_apr
revolving_balance
card_balance_limit

loan_arrangement_id
loan_opening_date
loan_review_date
original_loan_amount
loan_interest_rate
current_loan_balance

19

STS

19



Agility: Supertypes versus Subtypes

- Supertype / enterprise domain
  - Across all banking divisions
  - Enterprise schema
  - Reduce data redundancy across divisions
- Subtypes/ business area domains
  - Business area specific schemas
  - May choose to duplicate supertype data

Account

account_id (PK)
account_type_cd
open_dt
balance_amt
account_status_cd

Enterprise domain

account_id (PK)
annual_interest_amt
late_pay_fee_amt
reward_points_qty

Credit card domain

account_id (PK)
check_limit_amt
overdraft_fee_amt
min_balance_amt

Checking domain

one or more database instances

20

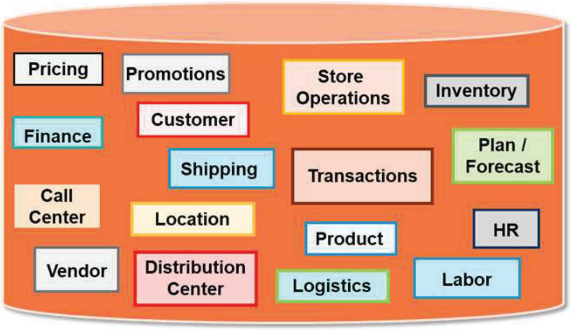
STS

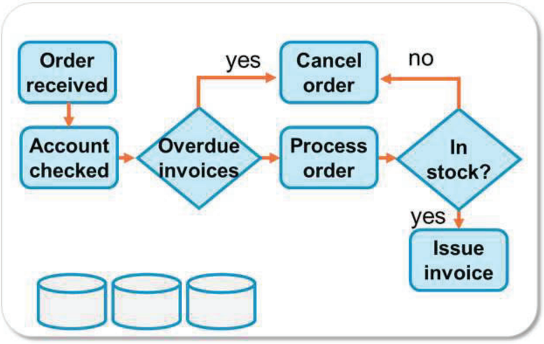
20

Subject Areas versus Business Process Domains

Subject Areas
Database concept
Spans multiple process domains
ELDM overwhelming to business
Data centric

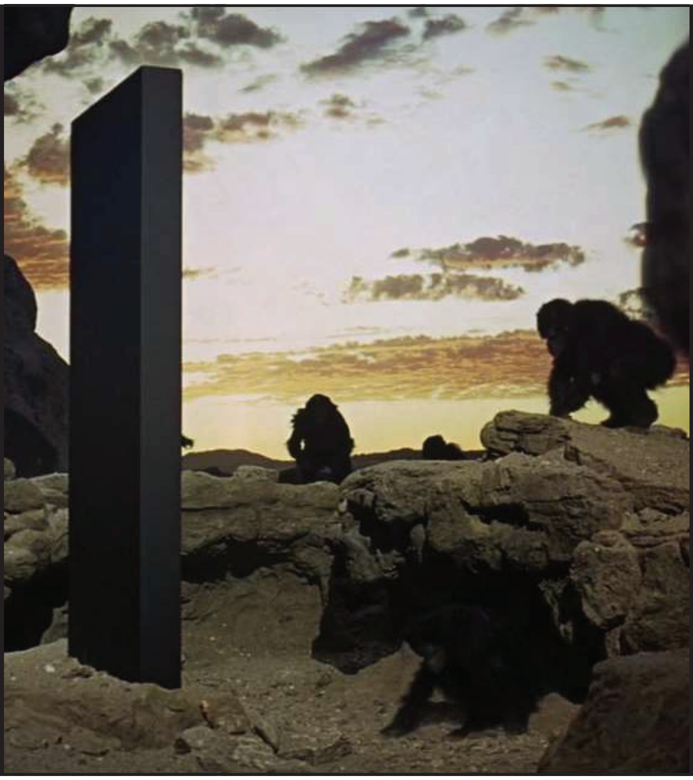
Business Process Domains
Business unit concept
Self-contained from business point-of-view
Small-to-medium size schemas
Business area centric



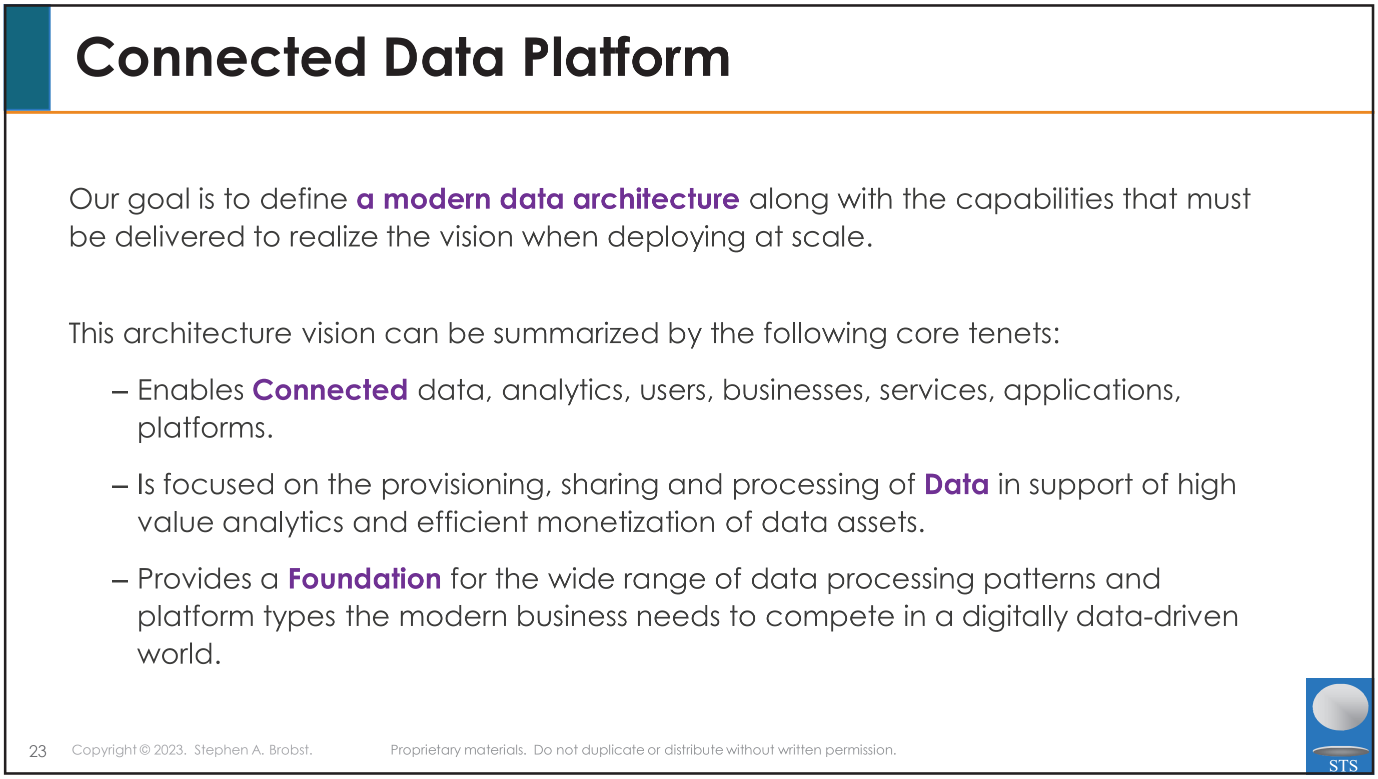


STS

Monolithic Data Warehousing Does Not Work



What can we do to be more agile?

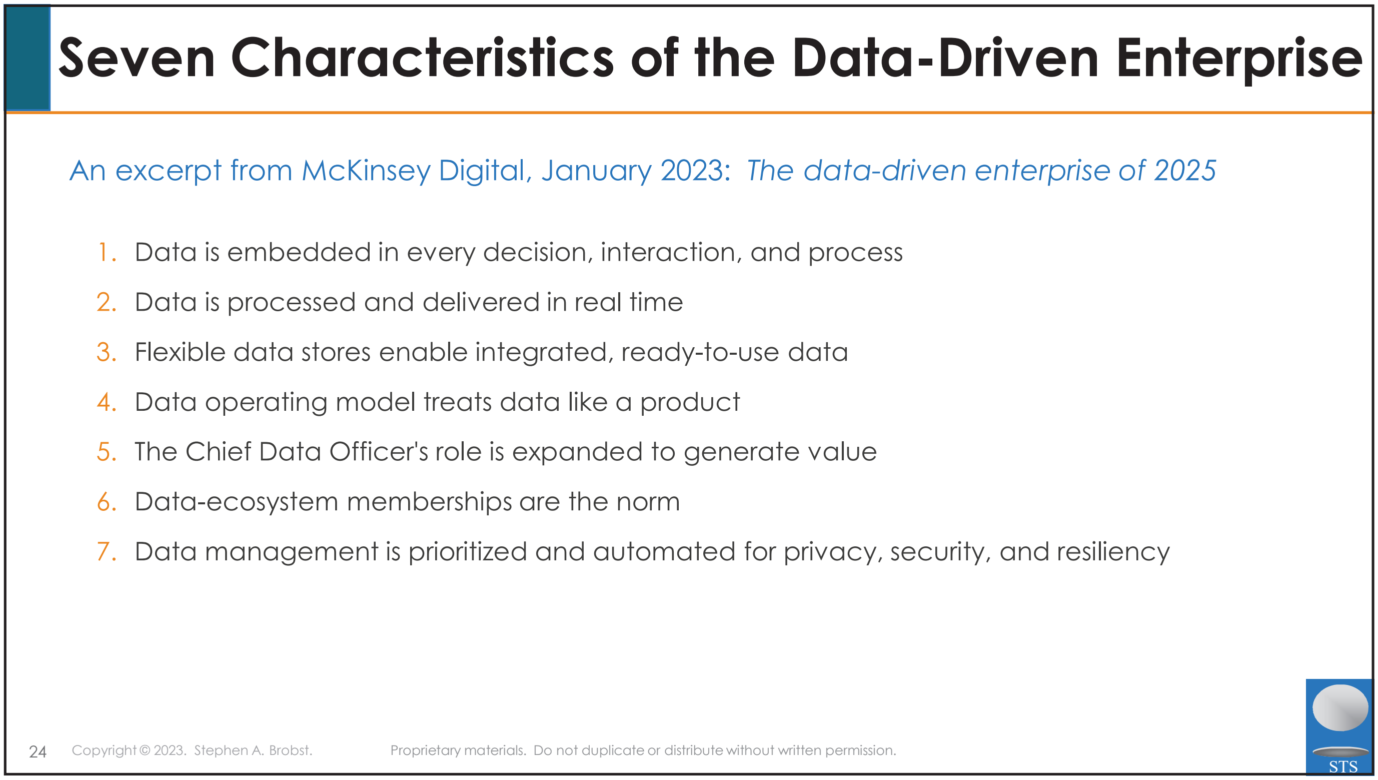


Connected Data Platform

Our goal is to define **a modern data architecture** along with the capabilities that must be delivered to realize the vision when deploying at scale.

This architecture vision can be summarized by the following core tenets:

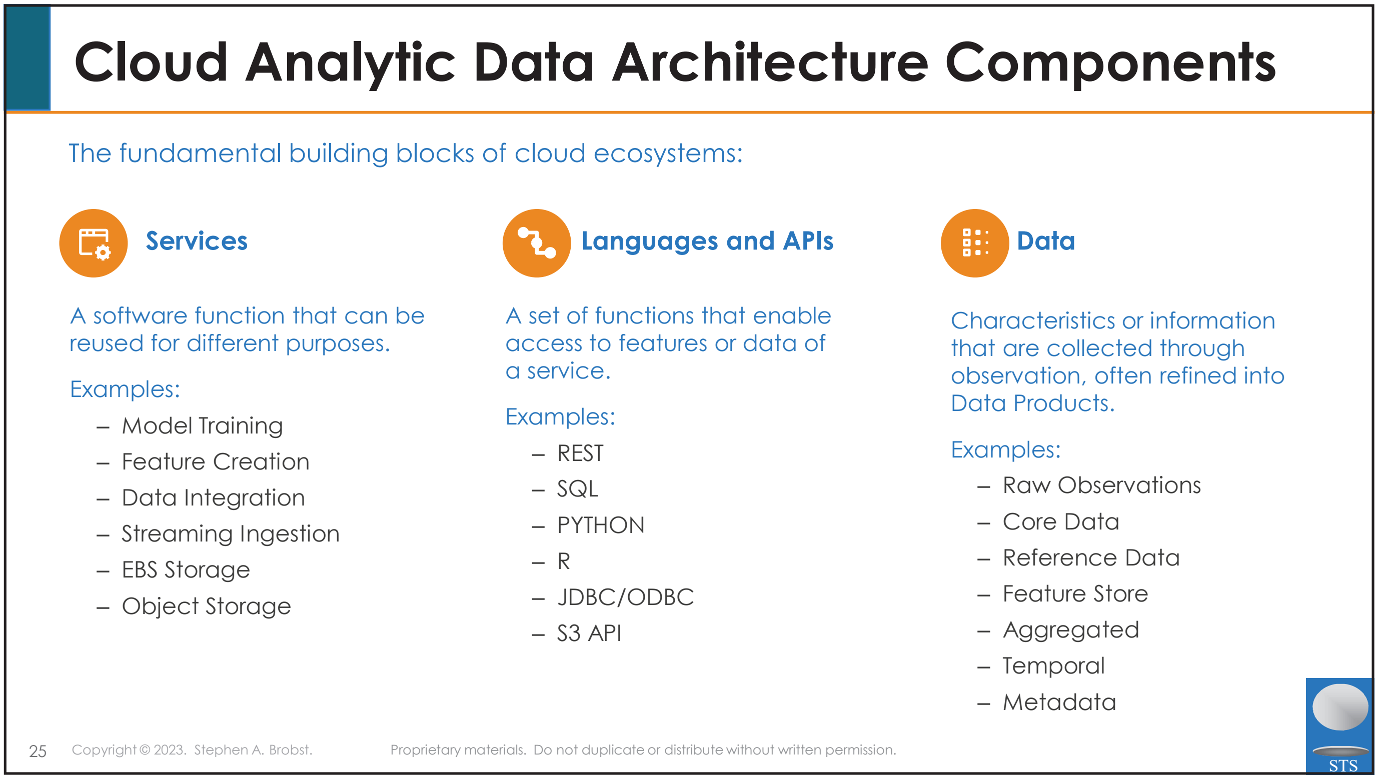
- Enables **Connected** data, analytics, users, businesses, services, applications, platforms.
- Is focused on the provisioning, sharing and processing of **Data** in support of high value analytics and efficient monetization of data assets.
- Provides a **Foundation** for the wide range of data processing patterns and platform types the modern business needs to compete in a digitally data-driven world.



Seven Characteristics of the Data-Driven Enterprise

An excerpt from McKinsey Digital, January 2023: *The data-driven enterprise of 2025*

1. Data is embedded in every decision, interaction, and process
2. Data is processed and delivered in real time
3. Flexible data stores enable integrated, ready-to-use data
4. Data operating model treats data like a product
5. The Chief Data Officer's role is expanded to generate value
6. Data-ecosystem memberships are the norm
7. Data management is prioritized and automated for privacy, security, and resiliency



Cloud Analytic Data Architecture Components

The fundamental building blocks of cloud ecosystems:

Services

A software function that can be reused for different purposes.

Examples:

- Model Training
- Feature Creation
- Data Integration
- Streaming Ingestion
- EBS Storage
- Object Storage

Languages and APIs

A set of functions that enable access to features or data of a service.

Examples:

- REST
- SQL
- PYTHON
- R
- JDBC/ODBC
- S3 API

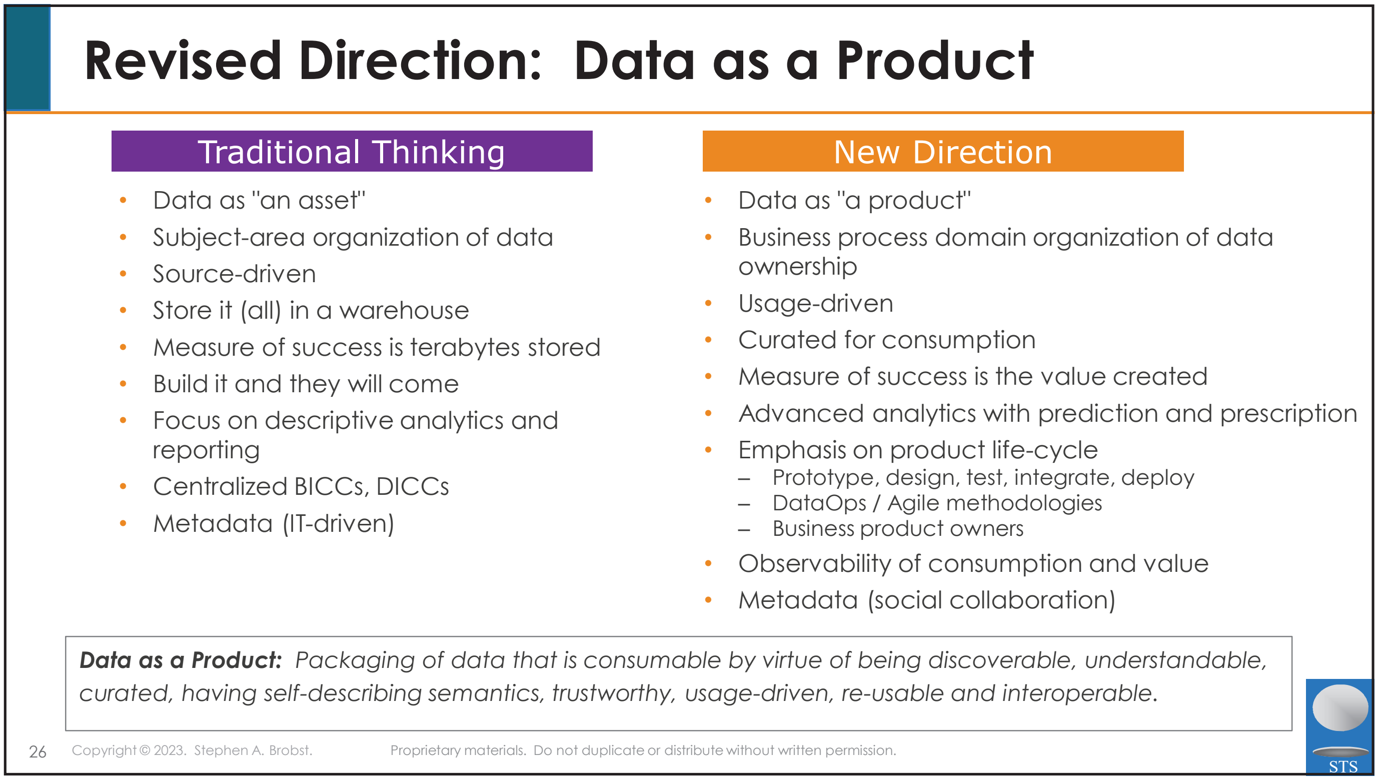
Data

Characteristics or information that are collected through observation, often refined into Data Products.

Examples:

- Raw Observations
- Core Data
- Reference Data
- Feature Store
- Aggregated
- Temporal
- Metadata

25



Revised Direction: Data as a Product

Traditional Thinking

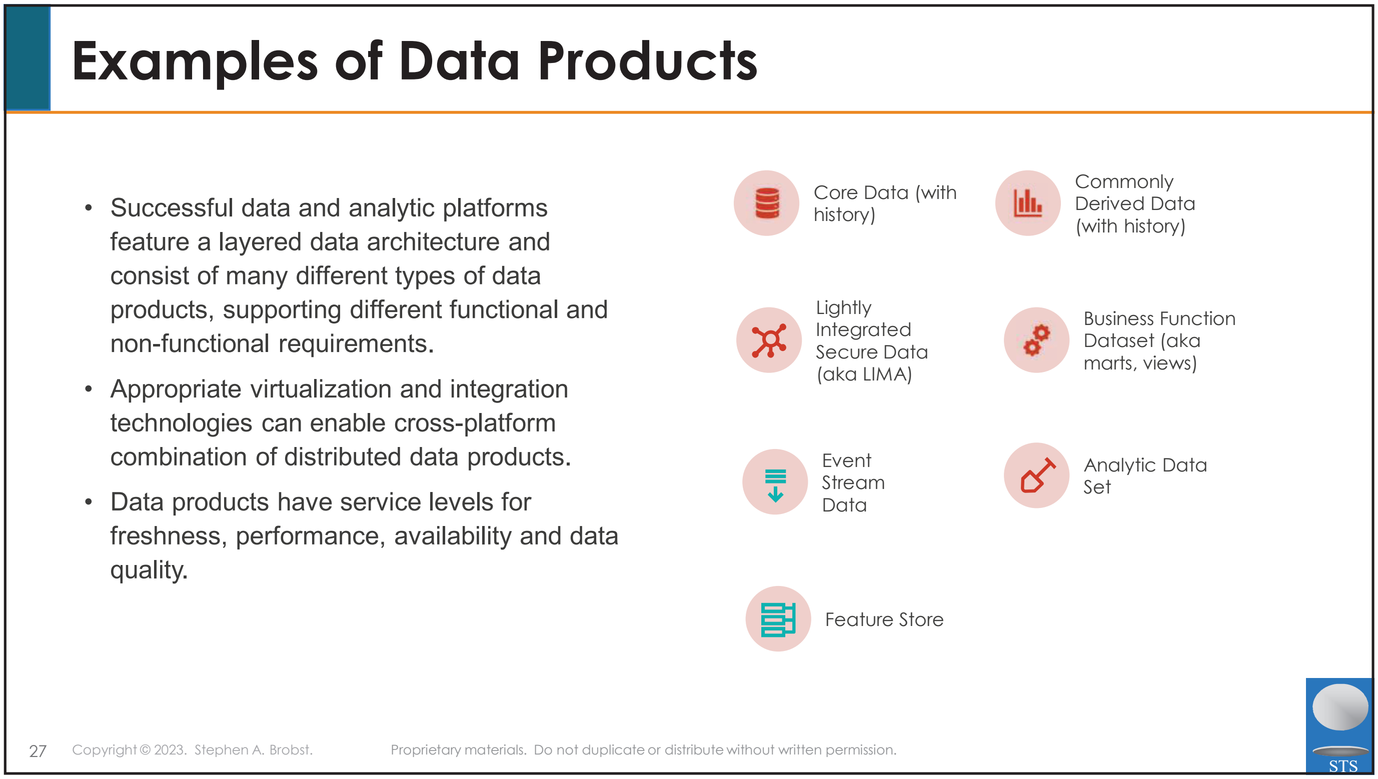
- Data as "an asset"
- Subject-area organization of data
- Source-driven
- Store it (all) in a warehouse
- Measure of success is terabytes stored
- Build it and they will come
- Focus on descriptive analytics and reporting
- Centralized BICCs, DICCs
- Metadata (IT-driven)

New Direction

- Data as "a product"
- Business process domain organization of data ownership
- Usage-driven
- Curated for consumption
- Measure of success is the value created
- Advanced analytics with prediction and prescription
- Emphasis on product life-cycle
  - Prototype, design, test, integrate, deploy
  - DataOps / Agile methodologies
  - Business product owners
- Observability of consumption and value
- Metadata (social collaboration)

Data as a Product: *Packaging of data that is consumable by virtue of being discoverable, understandable, curated, having self-describing semantics, trustworthy, usage-driven, re-usable and interoperable.*

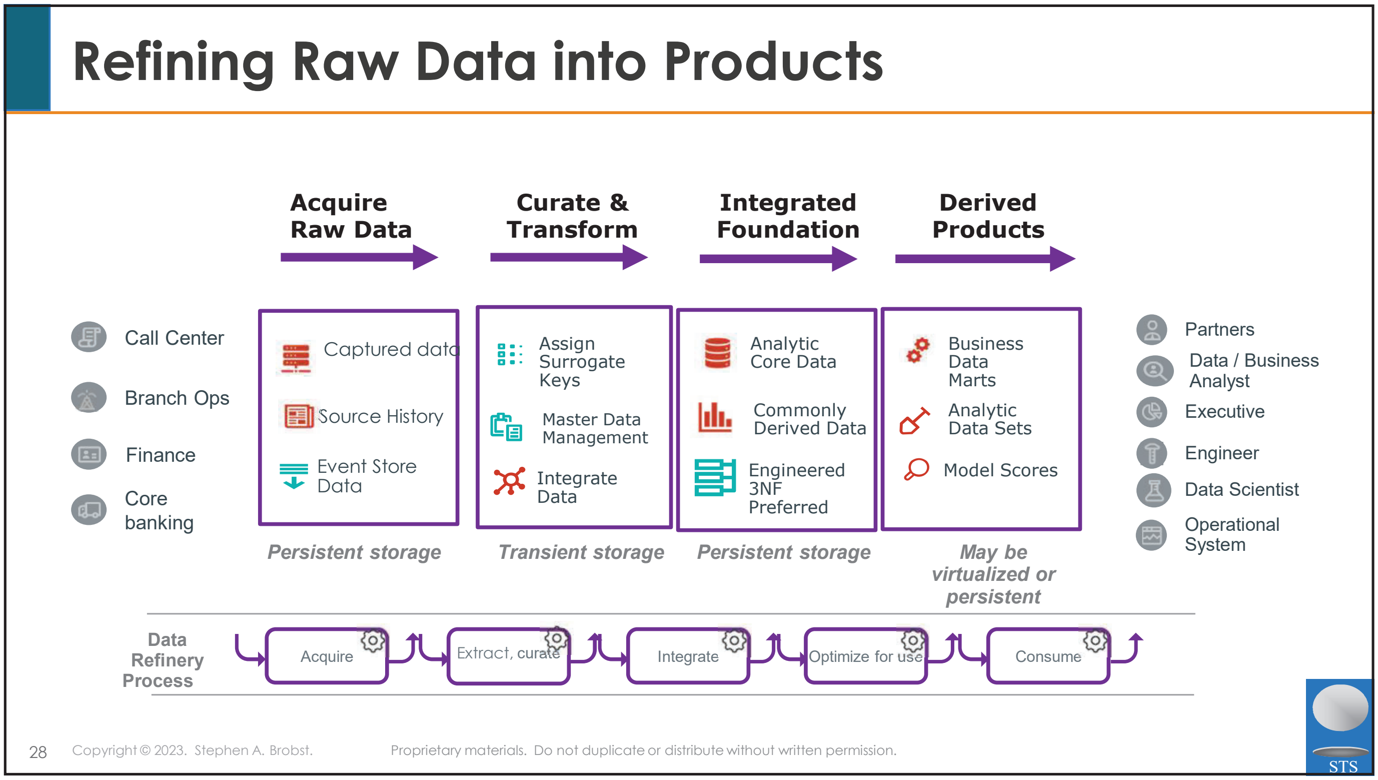
26



Examples of Data Products

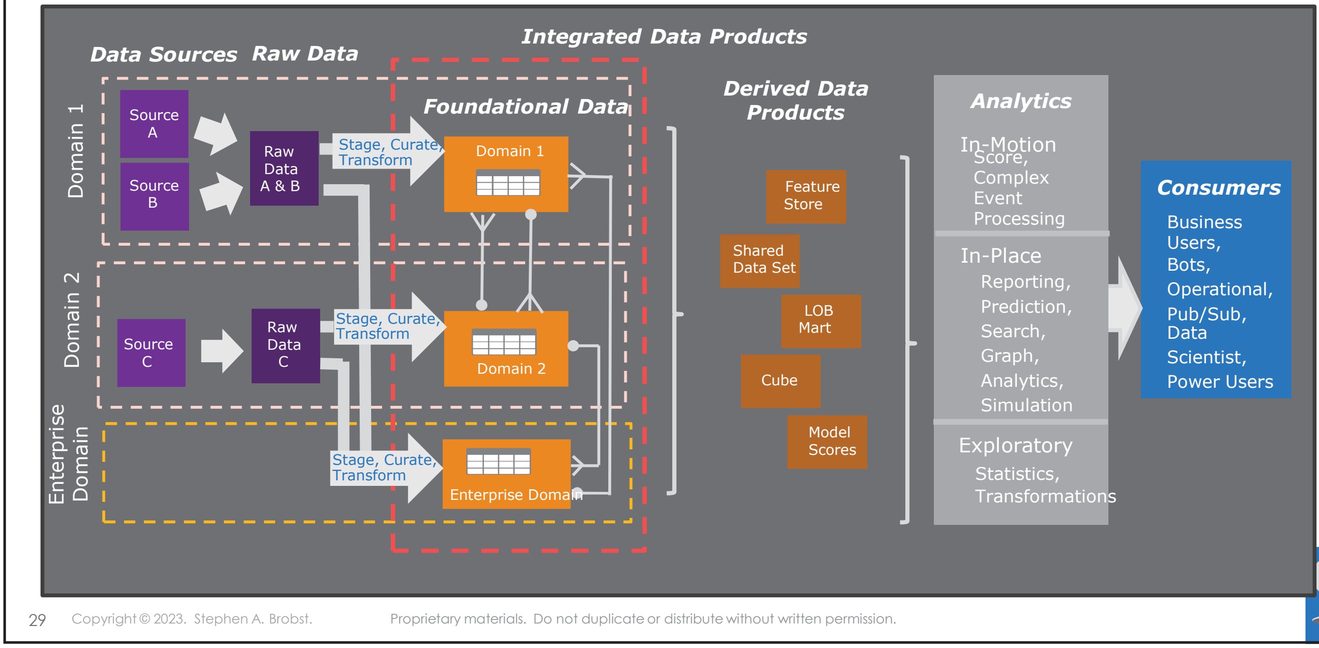
- Successful data and analytic platforms feature a layered data architecture and consist of many different types of data products, supporting different functional and non-functional requirements.
- Appropriate virtualization and integration technologies can enable cross-platform combination of distributed data products.
- Data products have service levels for freshness, performance, availability and data quality.

Core Data (with history)
Commonly Derived Data (with history)
Lightly Integrated Secure Data (aka LIMA)
Business Function Dataset (aka marts, views)
Event Stream Data
Analytic Data Set
Feature Store



Refining Raw Data into Products

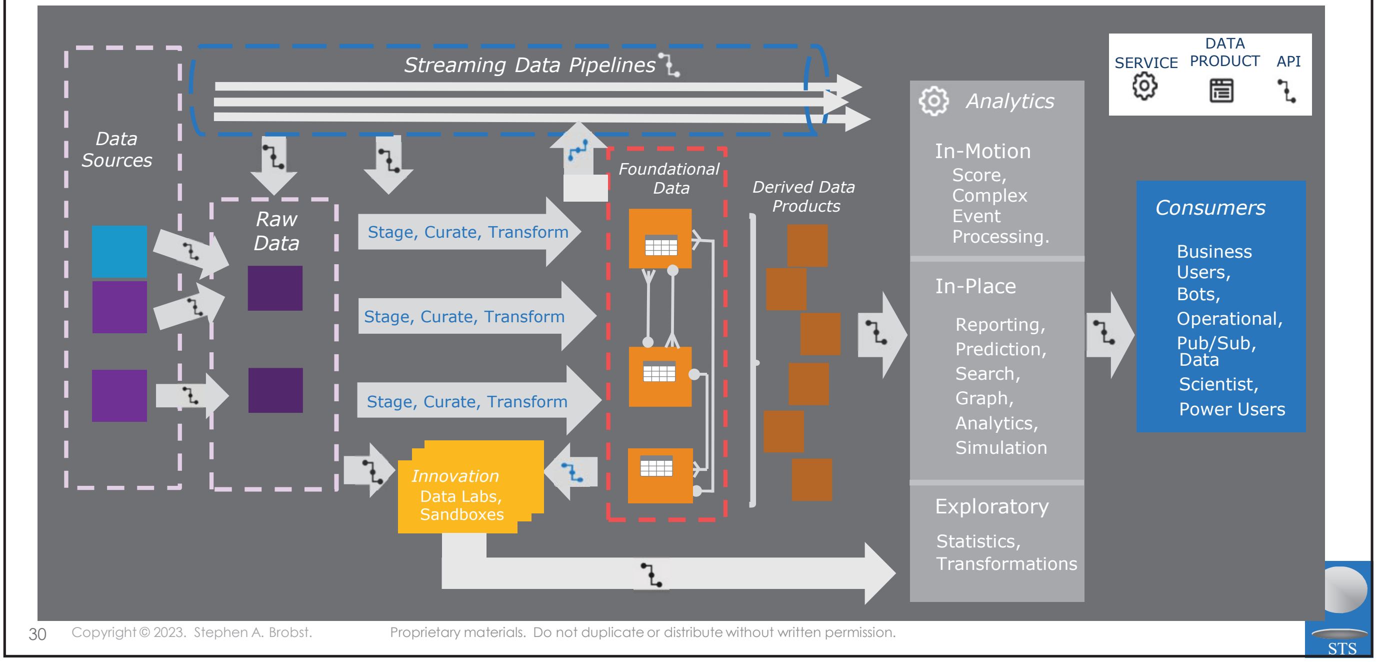
Connected Data Architecture: Distinct Domains with Referential Integrity



29 Proprietary materials. Do not duplicate or distribute without written permission.

STS

Connected Data Architecture: Extended with Innovation Data Labs

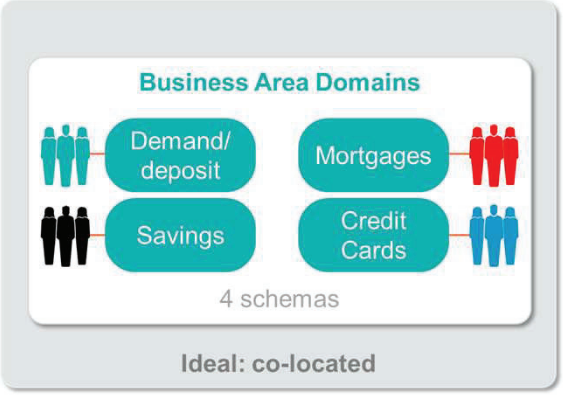


30 Proprietary materials. Do not duplicate or distribute without written permission.

STS

Implementation: Align Domains to Schemas

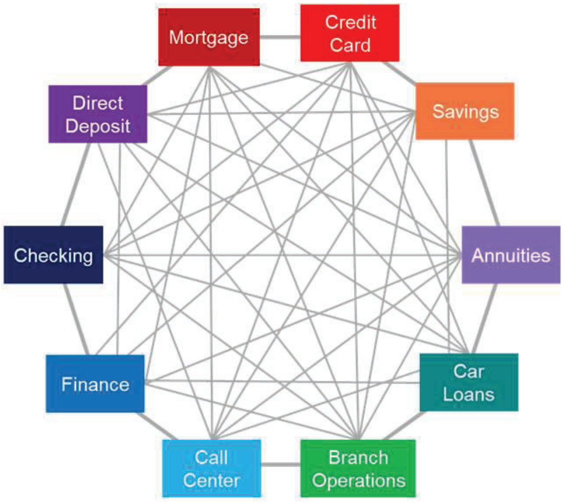
- Subject matter business experts
  - Detailed data model design
  - Populate models with data
  - Data quality ownership
  - Stewards of processes and data
  - Commonly, schema per domain
- Agility
  - Domains = process and data isolation
  - Loosely-coupled teams
  - Less friction, faster time-to-action



STS

Connected Schemas

- Individual schemas = linear value
- For each connected schema, value multiplies
  - Exponential value
  - PK-FK relationships required

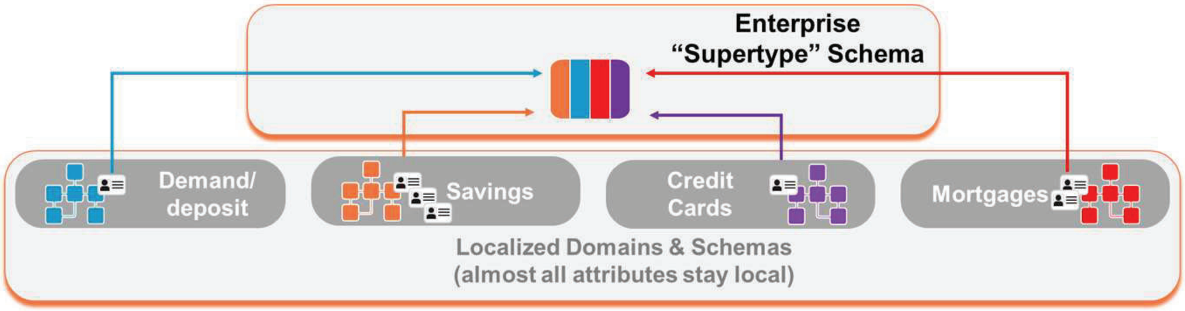


N * (N-1) / 2 connections

STS

Enterprise Schemas Span Domains

- Data integrated across domains
  - All account relationships
  - Supertype information beyond PKs/FKs
  - Bringing data together across domains
- Complete view of customer
- Consolidated view of risk



STS

Enterprise Governance is Still Desirable

Naming standards: Consistent naming, abbreviations, classification extensions, etc. increase usability for all data products.

Data typing: Well-thought and consistent data typing increases usability and performance for all data products.

Key construction: Domain neutral methods for surrogate key construction avoids PII and performance issues associated with the use of natural keys.

Quality metrics: Avoid re-inventing best practices for defining metrics and implementation of a continuous improvement program for data quality.

Access control: Leverage enterprise standards for global access control to data products.

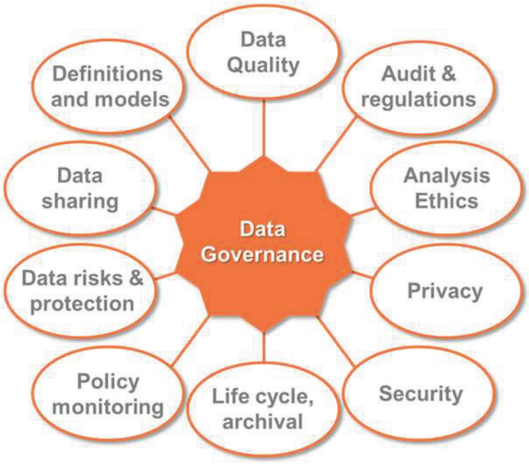
Technology choices: Encourage fit-for-purpose deployment without technology proliferation.

Autonomy ≠ Anarchy

STS

Agility: Light Governance and Standards

- Enterprise governance & standards
  - Light governance from a centralized team
  - Frameworks and blueprints, not detailed designs
- Critical for domain agility & independence
  - Agreed upon standards allows de-coupling of teams to execute autonomously
  - Detailed design owned by subject matter experts at a domain level
- Consolidate security across domains
  - Global access control = trust, simplicity



35

Three Types of Deployment

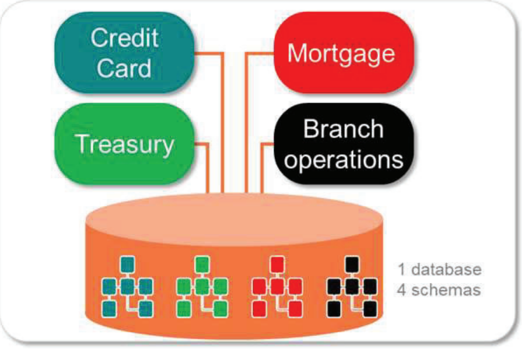
1. Connected and Co-located
2. Connected and Distributed
3. Isolated

Note that there is typically one schema per domain.

36

Connected and Co-Located Schemas

Every schema does not need a separate database instance!



Multiple schemas in one scalable database instance almost always outperforms a distributed database:

- Less data movement for joins across schemas
- Opportunities for optimizing access paths
- Opportunities for overall execution benefits

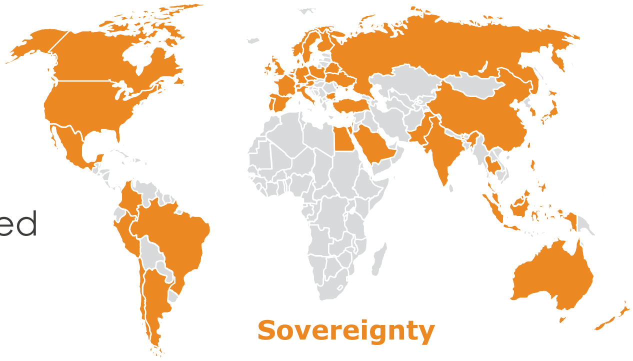
Lower TCO than proliferating database instances

37

STS

Connected and Distributed Schemas: Real-World Constraints

- Multiple clouds or multiple availability zones within a single cloud
- Primary to foreign key “connectedness” together with QueryGrid allows distributed schemas to be brought together across multiple geographies.

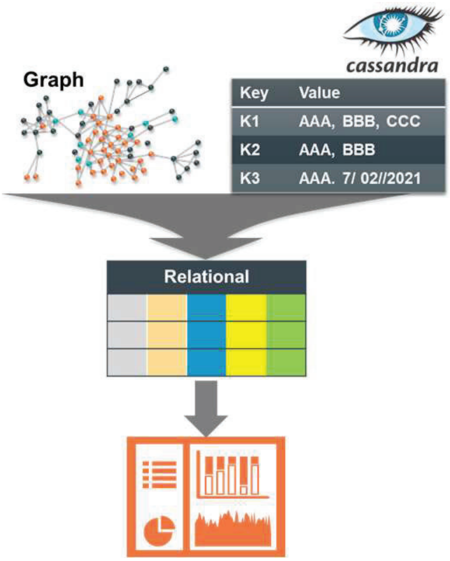


38

STS

Connected and Distributed Schemas
Fit for Purpose Tooling

- Heterogeneous data management platforms:
  - e.g., TigerGraph for product hierarchy, Cassandra for semi-structured weblog data and Teradata for relational data.
  - Data lakes versus data warehouses versus data marts.
- Primary to foreign key "connectedness" together with some form of Query Fabric allows distributed schemas to be brought together across multiple technologies on-demand.



Key	Value
K1	AAA, BBB, CCC
K2	AAA, BBB
K3	AAA. 7/ 02//2021

STS

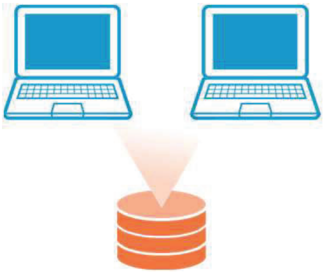
Isolated Schemas

Autonomous teams work independently on each schema:

- Self-contained governance.
- Localized standards, unique tooling.
- No PK-FK referential integrity across schemas.
- "Islands" of data with no re-use and no connectivity to other schemas.

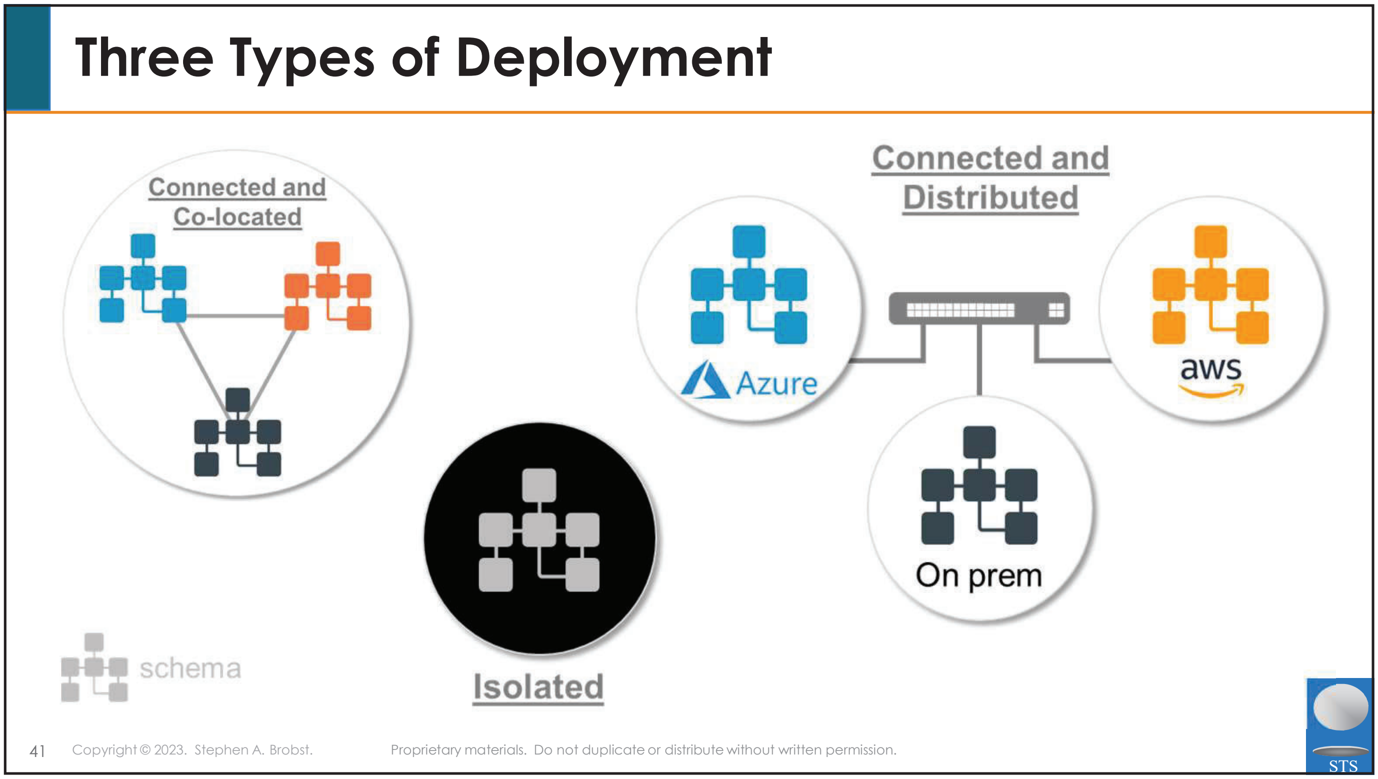
Advantages of isolated schemas:

- Detached requirements leads to better agility.
- Control (often masked as security).



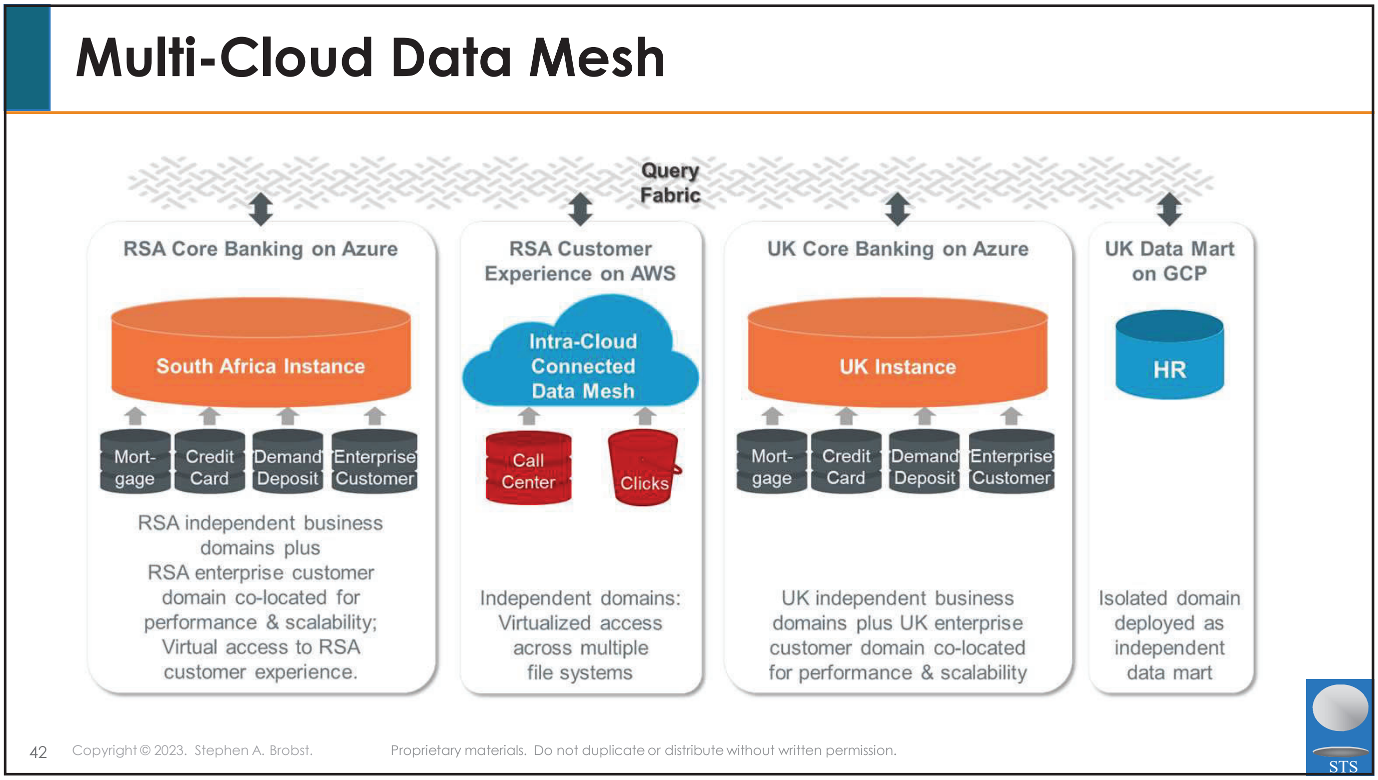
Employee Laptop XLS Data Sets
Human Capital Data Marts
Limited Scope Marketing Campaigns

STS



Three Types of Deployment

41 Copyright © 2023. Stephen A. Brobst. Proprietary materials. Do not duplicate or distribute without written permission.



Multi-Cloud Data Mesh

RSA independent business domains plus RSA enterprise customer domain co-located for performance & scalability; Virtual access to RSA customer experience.

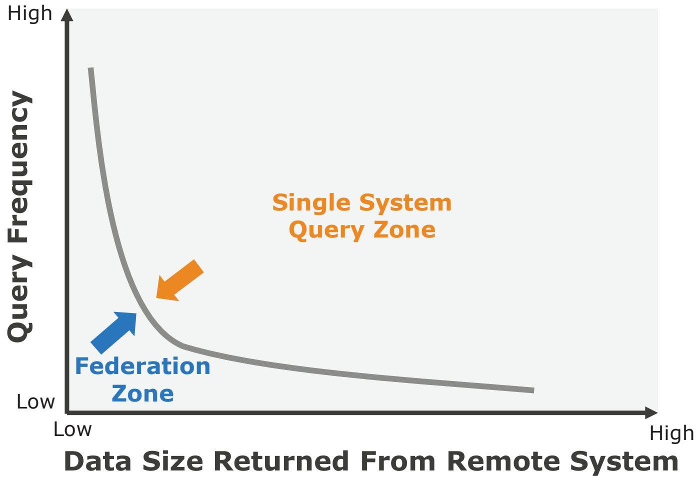
Independent domains: Virtualized access across multiple file systems

UK independent business domains plus UK enterprise customer domain co-located for performance & scalability

Isolated domain deployed as independent data mart

42 Copyright © 2023. Stephen A. Brobst. Proprietary materials. Do not duplicate or distribute without written permission.

Virtualization is NOT a Panacea



Considerations:

- Data requires specialized processing by another analytic engine.
- Query requires data that has not been integrated or requires schema-on-read access.
- Data location restrictions or sharing concerns; government, department, or BU.

- Network bandwidth/ latency concerns.
- Strict query performance SLAs or sub-second response-time.
- Source data requires extensive transformations before use.

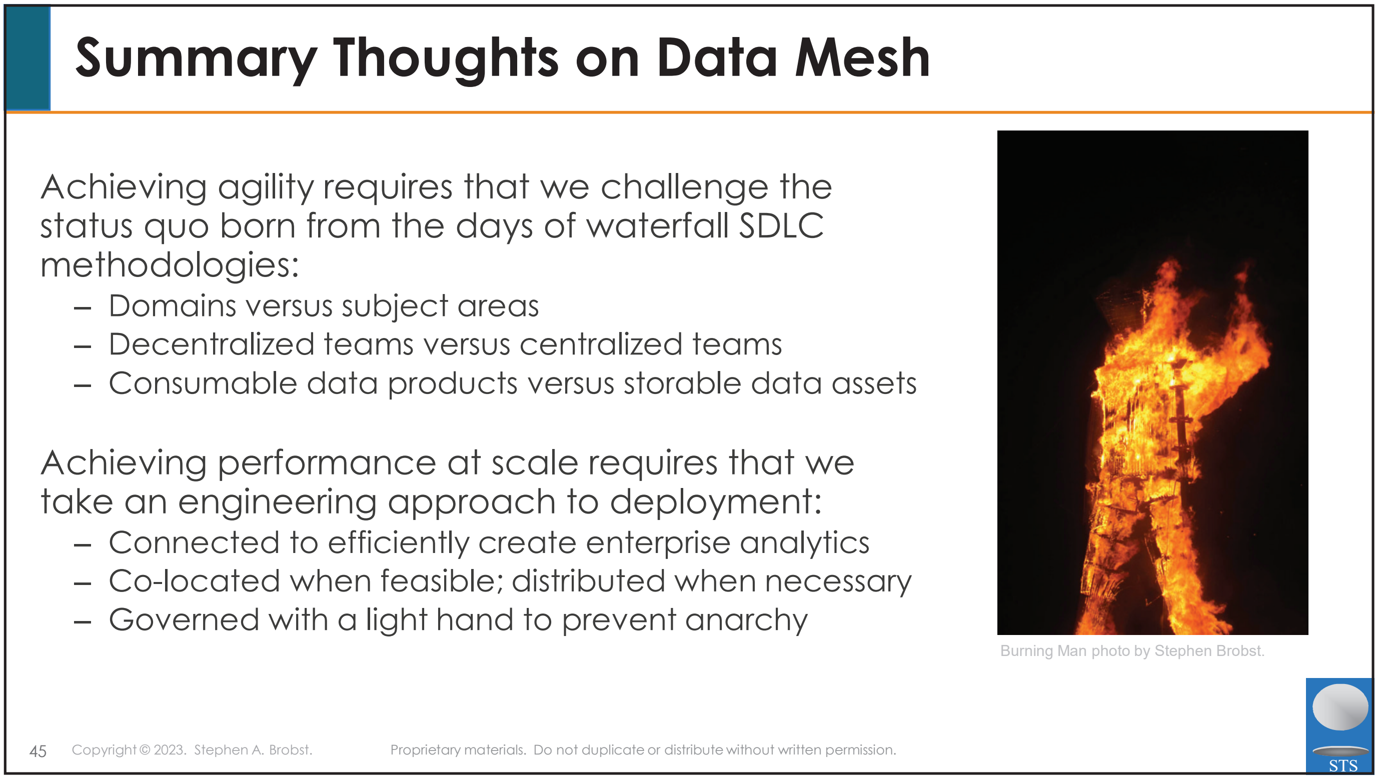
STS

Data Mesh and the Threads that Hold it Together

- Aligning to funded business initiatives is (still) the most critical success factor.
- Data must be semantically linked (connected) across domains to maximize value.
- Specialized expertise needed:
  - Successfully leverage data management tools
  - Effectively implement cross-domain data governance
- High-volume, complex, frequently joined data is best co-located whenever feasible.

Kevin Lewis
Blog: Data Mesh and the Threads that Hold it Together

STS



Summary Thoughts on Data Mesh

Achieving agility requires that we challenge the status quo born from the days of waterfall SDLC methodologies:

- Domains versus subject areas
- Decentralized teams versus centralized teams
- Consumable data products versus storable data assets

Achieving performance at scale requires that we take an engineering approach to deployment:

- Connected to efficiently create enterprise analytics
- Co-located when feasible; distributed when necessary
- Governed with a light hand to prevent anarchy

Burning Man photo by Stephen Brobst.

45

Thank you!

Stephen Brobst
sbrobst@alum.mit.edu

46