Using a Data Lake to Accelerating ETL Development

Part 2 of the Modern Data Series

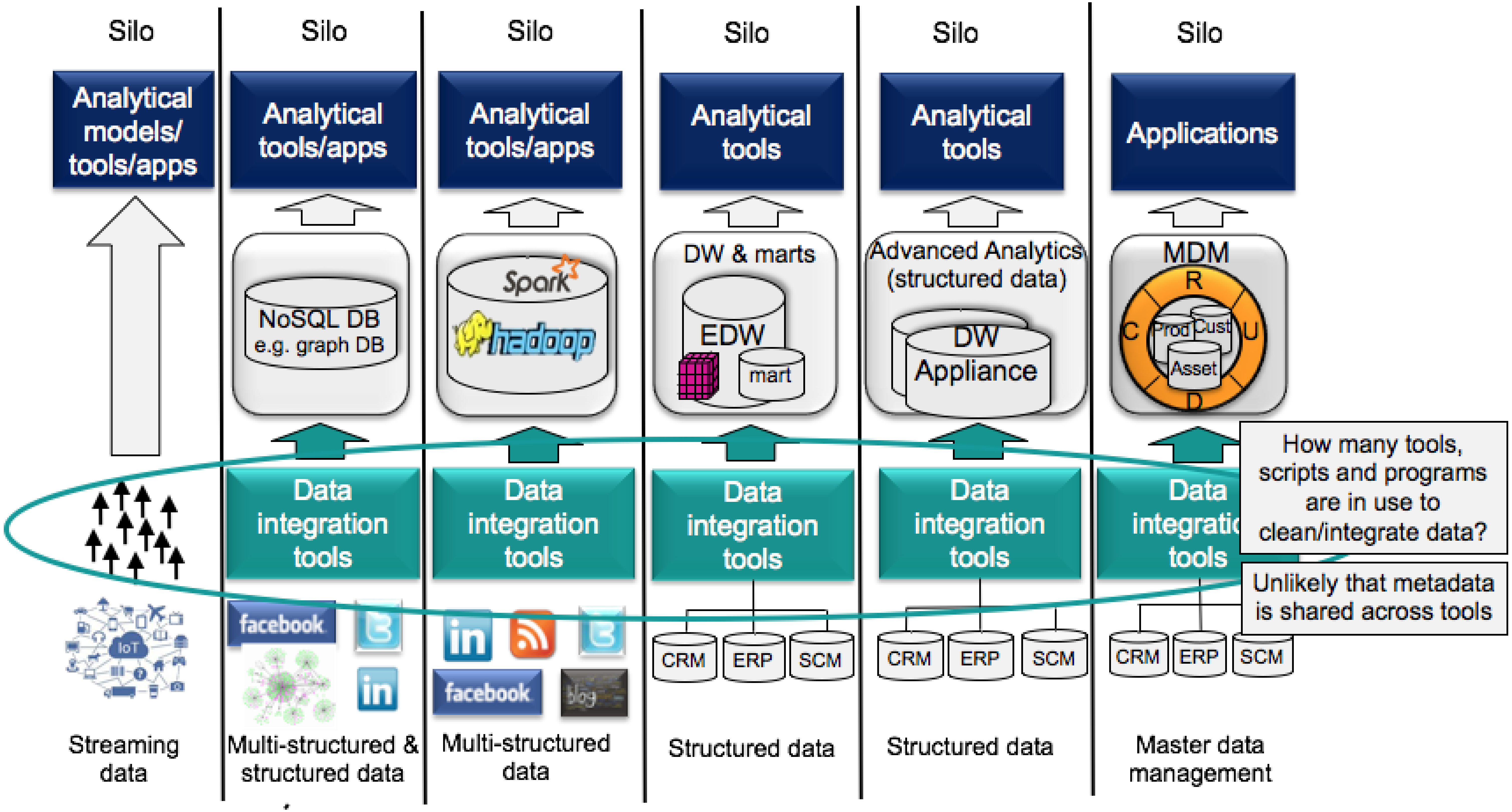
Mike Ferguson

quest for knowledge®

q4k.com

Many companies today have a siloed approach to data and analytics, with many tools, scripts and code in use to clean, transform and integrate data

—



Silo

Analytical models/ tools/apps

Streaming data

Silo

Analytical tools/apps

NoSQL DB
e.g. graph DB

Data integration tools

Multi-structured & structured data

Silo

Analytical tools/apps

Spark

hadoop

Data integration tools

Multi-structured data

Silo

Analytical tools

DW & marts

EDW

mart

Data integration tools

CRM ERP SCM

Structured data

Silo

Analytical tools

Advanced Analytics (structured data)

DW Appliance

Data integration tools

CRM ERP SCM

Structured data

Silo

Applications

MDM

R C U D

Prod Cust Asset

Data integration tools

CRM ERP SCM

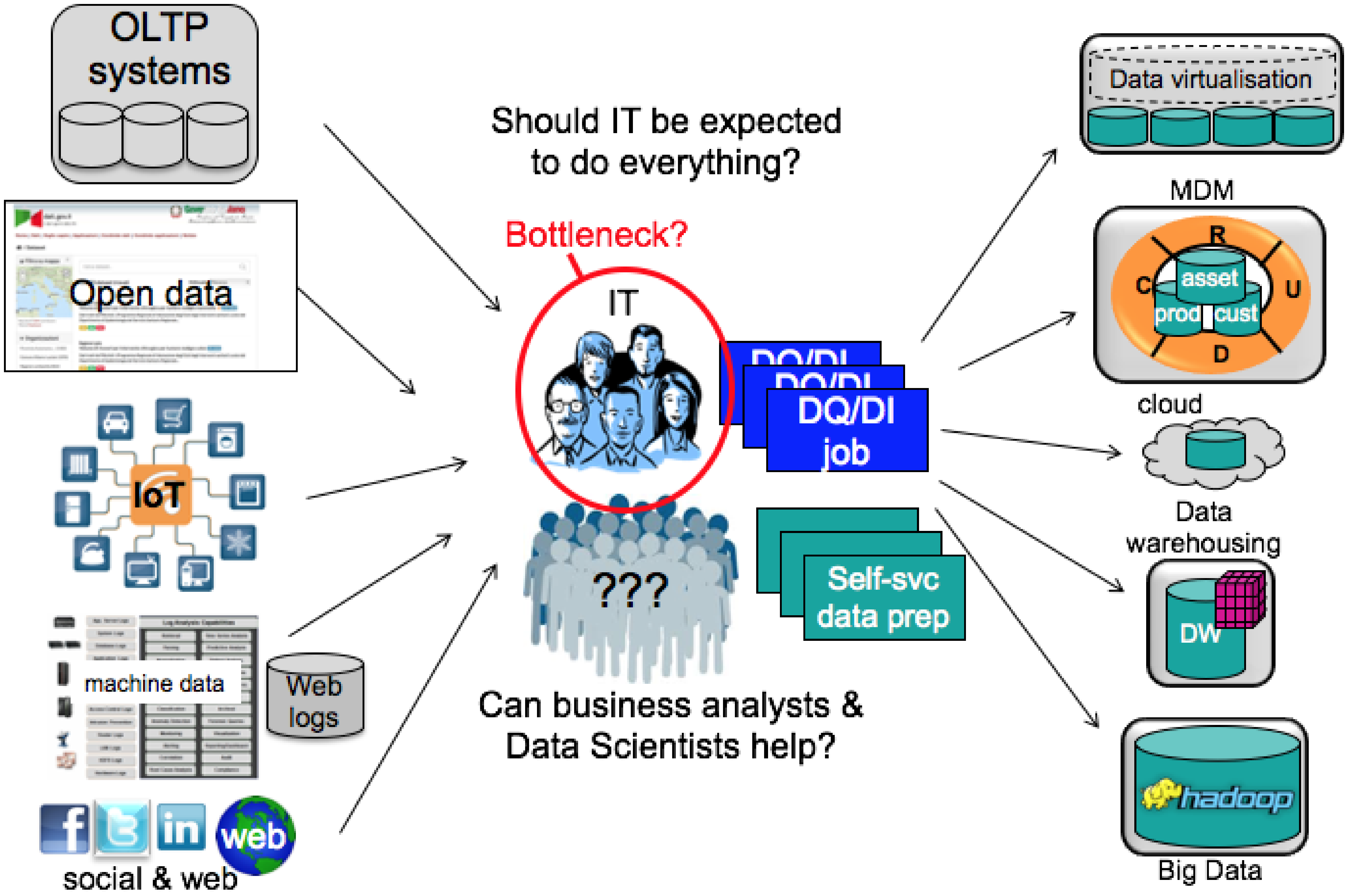
Master data management

How many tools, scripts and programs are in use to clean/integrate data?

Unlikely that metadata is shared across tools

But with 000's of data sources that business want to analyse, IT will likely become a bottleneck unless they can work with business to integrate data

—



OLTP systems

Open data

IoT

machine data

Web logs

social & web

Should IT be expected to do everything?

Bottleneck?

IT

DQ/DI

DQ/DI

DQ/DI job

???

Self-svc

Self-svc data prep

Can business analysts & Data Scientists help?

Data virtualisation

MDM

R

C

U

D

asset

prod

cust

cloud

Data warehousing

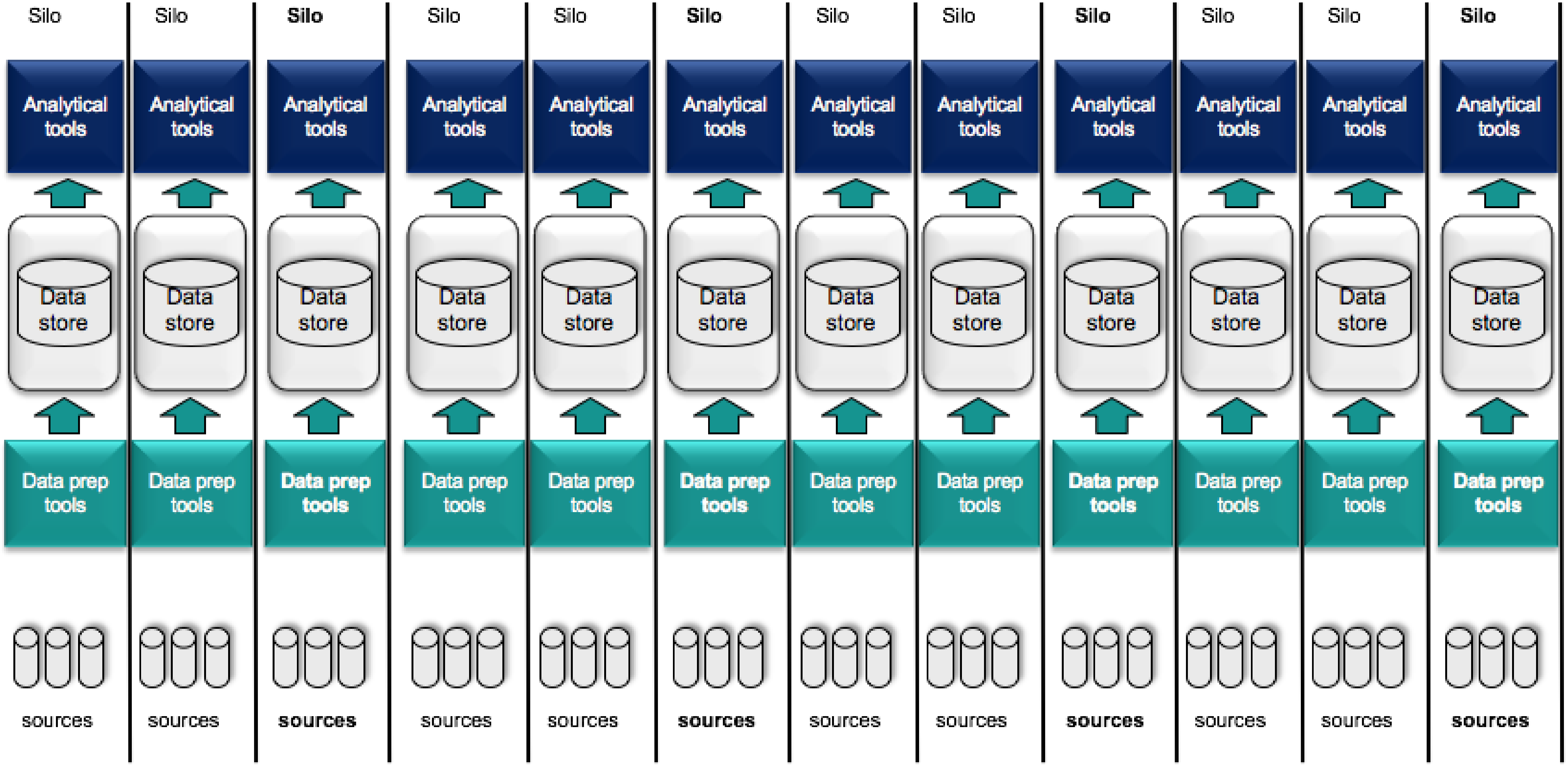
DW

hadoop

Big Data

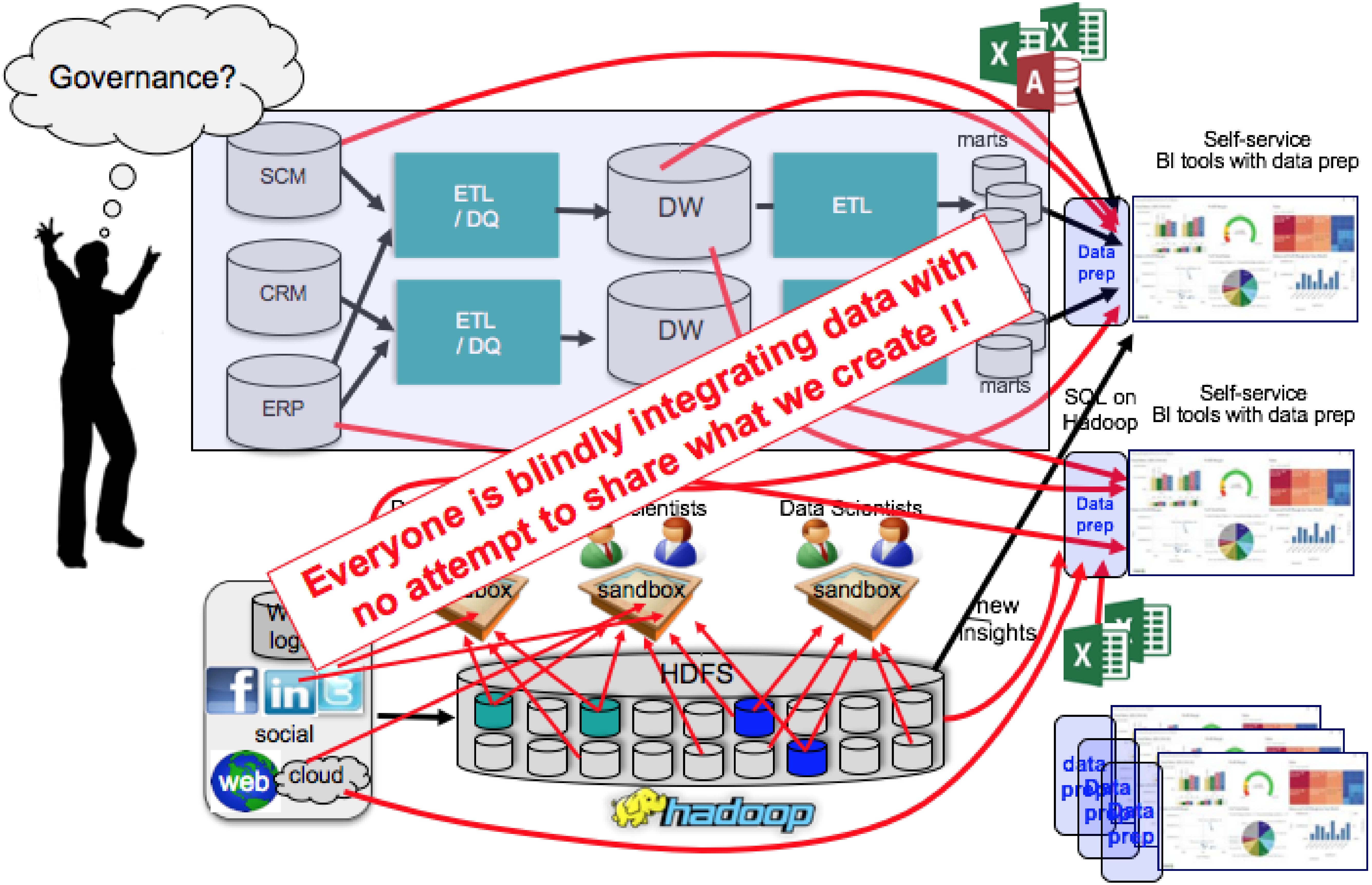
The danger of self-service data preparation

An explosions of personal silos!



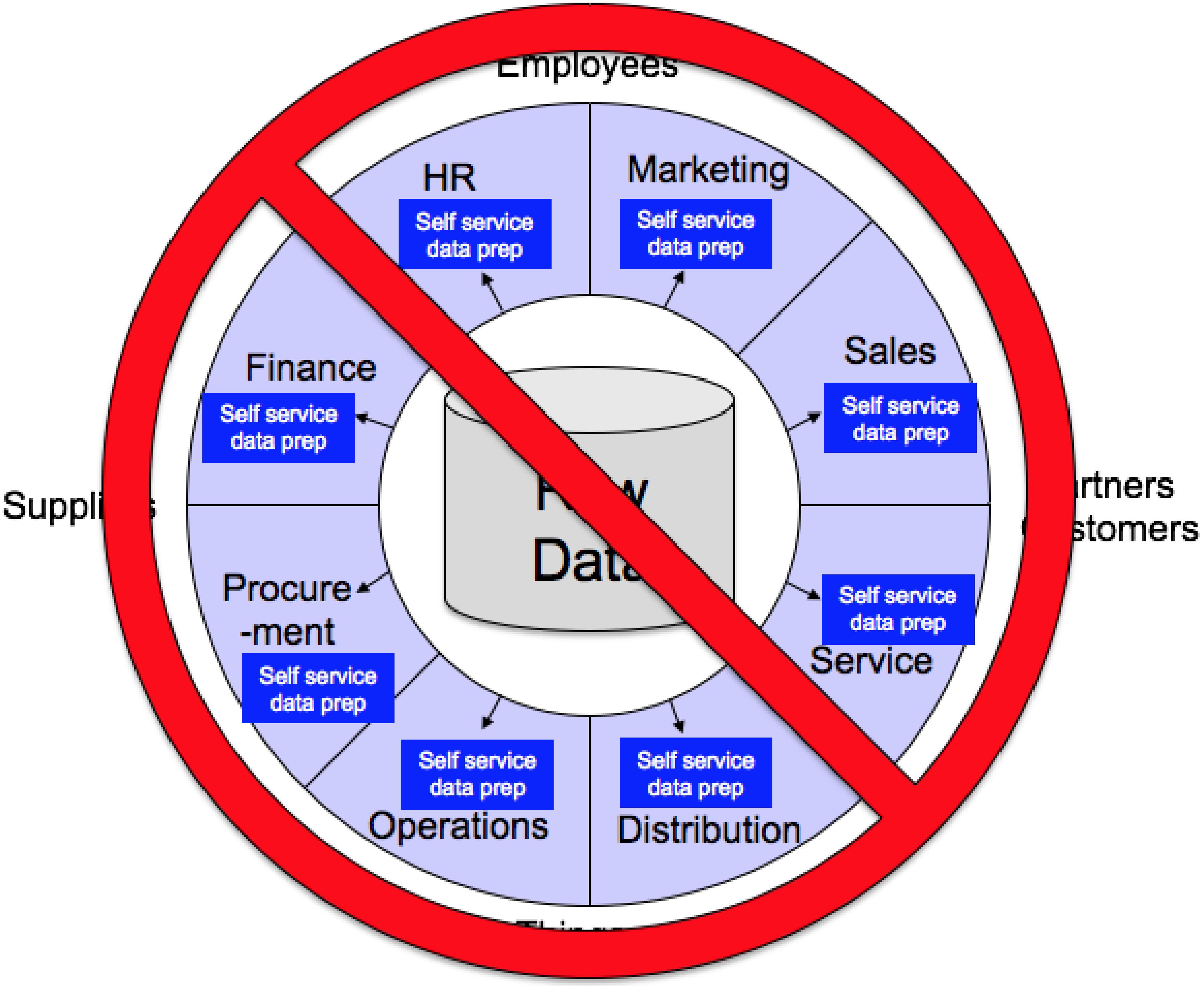
Silo	Silo	Silo	Silo	Silo	Silo	Silo	Silo	Silo	Silo	Silo	Silo
Analytical tools	Analytical tools	Analytical tools	Analytical tools	Analytical tools	Analytical tools	Analytical tools	Analytical tools	Analytical tools	Analytical tools	Analytical tools	Analytical tools
Data store	Data store	Data store	Data store	Data store	Data store	Data store	Data store	Data store	Data store	Data store	Data store
Data prep tools	Data prep tools	Data prep tools	Data prep tools	Data prep tools	Data prep tools	Data prep tools	Data prep tools	Data prep tools	Data prep tools	Data prep tools	Data prep tools
sources	sources	sources	sources	sources	sources	sources	sources	sources	sources	sources	sources

Challenges – How do you govern self-service data preparation to avoid chaos in the enterprise when departments are buying their own tools?



Everyone is blindly integrating data with no attempt to share what we create !!

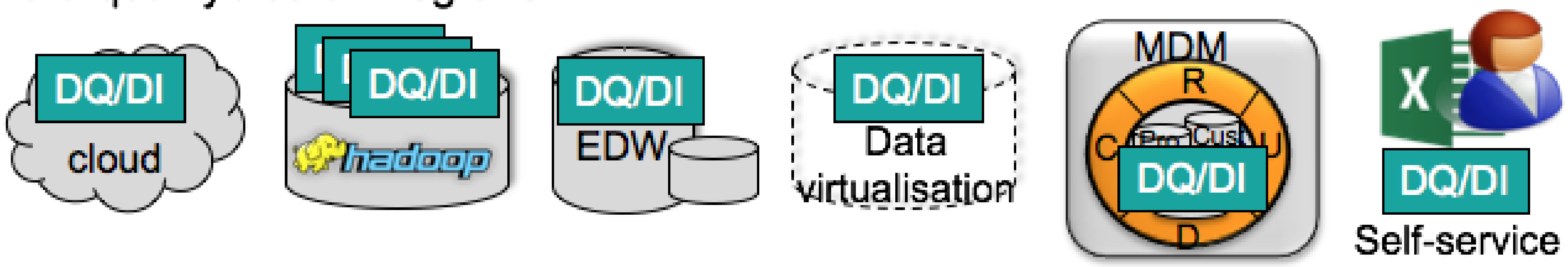
We do NOT want self-service chaos with many people creating inconsistent data


Employees
HR
Self service data prep
Marketing
Self service data prep
Sales
Self service data prep
Finance
Self service data prep
Raw Data
Suppliers
Partners
Customers
Procure-ment
Self service data prep
Self service data prep
Service
Self service data prep
Operations
Self service data prep
Distribution
Things

Problems with the current approach of many different tools and scripts being used by IT and business to integrate data

—

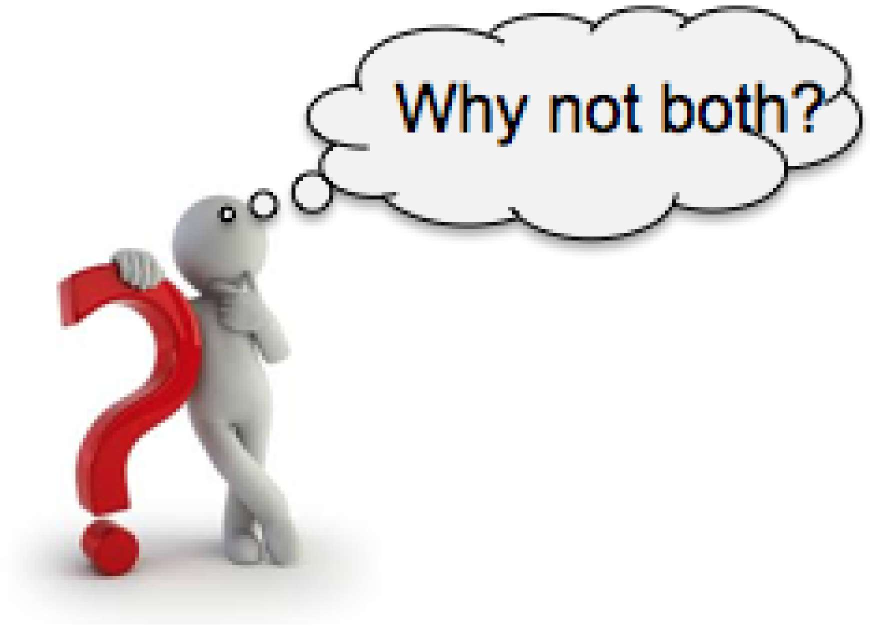
DQ/DI = Data quality / data integration



- Cost of data integration is much too high
- Slow speed of development
- Project oriented siloed approach to DQ/DI with limited collaboration
- Multiple DQ/DI technologies and techniques in use that are not integrated
- Multiple skill sets fractured across different tools and projects
- Lots of re-invention rather than re-use
- Fractured metadata across multiple tools or no metadata at all
- Risk of duplicate inconsistent DQ/DI rules for same data
- Metadata lineage is unavailable in many places
- Repetition of mistakes

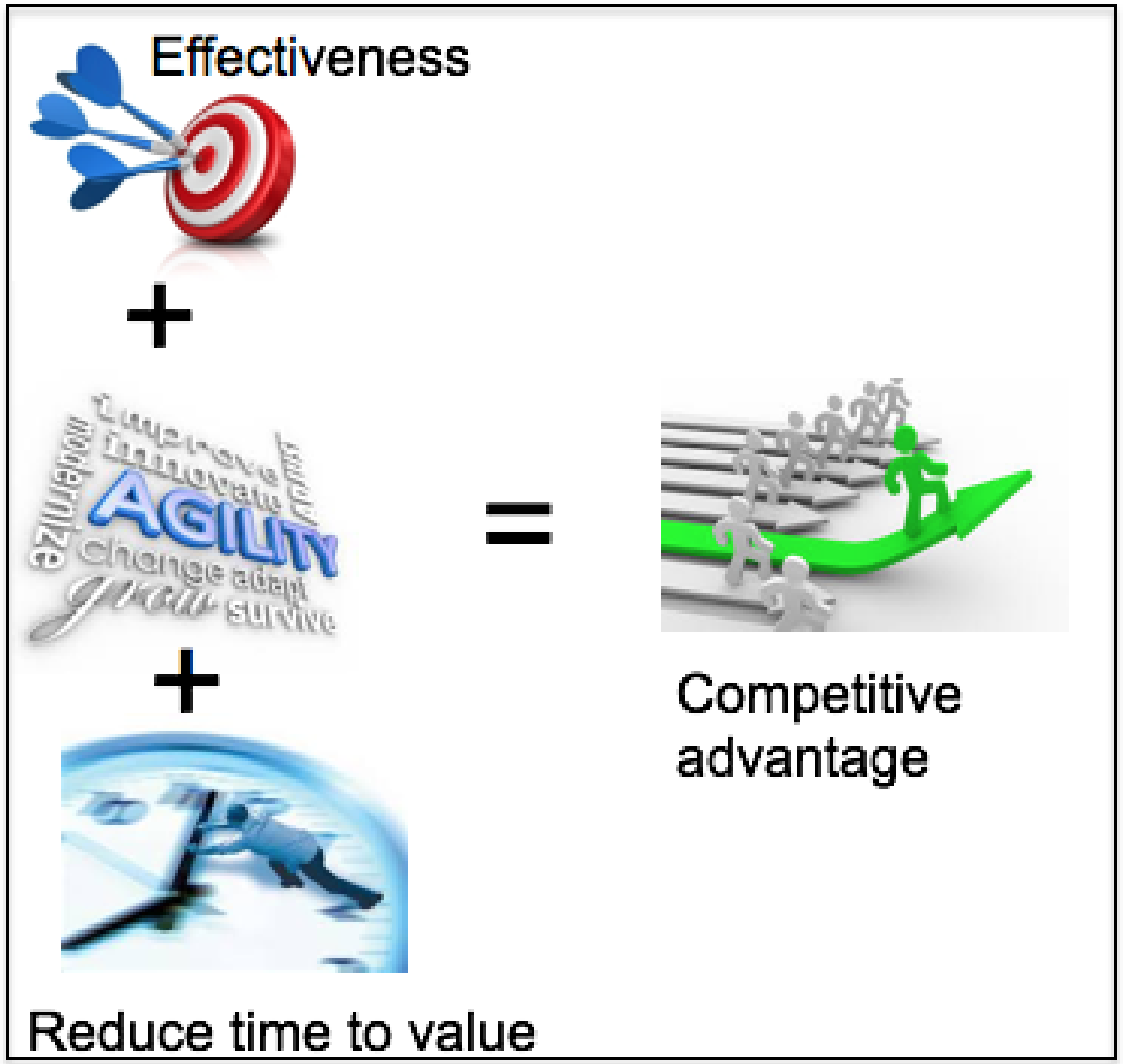
We need a way to accelerate ETL processing to be more agile, reduce time to value and improve effectiveness while also governing data

—



Data Governance

Vs



Effectiveness

+

+

=

Reduce time to value

Competitive advantage

Freedom to get ahead of the competition

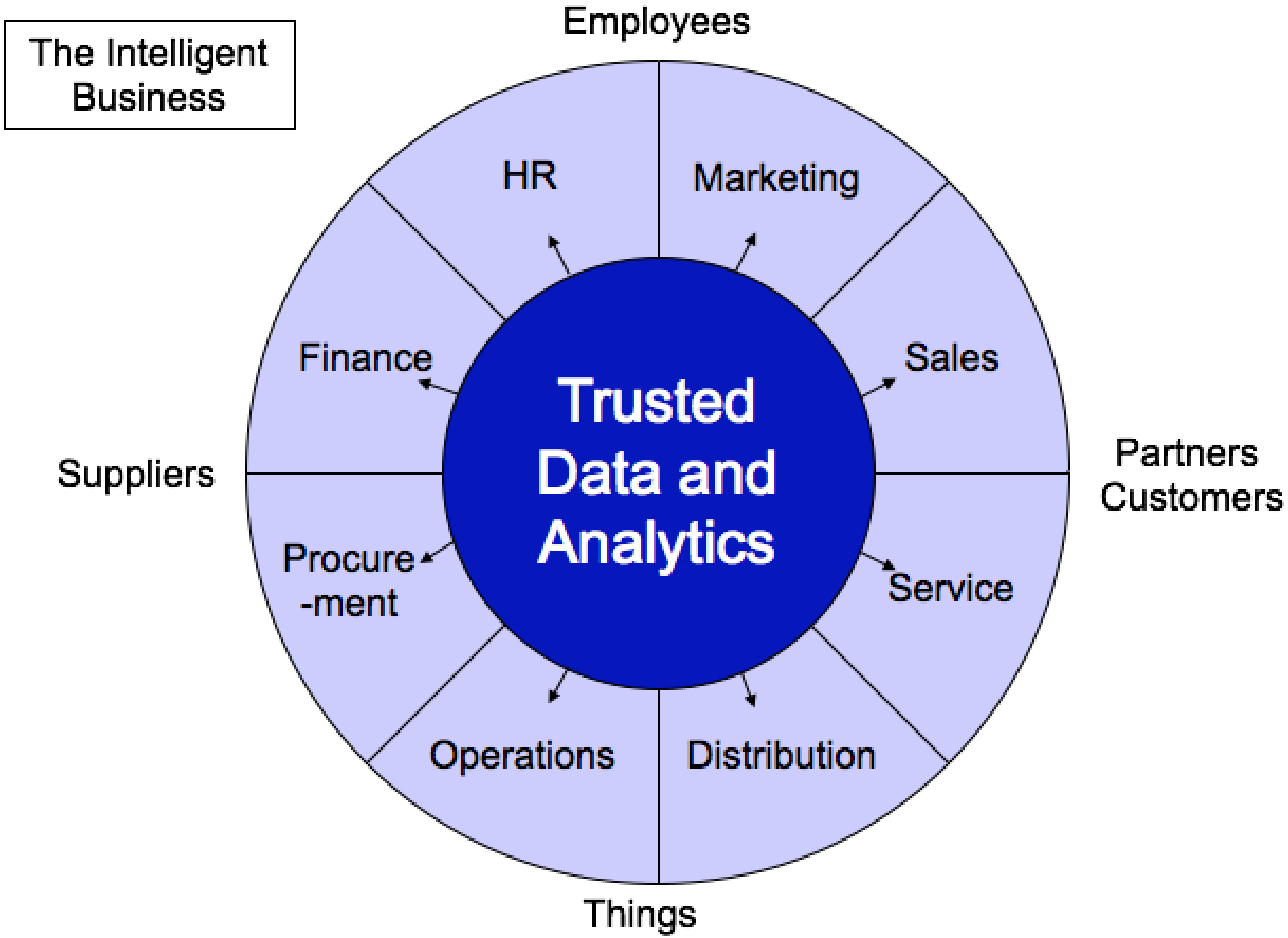
How can you accelerate ETL processing?

Build once, re-use everywhere

—

We want trusted data assets and analytics built once, reused everywhere in a data driven-enterprise

The Intelligent Business



Employees

HR

Marketing

Finance

Trusted Data and Analytics

Sales

Suppliers

Partners
Customers

Procure
-ment

Service

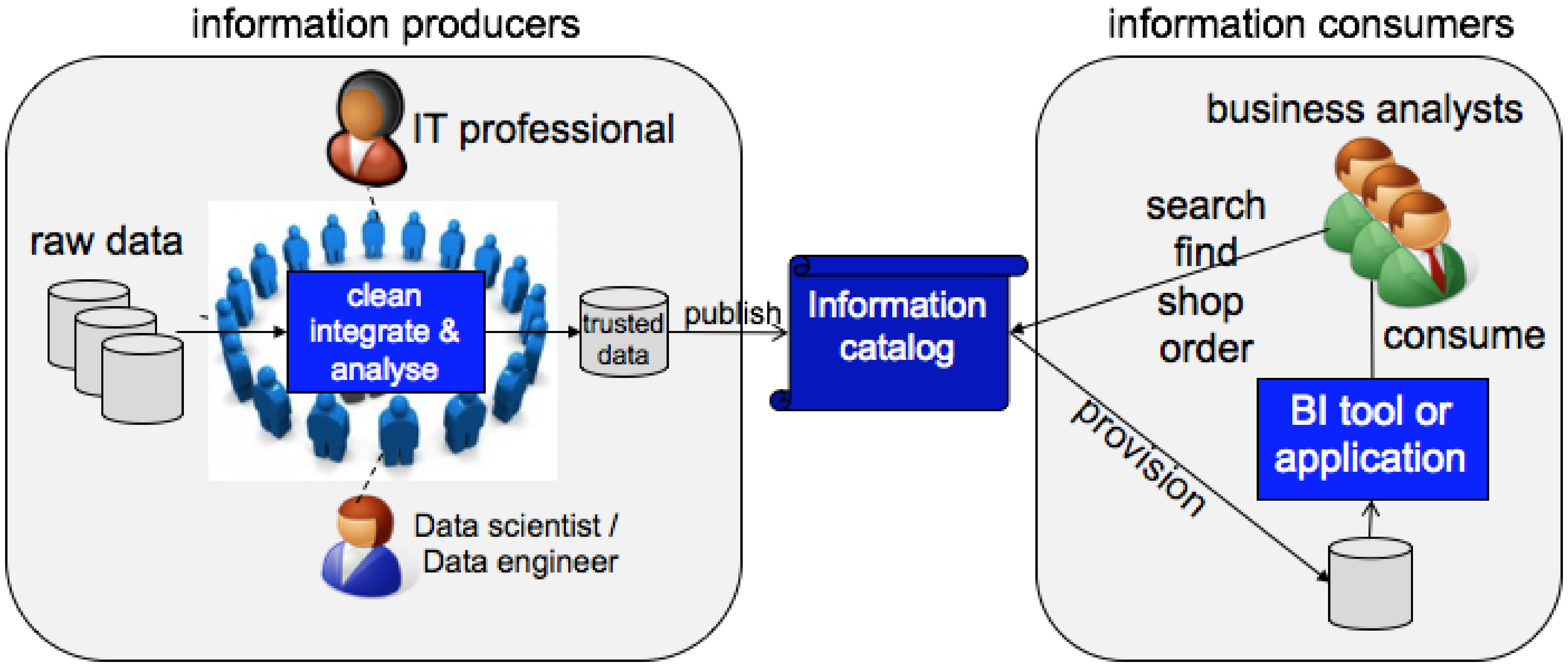
Operations

Distribution

Things

Organize to become data driven □– information producers and information consumers

—



Information producers and information consumers need to make use of

- A business glossary and information catalog
- Role-based data management tools aimed at IT AND business
- A collaborative approach by business and IT to produce data and analytical assets
- A catalog to quickly find reusable trusted assets to drive business value

We need to create trusted, business ready data for users to easily find, consume and use to drive value faster

Data available as a Service

BUSINESS READY

Business ready data products can be logical entities

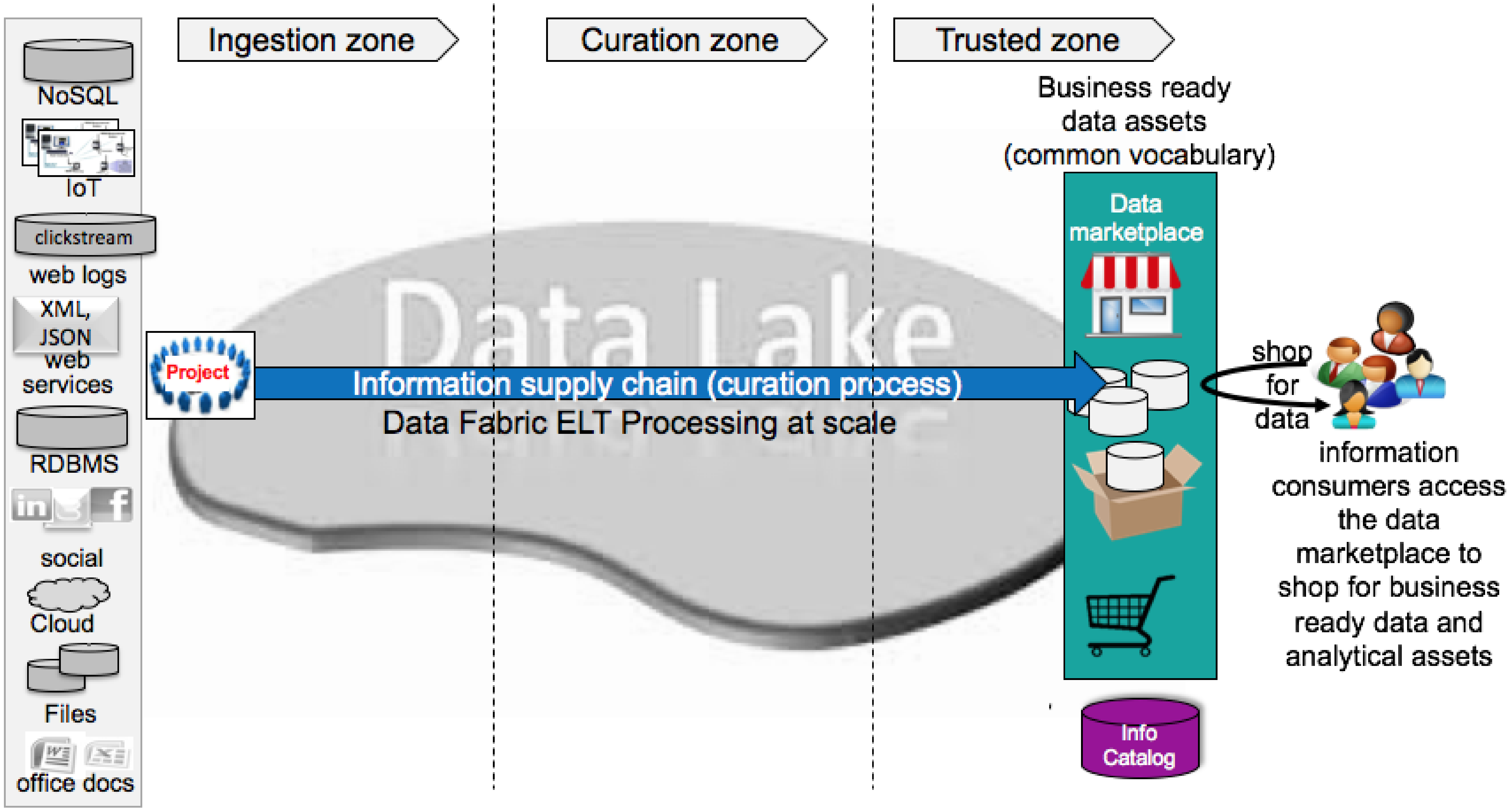
<u>Master Data</u>

- Customers
- Products
- Suppliers
- Assets
- Employees
- Materials

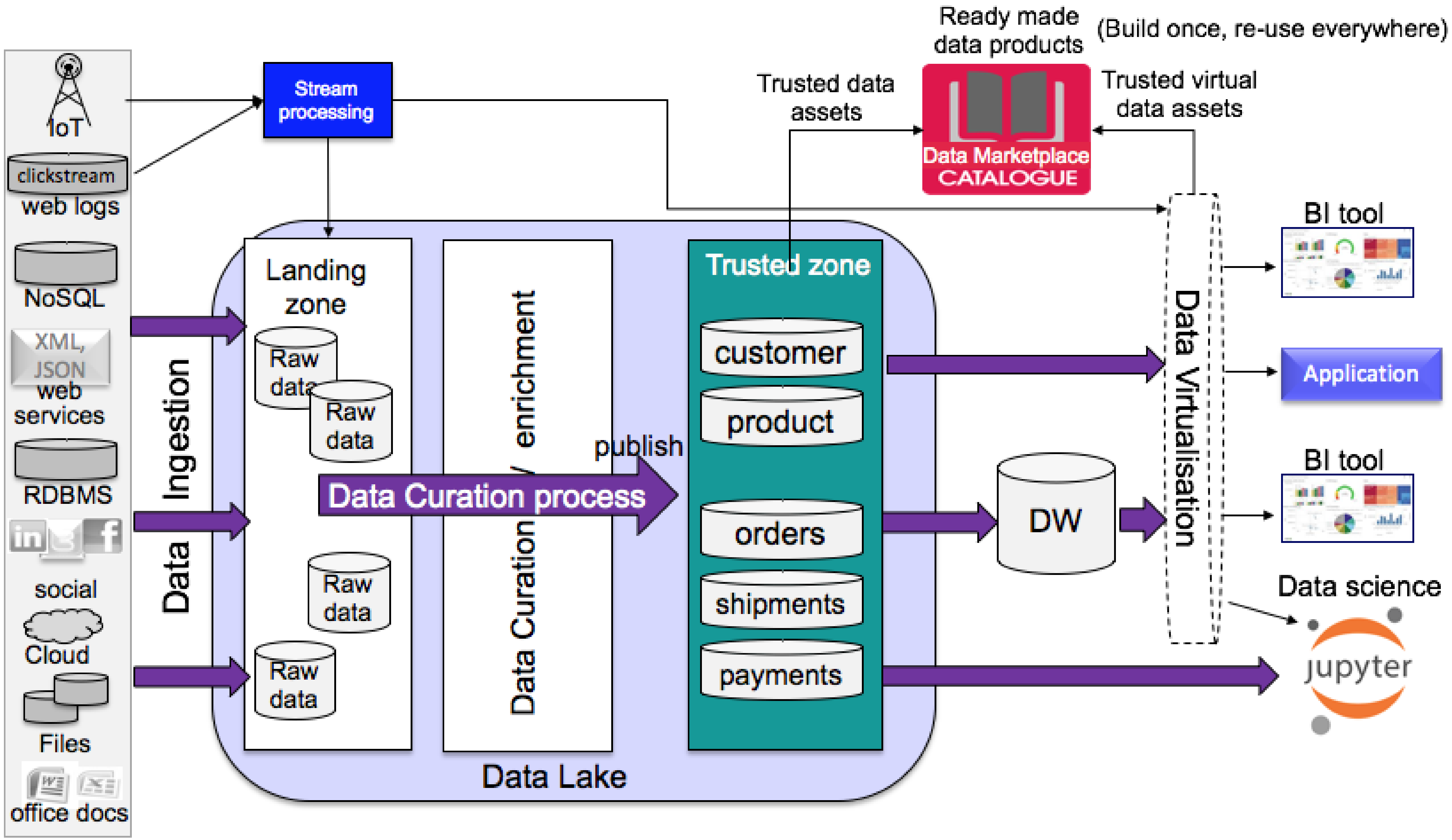
<u>Transaction Data</u>

- Orders
- Shipments
- Payments
- Adjustments
- Returns

Create a data lake to organize data so ‘business ready’ data and analytical assets can be produced and published in a marketplace for users to consume


NoSQL
IoT
clickstream
web logs
XML, JSON web services
RDBMS
social
Cloud
Files
office docs
Ingestion zone
Curation zone
Trusted zone
Data Lake
Project
Information supply chain (curation process)
Data Fabric ELT Processing at scale
Business ready data assets (common vocabulary)
Data marketplace
Info Catalog
shop for data
information consumers access the data marketplace to shop for business ready data and analytical assets

Use trusted data assets to consistently build MDM, DW, data sets for data science and virtual data services



Ready made data products

(Build once, re-use everywhere)

Trusted data assets

Data Marketplace CATALOGUE

Trusted virtual data assets

IoT

clickstream

web logs

NoSQL

XML, JSON

web services

RDBMS

social

Cloud

Files

office docs

Stream processing

Data Ingestion

Landing zone

Raw data

Raw data

Raw data

Raw data

Data Curation / enrichment

Data Curation process

publish

Trusted zone

customer

product

orders

shipments

payments

Data Lake

DW

Data Virtualisation

BI tool

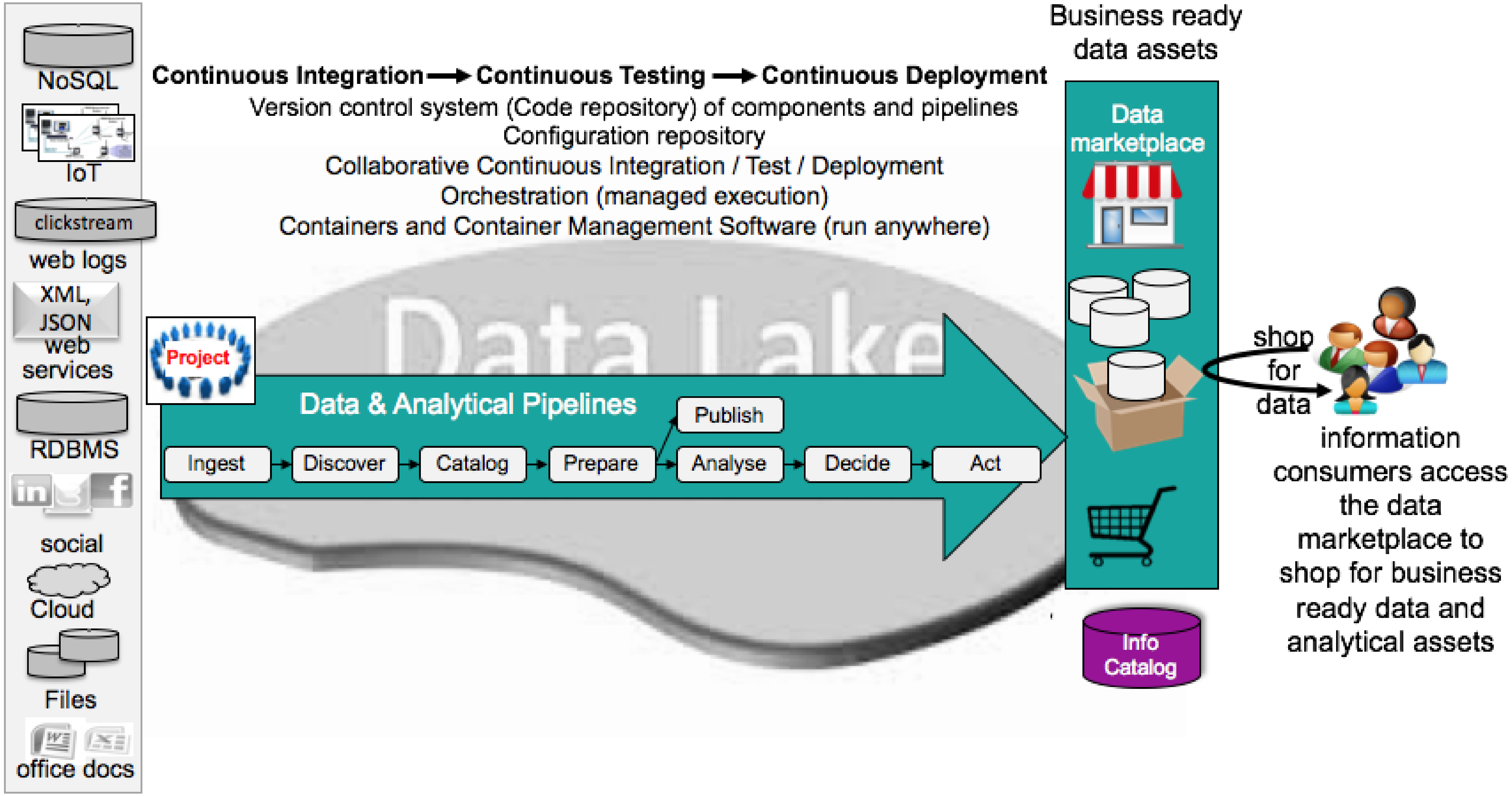
Application

BI tool

Data science

jupyter

Dataops – The continuous integration / continuous deployment (CI/CD) of component based pipelines to produce 'business ready' data & analytical assets



NoSQL
IoT
clickstream
web logs
XML, JSON web services
RDBMS
social
Cloud
Files
office docs

Continuous Integration → Continuous Testing → Continuous Deployment

Version control system (Code repository) of components and pipelines
Configuration repository
Collaborative Continuous Integration / Test / Deployment
Orchestration (managed execution)
Containers and Container Management Software (run anywhere)

Data Lake

Project

Data & Analytical Pipelines

Ingest → Discover → Catalog → Prepare → Publish
Prepare → Analyse → Decide → Act

Business ready data assets

Data marketplace

Info Catalog

shop for data

information consumers access the data marketplace to shop for business ready data and analytical assets

Upcoming Courses

Big Data Architecture and Technology for Analytics

Cloud Data Warehouse Migration

Data Warehouse ETL: The Kimball Approach

Data Warehouse Lifecycle: The Kimball Approach

Data Warehouse Modernization

Designing, Operating and Managing an Enterprise Data Lake

Hands-on Data Science for BI Professionals and Data Analysts

Dimensional Modeling: The Kimball Approach

Enterprise Data Governance & Master Data Management

STAY TUNED FOR PART 3

Modern Data Pipelines