Why Modernize Your Data Warehouse

Part 1 of the Modern Data Series

Mike Ferguson

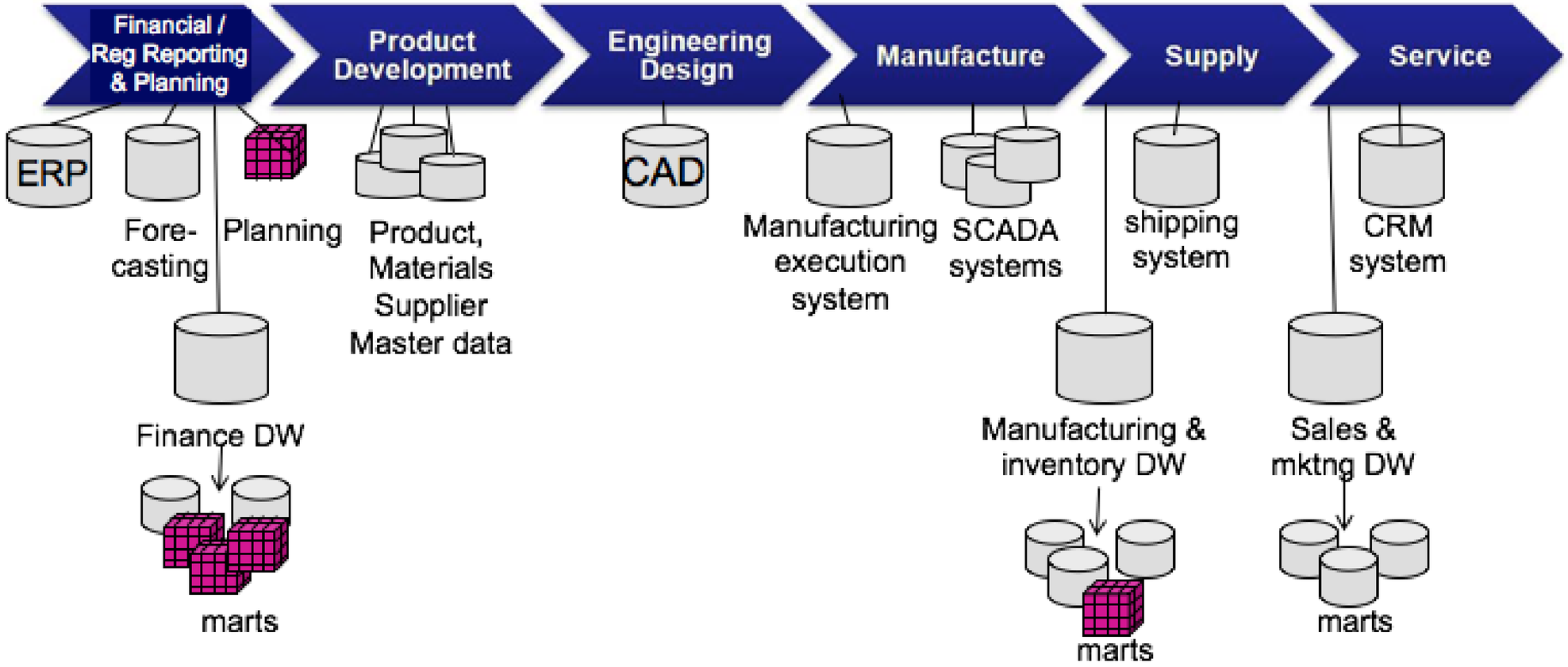
quest for knowledge®

q4k.com

Many companies today have built multiple data warehouses and data marts in different parts of their value chain

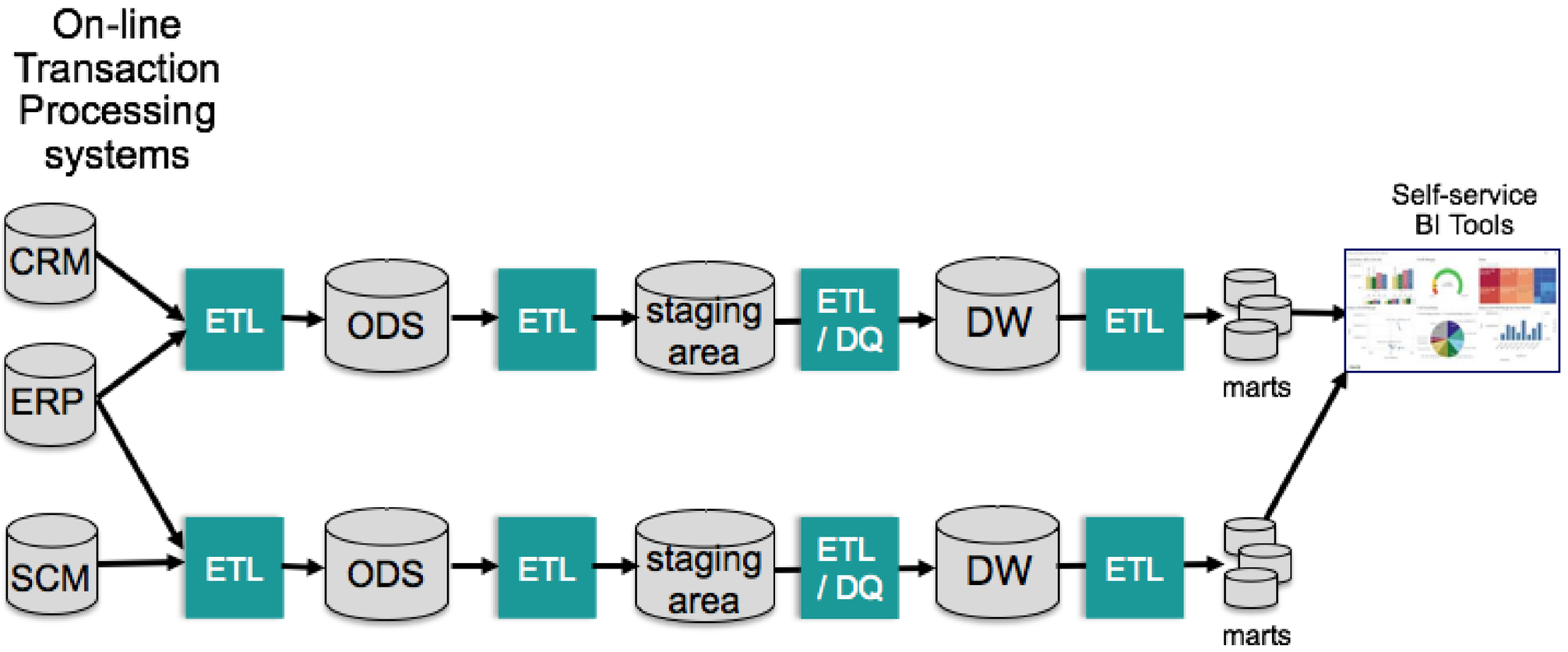
—

Makes management and regulatory reporting more challenging as data needs to be integrated to see across the value chain



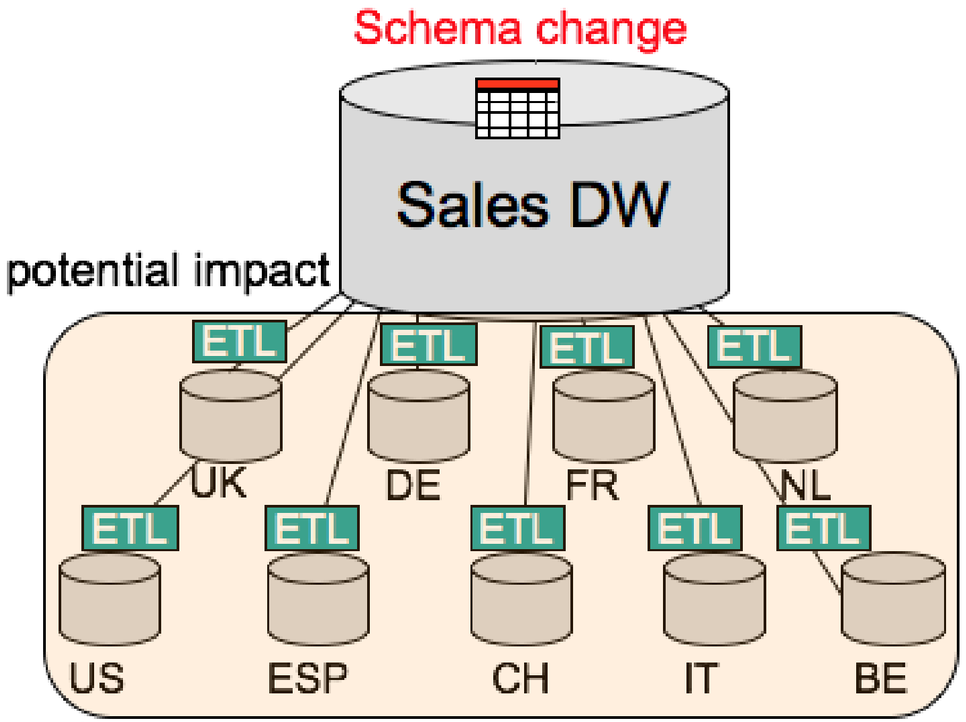
The traditional data warehouse has a multiple data store architecture and a waterfall approach to data flow

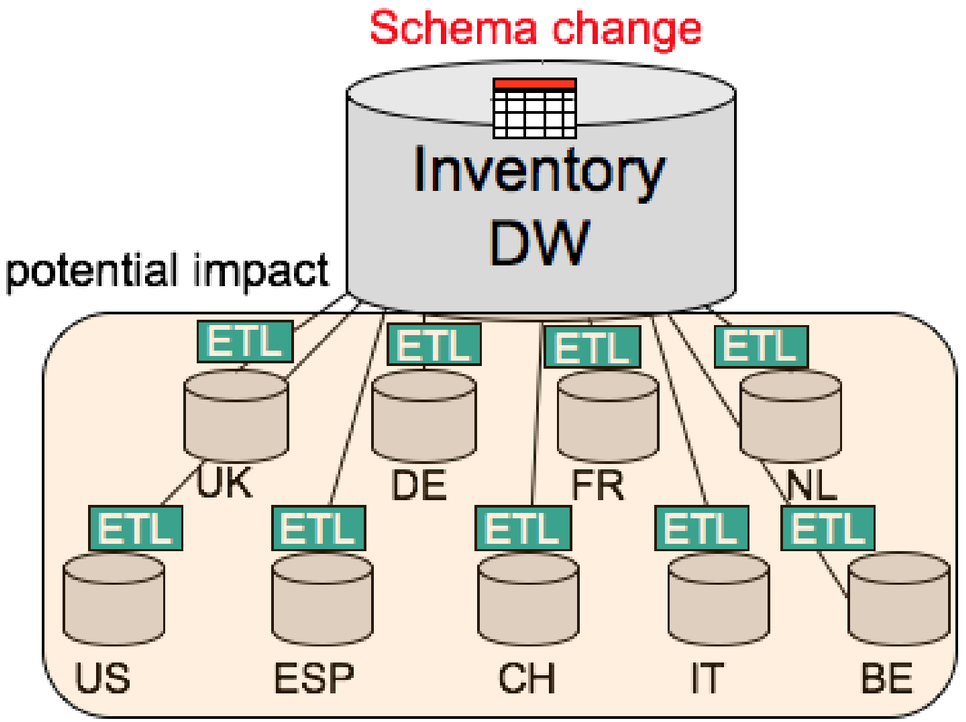
—


On-line
Transaction
Processing
systems
CRM
ERP
SCM
ETL
ODS
ETL
staging
area
ETL
/ DQ
DW
ETL
marts
ETL
ODS
ETL
staging
area
ETL
/ DQ
DW
ETL
marts
Self-service
BI Tools

Problems with existing data warehouses - the impact of change on ‘Production’ data warehouses

—





Change Impact

Potential changes to all data mart data models
Potential changes to all data mart ETL jobs
Potential changes to all BI tool semantic layers
Potential changes to many BI reports & dashboards

In some cases the same change may need to be made to multiple data warehouses!

Problems with existing data warehouse architectures - change management

—

How long does it take your organisation to make, test and deploy a schema change to your data warehouse?

1. One day or less?
2. Between one day and one week?
3. Between one week and one month?
4. Three months?
5. Six months or longer?

The demand for new data has increased rapidly with business creating other analytical systems outside of a data warehouse

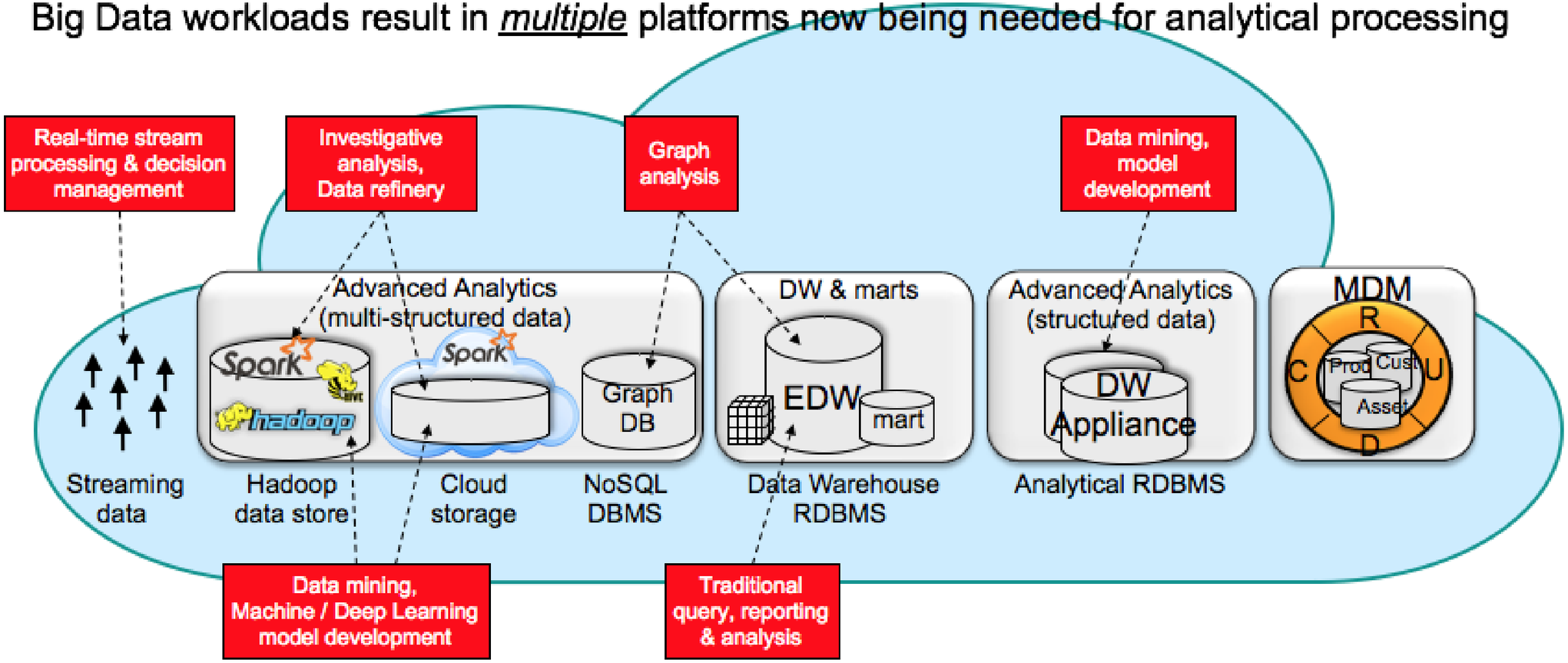
—

Type Of Data	Examples	Uses
Traditional structured data	• Master data • Transaction data	• Customer, product, employee, supplier, site,….. • Orders, shipments, returns, payments, adjustments…
Machine generated data	• Clickstream web server logs • IVR logs, App Server logs • DBMS logs	• On-line behaviour analysis • Cyber security
	• Consumer IoT (Sensor data) • Industrial IoT (Sensor data) • Location, temperature, movement, vibration, pressure	• Product usage behaviour • Product or equipment performance
Human generated data	• Social network data • Inbound email • Competitor news feeds • Documents • Voice interaction data	• Unstructured text, sentiment analysis 
External data	• Open government data • Weather data	• Structured data • Semi-structured data, e.g. JSON, XML, AVRO • Sales impact, distribution impact

The changing landscape – we now have different platforms optimized for different analytical workloads

—

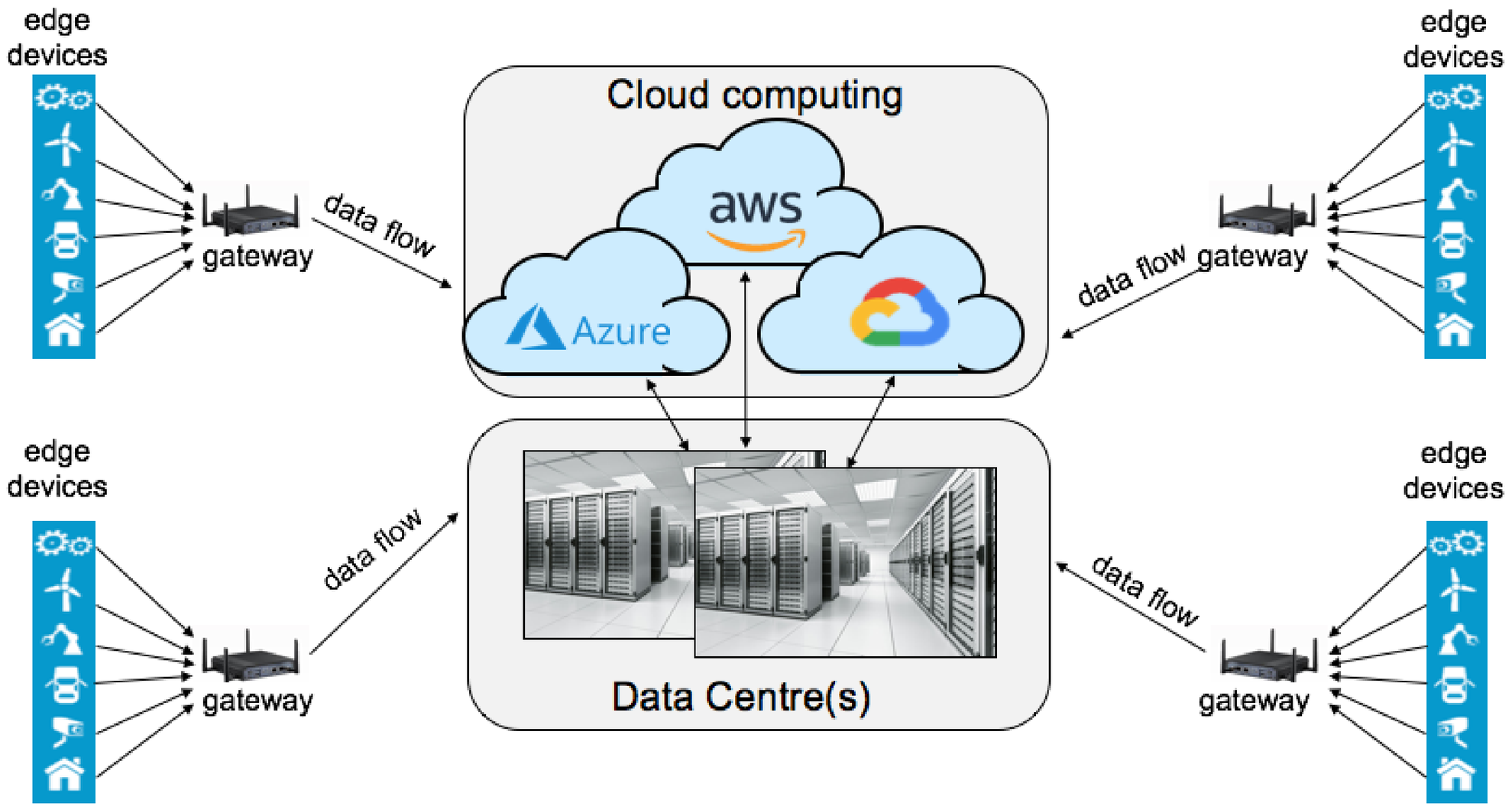
Big Data workloads result in *multiple* platforms now being needed for analytical processing



Several vendors now offer the entire analytical ecosystem on the cloud
Alternatively it can be a hybrid setup

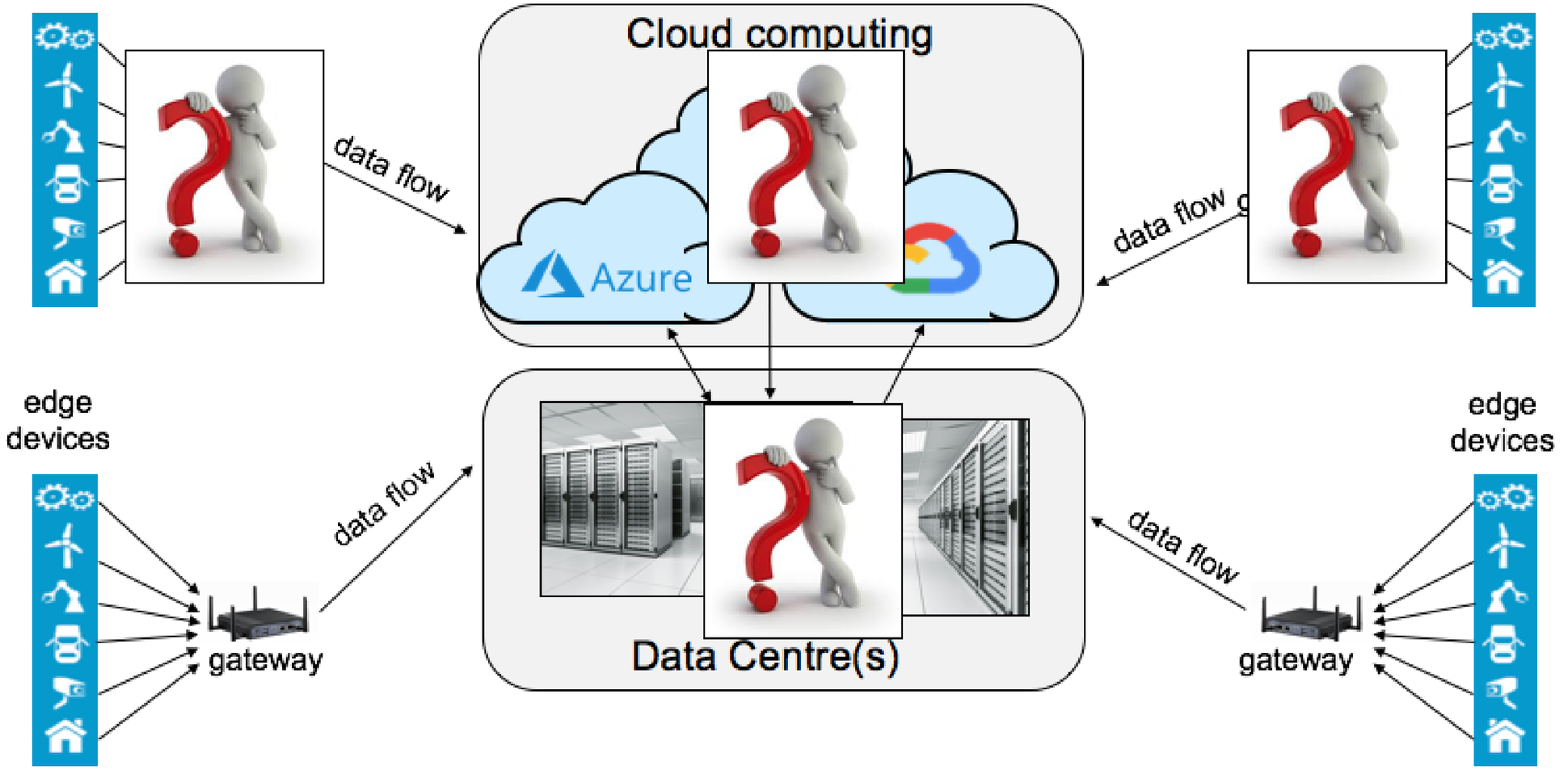
The operating environment that we have now created is more complex spanning the edge, multiple clouds and the data centre

—


edge
devices
gateway
data flow
Cloud computing
aws
Azure
edge
devices
data flow
gateway
edge
devices
gateway
data flow
Data Centre(s)
edge
devices
data flow
gateway

How will you capture, store, integrate and analyse data across the edge, multiple clouds and the data centre?

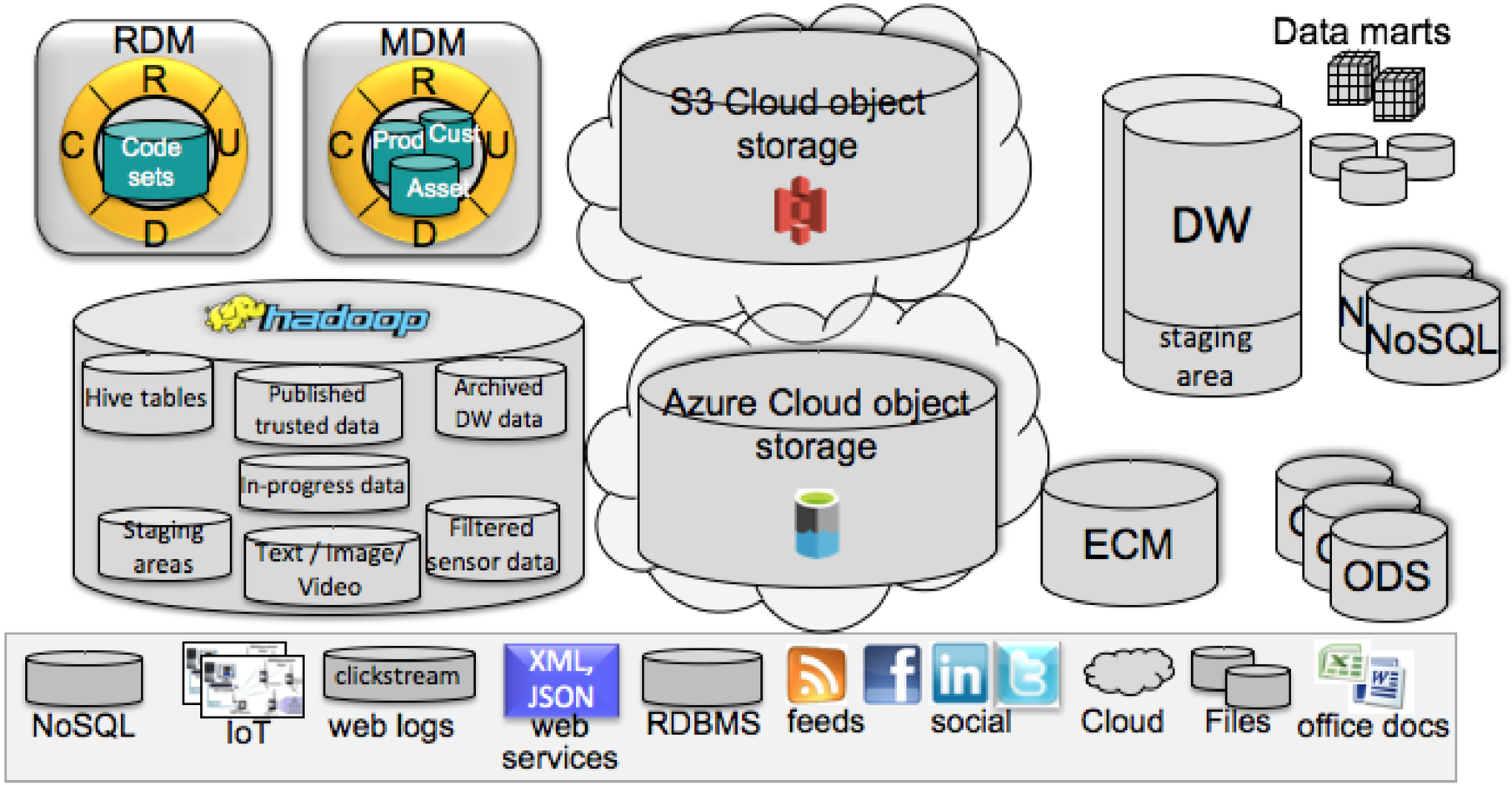
—


Cloud computing
Azure
data flow
data flow
data flow
data flow
edge
devices
edge
devices
gateway
gateway
Data Centre(s)

Problems caused by greater complexity - finding and governing data when it is stored in so many places

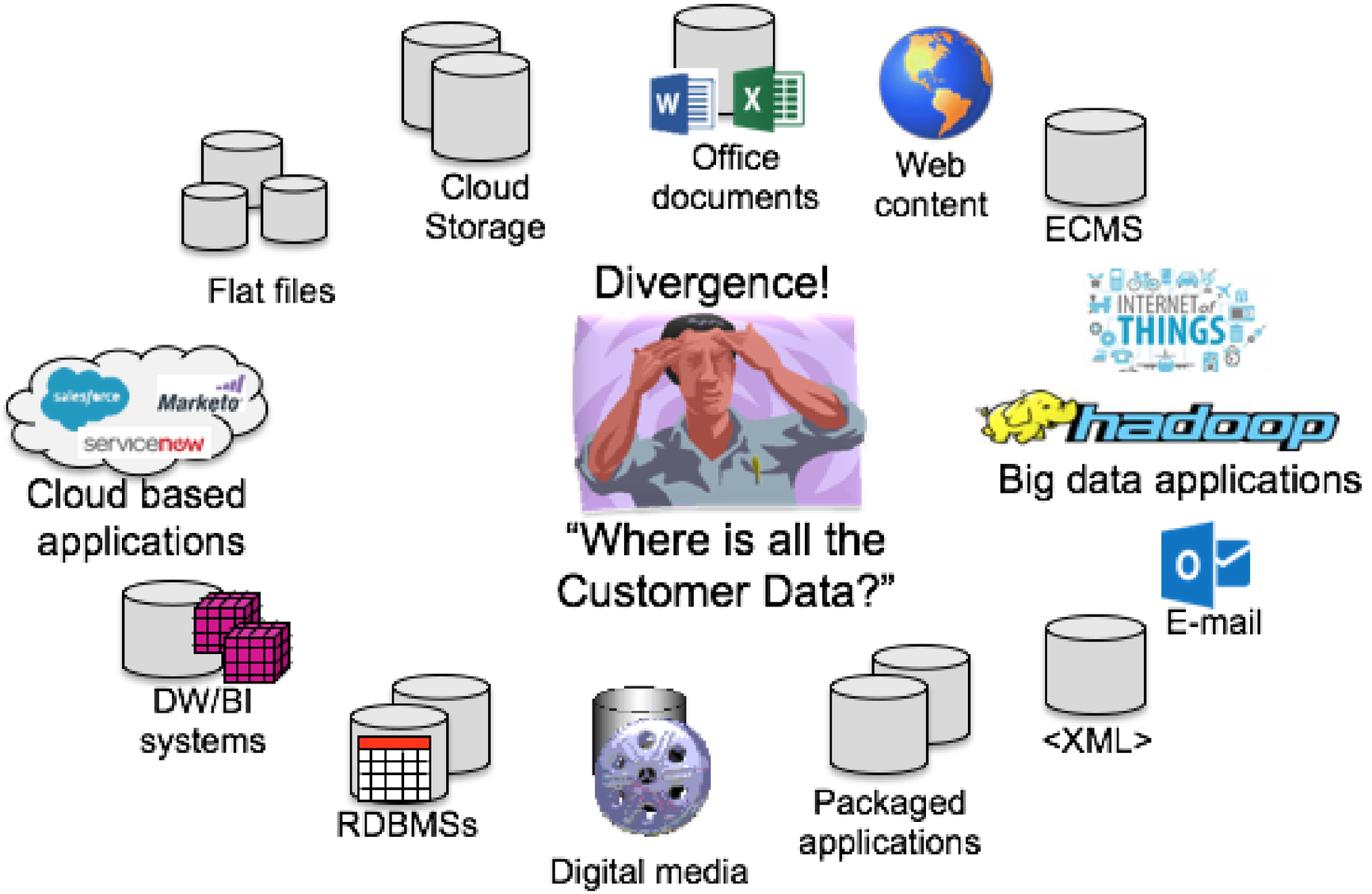
—

Data resides in multiple data stores
Data duplication, overlapping subsets and multiple versions of data
No idea what data is located where or what it means
No idea of data quality, where sensitive data, access security….
No view of what data is processed where and by whom….



Information overload – managing, governing and integrating data is becoming increasingly complex as data sources grow

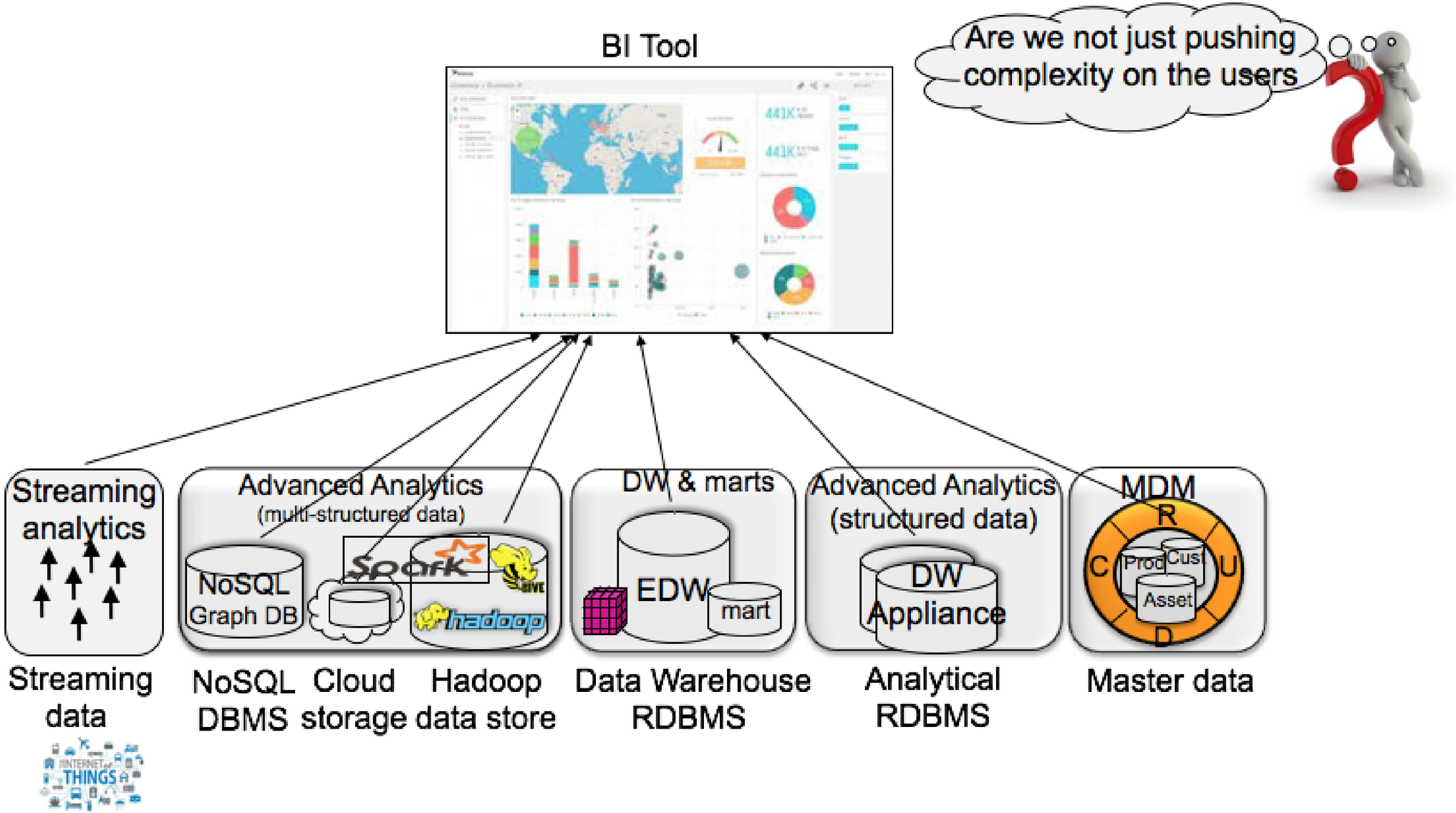
—



More and more data sources now need to be integrated to provide information for business use

Problems with existing data warehouse - multiple analytical systems are not integrated forcing bi users to connect to multiple data stores

—


BI Tool
Are we not just pushing complexity on the users
Streaming analytics
Advanced Analytics (multi-structured data)
NoSQL Graph DB
Spark
hadoop
DW & marts
EDW
mart
Advanced Analytics (structured data)
DW Appliance
MDM
R
C
U
D
Prod
Cus
Asset
Streaming data
NoSQL DBMS
Cloud storage
Hadoop data store
Data Warehouse RDBMS
Analytical RDBMS
Master data

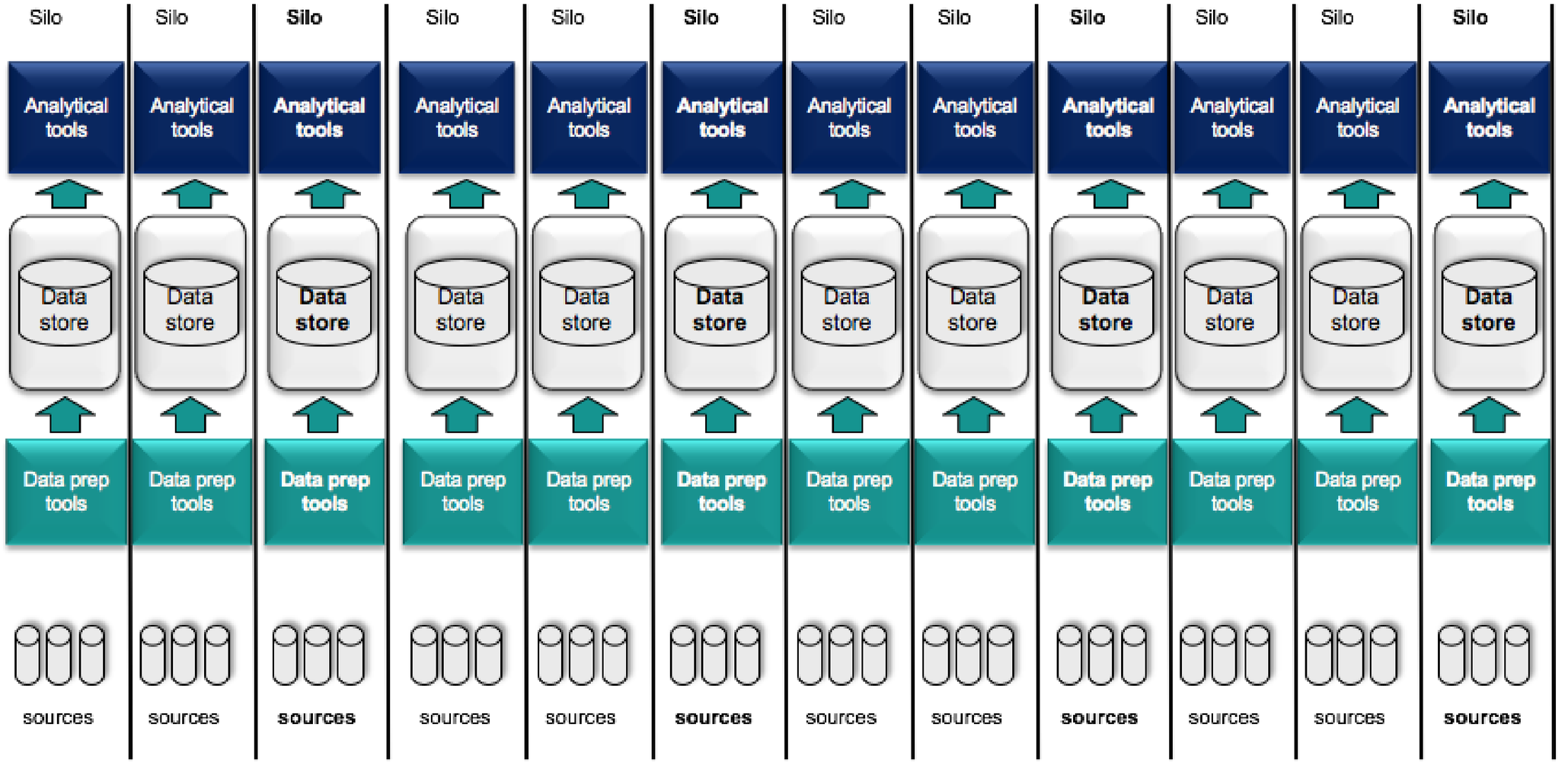
The danger of self-service data prep – an explosion of silos

—

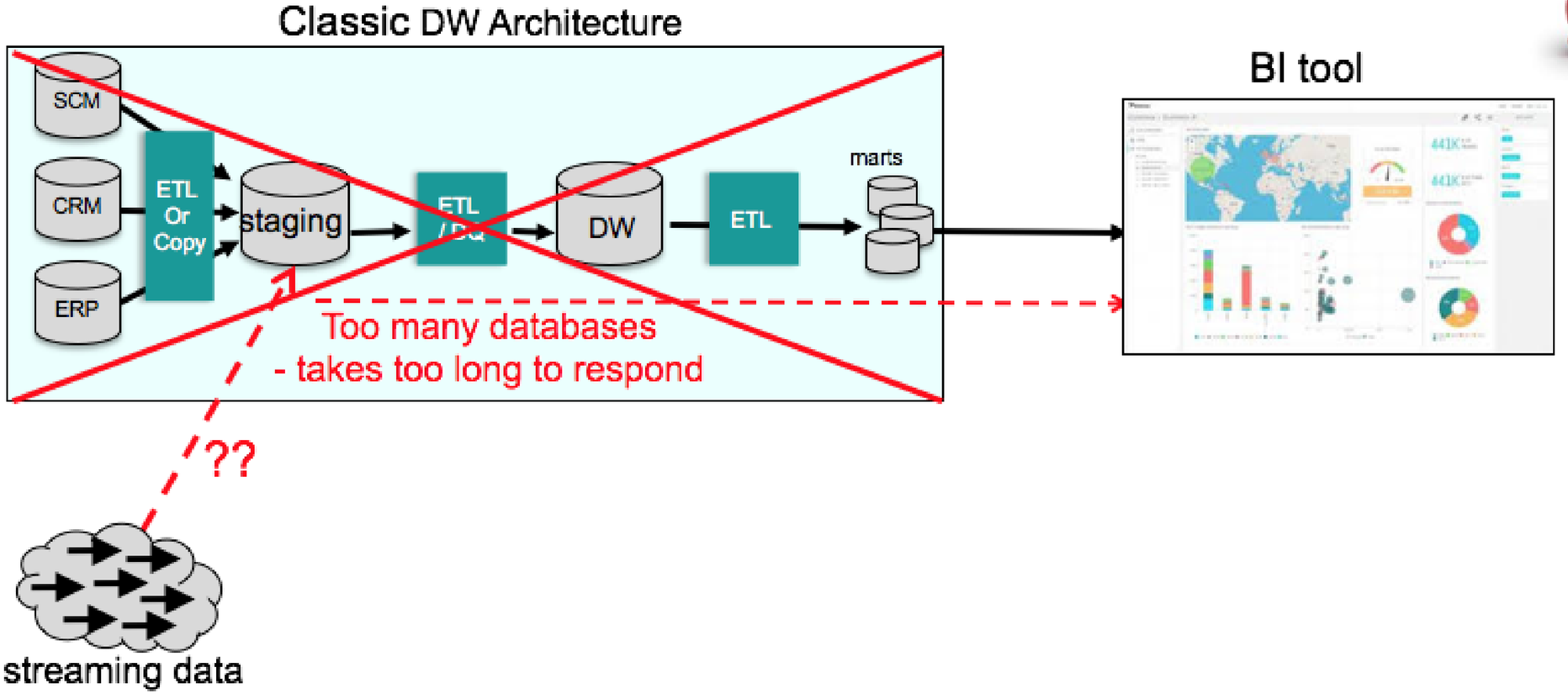
INTELLIGENT BUSINESS STRATEGIES

quest for knowledge

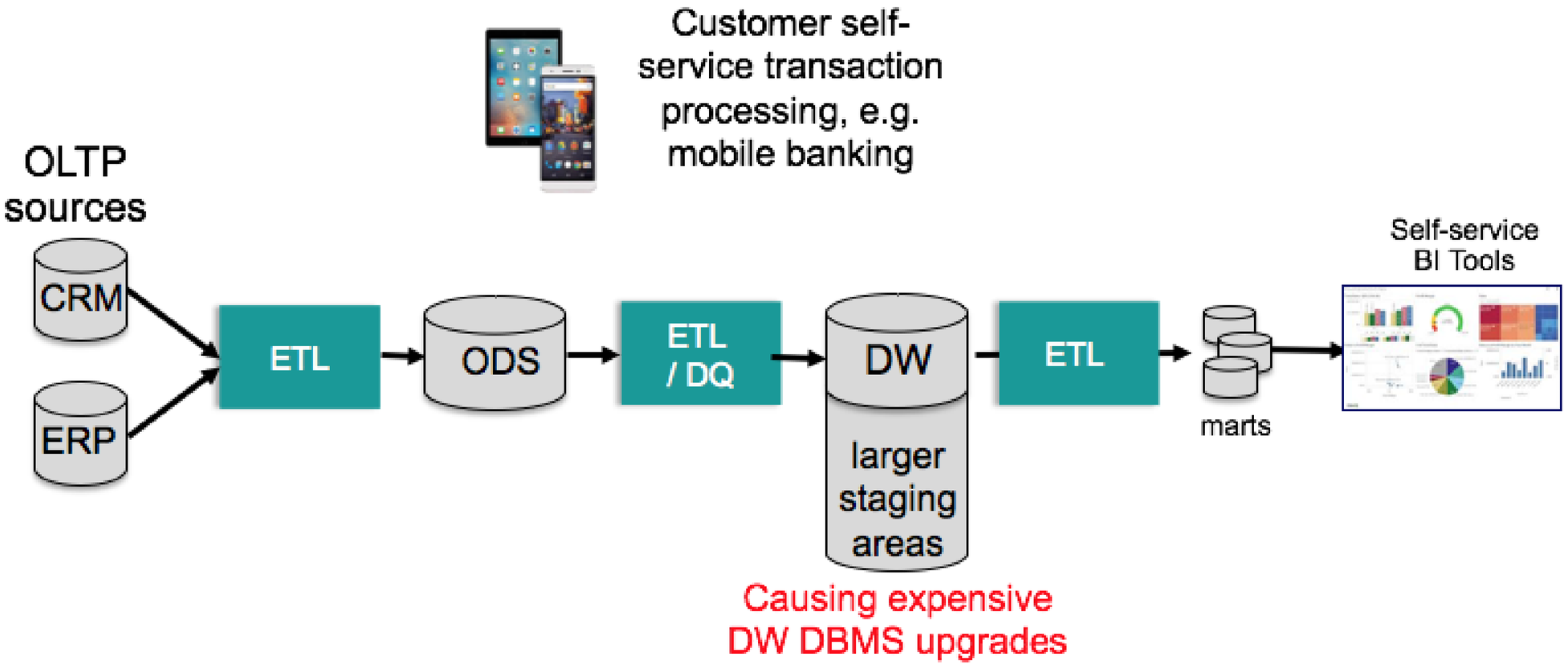
q4k.com


Silo
Analytical tools
Data store
Data prep tools
sources

Problems with existing data warehouse architecture – classic DW architecture is often unable to support access to streaming data


Classic DW Architecture
SCM
CRM
ERP
ETL
Or
Copy
staging
ETL
DW
ETL
marts
BI tool
Too many databases
- takes too long to respond
??
streaming data

Problems with existing data warehouse architecture - growth in transaction volumes is causing larger staging areas


Customer self-service transaction processing, e.g. mobile banking
OLTP sources
CRM
ERP
ETL
ODS
ETL / DQ
DW
larger staging areas
ETL
marts
Self-service BI Tools
Causing expensive DW DBMS upgrades

Modern Data Warehouse Requirements

—

INTELLIGENT BUSINESS STRATEGIES

quest for knowledge

q4k.com

Area	Requirements

Design	• Modern data modelling techniques
	• Rapid turnaround of new requirements (agility)
	  • Quick and easy to change (e.g. new source, new column)
	  • Very low impact of change
	• Reduced total cost of ownership
	• Easy to load
	• Multiple views of data that can co-exist

Data preparation and integration	• Business glossary • Automated data discovery, profiling and derivation of lineage • Information catalog • Offload of staged data from the DW to a data lake • Scalable ETL processing of multi-structured data • Integrated IT and self-service data preparation • Continuous integration / continuous delivery (CI / CD) • Reusable trusted data assets in a data marketplace • Fewer data copies
DBMS	• Cloud based analytical RDBMS • Integration with other analytical data stores • External tables to point to data in a data lake • In-database analytics, e.g. machine learning, graph analytics, etc.
Accelerated DW Development	• Business ready data assets published in a marketplace • Data warehouse automation for rapid development • Continuous integration / continuous delivery (CI / CD) • Common vocabulary
Simplified access and data independence	• Data Virtualisation for data independence, flexibility and agility and to integrate DW and other analytical data stores in a logical data warehouse

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 2

Accelerating ETL Development Using a Data Lake