

Driving better predictions in the oil and gas industry with modern data architecture

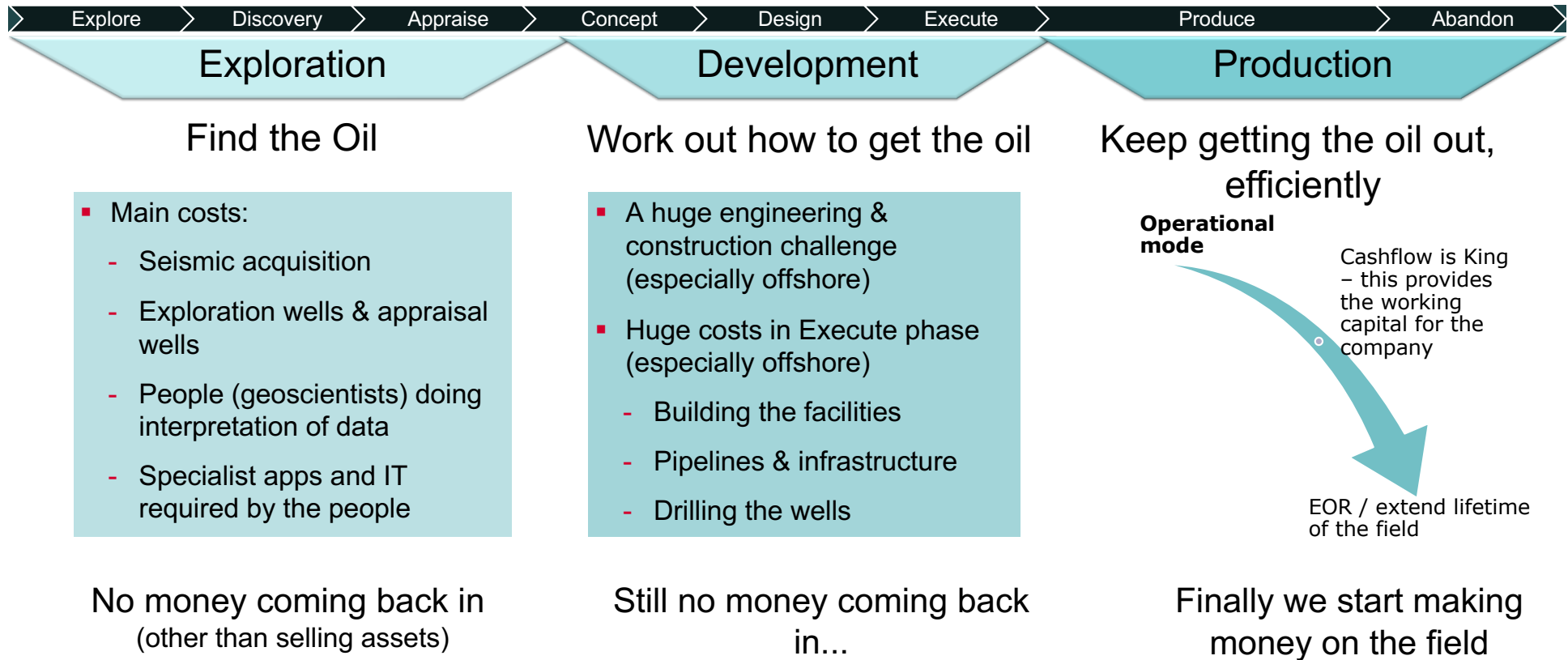
Jane McConnell and Paul Ibberson
Teradata

Strata
DATA CONFERENCE

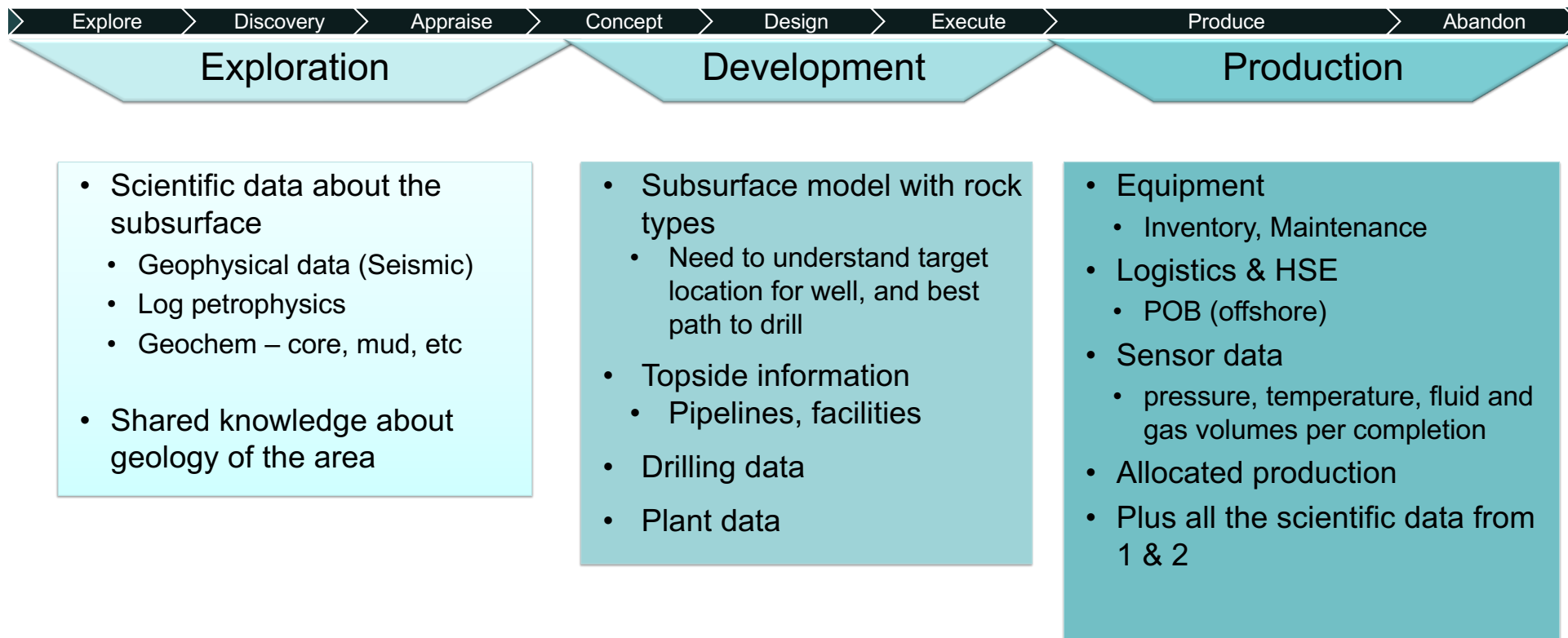
Upstream Oil and Gas Industry



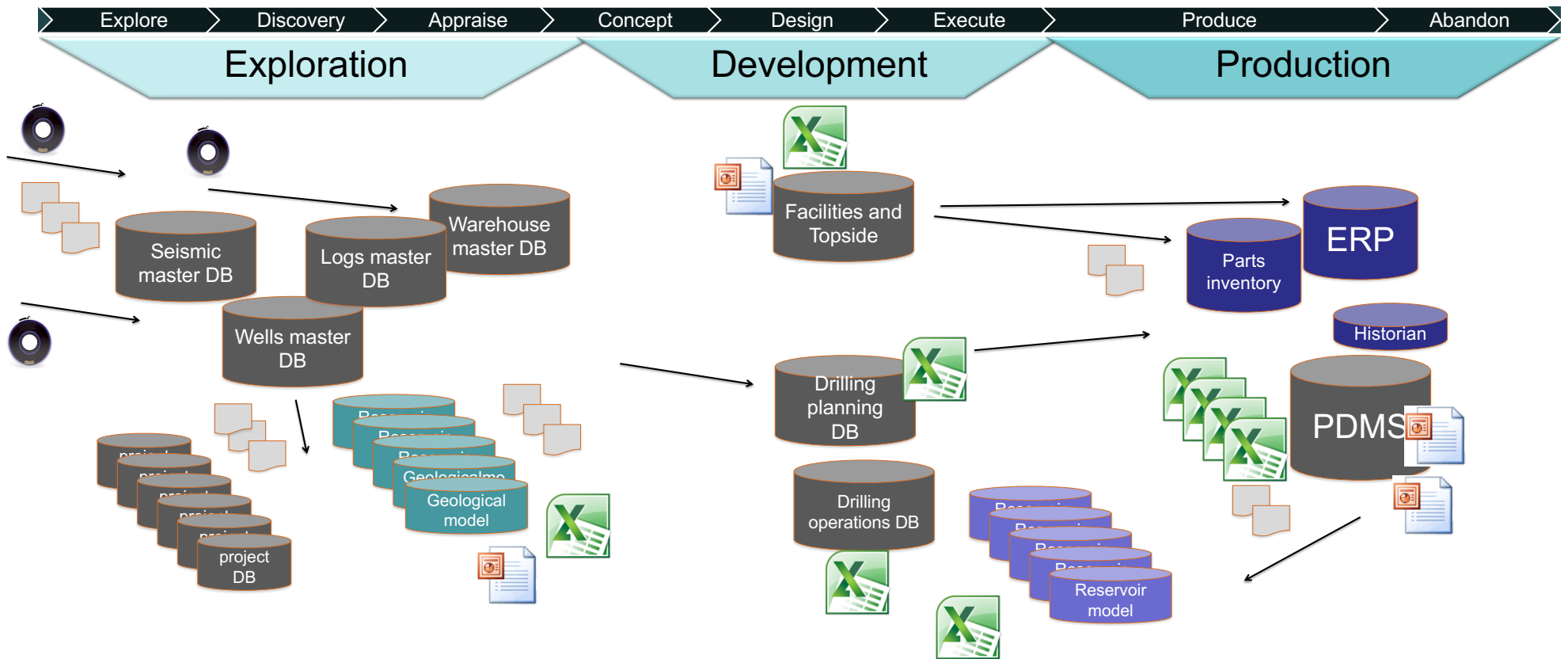
The Life of an Oil Field



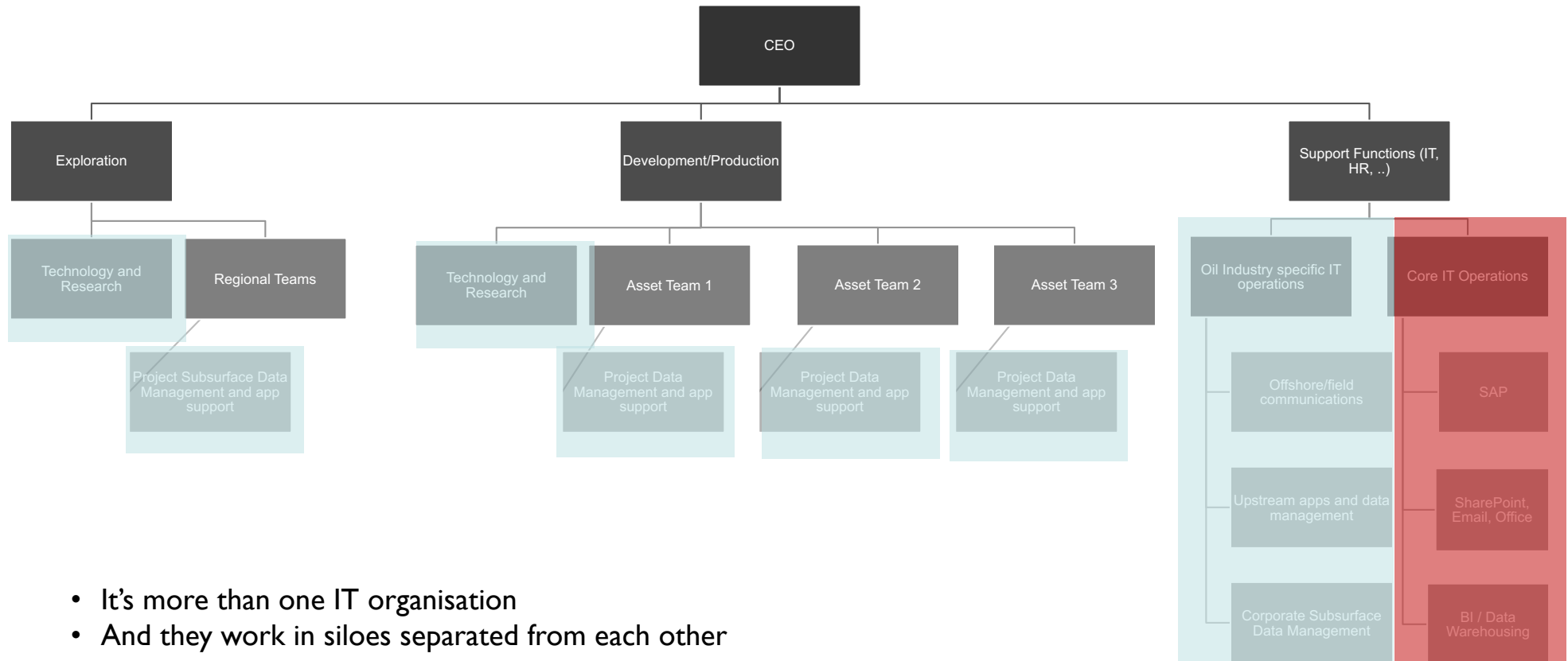
Types of Data



The Data Challenge



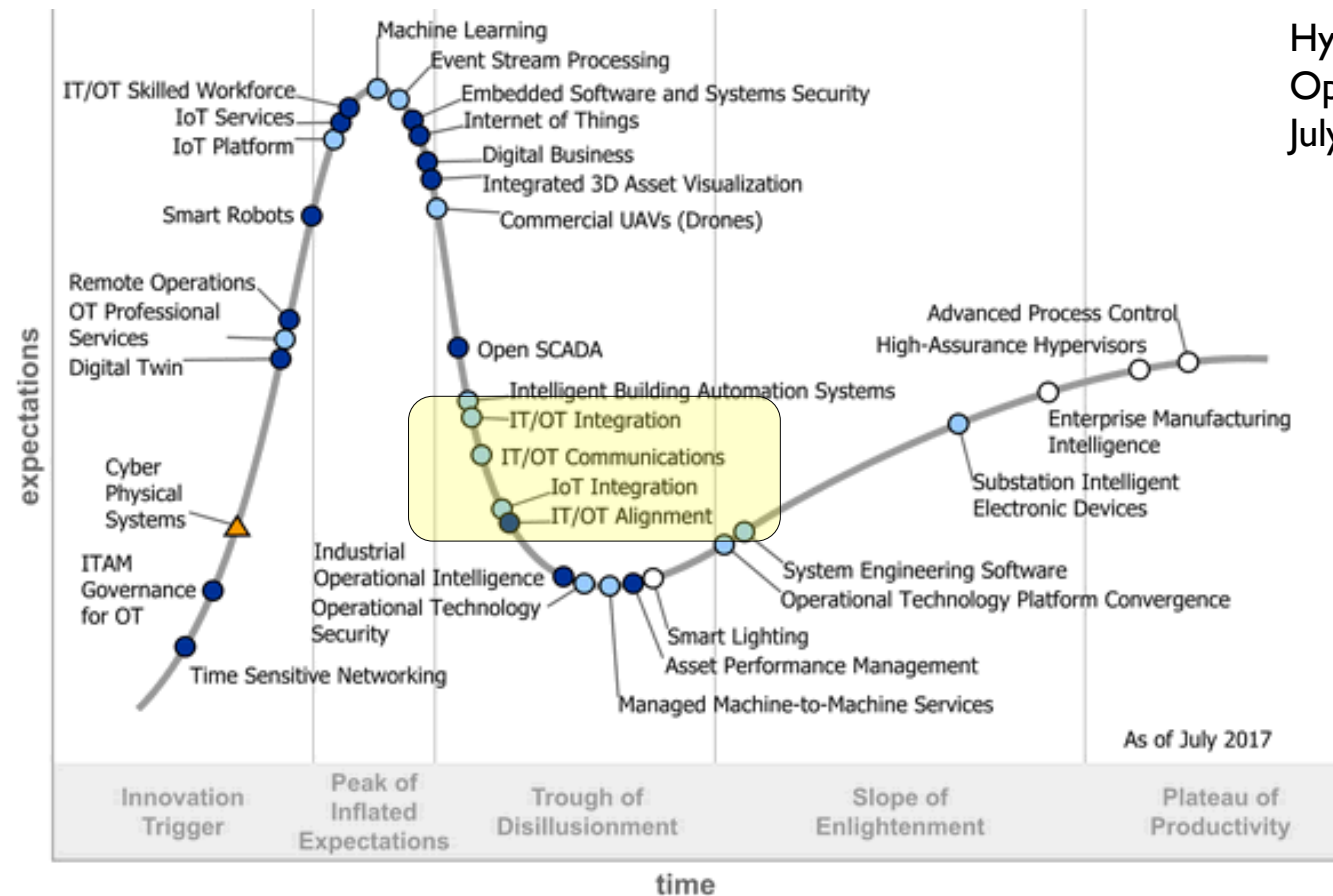
The Organisational Challenge



- It's more than one IT organisation
- And they work in siloes separated from each other

Hype Cycle for Managing Operational Technology, 2017, 25 July 2017

Gartner



Plateau will be reached:

- less than 2 years
- 2 to 5 years
- 5 to 10 years
- ▲ more than 10 years
- ⊗ obsolete before plateau

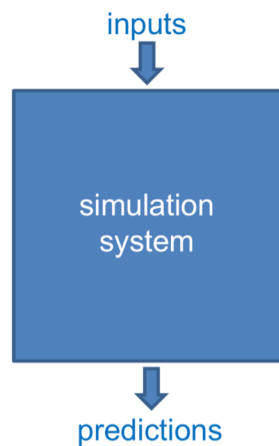
#StrataData

Strata
DATA CONFERENCE

The Analytic Challenge – Physics or Data driven?

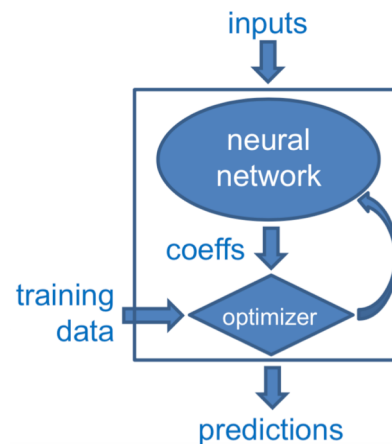


Physics-based
simulations



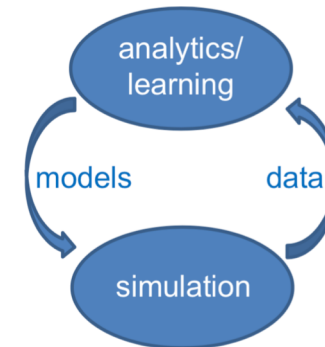
Simulations can't learn from observations

Data driven
learning



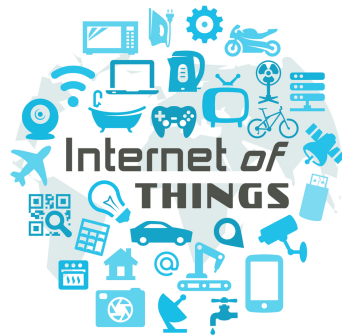
Data driven learning doesn't apply the laws of physics

Can we use them
together?



Can use observations when building models, use simulations to synthesise data or in feature engineering

Why the Digital Transformation Now?



Confused?? – Where do you start?



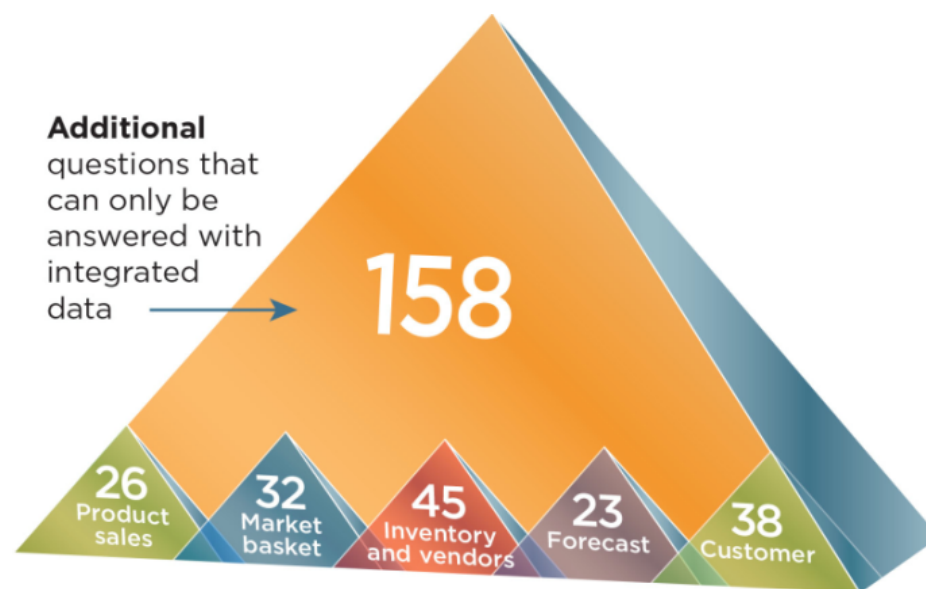
[This Photo](#) by Unknown Author is licensed under [CC BY](#)

Drive the architecture from the Business Requirements
(Experience shows us that you will fail if you don't!)

Integration – Bridge the Siloes

The academic definition of **integrated** is ‘to bring together or incorporate (parts) into a whole’

- Access and quickly integrate data outside of applications, to new sources to extract even greater value
 - Well operations linked to asset data in SAP
 - Daily drilling data linked to petrophysics data
 - Production data linked into SAP data for pricing

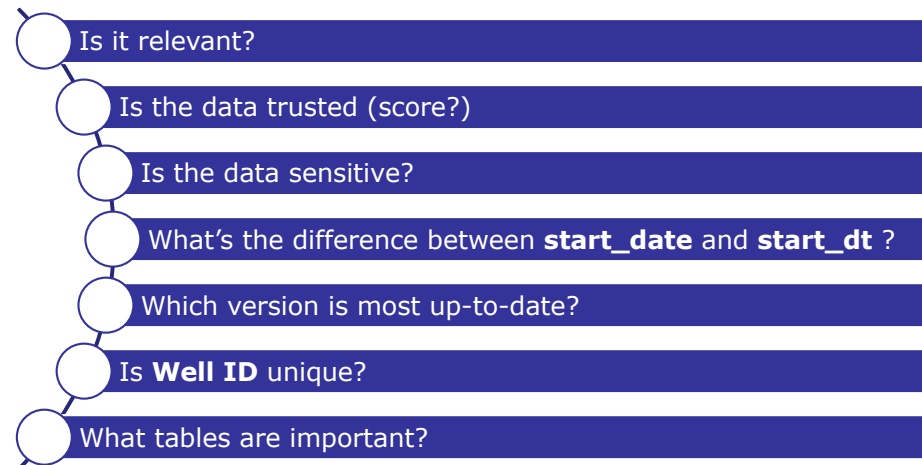


The Information Catalogue

An Information Catalogue is a **common repository** for all data and information related to Enterprise Data Management.



- **Inventory:**
 - What's there?
 - What's where?
- **Dictionary:**
 - What is this?
 - Type?
 - Meaning?



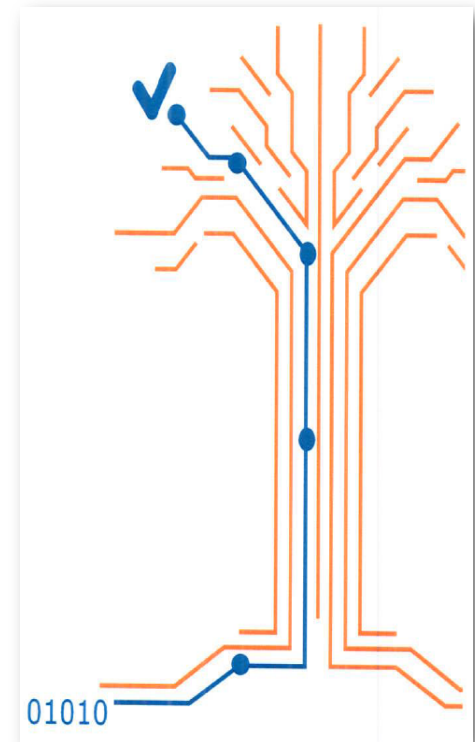
Basic Questions of Provenance

Who created the data asset and when?

- What is the source of the raw data used to create the asset?
- What processes were used to create the data asset?
- What are the known defects associated with the data asset?
- What algorithms were used to manipulate data?

Without provenance it is hard (sometimes impossible) to:

- Reproduce results
- Solve problems collaboratively
- Validate results with different input data
- Understand the process used to solve a particular problem
- Re-use the knowledge involved in the data analysis process



Source: Hansen, Johnson, Pascucci, and Silva. Visualisation for Data Intensive Science. The Fourth Paradigm. 2009 pp. 154;163
Slide Concept borrowed from Stephen Brobst – Teradata Universe 2015

Varying Data Quality Standards



- Highest provenance offering
- Well-defined governance program
- Minimises the risk of variable or inconsistent output
- Highest data protection and backup capabilities to protect from data loss



- Lower levels of provenance
- Subset of controls on data
- Still likely governed by a central body
- Not typically monitored and measured at the same levels as Gold



- Typically associated with user-defined data sets
- Potentially new, raw data feeds
- Can be elevated to Silver or Gold once the value and dependencies are understood

Data Placement Factors to Consider

Provenance

- Gold
- Silver
- Bronze

Integration

- Tightly Coupled
- Loosely Coupled
- Non-Coupled

Volume

- Extreme
- Moderate
- Low

Velocity

- Batch
- Streaming

Security

- Coarse Grain
- Fine Grain

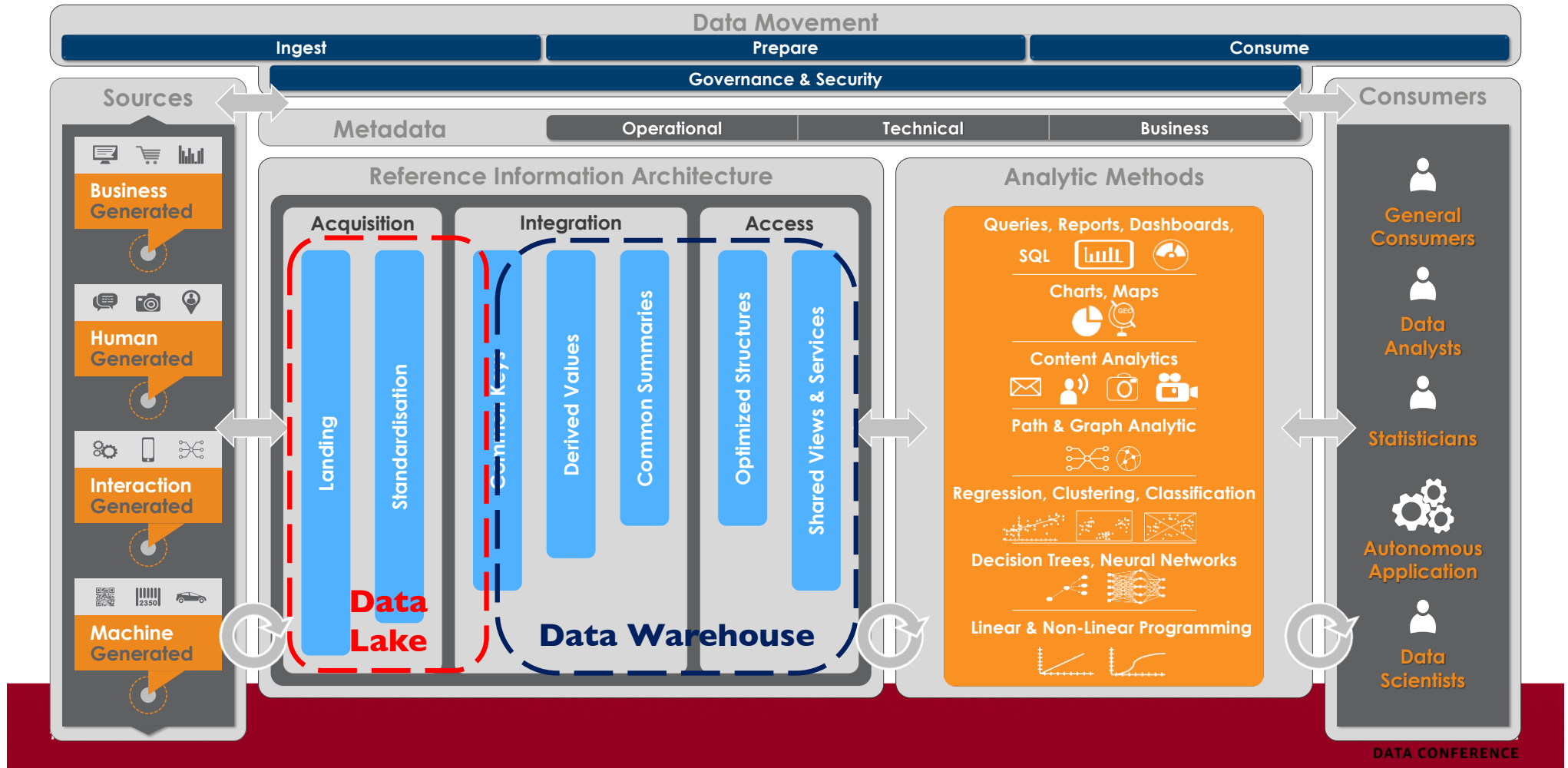
Schema

- Tabular
- Non-tabular

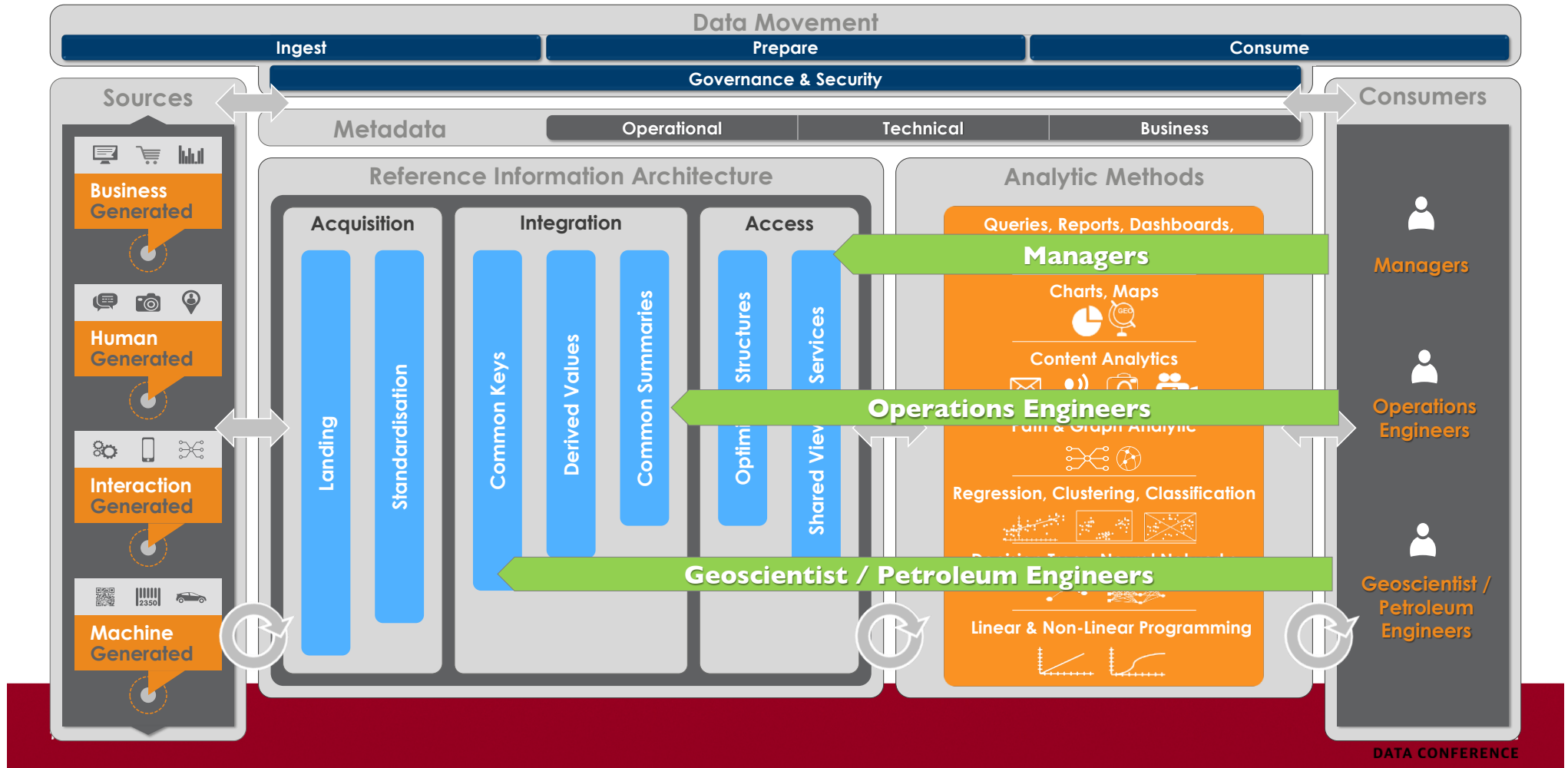
Information Classification

- Public
- Internal Business
- Confidential
- Secret

Reference Architecture



Business-Ready Data Products






Focus on the Data Management

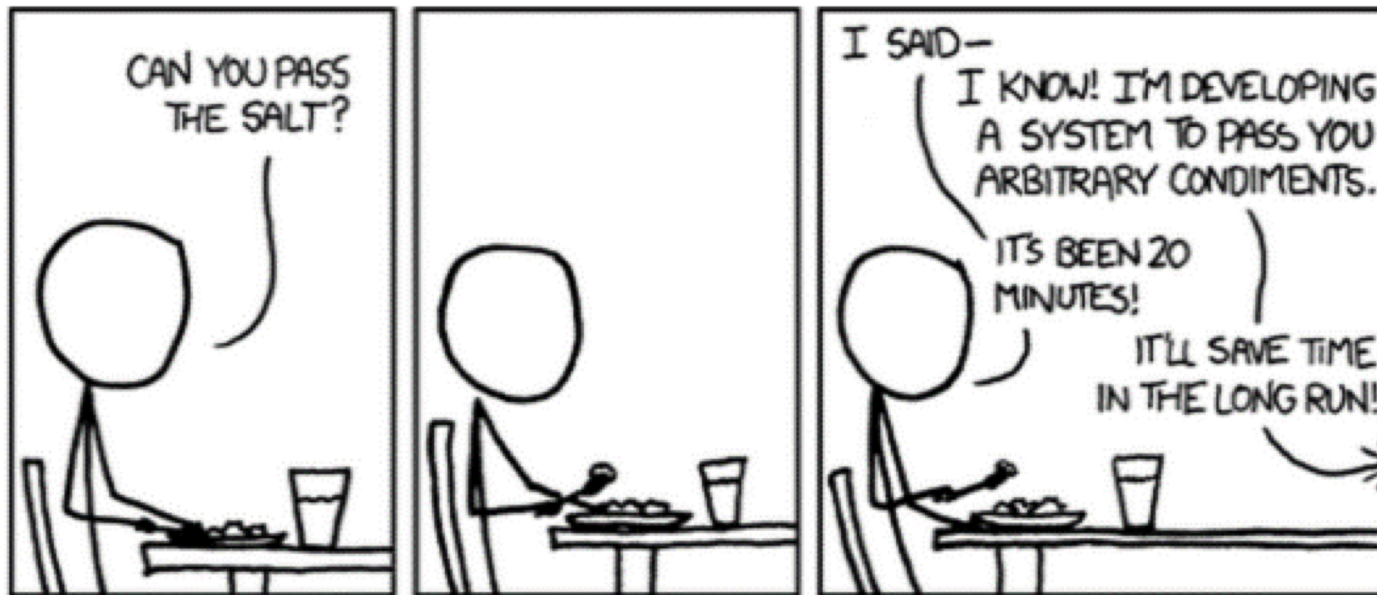
- It's all about the quality
- Degrees of Integration
- Master Data Management
- Data Provenance
- Metadata

Placement of data is important

- Volumes
- Velocity
- Information Classification

	Levels of data trust	Data integration	
	Certified	100%	
	Trustworthy	80%	
	Proven	60%	
	Experimental	40%	
	Raw/high risk	20%	

What NOT to do....



The General Problem. © xkcd.com

What you SHOULD do..

Iterate

Pick a business problem

Integrate Data

Deliver Often

Prioritise **knowing** the data quality vs perfecting the data quality

OK, maybe there is one
more problem...





Jane McConnell
Practice Partner O&G , Industrial IoT Group
Jane.mcconnell@teradata.com
+44 (0)7936 703343



My blog on [Teradata.com](https://www.teradata.com)



Follow me on Twitter [@jane_mcconnell](https://twitter.com/jane_mcconnell)



My [profile](#)

Paul Ibberson
Senior Architect (International Architecture CoE)
paul.ibberson@teradata.com
+44 (0)7803 231925



Follow me on Twitter [@paulibberson](https://twitter.com/paulibberson)



My [profile](#)