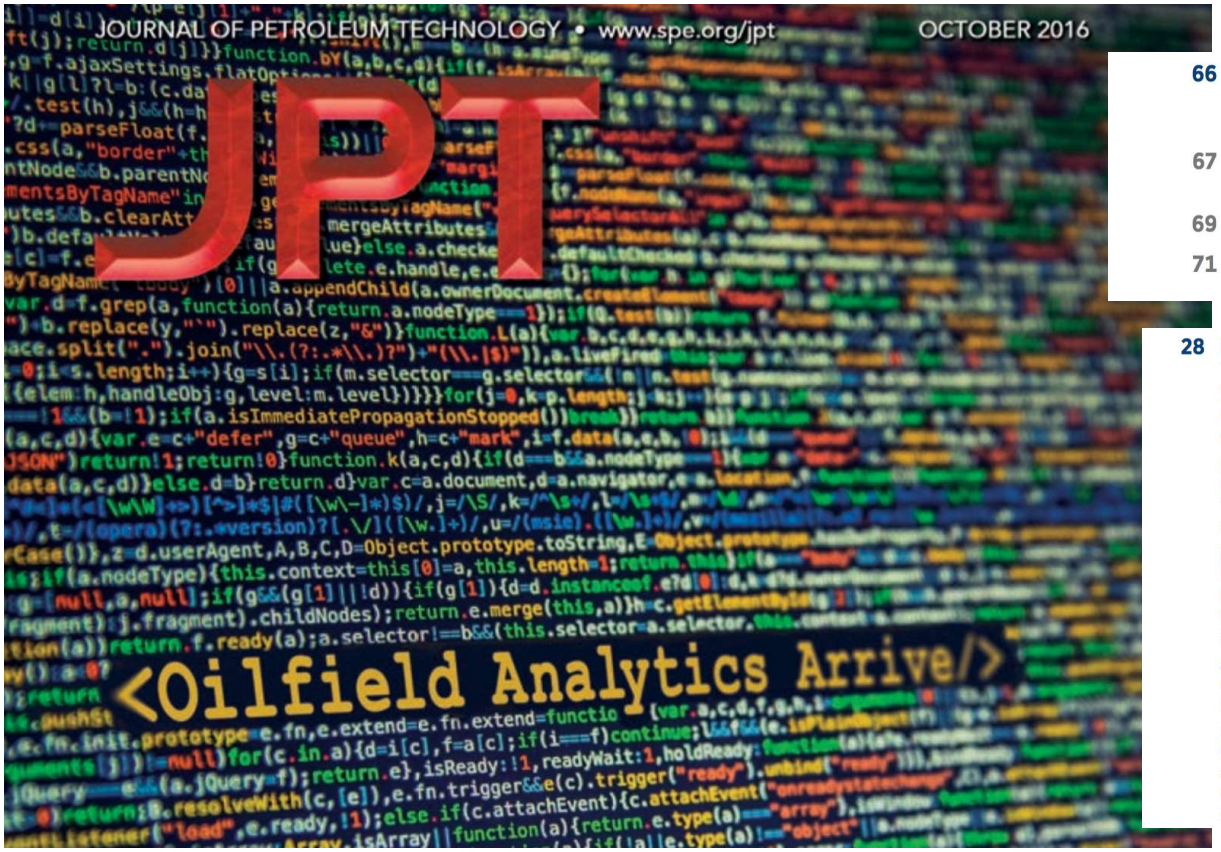
TERADATA

The future of Subsurface Data Management? Building a Data Science Lab Data Lake

Jane McConnell, Practice Partner Oil and Gas, Teradata

DEJ KL, 3 October 2017

Analytics and AI is gaining ground in Subsurface



...on the UKCS

OIL&GASUK

MEMBER LOGIN / REGISTER

ABOUT US | THE INDUSTRY | EFFICIENCY HUB | MEMBERSHIP | PUBLICATIONS | EVENTS | NEWS

Modern data science unlocks over 50 years of UKCS data

November 18, 2016

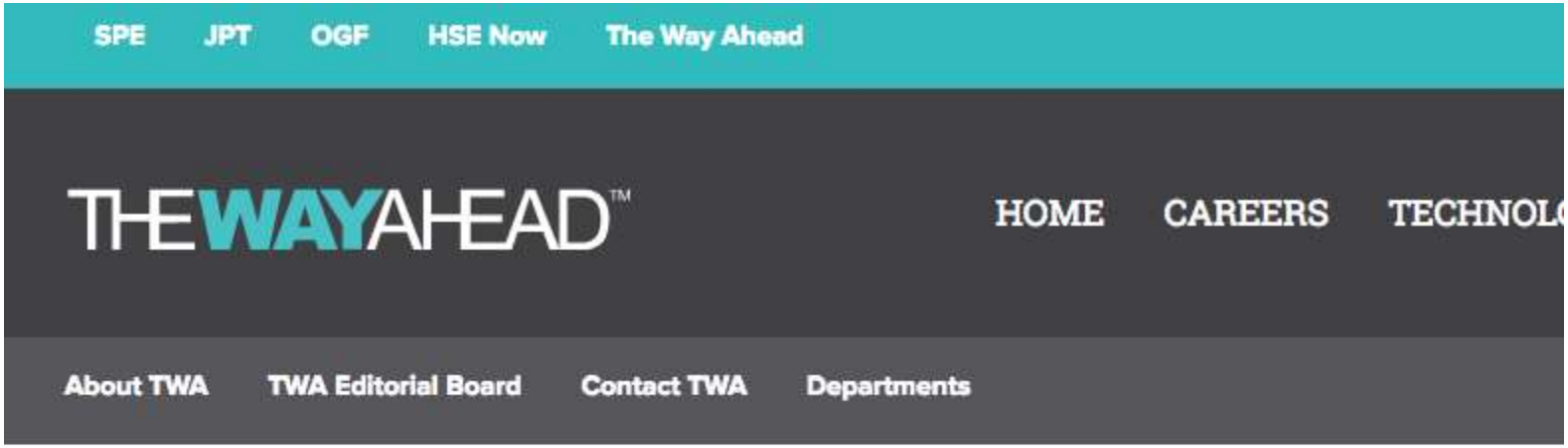
The results of the first Unstructured Data Challenge 2016 set by Common Data Access (CDA) to identify opportunities for extracting further value from UK North Sea exploration data will be shared at an industry workshop on November 30 at the Village Hotel in Aberdeen.

CDA, a subsidiary of Oil & Gas UK and provider of data management services for seismic and well information to the sector, issued the challenge in March. Its aim is to show how modern data and analytical techniques can yield valuable insights to assist industry efforts to maximise economic recovery from the UK Continental Shelf.

Nine companies: Agile Data Decisions; AGR Software; Cray Inc.; Flare Solutions; Hampton Data Services; Independent Data Services; KADME; New Digital Business; and Schlumberger Software Integrated Solutions took up the challenge. CDA gave the companies bulk access to more than 50 years of released data stored in its UKOilandGasData repository allowing them the opportunity to demonstrate how applying modern data science and data analytics techniques to sub-surface data sets could add value to current understanding of the subsurface.

TERADATA.

…with the younger generation



YP's Guide To...

Oil in the Digital Age: A Young Professional's Guide to Petroleum Data Science Organizations in Silicon Valley

…at the EAGE this year

Le grand hack!

June 13, 2017 / Matt Hall

It happened! The Subsurface Hackathon drew to a magnificent close on Sunday, in an intoxicatir coffee, and collaboration. It will take some beating.

EAGE 2017 - WS01: A REVOLUTION ?

14/06/2017 | Henri

Neural style transfer, github.com/jcjohnson/neural-style

Gram Gansslé

Illustration from Matt Hall keynote available here
GitHub for this mix of face image and seismic here

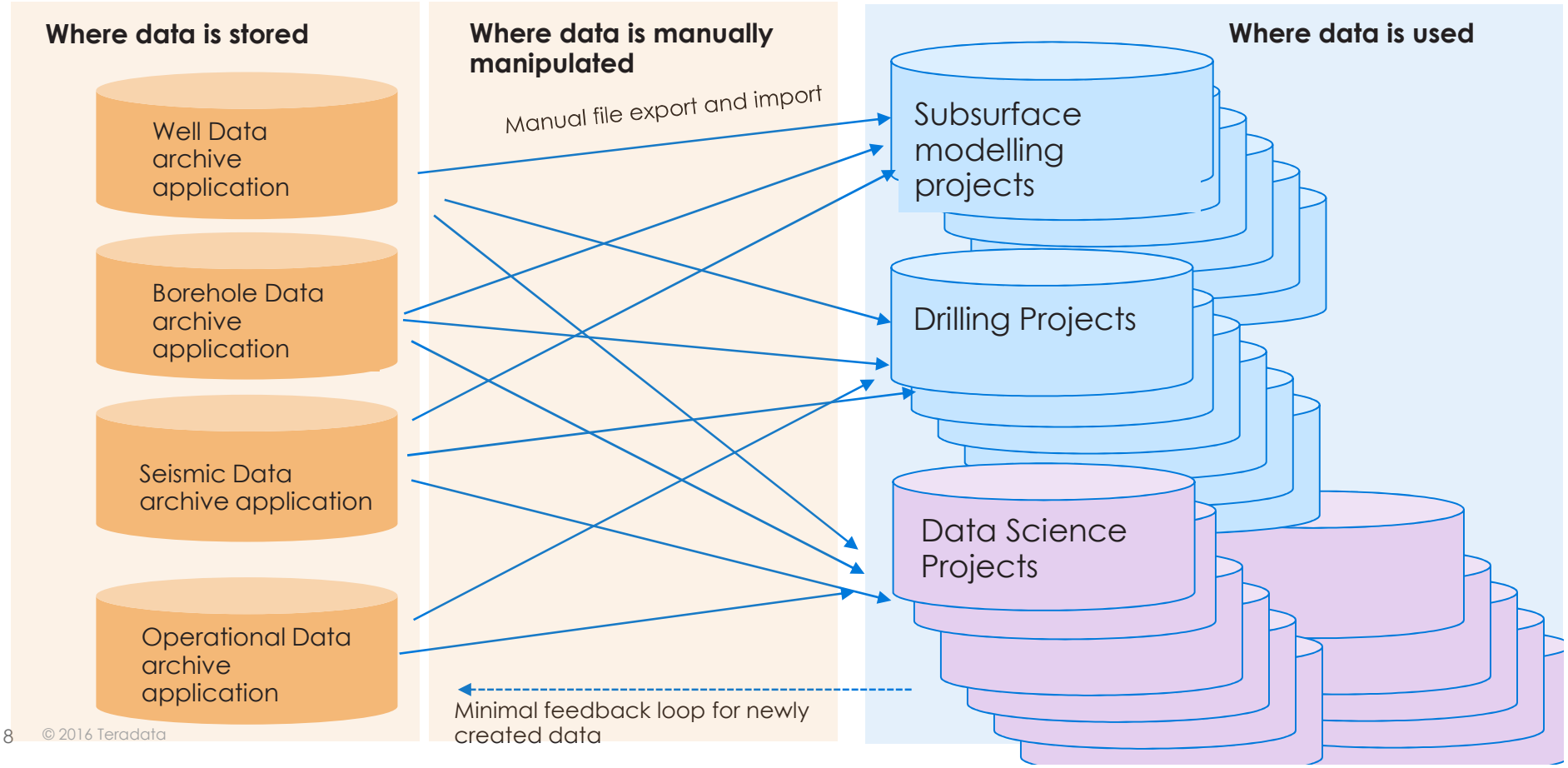
For the first time, EAGE organized a workshop about data sciences during its main annual event in Paris this year. We can measure the way done by the geoscientists for data science adoption in less than 2 or 3: our industry is going fast!
Remember, two or three years ago, very few data sciences papers were presented in conventions such as the EAGE, the SEG, the SPE or OTC and at this time, most of them were high level and philosophical papers about the 3 V and how our industry should embrace the Big Data (sorry for the authors of several exceptions to this rules!).

TERADATA.

The way people interact with subsurface data is changing.

The way we do data management needs to change too.

We can't keep doing this



But if we do nothing, it will be worse!

Some of what we have seen:

- Data associated with wrong well (or other entity)
- Units of measure not known or not consistent before combining data
- Incomplete historical data sets, no strategy to fill in the time gaps
- Data thrown into a machine learning model before it is understood

Leads to WRONG ANSWERS!

TERADATA.


Data lake, anyone?
© 2016 Teradata

What is a Data Lake?

Gartner

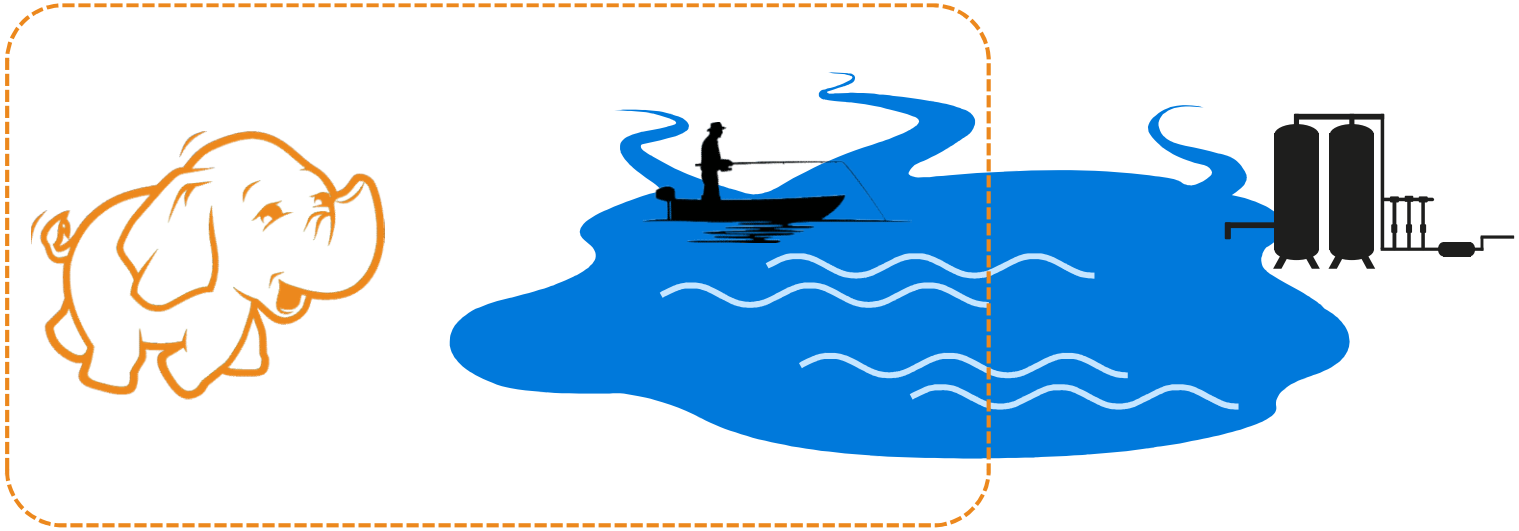
"A data lake is a collection of storage instances of various data assets. These assets are stored in a near-exact, or even exact, copy of the source format and are in addition to the originating data stores"

Source: Gartner, Three Architectural Styles for a Useful Data Lake

TERADATA.

Hadoop is more than a data lake.

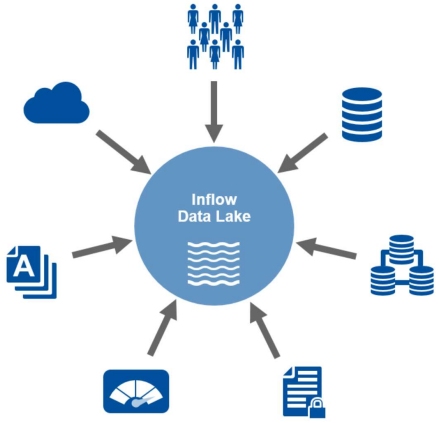
A data lake is more than Hadoop.



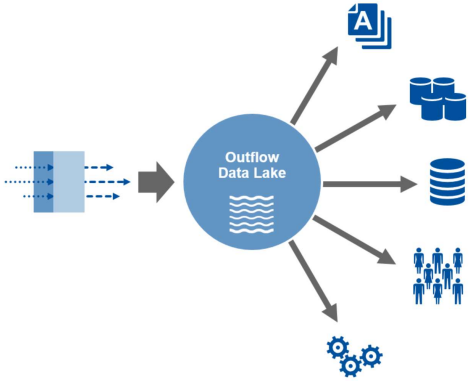
* Actually, a data lake is a design pattern, not a technology. A Data Lake doesn't need to use Hadoop at all. You could use an RDBMS, or a document store, or….

TERADATA.

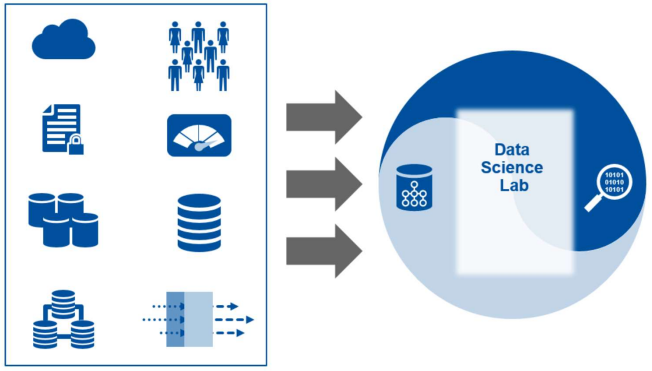
Gartner talks of 3 distinct styles of “useful” data lakes



Source: Gartner (July 2016)



Source: Gartner (July 2016)



Source: Gartner (July 2016)

In-flow Data Lake

A data hub, bringing together disparate sources of data. Close in use to a Data Warehouse, but built on cheaper technology for data infrequently used.

Substantial levels of data modelling required to make this easy to use for end users

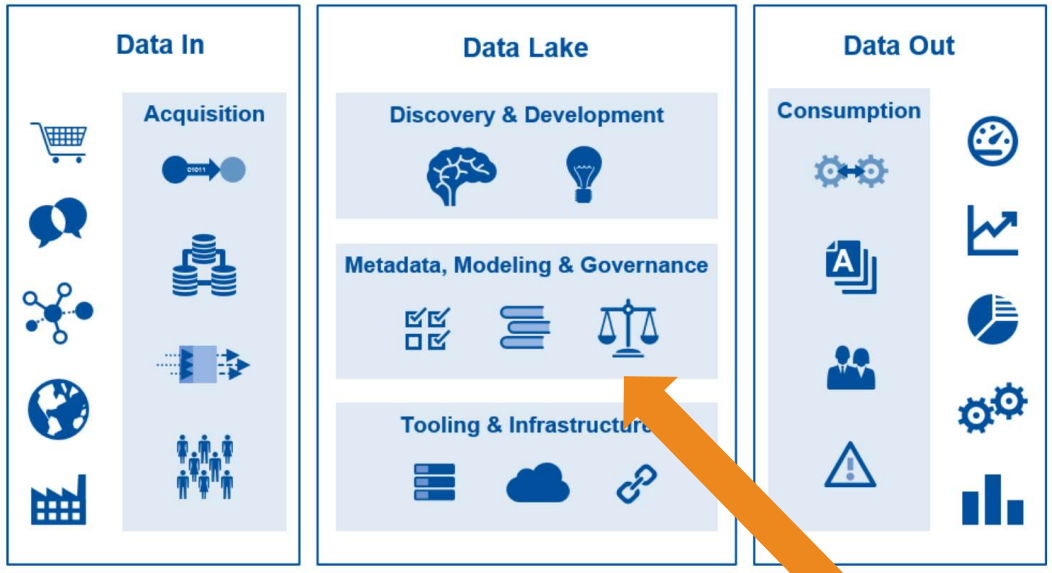
Out-flow Data Lake

Most commonly used for new-real-time data from operational data stores (historians etc) to manage the flow of new data in, and serve it up in the right format to the various systems that need that data.

Data Science Lab Data Lake

Best for enabling innovation. Little governance, just guard-rails. The data science lab lake is available only to a small, highly skilled user population of data scientists usually work side-by-side with specialists who can ask the right questions and interpret the outcomes

Avoiding a Data "Swamp"



Source: Gartner (July 2016)

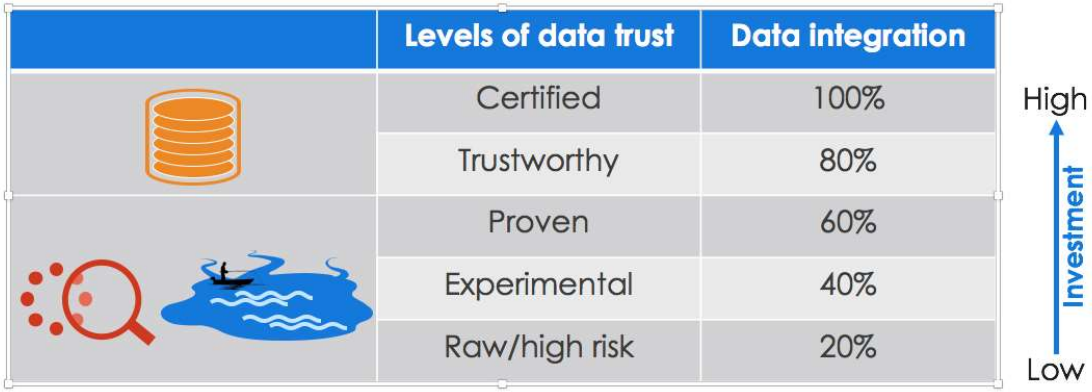
The need for metadata, modelling and governance does NOT go away

Data for Analytics: Define “Good Enough”

Prioritize Data Integration Labor and Investments

- Think about what level of data management is "good enough" for the task at hand

- Prioritise *knowing* the quality vs perfecting the data quality
- Think about the end user (and use) of the data, and the frequency of its use



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

High

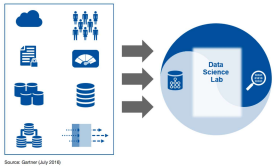
Investment

Low

TERADATA.

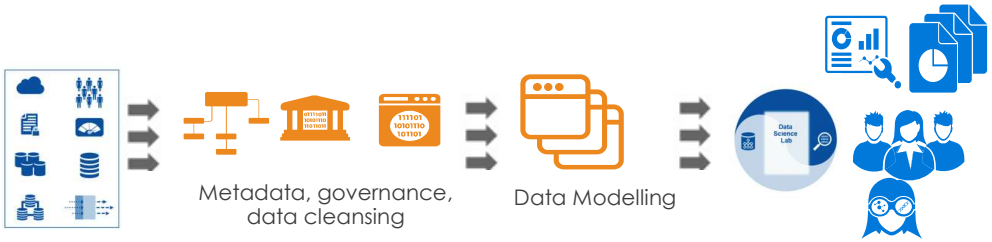
Define “Good enough” Data Management for your project

- Data lake to support a specific group of users eg petrophysicists for a one-off analytics project



- Data lake to support a specific group of users eg petrophysicists for all future analytics projects

- Data lake for wide use, with more data types supported



TERADATA.

How do you manage a data lake?

Automating Data Prep – Data Pipelines

Old World

- Data is manually loaded through "Import" options in petrotechnical software
  - Import procedure is fixed/proprietary and only perfect data will load
    - You need to change the data to match the software, rather than the other way round
  - Data loader must follow correct procedure
    - Human error can happen

New World

- Data is picked up from a directory and automatically ingested into the data lake through predefined data pipelines
  - Data manager defines the data flow
  - Data engineer builds the pipelines
  - Pipelines determine
    - How to parse or split files
    - What to index
    - What to load to databases
    - What quality checks to run
  - Pipelines automatically track lineage

TERADATA.

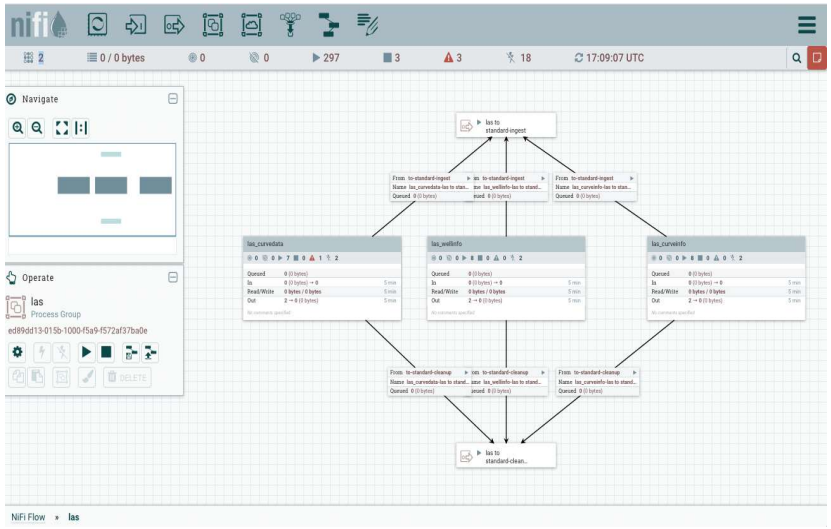
Use A Data Lake Management Software Platform

Kylo is a data lake management software platform and framework for enabling scalable enterprise-class data lakes on Apache Hadoop and Spark. Kylo is **licensed under Apache 2.0** and contributed by Think Big Analytics, A Teradata Company

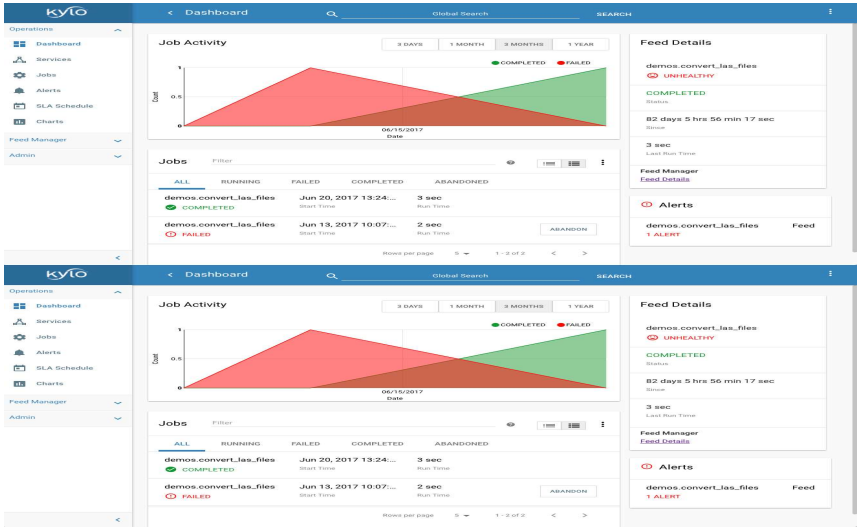
http://kylo.io

Here's one I built earlier – Data Pipeline for LAS

NiFi template to parse and load LAS files



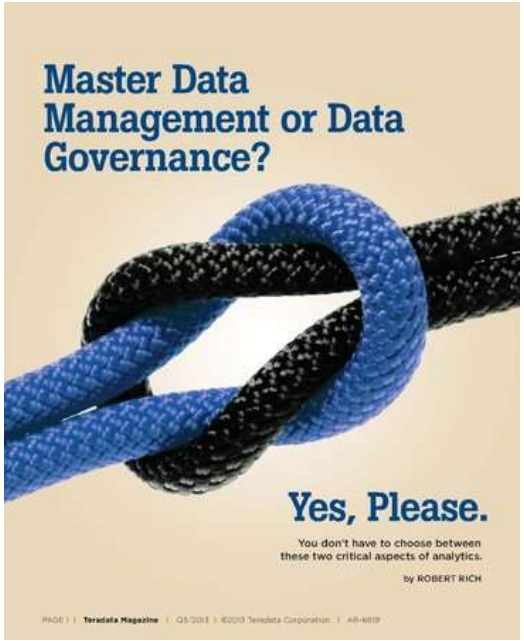
Installed as feed in Kylo to provide lineage



Some things always require a bit of human intervention

Master Data Management is one of those things

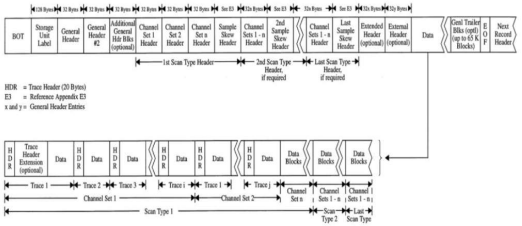
- It doesn't matter whether you are building a traditional data warehouse, or a data lake
- If you have data from multiple sources that you want to bring together, you need Master Data Management
- Master Data Management tools can support you in managing the Well List, the Curve Mnemonic List – whatever master and reference data you need
- Just get one.

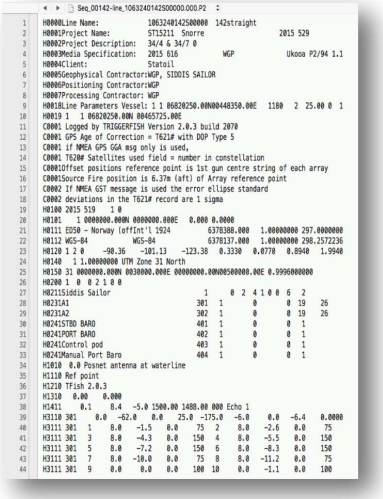


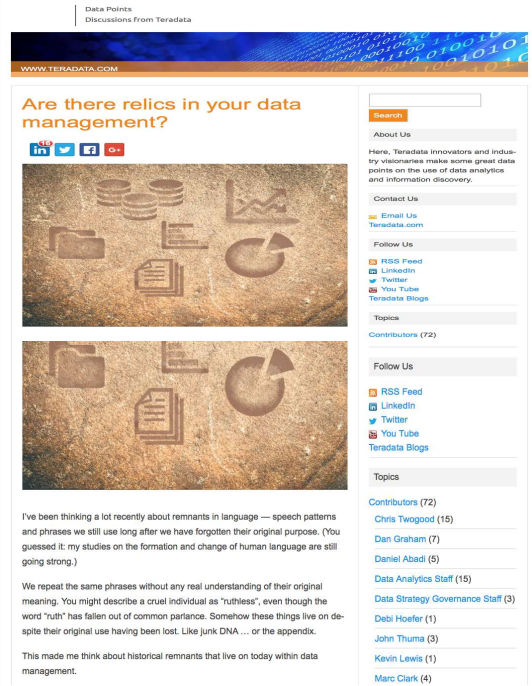
TERADATA.

OK, but isn't there just one more problem?

DVD Tape Disk

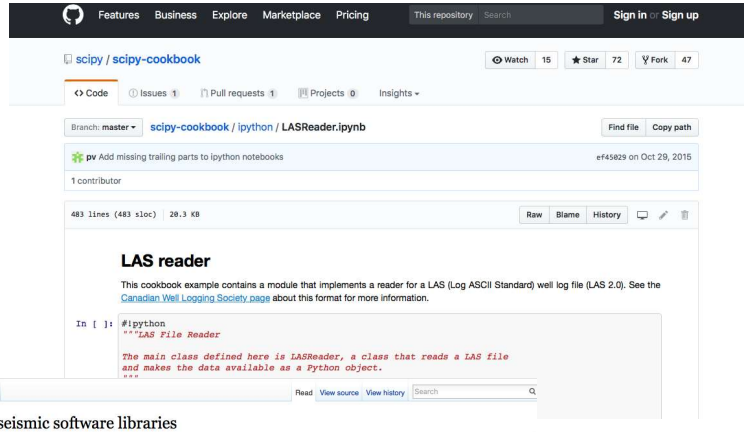


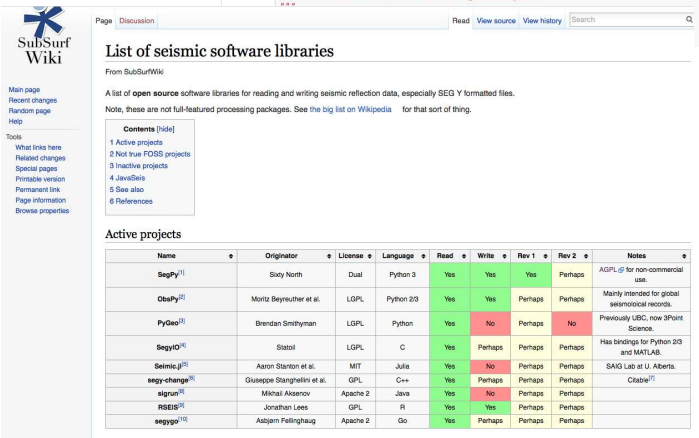




Open Source = Collaboration not obfuscation

- We have a strong community of Petroleum/Geo Data Managers
- There's a fair amount of open source code
- There are good standards in place
  - PPDM
  - WITSML
  - PRODML
  - RESQML





Open Source ("Free like Speech") data ingest

"

*Ability to access data should **not** be a competitive advantage*

"

* I can't remember who said it, but it was in the context of The Norwegian Model and DISKOS, back in the late '90s

TERADATA.

Working towards the O&G Managed Data Lake

Jane McConnell
Practice Partner O&G , Industrial IoT Group
Thing Big Analytics, a Teradata Company
Jane.mcconnell@thinkbiganalytics.com
+47 98282110

My blog on Forbes
My blog on Teradata.com
Follow me on Twitter @jane_mcconnell
My profile

TERADATA.