FLARE

Information Management

Increasing the ROI of your Data Lake

Dave Camden

DEJ London, 19-Nov-2018

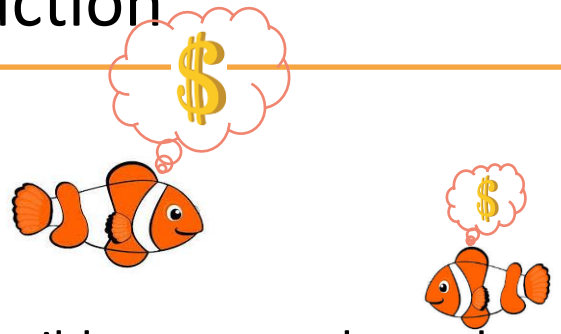
www.flare-solutions.com

“A Data Lake is a centralized repository that allows you to store all your structured and unstructured data at any scale.”

“A Data Lake is a system or repository of data stored in its natural format, usually object blobs or files.”

“A Data Lake is a place to store lots of data until you decide what to do with it…”

Data Lake Introduction



- Promise
  - Facilitate rapid, flexible access to huge datasets to support new value-creation analytical tools and methods
  - Remove data silos – a “data democracy”
  - Really cheap to load/store (schema-less, original format)
  - Enterprise scale
- Reality
  - A new technology, but familiar
    - it’s just a file system – and we all know what that can lead to
  - Not-so-cheap to use the data (finding, clean-up, schema-on-read)
  - Data quality, lack of skills, and governance issues
  - Are they manageable long-term?

Characteristics and Value Maximisation: An IM/DM Perspective

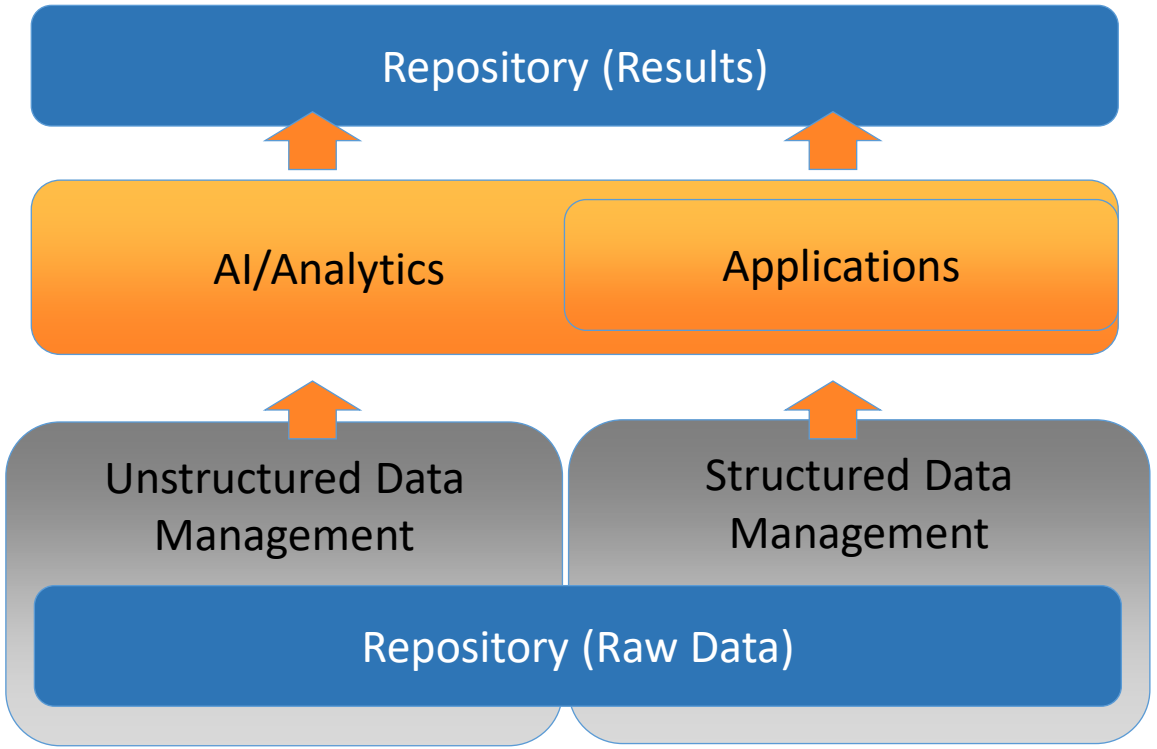
FLARE

New environment yet hauntingly familiar

Data and Analysis – supporting the technical professional

80% of many analytics projects are clean-up:

Good data management is critical



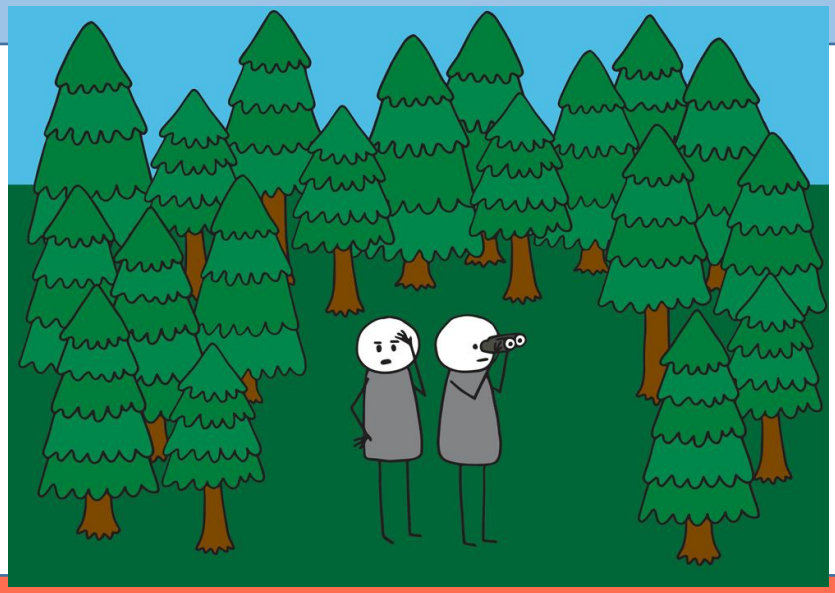
Good data management is critical

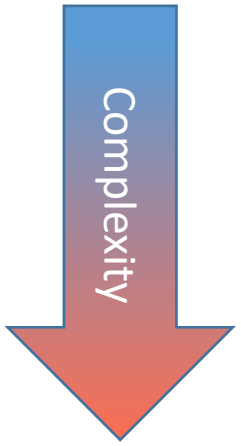
We still need to clean-up and organise our information – be it a database, data lake, doc-base or shared drive

Manage Complexity - Avoid Information Overload

- Think at Different Levels

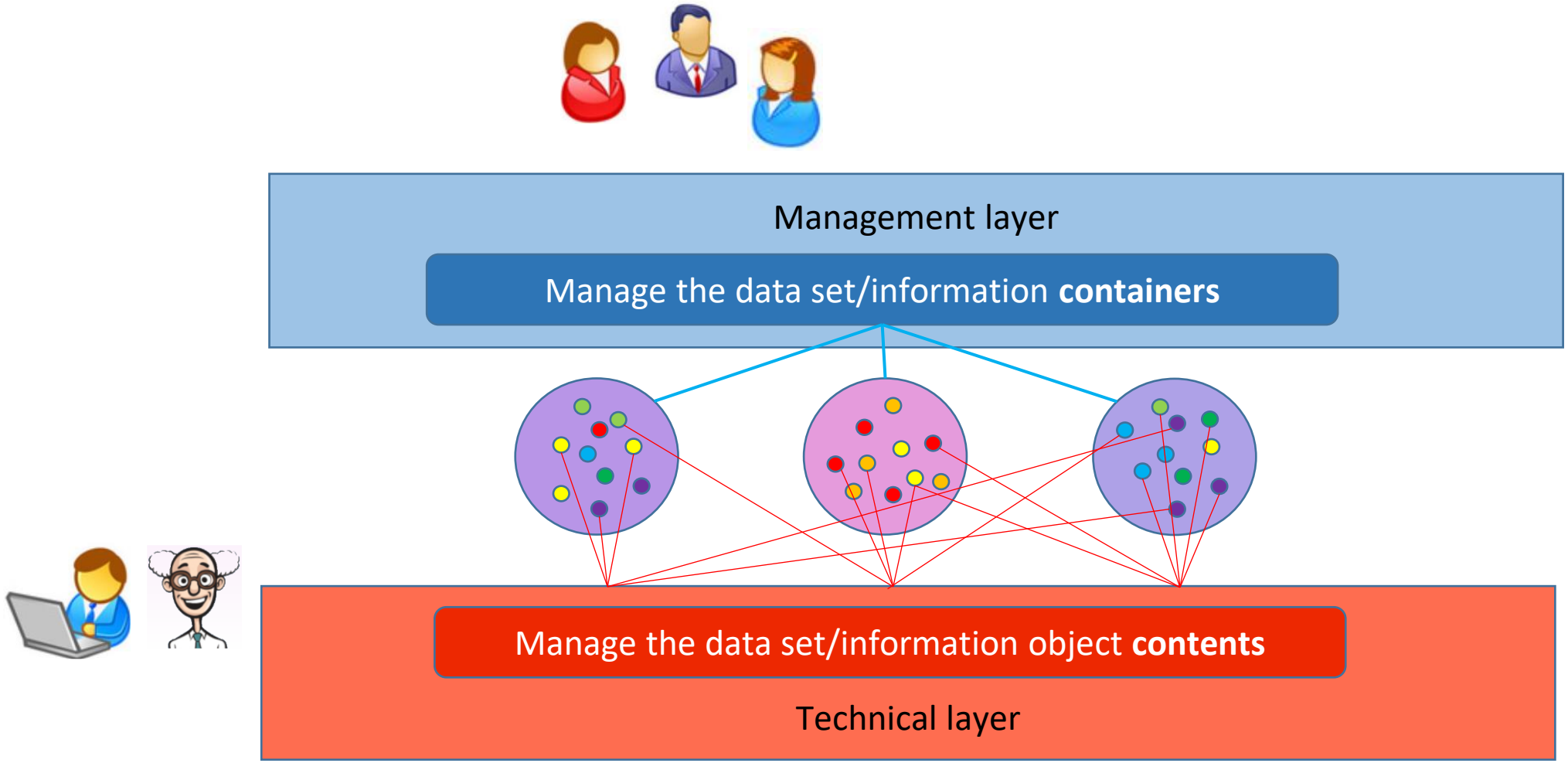
  - The woods – the Enterprise level – information catalogues, BI dashboards, wide visibility, federated searches





  - Trees - Deep down technical level – data catalogues, attributes, schema mappings, processing flows, detailed searches

Manage Complexity - Avoid Information Overload



Start Small and Grow

- Pilot projects – sandbox environment, prototyping, data exploration
- Development projects – mature techniques, develop work/dataflows, standards and governance
- Production environment – automated work/data flows, BI outcomes, widely searchable

Horses for Courses

It's not a 'one-size-fits-all' model

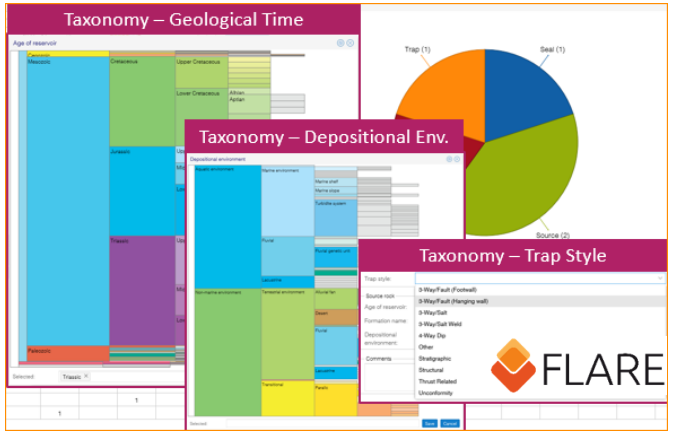
- Data Lakes may be hybrid constructs:
  - Unstructured and semi-structured data : original format files, schema-on-read
  - Structured time or depth series data: 'traditional' data stores
  - Hybrid data flows
    - Data Lake sandbox to feed Enterprise Data Warehouse (EDW)
    - Pre-cleaned, enrich with schema (AVRO, JSON, ORC…)

- All in one basket?
  - If you have 4 or 5Pb in one repository – no matter what it is – it's physically very hard to move it given current bandwidths and the annoyingly stubborn speed of light 

- A Data Lake is not for everyone
  - For some IoT and depth/time series analytical applications traditional EDW or relation databases may be more appropriate

Standards

- Consistent, standard metadata are essential to allow searching and data recovery at high and low levels
  - Standard industry formats like WITSML
  - Industry models for metadata reference values
    - For named entities (wells, fields, rigs, organisations etc) and relationships between them
    - Semantic taxonomies (common ‘language’) for information and data types
- Data Quality processes and standards
  - Data quality standards



An Interesting Read on Data Quality…..

Tom Redman

Harvard Business Review

If Your Data Is Bad, Your Machine Learning Tools Are Useless

by Thomas C. Redman

APRIL 02, 2018

SUMMARY SAVE SHARE 12 COMMENT TEXT SIZE PRINT $8.95 BUY COPIES

https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless

Governance

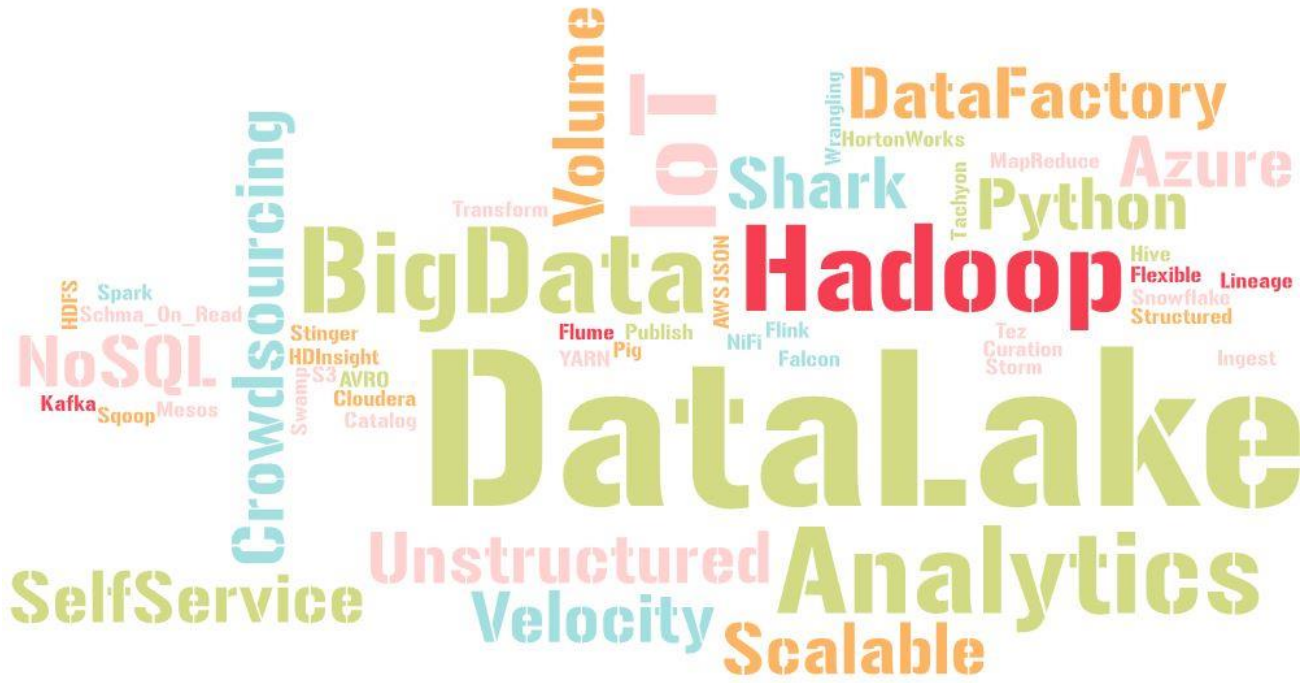
- High-level sponsorship and accountability
  - Provide feedback on ROI/results of projects

- Clear Data Lake management strategy and awareness
  - Ensure defined and implemented
    - Data ownership (authorisation)
    - Security model (access to content and metadata)
    - Metadata standards
    - Data loading, stewardship and delivery processes
    - Data quality standards and procedures

The Human Perspective

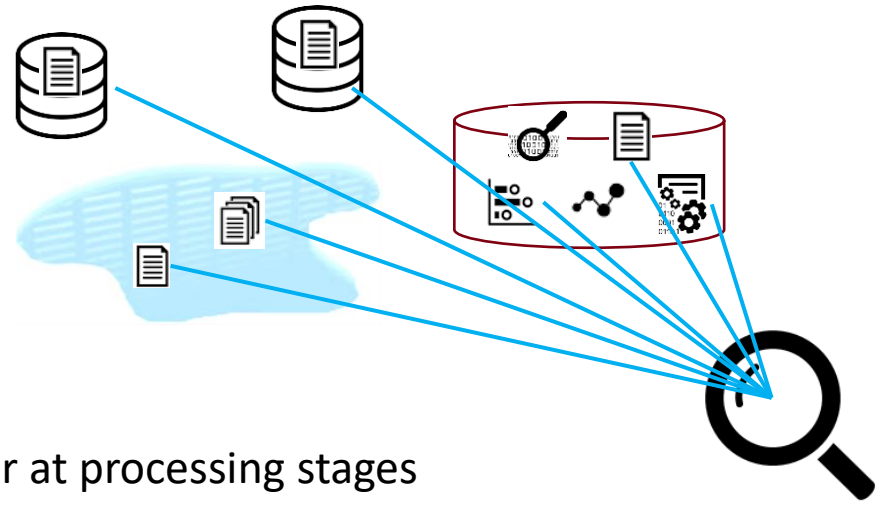
- Skills Shortages
  - Data-Information managers
  - Data scientists
  - IT (varying technology stacks)
- Culture
  - Bridging the IT-Business divide
  - “E&P is different”
  - Data are often with contractors – how to share it?
- Hype
  - free lunches for all

A complex, unfamiliar, dynamic environment
With lots of new buzzwords and ideas



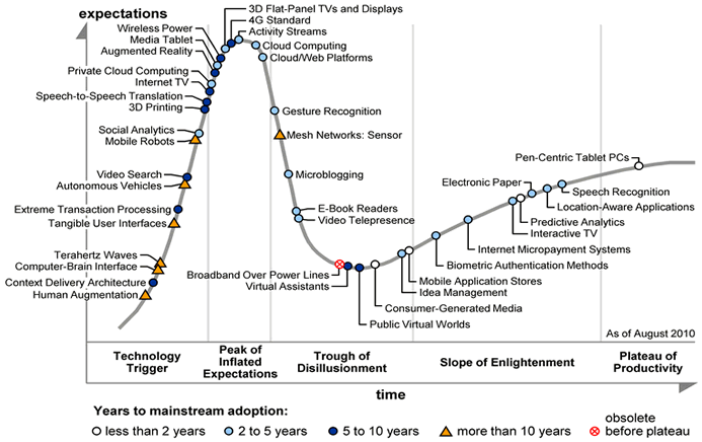
Information Integration

- If Data Lakes will hold significant amounts of persistent data, they should be seen as part of a company's Information Ecosystem
  - All data are somewhere
  - Some will be in the data lake
  - Need a consolidated view to avoid
    - Data silos
    - Duplication
    - Nearly the same data – but with different quality or at processing stages
    - Lost data



How Will it all Pan Out for Data Lakes?

- The hype cycle may have peaked



- Data Lakes are still in their infancy in the oil and gas industry – it's a journey
- The Jury is still out : will Data Lakes become a key component of our information infrastructure or have a limited role supporting analytics solutions?

d.camden@flare-solutions.com
+44 7703 234 891
+44 1892 875 007

www.flare-solutions.com

Flare Solutions Limited
3, Acorn Business Centre, Northarbour Road,
Cosham, Portsmouth, PO6 3TH,UK

Europe:
Tel: +44 1628 482 750
Fax: +44 8704 602 543

North America:
Tel: +1 403 932 4597
Fax: +1 403 932 6156

Email:
enquiries@flare-solutions.com