

The background is a solid teal color with a complex pattern of lighter teal hexagons and interconnected nodes and lines, resembling a molecular or network structure. The text is white and positioned in the upper left quadrant.

# ***Big Data and New Paradigms in Information Management***

**Vladimir Videnovic**  
**Institute for Information Management**



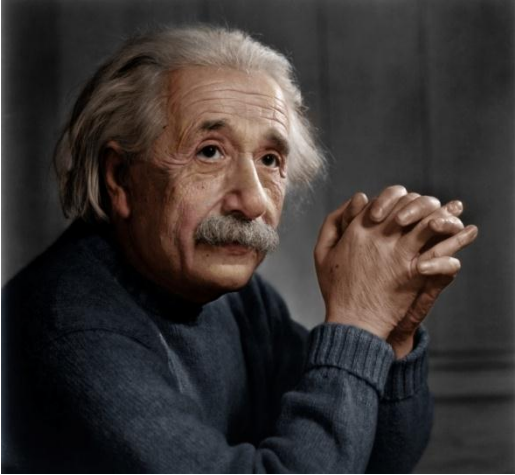
*"I am certainly not an advocate for frequent and untried changes ... laws and institutions must go hand in hand with the progress of the human mind. As that becomes more developed, more enlightened, as new discoveries are made, new truths disclosed manners and opinions change with the change of circumstances, institutions must advance also, and keep pace with the times..."*



**Thomas Jefferson, July 12, 1816**

# Introduction

## What is Big Data?



“If you can't explain it simply,  
you don't understand it well enough”

Albert Einstein

# Big Data

Term commonly used to describe phenomenon of  
large and/or complex datasets  
caused by the exponential growth and availability of data  
traditional data processing approaches cannot handle effectively



# When Data becomes Big Data?

**Volume, Velocity, Variety,**

level of structure and/or complexity of Data

exceeds organisations' abilities to

understand, ingest, store, process and analyse it in an efficient way.

new data sources, data types, data capture and storage mechanisms

new approaches to data analysis and realising its business value



# Big Data – Big Vs

- Volume
- Velocity
- Variety
- Visibility
- Vulnerability
- Veracity
- Variability
- Value
- V...
- Complexity
- Immaturity
- Scalability
- Performance
- Availability
- Manageability
- Trustworthiness
- Cost

# From Data To Decisions

Organisations today depend on their ability to realise potential of increasing volumes and complexity of **Data**, in order to transform it into valuable **Information** and derive actionable insights and **Decisions** in a timely manner



# Promise of Analytics

## An **Information-Powered Organisation**

is equipped with timely insights and knowledge to continuously optimise the way they

- derive answers, decisions and actions
- conduct their business
- streamline their operations
- deliver services and/or products
- adapt to new realities










# Why Big Data?

- make information transparent
- solve new business problems
- determine root causes of business process challenges, defects and inability to deliver desired outcomes
- provide valuable insights to increase efficiencies, minimise risks and improve outcomes
- support better products, services, policies by enabling optimised decisions
- deliver next generation of products and services



Organisation	Problem
 <p><b>Australian Government</b> Australian Taxation Office</p>	<ul style="list-style-type: none"> <li>• Use data in a smarter way to improve decisions, services and compliance</li> <li>• Develop difference-making insights, which proactively informs decisions that shape the tax and superannuation systems</li> <li>• Act with Agility and flexibility in driving continuous innovation</li> <li>• Make cohesive insight-led decisions to better achieve the ATO's goals</li> </ul>
	<ul style="list-style-type: none"> <li>• Collect criminal intelligence and combines it with intelligence from partner agencies to create a comprehensive national view of serious and organised crime</li> <li>• Securely store, retrieve, analyse and share criminal information and intelligence</li> </ul>
 <p><b>Australian Government</b> Australian Transaction Reports and Analysis Centre</p>	<ul style="list-style-type: none"> <li>• Analyse financial transaction data using sophisticated modelling and predictive analysis tools to identify money laundering and terrorism financing</li> <li>• 14,000 reporting entities across Australia</li> <li>• Large volume of financial transactions</li> <li>• Social-Network Analysis</li> <li>• Multi-structured data</li> </ul>
 <p><b>Australian Government</b> Bureau of Meteorology</p>	<ul style="list-style-type: none"> <li>• Store large volumes of unprocessed data</li> <li>• Unified datastore for disparate data</li> <li>• Single query over a range of data sets from disparate sources (sensors, manual data input, xml, email)</li> </ul>
 <p><b>Australian Government</b> Department of Social Services</p>	<ul style="list-style-type: none"> <li>• The Analytic Data Platform to enable effective Data acquisition, storage and analysis</li> <li>• flexible granular view of the data</li> </ul>

# Big Vs Survey






**Not such big data!**

---



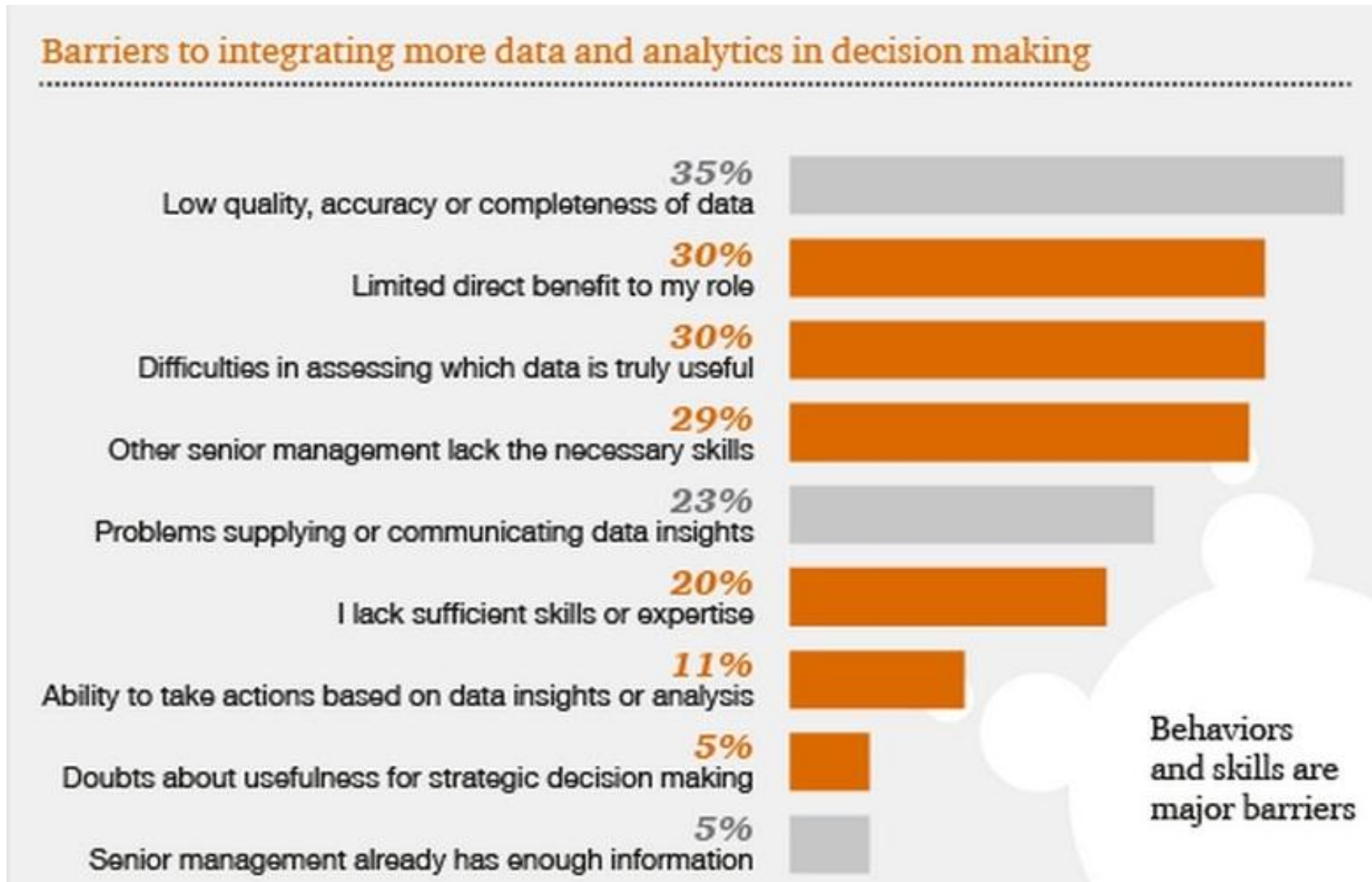
Work with data sets  
under 10,000 records

Source: Ernst & Young - *Global Forensic Data Analytics Survey 2014*

Organisation	Volume	Velocity	Variety	Visibility	Vulnerability	Veracity	Value
 <b>Australian Government</b> <b>Australian Taxation Office</b>			✓	✓	✓	✓	✓
		✓	✓		✓		✓
 <b>Australian Government</b> <b>Australian Transaction Reports and Analysis Centre</b>	✓	✓	✓		✓	✓	✓
 <b>Australian Government</b> <b>Bureau of Meteorology</b>	✓			✓			✓
 <b>Australian Government</b> <b>Department of Social Services</b>				✓	✓	✓	✓

# Barriers to Use Analytics for Making Decisions

## Survey



# Big Data Vs **Business Perspective**

## Value

- Enable smarter decisions, faster actions and optimised results

## Valuation

- Evaluation of outcomes and continuous optimisation
- Measure the impact on key performance indicators
- Assess the ability to forecast future outcomes

## Visibility

- Discover important insights – demand for good quality insights
- Study patterns, behaviours, trends and relationships
- The demand for data is highlighting the IT bottleneck – the ‘cash’ data economy



# Big Data Vs Information Management Perspective

## Variety

- Data Integration
- New data sources (Social Media, Data streams, Internet of Things)
- Lack of standards for managing various data types

## Velocity

- Data streaming, machine/sensor generated data
- Real-time insights

## Vulnerability

- Information Security
- Fraud discovery



# Big Data Vs Data Science Perspective

## Value

- Understanding what is possible
- Understanding the interestingness of data to make investment
- Prioritising the opportunities
- Analytics that doesn't support decision doesn't add value

## Variety

- Internal document/text/data sources may be inaccessible
- Need to prototype/test value add in adding new unstructured data from social media and other external sources

## Velocity

- Move to real-time discoveries and fraud detection





# What is a Data Scientist?



Programmer



Statistician



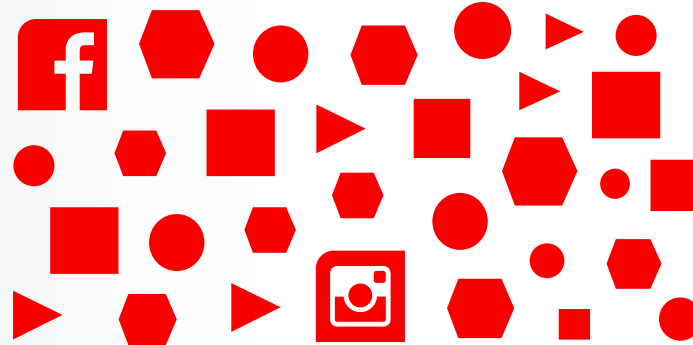
Business Analyst

# Need

## Analytic Value at Fast Pace

### Data Uncertainty

- Not familiar and overwhelming
- Potential value not obvious
- Requires significant manipulation



**80% effort typically  
spent on evaluating  
and preparing data**

### Tool Complexity

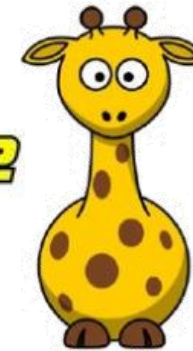
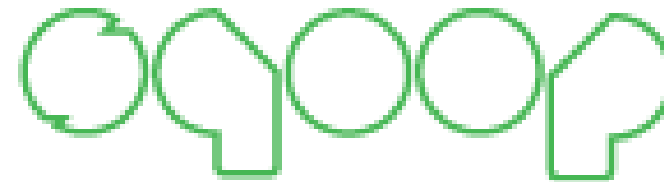
- Early Hadoop tools only for experts
- Existing BI tools not designed for Hadoop
- Emerging solutions lack broad capabilities



























**Overly dependent on  
scarce and highly  
skilled resources**

# Big Data Eco system

## The Hadoop Zoo



# Cloudera Hadoop Distribution

	 Hosts
	 HBase
	 HDFS
	 Hive
	 Hue
	 Impala
	 Oozie
	 Solr
	 Spark (Standalone)
	 Sqoop 2
	 YARN (MR2 Incorporated)
	 ZooKeeper

**HDFS** – Storage - distributed file system that provides high-throughput access to data

**MapReduce** – Process - framework for performing distributed data processing

**Hive** - Data Warehouse system for Hadoop

**HCatalog** - metadata abstraction layer for referencing data

**HBase** - Hadoop DataBase - distributed columnar database

**Hue** – user interface for managing and using Hadoop components

**Spark** - parallel data processing framework for developing Big Data applications

**YARN** – cluster management – reliable distributed processing of very large data sets

**Oozie** – workflow - coordination system for managing Hadoop jobs

**ZooKeeper** - centralised service for maintaining configuration information

**Sqoop** – efficient transfer of bulk data between Hadoop and structured datastores

**Impala** - analytic database designed to leverage the flexibility and scalability of Hadoop

**Solr** - reliable, scalable and fault tolerant search based on Lucene

**Pig** - platform for analysing large data sets



# People

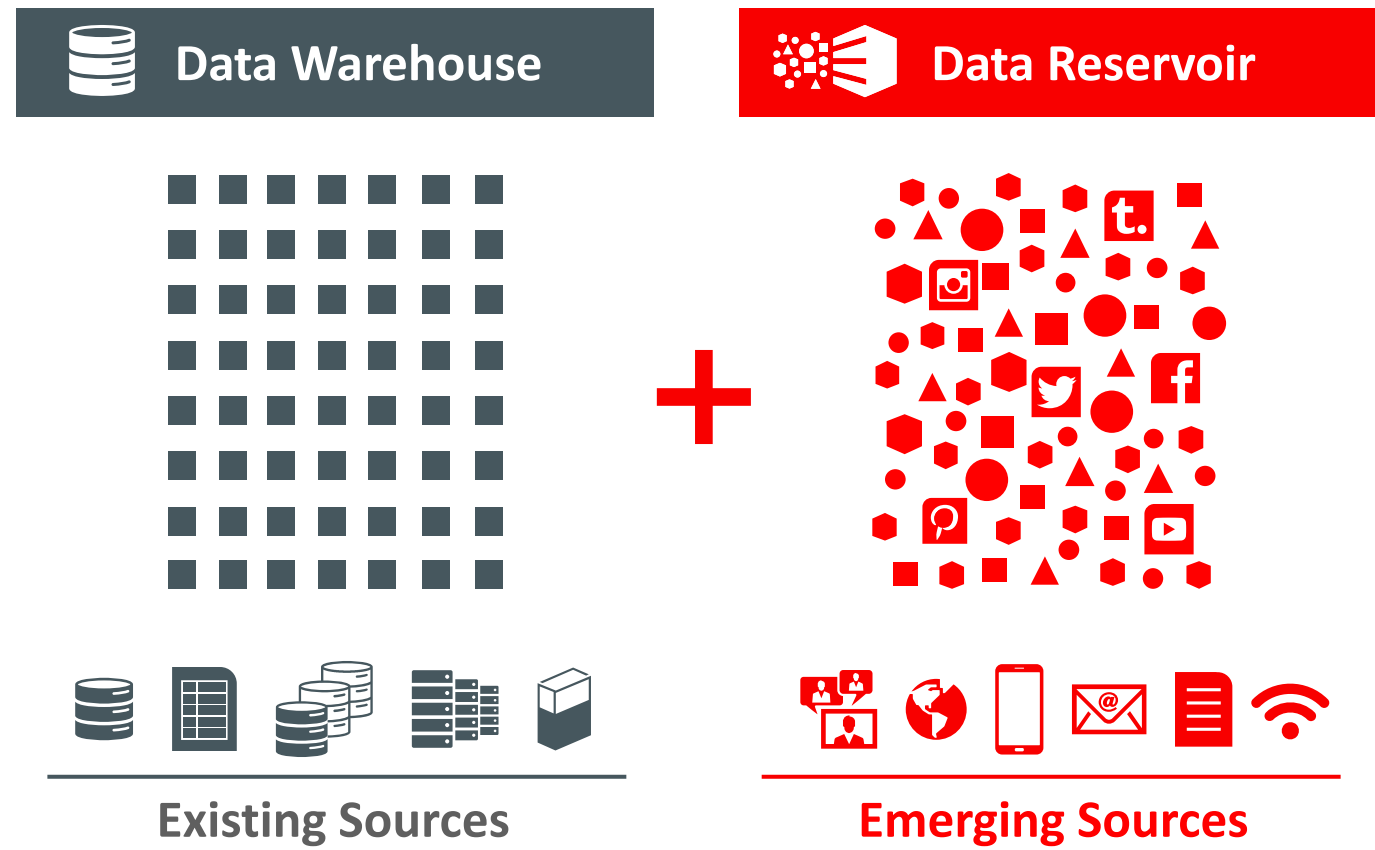
## Core Skills

- curiosity
- intellectual agility
- statistical fluency
- research stamina
- scientific rigour
- ability to visualise
- skeptical nature



# Data Reality

Existing sources + Emerging Sources



# Information Warehousing

approach that allows organisations to develop capabilities required to

- **acquire** relevant **Data** from multiple sources,
- **organise** it in a repository,
- **convert** it into **Information** by processing and analysing it and then
- utilise it to **derive** Intelligence to make smarter **Decisions**



# Big Information

Big Information drives organisations to the next level of success, assisting to:

- create more value from data
- generate opportunities for innovation
- help identify new services, policies and ways to deliver them more efficiently
- enable more effective delivery of programs across a range of government initiatives and operations





# Big Information

## Sample Use Cases

- Security Threats – relationship linking and behaviour analysis
- Crime Discovery – exploration of unrelated datasets in search for patterns
- Fraud Analysis – search for outliers and new patterns from millions of transactions
- Health – diagnostics and research of diseases based on analysis of genes
- Water Management – predictive maintenance based on sensor data
- Better Engagement with Citizens – sentiment analysis based on social media content



# Big Information Management

## End Goal – Big Data Roadmap (Example)



# Big Information Management

## Capable Organisation

### Elements of a Capable Organisation

Systems			Strategy		People		Infrastructure/ resources		Market	
1 Cm Communication	2 Lo Organisational Learning	3 Gv Governance	1 Vn Shared Vision	1 Mn Mission	4 Ld Leadership	1 Kn Knowledge	1 It ICT	1 Eq Equipment	5 Cu Customers	6 Pr Products
1 Co Collaboration	1 Pl Planning	1 Pm Performance Management	1 Ou Outcomes	1 Hy Hygiene Factors	2 Tr Trust	2 Ig Integrity	1 Fc Facilities	1 In Data & Information	Cl Clients	1 Sv Services
1 Pr Procedures	1 Si Simulation	1 Ch Manage Change	1 Pr Principles	1 St Stories & History	Mo Motivation	2 Tm Teamwork	1 To Tools	5 Cu Customers	Sp Suppliers	1 Qy Quality
1 Re Renewal	1 Re Recruit	1 Se Security	1 Dv Diversity	4 St Style	1 Rs Resilience	1 Be Beliefs	1 Fi Finances	Cl Clients	1 Sb Substitution or Disruption	1 Co Competitors
1 Ex Worked Examples	1 Te Templates	1 Sf Safe Fail	1 Sy Strategies & Plans	1 Po Policies	1 To Tolerance or Compassion	1 Nt Networks CoPs	1 Ip Intellectual Property	Sp Suppliers	1 Ne New entrants	1 Mi Market Intelligence
1 In Innovation	1 Bc Business Continuity	1 Ld Learning & Development	1 Sr Structure		1 Ad Adaptability	1 Ex Experience		Co Contractors	1 Re Relationships	1 Ma Materials
1 De Decision Making	1 Rm Resource management	1 Ci Continuous Improvement			1 Cr Creativity	1 Sk Skills			1 En Enablers	1 Ag Agreements Contracts

# Big Information Management

## Information Environment



# Big Information Management

## Data Exploration



### Definition

- Understand Data and Data Value
- Explore attributes/ metadata
- Understand data quality (needs/reality)
- Prioritise
- Identify custodians

### Required Capabilities

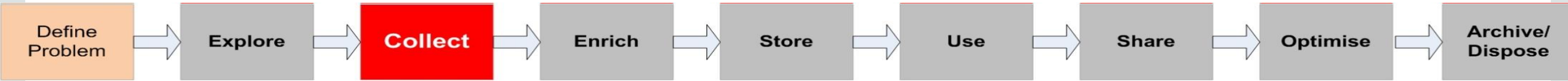
- Planning
- Experience
- Continuous Improvement
- Data and Information
- Enablers
- Tools

### Addressing Barriers

Access	
Efficient Use	✓
Quality	✓
Business Evolution	

# Big Information Management

## Data Collection



### Definition

- Acquisition/Ingestion/Extraction of data
- Quality audit
- Change Data Capture
- High-Volume Data Acquisition
- Stream/Event Data Capture

### Required Capabilities

- Resource Management
- Outcomes
- Skills
- Tools
- Enablers

### Addressing Barriers

Access	
Efficient Use	
Quality	✓
Business Evolution	



# Big Information Management

## Data Enrichment



### Definition

- Data Organisation
- Enrichment
- Processing
- Transformation
- Protection
- Data Lineage

### Required Capabilities

- Procedures
- Outcomes
- Hygiene Factors
- Tools
- Intellectual Property

### Addressing Barriers

Access	✓
Efficient Use	✓
Quality	✓
Business Evolution	

# Big Information Management

## Data Storage



### Definition

- Structured Data
- Content/Documents/Records
- Storage Strategy
- Data availability

### Required Capabilities

- Resource Management
- Business Continuity
- Knowledge
- Data and Information
- Intellectual Property
- Quality

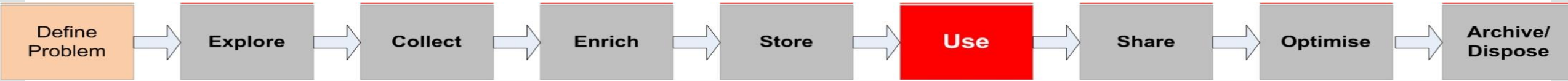
### Addressing Barriers

Access	✓
Efficient Use	✓
Quality	
Business Evolution	



# Big Information Management

## Data Use – Information Creation



### Definition

- Generate Information Assets
- Derive intelligence
- Actionable Insights
- Support Decisions
- Enable Innovation
- Support Business Processes

### Required Capabilities

- Decision Making
- Performance Management
- Change Management
- Strategy and Plans
- Resource Management
- Intellectual Property
- Policies
- Innovation
- Experience
- Skills
- Templates
- Creativity

### Addressing Barriers

Access	
Efficient Use	✓
Quality	
Business Evolution	✓

# Big Information Management

## Data / Information Sharing



### Definition

- Publish Information Assets
- Action on Intelligence
- Communicate Decisions
- Operationalise Innovation

### Required Capabilities

- Outcomes
- Decision Making
- Communication
- Collaboration
- Change Management
- Shared Vision
- Tools
- Experience
- Structure
- Relationships

### Addressing Barriers

Access	✓
Efficient Use	✓
Quality	
Business Evolution	✓

# Big Information Management

## Optimisation



### Definition

- Optimise Information Assets
- Optimise Business Processes
- Optimise Decisions

### Required Capabilities

- Performance Management
- Business Continuity
- Outcomes
- Shared Vision
- Data and Information
- Continuous Improvement

### Addressing Barriers

Access	
Efficient Use	
Quality	
Business Evolution	✓

# Big Information Management

## Data/Information Archiving/Disposal



### Definition

- Define and apply disposal authorities
- Preserve information of interest

### Required Capabilities

- Procedures
- Policies
- Security
- Resource Management

### Addressing Barriers

Access	✓
Efficient Use	✓
Quality	
Business Evolution	

# Big Information Management

## Governance

### Information Security

Snowden syndrome  
Filtering public data  
Derived assets  
Privacy

### Data Quality Management

Quality needs  
Quality assessment  
Real-time transformation

### Information Lifecycle Management

From Planning To Outcome  
Metadata  
History

### Data Modelling

From Model-centric  
To Model-on-demand

### Master Data Management

Golden Record  
Single View of...

### Metadata Management

Understanding assets  
Lost in Transformation  
Analysis

### Data Management

Proper data handling  
Data hygiene  
Standardisation  
Data Architecture

### Content Management

Web content  
Social-Media  
Documents  
Sensory Data  
Data feeds

### Information Environment Administration

Processes  
Info. Access  
Single point of  
Maintenance  
High-performing  
Fault-tolerant



# Why Big Data Projects Sometimes Fail?

- Poor Data Governance
- Lack of Big Data Strategy
- Complexity of siloed data, data integration/curation
- Building, optimisation and maintenance of in-house Big Data platforms
- Business Problem not clearly defined
- Insufficient skills and resources
- Inadequate understanding of data



**QA**