The Chartered Institute for IT

Enabling the information society

What is Big Data?

BCS Aberdeen Branch – 6 November 2014

Keith Gordon

- Soldier
- Engineer
- Teacher
- Information Systems Professional
- Data Manager
- Standards ‘Expert’
- Big Data ‘Sceptic’

What they say …

The overeager adoption of big data is likely to result in catastrophes of analysis comparable to a national epidemic of collapsing bridges.

…

Despite recent claims to the contrary, we are no further along with computer vision than we were with physics when Isaac Newton sat under his apple tree.

Michael I. Jordan
Pehong Chen Distinguished Professor
University of California, Berkeley.

Agenda

- **Defining Big Data**
- Storing Big Data
- Using Big Data
- Final thoughts

Towards a definition …

The New York Times

“It is only information”

Towards a definition …

The Institution of Engineering and Technology

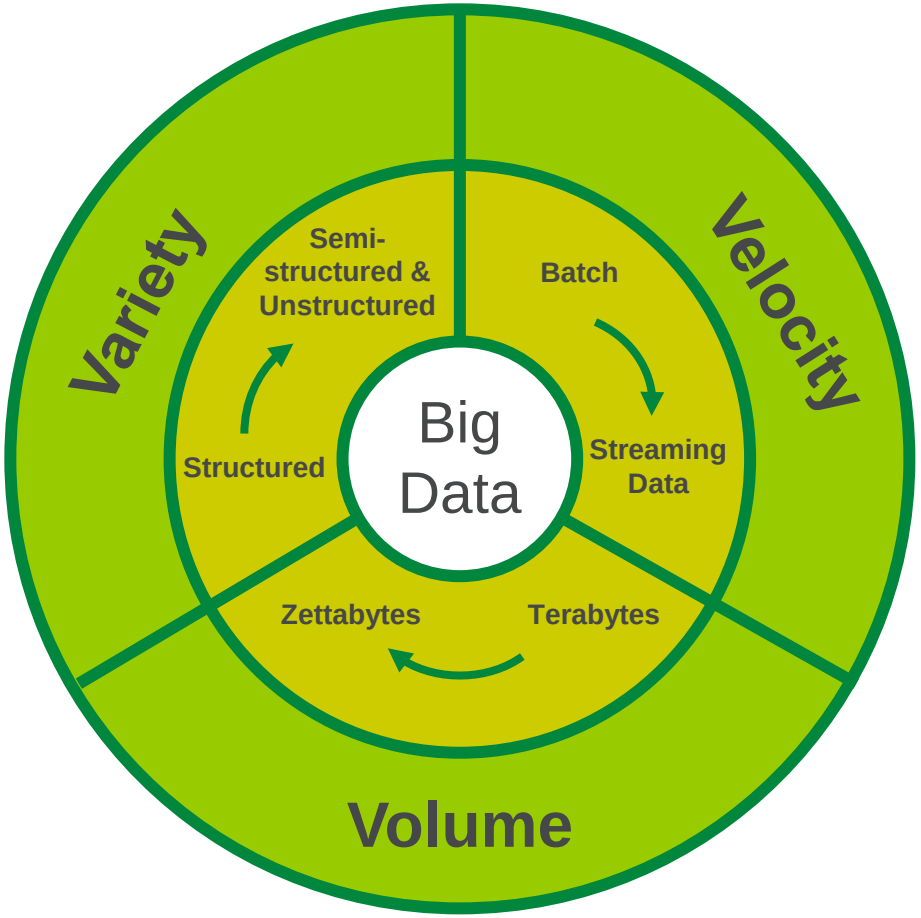
The IET Seminar on Big Data:
Turning big challenges into big opportunities

“I am confused!!”

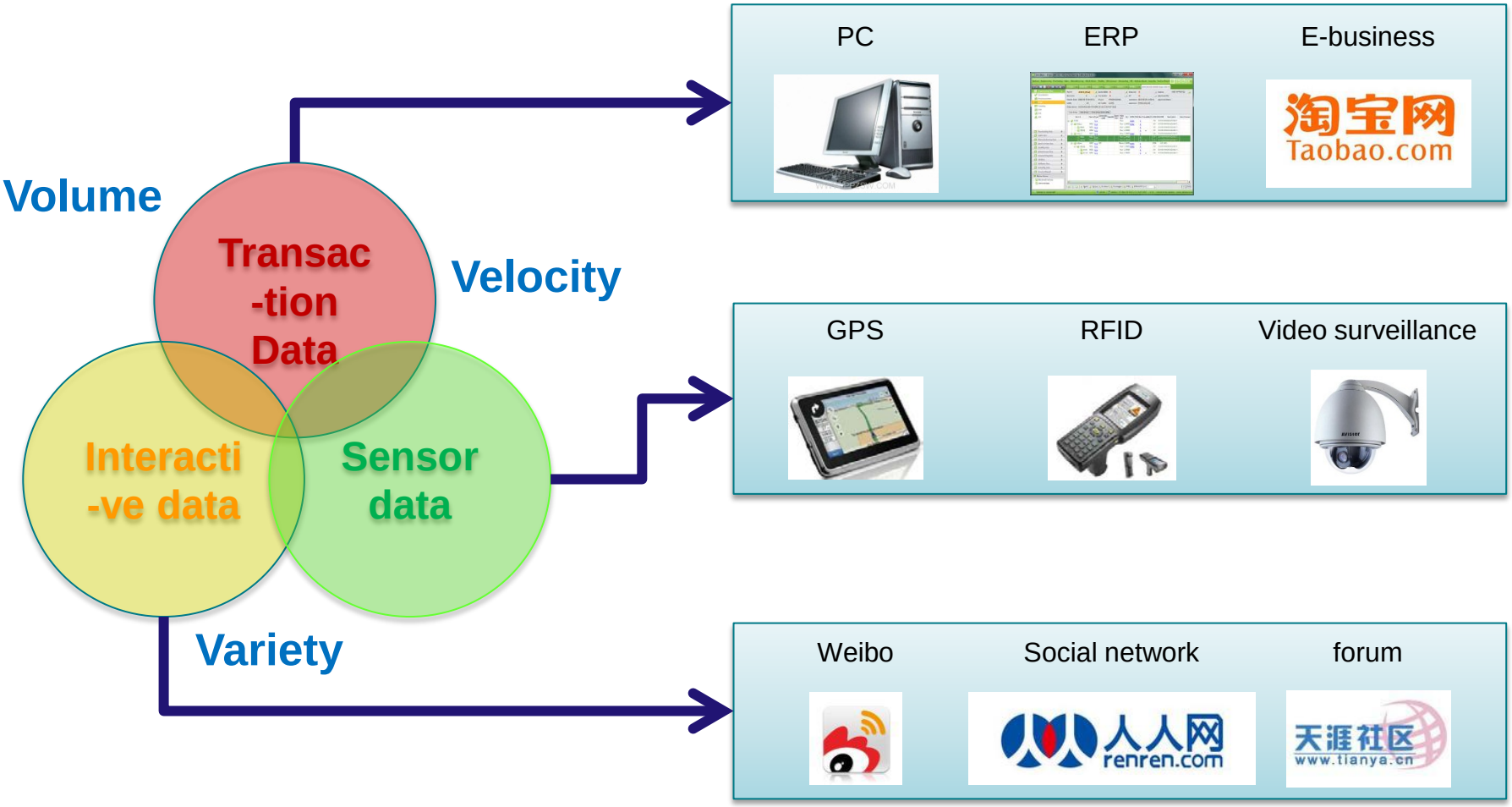
Towards a definition …

Gartner®

“The 3 Vs”



Towards a definition …



China Electronics Standardization Institute

Adding to the Vs

Volume – where the amount of data is sufficiently large so as to require special considerations

Variety – where the data consists of multiple types of data potentially from multiple sources; this variety can be combinations of structured data, semi-structured data and unstructured data

Velocity – where the data is produced at high rates and operating on 'stale' data is not valuable

Value – where the data has perceived or quantifiable benefit to the enterprise or organization using it

Veracity – where the correctness of the data can be assessed

Variability – where the change in velocity is important

We have a definition!

Big Data is a data set(s) with characteristics (e.g. volume, velocity, variety, variability, veracity, etc.) that for a particular problem domain at a given point in time cannot be efficiently processed using current/existing/established/ traditional technologies and techniques in order to extract value.

This definition distinguishes Big Data from business intelligence and traditional transactional processing while alluding to a broad spectrum of applications that includes them. The ultimate goal of processing Big Data is to derive differentiated value that can be trusted (because the underlying data can be trusted).

ISO/IEC JTC1 Study Group on Big Data

Agenda

Defining Big Data

Storing Big Data

Using Big Data

Final thoughts

SQL databases

Based on the relational model of data

- Grounded in set theory
- Mature and well understood
- Concurrency support and transaction support
- Fixed schema
- Now “object-relational”

But, there are problems

- Impedance mismatch
- Do not scale well
- Poor at semi-structured and unstructured data
- Cannot handle new datatypes

NoSQL databases

“Not only SQL” or “Never SQL”

Schema-less

Scalable (work well in clusters)

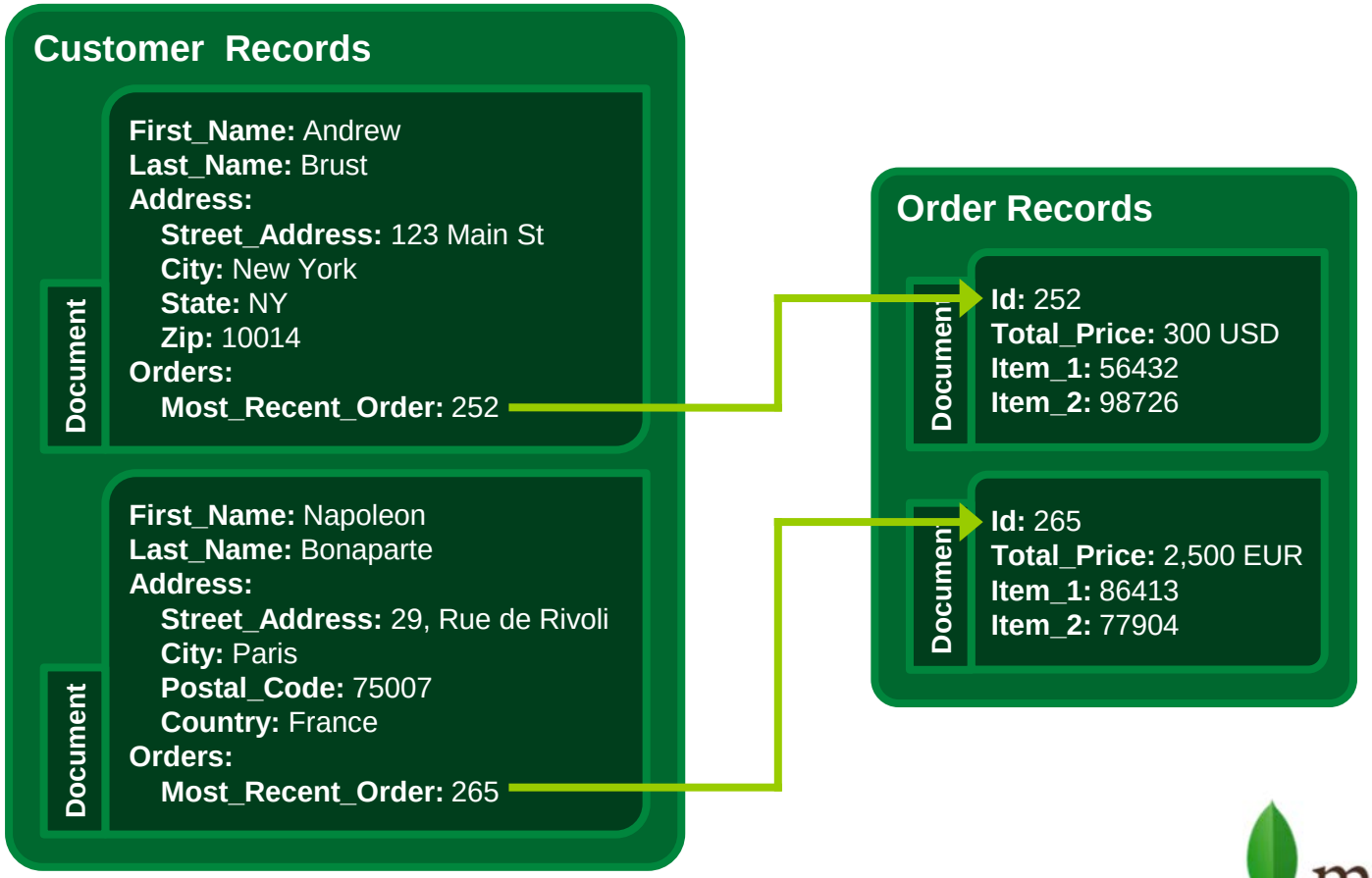
Four basic categories

- Key-value stores
- Document stores
- Wide column stores
- Graph databases

Key-value stores

aBcdeF27	Name: Arvind Age: 20 Type: Student Likes: play cricket Studies: History”
ghJK5L88	Name: Nitin Age: 25 Type: Student Likes: play cricket, play cards”

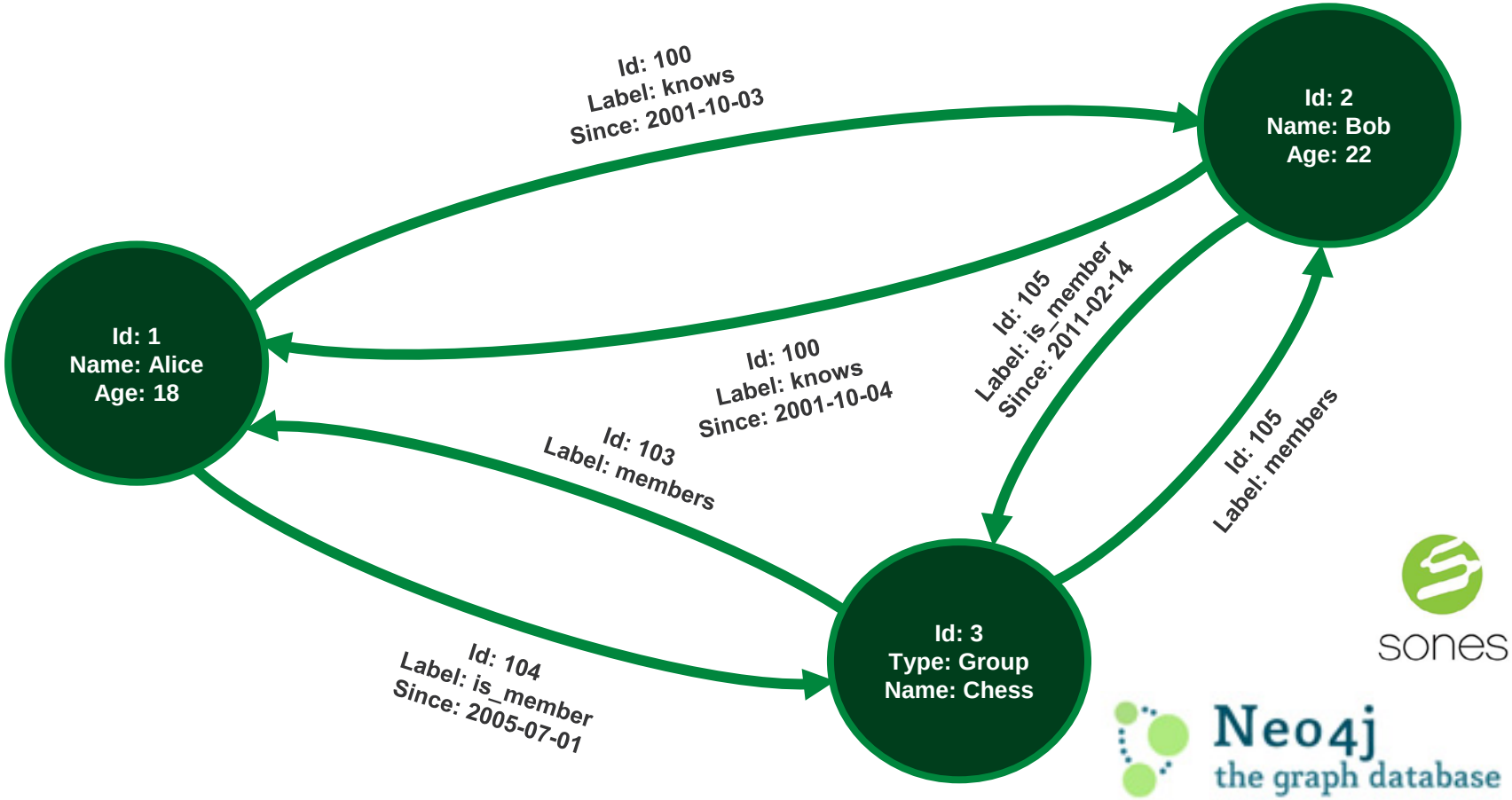
Document stores



Wide-column stores

7b976c48…	name: Bill Waterson	state: DC	birth_date: 1953
7c8f33e2…	name: Howard Tayler	state: UT	birth_date: 1968
7dd2a3630…	name: Randall Monroe	state: PA	
7da30d76…	name: Dave Kellett	state: CA	

Graph databases



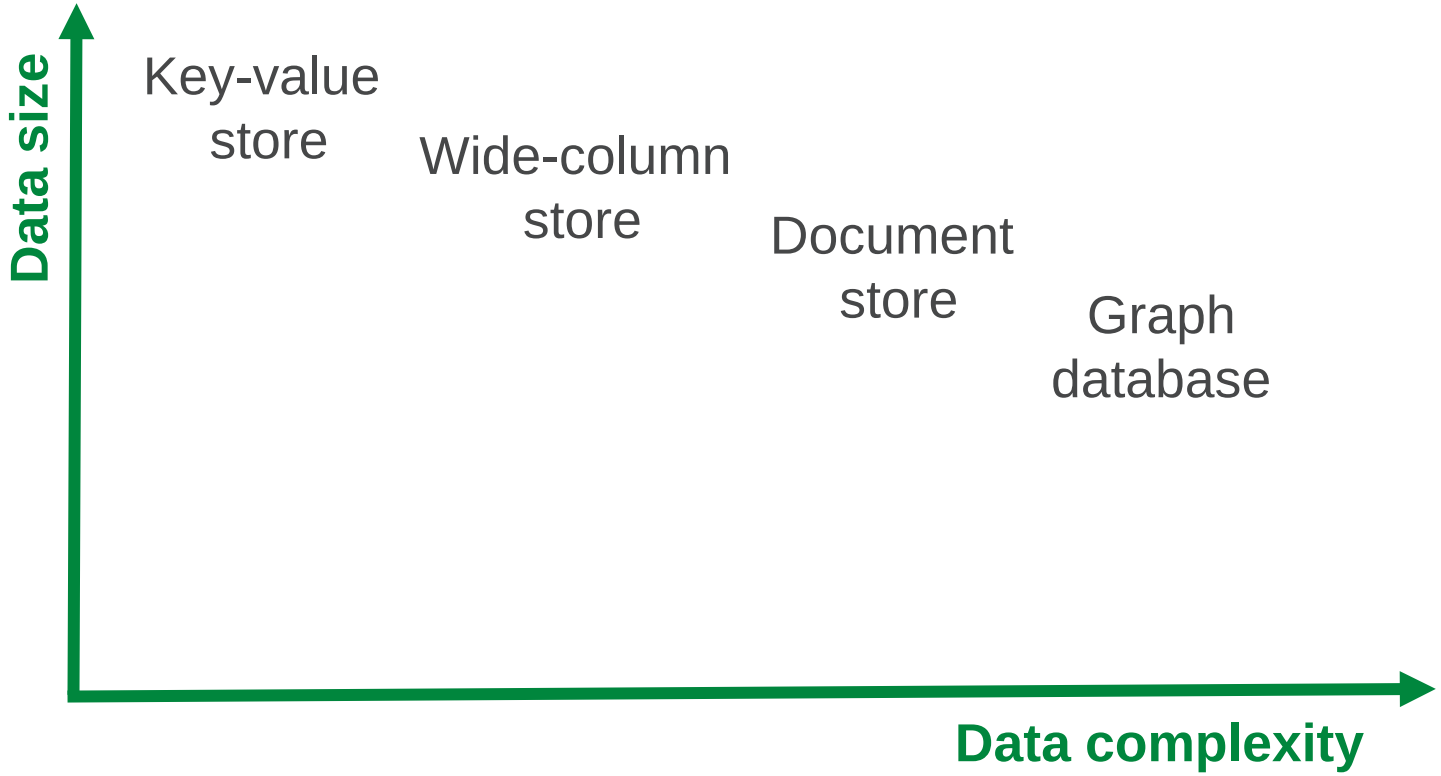
sones

Neo4j
the graph database

Performance comparison

Data Model	Performance	Scalability	Flexibility	Complexity
Relational database	variable	variable	low	moderate
Key-value store	high	high	high	none
Document store	high	variable	high	low
Wide-column store	high	high	moderate	low
Graph database	variable	variable	high	high

Size and complexity



Agenda

Defining Big Data

Storing Big Data

Using Big Data

Final thoughts

Initial players …

Google

amazon

facebook

Following closely …

Use cases – key-value stores

Suitable for:

- Session information
- User profiles, preferences
- Shopping basket/trolley information

Not suitable for:

- Relationships among data
- Multi-operation transactions
- Query by data

Use cases – document stores

Suitable for:

- Event logging
- Content management systems
- Real-time analytics
- E-commerce applications

Not suitable for:

- Complex transactions
- Queries against varying aggregates

Use cases – wide-column stores

Suitable for:

- Event-logging
- Content management systems
- Counters (particularly in web applications)

Not suitable for:

- Where ACID transactions are needed

Use cases – graph databases

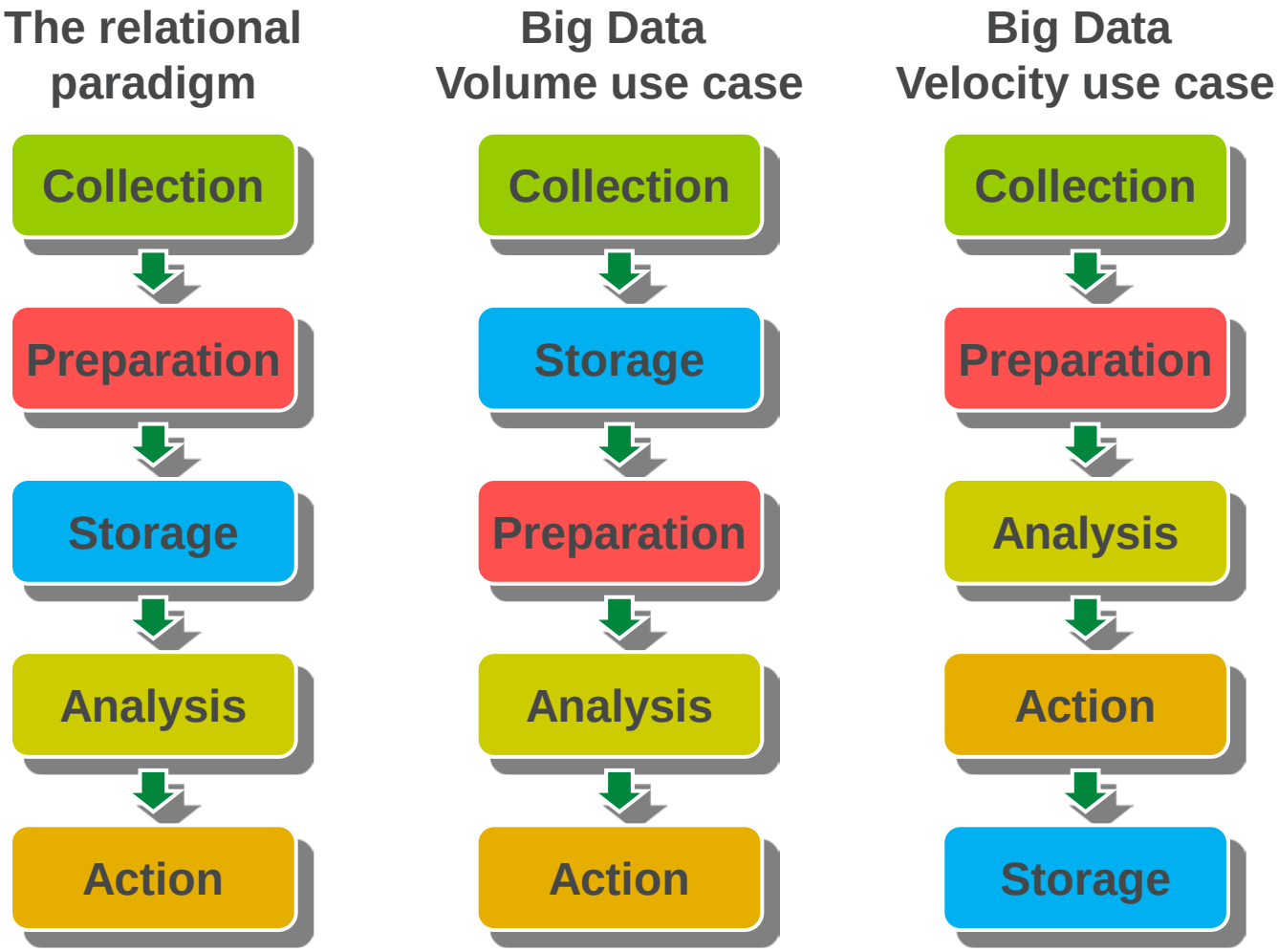
Suitable for:

- Connected data (friends, employees, etc)
- Routing, despatch, location-based services
- Recommendations

Not suitable for:

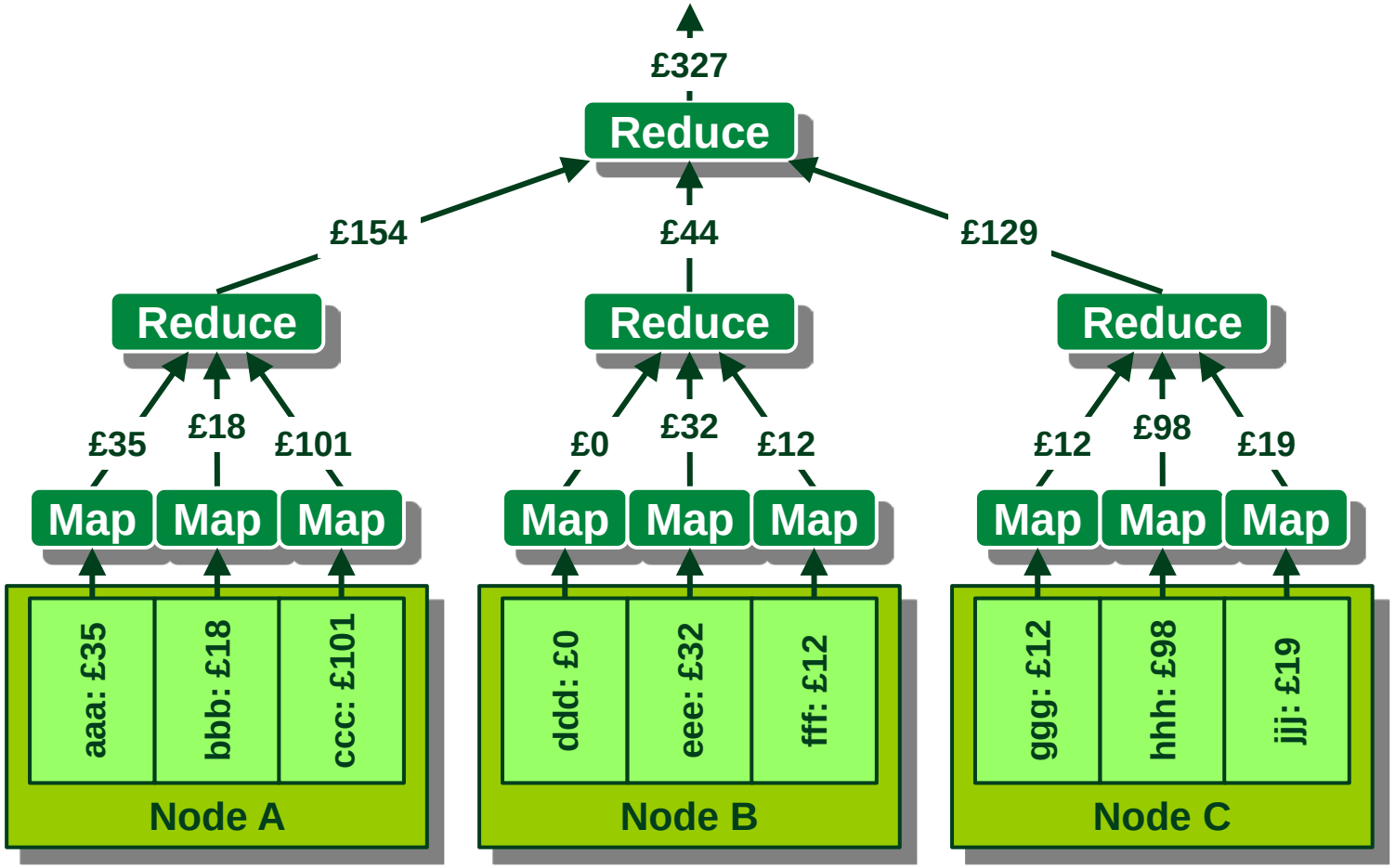
- Multiple updates

Changes in the data life cycle



MapReduce

“Google’s greatest lasting contribution to computer science”



Agenda

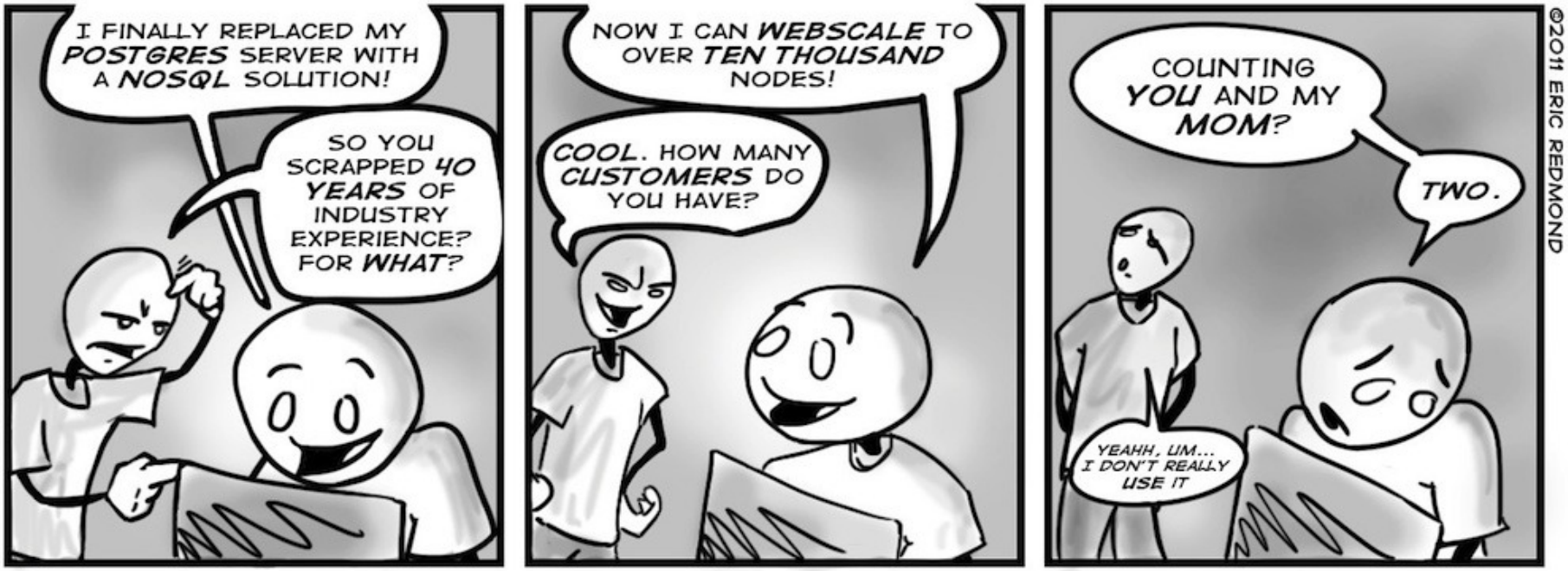
Defining Big Data

Storing Big Data

Using Big Data

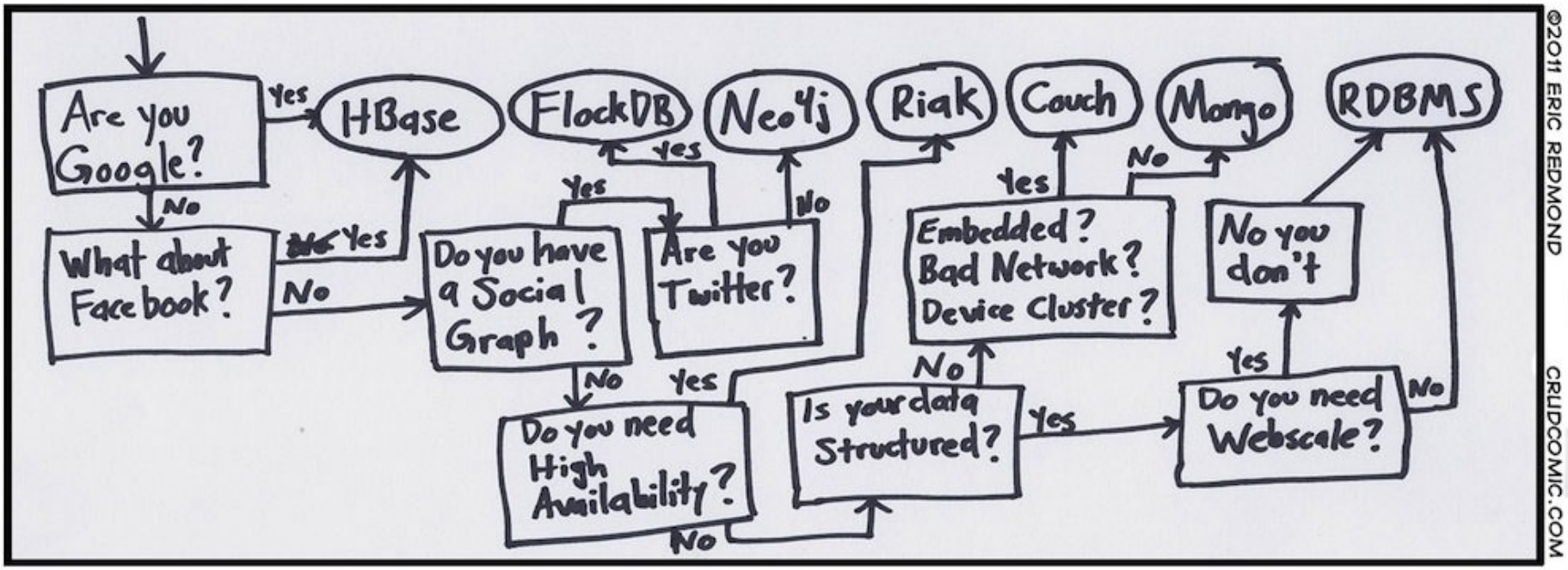
Final thoughts

Final thought #1



From Redmond E., Wilson J.R. “Seven Databases in Seven Weeks”

Final thought #2



Final thought #3

Do not try to solve a problem you do not have!

Agenda

- Defining Big Data
- Storing Big Data
- Using Big Data
- Final thoughts

The Chartered Institute for IT

Enabling the information society

What is Big Data?

BCS Aberdeen Branch – 6 November 2014