What is Big Data?

BCS Aberdeen Branch – 6 November 2014
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 …

“It is only information”
Towards a definition …

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

“I am confused!!”
Towards a definition ...

Gartner®

“The 3 Vs”

Big Data

Volume

Velocity

Variety

Structured

Semi-structured & Unstructured

Zettabytes

Terabytes

Batch

Streaming Data

Gartner is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and elsewhere.
Towards a definition …

Volume

Transac -tion Data

Velocity

Interacti -ve data

Sensor data

Variety

China Electronics Standardization Institute
Towards a definition …

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

1. Defining Big Data
2. Storing Big Data
3. Using Big Data
4. Final thoughts
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
<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
</table>
| 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

Customer Records
- First_Name: Andrew
- Last_Name: Brust
- Address:
  - Street_Address: 123 Main St
  - City: New York
  - State: NY
  - Zip: 10014
- Orders:
  - Most_Recent_Order: 252

Order Records
- Id: 252
  - Total_Price: 300 USD
  - Item_1: 56432
  - Item_2: 98726

- Id: 265
  - Total_Price: 2,500 EUR
  - Item_1: 86413
  - Item_2: 77904
## Wide-column stores

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>State</th>
<th>Birth Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>7b976c48...</td>
<td>Bill Waterson</td>
<td>DC</td>
<td>1953</td>
</tr>
<tr>
<td>7c8f33e2...</td>
<td>Howard Tayler</td>
<td>UT</td>
<td>1968</td>
</tr>
<tr>
<td>7dd2a3630...</td>
<td>Randall Monroe</td>
<td>PA</td>
<td></td>
</tr>
<tr>
<td>7da30d76...</td>
<td>Dave Kellett</td>
<td>CA</td>
<td></td>
</tr>
</tbody>
</table>
Graph databases

Id: 1
Name: Alice
Age: 18

Id: 2
Name: Bob
Age: 22

Id: 3
Type: Group
Name: Chess

Id: 103
Label: members
Since: 2001-10-04

Id: 104
Label: is_member
Since: 2005-07-01

Id: 100
Label: knows
Since: 2001-10-03

Id: 105
Label: is_member
Since: 2011-02-14
## Performance comparison

<table>
<thead>
<tr>
<th>Data Model</th>
<th>Performance</th>
<th>Scalability</th>
<th>Flexibility</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relational database</td>
<td>variable</td>
<td>variable</td>
<td>low</td>
<td>moderate</td>
</tr>
<tr>
<td>Key-value store</td>
<td>high</td>
<td>high</td>
<td>high</td>
<td>none</td>
</tr>
<tr>
<td>Document store</td>
<td>high</td>
<td>variable</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>Wide-column store</td>
<td>high</td>
<td>high</td>
<td>moderate</td>
<td>low</td>
</tr>
<tr>
<td>Graph database</td>
<td>variable</td>
<td>variable</td>
<td>high</td>
<td>high</td>
</tr>
</tbody>
</table>
Size and complexity

Data size

- Key-value store
- Wide-column store
- Document store
- Graph database

Data complexity
Initial players …
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

The relational paradigm

Collection → Preparation → Storage → Analysis → Action

Big Data Volume use case

Collection → Storage → Preparation → Analysis → Action

Big Data Velocity use case

Collection → Preparation → Analysis → Action → Storage
MapReduce

“Google’s greatest lasting contribution to computer science”

Node A
- Map
  - aaa: £35
  - bbb: £18
  - ccc: £101

Node B
- Map
  - ddd: £0
  - eee: £32
  - fff: £12

Node C
- Map
  - ggg: £12
  - hhh: £98
  - jjj: £19
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
What is Big Data?

BCS Aberdeen Branch – 6 November 2014