Big Data: Data Wrangling Boot Camp

Big Data Vs

Chuck Cartledge, PhD

27 January 2017
Table of contents

1. Introduction
2. Big Data's Vs
3. A laundry list of Vs
4. Q & A
5. Conclusion
6. References
Focusing on BD Vs

“What is Big Data? A meme and a marketing term, for sure, but also shorthand for advancing trends in technology that open the door to a new approach to understanding the world and making decisions.”

Lohr [15]

Image from [6].
Doug Laney, META Group

The origin of “Big Data” ideas and definitions.

- Started in the e-commerce Mergers and Acquisitions arena
- Used to explain why traditional Relational Database Management Systems (RDMS) wouldn’t scale
- Intended audience was non-technical management

<table>
<thead>
<tr>
<th>Volume</th>
<th>Velocity</th>
<th>Variety</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tiered storage/hub and spoke</td>
<td>Operational data stores</td>
<td>Inconsistency resolution</td>
</tr>
<tr>
<td>Selective data retention</td>
<td>Data caches</td>
<td>XML-based “universal” translation</td>
</tr>
<tr>
<td>Statistical sampling</td>
<td>Point-to-point data routing</td>
<td>Application-aware EAI adapters</td>
</tr>
<tr>
<td>Redundancy elimination</td>
<td>Balance data latency with decision cycles</td>
<td>Data access middleware and ETL/M</td>
</tr>
<tr>
<td>Offload “cold” data</td>
<td></td>
<td>Distributed query management</td>
</tr>
<tr>
<td>Outsourcing</td>
<td></td>
<td>Metadata management</td>
</tr>
</tbody>
</table>

Take away: traditional RDMS don’t/won’t scale and different approaches are needed.

Image from [12].
Laney’s original BD Vs

**Figure 1 — Data Management Solutions**

- **Volume**
  - Tiered storage/hub and spoke
  - Selective data retention
  - Statistical sampling
  - Redundancy elimination
  - Offload “cold” data
  - Outsourcing

- **Velocity**
  - Operational data stores
  - Data caches
  - Point-to-point data routing
  - Balance data latency with decision cycles

- **Variety**
  - Inconsistency resolution
  - XML-based “universal” translation
  - Application-aware EAI adapters
  - Data access middleware and ETLM
  - Distributed query management
  - Metadata management

*Extending data management options enables greater returns on information assets*

*Source: META Group*
Laney’s Vs recapped

1. Velocity
   - Frequency of data generation/delivery
   - Think of data from a device, or sensor, robots, clicklogs
   - Real-time analysis is small (9%) [19].
   - Most Big Data analytics is batch

2. Variety
   - Data from a multitude of different sources.
   - Not all data is useful.

   - Data is lost during “normalization”
   - Hopefully not important data, when in doubt: keep it somehow
   - Gets away from relational databases

3. Volume
   - Store relational records?
   - Store transactional records?
   - How long to keep data available?
   - How to access data?
   - How to migrate data?
Volume — what does it mean for Big Data?

How much is there? And, how do we store it?

- Store relational records?
- Store transactional records?
- How long to keep data available?
- How to access data?
- How to migrate data?

Figure 1
Data is growing at a 40 percent compound annual rate, reaching nearly 45 ZB by 2020


Image from [10].
Velocity — what does it mean for Big Data?

- Frequency of data generation/delivery
- Think of data from a device, or sensor, robots, clicklogs
- Real-time analysis is small (9%) [19].
- Most Big Data analytics is batch

Known as “Little’s Law” [13]

Take away: data is generated at a high speed, it must be analyzed before the next set of data is delivered.
Variety — what does it mean for Big Data?

Not all data is the same.

- Data from a multitude of different sources.
- Not all data is useful.
- Data is lost during “normalization”
- Hopefully not important data, when in doubt: keep it somehow
- Gets away from relational databases
The original Vs have been expanded

Lots more Vs.

1. Vagueness
2. Validity
3. Value
4. Variability
5. Variety
6. Velocity
7. Venue
8. Veracity
9. Viability
10. Vincularity
11. Virility
12. Viscosity
13. Visibility
14. Visible
15. Visualization
16. Vitality
17. Vocabulary
18. Volatility
19. Volume

We’ll delve into these now.
Big Data as 3 Vs

Image from [20].
Big Data as 4 Vs

Volume: Scale of Data
- 40 zettabytes (40 x 10^21 bytes) if data will be created by 2020, an increase of 300 times from 2005
- 2.5 quintillion bytes (2.5 x 10^18 bytes) of data are created each day
- 6 billion people have cell phones
- World population: 7 billion

Velocity: Analysis of Streaming Data
- The New York Stock Exchange captures 1 TB of trade information during each trading session
- By 2016, it is projected there will be 18.9 billion network connections – almost 2.5 connections per person on earth

Variety: Different Forms of Data
- As of 2011, the global size of data in healthcare was estimated to be 150 exabytes (1.5 x 10^18 bytes)
- 30 billion pieces of content are shared on Facebook every month
- 4 billion+ hours of video are watched on YouTube each month
- 400 million tweets are sent per day by about 330 million monthly active users

Veracity: Uncertainty of Data
- By 2015, 4.4 million IT jobs will be created globally to support big data, with 1.9 million in the United States
- 27% of respondents
- In one survey, 60% of survey respondents stated how much of their data was inaccurate

IBM

Image from [23].
Big Data as 5 Vs

Volume
- Terabytes
- Records/Arch
- Transactions
- Tables, Files

Velocity
- Batch
- Real/near-time
- Processes
- Streams

Variety
- Structured
- Unstructured
- Multi-factor
- Probabilistic

Value
- Statistical
- Events
- Correlations
- Hypothetical

Veracity
- Trustworthiness
- Authenticity
- Origin, Reputation
- Availability
- Accountability

Image from [3].
Big Data as 6 Vs

Image from [24].
Big Data as 7 Vs

Image from [21].
### Vs (part 1 of 7)

<table>
<thead>
<tr>
<th>Num.</th>
<th>Year</th>
<th>V</th>
<th>Definition</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2001</td>
<td>Variety</td>
<td>...no greater barrier to effective data management will exist than the variety of incompatible data formats, non-aligned data structures, and inconsistent data semantics.</td>
<td>[12, 16]</td>
</tr>
<tr>
<td>2</td>
<td>2001</td>
<td>Velocity</td>
<td>E-commerce has also increased point-of-interaction (POI) speed and, consequently, the pace data used to support interactions and generated by interactions.</td>
<td>[12]</td>
</tr>
<tr>
<td>3</td>
<td>2001</td>
<td>Volume</td>
<td>E-commerce channels increase the depth/breadth of data available about a transaction (or any point of interaction).</td>
<td>[12]</td>
</tr>
</tbody>
</table>
### Vs (part 2 of 7)

<table>
<thead>
<tr>
<th>Num.</th>
<th>Year</th>
<th>V</th>
<th>Definition</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2013</td>
<td>Validity</td>
<td>…is the data correct and accurate for the intended use.</td>
<td>[2, 14, 16, 17, 25]</td>
</tr>
<tr>
<td>5</td>
<td>2013</td>
<td>Value</td>
<td>How to determine the prescriptive value of data?</td>
<td>[2, 7, 14, 22, 25, 26, 11, 9, 4, 1]</td>
</tr>
<tr>
<td>6</td>
<td>2013</td>
<td>Variability</td>
<td>Many options or variable interpretations can confuse interpretation.</td>
<td>[2, 7, 16, 22, 26]</td>
</tr>
</tbody>
</table>
### Vs (part 3 of 7)

<table>
<thead>
<tr>
<th>Num.</th>
<th>Year</th>
<th>V</th>
<th>Definition</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>2013</td>
<td>Veracity</td>
<td>... to the biases, noise and abnormality in data.</td>
<td>[2, 7, 14, 17, 25, 26, 18, 9, 4, 5, 1]</td>
</tr>
<tr>
<td>8</td>
<td>2013</td>
<td>Viability</td>
<td>... can the data be analyzed in a way that makes it decision-relevant?</td>
<td>[7, 16]</td>
</tr>
<tr>
<td>9</td>
<td>2013</td>
<td>Virility</td>
<td>... Defined by some users as the rate at which the data spreads; how often it is picked up and repeated by other users or events.</td>
<td>[26]</td>
</tr>
</tbody>
</table>
### Vs (part 4 of 7)

<table>
<thead>
<tr>
<th>Num.</th>
<th>Year</th>
<th>V</th>
<th>Definition</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>2013</td>
<td>Viscosity</td>
<td>...used to describe the latency or lag time in the data relative to the event being described.</td>
<td>[26]</td>
</tr>
<tr>
<td>11</td>
<td>2013</td>
<td>Visibility</td>
<td>...the state of being able to see or be seen - is implied. [14, 25, 16]</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>2013</td>
<td>Visualization</td>
<td>Making all that vast amount of data comprehensible in a manner that is easy to understand and read. With the right analyses and visualizations, raw data can be put to use otherwise raw data remains essentially useless.</td>
<td>[22]</td>
</tr>
</tbody>
</table>
### Vs (part 5 of 7)

<table>
<thead>
<tr>
<th>Num.</th>
<th>Year</th>
<th>V</th>
<th>Definition</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>2013</td>
<td>Volatility</td>
<td>...how long is data valid and how long should it be stored.</td>
<td>[16, 17]</td>
</tr>
<tr>
<td>14</td>
<td>2014</td>
<td>Vagueness</td>
<td>...confusion over the meaning of big data (Is it Hadoop? Is it something that we've always had? What's new about it? What are the tools? Which tools should I use? etc.)</td>
<td>[2]</td>
</tr>
<tr>
<td>15</td>
<td>2014</td>
<td>Venue</td>
<td>...distributed, heterogeneous data from multiple platforms, from different owners systems, with different access and formatting requirements, private vs. public cloud.</td>
<td>[2]</td>
</tr>
</tbody>
</table>
## Vs (part 6 of 7)

<table>
<thead>
<tr>
<th>Num.</th>
<th>Year</th>
<th>V</th>
<th>Definition</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>2014</td>
<td>Vocabulary</td>
<td>. . . schema, data models, semantics, ontologies, taxonomies, and other content- and context-based metadata that describe the data's structure, syntax, content, and provenance.</td>
<td>[2]</td>
</tr>
<tr>
<td>17</td>
<td>2015</td>
<td>Vincularity</td>
<td>. . . it implies connectivity or linkage.</td>
<td>[16]</td>
</tr>
<tr>
<td>18</td>
<td>2015</td>
<td>Visible</td>
<td>We live in an increasingly visual world and the statistics of increase in the number of images and videos shared on the Internet is staggering.</td>
<td>[16]</td>
</tr>
</tbody>
</table>
Vs (part 7 of 7)

<table>
<thead>
<tr>
<th>Num.</th>
<th>Year</th>
<th>V</th>
<th>Definition</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>2015</td>
<td>Vitality</td>
<td>...criticality of the data is another concept that is crucial and is embedded in the concept of Value.</td>
<td>[16]</td>
</tr>
</tbody>
</table>
Big Data over time

Q & A time.

Q: What is the name of the world’s oldest kamikaze pilot?
A: Chicken Teriyaki.
What have we covered?

- Big Data Vs had a specific point of origin
- The list of Big Data continues to grow
- Big Data can be a very nebulous term

Next: Publicly available sources of Big Data.
References I


References II


References III


References IV


References V


References VI


References VII


References VIII
