

Big Data: Data Wrangling Boot Camp

Big Data Overview and Concepts

Chuck Cartledge, PhD

23 February 2018



Table of contents (1 of 1)

1 Introduction

- What we'll be covering

2 Big Data's Vs

- Classical definition
- Data sources and types

3 Concepts

- The Vs
- Lots of data
- What does data look like?

4 Virtualization

- Tricking hardware and software

- What is it good for?

- What is it not good for?

5 Q & A

6 Conclusion

7 References

8 Files



On the way to a working definition of BD.

*"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 [9]

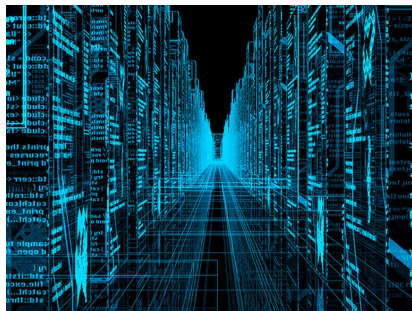


Image from [3].

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

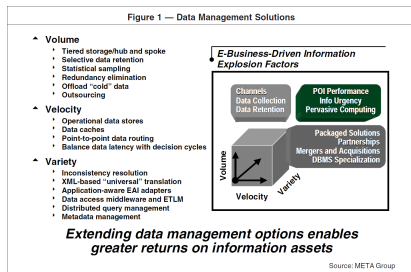


Image from [7].

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

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

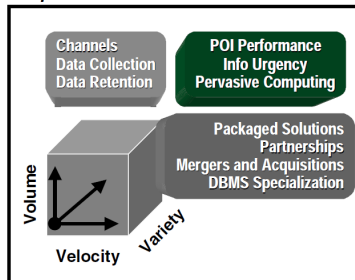
▲ Velocity

- ▶ 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

E-Business-Driven Information Explosion Factors



Extending data management options enables greater returns on information assets

Volume — what does it mean for Big Data?

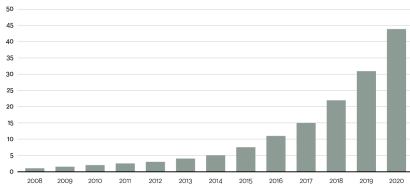
How much is there? And, how do we store it?

- 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

Data in zettabytes (ZB)



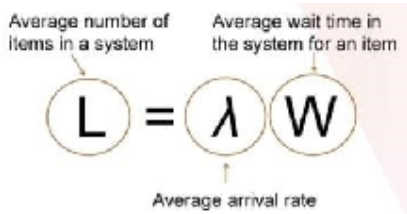
Source: Oracle, 2012

Image from [6].

See http://en.wikipedia.org/wiki/Metric_prefix for list of prefixes.

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%) [10].
- Most Big Data analytics is batch



Known as “Little’s Law” [8]

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

We'll talk about these later.

Our friends the Vs

- Classic Vs (Variety, Velocity, Volume)
- Additional Vs

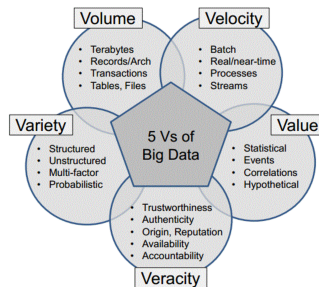


Image from [2].

The Vs tend to overlap.



What does data look like?

Data characteristics

- Formatted/unformatted
(even well-known numbers
can be very different)
- Bits, bytes, tagged, free
form
- Clean, messy
- Complete, fragmented

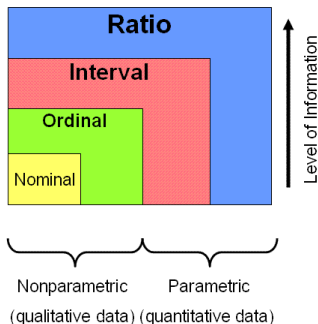
10000000
10000000 100000
<spaces> </spaces>
There are spaces.

We'll be looking at unformatted free form text.

What does data look like?

OBTW, there are different types of numbers.

- Categorical (Qualitative)[16]
 - Nominal – values are just different
 - Ordinal – values can order objects
- Numerical (Quantitative)
 - Interval – differences between values are important
 - Ratio – differences and ratios are important



***Nonparametric statistics may be used to analyze interval and ratio data measurements.**

Image from [14].



What does data look like?

Torrents of data

- Primary usage
- Secondary usage
- “Exhaust”
- Storage
 - 1 Accessibility
 - 2 Longevity
 - 3 Privacy



Image from [13].

Data can be intentional, or accidental, or by-products, but there is lots of it.

Big data players

- Visionaries – stand on the shoulders of giants and see new horizons
- Brokers – have seas and lakes of data at their disposal
- Scientists – dive into the seas and make the visions real



We will be performing a small part of the data scientist's labors.

A 50,000 foot view

What are the layers in this cake?

- User — the person (or thing) that want's something done
- Application — the program that does the work
- Operating system — arbitrates between multiple programs and limited resources
- Hardware — the silicone, copper, other tangibles that generate heat

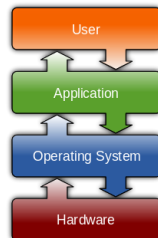


Image from [17].

Layering is a key concept.

Focusing on the OS

What does it do?

- Provides a user interface (maybe a Command Line Interface)
- Schedules access to the hardware
- Schedules the functions of the CPU

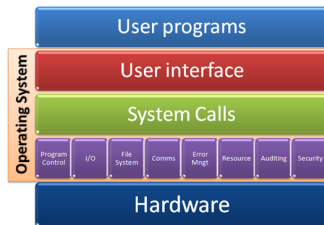


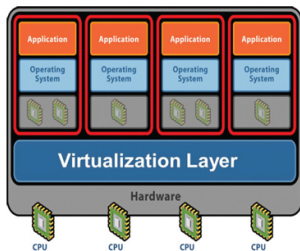
Image from [11].

An OS is a program (albeit, a large program). What if we could write a program that would run an OS as an application?



Tricking the upper layer

- Higher layers rely on lower layers for services
- Layers create interfaces
- Interfaces allow for hiding details



Virtualization software allows applications that previously ran on separate computers to run on one server machine.

Image from [5].

As long as the lower layer supplies all the services, the upper layer won't know where the services originated.



What is it not good for?

Anything that has to be fast.

- Underlying hardware suite is shared across all “machines”
- Mission critical applications





What is it not good for?

In summary.

- To use virtual machines, or
- To not use virtual machines.



It depends on what is important. Many BD tools and techniques make use of virtualization.



Q & A time.

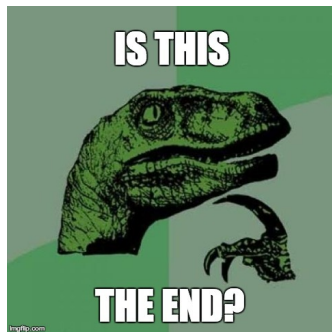
Q: How many existentialists does it take to screw in a light bulb?

A: Two. One to screw it in and one to observe how the light bulb itself symbolizes a single incandescent beacon of subjective reality in a netherworld of endless absurdity reaching out toward a maudlin cosmos of nothingness.



What have we covered?

- Big data Vs had a specific point of origin
- Big data has a list of challenges
- Big data can be very messy, and not neat and tidy
- Hinted at how BD tools and techniques use virtualization



Next: Understanding more about BD Vs.



References (1 of 6)

- [1] Divyakant Agrawal, Philip Bernstein, Elisa Bertino, Susan Davidson, Umeshwas Dayal, and Michael Franklin, Challenges and Opportunities with Big Data, Purde e-Pubs (2011).
- [2] Patrick Cheesman, How big data can transform your understanding of your customers, <http://www.patrickcheesman.com/how-big-data-can-transform-your-understanding-of-your-customers/>, 2106.



References (2 of 6)

- [3] David Gewirtz, Volume, velocity, and variety: Understanding the three v's of big data, <http://www.zdnet.com/article/volume-velocity-and-variety-understanding-the-three-vs-of-big-data/>, 2016.
- [4] Christian Hagen, KHalid Khan, Marco Ciobo, and Jason Miller, Big Data and the Creative Destruction of Today's Business Models, http://www.atkearney.com/strategic-it/ideas-insights/article/-/asset_publisher/LCcg0eS4t85g/content/big-data-and-the-creative-destruction-of-today-s-business-models/10192, 2013.



References (3 of 6)

- [5] Paul Hodge, Virtualization 101: Understanding how to do more with less, <https://www.isa.org/standards-and-publications/isa-publications/intech-magazine/2011/august/system-integration-virtualization-101-understanding-how-to-do-more-with-less/>, 2011.
- [6] Applied Innovations, Track website visitors, <http://www.appliedi.net/blog/track-website-visitors/>, 2010.
- [7] Doug Laney, 3D Data Management: Controlling Data Volume, Velocity and Variety, META Group Research Note **6** (2001).



References (4 of 6)

- [8] John DC Little, [A Proof for the Queuing Formula: \$L = \lambda W\$](#) , Operations Research **9** (1961), no. 3, 383–387.
- [9] Steve Lohr, [The age of big data](#), New York Times **11** (2012).
- [10] Philip Russom, [Big Data Analytics](#), TDWI Best Practices Report, Fourth Quarter (2011).
- [11] Willy-Peter Schaub, [UNISA Chatter Operating System Concepts: Part 2 System Structures](#)
http://blogs.msdn.com/b/willy-peter_schaub/archive/2010/01/07/unisa-chatter-operating-system-concepts-part-2-system-structures.aspx, 2010.



References (5 of 6)

- [12] NixOS Staff, [Nixos screenshots](https://nixos.org/nixos/screenshots.html), <https://nixos.org/nixos/screenshots.html>, 2016.
- [13] NYU Staff, [Nyu launches initiative in data science and statistics to push advances in medicine, science, technology, and other fields](https://www.nyu.edu/about/news-publications/news/2013/02/19/nyu-launches-initiative-in-data-science-and-statistics-to-push-advances-in-medicine-science-technology-and-other-fields.html), <https://www.nyu.edu/about/news-publications/news/2013/02/19/nyu-launches-initiative-in-data-science-and-statistics-to-push-advances-in-medicine-science-technology-and-other-fields.html>, 2013.
- [14] Six Sigma Staff, [Data Classification](#), 2017.

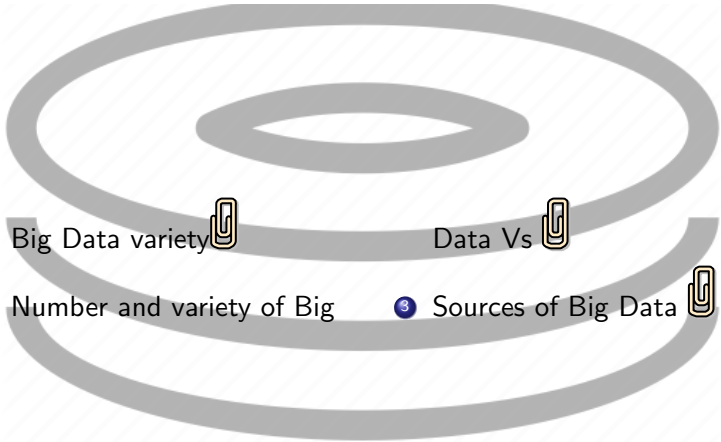



References (6 of 6)

- [15] Valsoft Services Staff, Challenges and Opportunities with Big Data, <http://valsoftservices.com/big-data-implementation/>, 2016.
- [16] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Introduction to Data Mining, Pearson Education India, 2006.
- [17] Wikipedia, Software — Wikipedia, The Free Encyclopedia, <http://en.wikipedia.org/wiki/Software>, 2015.



Files of interest

- 
- 1 Big Data variety 
 - 2 Number and variety of Big 
 - 3 Sources of Big Data 