

# What “Big Data” really means and where OpenVMS can play

Brett Cameron  
September 2014

## Abstract

“Big Data” seems to be an unavoidable business buzz phrase these days. According to Wikipedia, Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. This definition seems reasonable; however the term is badly overloaded and surrounded by considerable industry hype, and accordingly there continues to be considerable debate and confusion about what Big Data is and is not. For example, a recent survey of 154 companies found that nearly 70% used a volume-based definition for Big Data; 25% defined Big Data as 'massive growth of transaction data'; 24% thought Big Data referred to new technologies for managing massive data; and 19% defined it as the 'requirement to store and archive data for regulatory compliance'. In this talk the speaker will share his opinion (and those of a few noted luminaries) about what Big Data really means and will consider how OpenVMS might participate in this rapidly evolving area.

## About me

Brett Cameron currently works as a senior architect with HP's corporate Cloud Services group, focusing on the design and implementation of message queuing and related integration services for customers and for internal use. Brett lives in Christchurch, New Zealand, and has worked in the software industry for some 22 years. In that time he has gained experience in a wide range of technologies, many of which have long since been retired to the software scrapheap of dubious ideas. In recent years Brett has specialized in systems integration, and the design and implementation of large distributed systems for HP's enterprise customers. This work has seen Brett get involved in the research and development of low-latency and highly scalable messaging solutions for the Financial Services sector running on HP platforms, and as a consequence of this work, Brett has been involved in several interesting Open Source projects, and he has been responsible (or should that be irresponsible) for porting various pieces of Open Source software to the HP OpenVMS platform. Brett holds a doctorate in chemical physics from the University of Canterbury, and still maintains close links with the University, working as a part time lecturer in the Computer Science and Electronic and Computer Engineering departments. In his spare time, Brett enjoys listening to music, playing the guitar, and drinking beer.

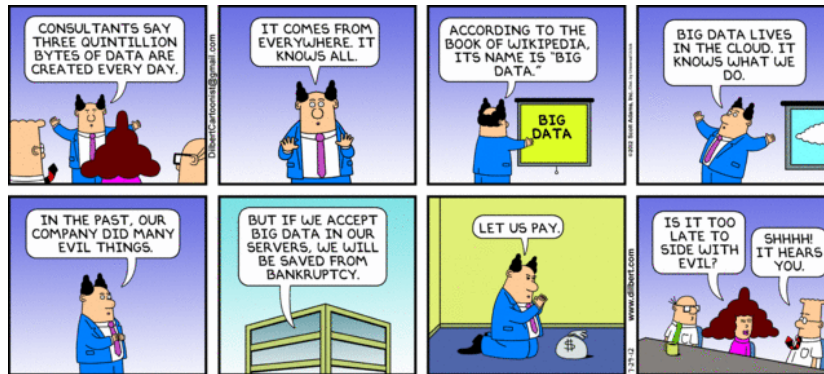


3

## AGENDA

- **Introduction**
- What is Big Data all about?
- Some case studies
- Technologies
- Where can OpenVMS play?
- Summary/conclusions
- Questions

## Introduction

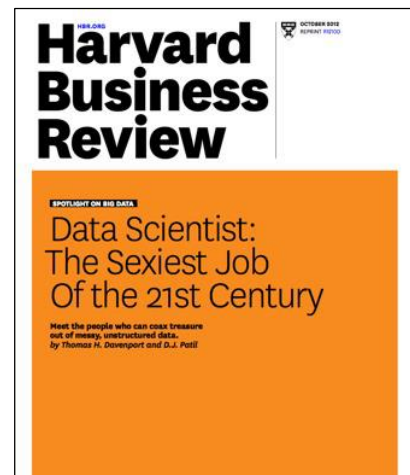


5

## Job prospects...

*"By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions."*


Seems a bit melodramatic. This stuff takes me back to my university days... scientific programming, we used to call it... number crunching using various statistical and numerical algorithms... used to be in FORTRAN; now we've got all sorts of other fancy options and more data to play with, but the underlying mathematical principles have really not changed that much (IMHO).



6

# The White House

- Big Data Research and Development Initiative
- \$200,000,000 funding (2012)



Office of Science and Technology Policy  
Executive Office of the President  
New Executive Office Building  
Washington, DC 20502

**FOR IMMEDIATE RELEASE**      Contact: Rick Weiss    202 456-6037    [weiss@ostp.eop.gov](mailto:weiss@ostp.eop.gov)  
March 29, 2012      Lisa-Joy Zupnik    703 292-8311    [lzupnik@ostp.gov](mailto:lzupnik@ostp.gov)

**OBAMA ADMINISTRATION UNVEILS "BIG DATA" INITIATIVE;  
ANNOUNCES \$200 MILLION IN NEW R&D INVESTMENTS**

Aiming to make the most of the fast-growing volume of digital data, the Obama Administration today announced a "Big Data Research and Development Initiative." By improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help solve some of the Nation's most pressing challenges.

To launch the initiative, six Federal departments and agencies today announced more than \$200 million in new commitments that, together, promise to greatly improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of digital data.

"In the same way that past Federal investments in information-technology R&D led to dramatic advances in supercomputing and the creation of the Internet, the initiative we are launching today promises to transform our ability to use Big Data for scientific discovery, environmental and biomedical research, education, and national security," said Dr. John P. Holdren, Assistant to the President and Director of the White House Office of Science and Technology Policy.

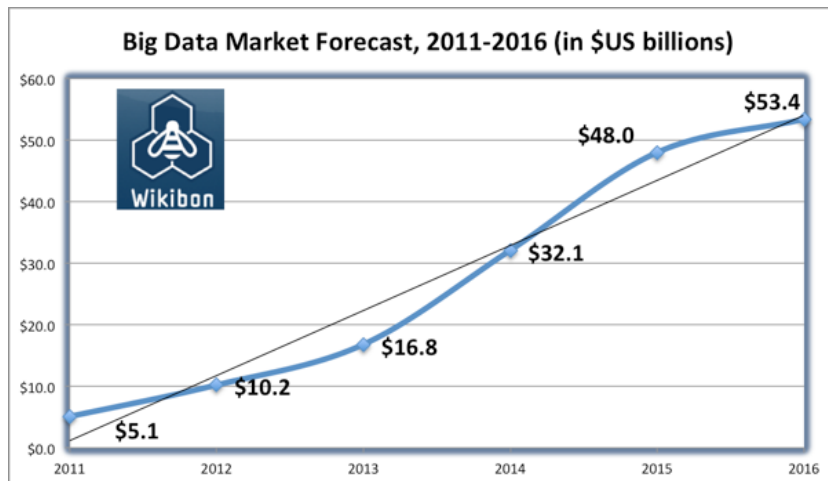
To make the most of this opportunity, the White House Office of Science and Technology Policy (OSTP)—in concert with several Federal departments and agencies—created the Big Data Research and Development Initiative to:

- Advance state-of-the-art core technologies needed to collect, store, preserve, manage, analyze, and share huge quantities of data.
- Harness these technologies to accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning; and
- Expand the workforce needed to develop and use Big Data technologies.

<http://www.whitehouse.gov>

7

# The market forecast



<http://www.wikibon.com>

*And \$US billions is being invested.*

8

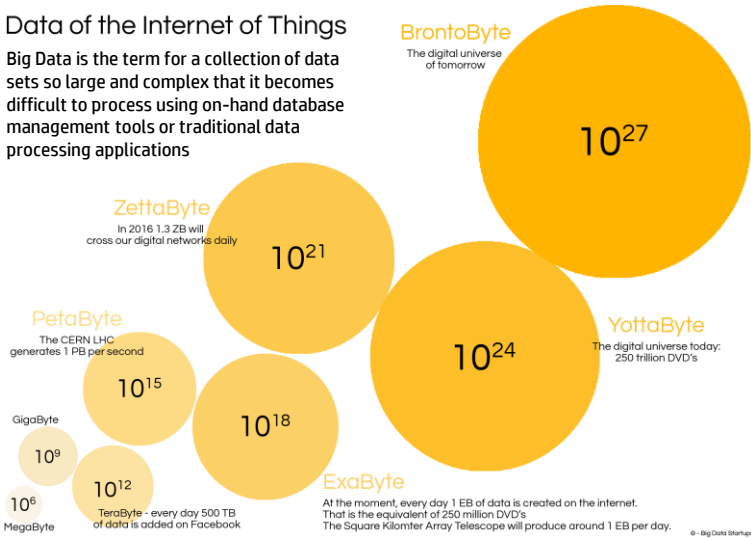
# AGENDA

- Introduction
- **What is Big Data all about?**
- Some case studies
- Technologies
- Where can OpenVMS play?
- Summary/conclusions
- Questions



# What is the problem?

Data of the Internet of Things  
Big Data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications



# Big Data – what is it?

Is Hadoop the death of data warehousing?

*Is big data just big hype?*

**The Age of Big Data**



**Is Big Data an Economic Big Dud?**

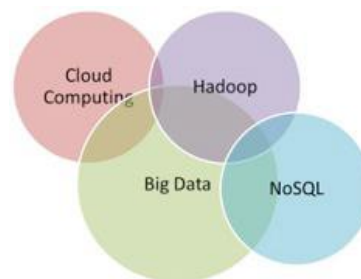
**Big Data Hasn't Jumped the Shark Yet**

**Big Data is the next Natural Resource**

**BIG DATA IS EVERYWHERE**

## Big Data – what is it?

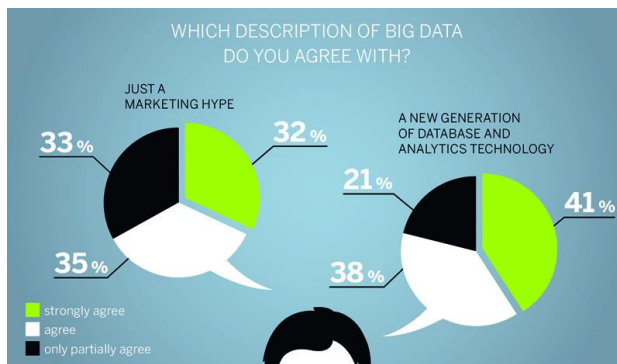
- Similar to Not-So-Big-Data ... but bigger ☺
- Data that cannot be processed using traditional tools
- Different approaches are required
- Different tools, new techniques, and technologies to facilitate working with data productively at any scale
  - Solving new problems and/or solving old problems more efficiently



13

## Big Data can be confusing

In 2012 the Experton Group carried out a study, on behalf of BT Germany GmbH & Co, on the question of how Big Data is changing business and IT. The study, "Data Explosion in Business IT", was carried out with 100 decision-makers working at companies with more than 500 employees.



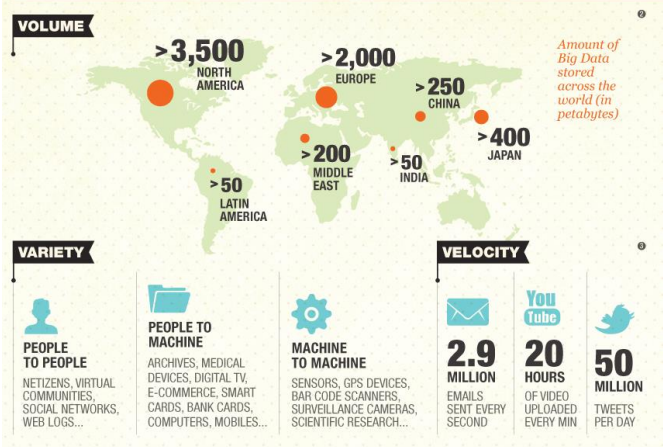
This is a somewhat small but suitably informed dataset. A total of 67% agree or strongly agree that Big Data is *Marketing Hype* yet in the same group 79% agree or strongly agree that Big Data is a *New Generation of Database and Analytics Technology*.

14

## Big Data standard definition (the 3 V's)

In 2001 Doug Laney defined data growth challenges and opportunities as being three-dimensional: Increasing **Volume** (amount of data), **Velocity** (speed of data in and out), and **Variety** (range of data types and sources). Since then many vendors and pundits have attempted to enhance his definition with clever(?) additional V's of their own. However, the three V's were intended to define the proportional dimensions and challenges specific to Big Data. Other V's like veracity, validity, value, viability, and so on are aspirational qualities of all data; they are not defining qualities of big data.

*The "three V's", i.e. the Volume, Variety and Velocity of the data coming in is what creates the challenge.*



15

## Some big numbers...

### Google

- 24 petabytes of data processed daily



### Twitter

- 400 million tweets per day
- 2.1 billion search queries per day
- 7+ terabytes of data added each day

### Facebook

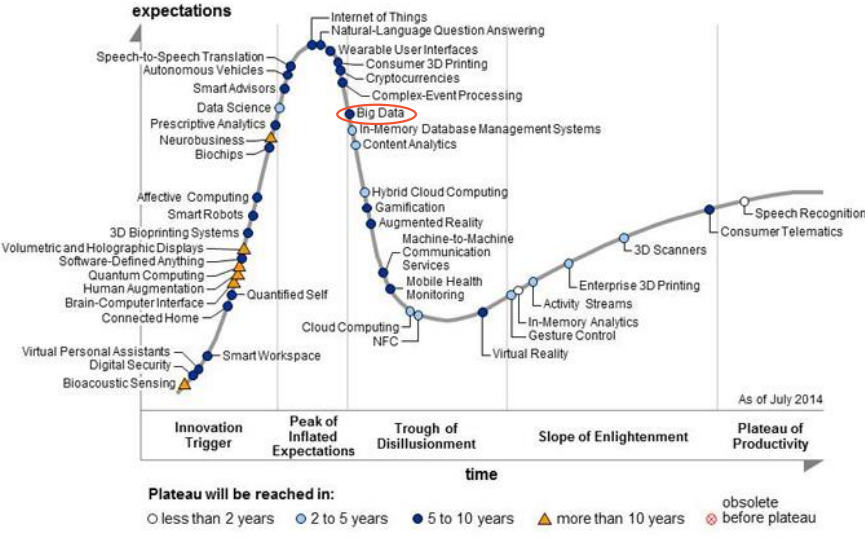
- More than 1.3 billion active users
- Processes more than 500 terabytes of data each day
- 3 billion likes and/or comments each day



16

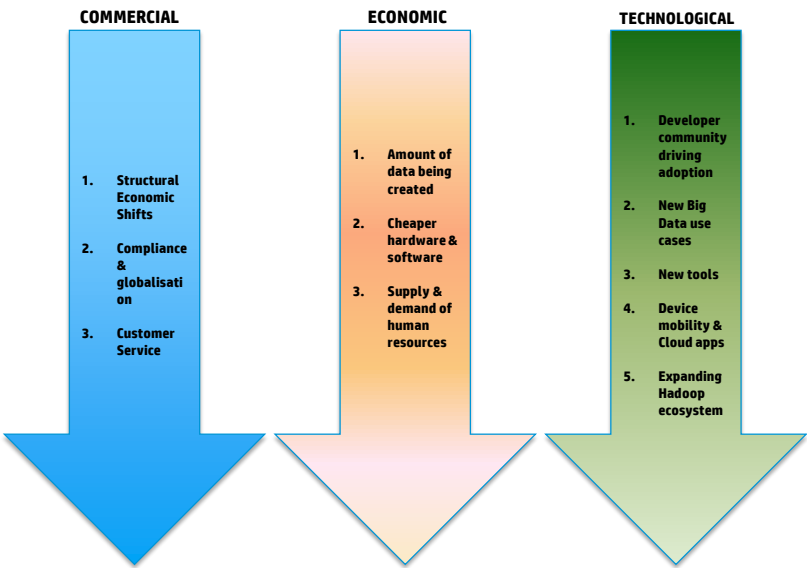


# Gartner Hype Cycle



This was as at August 2014. Big Data has now passed through the *Peak of Inflated Expectations*...

# Factors driving Big Data



## Key enablers

Some of the key enablers behind the appearance and growth of Big Data include:

- Increased storage capacities
- Increased processing power (distributed and otherwise)
- Better software tools and hardware for distributed computing
- Availability of data (the stuff is everywhere)



19

## AGENDA

- Introduction
- What is Big Data all about?
- **Some case studies**
- Technologies
- Where can OpenVMS play?
- Summary/conclusions
- Questions

# Application of Big Data analytics

Healthcare



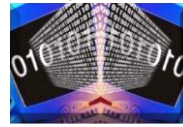
Multi-channel sales



Finance



Log Analysis



Security



Traffic Control



Mobile/Teleco



Search Quality



Manufacturing



Trading Analytics



Fraud and Risk



Retail



21

## A mobile apps example dear to my heart...

### Big Data using Erlang, C and Lisp to Fight the Tsunami of Mobile Data

*"BugSense, is an error-reporting and quality metrics service that tracks thousands of apps every day. When mobile apps crash, BugSense helps developers pinpoint and fix the problem. The startup delivers first-class service to its customers, which include VMWare, Samsung, Skype and thousands of independent app developers. Tracking more than 200M devices requires fast, fault tolerant and cheap infrastructure.*

*In the last six months, we've decided to use our BigData infrastructure, to provide the users with metrics about their apps performance and stability and let them know how the errors affect their user base and revenues.*

*We knew that our solution should be scalable from day one, because more than 4% of the smartphones out there, will start DDOSing us with data."*

See <http://highscalability.com/blog/2012/11/26/bigdata-using-erlang-c-and-lisp-to-fight-the-tsunami-of-mobi.html>



22