The World’s Fastest Big Data presentation

plus phunny pachyderm pics, and other things you should never do in a presentation

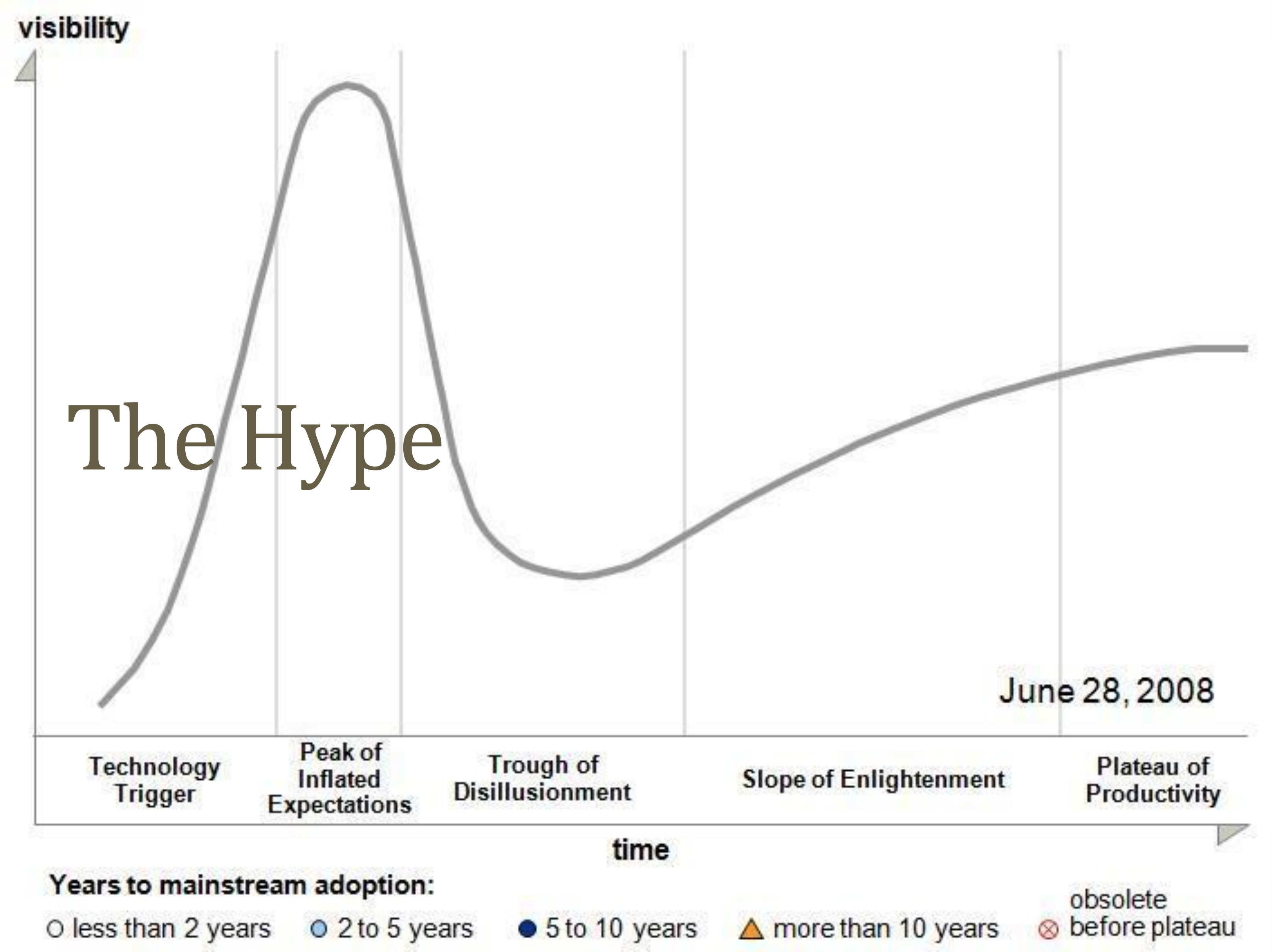
The Hype

What is it really?

~~4~~ 5 Use Cases

Hadoop Quickly

The Hype


visibility
June 28, 2008
Technology Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time
Years to mainstream adoption:
○ less than 2 years
● 2 to 5 years
● 5 to 10 years
▲ more than 10 years
⊗ obsolete before plateau

Buzzwords don't come any bigger than "Big Data," which promises to reveal the secrets hidden within big blocks of data held by companies, governments and musty old archives.

Dan Vergano, USA TODAY

http://usatoday30.usatoday.com/tech/science/columnist/vergano/story/2012-06-16/big-data/55628002/1

Big data is just about to hit its "peak of inflated expectations", according to Gartner's 2012 Hype Cycle for Emerging Technologies. Me? Expectations aren't inflated enough.

Irfan Khan, Vice President and Chief Technology Officer for Sybase

http://www.itworld.com/it-managementstrategy/293397/gartner-dead-wrong-about-big-data-hype-cycle

An interesting Google Trend Big Data factoid is that India and South Korea are rated with the highest interest, with the USA a distant third. So, all of the Big Data vendors should now focus on India and South Korea, and leave my email inbox clean.

Steve Hamby, CTO Orbis Technologies

http://www.huffingtonpost.com/steve-hamby/the-big-data-nemesis-simp_b_1940169.html

Gartner got its hype cycle wrong this time. Big data is already well along on the so-called Plateau of Productivity as its countless success stories already prove. …Today, it is those big data skeptics that we should not take too seriously.

Irfan Khan, Vice President and Chief Technology Officer for Sybase

http://www.itworld.com/it-managementstrategy/293397/gartner-dead-wrong-about-big-data-hype-cycle

Everything is on the Internet. The Internet has a lot of data. Therefore, everything is big data.

Alistair Croll

http://strata.oreilly.com/2012/08/three-kinds-of-big-data.html

What is it Really?

Is it a Technology?

Is it an Approach?

Is it Hadoop?

Is it funny elephant pictures?

Our working definition: (For Today)

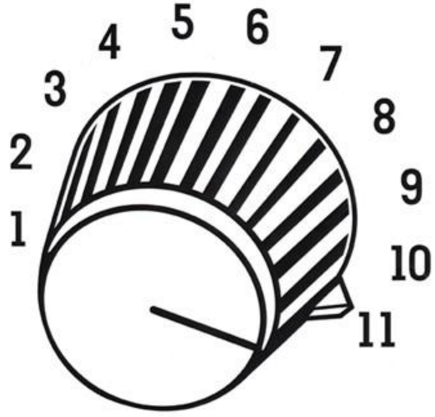
“The term ‘Big Data’ applies to information that can't be processed or analyzed using traditional processes or tools.”[1]

The 3 (sometimes 4) V’s

- Volume
- Variety
- Velocity
- Vagueness / Variability

Volume

Total amount of global data is expected to grow to 2.7 zettabytes during 2012
(2,700,000,000 TB)



35 zettabytes by 2020

FB generates 10TB daily

Twitter generates 7TB of data daily

Your organization?

Variety

Analysis includes different types of data

80% of the worlds data is not ‘traditionally structured’

Successfully leveraging of this data means an advantage

Velocity

"The real problem we're really trying to solve is fast data — a combination of large datasets, complex data models and a need to process that data at high frequency." [2]

Velocity

Not just growth rates of data repositories

The speed at which data is flowing

Key ability is to process data in motion

~~4~~ 5 Use Cases

Ad Targeting

- Based on preferences and behaviour
- Scale to millions of end users
- Data volume is enormous

Success Story (Advertising Exchange)

- Collects activity coming off of servers
- Continuously analyzes effectiveness
- Continuously refines models

Risk Modeling

- Analyze data/activity from many sources
- Consolidate large amounts of data in different silos

Success Story (Large Bank)

- A large bank took separate data warehouses from multiple departments and combined them into a single global repository in Hadoop for analysis
- Constructed a new and more accurate score of the risk in its customer portfolios.
- The more accurate score allowed the bank to manage its exposure better and to offer each customer better products and advice.
- Increased revenue and improved customer satisfaction.

Threat/Fraud Analysis

- Actual data (and all of it) vs. Samples of data
- Storage and processing intensive
- Estimated 20% of available (and useful) data being used in ‘Traditional’ approaches

Success Story (Credit Card Issuer)

- Faster refresh of models
- More encompassing (50% more broad)
- Not all of the available data was useful, but the data that was useful was able to be identified
- More accurate
- 3 Weeks → 2 Hours

Log Analysis

- Data Exhaust
- Some obviously have value (ie. clickstream)
- Some have value but it is not so obvious
- Correlate massive amount of data across enterprise for performance / security / insight / etc.
- IT for IT

Success Story (Financial Institution)

- 1 TB log data (daily amount) analyzed in 5 minutes
- Time is money
- Anticipate future problems and failures

The most important use of Big Data…

Counting Words

Hadoop Quickly

Core Hadoop

- Hadoop is a Big Data tool
- Scalable
- Uses many machines 2 – 1000's to create a distributed environment for data and processing
- Each node holds data and processes work
- Designed with hardware failure in mind
- Open Source
- Free
- HortonWorks / Cloudera / MapR / IBM / Zettaset / others

Core Hadoop

- HDFS (Hadoop Distributed File System)
  - The storage system for Hadoop
  - Each server/node stores just a small fragment of the complete data set, and each piece of data is replicated on more than one server. [3]
  - Scalable

Core Hadoop

- Distributed data processing [3]
  - Because Hadoop stores the entire dataset in small pieces across a collection of servers, analytical jobs can be distributed, in parallel, to each of the servers storing part of the data.
  - Each server evaluates the question against its local fragment simultaneously and reports its results back for collation into a comprehensive answer.
  - MapReduce is the plumbing that distributes the work and collects the results.

Core Hadoop

Moves the processing to the data instead of moving the data to the processing

MapReduce

- Jobs are written in MapReduce
- Higher level abstractions include Hive / Pig / Cascading
- Key Value pair concepts
- Datasets can be progressively mapped/shaped to desired results

Hadoop 2.x (YARN)

Questions?

Thank you BCforward for hosting, and providing lunch

BCforward®

join the evolution.

References

1. https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=sw-infomgt&S_PKG=500016891&S_CPM=is_bdebook1_biginsightsfp
2. http://www.itnews.com.au/News/317569,paypal-architect-warns-against-big-data-hype.aspx
3. https://www.cloudera.com/wp-content/uploads/2011/03/ten_common_hadoopable_problems_final.pdf

Recommended papers

- http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_exec_summary.ashx

Other Good Links

- http://www.cloudtweaks.com/2012/10/big-data-will-this-hype-last-long/
- http://nosql.mypopescu.com/post/6361838342/bigdata-volume-velocity-variability-variety
- http://www.forbes.com/sites/oreillymedia/2012/08/24/big-data-grows-up-three-spaces-to-watch-once-the-hype-subsides/