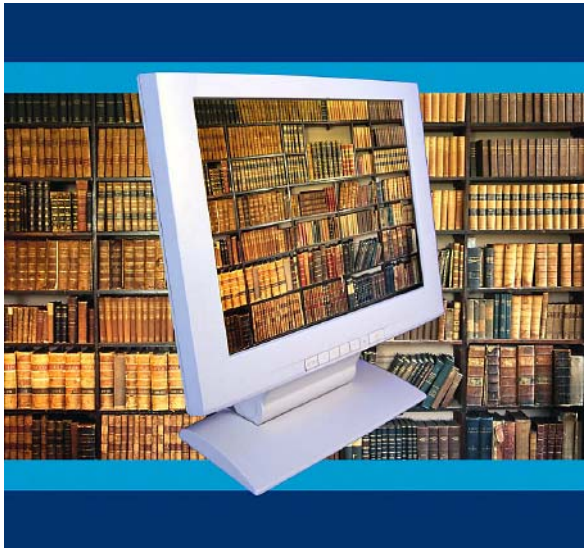
A Big Data Roadmap
For the DB2 Professional

Sponsored by:

CDB Software, Inc.



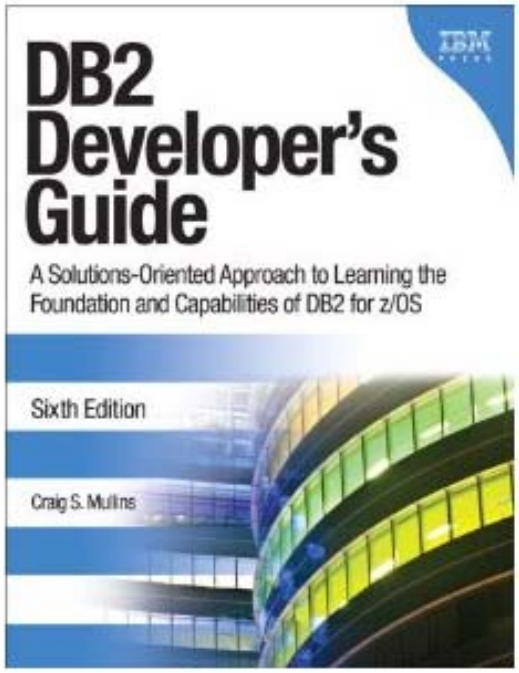
Craig S. Mullins
Mullins Consulting, Inc.
http://www.craigsmullins.com

Author

This presentation was prepared by:

Craig S. Mullins
President & Principal Consultant

Mullins Consulting, Inc.
15 Coventry Ct
Sugar Land, TX 77479
Tel: 281-494-6153
Fax: 281.491.0637
Skype: cs.mullins
E-mail: craig@craigsmullins.com
http://www.mullinsconsultinginc.com



Craig was named one of the Top 200 Thought Leaders in BigData & Analytics by AnalyticsWeek.

Agenda

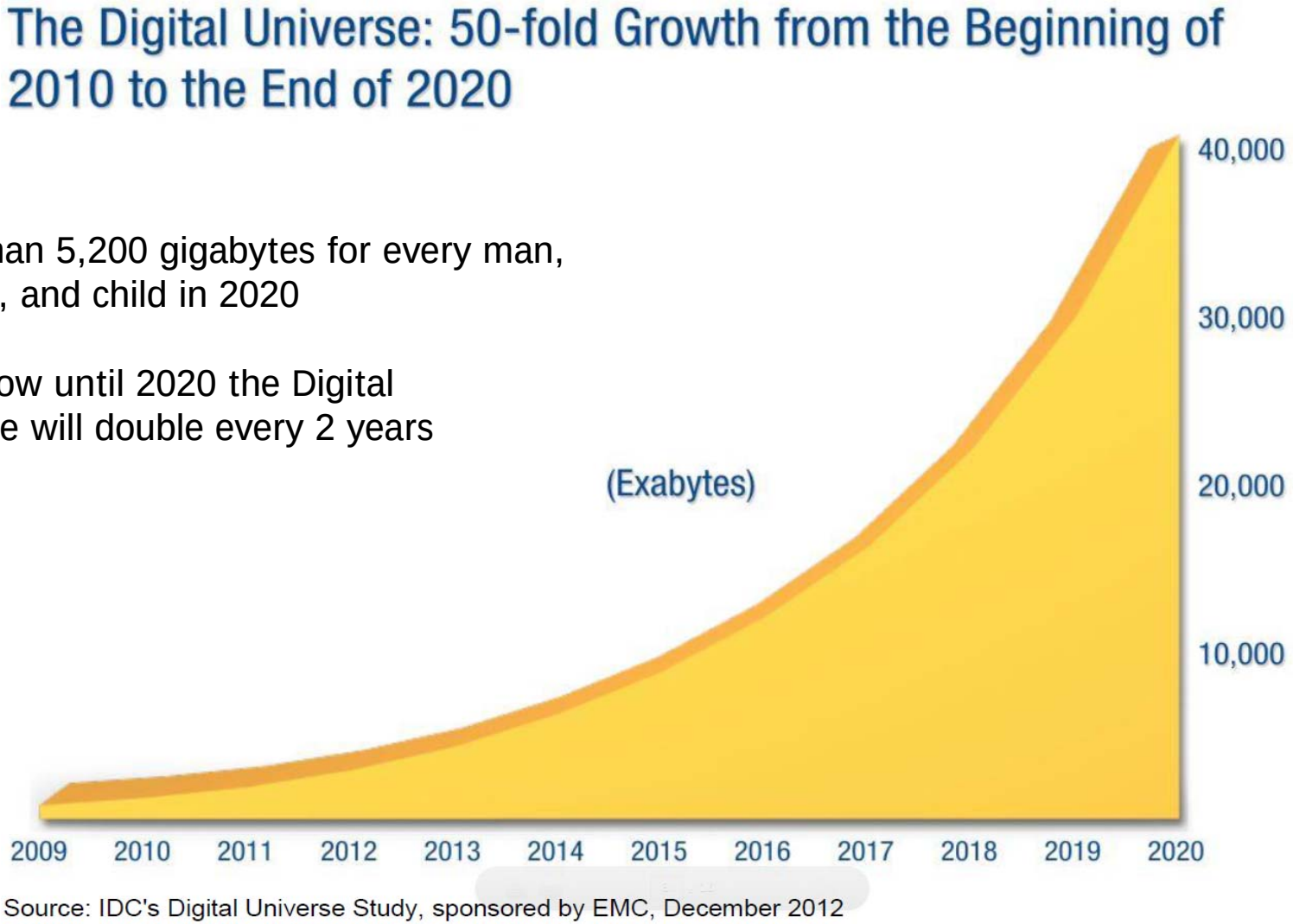
Uncover the roadmap to Big Data… the terminology and technology used, use cases, and trends.

- **Gain a working knowledge and definition of Big Data (beyond the simple three V's definition)**
- **Break down and understand the often confusing terminology within the realm of Big Data (e.g. polyglot persistence)**
- **Examine the four predominant NoSQL database systems used in Big Data implementations (graph, key/value, column, and document)**
- **Learn some of the major differences between Big Data/NoSQL implementations vis-a-vis traditional transaction processing**
- **Discover the primary use cases for Big Data and NoSQL versus relational databases**

Setting the Stage: Data Growth

More than 5,200 gigabytes for every man, woman, and child in 2020

From now until 2020 the Digital Universe will double every 2 years



Setting the Stage: Data to be Analyzed

As of December 2012, analysis by EMC and IDC estimated that there were 2.8 zettabytes of data "out there"

- For those of us zetta-challenged, that is 2.8 trillion gigabytes

Of those zettabytes of data, only 0.5% of it is analyzed in any way, shape, or form

- The analysts estimate that as much as 25% of the data has potential value though

"There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days."

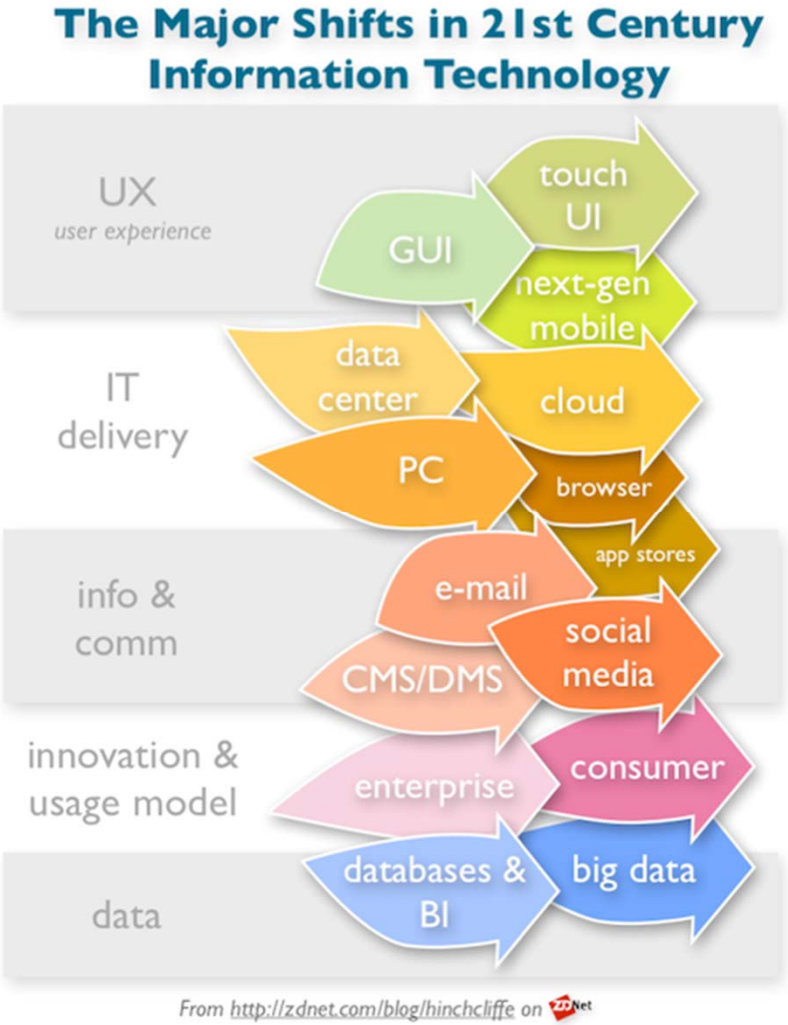
– Eric Schmidt, of Google, said in 2010

[1]http://www.emc.com/leadership/digital-universe/index.htm

Data Storage and Size Terminology

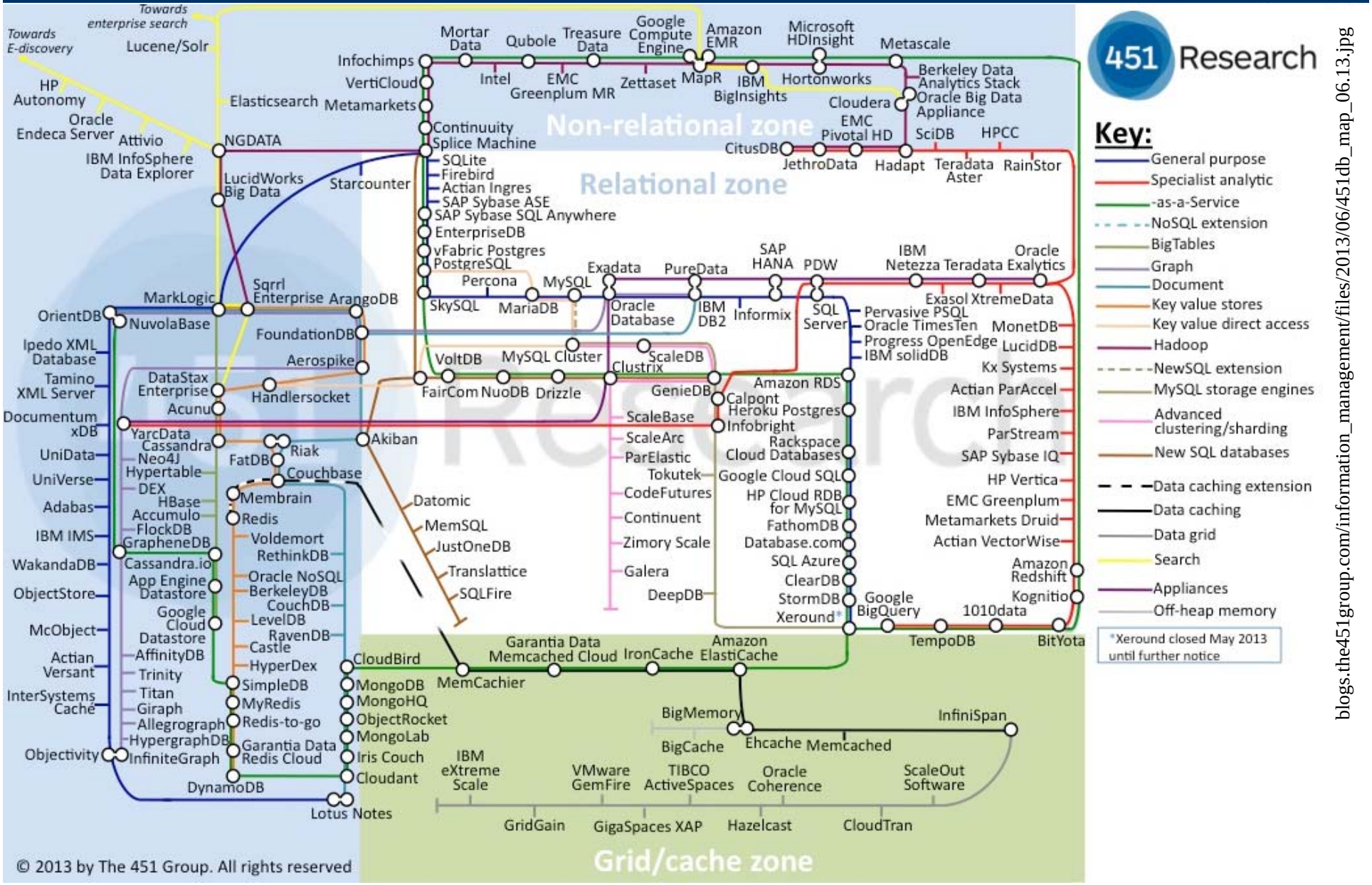
Abbreviation	Term	Size	Power of 2
B	Byte	8 bits	
KB	Kilobyte	1,024 bytes	2^{10} bytes
MB	Megabyte	1,024 KB	2^{20} bytes
GB	Gigabyte	1,024 MB	2^{30} bytes
TB	Terabyte	1,024 GB	2^{40} bytes
PB	Petabyte	1,024 TB	2^{50} bytes
EB	Exabyte	1,024 PB	2^{60} bytes
ZB	Zettabyte	1,024 EB	2^{70} bytes
YB	Yottabyte	1,024 ZB	2^{80} bytes
BB	Brontobyte	1,024 YB	2^{90} bytes

Big Data Represents a Major IT Shift



- Shift from mostly internal data to information from multiple sources
- Shift from transactional to analytical
- Shift from persistent data to data constantly on the move

The Database Landscape Map - June 2013



blogs.the451group.com/information_management/files/2013/06/451db_map_06.13.jpg

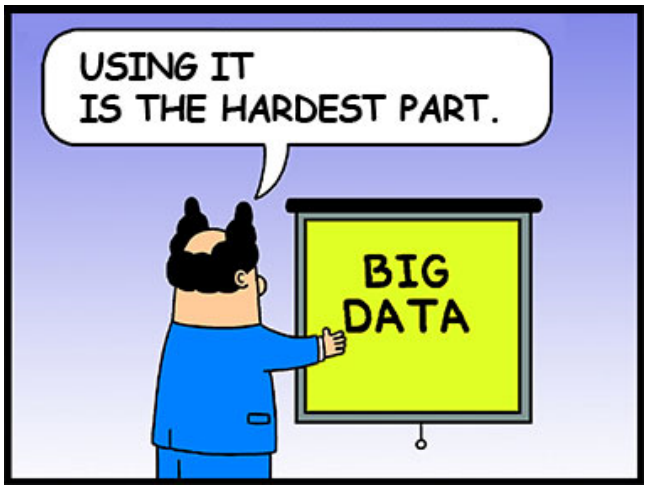
So What is Big Data?

The essence of the **Big Data** movement is being able to **derive meaning** quickly from **vast quantities of data** – both **structured and unstructured** – in order to improve business decision making.

- **Business Intelligence – structured queries**
- **Cloud Computing – access to large pools of computing power available as needed**
- **Distributed data - data is usually physically distributed across a network using inexpensive commodity hardware**
- **NoSQL and Hadoop – new data persistence methods geared for storing and processing large amounts of data**
- **Sensors – more sensors producing more data more frequently**
- **Analytical tools – for data from multiple sources and of variable types**
- **Networked devices – The number of networked devices overtook the global population of humans in 2011**
- **The Internet of Things – machine-generated data read and used by other machines**

But the Definition of "Big Data" Kinda Misses the Entire Point!

- **Requires large amounts and varieties of data…**
  - Social media
  - Website data
  - Streaming data
  - Machine-generated data
  - Etc.



- **…to be processed using Analytics**
  - Deriving useful observations from large pools of data
    - And this is the PRIMARY reason you'd ever want to attack Big Data
  - Perhaps a better definition would be Powerful Analytics?
  - But let's not quibble…

core business vs. analytical

Why Analytics is the Important Part

Big Data analytics can deliver capabilities that make your customers/users more willing to stick with your service/product

- LinkedIn – People You May Know
- Amazon – Books Recommended for You
- Netflix – Suggested Movies

Big Data Analytics can improve your profitability

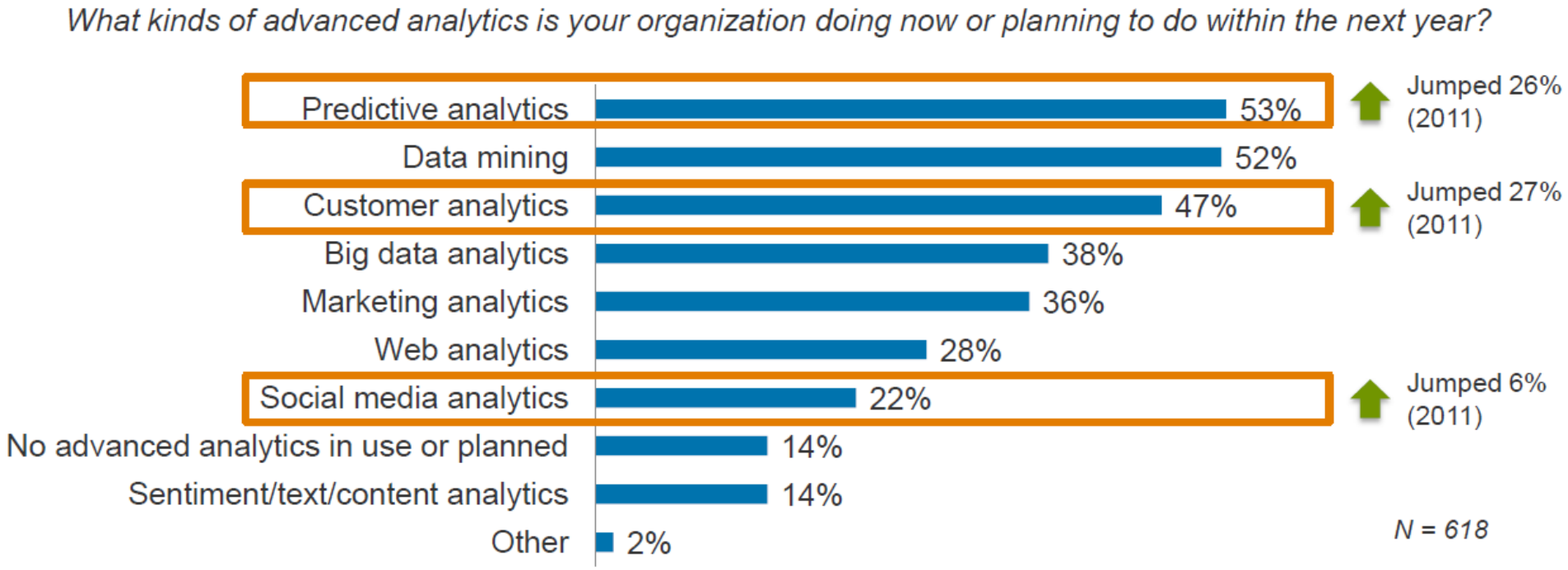
- Insurance companies send you a "gizmo" to track your driving behavior and base your payment rate on your actual driving metrics; good for company, good for "good" drivers

Big Data Analytics can uncover heretofore unknown trends and business patterns

Big Data: Embrace the Term!

- **But Big Data is a useful term for data professionals**
  - Hype and marketing, as with all "new" technologies
  - But this time it is a DATA thing!
    - So let's applaud the visibility of data in the executive ranks and try to label lots of data projects as Big Data!

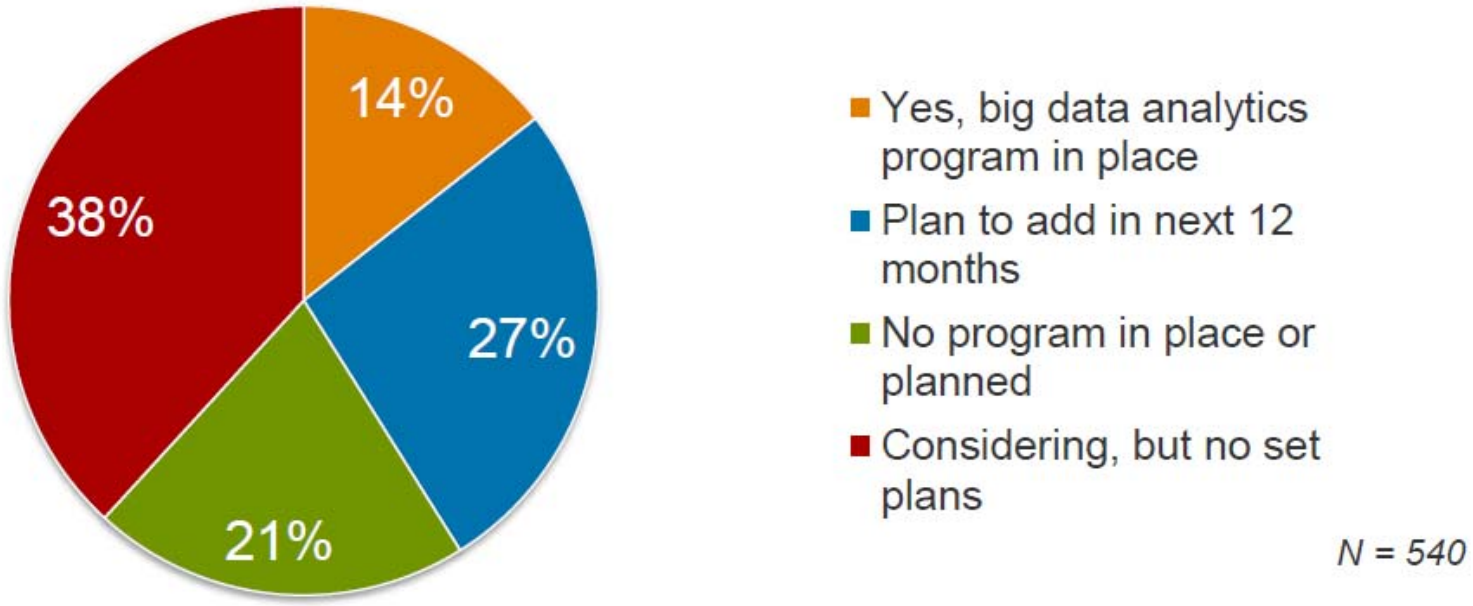
But Keep Analytics in Mind



Source: 2013 BI and Data Warehousing Survey
http://searchbusinessanalytics.techtarget.com/report/2013-BI-Data-Warehousing-Survey-Results

Big Data and Analytics Plans Increase

Do you have a big data management and analytics program under way in your organization or plan to implement one within the next year?



Source: 2013 BI and Data Warehousing Survey
http://searchbusinessanalytics.techtarget.com/report/2013-BI-Data-Warehousing-Survey-Results

IDC: HPDA and Big Data Market

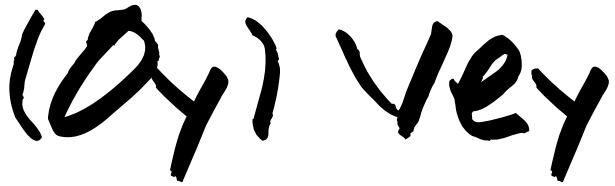
Analysts at IDC track the High Performance Data Analysis (HPDA) and Big Data Market

- **Data from the forecast for 2014 thru 2018 includes:**
  - The server market for HPDA will grow at a CAGR of 23.5%
  - The server market size will reach $2.7 billion by 2018
  - The storage market will expand to $1.6 billion by 2018

Source: Tools Journal, June 25, 2014
http://www.toolsjournal.com/integrations-articles/item/3282-idc-offers-new-forecast-for-worldwide-hpc-big-data-market

Step 1 Defining Big Data…
aka the "V"s of Big Data

- **Volume**
- **Velocity**
- **Variety**
- **Variability**
- **Verification**
- **Value**
- **Veracity**
- **Vicinity**
- **Vision**
- **Validation**



Nevertheless…

"There is no specific volume, velocity or variety of data that constitutes big. If a yottabyte is Big Data then that doesn't mean a petabyte is not?"

- Mike Gualtieri, Forrester analyst

- But let's dive in anyway and take a closer look at the Four Vs

Source for quote: http://slashdot.org/topic/bi/big-data-is-relative-term-forrester-analyst/

Volume

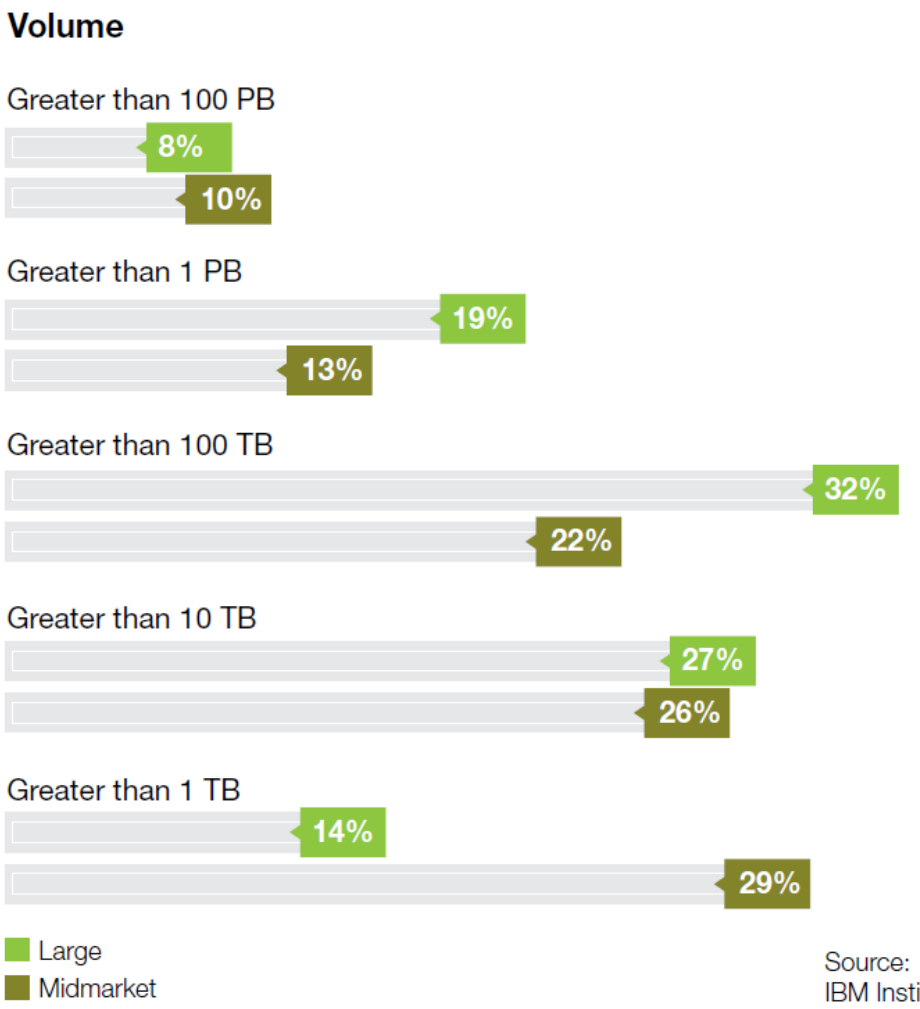
Large data volume driven by many factors

- Social media data on sites like Facebook, Twitter, and Instagram
- Data being collected from sensors, RFID, etc.
- System logs being mined for nuggets of information
- Larger unstructured data like images, audio, video, and other data (medical, seismic, genome, etc.)
- Streaming data

Internally and externally generated data

In some cases, it may be too voluminous to store for any length of time

How Big is Big?



Volume is the characteristic most associated with big data, but there is no set definition so drawing a line is arbitrary.

Source: Big Data @ Work survey, a collaborative research survey conducted by the IBM Institute for Business Value and the Saïd Business School at the University of Oxford. © IBM 2012

Velocity

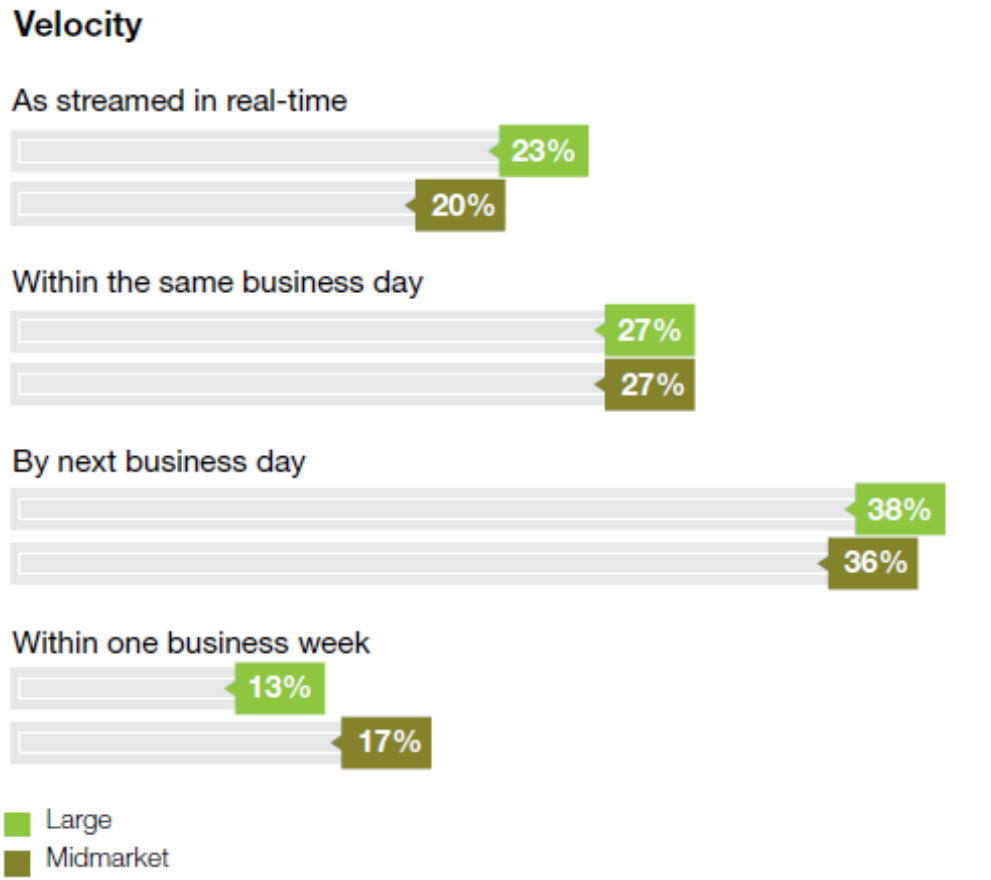
The speed at which data is being generated and collected has increased… and it is continuing to increase

One aspect of velocity is the progression from batch up to real-time

Another aspect to consider is the on-going, incessant generation of data:

- On social media
- From devices
- By sensors

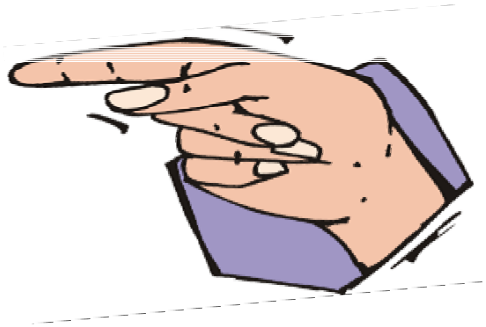
What Velocity is Big / Fast?



Source: Big Data @ Work survey, a collaborative research survey conducted by the IBM Institute for Business Value and the Saïd Business School at the University of Oxford. © IBM 2012

Variety

More types of data are being generated and stored than ever before

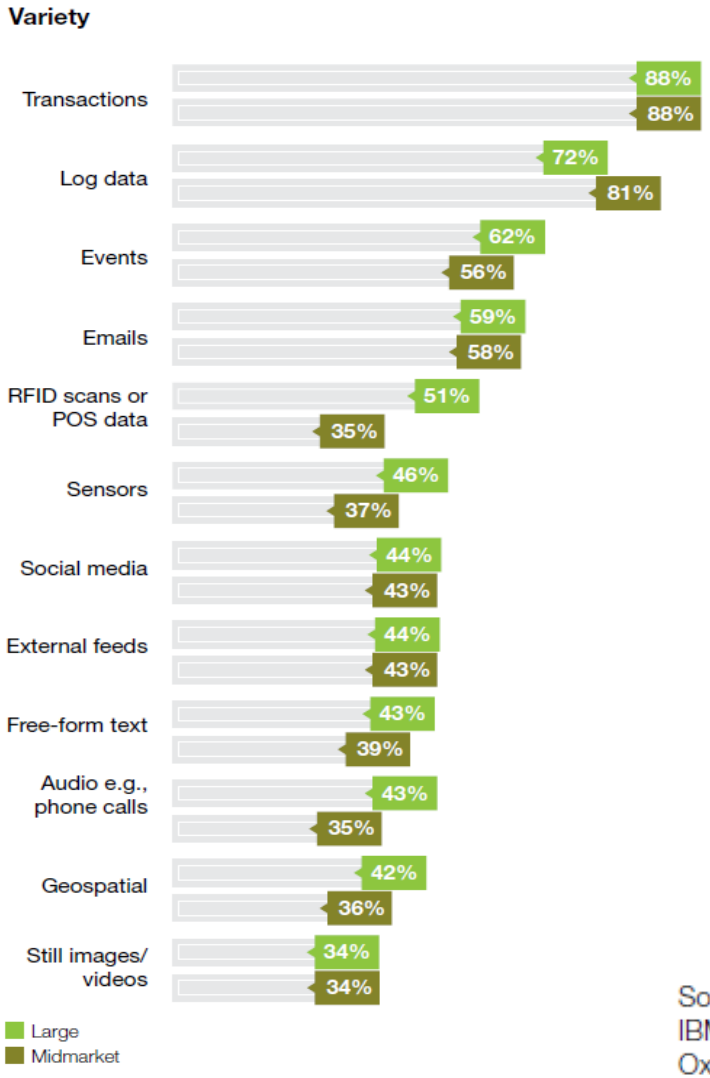


- Structured data
  - Relational database, flat files, VSAM, etc.
- Unstructured data
  - Audio, image/photo, video
  - Office documents (word processing, spreadsheets, presentations)
  - Social media
  - Web sites
  - Clickstream data
  - Wikis and blogs
  - IT data
    - Log files, Configuration, monitoring, audit and security data
  - Spatial and GPS coordinates
  - Machine-generated data

ETC.

Data Variety Example

What About Variety?



Source: Big Data @ Work survey, a collaborative research survey conducted by the IBM Institute for Business Value and the Saïd Business School at the University of Oxford. © IBM 2012

Variability

The fourth V is not as often discussed

The general idea here is that so much data must be processed that there are things that relational DBMS cannot easily accomplish

So we need to add new technology

Usually NoSQL… sometimes not even a DBMS (e.g. Hadoop)

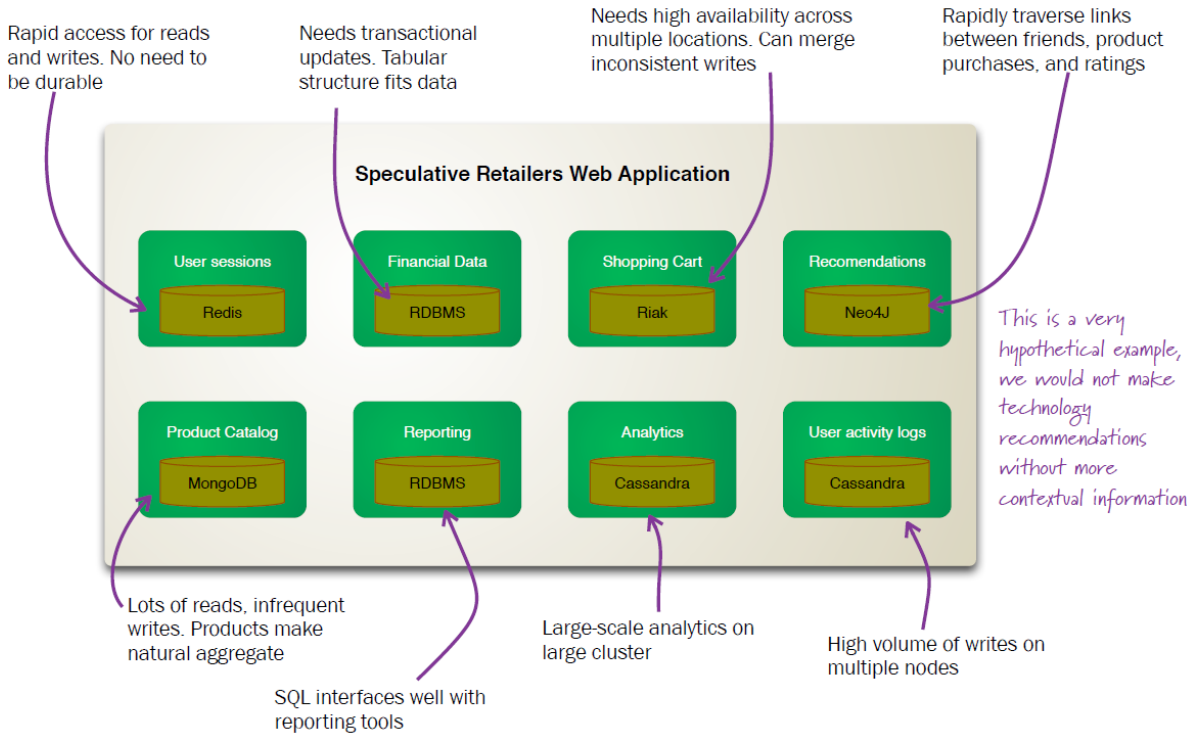
Which brings us to my "favorite" Big Data term →

- Polyglot Persistence

Polyglot Persistence

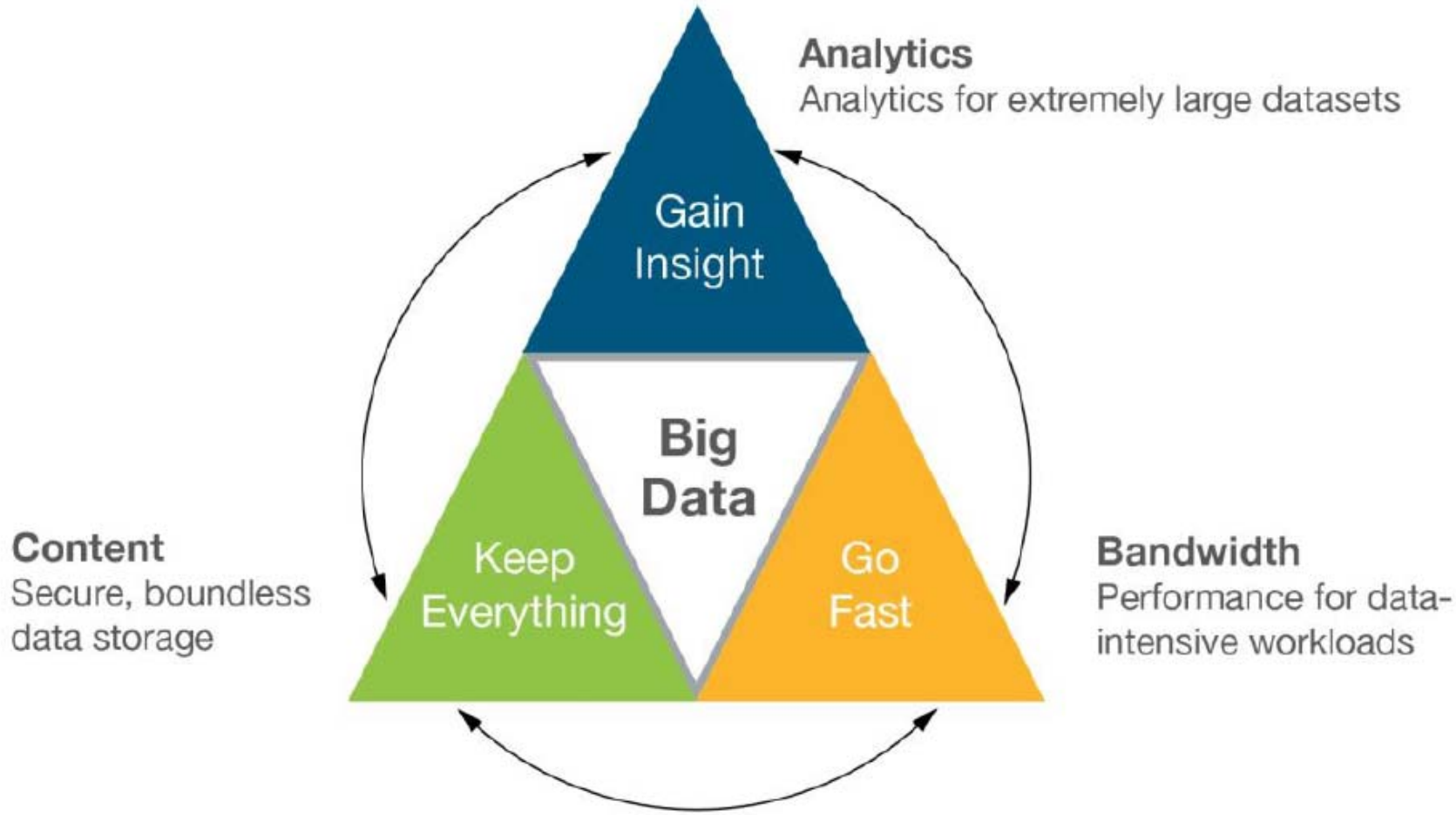
Polyglot Persistence is basically using multiple data storage technologies and techniques

- Better matching the application requirements to the data storage mechanism



Graphic source: The Future is Polyglot Persistence by Martin Fowler and Pramod Sadalage
http://martinfowler.com/articles/nosql-intro-original.pdf

Another Definition of Big Data



Source: The ABCs of Big Data by Richard Treadway & Ingo Fuchs NetApp, March 2012, WP-7147

So What is Big Data?

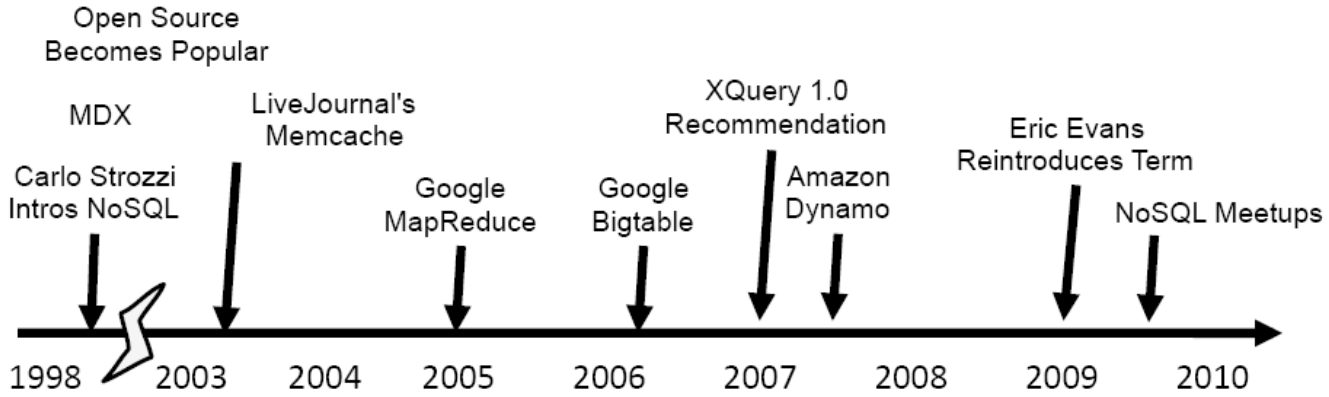
Do You Know Big Data When You See It?

What is NoSQL?

The NoSQL movement, which started out as "No SQL" has become "Not Only SQL"

- Non-relational and hybrid database systems
- NoSQL is based on the concept that relational databases are not the right database solution for ***all*** problems.

The World Wide NoSQL market is expected to reach $3.4 billion by 2018, with a compound annual growth rate of 21%.[1]



[1] Source: NoSQL Market Forecast 2013-2018, Tabular Analysis, Market Research Media, Ltd., July 2013

NoSQL Drivers

- **More users – 1000 users used to be a lot and 10000 was extreme; the web renders these numbers quaint**
- **More data – difficult to scale to terabytes of data for traditional relational database applications**
- **Different data – unstructured data is not easily handled by relational databases**
  - Documents, social media, etc.
- **More analytics – different use cases can require different technology**
- **Simplicity – at least in terms of the features supported by the applications/systems**
- **Rapid development – schema-free databases deliver flexibility for quicker development**

What Do You Mean "Schema Free"?

Of course, there must be some type of schema, or the data is not very useful, right?

But there need not be extensive knowledge of the schema before we get the data…

- The system automatically determines how the data should be indexed as it is loaded into the database

No in-depth up-front logical data modeling like with relational/SQL database systems

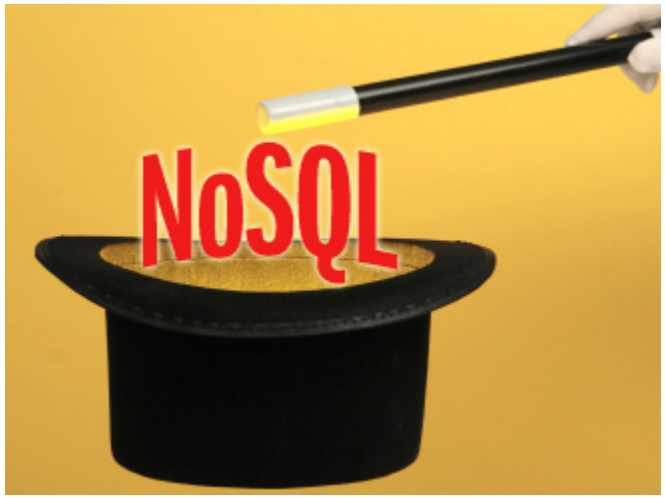
- But you do need to know some things about the data

Adding or changing data elements is not disruptive

- Different records (objects) may have different fields that are not in every record

Types of NoSQL Database Systems

1. Column Store
2. Document Store
3. Key/Value
4. Graph



Column Store Databases

Relational databases focus on access by row…

- Column stores focus on the column.

Column Store Use Cases

- Applications that count and categorize data
- Event logging applications
- High-speed queries are required

Document Store Databases

Document database systems store and retrieve at the document level

- A document is an object
  - For example XML or JSON
- The document is self-describing
- Every document does not have to be exactly the same
  - Table == Collection
  - Row == Document

```
{
    "firstName": "John",
    "lastName": "Smith",
    "age": 25,
    "address": {
        "streetAddress": "21 2nd Street",
        "city": "New York",
        "state": "NY",
        "postalCode": 10021
    },
    "phoneNumbers": [
        {
            "type": "home",
            "number": "212 555-1234"
        },
        {
            "type": "fax",
            "number": "646 555-4567"
        }
    ]
}
```

MarkLogic

RAVENDB

Couchbase

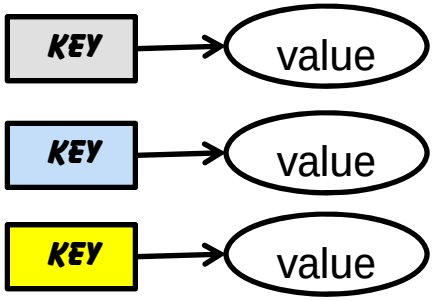
mongoDB

Document Store Use Cases

- Event logging applications
- Content management systems
- Blogging platforms
- Web analytics and real-time analytics
- E-Commerce applications

Key Value (K/V) Databases

Simplicity



- Key + Rest of Data
  - Find the Rest of Data using the Key
  - No alternate keys or indexing
- Values can be any type of data
- Scalable, Fast, simple API
- Query only by key – cannot query based on the rest of content
  - All queries return the value in a lump, no way to just return some of the "value"
  - ...think about a dictionary

Example K/V databases...

riak

AEROSPIKE

APACHE HBASE

redis

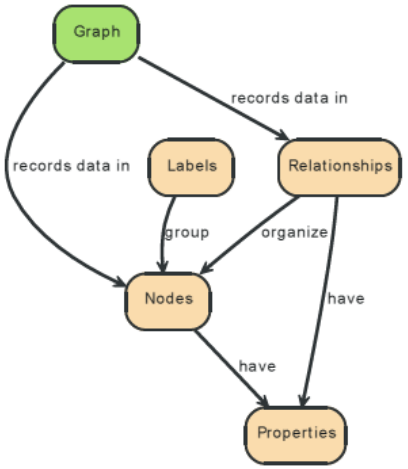
KV Use Cases

- Managing session information in web applications
- Managing player session details in massive multi-player games
- Managing the shopping cart for an online buyer

Graph Databases

Good for storing information about relationships between things where the relationship between two items in the database is at least as important as the items themselves.

- Graph databases are very good for analyzing how closely things are related, how many steps are required to get from one point to another.



GraphBase

Neo4j

SPARKLE DB

Graph Use Cases

Analyzing relationships between people in social media such as in LinkedIn, Facebook, and Twitter are typical use cases for graph databases.

How many "degrees of separation" are there between two people?

- Terrorist cell identification
- Organizational social media

NewSQL

There is also this "thing" called NewSQL defined as:

- Relational/SQL DBMS
- Scalable like NoSQL but ACID like SQL

Oftentimes coupled with other newer capabilities like:

- In-memory database
- Transparent sharding

Examples include:

- VoltDB
- TransLattice
- NuoDB
- Clustrix
- SAP HANA

SAP HANA

NoSQL Worries of the Relational Pro

Most NoSQL systems take advantage of many low-cost computers tied together with high-speed networking

Because of their distributed nature there are challenges involved with managing system failures and ensuring reliability

- Requires continuous management of components

Can promote a lack of planning about the database schema (may, or may not, be an issue)

- May wind up with lots of users with different fields and a mess of data that is hard to understand, let alone control

Consistency is an issue… Does it support ACID?

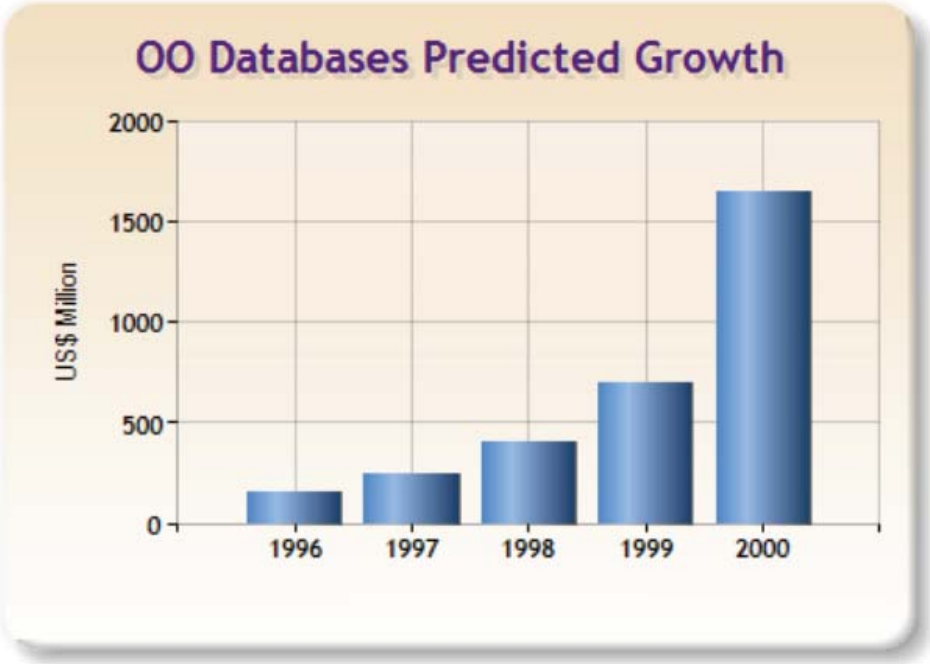
- Some NoSQL systems support ACID, but not most. We'll discuss this in more detail later in the presentation.

But What is "Wrong" With Relational?

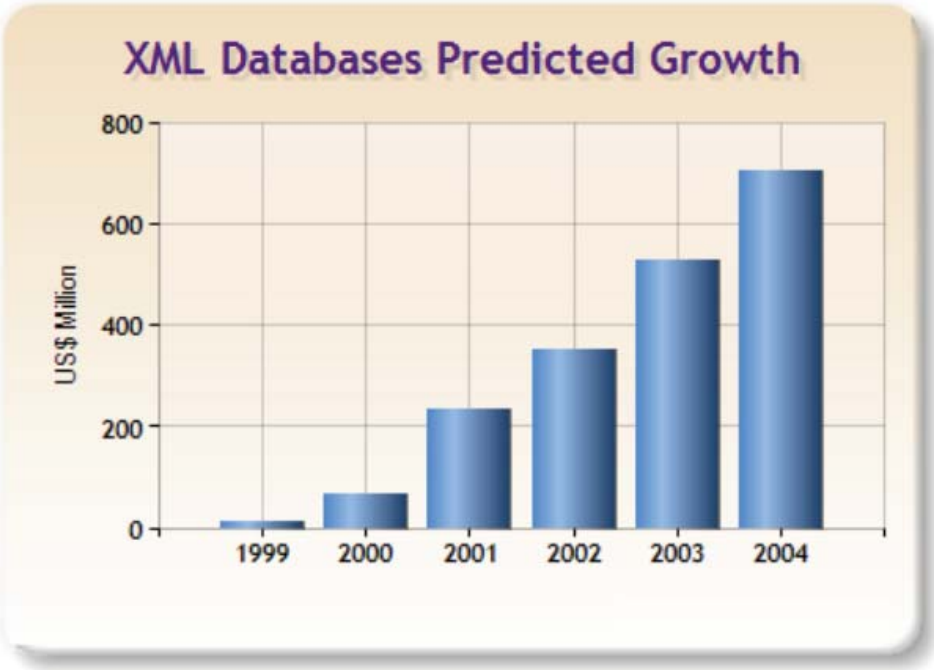
Nothing, but think about the relational world today…

- Relational databases are so ubiquitous in most organizations these days that many people may not even be aware that there are other types of databases, let alone when using another database *might* be preferable.
- Relational databases perform transaction update functions very well, particularly handling the difficult issues of consistency during update (ACID).
- Production strength relational databases can handle the complexity of two phase commit capability, where one business transaction affects multiple databases and tables, and all updates have to be effected at the same moment.
- However, relational databases apply much of the same overhead required for complex update operations to every activity, and that can handicap them for other functions.

Remember Your Database History



1990s ▸ Object Databases



2000s ▸ XML Databases

Not Replacement, but Augmentation

NoSQL and other Big Data technology is not going to replace relational and SQL database systems

- Still best technology for traditional OLTP systems

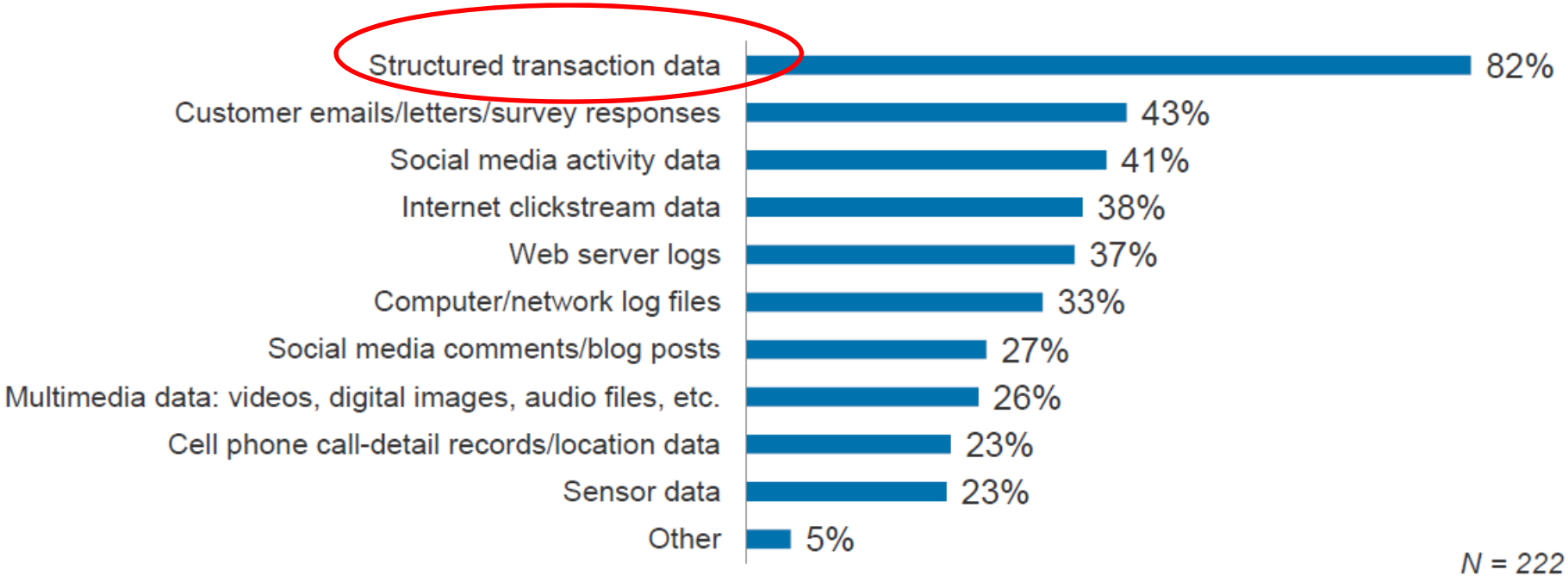
The idea is to augment, adding NoSQL and other technologies where and when they make sense

But realize also that relational has its place in the Big Data and Analytics world…

Remember Polyglot Persistence

Types of Data in Big Data Projects

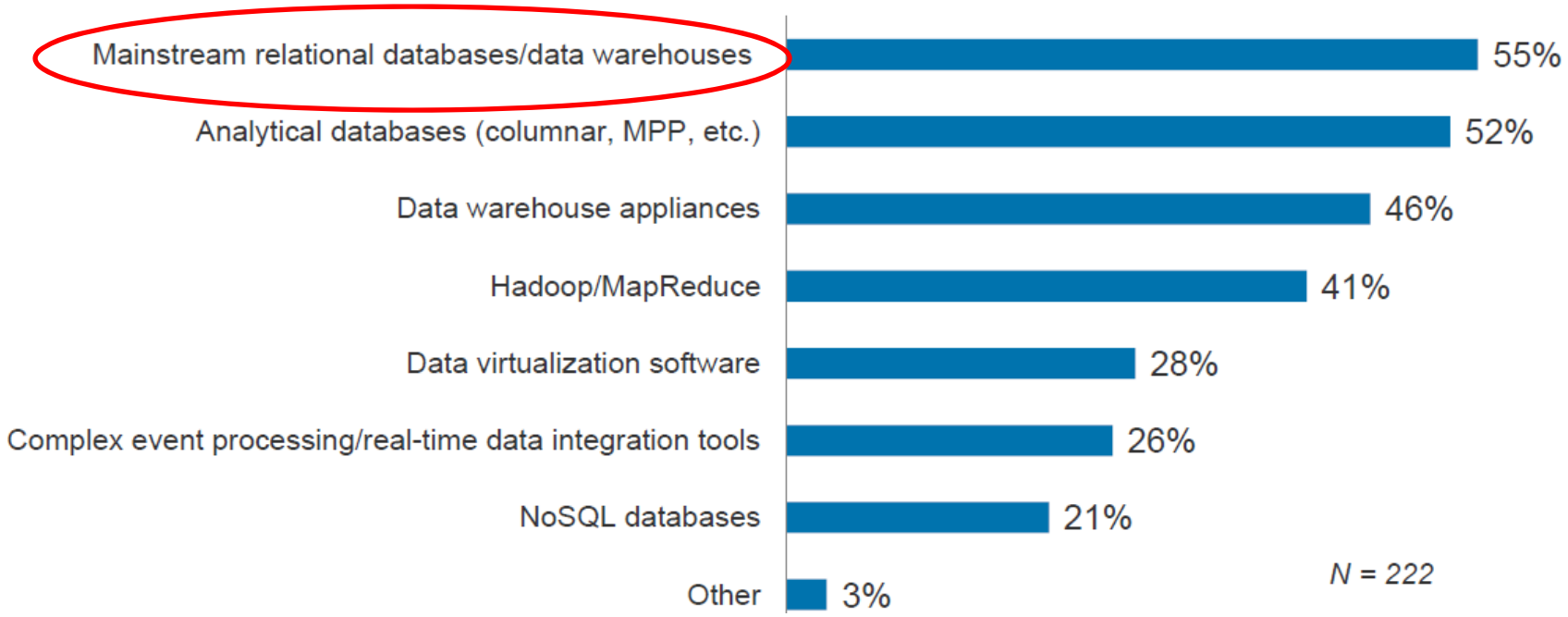
What types of data does your organization collect or plan to collect as part of its big data program?



Source: 2013 BI and Data Warehousing Survey
http://searchbusinessanalytics.techtarget.com/report/2013-BI-Data-Warehousing-Survey-Results

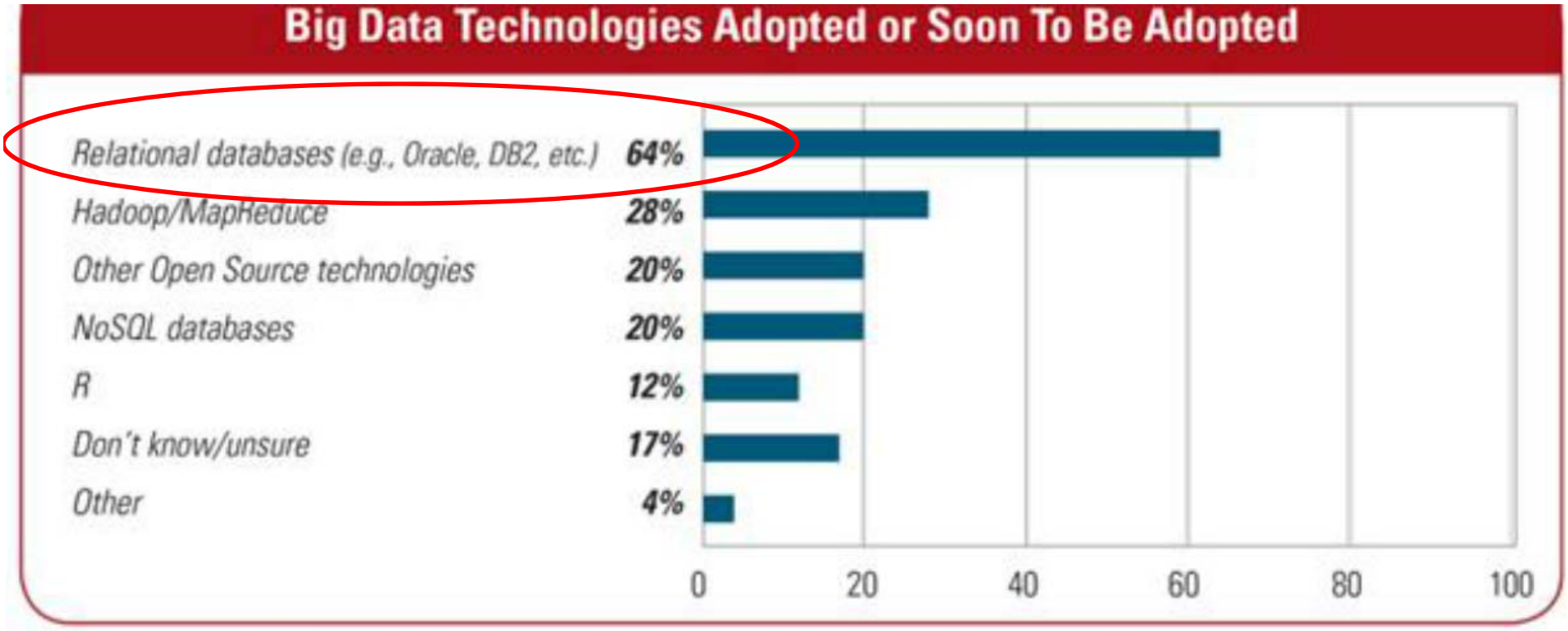
Architecture Used for Big Data Projects

What technologies does your organization use or plan to use to support its big data environment?



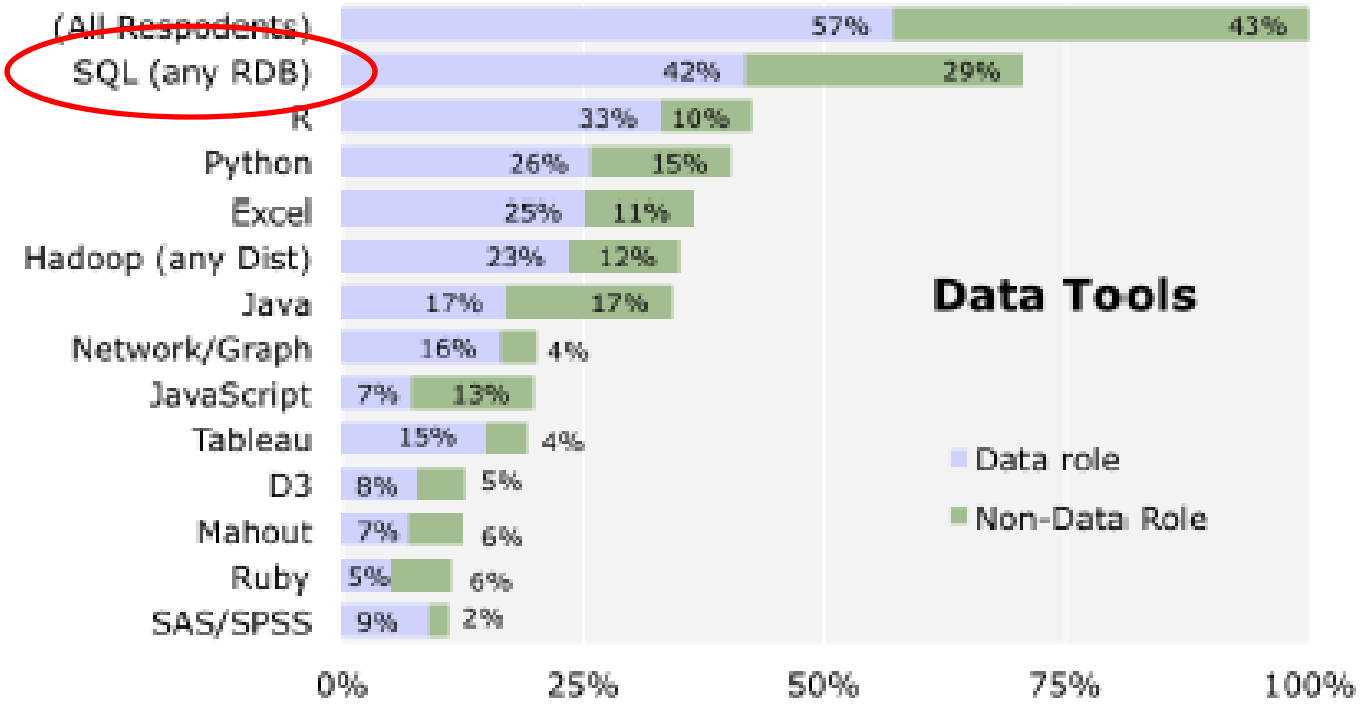
Source: 2013 BI and Data Warehousing Survey
http://searchbusinessanalytics.techtarget.com/report/2013-BI-Data-Warehousing-Survey-Results

Technologies Adopted for Big Data Projects



Source: Survey of 304 data managers and administrators who are subscribers to *Database Trends & Applications*, 2013 BIG DATA OPPORTUNITIES SURVEY, Unisphere Research, May 2013.

SQL Still Top Tool of Data Scientists



Source: 2013 Data Science Salary Survey, by O'Reilly conducted at the Strata Conference

So When Does NoSQL Make Sense?

NoSQL will NOT replace relational/SQL database systems

- The major DBMSes (DB2, Oracle, SQL Server) are entrenched in most organizations and adeptly handle OLTP requirements (as well as many analytical reqmts)
- NoSQL can be added on a project basis where and when it makes sense

Adding NoSQL database systems can make sense as part of an enterprise infrastructure that can handle unstructured and structured data

- Remember when Object databases were going to replace relational?

Major relational/SQL database systems (like DB2) will incorporate NoSQL capabilities over time

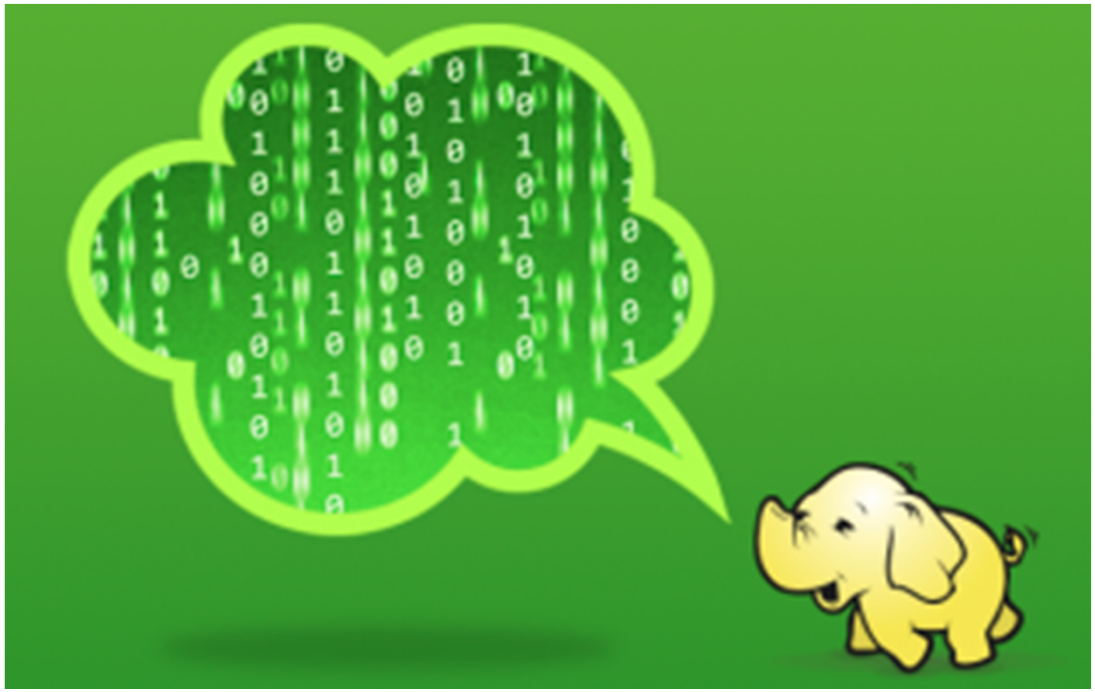
- Note the column store capabilities of BLU -- DB2 10.5 for LUW

Okay, Let's (Briefly) Look at Hadoop

Hadoop

HDFS

MapReduce



What is Hadoop?

Hadoop *is not* a DBMS

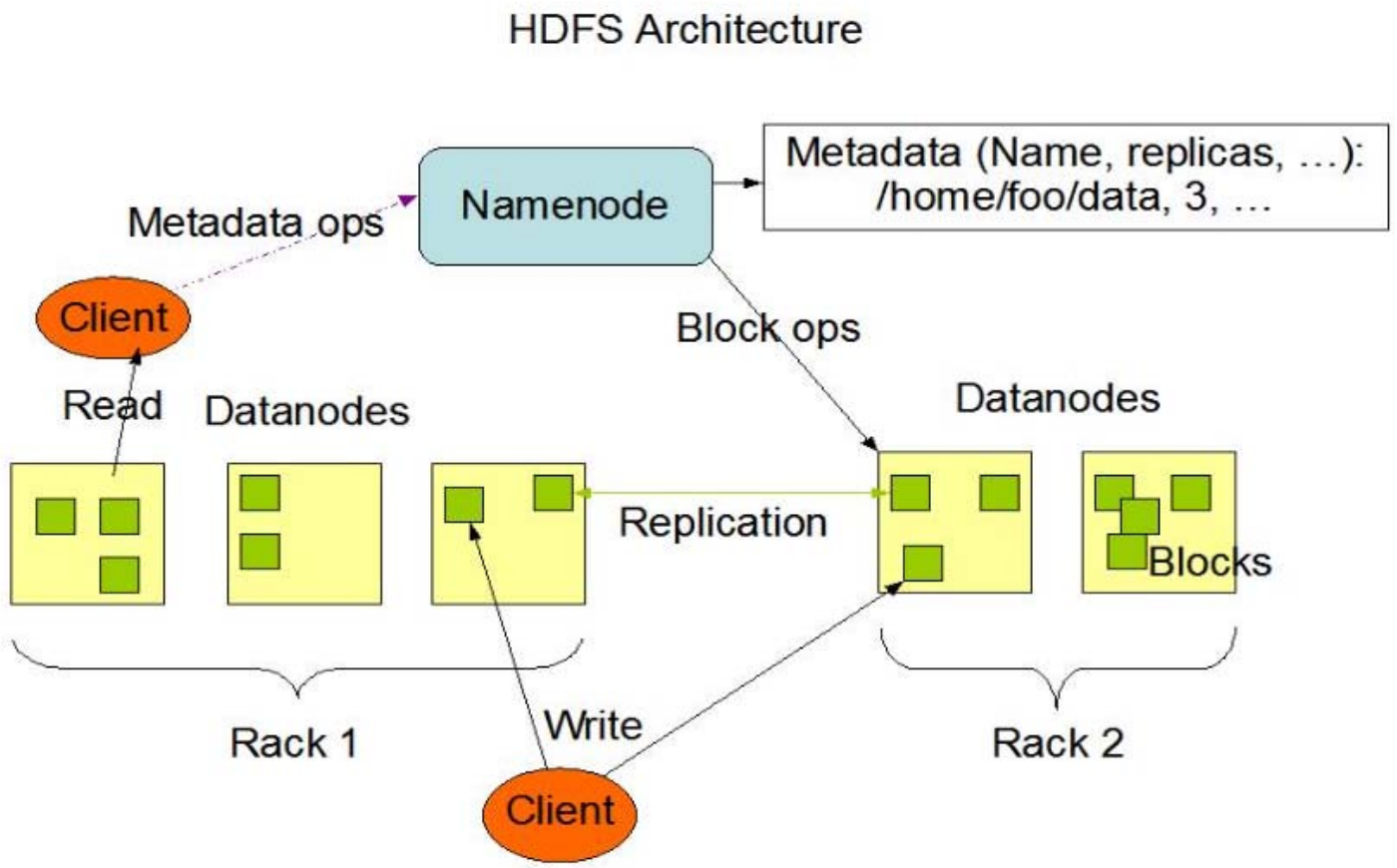
- The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across **clusters of computers** using simple programming models.
- It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.
  - Fault tolerant – if one node goes down processing can continue
  - Hadoop is not designed for real-time access → batch
  - Many vendors sell commercial implementations

But HBase *is* a DBMS built on Hadoop

- Cloudera is also built on top of Hadoop; Cloudera touts its offering as an "enterprise data hub" instead of a DBMS
  - Other commercial companies incorporate Hadoop into their commercial offerings including Greenplum, Hortonworks, IBM, Intel, and many others

http://hadoop.apache.org/

The HDFS Architecture
(Hadoop Distributed File System)



Source: *Why Hadoop is important in handling Big Data?* by Jagadish Thaker
http://bigdataweek.com/2013/11/26/why-hadoop-is-important-in-handling-big-data/

OK, But I'm Still not Clear… How Would I Use Hadoop?

The HDFS spans all the nodes of a Hadoop cluster

- In essence, making it one big file system

Hadoop uses the MapReduce framework to understand and assign work across a network of machines

- Parallel tasks across many distributed nodes

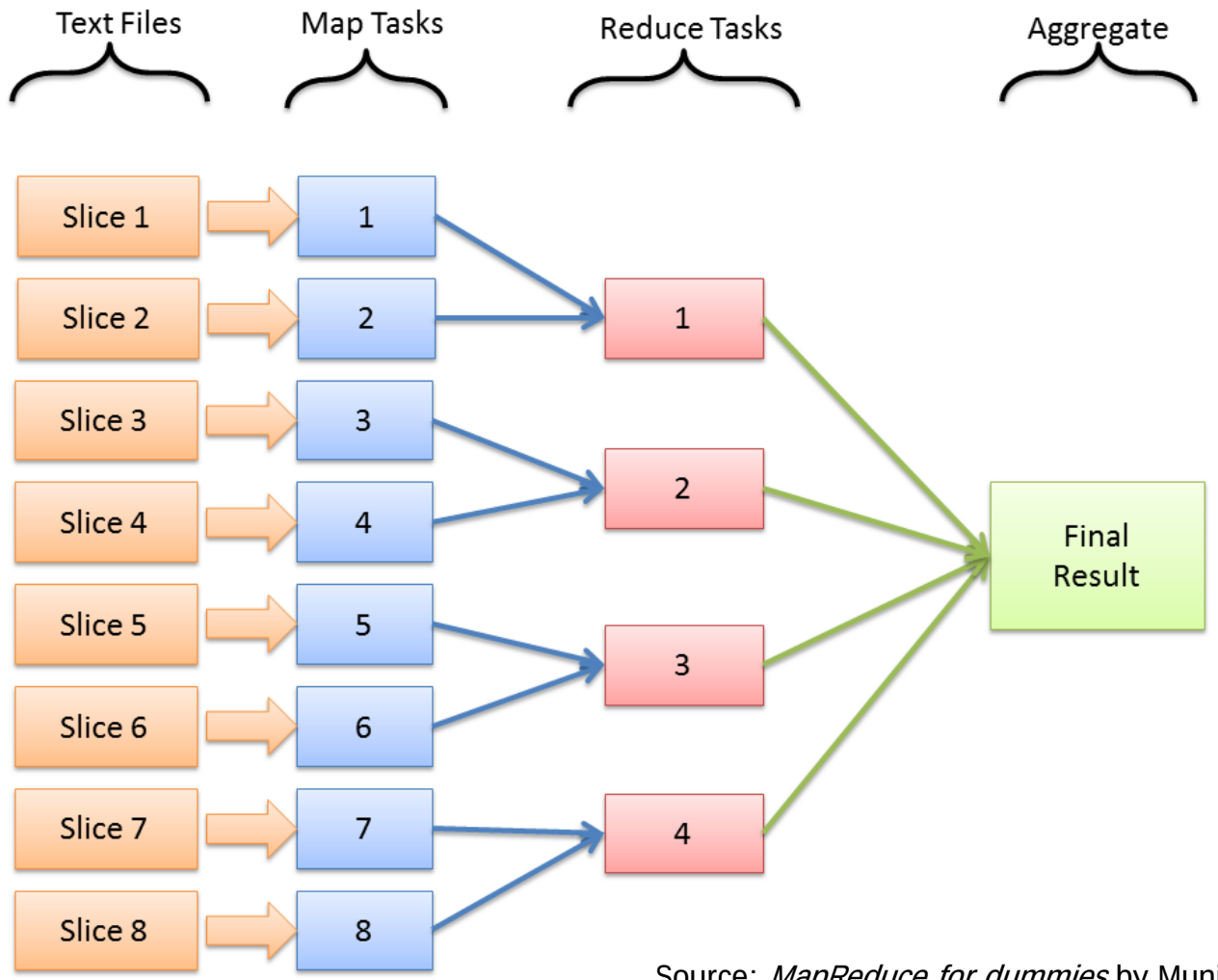
Consists of two basic steps: Map and Reduce

- The **Map** step splits the input into pieces
  - Worker nodes process individual pieces in parallel
  - Each worker node stores its results in the local file system
- The **Reduce** step aggregates the data from the multiple worker nodes
  - The reduce tasks can run in parallel

Nevertheless, remember Hadoop is batch, not real time

http://searchbusinessanalytics.techtarget.com/feature/Hadoop-still-too-slow-for-real-time-analysis-applications

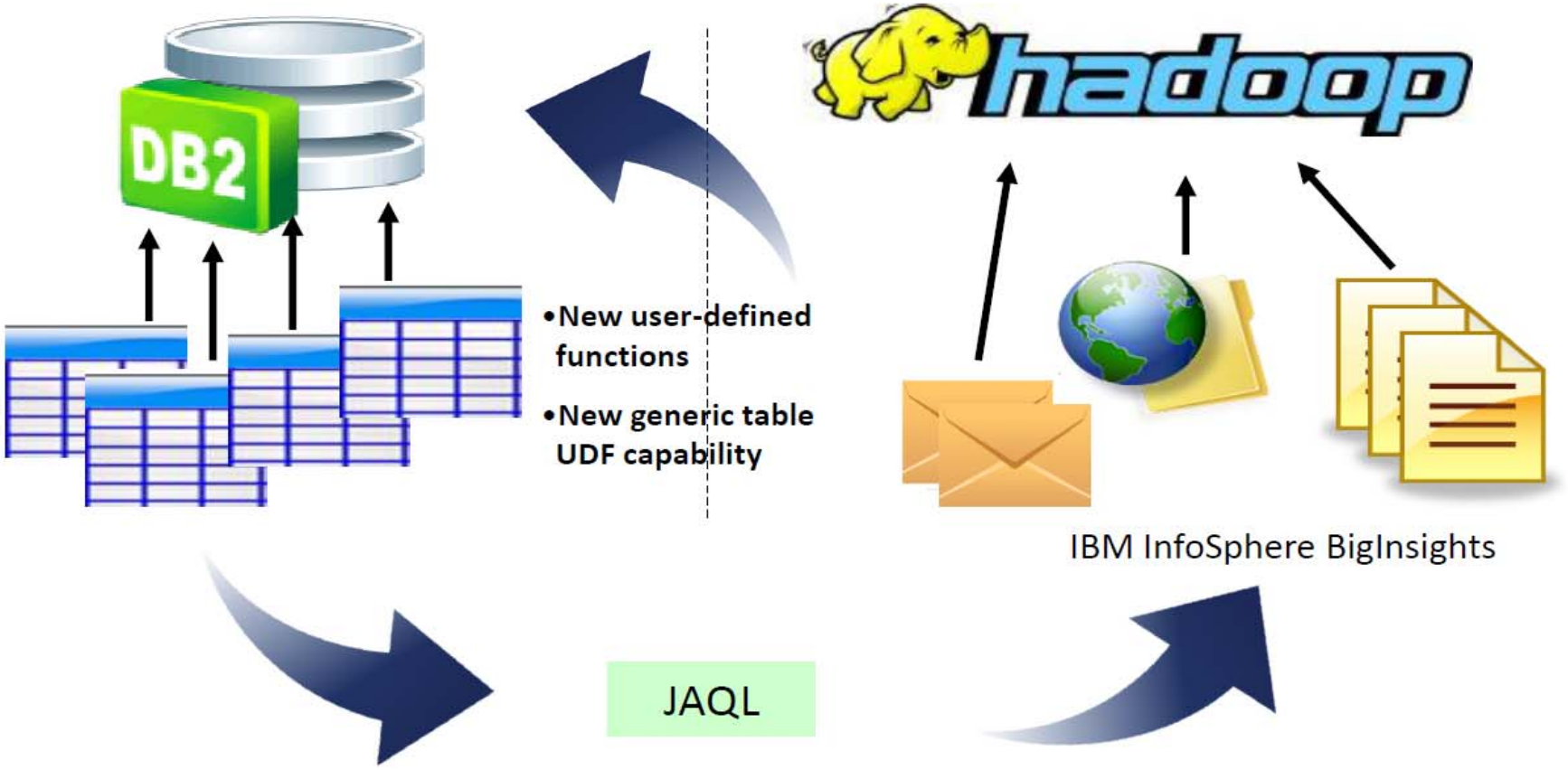
MapReduce



Source: *MapReduce for dummies* by Munish K. Gupta
http://www.techspot.co.in/2011/07/mapreduce-for-dummies.html

DB2 Connectors to Hadoop

DB2 is providing the connectors and the DB capability to allow DB2 applications to access data easily and efficiently in Hadoop

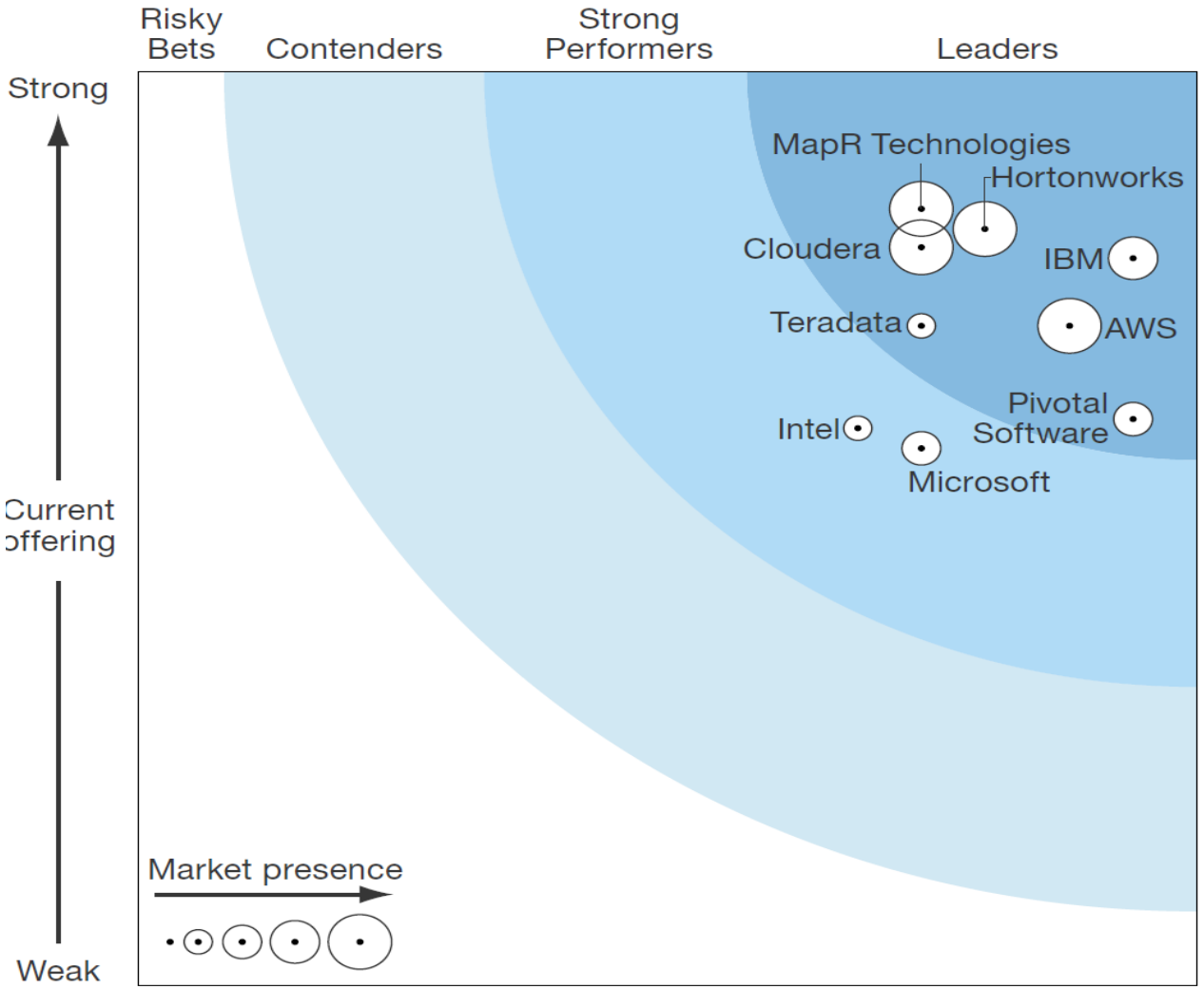


Source: DB2 11 for z/OS: Technical overview - Part 2
IBM presentation by John Campbell, Feb 19, 2014

IBM is Strong on Hadoop

Forrester Wave

- Lots of leaders
- None dominate
- But…



"IBM has more than 100 Hadoop deployments, some of which are fairly large and run to petabytes of data."

Source: The Forrester Wave: Big Data Hadoop Solutions Q1 2014
Mike Gualtieri and Noel Yuhanna, Feb 27, 2014

Quick Introduction to Other Pertinent Big Data, NoSQL and Related Terminology

- BASE
- CAP Theorem
- Sharding
- Stream Computing
- Visualization
- JSON
- R
- Hive
- Pig

ACID versus BASE

ACID

- **Atomic** – every transaction either completes entirely or fails
- **Consistent** – data is always in a valid state
- **Isolated** – concurrent transaction execution results in same system state as serial transaction execution
- **Durable** – once committed, data is always available (even in the event of partial system failures)

BASE

- **B**ase **A**vailability of **S**oft state, **E**ventually consistent...

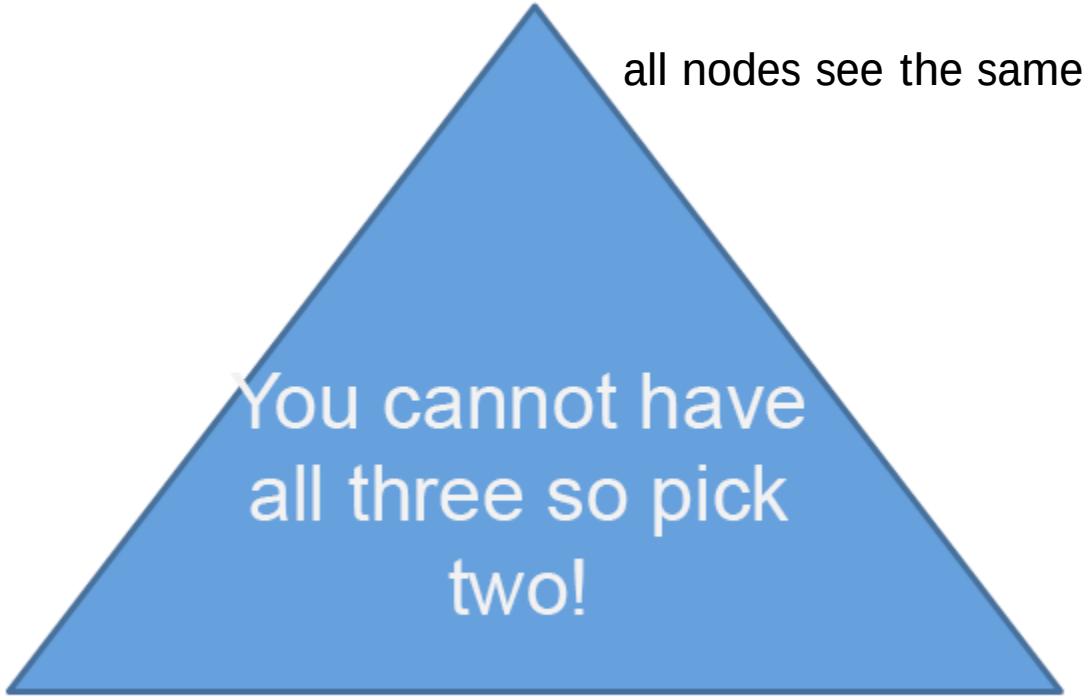
Keep in Mind
an eventually consistent system can return *any* value before it converges.

CAP Theorem

aka Brewer's theorem

Consistency

all nodes see the same data at the same time



Availability

a guarantee that every request receives a response (whether successful or not)

Partition Tolerance

system continues to operate despite arbitrary message loss or failure of part of the system

CAP Systems

CA – Consistent and Available

- Example: a single site, single database with no auto-sharding
- If you must partition: stop, partition, restart

CP – Consistent with Partition Tolerance

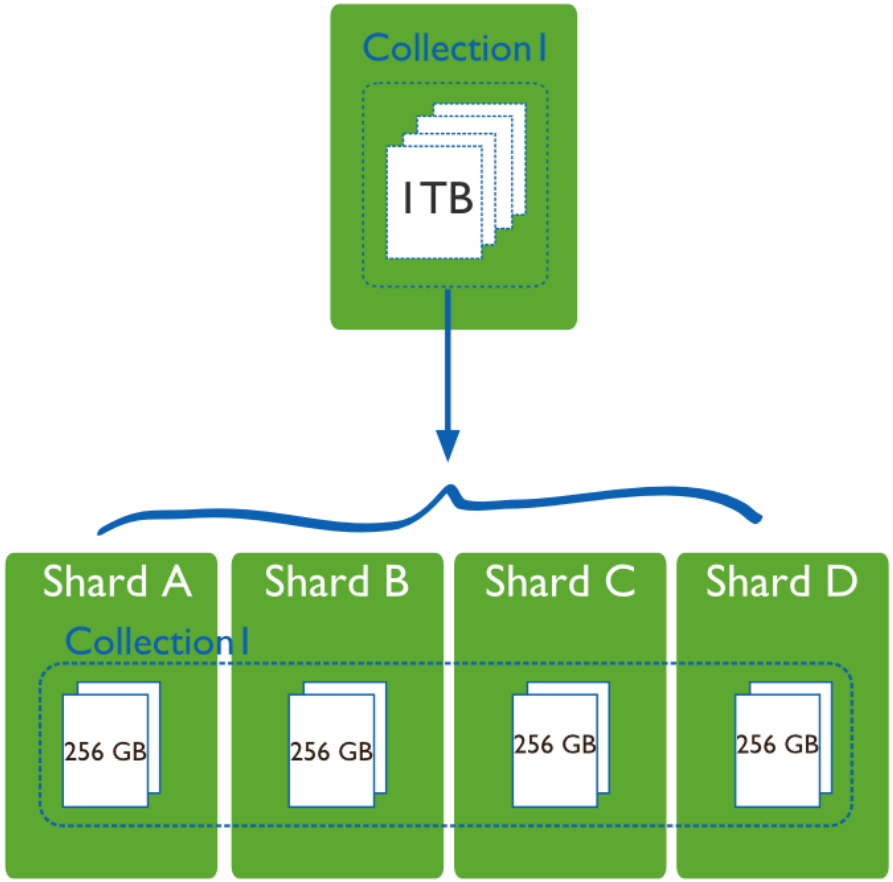
- Some data not available while partitioning
- But data is always **consistent**

AP – Available with Partition Tolerance

- Always on, but during partitioning, some data may be inconsistent
- After sharding, eventual consistency

What is Sharding?

Spreading data, and thereby workload, across nodes in a cluster



Stream Computing

Stream computing involves ingesting data (structured or unstructured) from arbitrary sources and analyzing it

- Without necessarily persisting it.

Applications:

- Real-time sensor output, stock ticker, medical devices etc.

Example

- IBM InfoSphere Streams

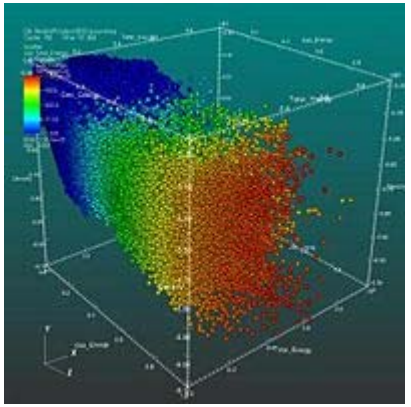
More Details

- "Analyzing Any Data, Anywhere, All the Time"
  - http://www.dbta.com/Columns/DBA-Corner/Analyzing-Any-Data-Anywhere-All-the-Time-67049.aspx

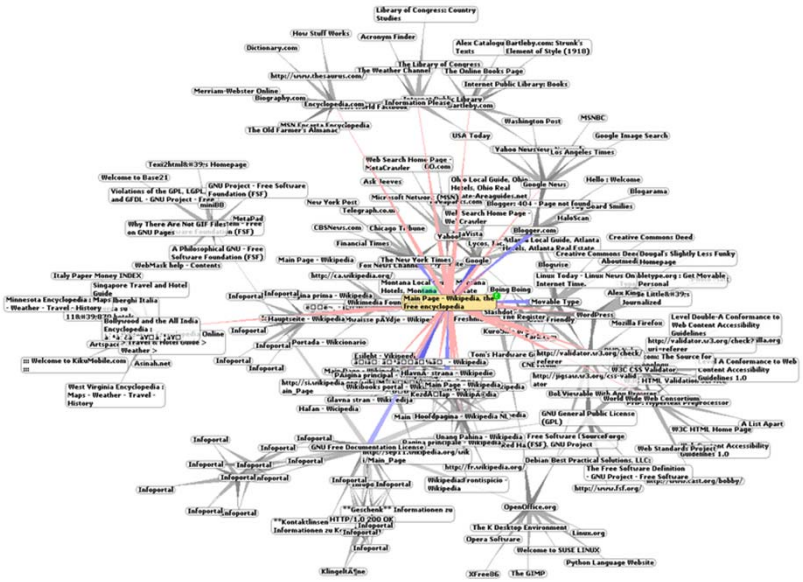
Data Visualization

Uncovering patterns and trends hidden in data using visual representations of the data

- Information that has been abstracted in some schematic form



Scatter Plot



A data visualization of Wikipedia as part of the World Wide Web,

Word cloud of my blog
http://db2portal.blogspot.com

Performance Benefits of Visualization

Performance Metrics (YoY Change)	Use Visualization Tools	Don't Use Visualization Tools	Performance Difference
Time-to-information	21% improvement	11% improvement	**1.9-times greater increase**
Accuracy of business decisions	22% improvement	12% improvement	**1.8-times greater increase**
Time-to-decision	20% improvement	7% improvement	**2.9-times greater increase**
Visibility / searchability of business data	27% improvement	6% improvement	**4.5-times greater increase**
Quality of analysis	22% improvement	2% improvement	**11-times greater increase**

Source: ***Analyst Insight:*** Seeing the Big Picture: Visualization for Big Data Aberdeen Group, April 2013

What is JSON?

JSON, or JavaScript Object Notation…

- An open standard format for data interchange that uses human-readable text to transmit data objects consisting of name–value pairs.
- Language-independent but uses conventions that are familiar to programmers of the C-family of languages
- JSON is built on two structures:
  - A collection of name/value pairs
  - An ordered list of values
  - JSON has no tags – not self-descriptive
- It is used primarily to transmit data between a server and web application, as an alternative to XML.

```
{
    "firstName": "John",
    "lastName": "Smith",
    "age": 25,
    "address": {
        "streetAddress": "21 2nd Street",
        "city": "New York",
        "state": "NY",
        "postalCode": 10021
    },
    "phoneNumbers": [
        {
            "type": "home",
            "number": "212 555-1234"
        },
        {
            "type": "fax",
            "number": "646 555-4567"
        }
    ]
}
```

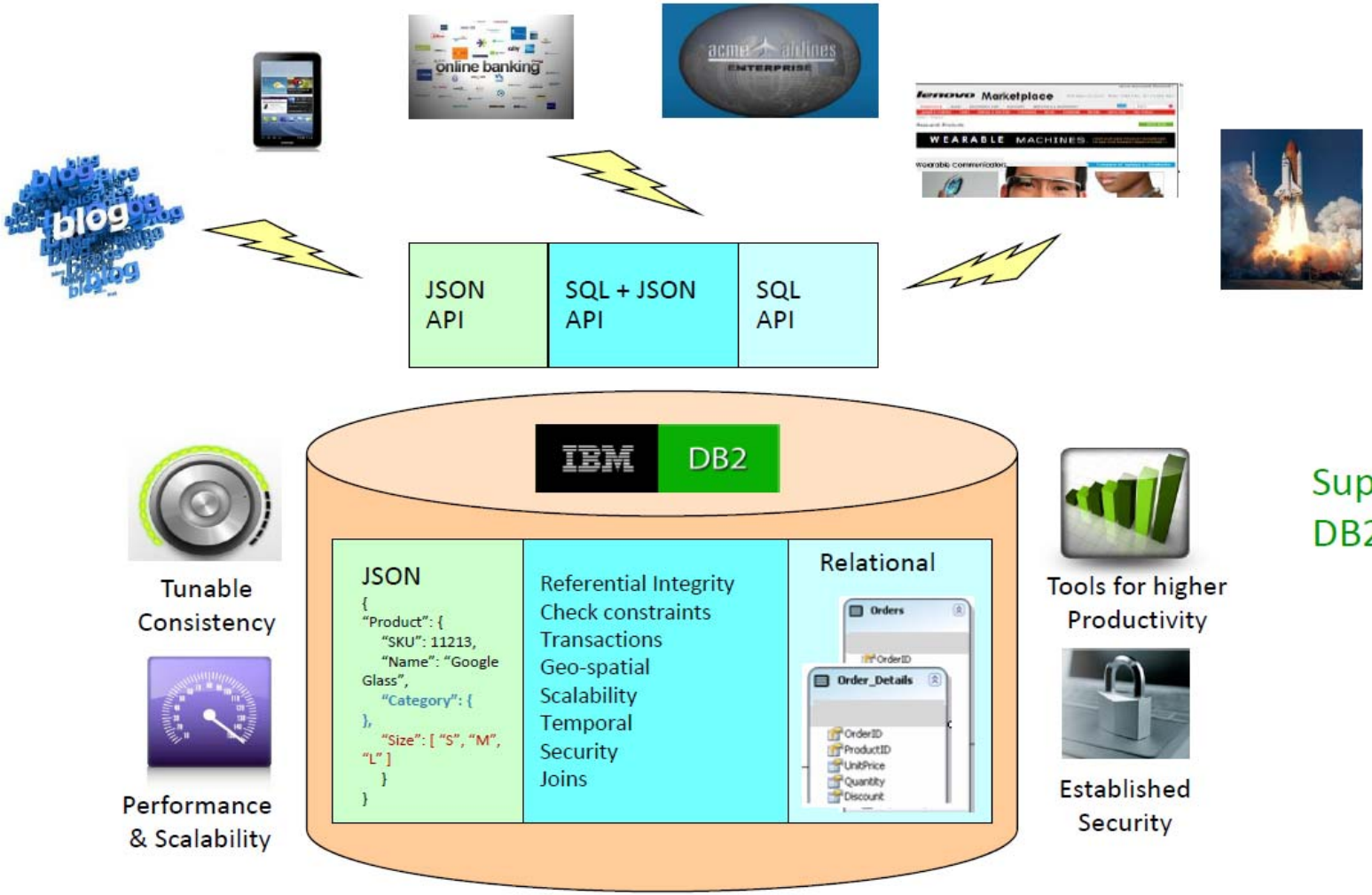
JSON representation of an object that describes a person

http://www.json.org/

JSON and DB2

http://www.ibm.com/developerworks/data/library/techarticle/dm-1306nosqlforjson1/

http://www.ibm.com/developerworks/data/library/techarticle/dm-1306nosqlforjson2/



Source: DB2 11 for z/OS: Technical overview - Part 2
IBM presentation by John Campbell, Feb 19, 2014

What is R?

R is a software programming language and software environment for statistical computing and graphics.

- an interactive environment for doing statistics
- widely used among statisticians, data scientists, and data miners for developing statistical software and data analysis.

Open Source

- R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems.

R uses a command line interface; however, several graphical user interfaces are available for use with R.

http://www.r-project.org/

What is Python?

Python is a general-purpose, high-level programming language.

- Its design philosophy emphasizes code readability
- Its syntax generally enables programmers to express concepts in fewer lines of code

Python has a large standard library

- One of Python's greatest strengths

Currently, Python does not match R's data analysis, data modeling and machine learning capabilities

- But it is used by data scientists, sometimes in conjunction with R

https://www.python.org/

What is Hive?

Apache Hive is an open source data warehouse system for Hadoop

Using Hive you can create a structure to the Hadoop data and then query it using an SQL-like language

- HiveQL
  - Converts into Java MapReduce programs
- Traditional MapReduce programs can be used in conjunction with HiveQL when it makes sense to do so

More details at:

- http://hive.apache.org/
- http://www.aptibook.com/Articles/Pig-and-hive-advantages-disadvantages-features

What is Pig?

Apache Pig is an open source platform for analyzing large data sets

- High level language for expressing data analysis programs
  - Pig Latin
- Infrastructure for evaluating programs
  - Compiler that produces MapReduce programs

Pig programs lend themselves to being run in parallel

More details at:

- http://pig.apache.org/
- http://www.aptibook.com/Articles/Pig-and-hive-advantages-disadvantages-features

What Does Big Data Mean for DB2?

- **IBM DB2 Analytics Accelerator for z/OS (IDAA)**
  - Powered by Netezza technology
- **BLU Acceleration**
  - Available in DB2 10.5 for LUW; not yet in DB2 for z/OS
  - Column Store
  - Three capabilities
    - Actionable compression
    - SIMD (single instruction multiple data)
    - Data skipping

Some Thoughts on What is Big…

- **Do we count number of rows, number of pages, or disk space consumed?**
- **Do we count just the base data or add up the space used by indexes on that data as well?**
  - What about compressed data?
- **Does type of data matter?**
  - Traditional vs. multimedia
- **All we really should care about is how does the large amount of data impact our job.**
  - Think in terms of how it complicates database administration and data availability
  - Compare and contrast using the number of pages, not the number of rows. (Easier to compare the size of one table space to another)

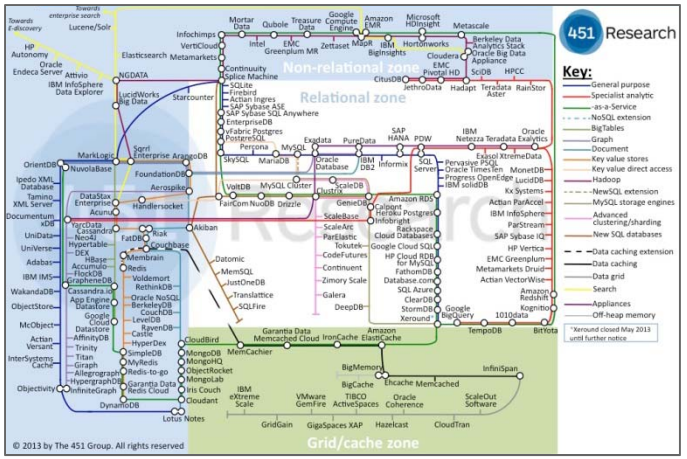
Key Takeaways

Big Data and its related technologies are not here to replace your entire existing data infrastructure

But embracing Big Data will cause you to change your data infrastructure to add new capabilities

- **Different data persistence technologies**
  - NoSQL DBMS
  - Hadoop
- **Analytic capabilities**
  - Traditional: SAS, SPSS, R
  - New: visualization, etc.
- **New technologies**
  - Streaming data

Market consolidation *will* happen



See Slide 8

Prepare to Avoid Failure

Know why you are pursuing Big Data / Analytics

- Make sure you have a **business case** for your Big Data projects

Do not ignore data quality

- Poor quality Big Data can produce Big Errors

Metadata is still important (as is DBA)

- If you don't know what the data is, you cannot properly analyze and interpret results

Technology is NOT a silver bullet

- That is, you cannot just implement Hadoop (for example) and expect to achieve results

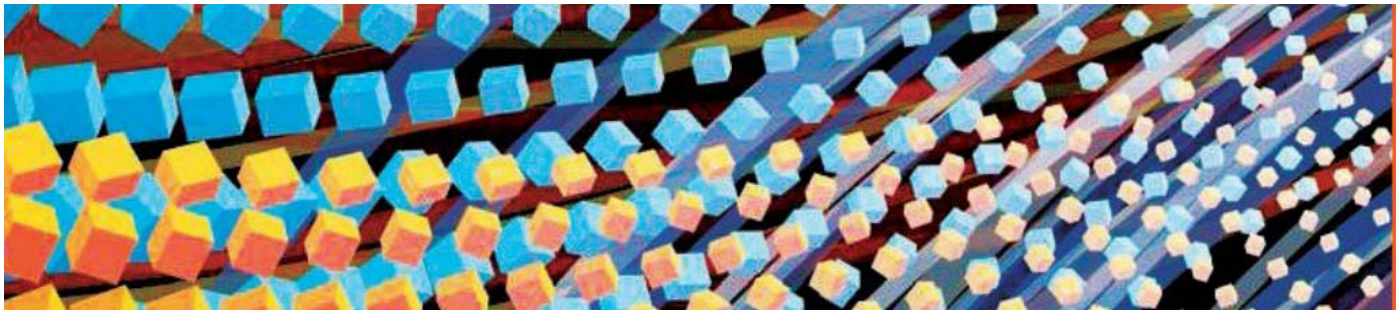
Focus on BOTH data AND analytics

- Accumulating a LOT of data provides no real benefits without the ability to process and analyze it

Embrace the "Fuzzy"-ness

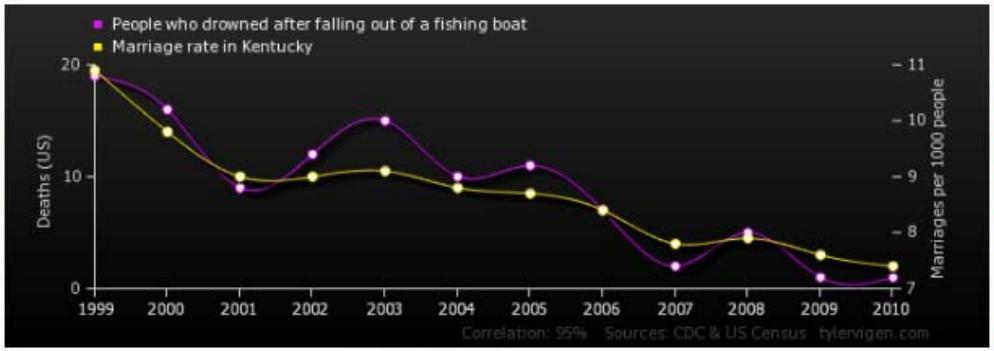
Big Data can be difficult to comprehend because it differs so substantially from what we are accustomed to doing:

- **Data not always clearly defined**
- **Data not always accurate**
- **Looking for insight in patterns and correlation**
  - As opposed to causality
- **Usefulness often requires combination of data sources**



Correlation ≠ Causality

People who drowned after falling out of a fishing boat
correlates with
Marriage rate in Kentucky



Upload this chart to imgur

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
People who drowned after falling out of a fishing boat Deaths (US) (CDC)	19	16	9	12	15	10	11	7	2	5	1	1
Marriage rate in Kentucky Marriages per 1000 people (US Census)	10.9	9.8	9	9	9.1	8.8	8.7	8.4	7.8	7.9	7.6	7.4

Correlation: 0.952407

Source: Spurious Correlations web site – www.tylervigen.com

So What Should a DB2 Professional Take Away from All of This?

What is big for you may differ from what is big for another shop

- Remember it is not just about big, but also about different, rapidly changing data

DB2 is gaining capabilities for ingesting, storing, and processing Big Data

Don't jump into a Big Data initiative without first identifying a specific business problem or need that the effort could help solve

- Start small and simple if possible

Centralized data management is a benefit when implementing Big Data Analytics

- More difficult to pull data together with siloed, disparate data management efforts

Some Additional Advice

You will likely need to augment your skillset to add Big Data and analytics skills

Beware of immature technology… much of the software is open source and, in some cases, not even Version 1!

- e.g.) Apache HBase v0.94.12; Pig and Hive are both at v0.12.0

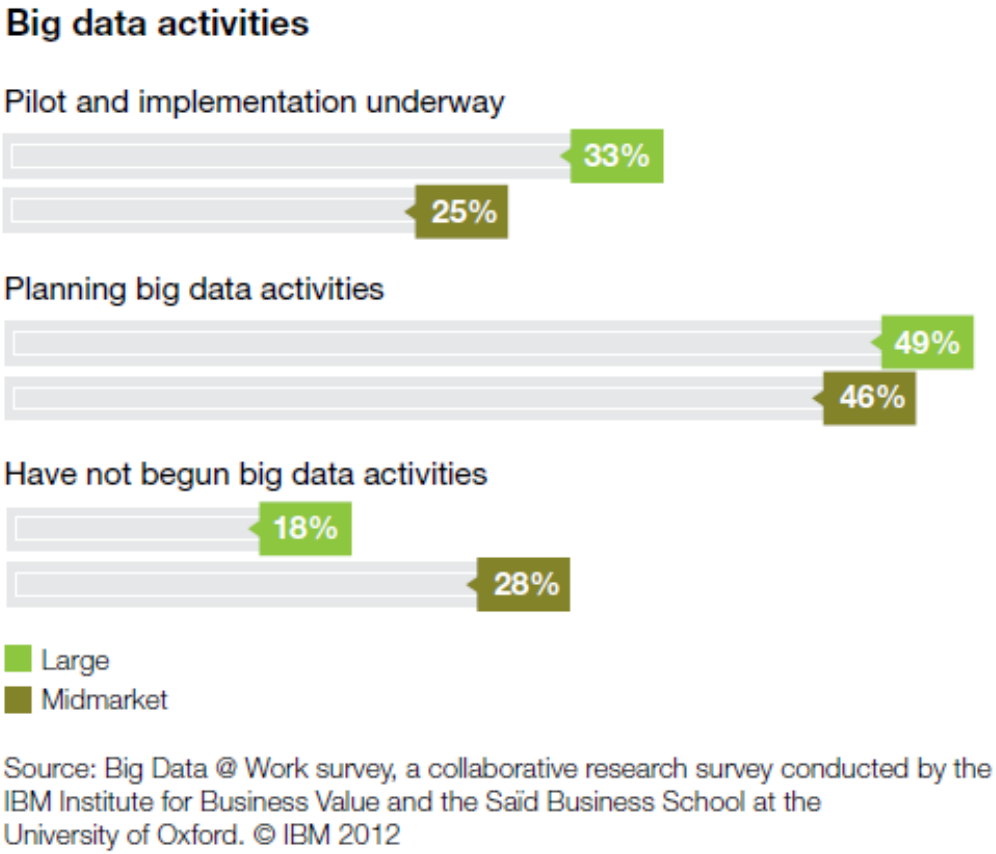
Analytics favors agility over stability

- Prepare models, test, refine…

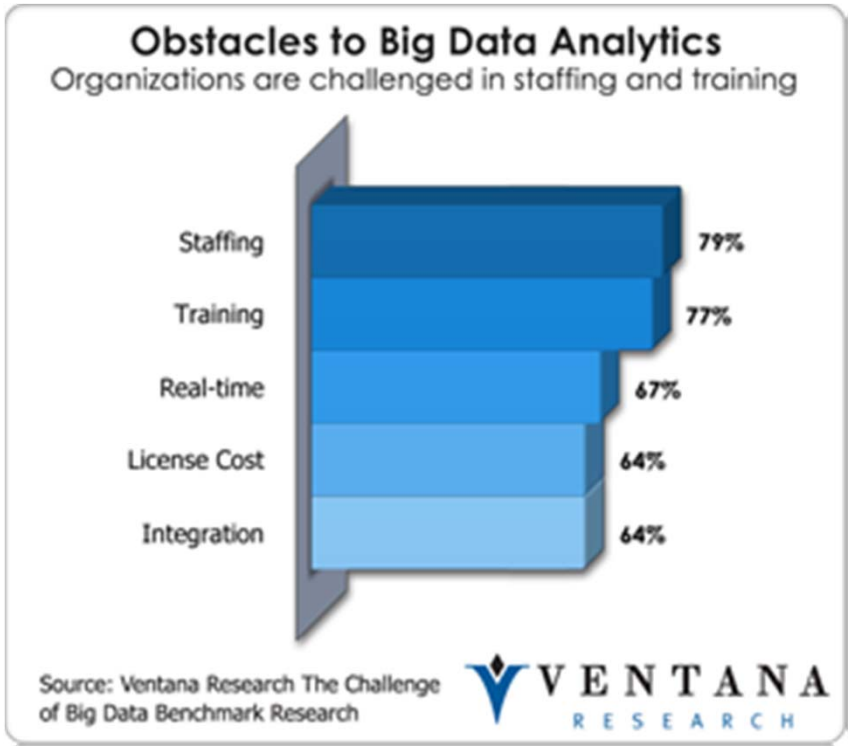
Ultimate Goal?

- To move from gut-based executive decision-making to **data-based decision-making**
- That is, base important business decisions on actual data instead of on the **HiPPO**

Are You Behind the Market?



Source: Big Data @ Work survey, a collaborative research survey conducted by the IBM Institute for Business Value and the Saïd Business School at the University of Oxford. © IBM 2012



Biggest obstacle orgs have with big data activities is finding the skills.

Three out of four organizations have big data activities underway; and one in four are either in pilot or production.

Some Books for Additional Research

Davenport, Thomas H., *Big Data @ Work: Dispelling the Myths, Uncovering the Opportunities*, Harvard Business Review Books (2014), ISBN 978-1-4221-6816-5

IBM RedBook, *Performance and Capacity Implications for Big Data*, January 2014: REDP-5070-00

Redmond, Eric and Jim Wilson. *Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement*, Pragmatic Bookshelf (2012), ISBN 978-1-93435-692-0

Sadalage, Pramod J. and Martin Fowler. *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*, Addison-Wesley (2013), ISBN 978-0-321-82662-6

Santhi, Dr. Arvind, *Big Data Analytics*, MC Press (2012), ISBN 978-1-58347-380-1

Silver, Nate. *The Signal and the Noise: Why So Many Predictions Fail-but Some Don't*, Penguin Group (2012), ISBN 978-1-59-420411-1

Smolan, Rick and Jennifer Erwitt. *The Human Face of Big Data*, AAOP (2013), ISBN 978-1-4549-0827-2

White, Tom. *Hadoop: The Definitive Guide, 3rd edition*, O'Reilly (2012) ISBN 978-1-449-31152-0

Zikopoulos, Paul, et al. *Harness The Power of Big Data: The IBM Big Data Platform*, McGraw-Hill (2013), ISBN 978-0-07-180817-0

Web Sites for Additional Research

Big Data University

- http://bigdatauniversity.com

Analytics Week

- http://analyticsweek.com

List of NoSQL Databases

- http://nosql-database.org/

IBM Analytics and Big Data

- http://www.ibm.com/smarterplanet/us/en/smarter-enterprise/solutions/big-data-and-analytics

Apache Hadoop and Related Projects

- http://hadoop.apache.org/
- http://hive.apache.org/
- http://pig.apache.org/
- http://cassandra.apache.org/

Contact Information

CDB Software, Inc.

Phone: 281-920-3305
cdbsales@cdbsoftware.com

http://www.cdbsoftware.com

Questions?
info@cdbsoftware.com

Craig S. Mullins
Mullins Consulting, Inc.
15 Coventry Court
Sugar Land, TX 77479
Phone: (281) 494-6153

craig@craigsmullins.com

http://www.mullinsconsulting.com