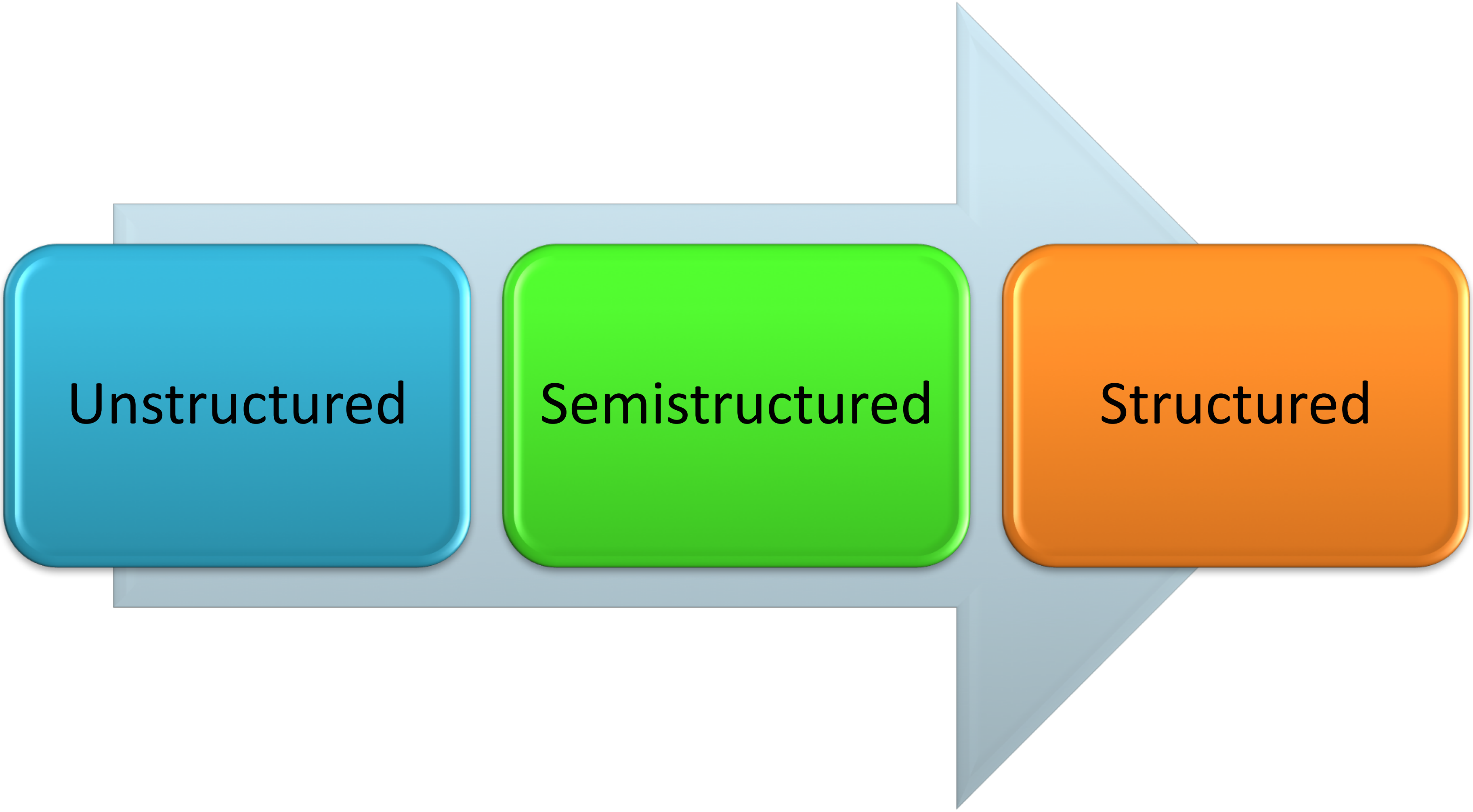
Data, Data, and More Data…

(Rao)

Types of Data



Unstructured

Semistructured

Structured

Unstructured Data

- Textual content – mainly for human understanding/cognition
  - HTML web pages
  - PDF files, MS Word files
  - Emails
  - Posts on social media sites such as Facebook and Twitter
  - Blogs
  - Text messages

Structured Data

- There is a defined structure on how the data should be stored and represented
- Relational databases store data in tables

SSN	Name	Age	Salary	Phone
1234	John Doe	25	100000	123-45678
2345	Jim Doe	35	…	…
…	…	…	…	…

Employee

Tables

Course	Department	SSN
CS490JU	CSEE	5678
…	…	…
…	…	…
…	…	…

Courses

Example

- List the SSN and salary of employees who teach a course along with the department offering the course, and sort the results by salary -- low to high

Employee(SSN, Name, Age, Salary, Phone)

Courses(Course, Department, SSN)

Example

Employee(SSN, Name, Age, Salary, Phone)
Courses(Course, Department, SSN)

```
SELECT E.SSN, E.Salary, C.Course
FROM Employee as E, Courses as C
WHERE E.SSN = C.SSN
ORDER BY E.Salary
```

ACID Transactions

- A transaction is a logical unit of work that contains a set of SQL statements

- A – atomicity
  - All or nothing (indivisible)
- C – consistency
  - Preserves database integrity (one valid state to another)
- I – isolation
  - Execute as if they were run alone/sequentially
- D – durability
  - Changes made by a committed transaction are not lost due to failures

Simple Example

Account number	Amount
101	1000
102	500

$\xrightarrow{T_1}$

Account number	Amount
101	500
102	1000

Transaction T_1: Transfer $500 from **101** to **102**

Operations

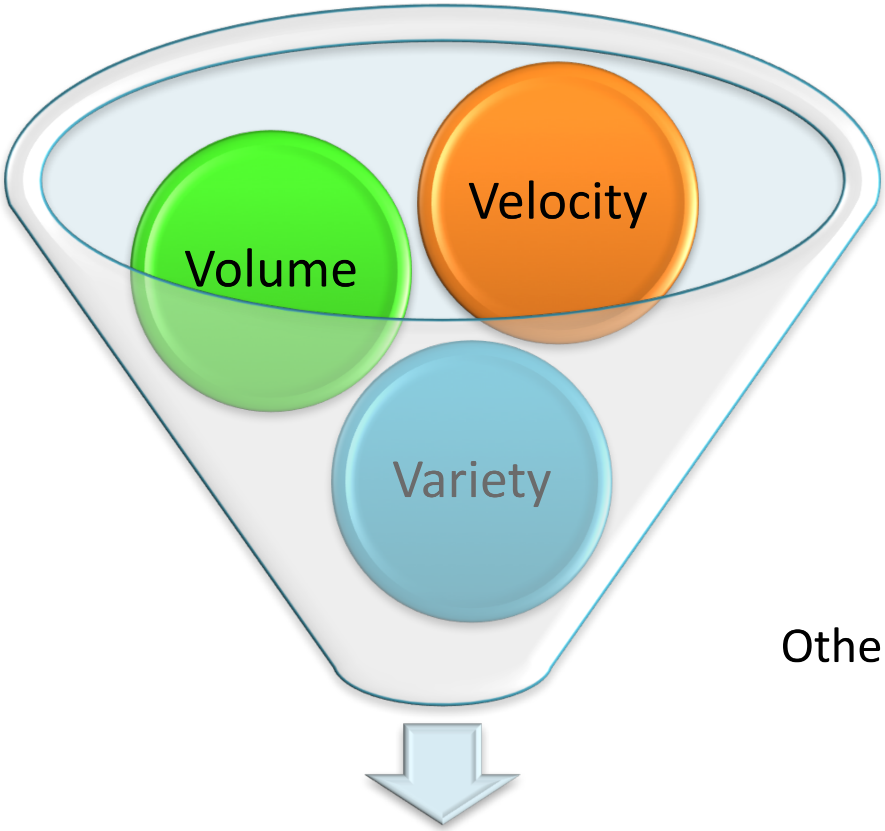
1. Subtract $500 from **101**
2. Add $500 to **102**

Transaction T_2: Transfer $300 from **102** to **101**

Semistructured Data

- No need of a fixed structure/schema
- Data can have partial/loose structure
- Example
  - XML data model
  - RDF data model; also referred to as “schema-free”
  - JSON

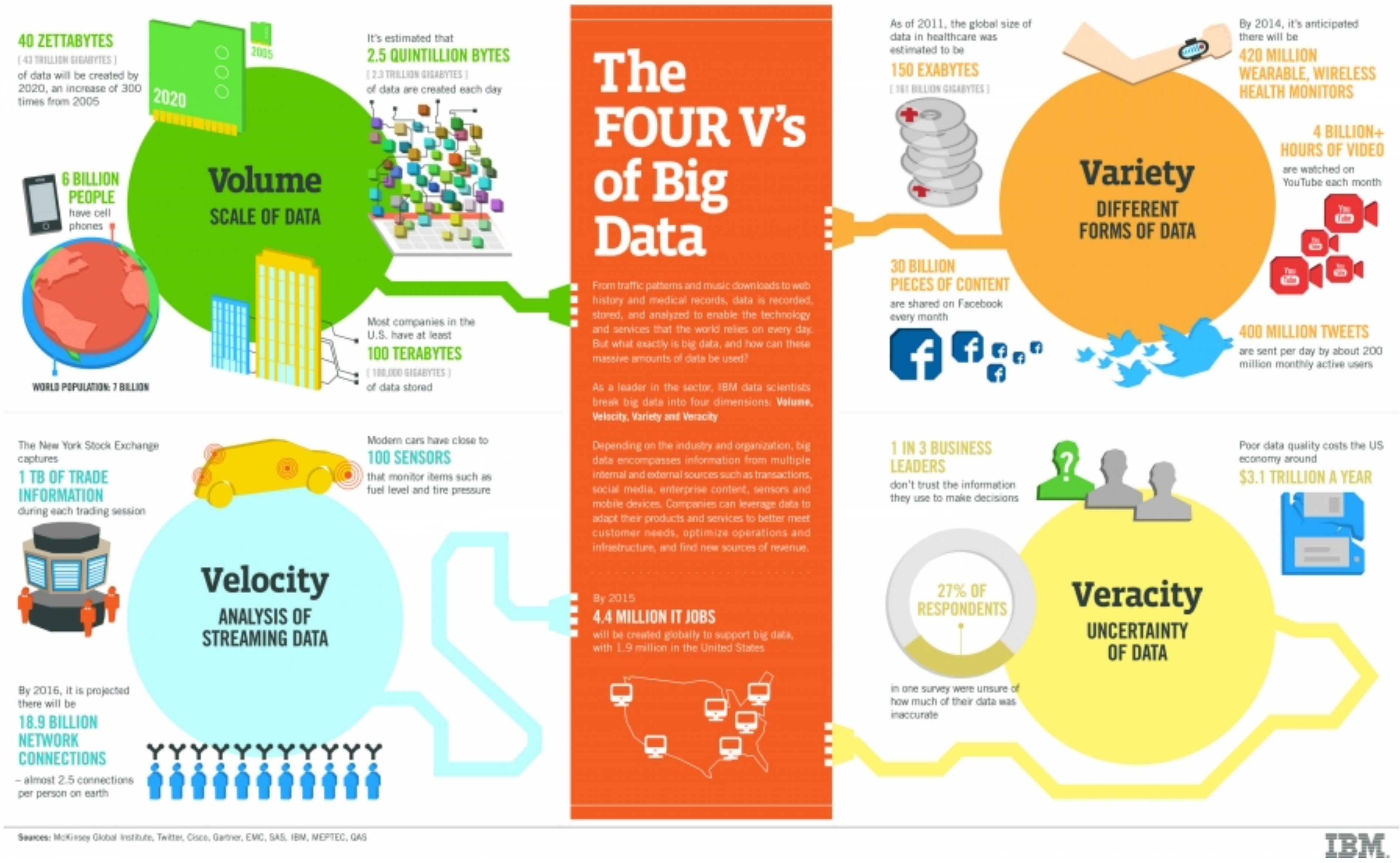
What is Big Data?



Other V’s: Veracity, Value, …

Big Data!

Nice Illustration



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005

It's estimated that **2.5 QUINTILLION BYTES** [2.3 TRILLION GIGABYTES] of data are created each day

6 BILLION PEOPLE have cell phones

Most companies in the U.S. have at least **100 TERABYTES** [100,000 GIGABYTES] of data stored

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES** [161 BILLION GIGABYTES]

By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

30 BILLION PIECES OF CONTENT are shared on Facebook every month

400 MILLION TWEETS are sent per day by about 200 million monthly active users

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions

Poor data quality costs the US economy around **$3.1 TRILLION A YEAR**

27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

Key Points

- Volume
  - Very large amounts of data
    - Petabytes (10^{15} bytes) and more
- Variety
  - Structured + unstructured + semistructured data
- Velocity
  - High rate of arrival of data
    - Stock quotes, Twitter tweets, sensor readings, web clicks, and many more

Impact?

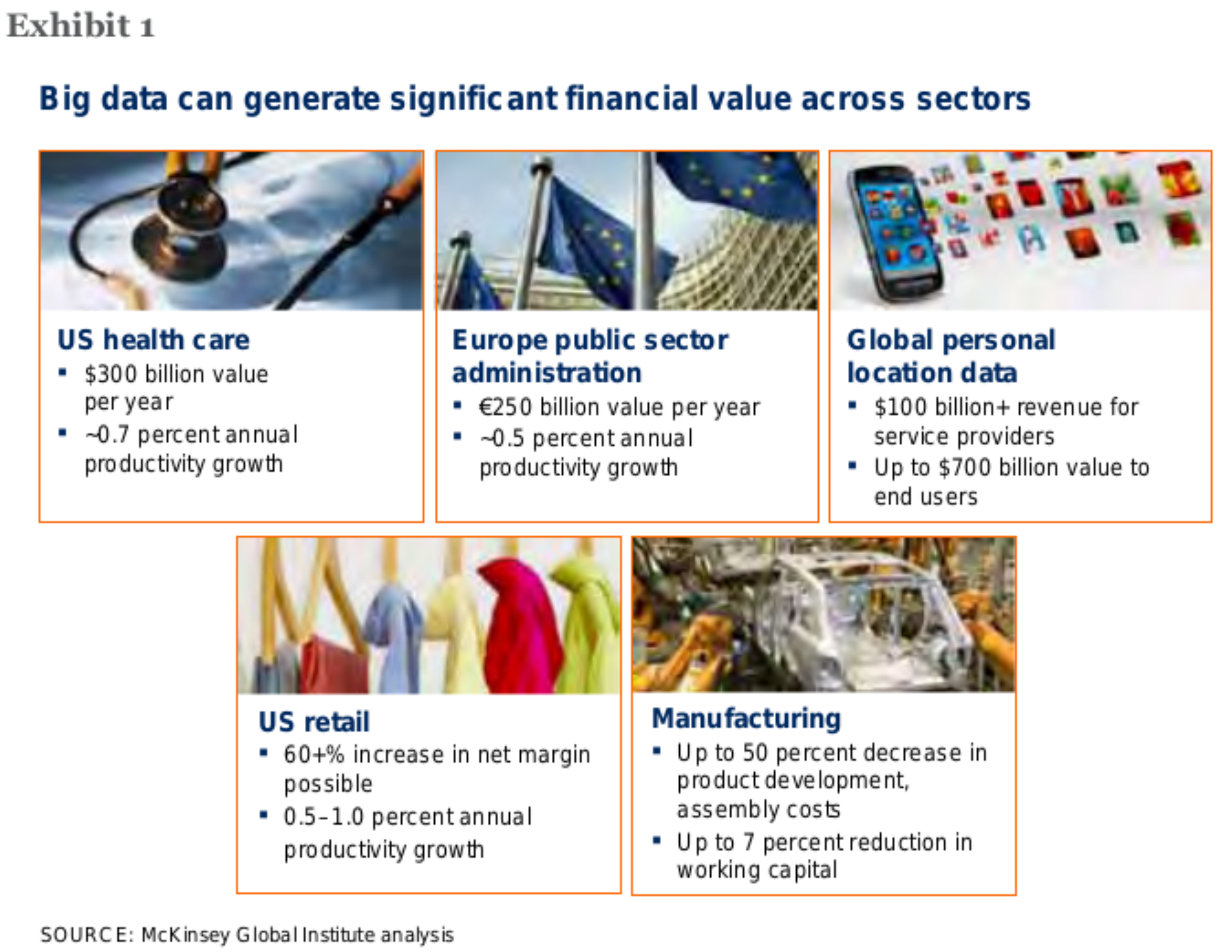


Exhibit 1

Big data can generate significant financial value across sectors

US health care
- $300 billion value per year
- ~0.7 percent annual productivity growth

Europe public sector administration
- €250 billion value per year
- ~0.5 percent annual productivity growth

Global personal location data
- $100 billion+ revenue for service providers
- Up to $700 billion value to end users

US retail
- 60+% increase in net margin possible
- 0.5–1.0 percent annual productivity growth

Manufacturing
- Up to 50 percent decrease in product development, assembly costs
- Up to 7 percent reduction in working capital

SOURCE: McKinsey Global Institute analysis

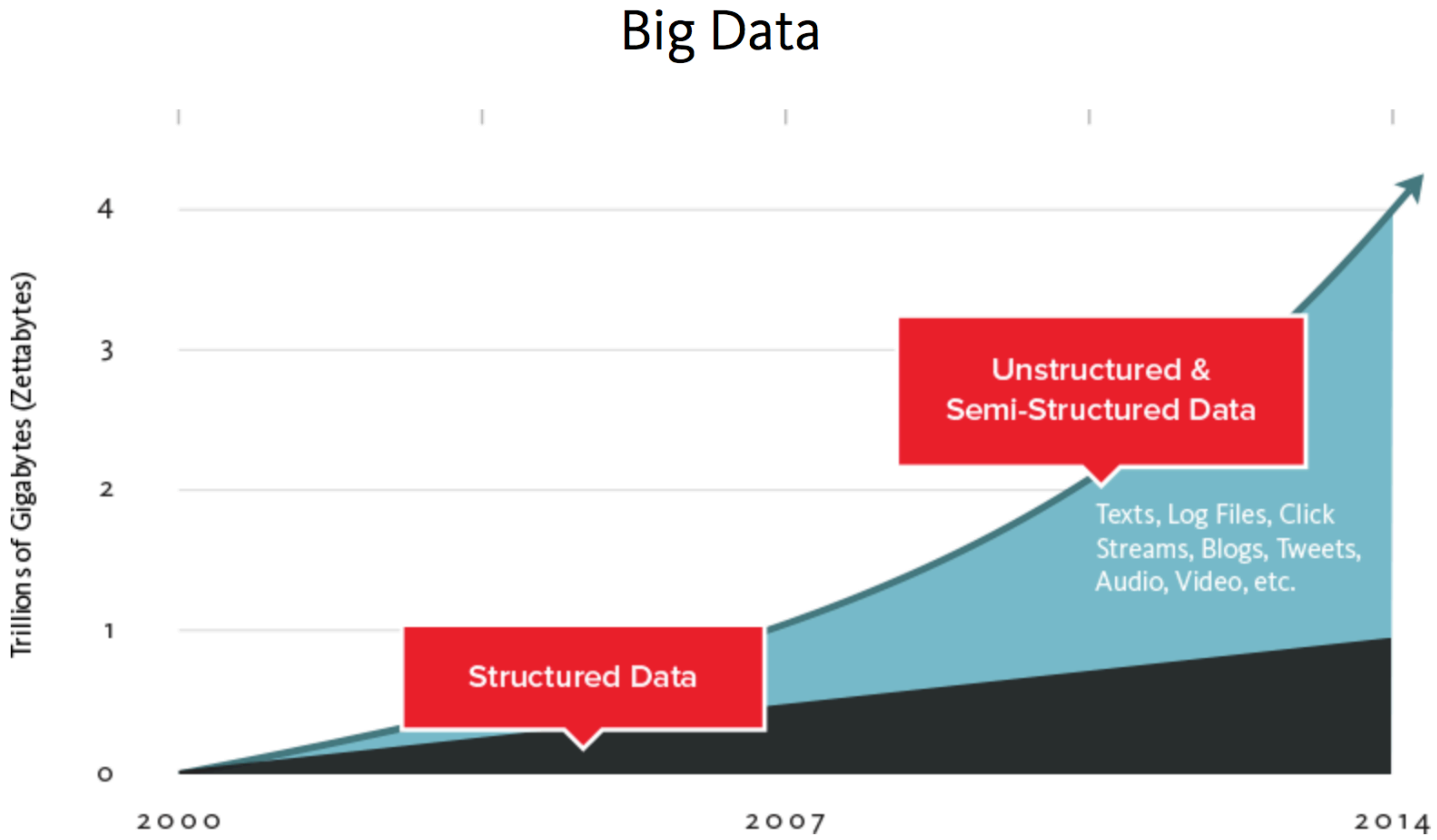
A Tweet

```
dbs@DBS:~$ cat ~/Tweets/json/test.json
{"contributors":null,
"text":"ポカーン。",
"geo":null,
"retweeted":false,
"in_reply_to_screen_name":null,
"truncated":false,
"lang":"ja",
"entities":{"urls":[],"hashtags":[],"user_mentions":[]},
"in_reply_to_status_id_str":null,
"id":289429398778687488,
"source":"<a href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twitter for iPhone<\/a>",
"in_reply_to_user_id_str":null,
"favorited":false,
"in_reply_to_status_id":null,
"retweet_count":0,
"created_at":"Thu Jan 10 17:52:00 +0000 2013",
"in_reply_to_user_id":null,
"id_str":"289429398778687488",
"place":null,
"user":{
  "location":"",
  "default_profile":true,
  "statuses_count":1885,
  "profile_background_tile":false,
  "lang":"ja",
  "profile_link_color":"0084B4",
  "profile_banner_url":"https://si0.twimg.com/profile_banners/459434688/1357637886",
  "id":459434688,
```

JSON: name/value pairs, arrays

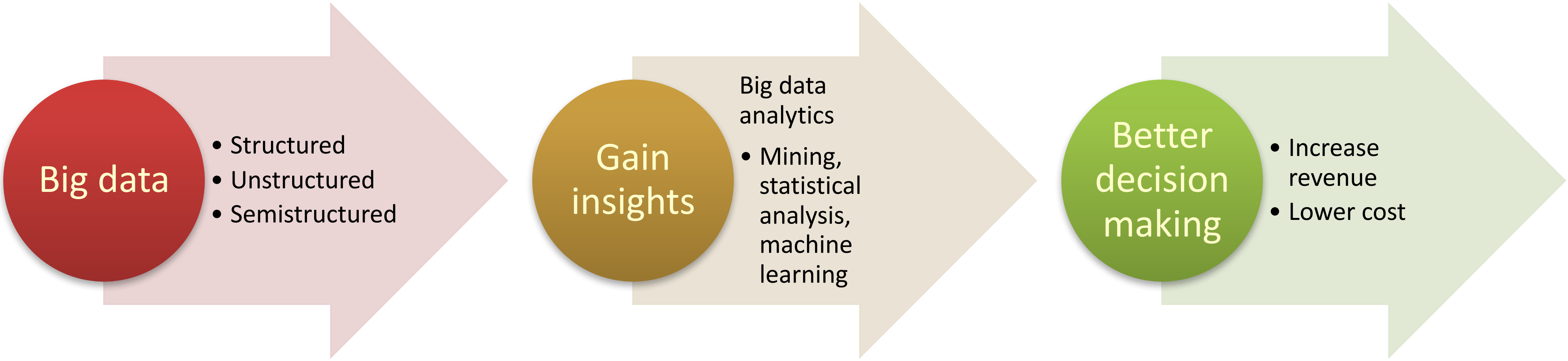
```
 "following":null,
  "favourites_count":0,
  "protected":false,
  "profile_text_color":"333333",
  "description":"そば屋にいるね",
  "verified":false,
  "contributors_enabled":false,
  "profile_sidebar_border_color":"C0DEED",
  "name":"まさひろ。",
  "profile_background_color":"C0DEED",
  "created_at":"Mon Jan 09 17:22:25 +0000 2012",
  "default_profile_image":false,
  "followers_count":113,
  "profile_image_url_https":"https://si0.twimg.com/profile_images/2093645100/image_normal.jpg",
  "geo_enabled":false,
  "profile_background_image_url":"http://a0.twimg.com/images/themes/theme1/bg.png",
  "profile_background_image_url_https":"https://si0.twimg.com/images/themes/theme1/bg.png",
  "follow_request_sent":null,
  "url":null,"utc_offset":null,z
  "time_zone":null,
  "notifications":null,
  "profile_use_background_image":true,
  "friends_count":104,
  "profile_sidebar_fill_color":"DDEEF6",
  "screen_name":"HiSoftbank",
  "id_str":"459434688",
  "profile_image_url":"http://a0.twimg.com/profile_images/2093645100/image_normal.jpg",
  "listed_count":0,
  "is_translator":false},
"coordinates":null
}
```

Data Explosion



http://www.couchbase.com/nosql-resources/what-is-no-sql

Benefit of Big Data



Big data

- Structured
- Unstructured
- Semistructured

Gain insights

Big data analytics

- Mining, statistical analysis, machine learning

Better decision making

- Increase revenue
- Lower cost

Impact of Big Data

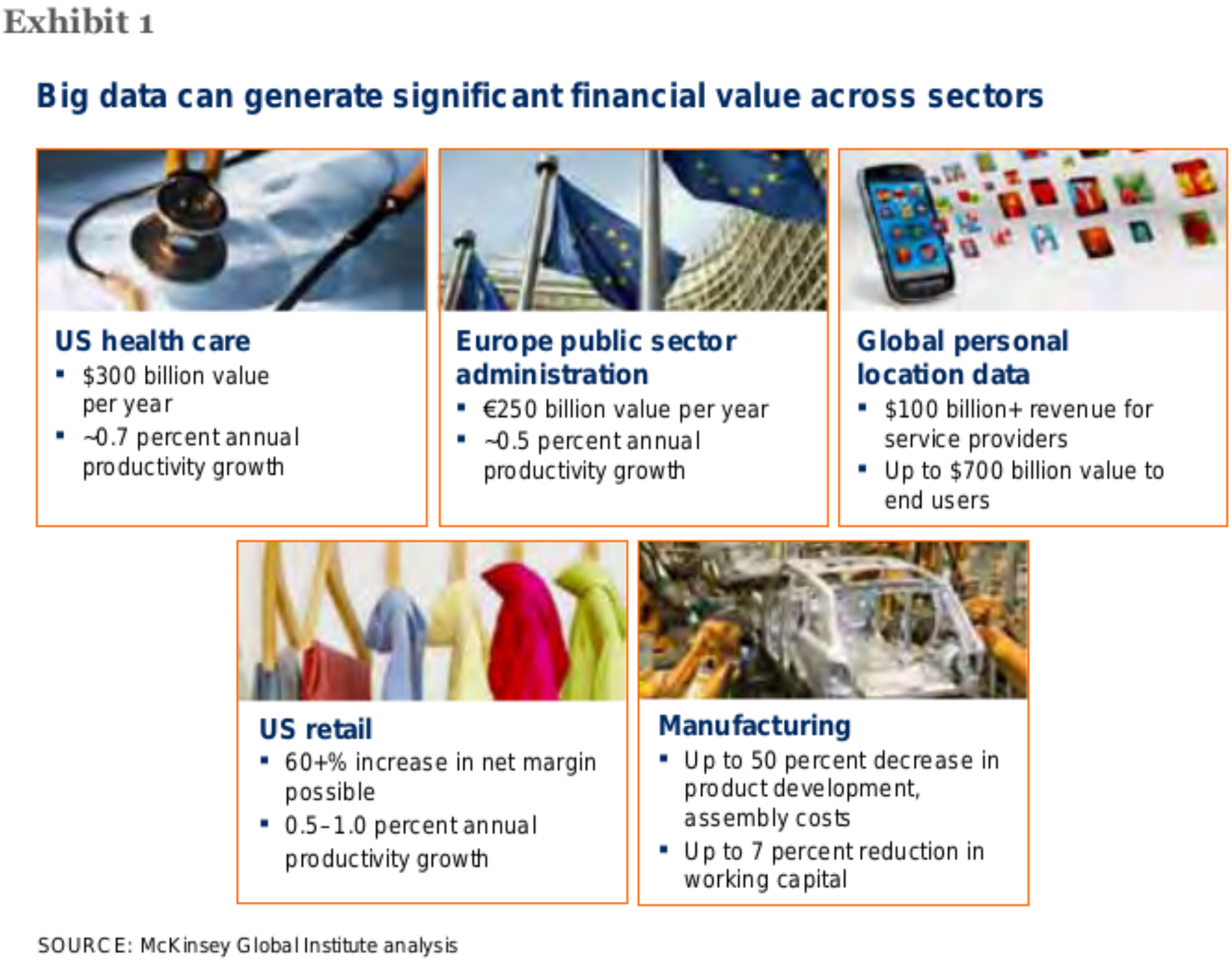


Exhibit 1

Big data can generate significant financial value across sectors

US health care

- $300 billion value per year
- ~0.7 percent annual productivity growth

Europe public sector administration

- €250 billion value per year
- ~0.5 percent annual productivity growth

Global personal location data

- $100 billion+ revenue for service providers
- Up to $700 billion value to end users

US retail

- 60+% increase in net margin possible
- 0.5–1.0 percent annual productivity growth

Manufacturing

- Up to 50 percent decrease in product development, assembly costs
- Up to 7 percent reduction in working capital

SOURCE: McKinsey Global Institute analysis

Impact of Big Data

- "Big data technologies will be transformative in every sphere of life."[1]
- According to IBM[2]

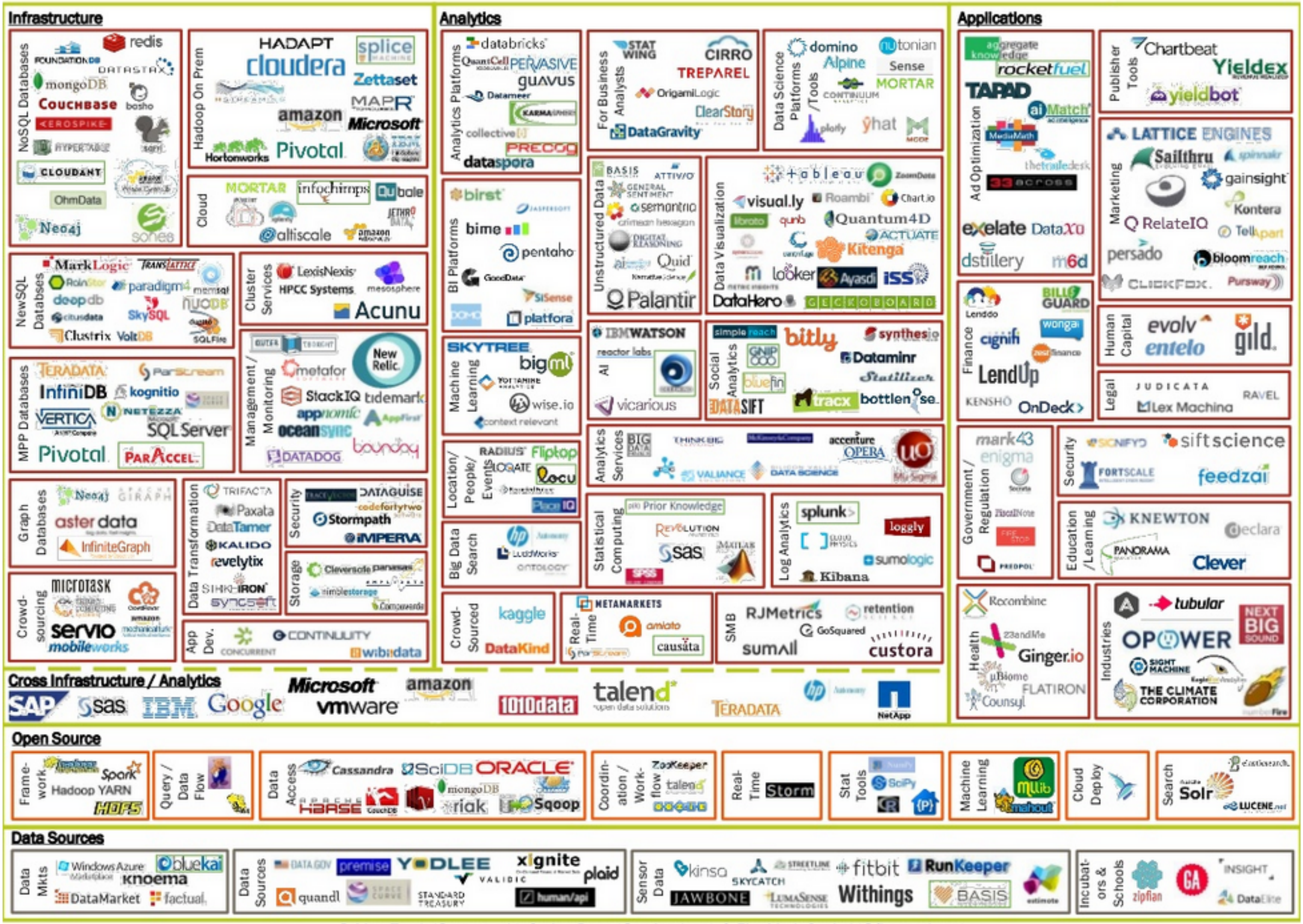
"Healthcare: 20% decrease in patient mortality by analyzing streaming patient data"
"Telco: 92% decrease in processing time by analyzing networking and call data"
"Utilities: 99% improved accuracy in placing power generation resources by analyzing 2.8 petabytes of untapped data"

[1] J. Podesta, P. Pritzker, E. Moniz, J. Holdren, and J. Zients. Big Data: Seizing Opportunities, Preserving Values. http://www.whitehouse.gov/sites/default/les/docs/big_data_privacy_report_5.1.14_final_print.pdf, 2014.

[2] http://www-01.ibm.com/software/data/bigdata/industry.html

BIG DATA LANDSCAPE, VERSION 3.0

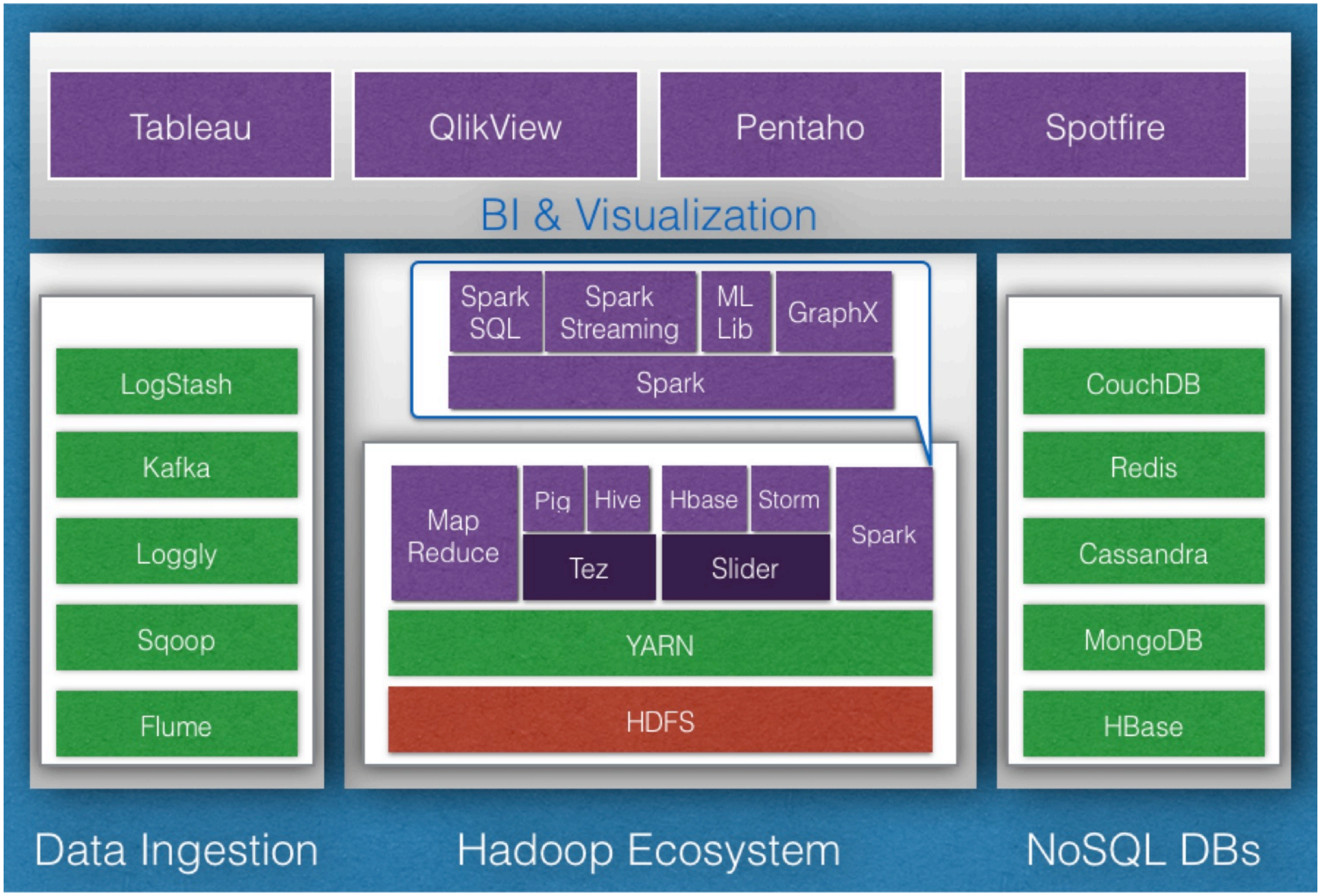
Exited: Acquisition or IPO



© Matt Turck (@mattturck), Sutian Dong (@sutiandong) & FirstMark Capital (@firstmarkcap)

http://www.slideshare.net/mjft01/big-data-landscape-matt-turck-may-2014

Bird’s Eye View of the Big Data Ecosystem



http://synerzip.com/wp-content/uploads/2013/08/Hadoop-Ecosystem1.jpg

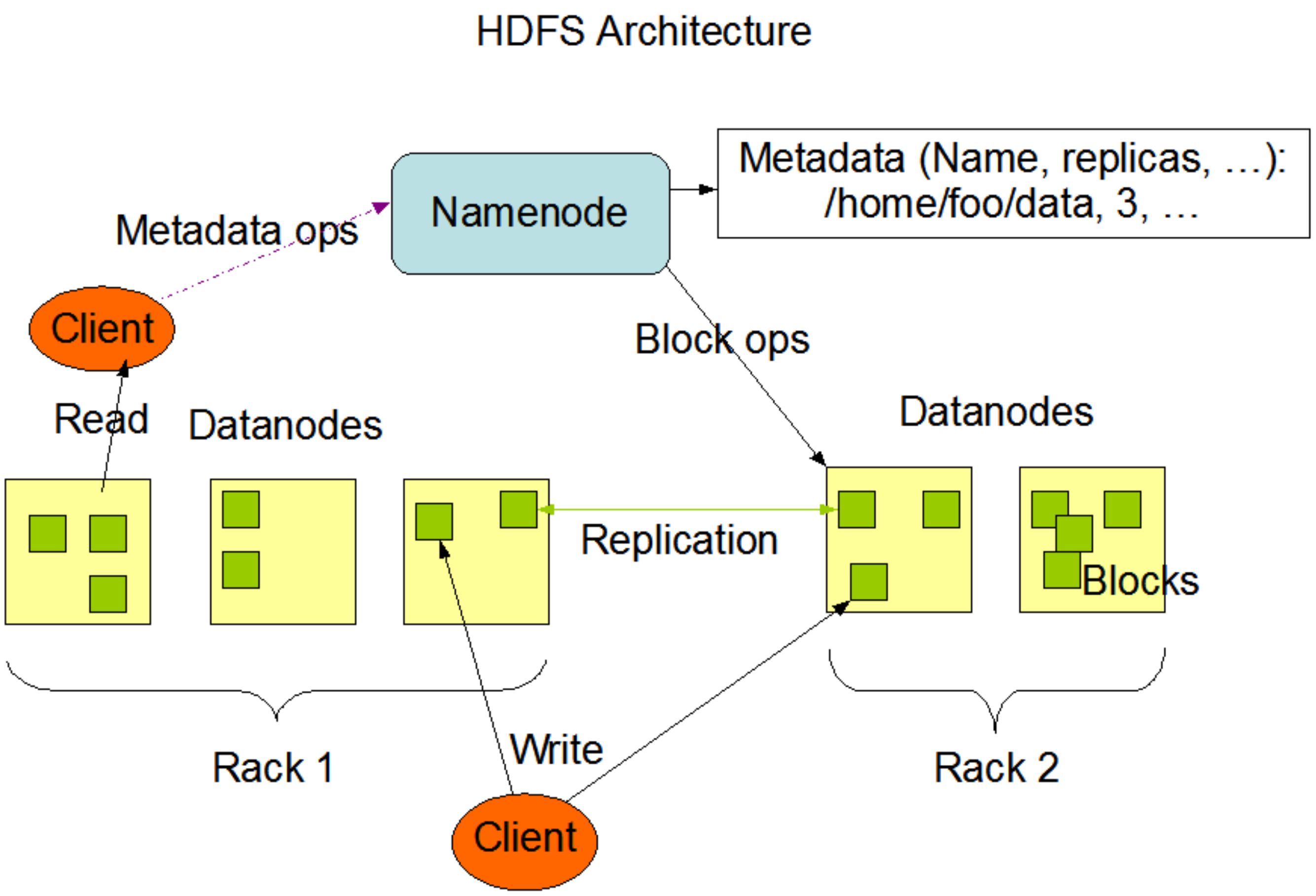
Apache Hadoop

- Open-source framework for storing and processing large amounts of data on a cluster of machines
  - http://hadoop.apache.org
- MapReduce Framework
  - Write parallel programs, execute the programs on a cluster of machines
- Hadoop Distributed File System (HDFS)
  - A distributed file system to store large number of large files using a cluster of machines (e.g., 2000 nodes)

File System Basics

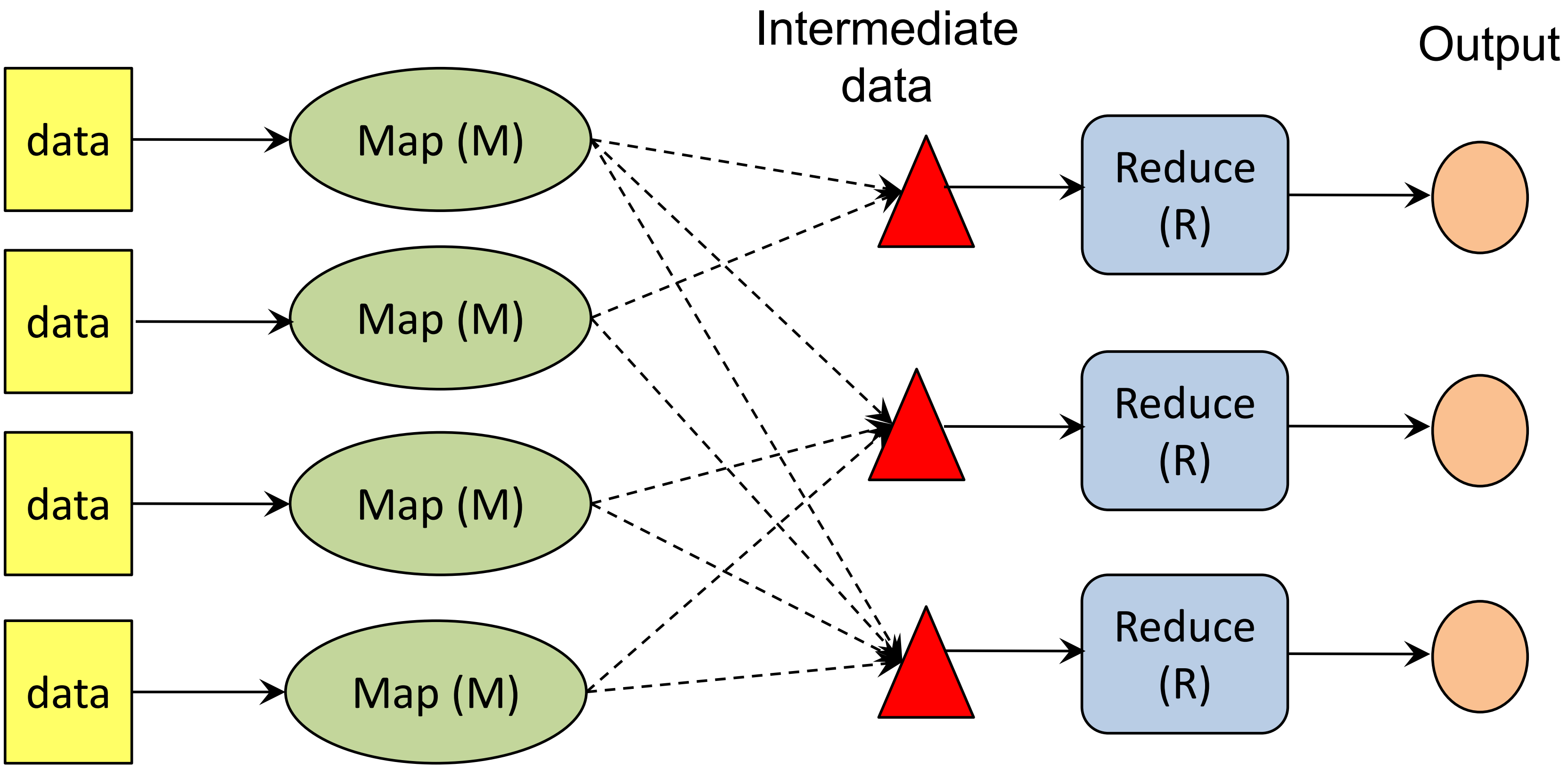
- A file has two main parts
  - Metadata
    - Name of the file, creation time, size, permissions, pointers to data blocks
  - Data blocks
    - Actual content of the file is broken down into equal-sized blocks

HDFS



Source: http://hadoop.apache.org/docs/stable/images/hdfsarchitecture.gif

MapReduce Model


Intermediate data
Output
data
data
data
data
Map (M)
Map (M)
Map (M)
Map (M)
Reduce (R)
Reduce (R)
Reduce (R)