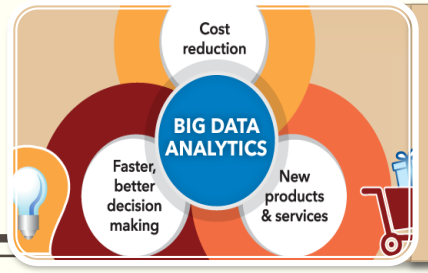
ภาษาการเขียนโปรแกรม 02-212-213

Week9

อ.ธิดาวรรร คล้ายศรี

บทที่3



Big Data Analytics

- Data Analytics Life Cycle
- Big data concepts & Hadoop ecosystem

3.1 Data Analytics Lifecycle (วงจรชีวิตการวิเคราะห์ข้อมูล)

3.1.1 Describing Problem (กำหนดปัญหา)

3.1.2 Identifying Data Sources (ระบุแหล่งข้อมูล)

3.1.3 Preparing Data (การเตรียมความพร้อมข้อมูล)

3.1.4 Planing Model (วางแผนโมเดล)

3.1.5 Building Model (สร้างโมเดล)

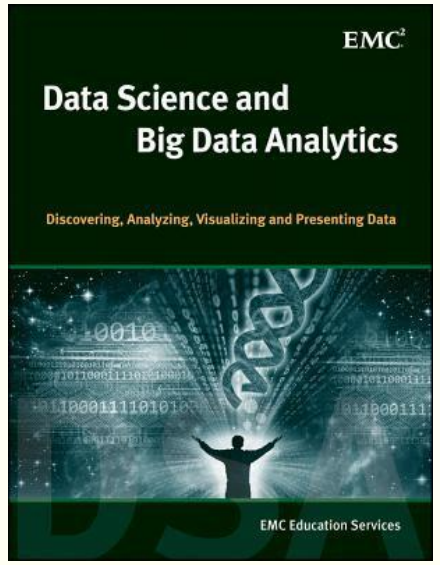
3.1.6 Visaulising Results (นำเสนอข้อมูลวิเคราะห์)

3.2 Big Data Concepts (แนวความคิดในเรื่อง big data)

3.3 Introductory Hadoop and Its Ecosystem (บทนำและวงชีวิตของฮาดูพ)

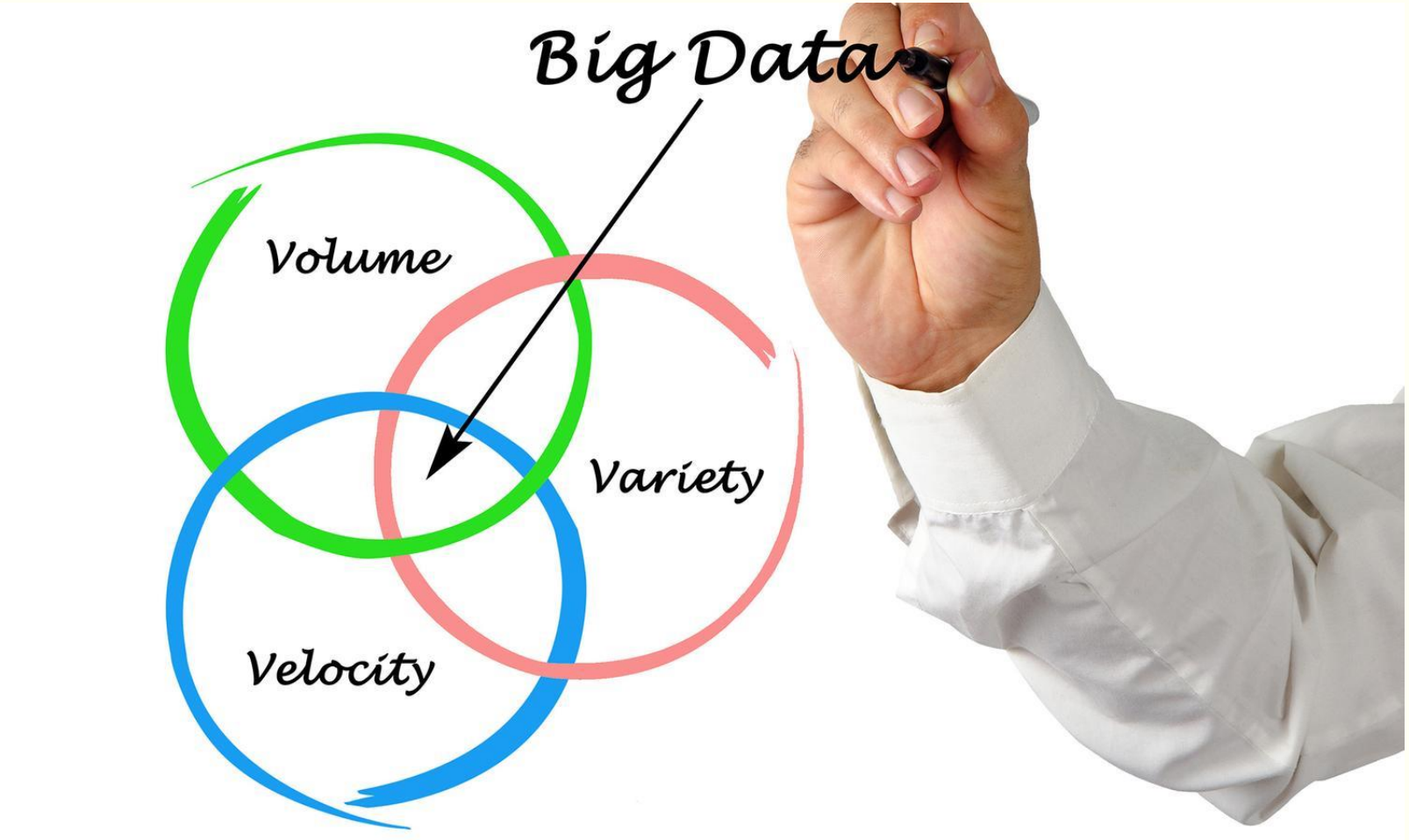
Source of the content:

Majority of the slides: "**Data Science and Big Data Analytic**" text book
by EMC Education Services
A few slides: DEPA Big data analytic

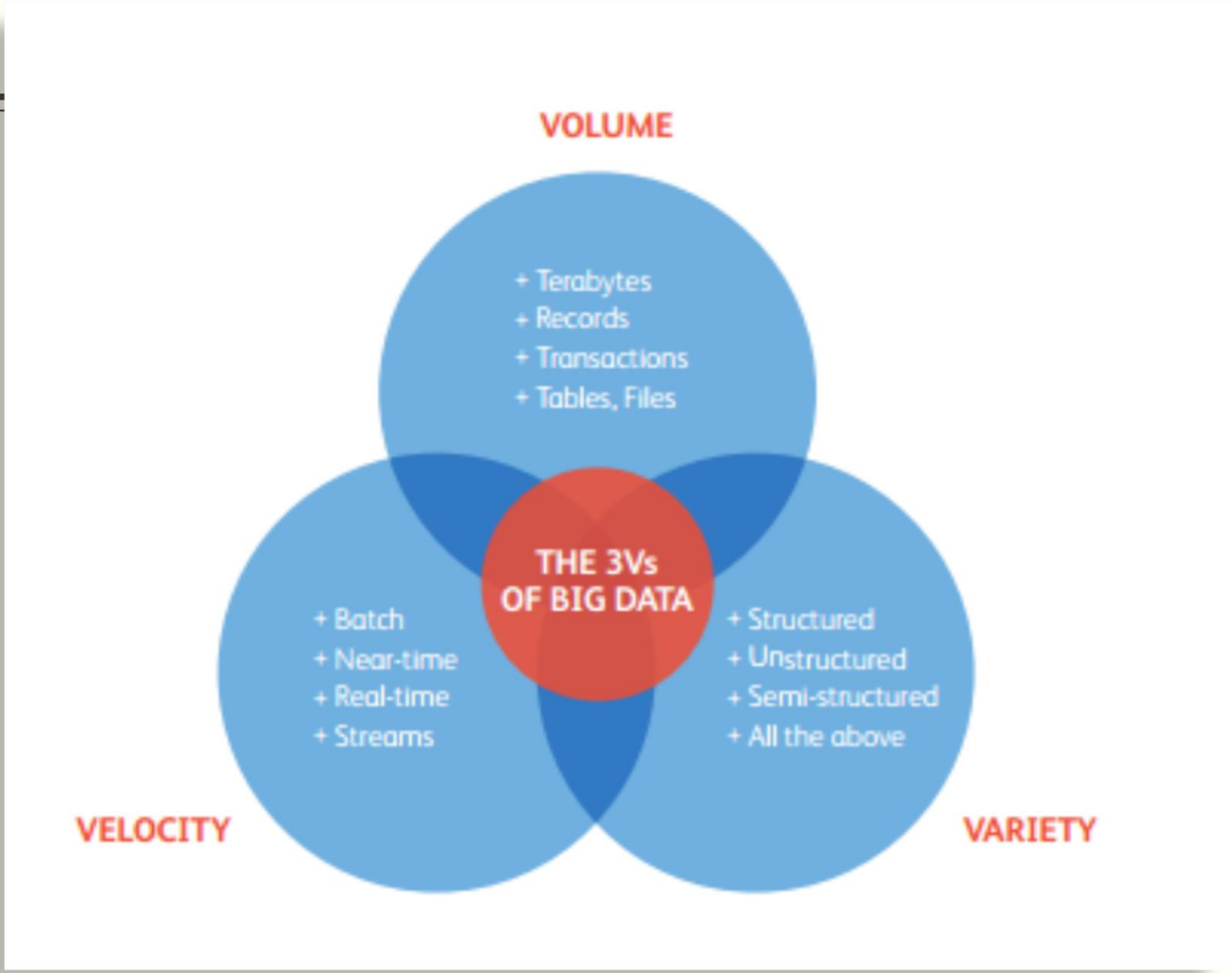


Part 2: 3.2 Big Data Analytic Concepts

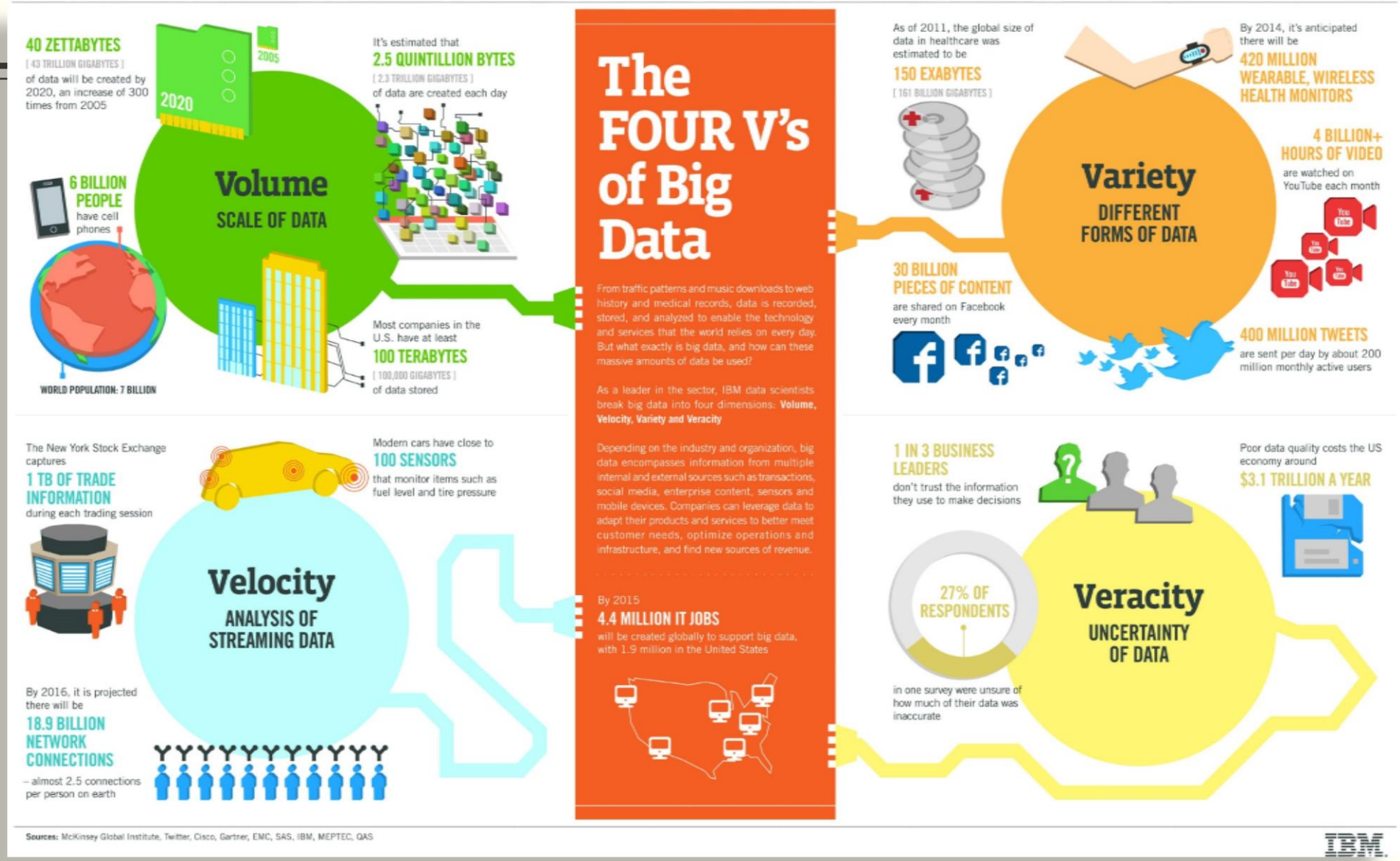
หลักในการวิเคราะห์ข้อมูลขนาดใหญ่


Big Data
Volume
Variety
Velocity

Big data


VOLUME
+ Terabytes
+ Records
+ Transactions
+ Tables, Files
THE 3Vs
OF BIG DATA
+ Batch
+ Near-time
+ Real-time
+ Streams
+ Structured
+ Unstructured
+ Semi-structured
+ All the above
VELOCITY
VARIETY

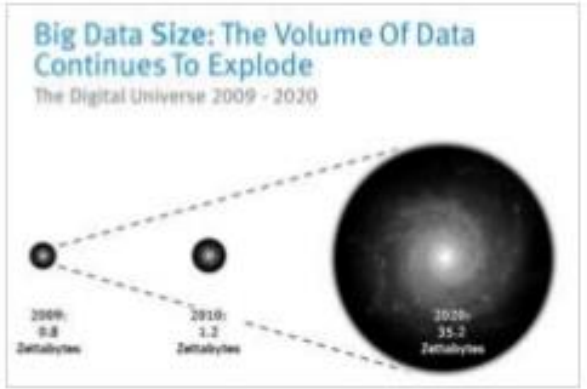
Big data


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2020
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
WORLD POPULATION: 7 BILLION
Volume
SCALE OF DATA
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
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
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
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
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
Variety
DIFFERENT FORMS OF DATA
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
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
Veracity
UNCERTAINTY OF DATA
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS
IBM

Big data

Key Characteristics of Big Data

1. **Data Volume**
   - 44x increase from 2010 to 2020 (1.2zettabytes to 35.2zb)



2. **Processing Complexity**
   - Changing data structures
   - Use cases warranting additional transformations and analytical techniques

3. **Data Structure**
   - Greater variety of data structures to mine and analyze

Copyright © 2012 EMC Corporation. All Rights Reserved.

Big data

What's Driving The Data Deluge?

Mobile Sensors

FACEBOOK UPLOADS
300 MILLION
PHOTOS EACH DAY

Social Media

Video Surveillance

Video Rendering

READING SMART METERS
EVERY 15 MINUTES IS
3000X MORE
DATA INTENSIVE

Smart Grids

OIL RIGS GENERATE
25000
DATA POINTS
PER SECOND

Oil Exploration

Medical Imaging

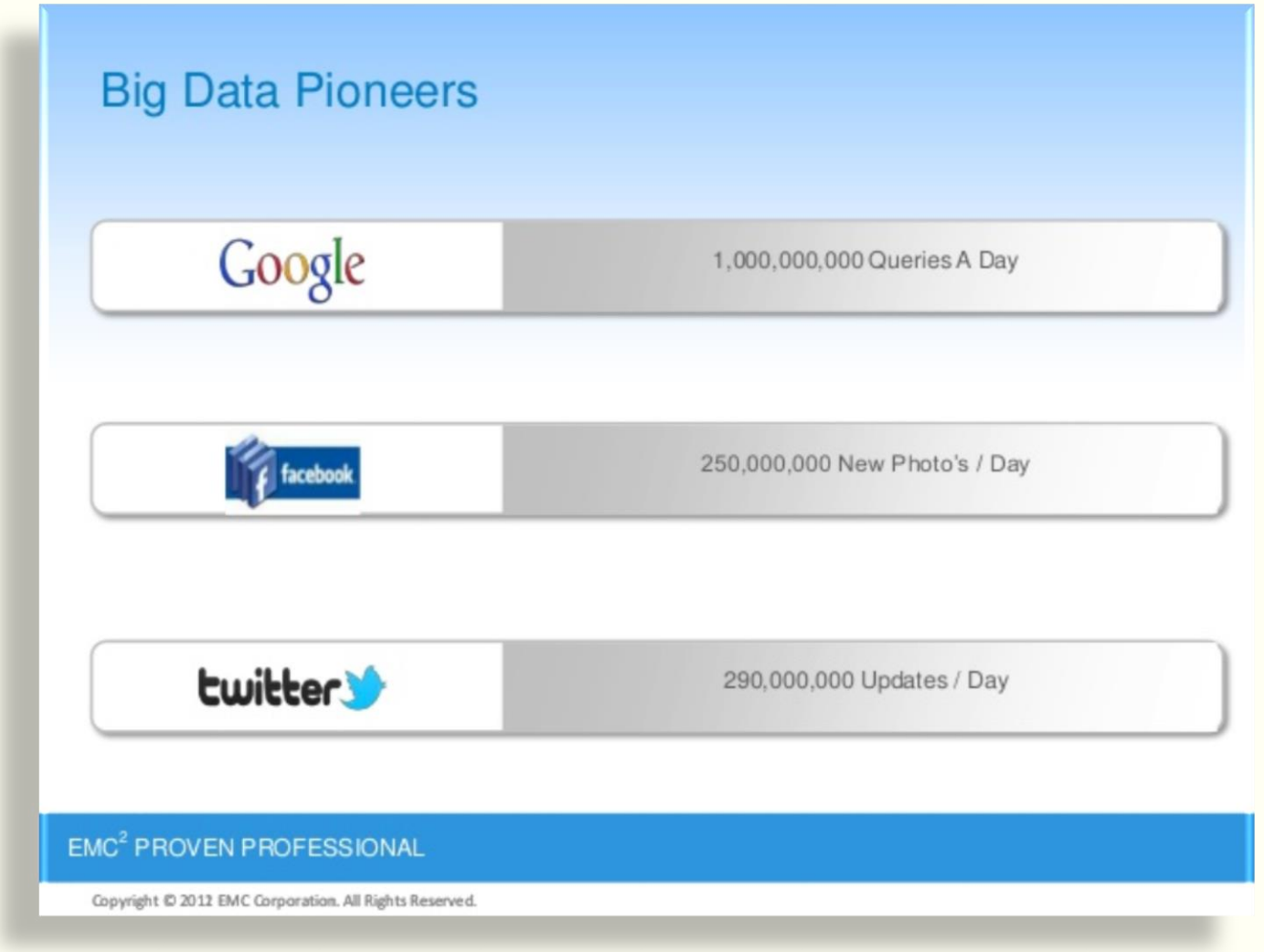
COST TO SEQUENCE
ONE
GENOME
HAS FALLEN FROM
$100M IN 2001
TO $10K IN 2011

Gene Sequencing

EMC²

© Copyright 2012 EMC Corporation. All rights reserved.

Big data


Big Data Pioneers
Google
1,000,000,000 Queries A Day
facebook
250,000,000 New Photo's / Day
twitter
290,000,000 Updates / Day
EMC² PROVEN PROFESSIONAL
Copyright © 2012 EMC Corporation. All Rights Reserved.

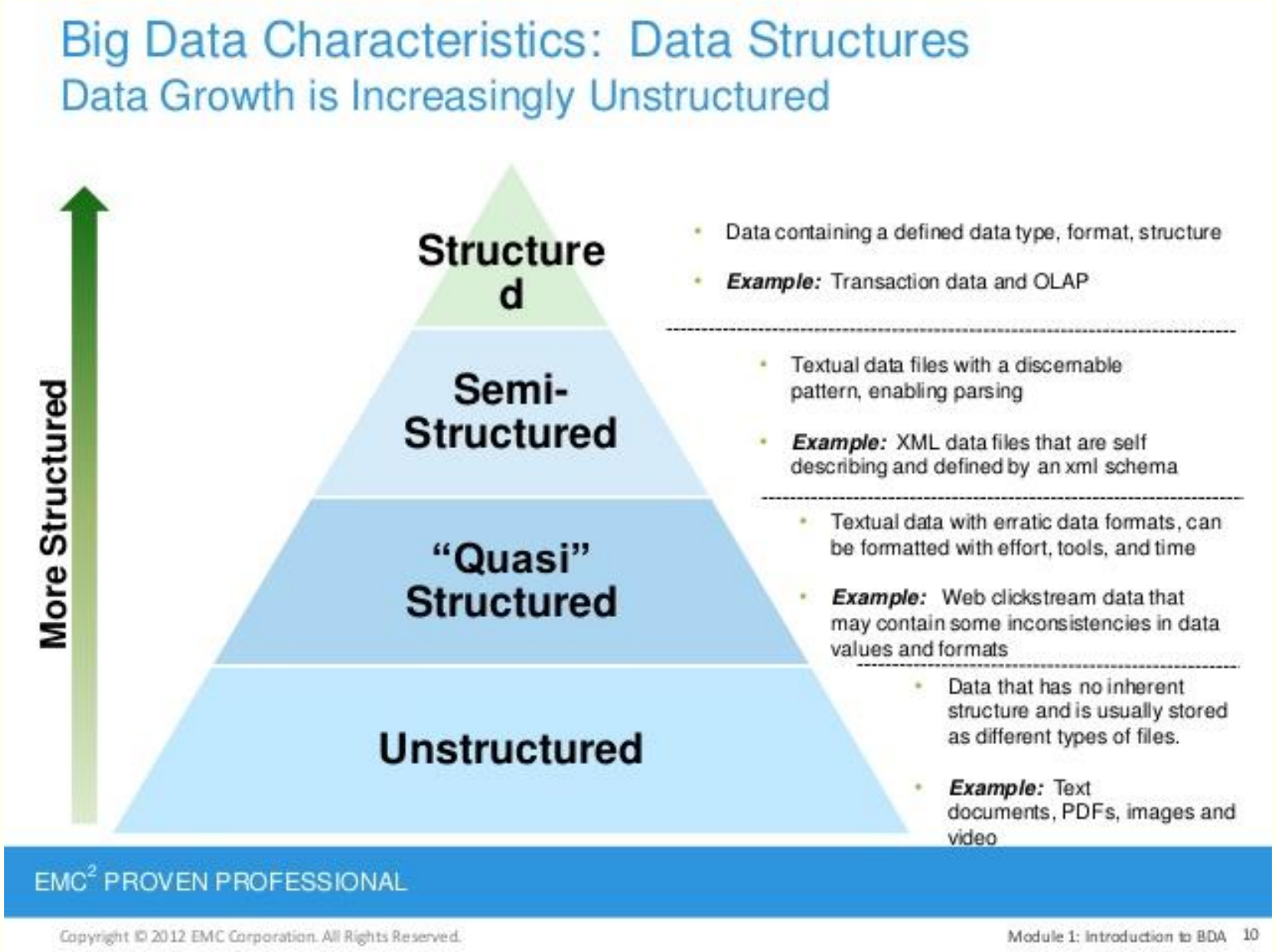
Big data


44,000,000
MESSAGES PROCESSED
486,000
PHOTOS
26
NEW REVIEWS
POSTED ON YELP
120
NEW ACCOUNTS
OPENED ON
LINKEDIN
MORE THAN
140
SUBMISSIONS
ON REDDIT
MORE THAN
2,315,000
SEARCHES
3,125,000
243,055
MORE THAN
21,000,000
MESSAGES SENT
70,000
VIDEO MESSAGES
SHARED
MORE THAN
195,000
MINUTES OF AUDIO CHATTING
ON WECHAT
Share
MORE THAN
3,000,000
ITEMS ARE
SHARED
MORE THAN
18,000
MATCHES MADE
972,000
DAILY SWIPES
ON TINDER
MORE THAN
69,500
HOURS OF
VIDEO WATCHED
ON NETFLIX
NETFLIX
IN
60
SECONDS
GO-Globe
CUSTOM WEB DEVELOPMENT
MORE THAN
150,000,000
E-MAILS ARE SENT
MORE THAN
48,000
APPS DOWNLOADED
ON IPHONE
MORE THAN
95,000
APPS DOWNLOADED
ON ANDROID
MORE THAN
430,000
TWEETS SENT
MORE THAN
2,700,000
VIDEO VIEWS AND
139,000 HOURS
OF VIDEO WATCHED
MORE THAN
39,300
HOURS OF MUSIC
LISTENED
AROUND
56,000
PHOTOS
UPLOADED
9,800
ARTICLES PINNED
ON PINTEREST
MORE THAN
280,000
SNAPS SENT
ON SNAPCHAT
MORE THAN
100
NEW DOMAINS
REGISTERED
14 NEW
SONGS ADDED
ON SPOTIFY
MORE THAN
300 HOURS
OF VIDEO ARE UPLOADED

Big data

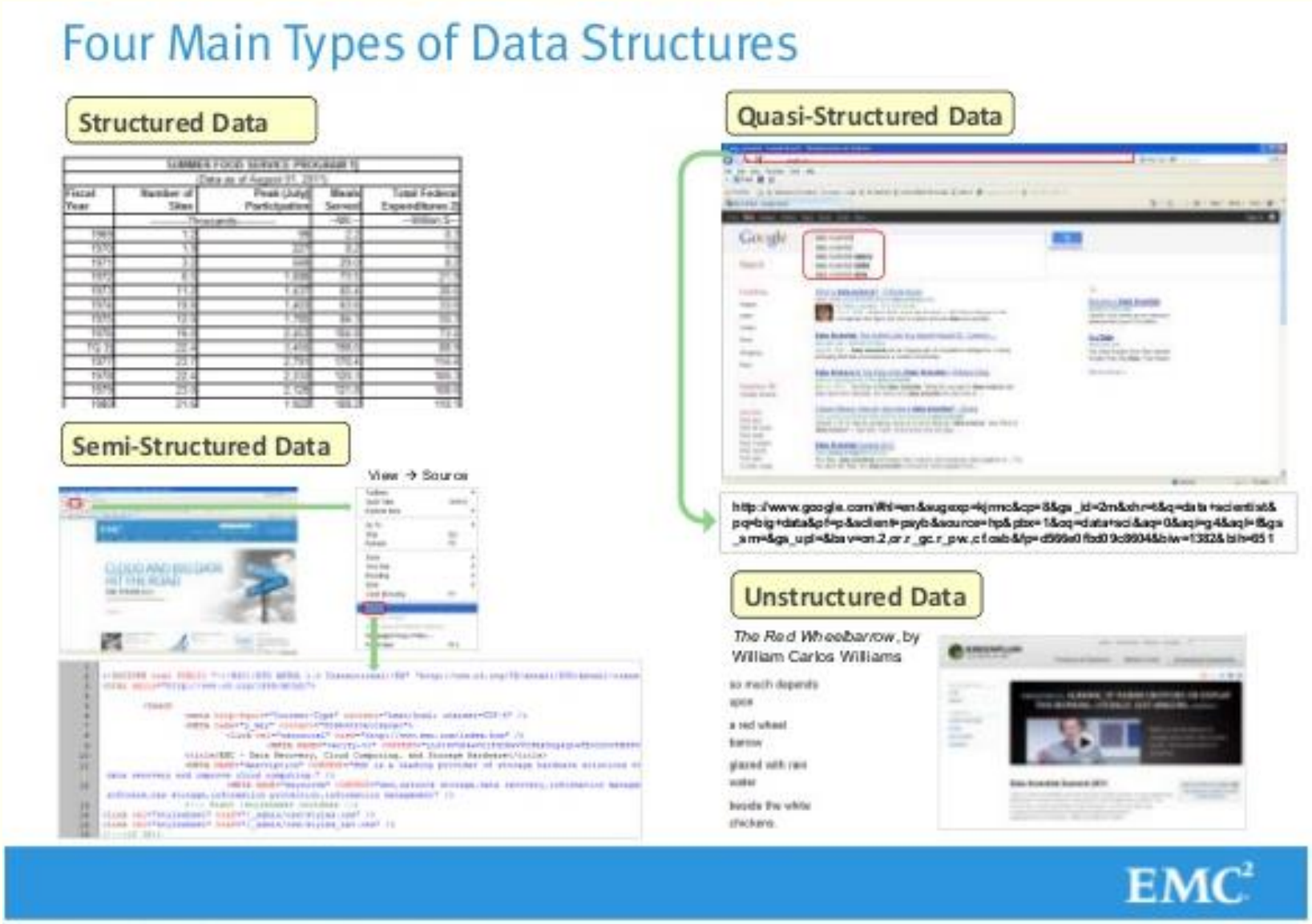

BIG DATA NOW
AND IN THE NEAR FUTURE
MB
MEGABYTE
(1 million bytes)
2MB
High Resolution Photograph
GB
GIGABYTE
(1 billion bytes)
1GB
Movie at TV Quality
TB
TERABYTE
(1 trillion bytes)
1TB
All X-ray Films in a Hospital
PB
PETABYTE
(1 million gigabytes)
2PB
All U.S. Academic Research Libaries
EB
EXABYTE
(1 quintillion bytes)
5EB
All words ever spoken by human beings
ZB
ZETTABYTE
(1 sextillion bytes)
.432ZB
World's technological Capacity to Receive Information Through One-Way Networks
YB
YOTTABYTE
named after yoda
Requires a Data Center the Size of Delaware AND Rhode Island
SOURCE: DEVON ENERGY

4 Types of Data (structures) in Big data era



Copyright © 2012 EMC Corporation. All Rights Reserved.

4 Types of Data (structures) in Big data era



Why–advanced data analytics?

Business Drivers for Advanced Analytics

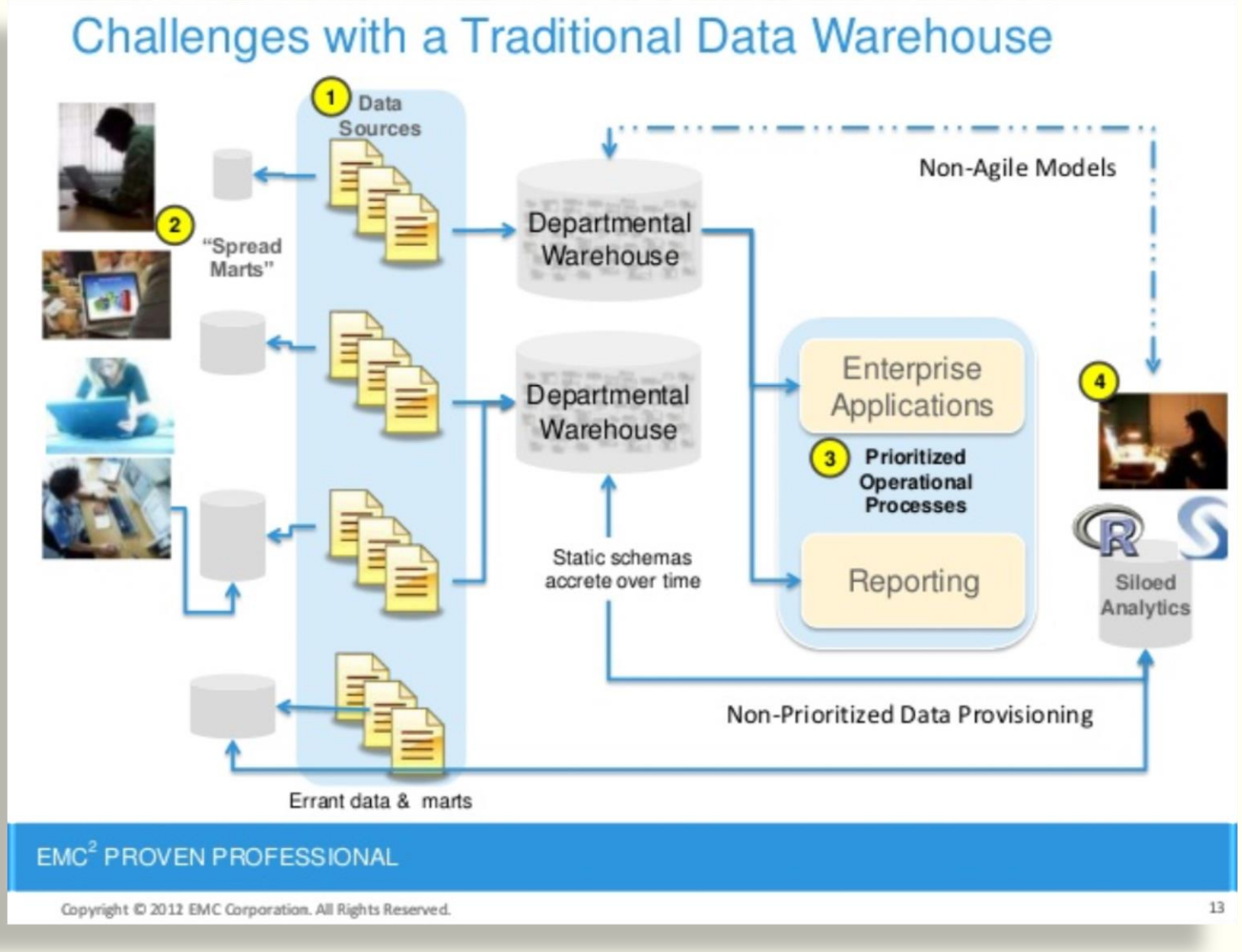
Current Business Problems Provide Opportunities for Organizations to Become More Analytical & Data Driven

	Driver	Examples
1	Desire to optimize business operations	Sales, pricing, profitability, efficiency
2	Desire to identify business risk	Customer churn, fraud, default
3	Predict new business opportunities	Upsell, cross-sell, best new customer prospects
4	Comply with laws or regulatory requirements	Anti-Money Laundering, Fair Lending, Basel II

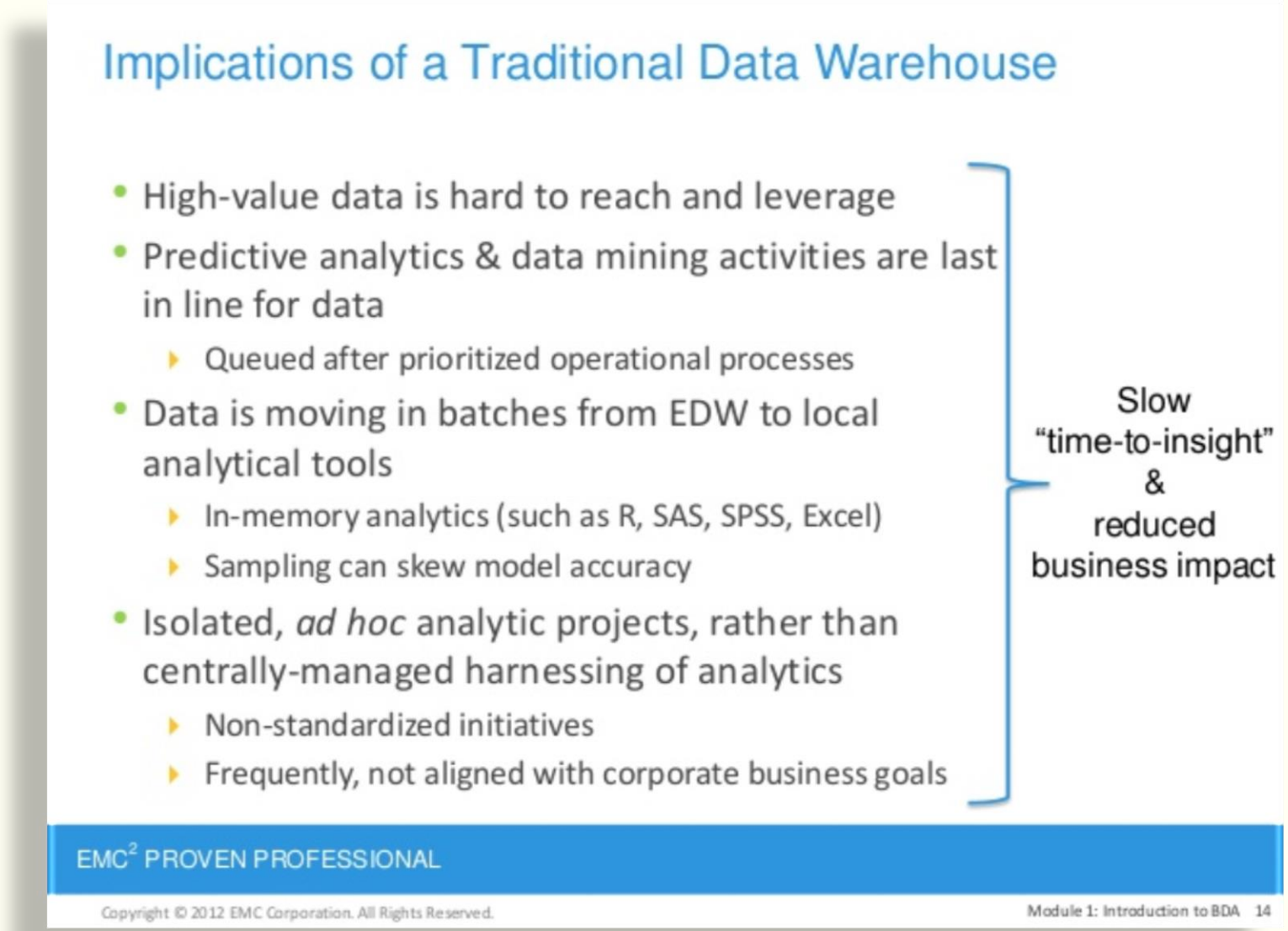
EMC²

Copyright © 2013 EMC Corporation. All Rights Reserved.

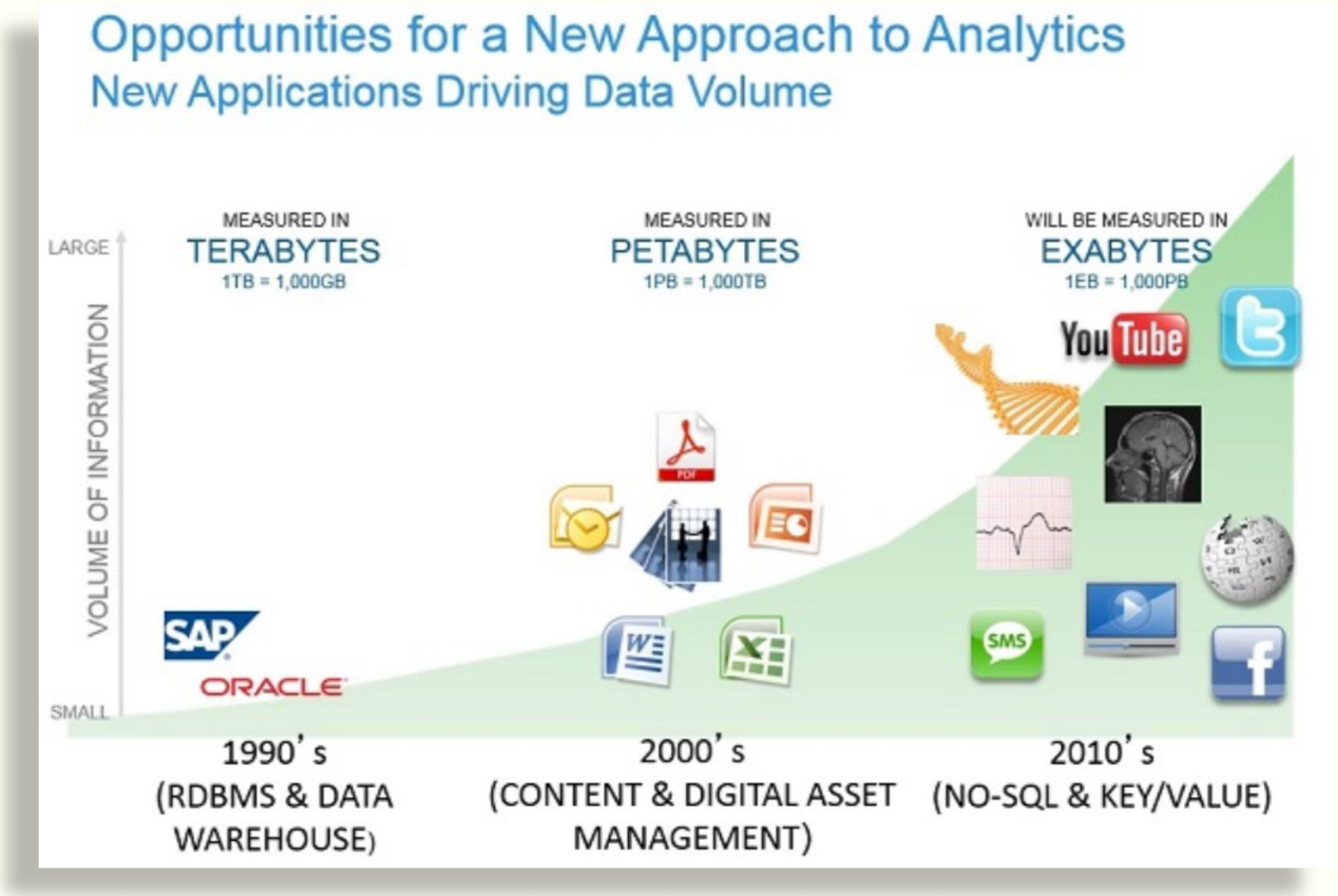
Why–big data analytics?



Why–big data analytics?



Why–big data analytics?



Why–big data analytics?

Considerations for Big Data Analytics

Criteria for Big Data Projects

1. Speed of decision making
2. Throughput
3. Analysis flexibility

New Analytic Architecture

Analytic Sandbox

Data assets gathered from multiple sources and technologies for analysis

- Enables high performance analytics using in-db processing
- Reduces costs associated with data replication into "shadow" file systems
- "Analyst-owned" rather than "DBA owned"

EMC² PROVEN PROFESSIONAL

Copyright © 2012 EMC Corporation. All Rights Reserved.

Big Data Analytics-use cases

Big Data Analytics: Industry Examples

1. Health Care
   - Reducing Cost of Care
2. Public Services
   - Preventing Pandemics
3. Life Sciences
   - Genomic Mapping
4. IT Infrastructure
   - Unstructured Data Analysis
5. Online Services
   - Social Media for Professionals

Copyright © 2012 EMC Corporation. All Rights Reserved.

1 Big Data Analytics: *Healthcare*

Situation

- Poor police response and problems with medical care, triggered by shooting of a Rutgers student
- The event drove local doctor to map crime data and examine local health care

Use of Big Data

- Dr. Jeffrey Brenner generated his own crime maps from medical billing records of 3 hospitals

Key Outcomes

- City hospitals & ER's provided expensive care, low quality care
- Reduced hospital costs by 56% by realizing that 80% of city's medical costs came from 13% of its residents, mainly low-income or elderly
- Now offers preventative care over the phone or through home visits

Copyright © 2012 EMC Corporation. All Rights Reserved.

2 Big Data Analytics: *Public Services*

Situation

- Threat of global pandemics has increased exponentially
- Pandemics spreads at faster rates, more resistant to antibiotics

Use of Big Data

- Created a network of viral listening posts
- Combines data from viral discovery in the field, research in disease hotspots, and social media trends
- Using Big Data to make accurate predications on spread of new pandemics

Key Outcomes

- Identified a fifth form of human malaria, including its origin
- Identified why efforts failed to control swine flu
- Proposing more proactive approaches to preventing outbreaks

Copyright © 2012 EMC Corporation. All Rights Reserved.

3 Big Data Analytics: *Life Sciences*

Situation

- Broad Institute (MIT & Harvard) mapping the Human Genome

Use of Big Data

- In 13 yrs, mapped 3 billion genetic base pairs; 8 petabytes
- Developed 30+ software packages, now shared publicly, along with the genomic data

Key Outcomes

- Using genetic mappings to identify cellular mutations causing cancer and other serious diseases
- Innovating how genomic research informs new pharmaceutical drugs

Copyright © 2012 EMC Corporation. All Rights Reserved.

4 Big Data Analytics: *IT Infrastructure*

Situation

- Explosion of unstructured data required new technology to analyze quickly, and efficiently

Use of Big Data

- Doug Cutting created Hadoop to divide large processing tasks into smaller tasks across many computers
- Analyzes social media data generated by hundreds of thousands of users

Key Outcomes

- New York Times used Hadoop to transform its entire public archive, from 1851 to 1922, into 11 million PDF files in 24 hrs
- Applications range from social media, sentiment analysis, wartime chatter, natural language processing

Copyright © 2012 EMC Corporation. All Rights Reserved.

5 Big Data Analytics: *Online Services*

Situation

- Opportunity to create social media space for professionals

Use of Big Data

- Collects and analyzes data from over 100 million users
- Adding 1 million new users per week

Key Outcomes

- LinkedIn Skills, InMaps, Job Recommendations, Recruiting
- Established a diverse data scientist group, as founder believes this is the start of Big Data revolution

Copyright © 2012 EMC Corporation. All Rights Reserved.

Case study: Starbuck-customer

Case study: Real time-traffic

Case study: Wholesale

Case study: Hospital

Case study: Banking