




Big Data: Concepts, Challenges & Opportunities

Shiow-yang Wu (吳秀陽)
CSIE, NDHU, Taiwan, ROC

Lecture material is mostly home-grown, partly
taken with permission and courtesy
from Professor Shih-Wei Liao of NTU.



Outline

- What is Big Data?
- Big data examples
- Big data concepts
- Challenges and opportunities
- Summary
- Next: Big Data Computing & Systems

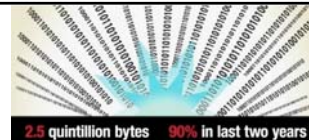
Simple to Start



- What is the maximum file size you have dealt with so far?
 - Movies/Files/Streaming video that you have used?
 - What have you observed?
- What is the maximum download speed you get?
- Simple computation
 - How much time to just transfer?

Memory unit	Size	Binary size
kilobyte (kB/KB)	10^3	2^{10}
megabyte (MB)	10^6	2^{20}
gigabyte (GB)	10^9	2^{30}
terabyte (TB)	10^{12}	2^{40}
petabyte (PB)	10^{15}	2^{50}
exabyte (EB)	10^{18}	2^{60}
zettabyte (ZB)	10^{21}	2^{70}
yottabyte (YB)	10^{24}	2^{80}

What is Big Data?



“Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few.”

This data is “**big data**.”

(<https://www-01.ibm.com/software/in/data/bigdata/>)

(<http://www.bbc.com/news/technology-20120816-every-day-we-create-2-5-quintillion-bytes-of-data-created-daily/>)

Big Data EveryWhere!




- Lots of data is being collected and warehoused
 - Web data, e-commerce
 - Purchases at department/grocery stores
 - Bank/Credit Card transactions
 - Social Network
 - Sensors
 - Internet of things
 - ...



OSIE59830 Big Data Systems Introduction 5

Huge Amount of Data



- There are huge volumes of data in the world:
 - From the beginning of recorded time until 2003,
 - We created 5 billion gigabytes (exabytes) of data.
 - In 2011, the same amount was created every **two days**.
 - In 2013, the same amount of data is created every **10 minutes**.
 - In 2017 ?

OSIE59830 Big Data Systems Introduction 6

Data/Minute on the Internet



- 2014, 2.4 billion users. 3.4 billion by 2016. In 2017, **3.8 billion Internet users**.
- **Every minute**, the following happens on the internet:
 - **Social media** gains 840 new users.
 - 455,000+ **Tweets** !
 - 400+ hours new video on **Youtube**! Watching 4,146,600 videos.
 - 46,740 million posts on **Instagram**!
 - 3 million posts, 510,000 comments, 293,000 status updates, 136,000 photos, 4+ million like button clicks on **Facebook**.
 - 3,607,080 **Google** searches.
 - Worldwide, 15,220,700 **texts** are sent every minute!

(<https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/>)

How much data per day?



- 1,209,600 new data on social media
- 656 million tweets
- 4+ million hours of video upload, 5.97 billion hours watched on Youtube
- 67,305,600 Instagram posts
- 2+ billion monthly active and 1.32 billion daily active Facebook users (June 2017)
- 4.3 billion Facebook messages posted
- 5.75 billion Facebook likes
- 22 billion texts sent
- 5.2 billion daily Google searches in 2017.

640K ought to be enough for anybody.



How Big is Big?



- Stu Feldman (head of Schmidt Sciences, ex Google VP) says at least 10TB in terms of data rate.
 - PB level is now commonly recognized as big
- “Please call your work Small Data, NOT Big Data.”
- Should we rename this course to “Small Data Systems” ?
- Luckily, Google agrees that their work is Extreme Data.

Is Big Data a real discipline standing on its own?



- Some heavyweights said “Big Data is not new. Database and data mining have been around for more than 30 years.”
- Rebuttal: Big data already disrupted the field of data models and relational database and demanded new ways of building systems. (量變造成質變)
- In the case of data mining above, see the free book on Professor Ullman: “*Mining of Massive Datasets*” (<http://www.mmds.org/>)

Is Big Data just a small part of Cloud Computing?




- Some heavyweights said “Big Data is just a small part of Cloud Computing. Don’t make a big deal out of it.”
- Rebuttal:
 - They are probably just different perspectives:
 - Cloud focuses more on elastic computing and warehouse computing.
 - Big Data focuses more on enterprise cloud and possible-time and real-time analytics.
 - It’s a **big deal**:
 - Many impossible business model → possible now.

Big Data Everywhere



- Big Data Everywhere :
 - E.g., Google indexed World Wide Web pages
 - Which demanded the creation of MapReduce and NoSQL.
 - Big Data is not just an isolated discipline: Big Data is not just **red hot** in one discipline.
 - When **data explodes**, new data properties appear, business applications emerge, which solidify the discipline
 - E.g., We now know that not all Big Data Applications are like indexing WWW.


Big Data Everywhere



Applications Everywhere: Domain knowledge, Business models
Analytics: Ad-hoc analytics, statistics, machine learning
Systems: Tools, Infrastructure

CSIE59830 Big Data Systems Introduction 13

Big Data from CSO's Perspective



- New revenue vs. cost reduction
 - Top-line vs. Bottom-line
 - CSO (Chief Strategy Officer) vs. CIO (Chief Information Officer)
- First, here we talk about Big Data from CSO's perspective
- IBM says, **Big Data = Big Business**
- Many impossible business model → possible now.

CSIE59830 Big Data Systems Introduction 14

Big Data from CIO's Perspective

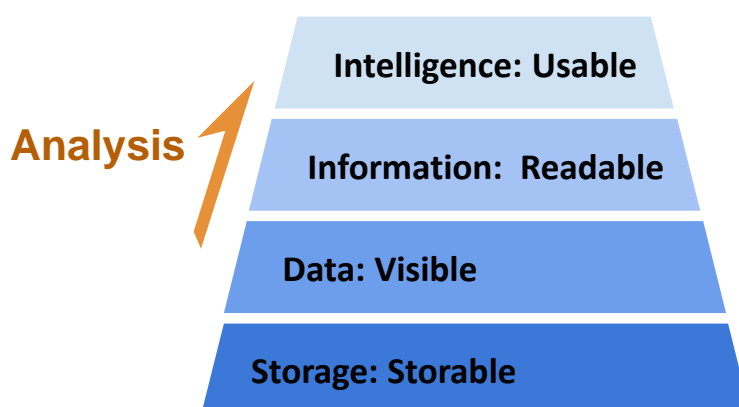


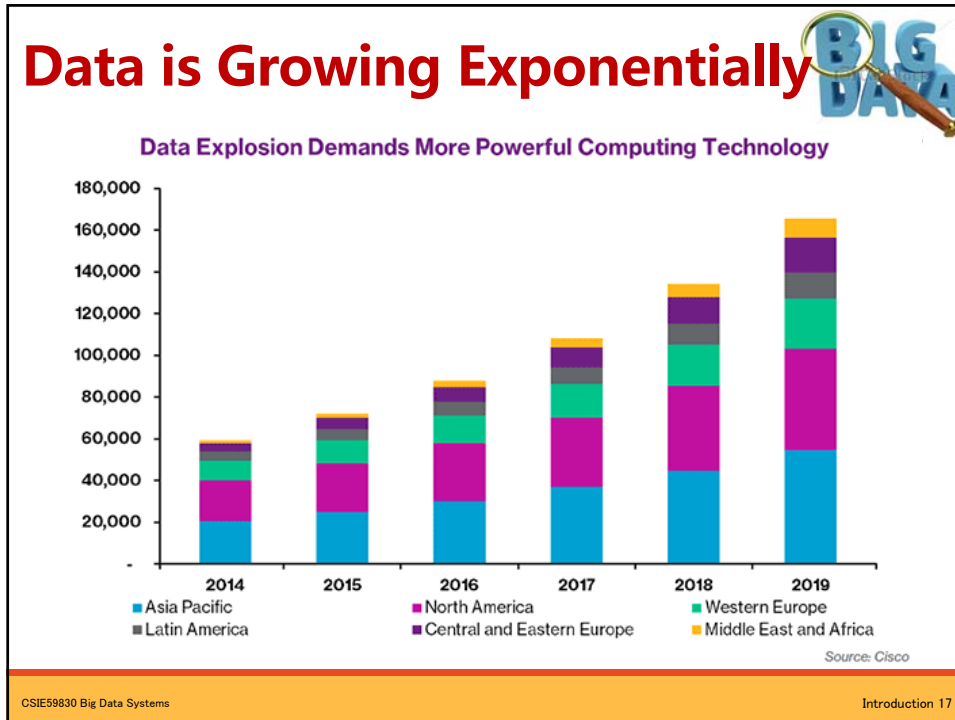
- CIO = Chief Information Officer
- To CIOs, Big Data means no more **IOE** :
 - No more expensive **I**BM machines : PC is enough
 - No more expensive **O**racle software : open-source, NoSQL
 - No more expensive **E**MC storage: No RAID. Just common hard drives
- To CIOs, Big Data means **large scale distributed computing** with **commodity systems** on **open-source software**

Back to Basics: What is Data?



Texts, Records, Statistics...





Big Data Examples

Large Hydron Collider (LHC)



Large Hydron Collider (LHC)



- 150 millions sensors
- 600 million collisions per second
- Recording all experimental data takes 500EB(1EB=1000PB) per day
- Annual rate of 150 million PB (before replication)
- 200 times higher than all the other sources combined in the world

The Earthscope (地球鏡)



- The Earthscope is one of the world's largest science project. Designed to track North America's geological evolution, this observatory records data over 3.8 million square miles, amassing 67 terabytes of data. It analyzes seismic slips in the San Andreas fault, sure, but also the plume of magma underneath Yellowstone and much, much more. (http://www.msnbc.msn.com/id/44363598/ns/technology_and_science-future_of_technology/#.TmetOdQ--ul)



Annual budget: \$25,000,000
 Construction cost: \$197,000,000
 Staff: 110
 Physical size: 3.8 million square miles
 Scientific utility: 10
 WAFI: 10
 Web factor: 10

WalMart




- The world's biggest retailer with over 20,000 stores in 28 countries.
- Building the world's biggest private cloud, to process 2.5 petabytes of data every hour.
- Merely storing them is already a big problem.
- Let alone processing and analyzing them.

Facebook




- 2.13 billion monthly active users (1.37 billion daily active users)
- 500,000 new users every day; 6 new profiles every second
- Every 60 seconds on Facebook: 510,000 comments are posted, 293,000 statuses are updated, and 136,000 photos are uploaded.
- 4 new PBs of data per day
- Hive data warehouse with 300 PBs of data
- 100 million hours of daily video watch
- 4 million likes every minute
- 250 billion photos (350 million per day)


Who's Generating Big Data




Social media and networks
(all of us are generating data)



Scientific instruments
(collecting all sorts of data)



Mobile devices
(tracking all objects all the time)



Sensor technology and networks
(measuring all kinds of data)



- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

CSIE59830 Big Data Systems
Introduction 25



The Model Has Changed...

- **The Model of Generating/Consuming Data has Changed**

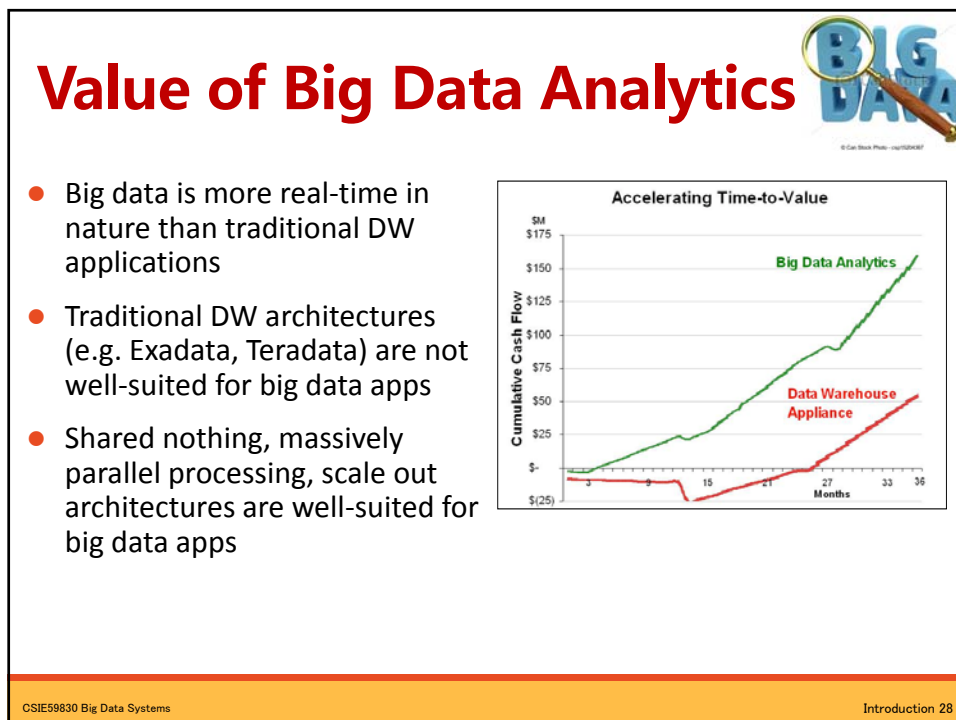
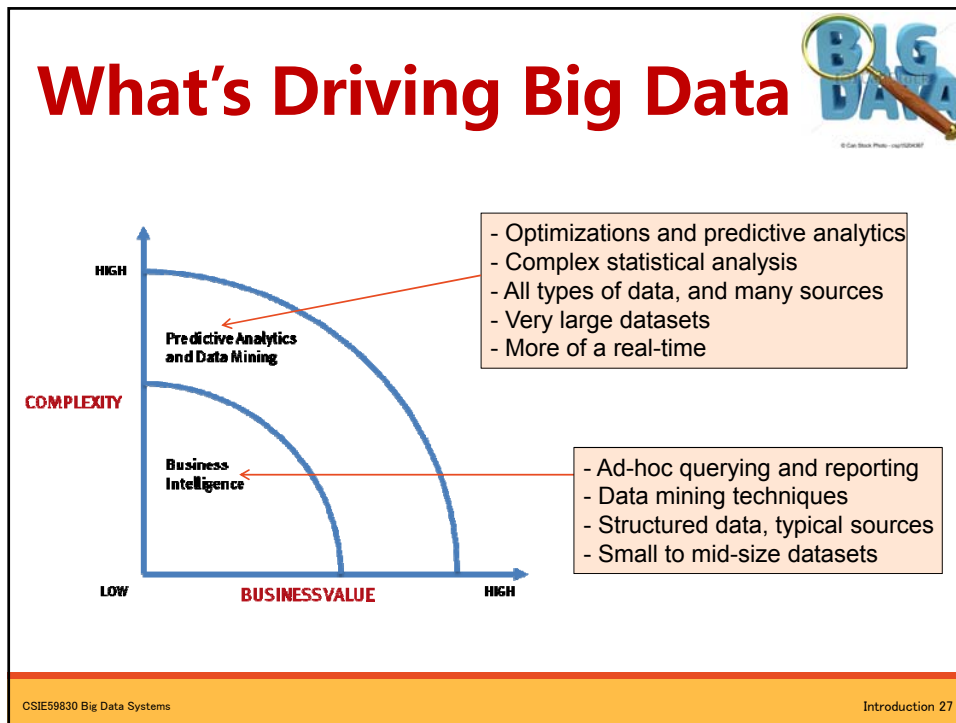
Old Model: Few companies are generating data, all others are consuming data


→


New Model: all of us are generating data, and all of us are consuming data


→


CSIE59830 Big Data Systems
Introduction 26



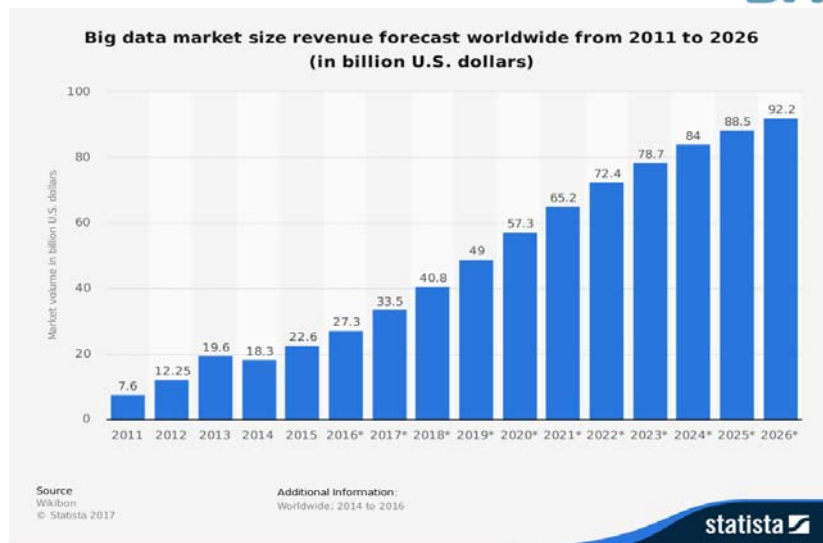
Data Business



- Retails(零售)
- Traffic(交通)
- Health(健康)
- Education(教育學習)
- Manufacturing(製造)
- Contents(數位內容)
- Agriculture(農業)
- Advertising(廣告)
- Telecommunication(電信)
- Finance(金融)
- Smart grid(智慧電網)



Big Data Market Forecast



Types of Data



- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Multimedia Data (images, audio, video)
- Graph Data
 - Social Network, Semantic Web (RDF), ...
- Streaming Data
 - You can only scan the data once

Value of Big Data



- When the data size grows · intrinsic properties of data emerges
- “A kilo of data is worth more than a gram of algorithm”
- A small improvement of algorithm is no longer important
- The key is on whether the algorithm **scale** !!

What to do with these data?

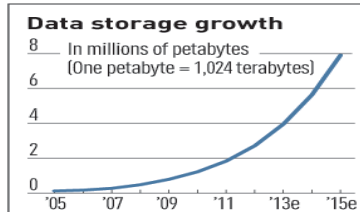


- Aggregation and Statistics
 - Data warehouse and OLAP
- Indexing, Searching, and Querying
 - Keyword based search
 - Pattern matching (XML/RDF)
- Knowledge discovery
 - Data Mining
 - Statistical Modeling

Challenges in Handling Big Data

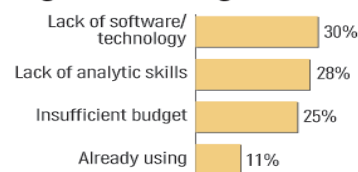


Big Data Boom



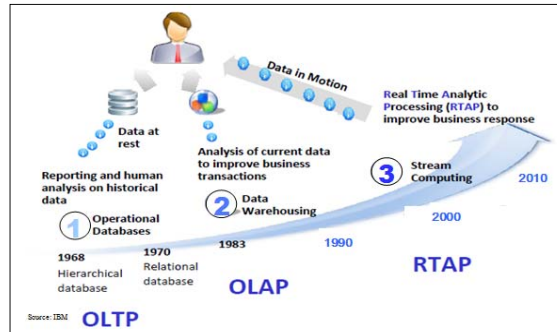
Sources: IDC, DataXu

Big data challenge



- **The Bottleneck is in technology**
 - New architecture, algorithms, techniques are needed
- **Also in technical skills**
 - Experts in using the new technology and dealing with big data

Harnessing Big Data



- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

Myths About Big Data



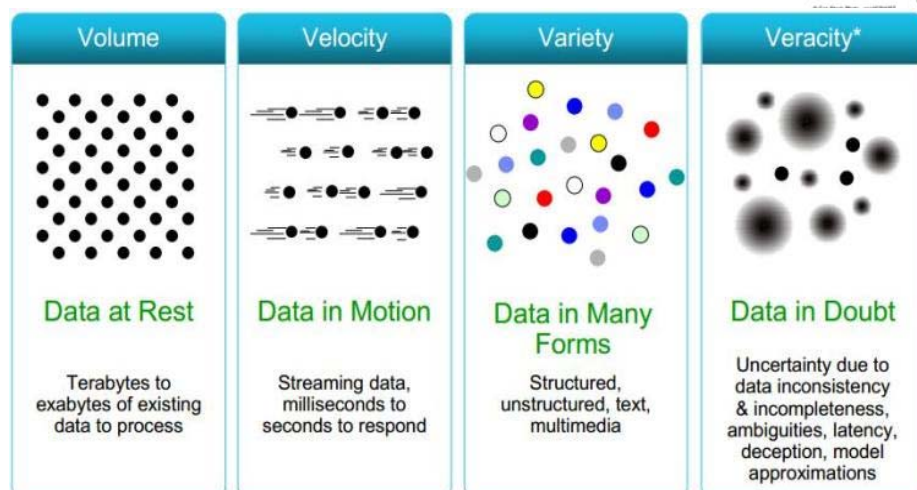
- **“Big Data is Only About Massive Data Volume”**
 - Volume is just an element of Big Data
- **“Big Data is all-powerful”**
 - Can get All Of The Data
 - Big Data Yields Certainty
 - Can answer WHY

“Big Data is Only About Massive Data Volume”



- **4 Vs**
 - **Volume** : The start point of Big Data, but the least important of 4 elements.
 - **Variety** : Traditional data management processes can't cope with the heterogeneity of big data.
 - **Velocity** : Data is generated in real time, with demands for usable information to be served up immediately.
 - **Veracity** : Refers to the biases, noise and abnormality in data. How to make data to be trusted for the organization to make crucial decision?

4Vs of Big Data



4Vs of Big Data

Big data Expands on 4 fronts

<http://whatis.techtarget.com/definition/3Vs>

OSIE59830 Big Data Systems Introduction 39

4Vs of Big Data

The FOUR V's of Big Data

Volume
SCALE OF DATA

- 40 ZETTABYTES (140 TRILLION GIGABYTES) of data will be created by 2020, an increase of 300 times from 2009.
- It's estimated that 2.5 QUINTILLION BYTES (2.5 TRILLION GIGABYTES) of data are created each day.
- Most companies in the U.S. have at least 100 TERABYTES (100,000 GIGABYTES) of data stored.

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.

Variety
DIFFERENT FORMS OF DATA

- As of 2013, the global size of data in healthcare was estimated to be 150 EXABYTES (150 TRILLION GIGABYTES).
- By 2014, it's anticipated there will be 420 MILLION WEARABLE, WIRELESS HEALTH MONITORS.
- 30 BILLION PIECES OF CONTENT are shared on Facebook every month.
- 4 BILLION+ HOURS OF VIDEO are watched on YouTube each month.
- 400 MILLION TWEETS are sent per day by about 200 million monthly active users.

Veracity
UNCERTAINTY OF DATA

- 1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions.
- 27% OF RESPONDENTS don't trust the information they use to make decisions.
- 50 one survey were unsure of how much of their data was inaccurate.
- 1 in 3 BUSINESS LEADERS don't trust the information they use to make decisions.
- Flaw data quality costs the US economy around \$3.1 TRILLION A YEAR.

By 2013, 4.4 MILLION IT JOBS will be created globally to support big data, with 1.5 million in the United States.

Depending on the industry and organization, big data encompasses information from multiple internal and external sources in marketing, social media, operational systems, sensors, and mobile devices. Companies can leverage data to expand their products and services to better meet customer needs, optimize operational and infrastructure, and find new sources of revenue.

Source: McKinsey Global Institute, Forrester, Cisco, Statista, EMC, IDC, IHS, NIST, PCAST, GSA

OSIE59830 Big Data Systems Introduction 40

The 5 Vs of BigData

The diagram features five horizontal arrows of different colors pointing outwards from a central vertical line. From top to bottom, they are: a blue arrow for Volume, a green arrow for Velocity, an orange arrow for Variety, a red arrow for Veracity, and a pink arrow for Value. Each arrow is accompanied by a brief definition.

- Volume**: The size of the data
- Velocity**: The speed at which the data is generated
- Variety**: The different type of data
- Veracity**: The trustworthiness of the data in terms of accuracy
- Value**: Just having BigData is of no use unless we can turn it into value

Powered by StackDataLabs

CSIE59830 Big Data Systems Course Information 41

5 V's of Big Data

The diagram consists of five overlapping circles arranged in a circle around a central pentagon. The pentagon is labeled '5 Vs of Big Data'. Each circle contains a list of characteristics for its respective 'V'.

- Volume**
 - Terabytes
 - Records/Arch
 - Transactions
 - Tables, Files
- Velocity**
 - Batch
 - Real/near-time
 - Processes
 - Streams
- Value**
 - Statistical
 - Events
 - Correlations
 - Hypothetical
- Veracity**
 - Trustworthiness
 - Authenticity
 - Origin, Reputation
 - Availability
 - Accountability
- Variety**
 - Structured
 - Unstructured
 - Multi-factor
 - Probabilistic

CSIE59830 Big Data Systems Introduction 42

Other V's









- **Validity:** The issue of collecting data which is correct and accurate for the intended use.
- **Volatility:** How long is data valid and how long should it be stored.
- **Variability:** Big data is variable, i.e. variance in meaning, changing of meaning (rapidly).\
- **Visualization:** Making data comprehensible, easy to understand and read.

The V's of Big Data



Big Data is not just Hadoop

Understand and navigate federated big data sources		Federated Discovery and Navigation
Manage & store huge volume of any data		Hadoop File System MapReduce
Structure and control data		Data Warehousing
Manage streaming data		Stream Computing
Analyze unstructured data		Text Analytics Engine
Integrate and govern all data sources		Integration, Data Quality, Security, Lifecycle Management, MDM

CSIE59830 Big Data Systems Introduction 45

Emerging Technologies for Managing Big Data

- Architecture
- Storage
- Computing
- Graph
- Database/Data warehousing
- Stream processing
- Real-time Analytics & Business knowledge
- Big data as a service

CSIE59830 Big Data Systems Introduction 46

Problems when data is BIG



- How to store ?
- How to retrieve ?
- How to process ?
- How to analyze ?

How to store ?



- It's not likely to store it on a single machine
 - Facebook generates TBs of data every day
 - 400+hr. of video is uploaded to Youtube every minute
- Distributed File System
 - Google File System (GFS)
 - Hadoop Distributed File System (HDFS)

How to retrieve ?



- You may want to use a database system to organize data
 - MySQL, PostgreSQL....
- Unfortunately, they don't scale well to this level...
 - One naive reason is that they usually run on only 1 machine.

Structure of Data



- The structure of data can be classified into:
 - **Structured data**: Data with a defined format and structure (RDB, spreadsheets, CSV, ...)
 - **Semi-structured data**: Textual data files with a flexible structure that can be parsed (XML, ...)
 - **Quasi-structured data**: Textual data with erratic data formats (Web click stream data, ...)
 - **Unstructured data**: Data that have no inherent structure (text docs, PDF files, images, videos, ...)
- Use different tools for different cases.

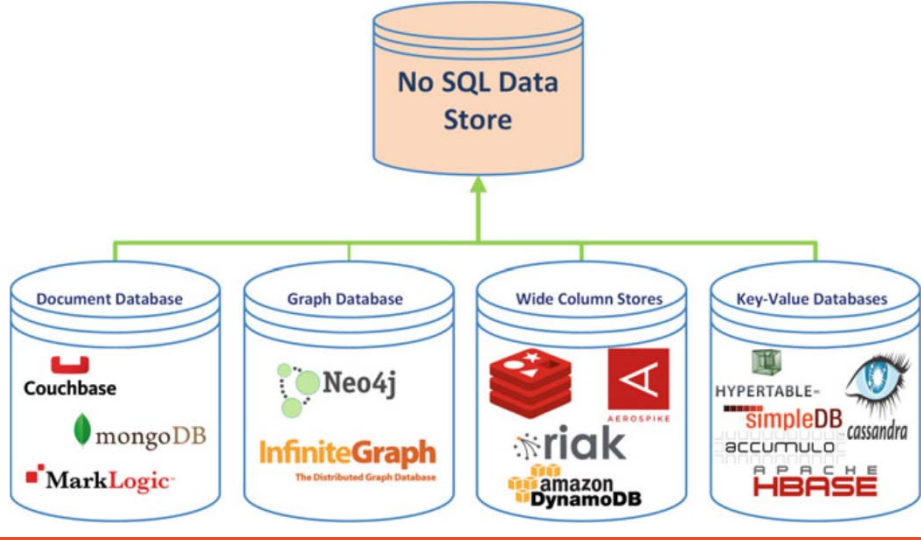
Storage & Warehousing

- BigTable / HBase
- Cassandra
- MongoDB
- Hive / Spark SQL



OSIE59830 Big Data Systems Introduction 51

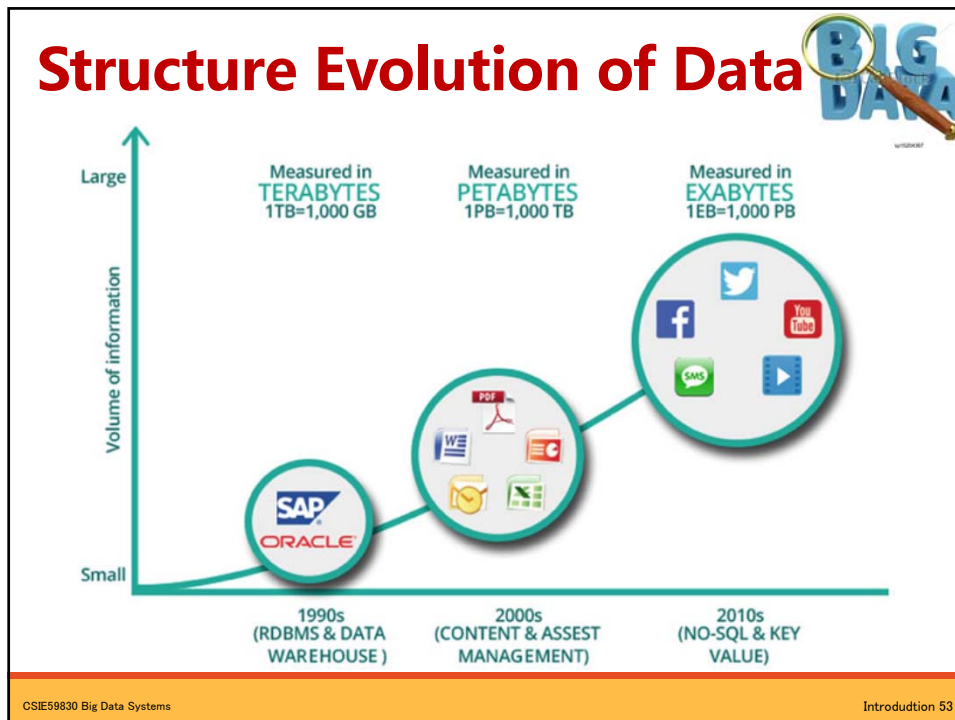
Types of NoSQL Data Stores



No SQL Data Store

- Document Database**: Couchbase, mongoDB, MarkLogic
- Graph Database**: Neo4j, InfiniteGraph
- Wide Column Stores**: riak, amazon DynamoDB
- Key-Value Databases**: HYPERTABLE, SimpleDB, accumulo, cassandra, APACHE HBASE

OSIE59830 Big Data Systems Introduction 52






How to process ?


- Finally, we can get the data we need efficiently from the monster-like data set.
- And it's time to do something cool now
 - Learning, mining, retrieval...
- But you'll soon face some trouble...
 - Data can't fit in memory / disk on a single machine
 - Not powerful enough with a single machine

CSIE59830 Big Data Systems Introduction 54


Computing



- MapReduce 
- Massive Parallel Processing
- Spark: in-memory computing 
- Storm: real-time streaming
- ...




APACHE
STORM[™]
Distributed • Resilient • Real-time



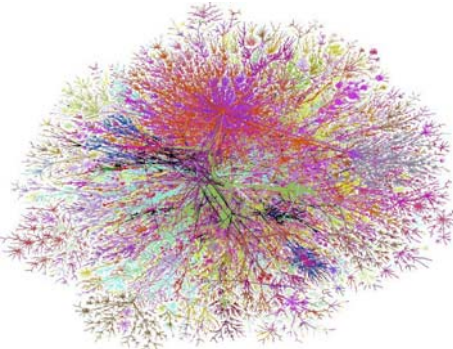
Spark
Lightning-Fast Cluster Computing

CSIE59830 Big Data Systems Introduction 55

Graph Computing




- Why graph?
 - Graphs abstract application-specific algorithms into generic problems represented as interactions using vertices and edges
 - Algorithms vary in their computational requirements





CSIE59830 Big Data Systems Introduction 56

Big Graph Processing

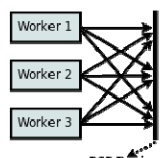


- Pregel: from Google
- Apache Hama
- GraphLab

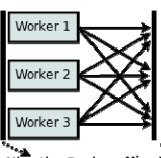
The Apache Hama Project
<http://hama.apache.org/>

Superstep 1



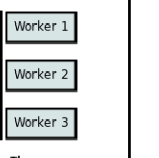
BSP Barrier

Superstep 2



Migration Barrier

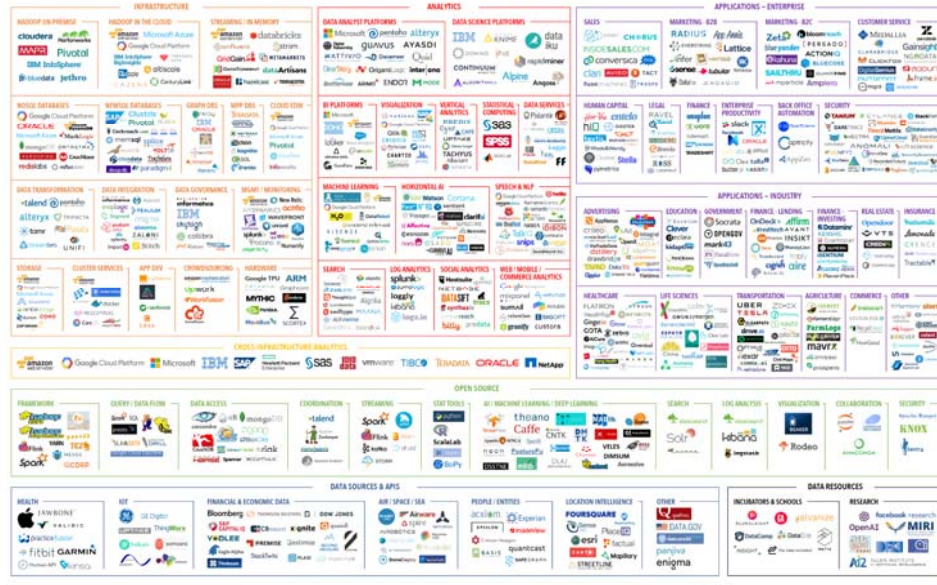
Superstep 3



Migration Planner

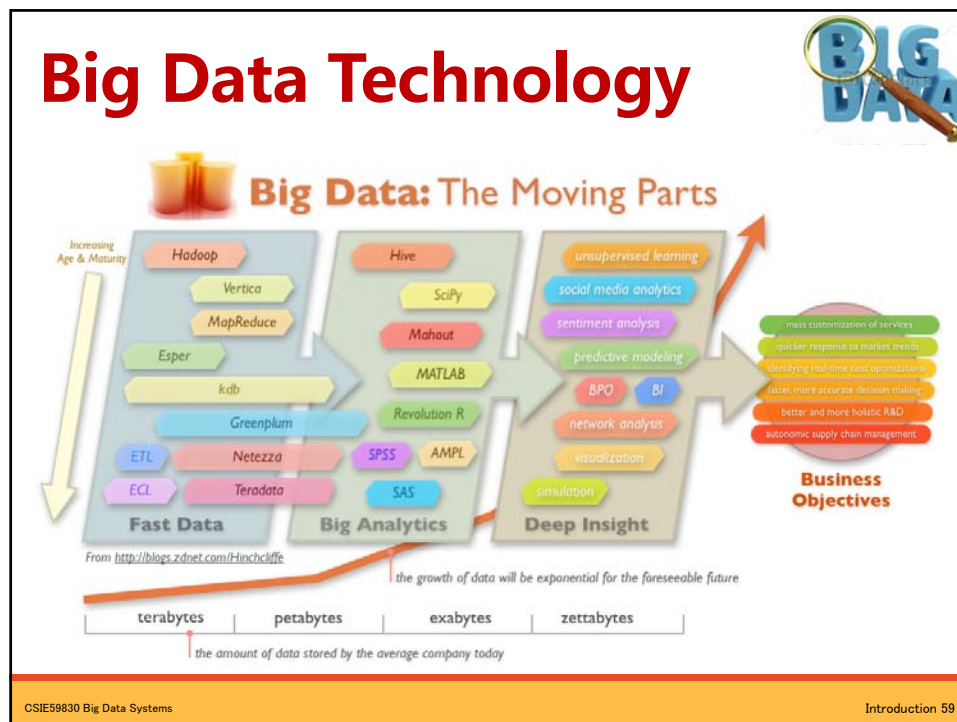
OSIE59830 Big Data Systems Introduction 57

Big Data Landscape 2017



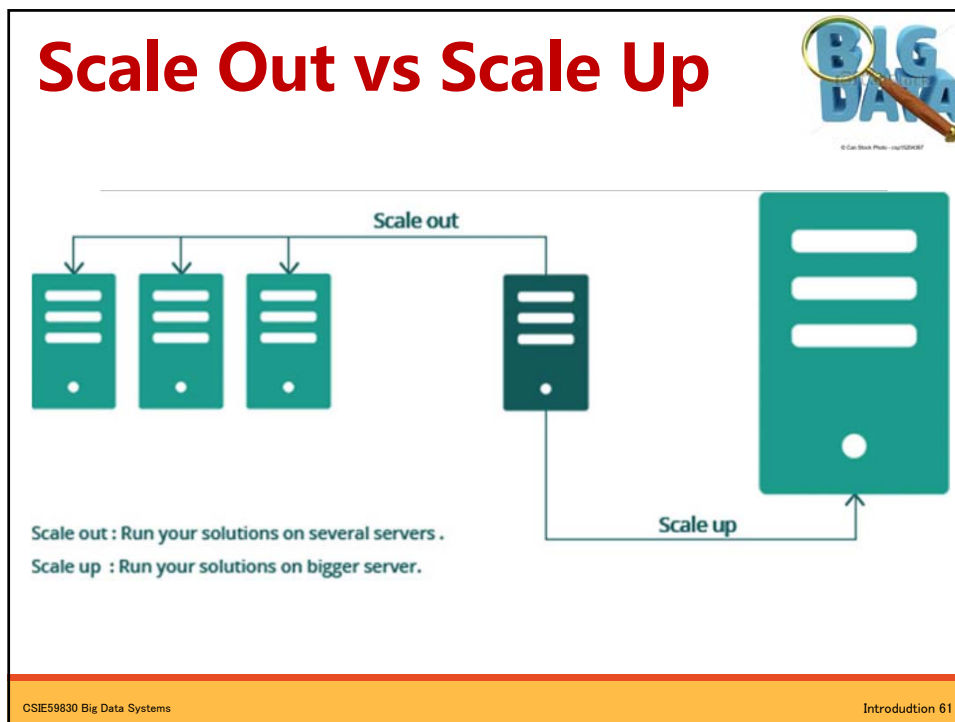
V2 - Last updated 5/3/2017 © Matt Turck (@matturck), Jim Hao (@jimhao), & FirstMark (@firstmarkcap) matturck.com/bigdata2017 FIRSTMARK

OSIE59830 Big Data Systems Introduction 58



Summary

- Big data era is coming!
- Big data calls for completely new models of storing, managing, processing, and analyzing data.
- The primary concern is to **scale out** rather than **scale up**.
- Big data means big business!



Coming Lectures

- Big data processing architecture
 - Hadoop
- General purpose big data processing system
 - MapReduce
 - Spark
- Data mining algorithms based on MapReduce and Spark

OSIE59830 Big Data Systems

Introduction 62

Coming Lectures



- Storage systems for big data processing
 - Google File System
 - Hadoop Distributed File System
 - Google BigTable
- NoSQL database systems
 - HBase
 - Cassandra
 - MongoDB
- Data warehousing systems
 - Google BigQuery
 - Apache Hive
 - Spark SQL

Coming Lectures



- Systems for big graph processing
 - Google Pregel
 - Apache Hama
 - GraphLab
- Systems for stream processing
 - Apache Flink
 - Apache Storm
 - Spark Streaming
- Big data analytics**
 - Google Dremel, Apache Drill and Apache Impala
 - Google Cloud Platform vs Amazon Web Services
 - Beyond Hadoop

