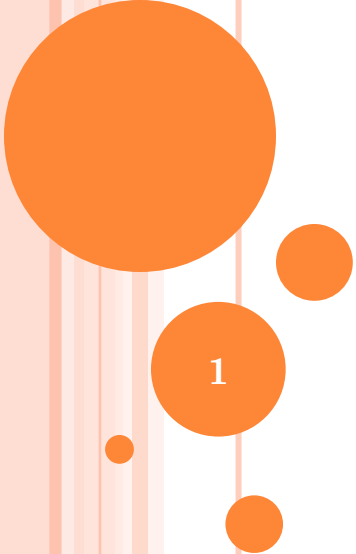




EE4221: LECTURE 5 – BIG DATA AND COULD

Dr Eric Wong

1

Big Data and
Cloud, isn't
these two
different areas
all together?
How do they
really come
together?



Big Data and Cloud Computing

- Big Data is all about extracting 4Vs
- Cloud focuses on On-Demand, Elastic, Scalable, Pay-Per use Self Service models



40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Volume SCALE OF DATA

It's estimated that 2.5 QUINTILLION BYTES

[2.3 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

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



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

Velocity ANALYSIS OF STREAMING DATA



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



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]



**30 BILLION
PIECES OF CONTENT**
are shared on Facebook every month



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



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

The Evolution of Big Data

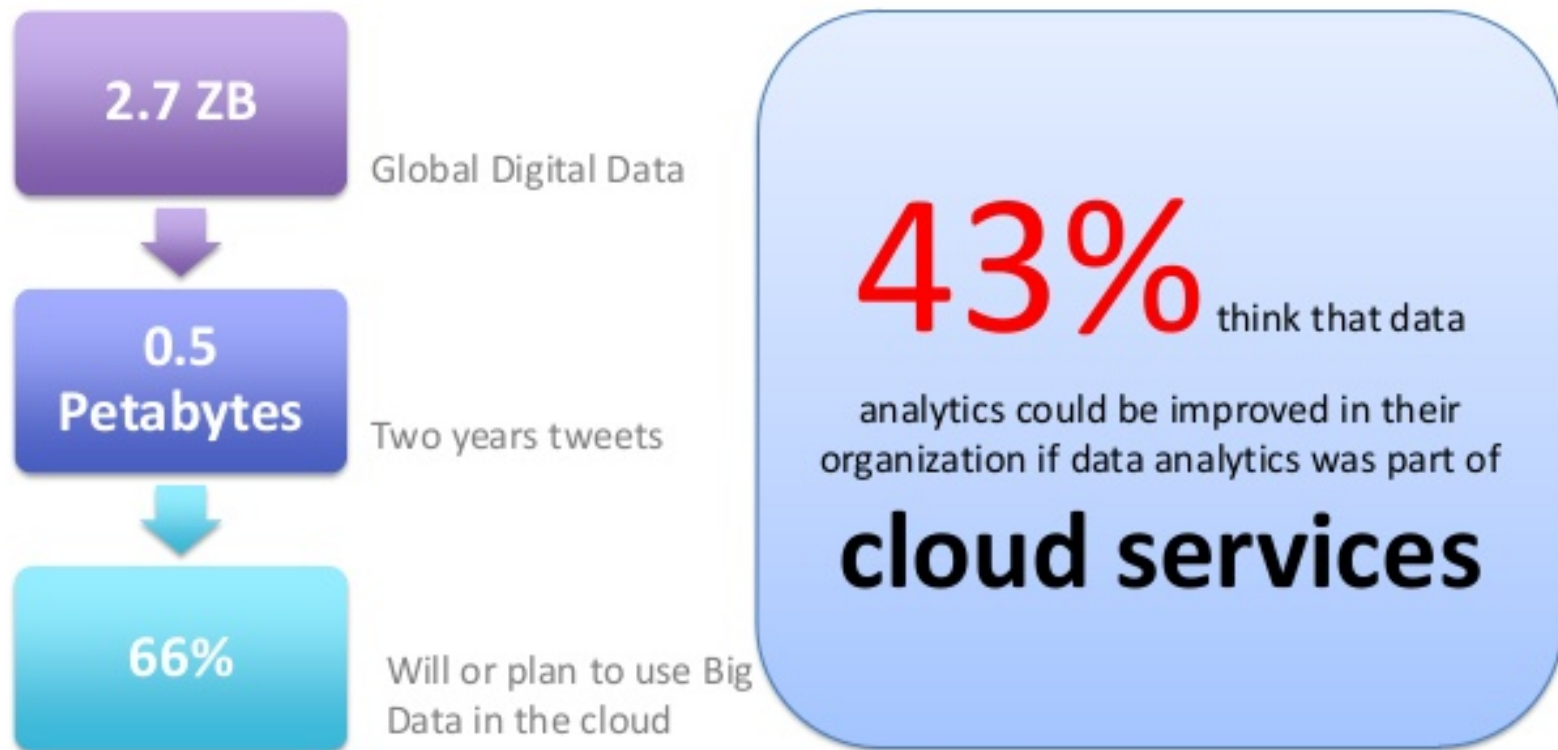
“Big data is not a stand-alone technology; rather, it is a combination of the last 50 years of technology evolution”

Alan Nugent

- Wave 1: Creating manageable data structures
- Wave 2: Web and content management
- Wave 3: Managing big data



Big Data In The Cloud



1 Zettabyte= 10^{21} bytes , 1Petabyte= 10^{15} bytes

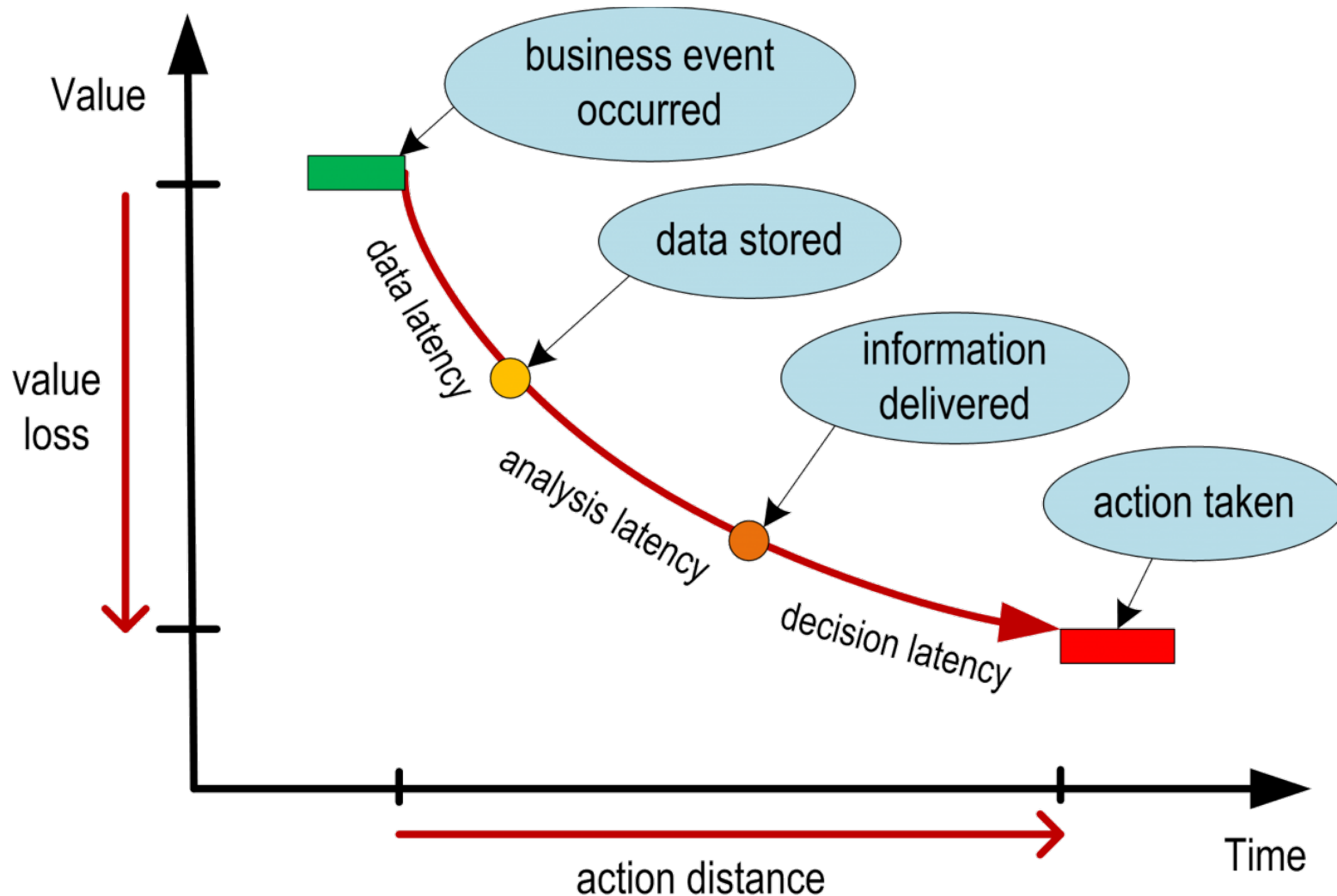
Source: <http://www.slideshare.net/>

DRIVERS FOR BIG DATA ON CLOUD ADOPTION

- Cost reduction
 - Managing cloud-based big data is cost-effective, scalable and fast to build.
- Rapid provisioning and fast time to market
 - Faster provisioning is important for big data applications since the value of data reduces quickly as time goes by.
- Flexibility/scalability
 - Big data analysis, especially in the life sciences industry, requires huge compute power for a brief amount of time. For this type of analysis, servers need to be provisioned in minutes.

TIME REALLY IS MONEY:

THE TIME VALUE CURVE



THE CHALLENGES OF BIG DATA

- They are growing
- They are deeply correlated
- They are expected to be used in a way with a low cost/performance ratio

The Challenge



THREE REASONS FOR BIG DATA IN THE CLOUD

- Skills (e.g. big data analytics)

Do you really need all this in your company?

- Vast amounts of data

Is it better to keep and manage all this data outside your company to the cloud?

- Value and management of data

Do you prefer focusing on the value of your data rather than managing them?

Managing Big Data on the Cloud



- Auto start VMs
- Install and configure app components
- Monitor
- Repair
- (Auto) Scale
- Burst...

DO WE REALLY NEED CLOUD FOR BIG DATA?

Before we begin, we need to break some common myth's on portability, a key feature that Cloud well provides.

Cloud Portability Myth #1

No one really needs
cloud portability

Cloud Portability Facts



Zynga moved ~80% of
their workload from
Amazon to their private
zCloud

*“own the base, rent the
spike”*

CLOUD PORTABILITY FACTS

- You really need the flexibility to choose
 - What is right for you
 - When it is right for you
- All your choices are based on the price, features, availability, performance, etc.

Cloud
Portability
Myth #2

Cloud Portability
==
Cloud API
Standardization

Cloud APIs, Today

Standard APIs

Not practical in the
foreseeable future

OSS Projects

Need a couple more
years to converge & mature

Abstraction Frameworks

Probably the only practical
(near-term) option

Realization:

What You
Really Care
about Is
**App
Portability**

OS is the same on any cloud

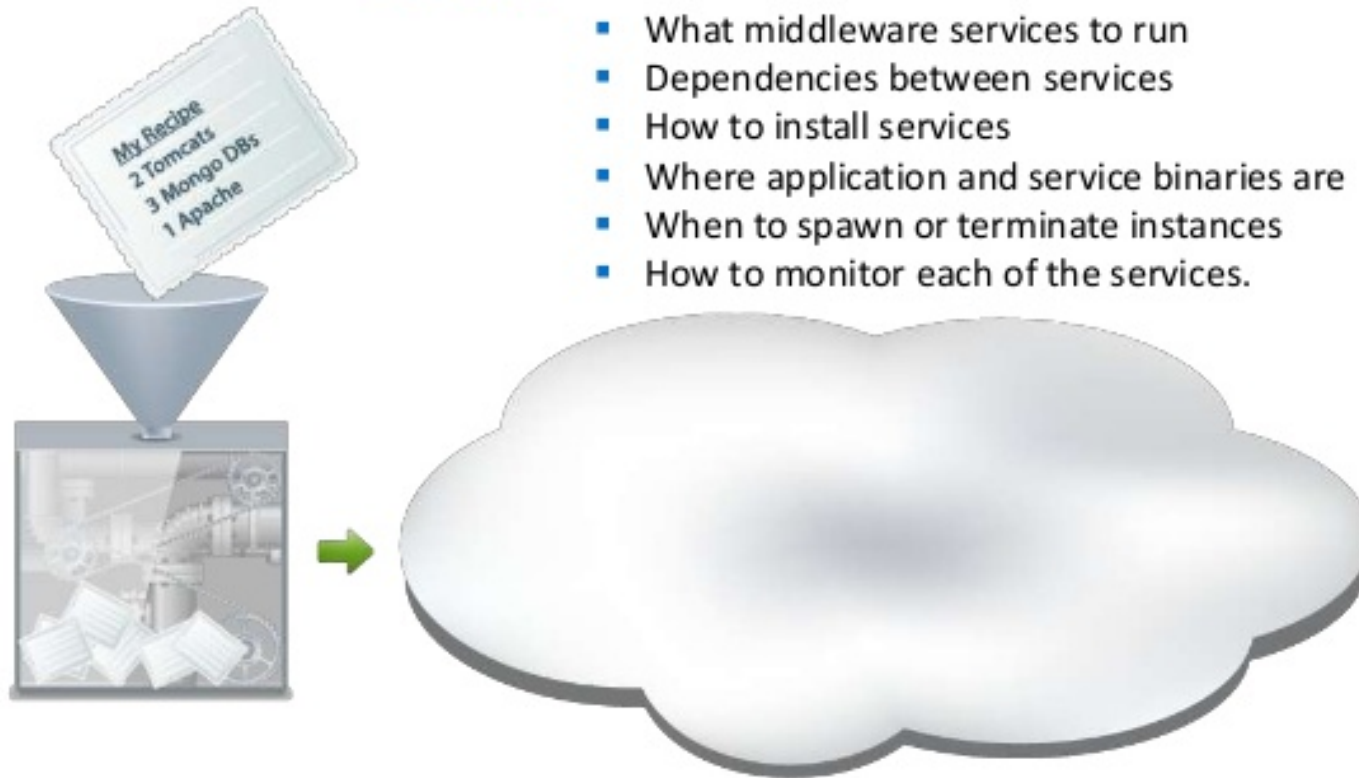
Most clouds have compute & storage

Elasticity & scaling have same effects on the app, regardless of the cloud

Consistent Management

Recipes consistent description for running **any app**:

- What middleware services to run
- Dependencies between services
- How to install services
- Where application and service binaries are
- When to spawn or terminate instances
- How to monitor each of the services.



© Copyright 2012 GigaSpaces Ltd. All Rights Reserved

The Right Cloud for the Job (Cloud Portability)



Automation across the stack





Big Data Apps,
on Any Cloud,
Your Way

Open source
(Apache2)

How is Big Data Different?

Traditional Data

- Large scale
- Highly centralized
- Structured
 - Files
 - Records
 - Databases
- Sequential
- Indexed
- **Processing transactions**

Big Data

- Massive scale
- Highly distributed
- Unstructured
 - Emails
 - Audio/Video
 - Documents
 - Spreadsheets
 - Log files
 - Sensor data
 - Geo-spatial data
 - Books
 - Journals
 - Blogs
 - Text messages
 - Chat sessions
 - Search data
- Random
- **Looking for patterns and relationships**

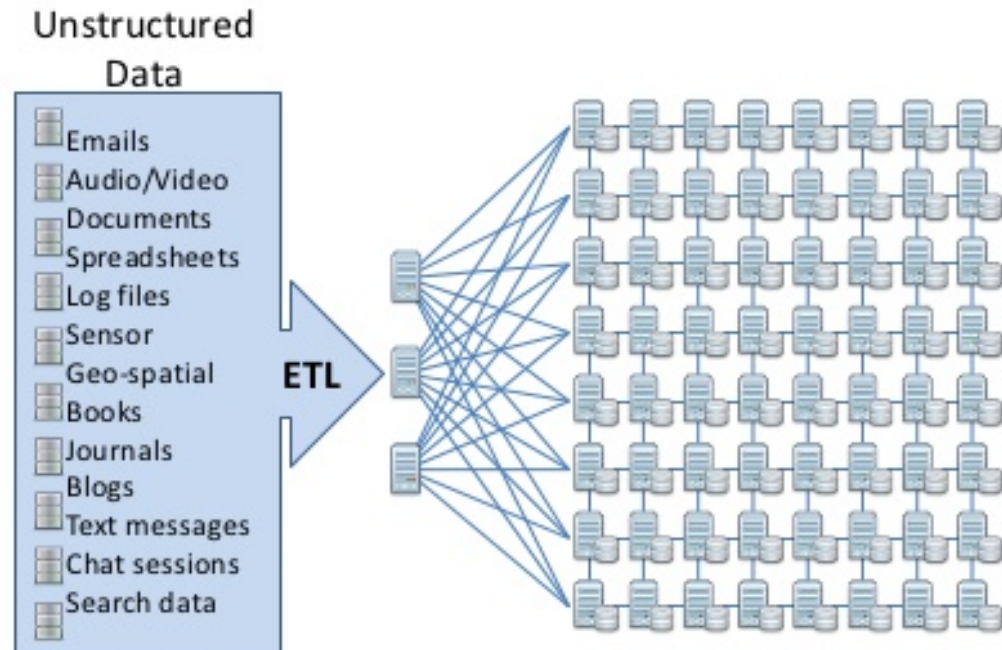
Dealing with Structured Data



- Very large databases stored in a central location – data warehouse
- Attached to very large, very powerful computers
- Accessed by structured queries
- Continually updated
- Used for “real time” transactional processing
- Reports created by a “batch” process

Dealing with Big Data

- Unstructured data retrieved from variety of sources
- Data is **Extracted, Translated and Loaded** into Big Data system
- Small parts of data are distributed by master nodes to hundreds (thousands) of small networked nodes
- Each node processes a part of the data and returns an answer
 - MapReduce
- Process is repeated until all data is analyzed
- Results are then used for further analysis



The Future

Cloud Computing + Big Data

- Stream computing
- Dramatically improved forecasting and predictive analysis across all scientific disciplines
- The rise of the Social Graph
 - Battle lines are drawn
- Individually tailored and personalized solutions, services and experiences
 - Medical diagnosis and treatment
 - Lifestyle management
 - Targeted marketing and advertising



Q & A