Web Computing for Information Island Crisis in the Era of Big Data

数据孤岛的Web开放之道

Gang Huang
Peking University
2016.10.20, Taiyuan, China
Agenda

- Information Island Crisis in the Era of Big Data
- Web Computing Paradigm as a Silver Bullet
- 10 Years Research on Web Computing
- Future of Web Computing for Big Data
Data as a Resource

Big Data is generated by billions of Information Systems

“Surface” Data from World Wide Web
- Data can be retrieved by standard web crawlers or search engines such as Google, Baidu, Bing, etc.
- Till June 2016, 4.5+ million web sites with 200+ billion pages

“Deep” Data from Service-Oriented Web
- Source: enterprise/organization information systems, business systems like Amazon, Ctrip, CRM, SCM, and zillions of desktop/mobile apps
- Such data is dynamically generated with the service interaction, but CANNOT be accessed via crawler!
- Volume: 10x-100x compared to surface data (excluding video/audio)
- Value: pretty higher than surface data
Deep Data Collection

Surface data collection is the core competence of WWW
Deep data collection is the core competence of Big Data

In 2012, Google announced the “In-App Search” for deep data exploration

In 2015, Apple iOS 9 supports deep data search for Apple APPs and cached data search of other APPs.
Information Island Crisis in the Era of Big Data

The In-App Search can support only 1000+ apps.

The iOS supports only local cache of third-party APPs.

Collecting data from 100,000 e-Gov Systems:

- 50,000,000 Man-Months
- 100,000,000,000 RMB

* from Digital China, Neusoft, Taiji, CS&S, etc.
Silver Bullet to Information Islands

- Specific or ad-hoc solutions for different levels and scenarios.
- Typically include DB exporter/importer, crawler, refactoring,
- Heavily depend on the application infrastructure, e.g., hardware, OS, security policies.
- High difficulty, risk, cost, labor-based, error-prone.

Web Computing for Big Data - Gang Huang
Agenda

• Information Island Crisis in the Era of Big Data
• Web Computing Paradigm as a Silver Bullet
• 10 Years Research on Web Computing
• Future of Web Computing for Big Data
YanCloud for Data as a Service

Desktop/Web/Mobile Application Systems

Data API Learning and Construction Platform

- General at memory level
- Read, but can write back
- Real-time data manipulation
- WYSWYG data visualization

Data API Store

- Domain-Specific API
- API production, consumption

Data API Runtime and Management Platform

- Data Catalog
- API composition
- Online deployment and evolution
- Data accounting

- The unique PRODUCT supports deep data collection of Web/PC/App
YanCloud Applications on Smart City

315 data APIs for 121 systems from 43 cities

From 60 man-months
To 1 man-day
YanCloud Applications on Data Collection

Tax Management Systems

From impossible
To 5 man-days using YanCloud

HR Management System

From unsolvable
To 3 man-days using YanCloud
YanCloud Applications on Data Collection

500+ Systems in 20+ Provinces and Ministries across China in 2016

Engineering Efficiency \( \uparrow 100\times \) Labor Cost \( \downarrow 90\% \)

Sharing and Crowdsourcing of data, algorithms, applications and stakeholders
YanCloud Applications on Mobilization

Generate mobile APP from legacy Visa Application System

Generate WeChat Public Account from legacy Journal Portal
DaaS Applications on Mobile Intelligence

Deep Sensing
Deep Searching
Deep Linking
Agenda

• Information Island Crisis in the Era of Big Data
• Web Computing Paradigm as a Silver Bullet
• 10 Years Research on Web Computing
• Future of Web Computing for Big Data
Our Vision on Internet Computing

Technical Trend
- Semantic Web
- Social Computing
- Service Computing
- System of Systems
- Pervasive Computing
- Grid/Cloud Computing
- Internet of Things

Big Trend
- Internet as a Computer

Business Trend
- Digital Economy
- E-government
- Internet Culture
- Social Network
- Modern Service
- Virtual World
- Smarter Planet

• Grid/Cloud computing proposes a new model of networked applications from the perspective of resource sharing and management.
• Pervasive computing discusses a new situation of networked applications from the perspective of human computer interaction.
• Service Oriented Computing focuses on a new form of software with emphasis on collaboration and dynamism from the philosophy of software as a service.
• …
“Internet Computer” requires substantial improvements in software characteristics for implementing new business naturally with new technology.

Internetware: A New Software Paradigm for Internet Computing, IEEE Computer 2012
IBM GTO (Global Technology Outlook) 2012
Web Pages as Web Services

Service mashup is a data flow integrating multiple interactive web services

Q 1: Very few service mashup components?
A: Any Web page can become a web mashup component if we break the security mechanisms of standard web pages, i.e., sandbox.

Silver Bullet Part 1: We control the web pages for opening information island!

CyberC 2009 Best Paper
IEEE Transactions on Services Computing 2009
In-Depth Analysis on Services Mashups

Q 2: Web browser controls the behavior of web pages?
A: We analyze the source code of web browser and model checks its runtime behavior for understanding the whole browser-based service mashups
Data Cache for Services Mashup

### Desired cache strategy

<table>
<thead>
<tr>
<th>Context</th>
<th>Desired cache strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>The same as from Google.</td>
</tr>
<tr>
<td>B</td>
<td>Frequency: cache data does not expire within five minutes from the last response.</td>
</tr>
<tr>
<td>C</td>
<td>Granularity: cache should be done on fine-grained structures within the responses.</td>
</tr>
</tbody>
</table>

#### Google Weather’s cache strategy:
Cache expires immediately.

A uses the weather data in a real time application.

B uses the weather data to feed other services which care less about the accuracy.

C only needs today’s weather from the responses, which varies less frequent.

**Q 3: Standard Browser/Server interactions unfit service mashup?**

A : We control the cache strategies of HTTP and HTML.

Silver Bullet Part 3: We control the interaction between web browser and web server for opening information island!

SOCA 2010 Best Paper Nomination, WWW 2015

Web Computing for Big Data - Gang Huang
Offloading Javascript Programs

• Rich Web mashups cannot work well on mobile devices
  • Chess games, 3D Graphics, RPGs
• Mobile Web can leverage the cloud-side resources

Q 4: Javascript make web pages much more complex to understand and control?
A : We make the Javascript programs offloaded from mobile browser to cloud.

Generally applied to major browsers Chrome, Safari, and FireFox

49x page load time improvement
92% Energy saving

All-in-One by SM@RT

SM@RT Client-Cloud-Convergence Platform

SM@RT SMVC Model

Science China 2013 & IEEE Transactions on Services Computing 2016
SM@RT for Java-based Information Islands

Re-implement the silver bullet for opening Java-based information island

- Java bytecode
- Java VM
- Java Invocation
- VM in Cloud

Runtime model of an offloaded Android app

97% execution time and 83% energy saving

OOPSLA 2012
Our Silver Bullet to Information Island Crisis

Service Oriented Software Architecture

Data and Service Innovation
- UE Optimization
- Process Optimization
- Service Integration
- Data Collection

Micro Service
- View
- Controller
- Model

Recovery and Refactoring

SM@RT

Networked Software Architecture
- Apps/Server, Client/Server, Browser/Server

HTML/CSS
- Javascript
- Java bytecode
- Assembly

Code/Data Analytics
- Browser
- JDK/JVM
- GUI Widget
- HTTP Stack
- Android/Linux

Dev Framework

Presentation

Business

Data

Middleware

Host OS

Network
Summary of Our Web Computing for Big Data

• The ONLY Silver Bullet for Web/Desktop/Mobile Information Islands

• 500+ Government and Enterprise Applications

• 100X Engineering Efficiency Improvement

• 90% Labor Cost Saving

• 80,000,000 RMB Patent Royalties

• 10 years research and practice
Agenda

• Information Island Crisis in the Era of Big Data
• Web Computing Paradigm as a Silver Bullet
• 10 Years Research on Web Computing
• Future of Web Computing for Big Data
Intra-Organization Deep Data Sharing

Palantir: real-time inspection of tens of government system
20 billion $ assessed value

DOMO: Real-time collection of hundreds of EIS and BI support
2 billion $ Assessed Value
API Economy for Big Data

**Intra-Organization Deep Data**

Palantir: real-time inspection of tens of government system, 20 billion $ assessed value

DOMO: Real-time collection of hundreds of EIS and BI support, 2000M $ Assessed Value

**Inter-Organization Deep Data**

API-based data trading (10+ billions of RMB market)

**Situational Deep Data**

API Economy for situational applications
- 5 Billion API request of Google and Facebook
- 3 Billion API request of Twitter (75% of total traffic)
- 25+ billion USD market (Gartner)
API Economy by Web Computing

Web 1.0 (HTML+HTTP) → Web 2.0 (REST+XML) → Web 3.0 (Semantics)

Web 1.0 (HTML+HTTP) → Web 2.0 (REST+API) → Web 3.0 (Big Data)

HTML v.s. API/Data Spec

Web Search v.s. API/Data Search

HTTP/SSL for Web Pages v.s. HTTP/Block-chain for Data

RESTful v.s. Micro-Services

API Specification (Data)

API Management (Data)

API Invocation (Data)

API Consumption (Data)
Thanks

Web Computing for Information Island Crisis in the Era of Big Data

Gang Huang, Peking University, hg@pku.edu.cn

数据孤岛的Web开放之道 黄罡 北京大学