Big Data
Technological Challenges

François Bourdoncle
Co-founder & CTO
3DS Exalead
Sept. 25th 2012
Big Data

Context & Main Challenges
Why Big Data? Why Now?

▶ More data available now
  ▶ HTML Web pages
    ▶ 32 billion pages = 1 Peta-bytes = 1,000 Tera-bytes = 1 million Giga-bytes
  ▶ Social networks (Facebook, Twitter)
  ▶ Transaction records
  ▶ Application logs
  ▶ Internet of Things (including smartphones)

▶ More processing power
  ▶ Cloud Computing
  ▶ Google: 1 million servers
  ▶ Distributed computing frameworks (e.g., Hadoop Map/Reduce)
  ▶ IBM « Watson »
    ▶ Beats humans at « Jeopardy »
    ▶ Reasoning on Wikipedia contents
How Big is “Big”? 

- **Two approaches**
  - When the volume of data becomes larger than usual
    - Tera, Peta (Web), Exa, Zetta
  - Or when the problem becomes untractable using traditional tools & techniques
    - The end of relational databases supremacy
    - The new NoSQL trend (No-SQL / Not-only SQL)

- **Moore’s law**
  - The volume of data grows much faster than our ability to process it
  - Exponential laws
  - Difference in parameters
  - Massively parallel processing
Big Data and “Emergence”

- The business approach
  - More hardware, new software, new services, etc.
  - Very limited way of looking at the problem

- A more philosophical approach
  - Self-organization from noise
  - Ilya Prigogine, Isabelle Stengers, Francisco Varela, etc.

- The scientific approach
  - Physics laws and emergence
  - The end of reductionism?

- The end of social sciences?
  - The end of the normative and inductive approaches?
  - The new study of emerging social laws (with the proper tools)
Exalead

Some background
Exalead at a Glance

- **European leader of Enterprise Search platforms**
  - 1/3 customers are « online » customers
  - 2/3 customers are « entreprise » customers
- Created 12 years ago
  - By AltaVista veterans
  - Acquired by Dassault Systèmes in June 2010
- 170 employees worldwide
Exalead Application Domain

- **Software vendor (B2B)**
  - Infrastructure software for unified information access
  - Licenses, maintenance & services

- **Next-generation business applications**
  - Beyond classical « Enterprise Search »
  - Search-Based Applications (SBA)
  - CRM, BI, logistics, etc.
  - Machine-generated & social data

- **Large online applications**
  - Skyrock, 118218.fr, ViaMichelin, Lagardère, Rightmove, etc.
  - 110 million unique visitors./month

- **www.exalead.com**
  - 16 billion Web pages (soon to be 32 billion)
  - 1 million unique visitors/month
B2C Applications

Restminer

Pure Food And Wine
54 Irving Pl, New York, NY 10003, USA
Phone: (212) 477-1010
Subway: A, C, E, L, N, Q, R, W at 14th St-Union Sq.
Website: http://www.purefoodandwine.com - Book Online!
Kind of food: Vegetarian, Health Food, Vegetarian/Vegan

Overview
Review

Pure food and wine serves upscale raw vegan cuisine - food in its natural and purest state, in order to preserve the food’s natural enzymes, vitamins and minerals, nothing is heated above 118 degrees. Only the freshest seasonal, organic ingredients are used to produce handcrafted flavors that rejuvenate the mind, body and spirit. An extensive collection of organic and biodynamic wines and sakes are available.

Most viewed comment

Having hesitated for months to try a raw dining experience, we finally made our way to PFW, with hesitant optimism. What we found was a restaurant with a pleasant atmosphere, warm staff, attentive service, and tasteful food delicately prepared. The cocktails were fun—a blackberry mojito-style, and a glass of vegan Pinot Grigio—and generously large. An appetizer of the spicy Thai Lettuce Wraps was refreshing and stimulating. And entrees of the Tamales and Ravioli were bursting with flavor and beautifully served. Light but filling, both main dishes boast fillings and sauces that... More >>

Bloggers entries about Pure Food And Wine

- Raw-Food Guru Matthew Kenney Tries His Luck in Orlando (Grub Street)

Details
Payment options: AMEX, Diners Club, MasterCard, Visa
Chef: Matthew Kenney and Sarma Melngals
Price range: Moderate, Expensive
Dining style: Casual Elegant

Special features: Bar/Lounge, Patio/Outdoor Dining, Personal wines welcome (corkage fee applies), Takeout, Good for Groups, Outdoor Dining, Take-Out, Outdoor Dining, Tasting Menu, Online Reservations

Opening Hours

<table>
<thead>
<tr>
<th>Day</th>
<th>Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td>12pm - 3pm</td>
</tr>
<tr>
<td>Tuesday</td>
<td>12pm - 3pm</td>
</tr>
<tr>
<td>Wednesday</td>
<td>12pm - 3pm</td>
</tr>
<tr>
<td>Thursday</td>
<td>12pm - 3pm</td>
</tr>
<tr>
<td>Friday</td>
<td>12pm - 3pm</td>
</tr>
<tr>
<td>Saturday</td>
<td>5:30pm - 12am</td>
</tr>
<tr>
<td>Sunday</td>
<td>5:30pm - 11pm</td>
</tr>
</tbody>
</table>

People sentiments

exceptional/extraordinary/favorable/disappointing/bad
flanking/inspiring/refreshing/disappointing/cheap
pleasent/disappointing/unpleasant/responsible/unbelievable

spicy/healthy/vegetarian/locally grown/sustainable

From OpenTable:

From New York Magazine:

good/favourable/high

great

southeast/nice/cheap/cheering/healthy

greenery/fresh/vegetarian/locally grown/unsustainable

interesting/attentive/excellent/acceptable/decorated

hiding/demanding/big/comfortable/delicious/beautiful
advanced/gorgeous/high-class/exclusive/outstanding
fantastic/courteous/undiscovered/impeccable

DASSAULT SYSTEMES

IF WE ask the right questions, we can change the world.
Market Monitoring
Search-Based Applications

Legacy Approach
High complexity/costs, Low performance/reusability

SBA Approach
Low complexity/costs, High performance/reusability
Some Exalead Big Data Applications

- **Anti-piracy** applications
  - Software fraud (Dassault Systèmes)
  - Fiscal fraud (EQUITALIA)
- **Automotive** industry
« Big Data » Key Technologies
Storage and Indexing Technologies

- A new way to **structure and store information**
  - Semi-structured nodes (free text, business items)
  - Relations (semantic, business, etc.)

- A new type of **software infrastructure**
  - Massively distributed infrastructure
  - (Quasi-)linear scaling
  - Hadoop Map/Reduce
  - Analytics and aggregation capabilities

- As « **ACID » as possible**
  - **ATOMICITY**
  - **CONSISTANCY**
  - **ISOLATION**
  - **DURABILITY**
Analytics and Discovery Technologies

- **Text-mining**
  - Extraction of noun groups
  - Extraction of semantic relations

- **Graph-mining**
  - Isolation of « clusters »
  - Computation of « hubs » & « authorities »
  - Computation of « Page rank »
  - The Web graph looks like a bowtie

- **Machine learning**
  - Core technology for this domain
  - Neural networks, hidden markov models, conditional random fields, etc.
  - Numerous and diverse applications

- **Data vzialisation**
Other Technologies

- **Natural Language Processing**
  - Speech-to-text
  - Machine translation

- **Image processing**

- **Knowledge representation**
  - Ontologies
    - Taxonomies (RDF)
    - Reasoning (OWL)
  - Ontology alignement
  - Ontology construction
    - Supervised
    - Non-supervised
A Few Challenges for the Future

- **Semantic Web**
  - **Normative approach** (Sir Tim Berners-Lee)
    - Meaning is defined by the “author”
  - **Emerging approach**
    - Meaning is defined by the “reader”

- **Graph analysis**
  - Facebook (social graph)
  - Business or design relationships (e.g., 3D design)

- **Scaling**
  - $O(n)$, or $O(n \log n)$ algorithms

- **Reasoning** (logical inferences, Datalog-style)
  - Example: WATSON d’IBM
  - Facts inferred from DBPedia (~ Wikipedia)
Reasoning

Example of inference rules

-- Define the notion “follows” of being a direct or indirect follower of someone

- $X$ follows $Y$ WHEN $X$ isFollowerOf $Y$
- $X$ follows $Y$ WHEN $X$ isFollowerOf $Z$ AND $Z$ follows $Y$

-- Define the notion of being a grand-parent from that of being the son or daughter of someone

- $X$ isChildOf $Y$ WHEN $X$ isSonOf $Y$
- $X$ isChildOf $Y$ WHEN $X$ isDaughterOf $Y$

- $X$ isGrandParentOf $Y$ WHEN $Y$ isChildOf $Z$ AND $Z$ isChildOf $X$