



Turning Web X.0 Data into Competitive Advantage

Tyrone Grandison PhD, Daniel Gruhl PhD
IBM Almaden Research Center
Silicon Valley, California

The Roadmap

- Background
 - People
 - Web 1.0 to the world
 - The Problem Space
- Project Description
 - Hardware Overview
 - Software Architecture
- Issues Encountered
- Pilot Feedback
- Current Status
- Contributions to the Field of Computer Science
- Lessons for Developing Countries

Take-Away Message

Targeting low-margin support services opportunities is not a sustainable course of action for the developing countries.

A diversified portfolio that focuses on high cognitive ICT services offered to Fortune 500 companies is within reach and very profitable.

People

Main Team – Intelligent Information Systems (IIS) Group

Present Family: Varun Bhagwan, Karen Brannon, Sangeeta Doraiswamy, Alexandre Evfimievski, Tyrone Grandison, Kun Liu, Evimaria Terzi

Immediate Extended Family: Alfredo Alba, Stefan Edlund, Ronald Fagin, Deon Glajchen, Niina Haiminen, Joshua Hui, Jerry Kiernan, Daniel Gruhl, Heikki Mannila, Jan Pieper, Christine Robson, Tarun Thakur, David Woodruff

Past Family Members: Rakesh Agrawal, Dmitri Asonov, Roberto Bayardo, Rafae Bhatti, Alvin Cheung, Christan Grant, Bryan Hickerson, Christopher Johnson, Theodoros Lappas, Kristen Lefevre, Karin Murthy, Ralf Rantzau, Stefan Schönauer, Ramakrishnan Srikant, Raja Velu, Niko Vuokko, Steve Watts, Yirong Xu



Work: We do fundamental and "visionary" research that closely ties with applications. The team has a solid track record of doing pioneering work: *Association Rule Mining, Hippocratic Databases, RFID Traceability, Disclosure-Compliant, Localized Mobile Device Information Sharing, Privacy-Preserving Social Network Analysis*, etc.

Group's Website: <http://www.almaden.ibm.com/cs/disciplines/iis/>

Project Team – BBC Sound Index

- **British Broadcasting Corporation:** Geoff Goodwin (Head of BBC Switch), Beth Garrod (BBC Sound Lead)
- **IBM:** Daniel Gruhl (ARC-Health Informatics), Varun Bhagwan (IIS), Alfredo Alba (Almaden Services Research), Jan Pieper (ARC-User Experience), Anna Liu (Almaden Services Research), Bill J Scott (IBM Global Business Services), Aidan Toase (IBM Global Business Services)
- **NovaRising**

Project's Website: <http://www.almaden.ibm.com/cs/projects/iis/sound/>

The Web

- Web 1.0
 - sites are static, sites tend not to be interactive, applications are normally proprietary.

- Web 2.0
 - websites build on the interactive facilities of "Web 1.0" to provide "network as platform" computing.
 - users can own the data on a site and may exercise control over that data.
 - sites are built on an "Architecture of Participation" that encourages users to add value to the application as they use it, e.g. social-networking sites, video-sharing sites, wikis, blogs, folksonomies, etc.

- Web 3.0
 - Many think Web 3.0 is going to be like having a personal assistant who knows practically everything about you and can access all the information on the Internet to answer any question.
 - Many compare Web 3.0 to a giant database. While Web 2.0 uses the Internet to make connections between people, Web 3.0 will use the Internet to make connections with information.

NB: The term Web 2.0 was coined around 2004 by business professionals. Since then Web 1.0 has been retrofitted and Web 3.0 has been on the tongues of many luminaries.

The Problem Space

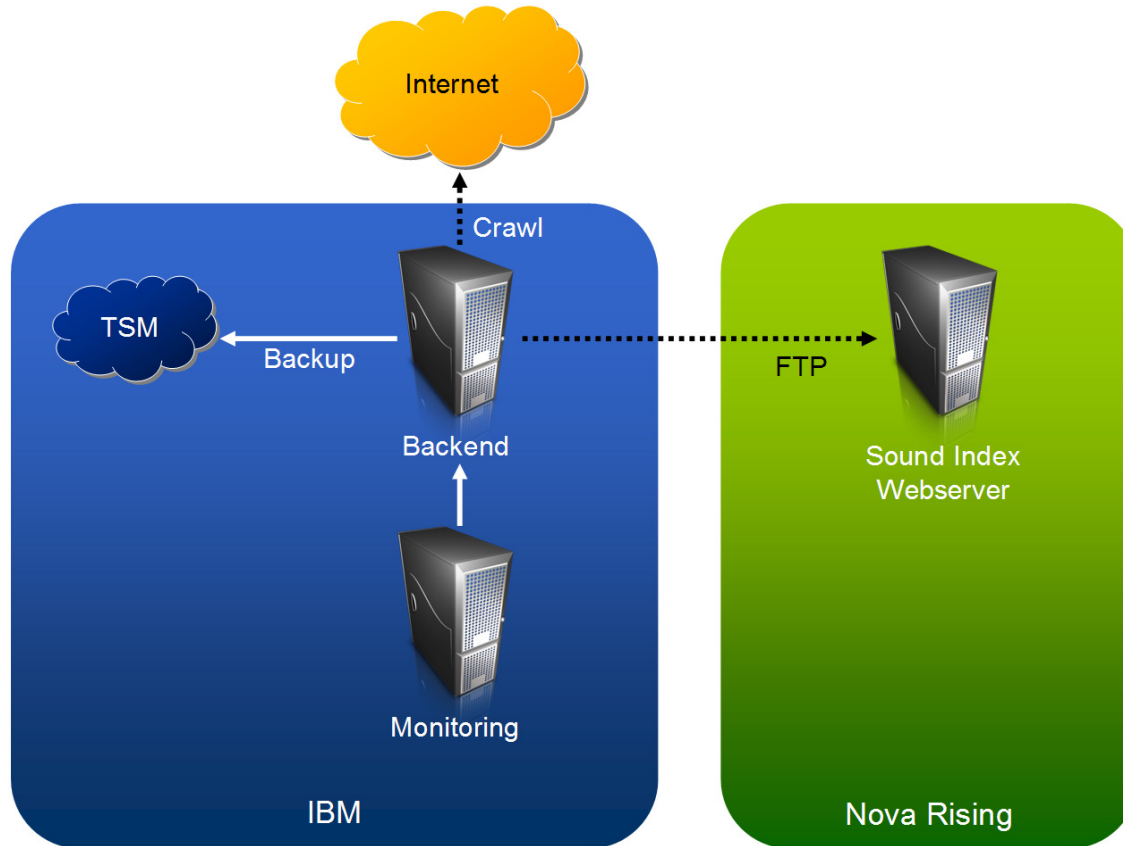
- British Broadcasting Corporation (BBC) Switch's mandate is to recapture the lost teen audience and engage with it across all platforms – Web, TV, Radio and mobile.
 - Music is viewed as one of the keys to this market. Unfortunately, the music charts do not necessarily reflect current teens' interests.
- The landscape in which they were operating was changing rapidly:
 - The core assumption has changed:
 - The internet is the new medium of preference for teens
 - 48% of teens did not buy a CD in 2007*
 - The current charts are slowly becoming irrelevant
 - For the most part, Billboard et al. are perceived as losing their touch with the pulse of the new generation
 - No measure for the *lead-up* or anticipation for an album release
- BBC Switch wanted to produce an engaging, interactive and immediate alternative to the traditional sales-based, weekly chart.

* Source: LA Times, Feb 27, 2008

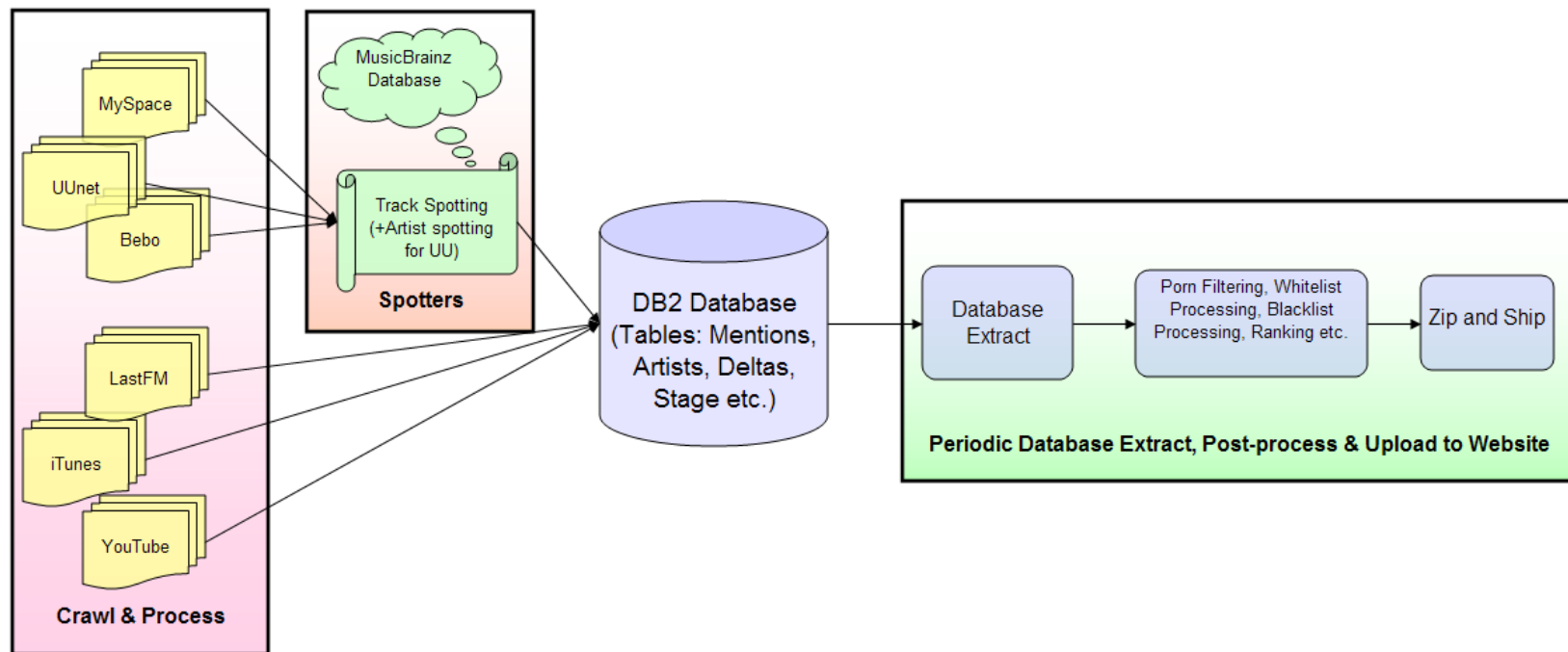
Project Description

- In late 2007, the BBC Switch division commissioned the IIS group at IBM Almaden Research Center to help them make a new platform for creating music charts, based on online buzz.
- The idea of the Sound Index was to capture the buzz around music from the major social networking and music sites on the Web and find the artists and tracks that were the most popular online, right now.
 - Which artists and tracks were people talking about, writing about, downloading and listening to?
 - Is the buzz on an artist or track positive or negative?
 - What demographic segment was feeling artist X or track Y?
- Solution Approach: Use IBM's semantic data extraction and analysis research technology (in this instance, MONGOOSE technology) to process data online in “real-time” and then send it to a presentation server for visualization.
 - Data Sources were: BBC, Bebo, LastFM, YouTube, MySpace, Google groups, iTunes
 - Images were retrieved from Amazon.com
 - Domain knowledge verified against MusicBrainz

Hardware Setup



Software Architecture



Issues Encountered

■ Noise versus Freshness

- There is a tension between the desire for rapid and frequent updates reflecting the very cutting edge of what is hot, and minimising the influence of noise in the charts due to short term spikes. Striking a balance here poses an interesting challenge.
- Effects such as weekends, nights and holidays need to be weighed against events such as new album releases, celebrity gossip events and award shows.
- Any such system will ultimately be a compromise between being too sensitive and not reactive enough and optimising this balance is a difficult research challenge.
- To solve this issue, we used a 24 hour window (that is 4 6-hour cycle periods) to smooth out some effects.

■ Spam and Off-topic Detection

- The tremendous popularity enjoyed by websites such as MySpace and YouTube also attracts undesirable attention. Spammers and other commercial sites regularly attempt to peddle their content via these sites, by masquerading as “bands” or “users”. This poses a challenge that has two distinct flavours.
- The first one is the ability to distinguish valid artists from those that are nearly product spam. A subject matter topical dictionary enables a first pass at this, as does a list of fairly common “spam” phrases, but the ultimate editorial adjudication at this point is subject matter expert driven.
- The second is the ability to filter spam and profanities.

■ Business Intelligence

- The front-end analysis engine had to be re-built several times utilising different BI environments before it was realised that existing BI technologies couldn't cope with our requirements for a mass-user online audience.
- We went back to the core theory behind BI and in there lay the reasons it wouldn't work in what is effectively still a primitive online environment.
- We hence built our own engine which refined how searches across the data cube are undertaken and pays more credence to optimising these for web technologies.

Pilot Feedback

- The pilot was more successful than expected
 - Without any marketing or promotion, the Sound Index went from a standing start in April 2008 to deliver 43,469 visits from 37,900 unique users.
 - In June 2008, there were 140,383 page views at an average of 3.67 per user. Each user spent an average time of 3 minutes 40 seconds on the site (53 seconds per page).
 - In August 2008, there were over 772,000 web page references to the Sound Index.
- The Sound Index was described as “the first definitive music chart for the internet age”⁺.
- The system was named Web 2.0 technology of the week by the UK Observer for several consecutive weeks and has been named the hottest thing in music (in March 2008) by the UK Guardian Music Monthly.
- There was a lot of positive comment from the web and from the traditional press.
 - The Sound Index generated a lot of debate about what constitutes popularity and how the results should be viewed.

⁺Chris Salmon. Click To Download. Guardian UK, <http://arts.guardian.co.uk/filmandmusic/story/0,,2274132,00.html>

Current Status



- The pilot has now closed and the results are under evaluation. The engine is applicable beyond the music domain, e.g. the television, politics or film domain.
- MONGOOSE technology is being used to build systems for other domains.

Live Demo Website <http://bbcindex.novarising.com/>

Contributions to the Field

- The Sound Index serves as a model for the new era in business innovation.
 - Demonstrates the next wave in delivering better services and products – the real-time integration of multiple, relevant online information with one’s own data to drive new and significant value for, re-invigorate connection to and strengthen brand affinity to one’s customer base.
 - The Sound Index is the first instance of a consumer-focused, Business Intelligence (BI) system delivered to the masses interested in music.

- The core innovations in the data extraction and synthesis technology were:
 - Creation of novel techniques to parse Broken English.
 - Holistic Disambiguation – allows Web comments like “U R 50 bad”, “the guitarist killed last night”, “you are sh*t”, “you are the sh*t” and “pink was off the chain” to be transformed into their intended equivalents,
 - Integration of pieces of information of different modalities
 - Given that we were collecting information from many sources, with different populations, it was necessary to create a balanced and tunable ranking algorithm.
 - A new Business Intelligence (BI) engine was required to model the data.
 - The site rebuilt itself every 6 hours based upon the new data ingested from the MONGOOSE platform.
 - There was an 'early-warning' capability to allow fast detection (and resolution) of changes on the websites.

Lessons for Developing Countries

- The Sound Index represents the first of its kind and showcases how enterprises can embrace the Internet to enhance their businesses.
- More importantly, it presents an opportunity to innovate and lead in an emerging field for whomever is willing to explore.
- The project shows how the power of harnessing Web X.0 data, which is freely available and becoming a critical resource for companies.
 - This data can be captured by anyone, harnessed for any particular application domain and then packaged and sold to corporations and governments.
- The mining of Web X.0 data is still an emerging field with industry and academic leaders just starting to explore the space.
 - There is room for all interested parties to enter into this arena.
 - The barrier to entry is currently quite low.
 - 1) skilled ICT practitioners & researchers with knowledge of text analytics techniques,
 - 2) cheap bandwidth, 3) COTS hardware

More Lessons

- Thus, there is a tremendous opportunity for developing countries to become IT Research Centers of Competency for Advanced Analytics.
- Investment in this course action involves:
 - Determining the sweet spot
 - Providing IT practitioners & researchers with the right skills
 - Creating an environment where they can create and innovate
 - Working with industrial partners with real problems. Preferably partners in crisis.
- The right ICT Research area will:
 - Be applicable across multiple domains, e.g. construction, public relations, automotive, government, healthcare, telecommunications etc.
 - Enable developing countries to be leaders in the space.
- Why is this necessary?
 - The outsourced services market, like telecommunications, are driven by the availability of cheap labor and government incentives. Most developing companies will never be lucrative or profitable options over India, China and Russia.
 - Too many eggs in one basket normally has disastrous result. Diversification is a less risky approach for continued long-term growth.
 - There is no one-size-fits-all for all developing nations. Each nation has different strengths, e.g. some small islands well-suited to Financial Services ICTs, large islands well-suited to large-scale, ground-breaking ICT Research Incubators.
- It is our hope that following a focused ICT Research track will enable a new era of innovation led by the developing world.



Questions ?

More Information on BBC Sound Index

- Flash Demo

http://www.almaden.ibm.com/cs/projects/iis/sound/BBCSoundIndex_Oct2008.swf

- Papers

- Varun Bhagwan, Tyrone Grandison, Daniel Gruhl, "**Sound Index: Music Charts By The People, For The People**". To appear in Communications of the ACM. September 2009. Vol 52, No 9.
- Alfredo Alba, Varun Bhagwan, Tyrone Grandison. "**Accessing The Deep Web: When Good Ideas Go Bad**". The Proceedings of the ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA). Nashville, Tennessee. October 2008.
- Alfredo Alba, Varun Bhagwan, Tyrone Grandison, Daniel Gruhl, Jan Pieper, "**Text Analytics and Integration of Web 2.0 Sources to Transform Media & Entertainment**". Information On Demand (IOD). Las Vegas, Nevada. October 2008.
- Alfredo Alba, Varun Bhagwan, Julia Grace, Daniel Gruhl, Kevin Haas, Meenakshi Nagarajan, Jan Pieper, Christine Robson, Nachiketa Sahoo. "**Applications of Voting Theory to Information Mashups**". The Proceedings of the Second IEEE International Conference on Semantic Computing. Santa Clara, CA, USA. Aug 2008.

- Live Demo Website <http://bbcindex.novarising.com/>