

Contact:  
Erica Topolski  
IBM Software Group  
+1-617 693 2816  
ericat@us.ibm.com

## IBM & Hadoop Media Background

May 2010

Companies are faced with a challenge today: how to get their arms around the massive amounts of data being generated by their internal business applications, their various IT systems and from external sources. But like so many things, within that challenge are the seeds of opportunity: big data gives companies big opportunities.

As massive amounts of data create significant business challenges--and opportunities, IBM has been investigating how distributed computing might address some of those needs. Hadoop, an open source Apache project, is a technology which IBM has been leveraging with clients who generate significant amounts of data--data which is not being leveraged as effectively as it could be.

From the acquisitions of Cognos and SPSS to the creation of InfoSphere Streams and a new line of global consulting services, IBM has been at the forefront of providing clients innovative analytics solutions. Today, IBM announced the next phase of its analytics story – analytics for internet-scale data.

IBM is working with leaders in the Big Data community, [Karmasphere](#) and [Cloudera](#) to expand and develop the use of Apache Hadoop for enterprises.

As part of today's news, Karmasphere announced that it will be supporting IBM's distribution of Apache Hadoop.

"IBM's involvement is an important milestone in the rapidly evolving Apache Hadoop ecosystem, and Karmasphere is pleased to support these efforts," said Martin Hall, co-founder and chief executive officer, Karmasphere. "We are also pleased to announce that Karmasphere Studio: Community Edition will soon be available for Eclipse."

Karmasphere is a business analytics software company that brings Apache Hadoop power to the desktop. Karmasphere enables companies to unlock the competitive advantages within their large datasets by providing an easy-to-use class of client-side software. Karmasphere's initial product, [Karmasphere Studio: Community Edition](#), is helping data processing developers around the world crunch massive data sets that run in public and private, on-premise, cloud-based distributed processing environments.

"Cloudera is excited to be working with IBM and partnering for success at joint customers. IBM's entry into the Apache Hadoop ecosystem is a strong endorsement of the value Hadoop brings to enterprise customers, said Michael Olson, chief executive officer, Cloudera.

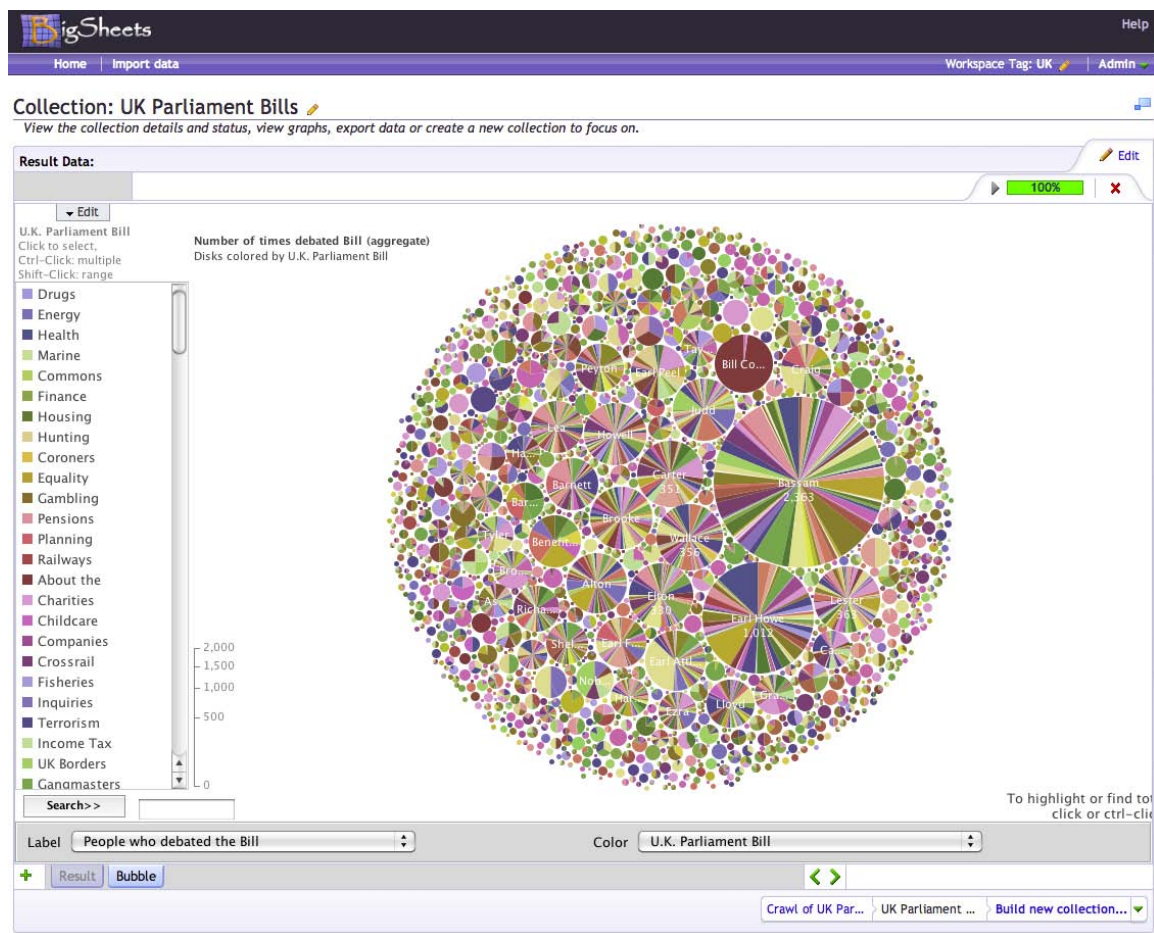
### **BigInsights**

The InfoSphere BigInsights portfolio of offerings is based off the Apache Hadoop distribution, or Apache Hadoop kernel. We refer to this as the IBM distribution of the Hadoop kernel as the

BigInsights Core, since it has additions to it above and beyond the base kernel from Apache. The offering codenamed BigInsights Core is compatible with the Apache Hadoop distribution but has been through our review and certification process and comes with installation, monitoring and configuration enhancements that are not available from Apache and includes IBM additions in querying, analytics and management. The offering codenamed BigInsights Core is available today and we use the core to help customers install, configure, deploy, manage, monitor and evolve fault tolerant and flexible Hadoop-based infrastructures.

## BigSheets

Part of IBM's new BigInsights Portfolio is BigSheets, a technology preview that helps business professionals extract, annotate and visually analyze vast amounts of Web information using a Web browser. IBM's new technology prototype is helping the British Library archive and preserve massive amounts of Web pages, and then unlock the virtual door to its archives for generations to come.



*Caption: IBM BigSheets allows analysts to sort through massive amounts of data such as the number of times a UK Parliament Bill was debated.*

BigSheets is an extension of the mashup paradigm that integrates gigabytes, terabytes, or petabytes of unstructured data from Web-based repositories; collects a wide range of unstructured Web data stemming from user-defined seed URLs; extracts and enriches that data using an unstructured information management architecture; and lets the user explore and visualize this data in specific, user-defined contexts. For example, users can see search results in a pie chart and look at the data in a tag cloud.

## **How Hadoop Helps Businesses: Storing and Exploring Data**

IDC estimates that 988 exabytes of data will be generated in 2010. That's more information than is contained in every book ever written...in fact, it's 18 million times more. The challenge of all this data is threefold:

1. Discovering the data.
2. Gathering and Storing the data.
3. Exploring/Analyzing the data.

Hadoop is a technology which can aid in the storing of the data as well as in exploring and analyzing it by allowing businesses to deploy distributed applications, running on thousands of nodes and sifting through petabytes of data.

## **What is Hadoop?**

How can business process tremendous amounts of data--and do so in an efficient and timely manner? Hadoop allows developers to create distributed applications--applications capable of running on clusters of computers. This infrastructure can then be leveraged to tackle very large data sets--by breaking up the data into "chunks" and coordinating the processing of the data out into the distributed, clustered, environment.

## **A History of Open Source**

IBM's enterprise Hadoop strategy is yet another example of IBM's commitment to the open source community. For example, IBM is the third largest contributor to Linux, a major contributor to Eclipse and significant contributor to more than 150 open source projects.

## **The Business Bottom Line**

Hadoop applications can then process your data rapidly and efficiently, .in fact, once the data has been distributed to the cluster, follow-up queries of the data can be handled efficiently since the data has already been distributed to the various nodes. The bottom line: businesses can finally get their arms around massive amounts of data, and mine that data for valuable insights, in a more efficient, optimized, and scalable way.

###

For more information:[www.ibm.com/software/data/infosphere/hadoop](http://www.ibm.com/software/data/infosphere/hadoop)