



The Next Frontier: Big Data

IBM Information Management Cloud Computing Center of Competence
IBM Canada Labs

The world is changing and becoming more...



INSTRUMENTED



INTERCONNECTED



INTELLIGENT



The resulting explosion of information creates a need for a new kind of intelligence

...to help build a Smarter Planet

05/30/2011

Template Documentation

2

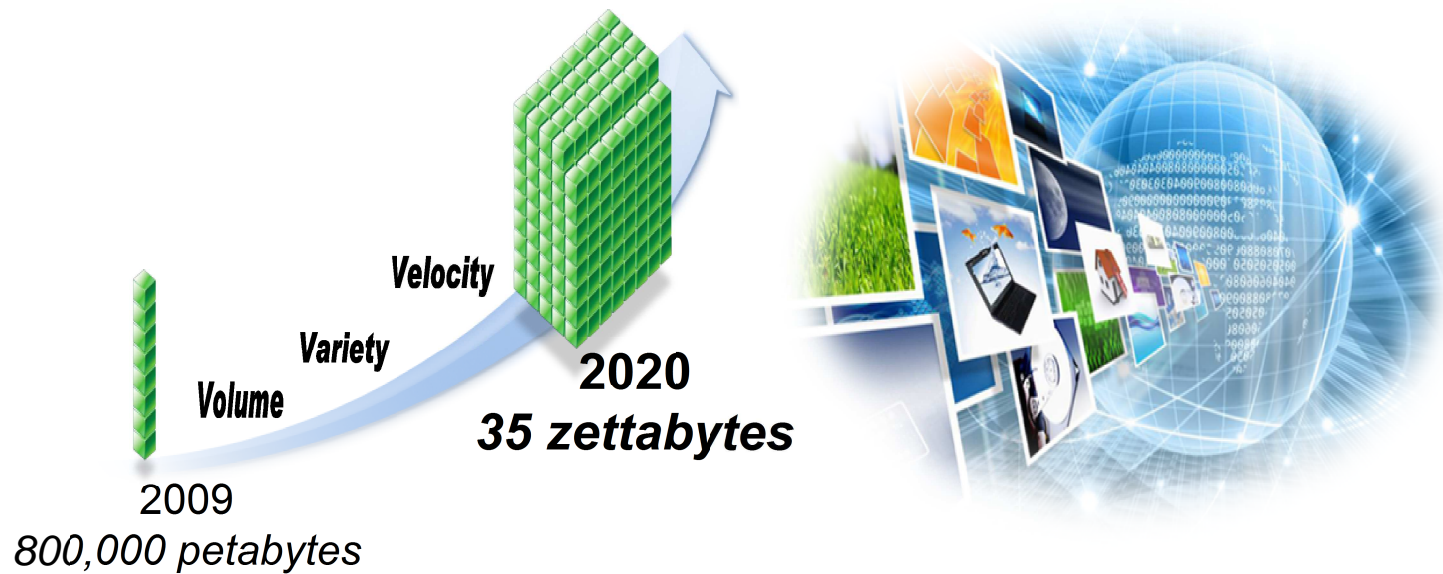
© 2011 IBM Corporation

Our world is becoming more instrumented, more interconnected and, just by virtue of those two items alone, more intelligent. We've got more data about things. You can put 32- and 64-bit microprocessor technology into lots of things these days and capture more information about physical infrastructures (with sensors), business processes, human interactions and so on. You can do things today that just were not possible to do a decade ago. The interconnect infrastructure that supports us all is extraordinary in the level of bandwidth that's available. In both instrumentation and connectivity, we're able to make things considerably more intelligent by taking advantage of mining this new wealth of information. This is really what the smarter planet idea is all about. So let's look at how we can make things smarter in the context of what can be done now in information management.

Information is growing at a phenomenal rate...

44x as much data and content
over coming decade

80% Of world's data
is unstructured



05/30/2011

3

Source: IDC, The Digital Universe Decade - Are You Ready?, May 2010

© 2011 IBM Corporation

This explosion is also characterized by massive amounts of unstructured information and that presents different challenges. We all understand the well-organized structured data world. We've dealt with it for decades. It's at the very core of what information technology and the advancement of programmable computing has brought to us over the last 60 years. We're really just at the very beginning of the unstructured space in terms of what the possibilities are, how we can get at that information and what we can do with it. Think about all the information being generated by social networking sites (Twitter, Facebook, LinkedIn, etc), web logs, click streams, instant messages, emails, electronic sensor data, and so on. How can we use that information if we could efficiently filter through that data, aggregate the right bits and pieces, combine it with existing operational data, and analyze it effectively?

IDC reference:

<http://idcdocserv.com/925>

<http://www.computer.org/portal/web/news/home/-/blogs/2613266;jsessionid=abbfded1402383e107abfa2641d6>

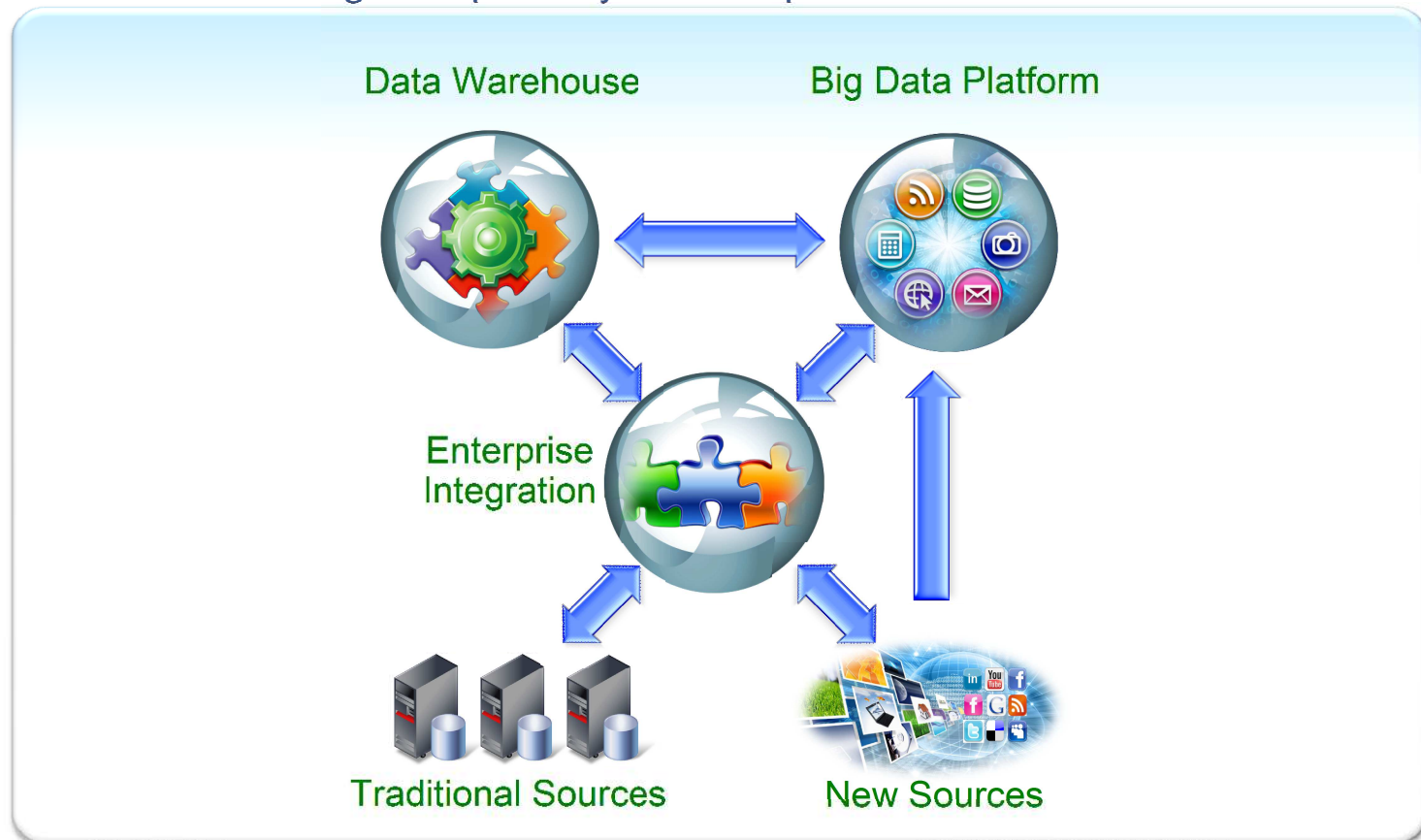
The challenge: Bring together a large volume and variety of data to find new insight



- Today organizations are only tapping in to a small fraction of the data that is available to them
- The challenge is figuring out how to analyze ALL the data, and find insights in these new and unconventional data types
- Imagine if you could analyze the 12B TB of tweets being created each day to figure out what people are saying about your products, figure out who the key influencers are within your target demographics. Can you imagine being able to mine this data to identify new market opportunities.
- What if hospitals could take the thousands of sensor readings collected every hour per patients in ICUs to identify subtle indications that the patient is becoming unwell, days earlier than is allowed by traditional techniques.
- Imagine if a green energy company could use petabytes of weather data along with massive volumes of operational data to optimize asset location and utilization, making these environmentally friendly energy sources more cost competitive with traditional sources.
- Imagine if you could make risk decisions, such as whether or not someone qualifies for a mortgage, in minutes, by analyzing many sources of data, including real-time transactional data, while the client is still on the phone or in the office.
- Imagine if law enforcement agencies could analyze audio and video feeds in real-time without human intervention to identify suspicious activity.
- As these new sources of data continue to grow in volume, variety and velocity, so too does the potential of this data to revolutionize the decision-making processes in every industry.

Big Data Shouldn't Be a Silo

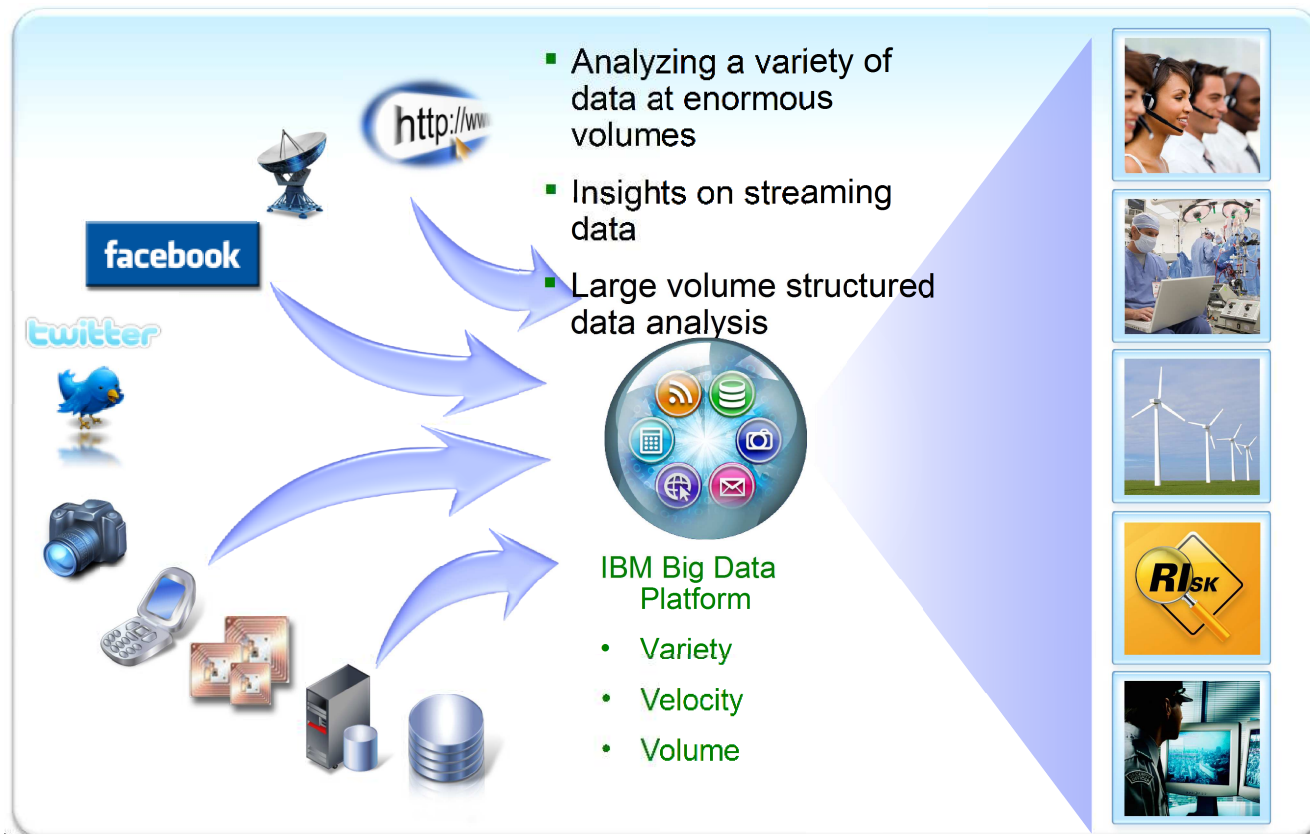
Must be an integrated part of your enterprise information architecture



- One critical component of a comprehensive Big Data strategy to mention: an effective Big Data platform and approach requires integration with the rest of your IT infrastructure.
- The last thing you need in your architecture is another technology or data silo.
- IBM's Big Data technologies should work in tandem with and extend the value of your existing data warehouse and analytics technologies.
- The value of traditional and non-traditional data sources, and traditional and non-traditional technologies increases in value when they are brought together.

The Solution – IBM's Big Data Platform

Bring together any data source, at any velocity, to generate insight



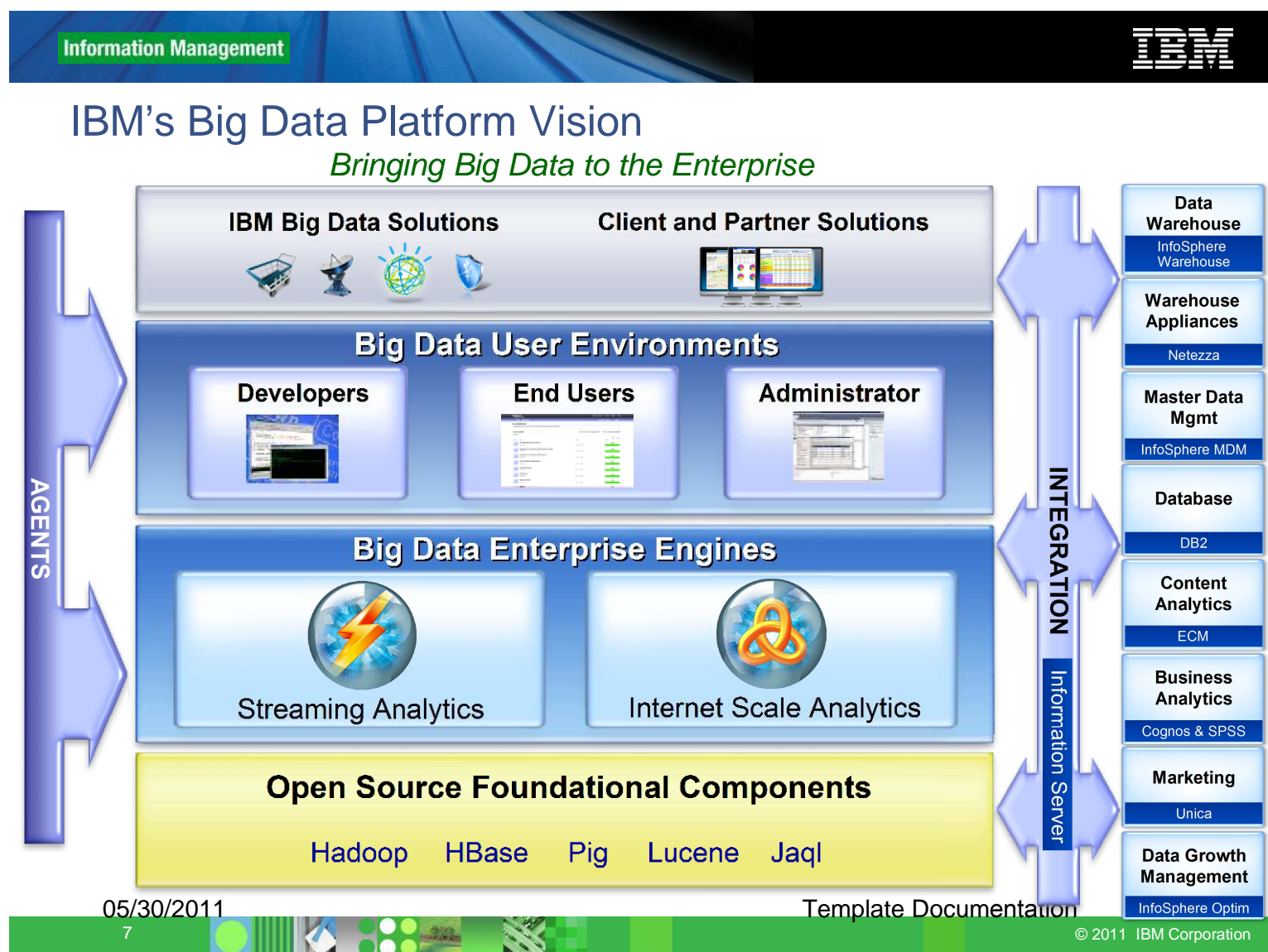
05/30/2011

Template Documentation

6

© 2011 IBM Corporation

IBM's Big Data platform is the solution to answering all your big data questions. It can help you bring together any number of data source, of any types, and at any speed to answer those business questions that had previously been beyond reach.



Key Points

- Big Data platform is built upon open source. IBM has embraced open source movement because they believe the Hadoop technology is the correct one to address internet scale analytics. But IBM's approach is to mature and build upon that technology for an enterprise class platform.
- Open source – built on Hadoop (map reduce, HDFS), HBase (Hadoop database), Pig (analysis of large data sets, high level language for data analysis programs), Lucene (full text search), Jawl (query language for Javascript Object Notation)
- IBM has matured the open source tools with two enterprise engines for processing large volumes of data and analyzing a variety of data. Streaming analytics is designed to manage stream flows and apply various analytics – mining, mathematical, video, etc. – against that streaming data. Internet scale analytics is designed to store data at rest, as-is, and apply analytics, such as text analytics against that data set.
- User Environments: This is an important part of maturing the platform and exposing the power of big data to existing resources, not just specialist programmers who can write map reduce programs. The develop environment is designed to provide a mature environment for developing and testing Big Data analytics and applications. There are end-user visualization capabilities to explore the data and analyze it.
- Integration – this is an important aspect of the Big Data platform – it had to be integrated in order to “bring big data to the enterprise” – the insight has to be integrated to warehouses, databases, applications, etc. One of the key vehicles for doing that is Information Integration – which includes governing that data.

Proof Points & Stats

- Tera Echos – ‘our developers can deliver apps 45% faster due to the agility of the streams processing language’ – shows how a mature development environment and language speeds development of BD apps.

How can you learn more about Big Data, including Hadoop?



INSTRUMENTED



INTERCONNECTED



INTELLIGENT



Learn more about IBM's Big Data Platform and Technology:
<http://www-01.ibm.com/software/data/bigdata/>

Hadoop: Part of the technology behind Big Data:
<http://www-01.ibm.com/software/ebusiness/jstart/hadoop/>

05/30/2011

Template Documentation

8

© 2011 IBM Corporation

Go to the links in the slide to learn more about Big Data. The future is now here for Big Data. Learn more to be prepared to use Big Data to solve real world programs to create a Smarter Planet.



Thank you!

Use the forum in the db2university.com course AA001EN if you have technical questions about the materials covered in this course. Fellow students, faculty and IBMers can help you!