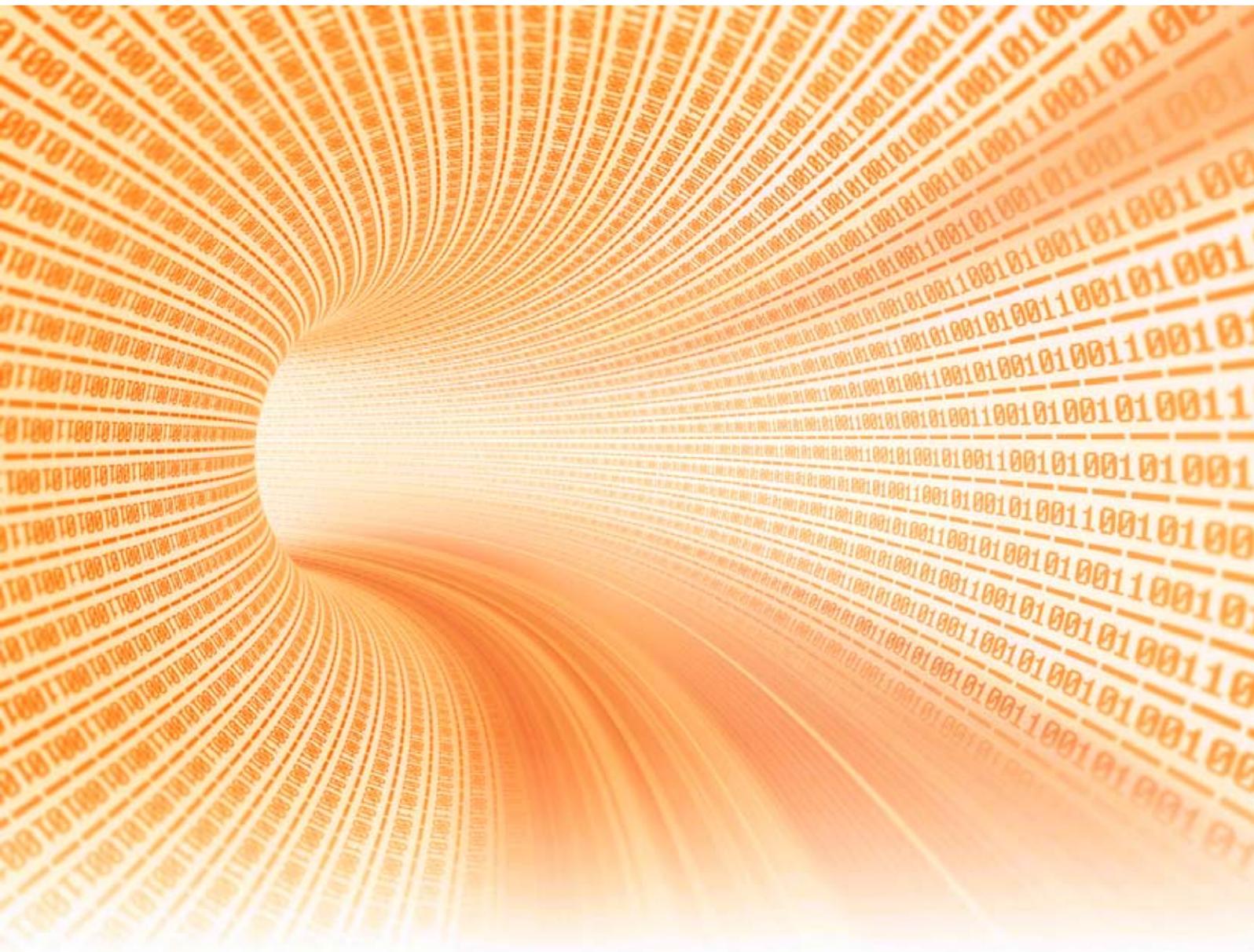# Big Data for Enterprise:
Technology, Strategy, Adoption and Outlook

*Practical advice, recommendations and predictions
on the expanding application of big data in enterprise*

Recent astronomic increases in the rate of data creation have left companies around the world grappling with the same conundrum: how to make what is immense, slow and unwieldy into what is small, nimble and usable. In 2011, it was said that an estimated 90% of the world's data (from the WWW and machine generated data from network nodes and applications) was created over the past two years. The world's data is reportedly doubling every two years and global annual data creation is set to leap from 1.2 zettabytes in 2012 to 35 zettabytes in 2020, according to IDC's 2011 Digital Universe Report. This data explosion, dubbed Big Data, is proving both a stimulating technical challenge and exciting business opportunity. Companies that have successfully invested in big data analytics are unlocking its business potential, and new technologies are storing and analysing data ever more affordably.

Database vendors have exploited these advances by developing new high speed platforms. 2011 saw big data technologies hit the mainstream, with Oracle, IBM, SAP, Teradata and others starting to integrate Hadoop into their enterprise data solutions. Hadoop is a software framework which supports data intensive, distributed applications under a free licence. It enables these applications to work with thousands of nodes and petabytes of data. Originating in the mid-2000s through technologies from Google (MapReduce), Yahoo and others, Hadoop is now central to the big data initiatives of enterprises, service providers and other organizations.

Decision makers across many industries, in particular telecoms, utilities, retailers, governments, healthcare and manufacturing, are beginning to apply new big data strategies and technologies to derive immediate and detailed data-driven insights. Enhanced data visualization; more sophisticated sales, marketing and other CRM applications analytics; rapid detailed, interactive, multidimensional statistical analysis; and improved forecasting are starting to transform the business world.

There's no doubt that having a coherent big data strategy is becoming a critical commercial imperative for enterprise. Yet whilst many companies are now recognising the power of big data analytics to deliver sustainable advantage, most are still struggling to manage this data tsunami and really leverage its true potential.

For this report, Data Driven Business asked three big data experts with different perspectives to share their insights on:

- the current technology landscape
- planning a cost effective and workable big data strategy
- where big data is currently being adopted and applied in business
- outlook for big data over the next three to five years

Contributors

KEITH MANTHEY, VP TECHNOLOGY, EQUIFAX
Keith Manthey is the Vice President of Technology on the Integrated Data Solutions team at Equifax responsible for Equifax's Big Data Platforms. He leads a large team that is focused on developing and managing Equifax's Big Data Platforms. The platforms run on GRID and Private Cloud Architectures that Equifax have been using and developing since 2002. Prior to Equifax, Keith worked at ChoicePoint, Capital One, Freddie Mac, and Prudential in various roles. He holds a Master in MIS from The University of Georgia and a Bachelors in Accounting from Virginia Tech. Keith holds one patent and is actively working on a patent for another technology innovation.

JAMES KOBELIUS, SENIOR ANALYST, BIG DATA, FORRESTER RESEARCH
James Kobelius is a Senior Analyst for Big Data and Andvanced Analytics at Forrester Research. He is a leading expert on data warehousing, predictive analytics, data mining, and complex event processing. In addition to his core areas, James contributes to Forrester's research inbusiness intelligence, data integration, data quality, and master data management. James is a widely published business/technology author and has spoken at many industry events.

KRISH KRISHNAN, CEO, SIXTH SENSE ADVISORS
Krish Krishnan is an expert in the strategy, architecture and implementation of high performance data warehousing solutions. He is a recognized data warehouse thought leader, writing and speaking at industry leading conferences, user groups and trade publications. In his 18 years of professional experience he has been solving complex solution architecture problems spanning all aspects of data warehousing and business intelligence for fortune 1000 clients.

Krish and Keith will be also be speaking at Data Driven Business's Big Data for Enterprise conference, taking place on June 27-28 in San Francisco. James is leaving Forrester in April but his colleague Brian Hopkins, a Principal Analyst in Forrester's Enterprise Architecture division, will be speaking at the event.

Big Data for Enterprise USA 2012 brings together innovative companies which are now recognizing the power of big data analytics to drive business growth and deliver sustainable advantage.

Full details can be found at:
www.datadrivenbiz.com/enterprise-big-data/index.php

Big Data for Enterprise USA 2012 brings together innovative companies which are now recognizing the power of big data analytics to drive business growth and deliver sustainable advantage.

Full details can be found at www.datadrivenbiz.com/ enterprise-big-data/index.php

Big Data for Enterprise USA 2012
Conference and Expo
June 27-28, San Francisco

Unlock the power of big data to maximize customer value, identify new revenue streams and ensure competitive advantage

www.datadrivenbiz.com/bigdata

What is happening in big data technology in 2012?

JAMES KOBELIUS: Adoption of big data technology has been staggeringly rapid. Batch loading is still typical, but less dominant. Big data technologies are enabling a shift away from batch-loading to data-streaming and real-time continuous analytics. Pilots, testing and development, and an expanding range of applications have recently progressed at a fever pace. Focus on big data technology has intensified over the past two years, with substantial interest from large enterprises and IT companies. In 2010, companies were asking about big data technology definitions. In 2011, they were asking about technology decisions. Over the past six to nine months they have a general idea of who or what, but want input on best practices, optimizing and insight on what's not quite working with Hadoop, and how they can learn from mistakes that have already been made. Hadoop is evolving rapidly in new directions. It is being widely commercialized and adopted in enterprises and its open-source nature will be the heart of the new frontier in big data. What's more, Hadoop and other approaches will merge and mash, which will be very exciting.

KRISH KRISHNAN: There are two things that will revolutionize big data technology in 2012. The first is in-memory processing for Hadoop, which I expect will happen sooner or later this year. The second is a strong emergence of powerful mobility based business intelligence (BI) solutions on mobile platforms. This is being driven by the arrival of LTE (4G) and will be a huge adoption driver for people to start utilising big data reports on mobile platforms. It will offer major benefits in terms of enabling easier big data visualization and personalizing big data. Tracking business performance from big data on social media and non-traditional solutions will radically change how people view and consume big data. An example of this is smart meters: energy and cost conscious end-users will be able to control domestic energy use from an app on their smartphones when out and about, depending on their needs that day. In terms of the technology solutions marketplace, this is an exciting phase. The key Hadoop start-up technology distributors (Cloudera, Hortonworks, Hadapt, Skytree etc.) are marketing themselves as a partnership point for the established enterprise data players. Lots of tie-ups have already happened in the first quarter of 2012 and there will be more over the rest of the year. There may well be significant mergers and acquisitions activity, especially towards the last quarter of 2012 with data leaders such as IBM, Oracle, SAP, Microsoft and Teradata looking to purchase the new start-up leaders. Or we could see solutions based consolidation with significant acquisitions of front-end enablers, leaving Hadoop alone for a while. The third thing to watch for is the emergence of cloud based development and deployment, especially for big data solutions. Now, with examples like Amazon Dynamo DP, you have a much more scalable database from the back-end that supports everything that you can throw at it.

KEITH MANTHEY: The past year has been characterised by the maturing of the reporting and BI tools that work on top of the existing big data tool sets. Now in 2012, tools such as Data Meer and JasperSoft provide a more robust reporting and BI capability for the big data ecosystem. The availability of these tools allows for reporting direct from the big data environment instead of having to 'spin a data mart' for every report. While

Hadoop is evolving rapidly in new directions. It is being widely commercialized and adopted in enterprises and its open-source nature will be the heart of the new frontier in big data.

this may seem minor, a group that generates 200 different types of reports may in fact find themselves spinning upwards of 50 or more marts to feed those reporting requirements. The development of new tools and the big data ecosystem build-out are providing a compelling reason for even companies like Equifax, who have traditionally 'rolled their own' big data platforms, to look towards the vendor offerings that are available on the market. The ecosystems surrounding the big data tools will continue to evolve over 2012/13. My expectation is that the level of investment by vendors and consolidation of software providers will help grow the total ecosystem supporting tools like Hadoop. As the big data ecosystem grows, enhancements around data security, reporting, scalability, and auditing will help even heavily regulated industries to get a better comfort level with Hadoop and other big data tools. I see less of a focus on ground breaking technologies in the next year and a stronger focus on hardening the tools that already exist. Continued investment in tools like Cloudera and Mapr that are working on hardening the entire big data ecosystem will continue to help adoption at companies wanting that level of security and hardening instead of them having to put investment into building those capabilities themselves.

As the big data ecosystem grows, enhancements around data security, reporting, scalability, and auditing will help even heavily regulated industries to get a better comfort level with Hadoop and other big data tools.

**Big Data for Enterprise
USA 2012**
Conference and Expo
June 27-28, San Francisco

Unlock the power of big data to maximize customer value, identify new revenue streams and ensure competitive advantage

www.datadrivenbiz.com/bigdata

What is your strategic advice to companies considering implementing new big data technology?

JAMES KOBELIUS: Early adopters have implemented big data technology to support customer-facing applications such as customer experience optimization as well as IT functions such as root-cause analysis, event processing and software usage compliance. There are Hadoop applications that specifically meet the needs of web-facing businesses in terms of optimizing ad placement, monitoring content republishing and identifying influential customers. Companies considering adopting Hadoop technologies should consider it as a possible approach for the most demanding analytic applications that would otherwise run on their massively parallel processing (MPP) enterprise data warehouse (EDW). They should consider adopting the open source Apache Hadoop codebase and customizing it to their specific requirements. Many early implementers use Hadoop clusters as staging layers behind a RDBMS-based EDW or data mart. For all its strengths and potential, Hadoop is no panacea. It lacks some critical EDW features, such as real-time integration and robust high availability. Don't get too distracted by what's trendy; look at what you've invested in in terms of skill sets and your existing environment. Think carefully about what is the added value of going with a new approach. Don't adopt new technologies until you are completely comfortable with them, and have examined use cases. Reconnect with existing vendors and see how they have already adopted Hadoop. Compare new vendors in Hadoop with your incumbent vendors. Don't go big data if you don't have big data requirements!

KRISH KRISHNAN: If we take a step back, there are a number potentially disruptive influences that enterprises need to acknowledge before developing their big data strategy. The first of these is competition. Every company wants to achieve superior customer excellence through social CRM, every customer wants to be treated as the most important customer and every vendor wants to gain wallet share by offering competitive price-points. In addition, customer access to technology means that the norms of customer behaviour are changing. Customers can go into a store and simply by clicking on the product code can compare prices. This competition is making vendors think about how to get close to customers, and there is a constant pressure to innovate and provide new services or solutions while at the same time dealing with current business demands. The second is compliance. So many laws have come in over the past three to five years around the world, and this is creating huge disruption across many industry sectors.

KEITH MANTHEY: First and foremost, a company should make sure they understand why investment in a big data environment will provide them a with competitive business advantage. At Equifax, we have been dealing with tens of billions of pieces of unstructured data at the terabyte and petabyte scale for decades, and so the evolution to big data was relatively natural for us. The level of investment to manage the velocity, volume, or latency of update is up to the individual company. Beyond the investment level, the governance processes required to establish a large scale big data 'factory' are different than those for a company that stores all of its data in

Don't get too distracted by what's trendy; look at what you've invested in in terms of skill sets and your existing environment. Think carefully about the added value of going with a new approach.

Big Data for Enterprise USA 2012
Conference and Expo
June 27-28, San Francisco

Unlock the power of big data to maximize customer value, identify new revenue streams and ensure competitive advantage

www.datadrivenbiz.com/bigdata

relational data marts. Petabytes of data distributed across a large compute cluster further complicate data governance. Any company that wishes to entertain a big data environment needs to understand how the governance model should work surrounding the data. While there are many aspects of big data best practice to consider, another crucial thing to think about is infrastructure strategy. It is important for the potential big data company to understand how they will manage data and data sprawl. What will be the strategy for managing the distributed data sets? How will data growth be governed and scaled to handle its growth? Will the environment infrastructure be architected so that it can handle the input/output (I/O) requirements of the big data platform? I remember a conversation I had with another company looking to break into the big data space with a Massively Parallel Database provider. They had no experience in handling the scaling needs that a distributed environment would require. They thought their experience around managing large centralized disk towers would prepare them for managing a distributed compute environment. As such, they used a centralized SAN based environment to feed their distributed compute environment. While I didn't review their architecture, I can speculate that they didn't take account of the real access path needs to disk a distributed computing environment. The reason I can speculate this is because I heard feedback that their foray into the distributed environment computing was not very successful and the environment was slowed down waiting for disk access.

Any company that wishes to entertain a big data environment needs to understand how the governance model should work surrounding the data.

At what stage are enterprises looking to implement big data technology typically at?

KRISH KRISHNAN: While the more web-based customer-focused companies have already invested in big data solutions, most mid-sized organizations in industries such as manufacturing and healthcare are currently in the planning stage. They often know what they want to achieve, but are not sure how to get there. Many are using a partner, such as enterprise software consultants, to help build a road-map of how to select from the growing number of options out there, as it is very important to get it right and different technologies are required for different organizations and goals. The requisite expertise in adoption and roll-out will not usually be available in-house for such new technologies. As this is still the early adoption phase, the current lack of clarity in regulation and a fear of the unknown may tempt many analysts to wait for a different technology solution or a lower price-point. This is understandable. However, at some point sooner rather than later, they will need to make a leap of faith. Until then, as there is so much potential for business gain if well planned and implemented, it is crucial to start to working on your big data implementation plan now, so that when the time comes you know where you want to go and how you need to get there.

KEITH MANTHEY: We have watched big data technologies work on solving security and consistency challenges, and are now embarking on new technologies beyond research and development in lieu of developing our own technologies. In our case, it was advantageous for us to wait until the ecosystem was built out. The biggest challenge that most organizations will face in attempting to implement a big data strategy is finding technologists with experience in a big data environment. There are certainly plenty of training courses and vendor certifications that can educate and train staff for a big data environment. There is still no substitute for having your big data environment architect with real world big data experience. We have a three pronged approach to getting the right people: Equifax employees with years of experience, hiring qualified talent away from other established big data players, and grooming our own next generation of big data talent.

It is crucial to start to working on your big data implementation plan now, so that when the time comes you know where you want to go and how you need to get there.

In what way is big data technology currently being most widely applied?

JAMES KOBELIUS: Enterprises are increasingly evaluating big data technologies to enable performance-enhancing strategies across their organizations. Four key focus areas include faster queries, reports, dashboards, and key performance indicators (KPI) delivered from the database; a single view of everything consolidated into a high performance database; a single version of the truth (governed through tight controls on the trustworthiness of information in the database) and self-service information exploration (interactive visualisation, what-if analysis, simulation and modelling of information in the database). Currently, industry- specific applications of big data include: fraud detection and prevention in financial; geospatial analytics in telecoms and online services, supply-chain optimization in manufacturing and evidence based medicine in pharmaceuticals.

KRISH KRISHNAN: There are two areas of early focus, namely social CRM and 'inside use' of unstructured data. Without a doubt, social CRM provides the best bang for your buck right now in terms of big data application. Behavioural analysis means that you are able to look not only at what the customer wants, but what the customer does. You are going to make sure that you meet their expectations or beat their expectations. As an example, say a customer wants to purchase an SD card for her camera. She will go to the website, click at the ones of interest, and then look at the customer peer reviews. 99% of customers will choose the product with the highest rating, which is usually one of the first four that they see. Peer reviews, and especially a ratio of peer reviews (e.g. 'this product has had 200 positive reviews and 80 negative reviews') and personalization at a granular level provide the customer with a dashboard and enable the sellers to interact with and influence the customer to increase sales. 'Inside use' of unstructured data refers to the organisation and analysis of unstructured data lying in many different silos, such as content management farms, paper, spreadsheets etc., within an organisation. By scanning all internal documents, applying matrixes and using metadata to link it to your existing applications, enterprise searches can be oriented more towards a contextualised search. This will lead to better operational decisions, and will be good for compliance and auditing. Early adopters will be heavily regulated industries such as banks, healthcare and pharmaceuticals.

KEITH MANTHEY: Web clicks get press coverage. However, big data techniques are being used in a different way by companies that provide financial services from private data assets to render their formerly unwieldy large stores into historical/performance views and derive intelligence from trends in the data. Like web clicks, usage patterns in the data are also new forms of information on top of the asset catalogues. Companies that value big data and 'keeping everything they know' historically have justified expensive solutions. As capability costs fall due to new technologies and lower storage and hardware-performance price points, more companies are taking the leap into big data technologies, because what were prohibitively expensive or unwieldy assets suddenly have positive value. Adoption will be shockingly sudden. Marketing companies or companies

Without a doubt, social CRM provides the best bang for your buck right now in terms of big data application.

that want to in-source marketing studies will lead the charge to adoption because the relatively simple schemas of web activity or sales activity are well-known and common examples in the space. A full array of pre-canned big data tools are already popping up to support these domains. Innovations in scale will come from these markets, and maybe less so from organizations that support scientific computation, such as medical and atmospheric domains, that have developed mature parallel computation models already. Corporate and civil entities that build and serve applications on their own large stores and currently pay for extravagant data management solutions will drive innovation in schema, metadata management and a variety of storage formats and media to support transactional and batch throughput requirements.

Key obstacles to big data adoption will first and foremost be newcomers from buyer to software developer to CTO who don't appreciate or understand scale and the possibilities that emerge (as well as complexities of apparent nondeterministic behaviour and cleaning up processing errors) when you can divide the historical wall-clock runtime or computational expense by two orders of magnitude. Implications abound in costs, design and toolset choices, requiring behavioural changes in patterns, purchasing and operation. Scale must be learned, sometimes painfully. Failure to partner or staff with appropriate skills or unwillingness to recast the business problem will throw organizations over the cliff (cresting the 'hype curve') quickly. Some components of today's big data solutions are still immature because a fairly narrow example set is getting the lion's share of attention. Business domains that need to process simple, atomic units of work in large quantities ('map') - particularly those with business problems that don't require multi-core technology but have many 'units of work' in flight constantly throughout the processing day - will benefit most and quickest. Those that are heavy 'reduce' - needing monolithic processing or post-processing because of complex aggregation, unclear natural data partitions or stateful data processing may be disappointed until the more complex data domains get behind and help mature the open source and commercial offerings.

Failure to partner or staff with appropriate skills or unwillingness to recast the business problem will throw organizations over the cliff quickly.

What is the outlook for big data technology over the next decade?

JAMES KOBELIUS: Over the next decade, companies will move from the high gigabytes and low petabytes to operational analytics in the tens to hundreds of petabytes. Most industries are currently internally testing and piloting Hadoop as an adjunct to their EDW. Hadoop is a growing community that is already universal in terms of industries and geography. But it is not the only open source approach. Hadoop and other approaches will continue to mature and evolve parity of features with EDW. In 2010, cost per petabyte of raw storage was US$80,000. In 2020, it is not far-fetched to assume a cost as low as US$4 per petabyte of usable storage. By the end of this decade, people throughout the world from every socio-economic class will cost-effectively be able to fit one hundred times the size of a current EDW on their hand-held device.

KRISH KRISHNAN: Five years from now every segment of a large multinational company will have a very streamlined end-to-end data strategy that will integrate not only existing data but will include new data types and new data sets. It will have reached a level of maturity. However, we will face a number of challenges along the way. Globalization -the use of the same products across many different countries - will require an ability to support data strategies at a regional level that is decentralised and self-contained. It will also require the ability to scale up and scale out solutions. A third issue will be seamless security in the cloud, and security in general. This big data technology journey is just beginning. In a few years' time there will be major challenges, as I have just explained, but by then the market will evolved enough to tackle those challenges. And two years on from then, it will be really well on its way.

KEITH MANTHEY: I think the most impactful big data analysis will actually emerge surreptitiously. I am not talking about wilful secrecy. There are a plethora of big data solutions in the cloud that allow business users to perform big data analytics with no involvement from IT. This problem has loomed before in other spaces and corporate IT has had to play catch up. The most recent was the governance and management of social media and its impact on companies. Without proper governance and approval, organizations may find that their data has been moved to the 'big data cloud' without their knowledge. The ease with which clusters of computers are ramped up, burdened with work, and then ramped down, is amazing. The cost factor for provisioning compute cycles on demand and not incurring the sunk cost for assets when they are not in use is compelling. However, this isn't always the right solution. I do think that the continued expansion and use of cloud-based big data platforms will drive organizations to develop solutions, governance, and usage models for big data platforms. This is especially the case for companies that haven't viewed themselves as players in the big data space, but have business problems that big data platforms can lend themselves to.

The 'next big thing' in data will be tools to manage the governance and data sprawl caused by the adoption of big data environments. As organizations struggle to deal with rapid adoption, they will be looking for

By the end of this decade, people throughout the world from every socioeconomic class will cost-effectively be able to fit one hundred times the size of a current EDW on their hand-held device.

tools to help them manage the proliferation of big data environments. I view this as a huge opportunity.

Big data is still in its infancy on the 'hype curve'. The biggest opportunity will be amongst organizations that have not traditionally considered themselves big data companies. As the implementation costs either via cloud models or of on-site appliance decline, companies will look to leverage their data assets in new ways. This creates a boom for big data employment, tool vendors, hardware providers (either cloud or appliance) as well as the big data industry as a whole. On the flip side, the challenge will be for the vendors to continue to harden solutions to help with real world governance issues. Companies that need to deal with regulatory issues such as PCI, Gramm Leach Bliley, or any other situation will be looking for extra compliance modules. Without these being available in all markets for all compliance related issues, companies will not necessarily leap head long into adoption. There are still some reasons why certain companies may find they aren't ready to leap head long into the bright new world of big data.

The 'next big thing' in data will be tools to manage the governance and data sprawl caused by the adoption of big data environments.

Big Data for Enterprise USA 2012
Conference and Expo
June 27-28, San Francisco

Unlock the power of big data to maximize customer value, identify new revenue streams and ensure competitive advantage

www.datadrivenbiz.com/bigdata

Conclusions

1. Hadoop is evolving rapidly in new directions. It is being widely commercialized and adopted in enterprises and its open-source nature will be the heart of the new frontier in big data.

2. In-memory processing for Hadoop and a strong emergence of powerful mobility based business intelligence (BI) solutions on mobile platforms will take the big data technology world by storm in 2012.

3. As the big data ecosystem grows, enhancements around data security, reporting, scalability, and auditing will help even heavily regulated industries to get a better comfort level with Hadoop and other big data tools.

4. For those companies wishing to implement a big data strategy in due course, it is crucial to start to working on your big data implementation plan now.

5. The biggest challenge that most organizations will face in attempting to implement a big data strategy is finding technologists with experience in a big data environment.

6. Big data technologies enable performance-enhancing strategies in four key focus areas:
   a. Faster queries, reports, dashboards, and key performance indicators (KPI) delivered from the database
   b. A single view of everything consolidated into a high performance database
   c. A single version of the truth (governed through tight controls on the trustworthiness of information in the database)
   d. Self-service information exploration (interactive visualisation, what-if analysis, simulation and modelling of information in the database)

7. There are two areas of early focus, namely social CRM and 'inside use' of unstructured data.

8. Marketing companies or companies that want to in-source marketing studies will lead the charge to adoption because the relatively simple schemas of web activity or sales activity are well-known and common examples in the space.

9. Key obstacles to big data adoption will first and foremost be newcomers from buyer to software developer to CTO who don't appreciate or understand scale and the possibilities that emerge.

10. By the end of this decade, people throughout the world from every socio-economic class will cost-effectively be able to fit one hundred times the size of a current EDW on their hand-held device.

11. Five years from now every segment of a large multinational company will have a very streamlined end-to-end data strategy that will integrate not only existing data but will include new data types and new data sets.

12. The 'next big thing' in data will be tools to manage the governance and data sprawl caused by the adoption of big data environments.

13. Continued expansion and use of cloud-based big data platforms will drive organizations to develop solutions, governance, and usage models for big data platforms, especially companies that haven't viewed themselves as players in the big data space, but have business problems that big data platforms can lend themselves to.

All of the industry experts interviewed in this report are speaking at the 'Big Data for Enterprise' event in **San Francisco on June 27—28, 2012**.

They will be joined by more than 20 other speakers, who will present on all aspects of big data technology, strategy and application for enterprise.

More information and opportunities to reserve places at this landmark event are available on the meeting website, **www.datadrivenbiz.com/bigdata**

Exclusive ticket offer: As you have downloaded this report, just use discount code "REPORT" and receive an extra $100 off the current listed prices when you reserve your place at the conference through the event website — www.datadrivenbiz.com/bigdata