**Technology Forecast: Rethinking integration**
*Issue 1, 2014*

# *The enterprise data lake: Better integration and deeper analytics*

**By Brian Stein and Alan Morrison**

Data lakes that can scale at the pace of the cloud remove integration barriers and clear a path for more timely and informed business decisions.

pwc

*Enterprises across industries are starting to extract and place data for analytics into a single, Hadoop-based repository.*

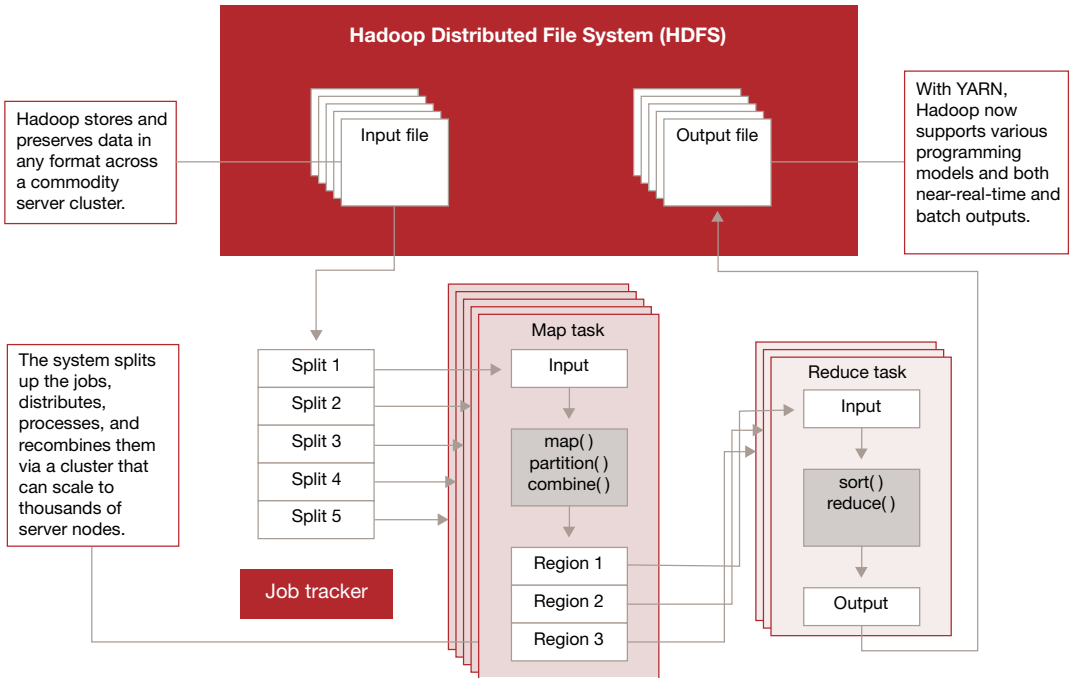
## *Data lakes: An emerging approach to cloud-based big data*

UC Irvine Medical Center maintains millions of records for more than a million patients, including radiology images and other semi-structured reports, unstructured physicians' notes, plus volumes of spreadsheet data. To solve the challenge the hospital faced with data storage, integration, and accessibility, the hospital created a data lake based on a Hadoop architecture, which enables distributed big data processing by using broadly accepted open software standards and massively parallel commodity hardware.

Hadoop allows the hospital's disparate records to be stored in their native formats for later parsing, rather than forcing all-or-nothing integration up front as in a data warehousing scenario. Preserving the native format also helps maintain data provenance and fidelity, so different analyses can be performed using different contexts. The data lake has made possible several data analysis projects, including the ability to predict the likelihood of readmissions and take preventive measures to reduce the number of readmissions.[1]

Like the hospital, enterprises across industries are starting to extract and place data for analytics into a single Hadoop-based repository without first transforming the data the way they would need to for a relational data warehouse.[2] The basic concepts behind Hadoop[3] were devised by Google to meet its need for a flexible, cost-effective data processing model that could scale as data volumes grew faster than ever. Yahoo, Facebook, Netflix, and others whose business models also are based on managing enormous data volumes quickly adopted similar methods. Costs were certainly a

### A basic Hadoop architecture for scalable data lake infrastructure




Source: Electronic Design, 2012, and Hortonworks, 2014

1 "UC Irvine Health does Hadoop," Hortonworks, http://hortonworks.com/customer/uc-irvine-health/.

2 See Oliver Halter, "The end of data standardization," March 20, 2014, http://usblogs.pwc.com/emerging-technology/the-end-of-data-standardization/, accessed April 17, 2014.

3 Apache Hadoop is a collection of open standard technologies that enable users to store and process petabyte-sized data volumes via commodity computer clusters in the cloud. For more information on Hadoop and related NoSQL technologies, see "Making sense of Big Data," PwC *Technology Forecast* 2010, Issue 3 at http://www.pwc.com/us/en/technology-forecast/2010/issue3/index.jhtml.

*Hadoop can be 10 to 100 times less expensive to deploy than conventional data warehousing.*

factor, as Hadoop can be 10 to 100 times less expensive to deploy than conventional data warehousing. Another driver of adoption has been the opportunity to defer labor-intensive schema development and data cleanup until an organization has identified a clear business need. And data lakes are more suitable for the less-structured data these companies needed to process.

Today, companies in all industries find themselves at a similar point of necessity. Enterprises that must use enormous volumes and myriad varieties of data to respond to regulatory and competitive pressures are adopting data lakes. Data lakes are an emerging and powerful approach to the challenges of data integration as enterprises increase their exposure to mobile and cloud-based applications, the sensor-driven Internet of Things, and other aspects of what PwC calls the New IT Platform.

## *Issue overview: Integration fabric*

The data lake topic is the first of three topics as part of the integration fabric research covered in this issue of the PwC *Technology Forecast*. The integration fabric is a central component for PwC's New IT Platform.*

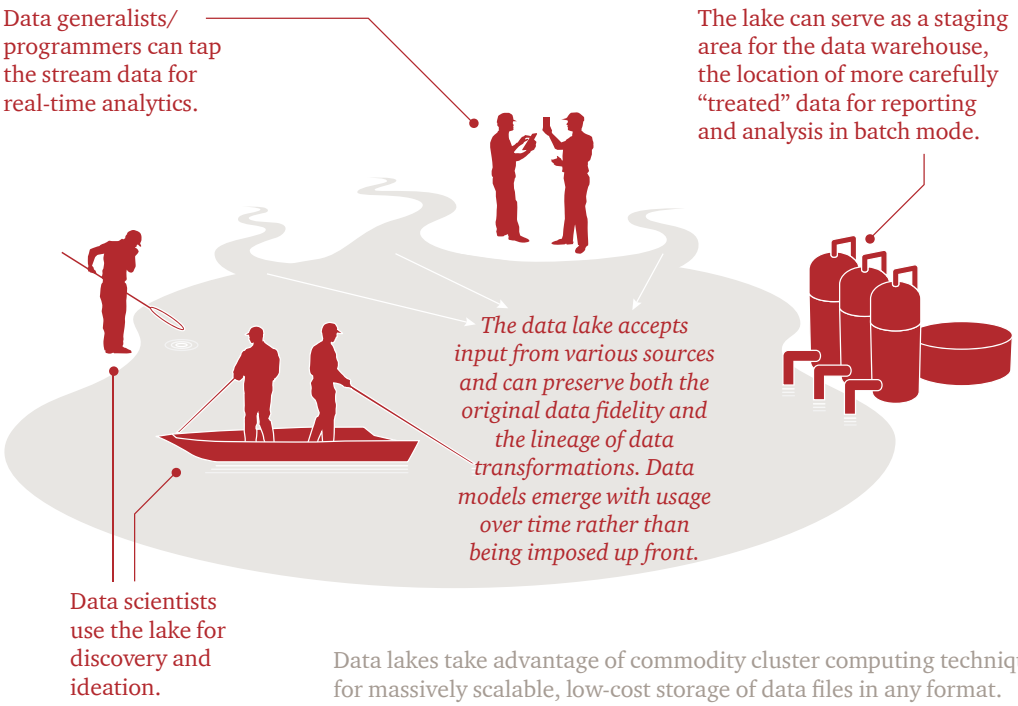
Enterprises are starting to embrace more practical integration. A range of these new approaches is now emerging, and during the next few months we'll ponder what the new cloud-inspired enterprise integration fabric looks like. The main areas we plan to explore include these:

| Integration fabric layers | Integration challenges | Emerging technology solutions |
|---|---|---|
| **Data** | Data silos, data proliferation, rigid schemas, and high data warehousing cost; new and heterogeneous data types | Hadoop data lakes, late binding, and metadata provenance tools |
| | *Enterprises are beginning to place extracts of their data for analytics and business intelligence (BI) purposes into a single, massive repository and structuring only what's necessary. Instead of imposing schemas beforehand, enterprises are allowing data science groups to derive their own views of the data and structure it only lightly, late in the process.* | |
| **Applications and services** | Rigid, monolithic systems that are difficult to update in response to business needs | Microservices |
| | *Fine-grained microservices, each associated with a single business function and accessible via an application programming interface (API), can be easily added to the mix or replaced. This method helps developer teams create highly responsive, flexible applications.* | |
| **Infrastructure** | Multiple clouds and operating systems that lack standardization | Software containers for resource isolation and abstraction |
| | *New software containers such as Docker extend and improve virtualization, making applications portable across clouds. Simplifying application deployment decreases time to value.* | |

* See http://www.pwc.com/us/en/increasing-it-effectiveness/new-it-platform.jhtml for more information.

## What is a data lake?

A repository for large quantities and varieties of data, both structured and unstructured.




Data lakes take advantage of commodity cluster computing techniques for massively scalable, low-cost storage of data files in any format.

## *Why a data lake?*

Data lakes can help resolve the nagging problem of accessibility and data integration. Using big data infrastructures, enterprises are starting to pull together increasing data volumes for analytics or simply to store for undetermined future use. (See the sidebar "Data lakes defined.") Mike Lang, CEO of Revelytix, a provider of data management tools for Hadoop, notes that "Business owners at the C level are saying, 'Hey guys, look. It's no longer inordinately expensive for us to store all of our data. I want all of you to make copies. OK, your systems are busy. Find the time, get an extract, and dump it in Hadoop.'"

Previous approaches to broad-based data integration have forced all users into a common predetermined schema, or data model. Unlike this monolithic view of a single enterprise-wide data model, the data lake relaxes standardization and defers modeling, resulting in a nearly unlimited potential for operational insight and data discovery. As data volumes, data variety, and metadata richness grow, so does the benefit.

Recent innovation is helping companies to collaboratively create models—or views—of the data and then manage incremental improvements to the metadata. Data scientists and business analysts using the newest lineage tracking tools such as Revelytix Loom or Apache Falcon can follow each other's purpose-built data schemas. The lineage tracking metadata also is placed in the Hadoop Distributed File System (HDFS)—which stores pieces of files across a distributed cluster of servers in the cloud—where the metadata is accessible and can be collaboratively refined. Analytics drawn from the lake become increasingly valuable as the metadata describing different views of the data accumulates.

Every industry has a potential data lake use case. A data lake can be a way to gain more visibility or put an end to data silos. Many companies see data lakes as an opportunity to capture a 360-degree view of their customers or to analyze social media trends.

In the financial services industry, where Dodd-Frank regulation is one impetus, an institution has begun centralizing multiple data warehouses into a repository comparable to a data lake, but one that standardizes on XML. The institution is moving reconciliation, settlement, and Dodd-Frank reporting to the new platform. In this case, the approach reduces integration overhead because data is communicated and stored in a standard yet flexible format suitable for less-structured data. The system also provides a consistent view of a customer across operational functions, business functions, and products.

Some companies have built big data sandboxes for analysis by data scientists. Such sandboxes are somewhat similar to data lakes, albeit narrower in scope and purpose. PwC, for example, built a social media data sandbox to help clients monitor their brand health by using its SocialMind application.[4]

### *Data lakes defined*

Many people have heard of data lakes, but like the term *big data*, definitions vary. Four criteria are central to a good definition:

- **Size and low cost:** Data lakes are big. They can be an order of magnitude less expensive on a per-terabyte basis to set up and maintain than data warehouses. With Hadoop, petabyte-scale data volumes are neither expensive nor complicated to build and maintain. Some vendors that advocate the use of Hadoop claim that the cost per terabyte for data warehousing can be as much as $250,000, versus $2,500 per terabyte (or even less than $1,000 per terabyte) for a Hadoop cluster. Other vendors advocating traditional data warehousing and storage infrastructure dispute these claims and make a distinction between the cost of storing terabytes and the cost of writing or written terabytes.*
- **Fidelity:** Hadoop data lakes preserve data in its original form and capture changes to data and contextual semantics throughout the data lifecycle. This approach is especially useful for compliance and internal audit. If the data has undergone transformations, aggregations, and updates, most organizations typically struggle to piece data together when the need arises and have little hope of determining clear provenance.
- **Ease of accessibility:** Accessibility is easy in the data lake, which is one benefit of preserving the data in its original form. Whether structured, unstructured, or semi-structured, data is loaded and stored as is to be transformed later. Customer, supplier, and operations data are consolidated with little or no effort from data owners, which eliminates internal political or technical barriers to increased data sharing. Neither detailed business requirements nor painstaking data modeling are prerequisites.
- **Late binding:** Hadoop lends itself to flexible, task-oriented structuring and does not require up-front data models.

*For more on data accessibility, data lake cost, and collective metadata refinement including lineage tracking technology, see the interview with Mike Lang, "Making Hadoop suitable for enterprise data science," at www.pwc.com/technologyforecast/mike-lang. For more on cost estimate considerations, see Loraine Lawson, "What's the Cost of a Terabyte?" *ITBusinessEdge,* May 17, 2013, at http://www.itbusinessedge.com/blogs/integration/whats-the-cost-of-a-terabyte.html.

## *Motivating factors behind the move to data lakes*

Relational data warehouses and their big price tags have long dominated complex analytics, reporting, and operations. (The hospital described earlier, for example, first tried a relational data warehouse.) However, their slow-changing data models and rigid field-to-field integration mappings are too brittle to support big data volume and variety. The vast majority of these systems also leave business users dependent on IT for even the smallest enhancements, due mostly to inelastic design, unmanageable system complexity, and low system tolerance for human error. The data lake approach circumvents these problems.

### Freedom from the shackles of one big data model

Job number one in a data lake project is to pull all data together into one repository while giving minimal attention to creating schemas that define integration points between disparate data sets. This approach facilitates access, but the work required to turn that data into actionable insights is a substantial challenge. While integrating the data takes place at the Hadoop layer, contextualizing the metadata takes place at schema creation time.

Integrating data involves fewer steps because data lakes don't enforce a rigid metadata schema as do relational data warehouses. Instead, data lakes support a concept known as *late binding*, or *schema on read*, in which users build custom schema into their queries. Data is bound to a dynamic schema created upon query execution. The late-binding principle shifts the data modeling from centralized

4 For more information on SocialMind and other analytics applications PwC offers, see http://www.pwc.com/us/en/analytics/analytics-applications.jhtml.

data warehousing teams and database administrators, who are often remote from data sources, to localized teams of business analysts and data scientists, who can help create flexible, domain-specific context. For those accustomed to SQL, this shift opens a whole new world.

In this approach, the more that is known about the metadata, the easier it is to query. Pre-tagged data, such as Extensible Markup Language (XML), JavaScript Object Notation (JSON), or Resource Description Framework (RDF), offers a starting point and is highly useful in implementations with limited data variety. In most cases, however, pre-tagged data is a small portion of incoming data formats.

## Early lessons and pitfalls to avoid

Some data lake initiatives have not succeeded, producing instead more silos or empty sandboxes. Given the risk, everyone is proceeding cautiously. "We see customers creating *big data graveyards*, dumping everything into HDFS and hoping to do something with it down the road. But then they just lose track of what's there," says Sean Martin, CTO of Cambridge Semantics, a data management tools provider.

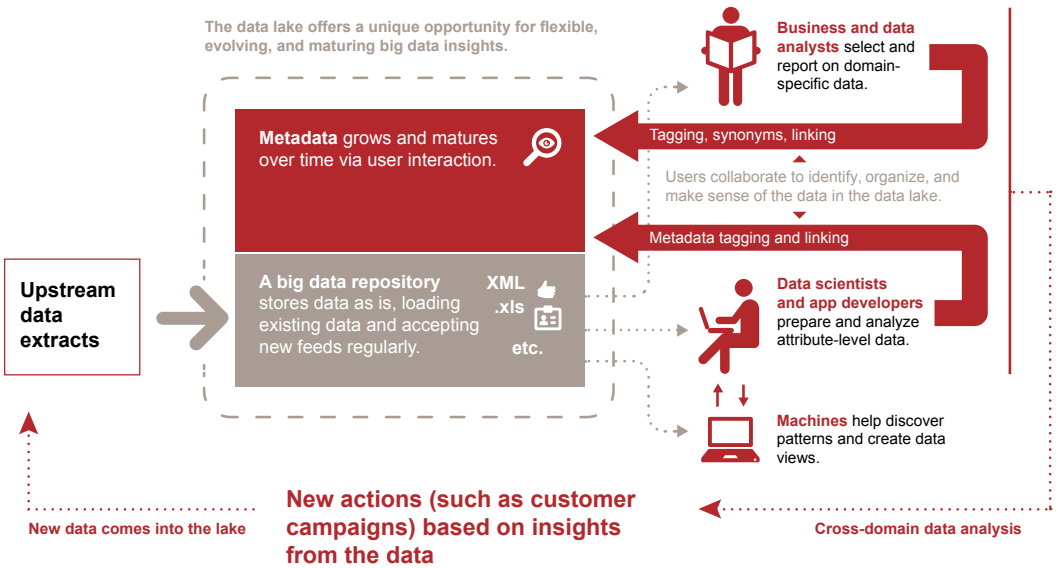
Companies avoid creating big data graveyards by developing and executing a solid strategic plan that applies the right technology and methods to the problem. Few technologies in recent memory have as much change potential as Hadoop and the NoSQL (Not only SQL) category of databases, especially when they can enable a single enterprise-wide repository and provide access to data previously trapped in silos. The main challenge is not creating a data lake, but taking advantage of the opportunities it presents. A means of creating, enriching, and managing semantic metadata incrementally is essential.

*"We see customers creating big data graveyards, dumping everything into HDFS and hoping to do something with it down the road. But then they just lose track of what's there."*

*—Sean Martin, Cambridge Semantics*

### Data flow in the data lake

The data lake loads data extracts, irrespective of format, into a big data store. Metadata is decoupled from its underlying data and stored independently, enabling flexibility for multiple end-user perspectives and incrementally maturing semantics.



*With the data lake, users can take what is relevant and leave the rest. Individual business domains can mature independently and gradually.*

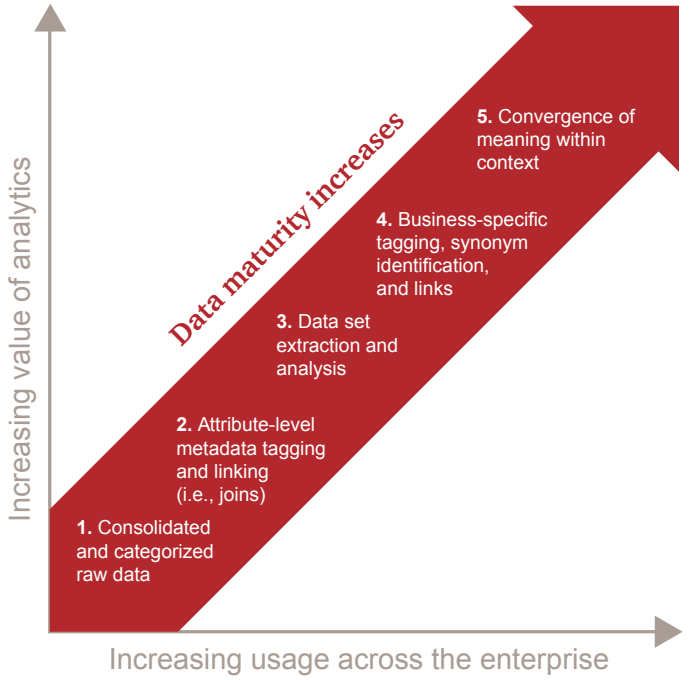
## *How a data lake matures*

Sourcing new data into the lake can occur gradually and will not impact existing models. The lake starts with raw data, and it matures as more data flows in, as users and machines build up metadata, and as user adoption broadens. Ambiguous and competing terms eventually converge into a shared understanding (that is, semantics) within and across business domains. Data maturity results as a natural outgrowth of the ongoing user interaction and feedback at the metadata management layer—interaction that continually refines the lake and enhances discovery. (See the sidebar "Maturity and governance.")

With the data lake, users can take what is relevant and leave the rest. Individual business domains can mature independently and gradually. Perfect data classification is not required. Users throughout the enterprise can see across all disciplines, not limited by organizational silos or rigid schema.

### Data lake maturity

The data lake foundation includes a big data repository, metadata management, and an application framework to capture and contextualize end-user feedback. The increasing value of analytics is then directly correlated to increases in user adoption across the enterprise.



## *Maturity and governance*

Many who hear the term *data lake* might associate the concept with a big data sandbox, but the range of potential use cases for data lakes is much broader. Enterprises envision lake-style repositories as staging areas, as alternatives to data warehouses, or even as operational data hubs, assuming the appropriate technologies and use cases.

A key enabler is Hadoop and many of the big data analytics technologies associated with it. What began as a means of ad hoc batch analytics in Hadoop and MapReduce is evolving rapidly with the help of YARN and Storm to offer more general-purpose distributed analytics and real-time capabilities. At least one retailer has been running a Hadoop cluster of more than 2,000 nodes to support eight customer behavior analysis applications.*

Despite these advances, enterprises will remain concerned about the risks surrounding data lake deployments, especially at this still-early stage of development. How can enterprises effectively mitigate the risk and manage a Hadoop-based lake for broad-ranging exploration? Lakes can provide unique benefits over traditional data management methods at a substantially lower cost, but they require many practical considerations and a thoughtful approach to governance, particularly in more heavily regulated industries. Areas to consider include:

- **Complexity of legacy data:** Many legacy systems contain a hodgepodge of software patches, workarounds, and poor design. As a result, the raw data may provide limited value outside its legacy context. The data lake performs optimally when supplied with unadulterated data from source systems, and rich metadata built on top.
- **Metadata management:** Data lakes require advanced metadata management methods, including machine-assisted scans, characterizations of the data files, and lineage tracking for each transformation. Should schema on read be the rule and predefined schema the exception? It depends on the sources. The former is ideal for working with rapidly changing data structures, while the latter is best for sub-second query response on highly structured data.
- **Lake maturity:** Data scientists will take the lead in the use and maturation of the data lake. Organizations will need to place the needs of others who will benefit within the context of existing organizational processes, systems, and controls.
- **Staging area or buffer zone:** The lake can serve as a cost-effective place to land, stage, and conduct preliminary analysis of data that may have been prohibitively expensive to analyze in data warehouses or other systems.

To adopt a data lake approach, enterprises should take a full step toward multipurpose (rather than single purpose) commodity cluster computing for enterprise-wide analysis of less-structured data. To take that full step, they first must acknowledge that a data lake is a separate discipline of endeavor that requires separate treatment. Enterprises that set up data lakes must simultaneously make a long-term commitment to hone the techniques that provide this new analytic potential. Half measures won't suffice.

* Timothy Prickett Morgan, "Cluster Sizes Reveal Hadoop Maturity Curve," *Enterprise Tech: Systems Edition*, November 8, 2013, http://www.enterprisetech.com/2013/11/08/cluster-sizes-reveal-hadoop-maturity-curve/, accessed March 20, 2014.

## *To have a deeper conversation about rethinking integration, please contact:*

Tom DeGarmo
Global and US Advisory Technology Consulting Leader
+1 (267) 330 2658
thomas.p.degarmo@us.pwc.com

Chris Curran
Chief Technologist
+1 (214) 754 5055
christopher.b.curran@us.pwc.com

Michael Pearl
Principal
New IT Platform Leader
+1 (408) 817 3801
michael.pearl@us.pwc.com

Bo Parker
Managing Director
Center for Technology and Innovation
+1 (408) 817 5733
bo.parker@us.pwc.com

## *About PwC's* Technology Forecast

Published by PwC's Center for Technology and Innovation (CTI), the *Technology Forecast* explores emerging technologies and trends to help business and technology executives develop strategies to capitalize on technology opportunities.

Recent issues of the *Technology Forecast* have explored a number of emerging technologies and topics that have ultimately become many of today's leading technology and business issues. To learn more about the *Technology Forecast*, visit *www.pwc.com/technologyforecast*.

## *About PwC*

PwC firms help organizations and individuals create the value they're looking for. We're a network of firms in 157 countries with close to 184,000 people who are committed to delivering quality in assurance, tax and advisory services. Tell us what matters to you and find out more by visiting us at *www.pwc.com*.