*An eBook with big data stories and stats*

# MOVE IT DON'T LOSE IT

## IS YOUR BIG DATA COLLECTING DUST?

ATTUNITY

# CONTENTS


Big Data Growth Like a Runaway Train 2

Getting Insight from Everywhere: The Growth in Big Data Sources 3

Big Data, Big Value 3

Use It or Lose It: The Importance of Fresh Data 4

Moving Big Data Old School 5

Pervasiveness 101 - How to Be Everywhere at the Same Time 5

Making ETL More Efficient 6

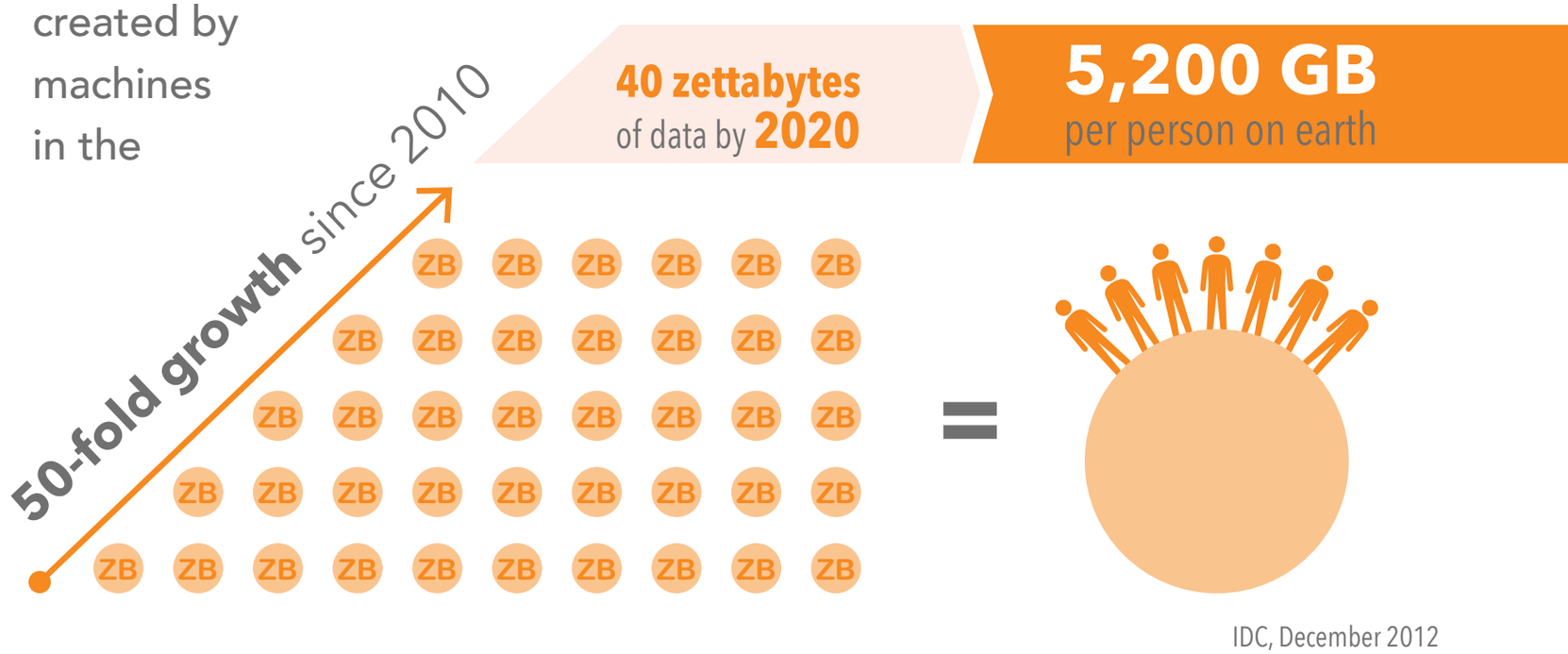
The Reality of Real Time Data 7

# **BIG DATA GROWTH** LIKE A RUNAWAY TRAIN

The growth in big data is awe-inspiring. By 2020, IDC projects that the digital universe will reach 40 zettabytes of data. That's 5,200 GB of data for every person on earth. Given that growth rate, it's more challenging than ever to gain real-time insights using that data, especially when we factor in an expected 40 percent compound annual growth rate in global data generated through 2020.

On top of the data volume challenge, the "big" in big data is not matched by big budgets. For example, McKinsey anticipates growth in global IT spend during the same period will only be 5 percent per year.

Where is all this data coming from? In large part, data is being created by machines in the physical world, many of which are not natively "digital" and relative newcomers to the Internet. The universe of connected devices—virtually anything with an "on" switch—is the "Internet of Things."

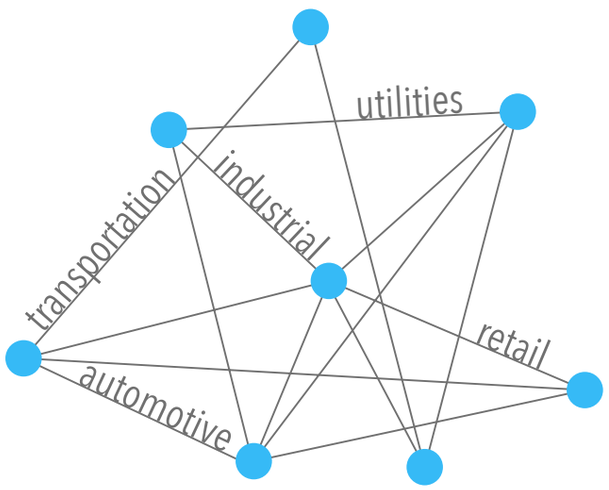



IDC, December 2012

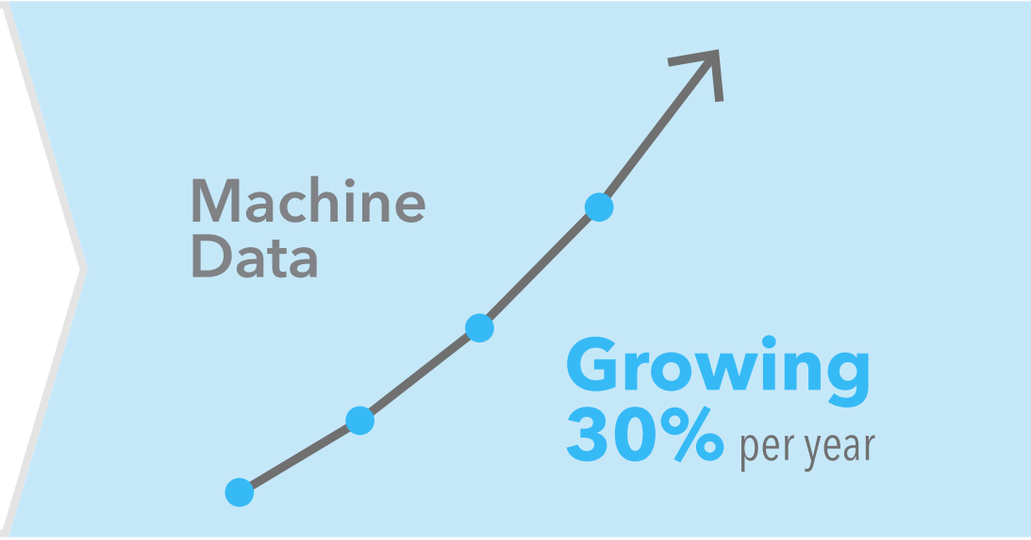
Utility meters, transportation signals and control devices, automotive parts, industrial machines and radio-frequency identification (RFID) tags on a huge range of consumer goods—all of these "things" now send out data about themselves. There are more than 30 million connected sensor nodes, and that number is growing 30 percent per year. That's in addition to the mobile devices that have become ubiquitous in our daily lives.

Many businesses are making substantial investments to get a handle on all of this data because of the enormous opportunity to learn from it. It's safe to say that if your business doesn't get on top of the deluge of big data soon, you'll be buried under it before you know what happened.

In addition to this rapid, runaway growth in individual data sources, the number of data sources itself is multiplying, as described next.




> **30** Million Sensor Nodes




McKinsey, June 2011

# GETTING **INSIGHT** FROM EVERYWHERE: THE GROWTH IN **BIG DATA SOURCES**

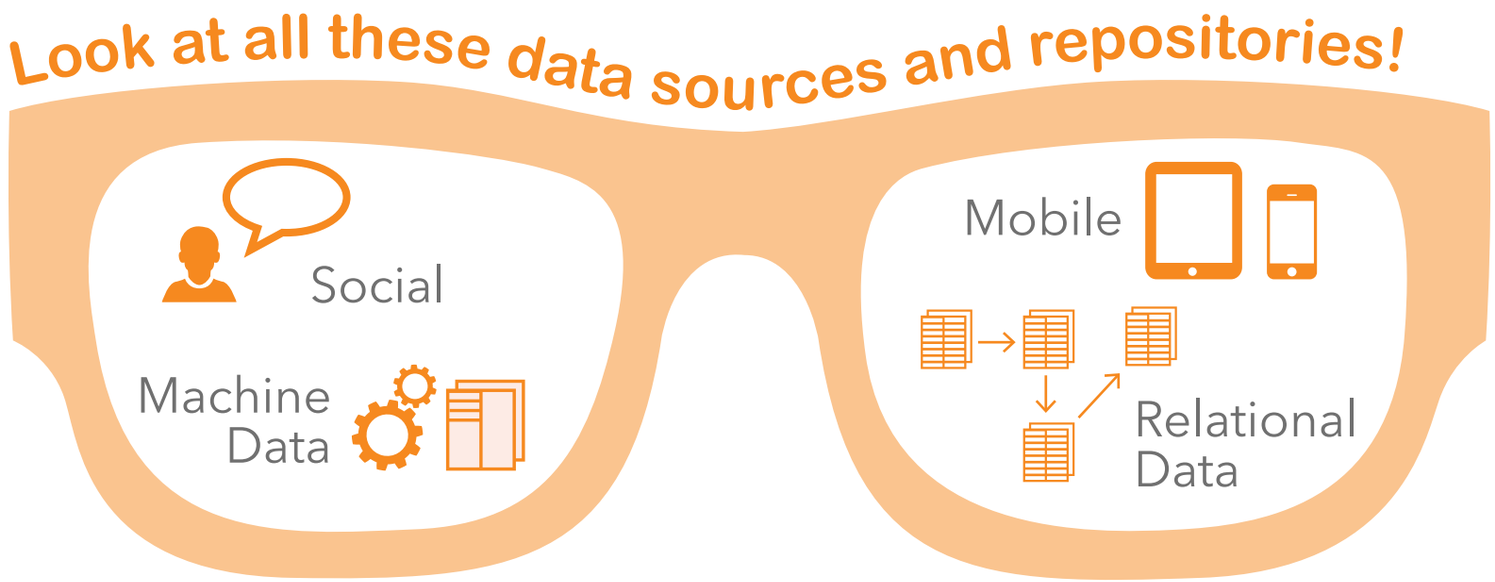
At the same time big data is seeing runaway growth, so are the number of data sources, both internal and external, from which companies are gaining insight. It's no longer enough for most companies to draw from the same internal data stores they've been relying on for years in traditional relational databases – customer receipts, manufacturing manifests, inventory, and the like. So much big data now comes from outside the business, and it has to be sorted, cleansed, analyzed and reconciled with traditional data to generate useful insights.

The new sources include now-ubiquitous smartphones and tablets, machine data (both from the Internet of Things and web log files), weather data, consumer sentiment data, and social media, to name a few. Even five years ago, most of these sources were not widely used. Making things more complicated, much of this data is unstructured and was never meant to be consumed by traditional business intelligence (BI) systems, which rely on rigidly structured data and "cubes" – pre-determined queries – to function.

Businesses have found some success with distributed file systems such as Hadoop – which has no requirements about the structure of data before storing it – but that only gets users halfway to the insights they

need. Hadoop functions well as a pool for catching heterogeneous forms of data, but it is not a complete analytics solution. For example, consider the need to create a multi-channel analysis of customer interactions, which requires correlating information from systems of engagement (clickstream and textual comments stored in Hadoop) with information from systems of record.

Therefore, many companies prefer to move the data to highly parallelized computing engines like Teradata, HP Vertica, Pivotal and other big data warehouses that are built for performing robust analysis. But moving big data to and from these platforms is where it gets challenging. Luckily, in the short, recent history of big data analysis, new types of optimized solutions such as real-time data loading and replication have developed to overcome these hurdles and enable swift movement of big data. These trends are changing and leading companies are leveraging them to gain competitive advantage.




**Many sources were not widely used 5 years ago**

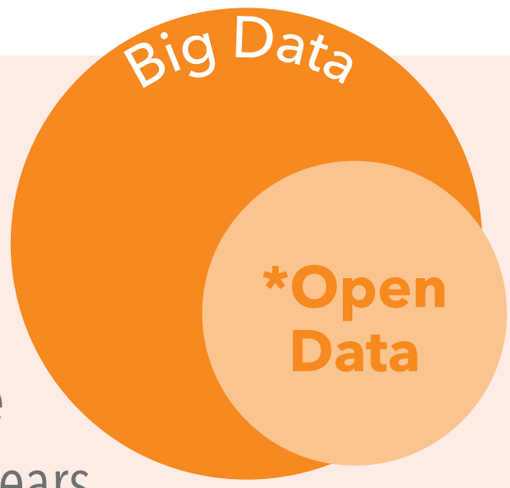
# BIG DATA, **BIG VALUE**

Big data projects are under scrutiny these days, and with good reason: nearly half of them fail.

The key question is whether they drive business value. Enterprises have gotten fairly adept at capturing big data, but moving it to turn it into value has proven difficult. That said, there is good reason to try: the opportunity to improve insights, drive decision making, and even boost revenue as a result is tremendous.

Businesses have the potential to reap $3 trillion in business value from open data in the next few years. "Open data" is non-proprietary data that is free for anyone to use, reuse, and redistribute—statistics on transportation, electricity, healthcare, and economic trends are included in this group, as is the vast world of tweets. Retailers, as an example, stand to improve operating margins by as much as 60 percent using big data, according to McKinsey.

Businesses will get **$3 trillion** in business value from **open data** alone over the next several years




McKinsey, October 2013

McKinsey, June 2012

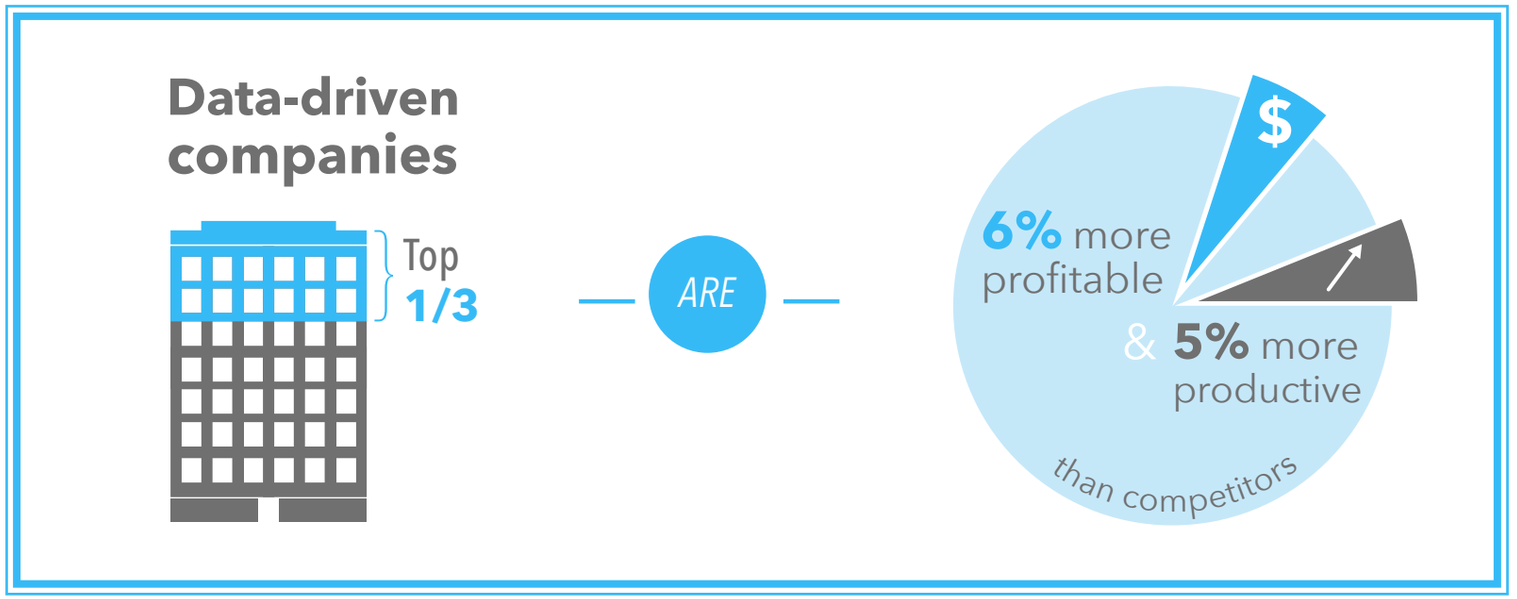
**60%** potential increase in

**RETAILERS'** Operating Margins WITH **BIG DATA**

Case in point: Harvard Business Review has found that the top third of data-driven companies are 6 percent more profitable and 5 percent more productive than their competitors.

Want to know how your company can get into the top third? Much of it comes down to having great business instincts, but to level out the playing field, the best thing most businesses can do is improve their insights and secure the right solutions to make big data easier to analyze. Without adequate data replication tools, a business simply cannot get data where it needs to go, for both storage and analytics. When data gets where it needs to go, businesses can take control of their analytics and start delivering on the big promises of big data.

One key to data's value is its freshness. Like yesterday's weather, the value of information is reduced the longer that you wait to use it, a topic addressed in the next section.



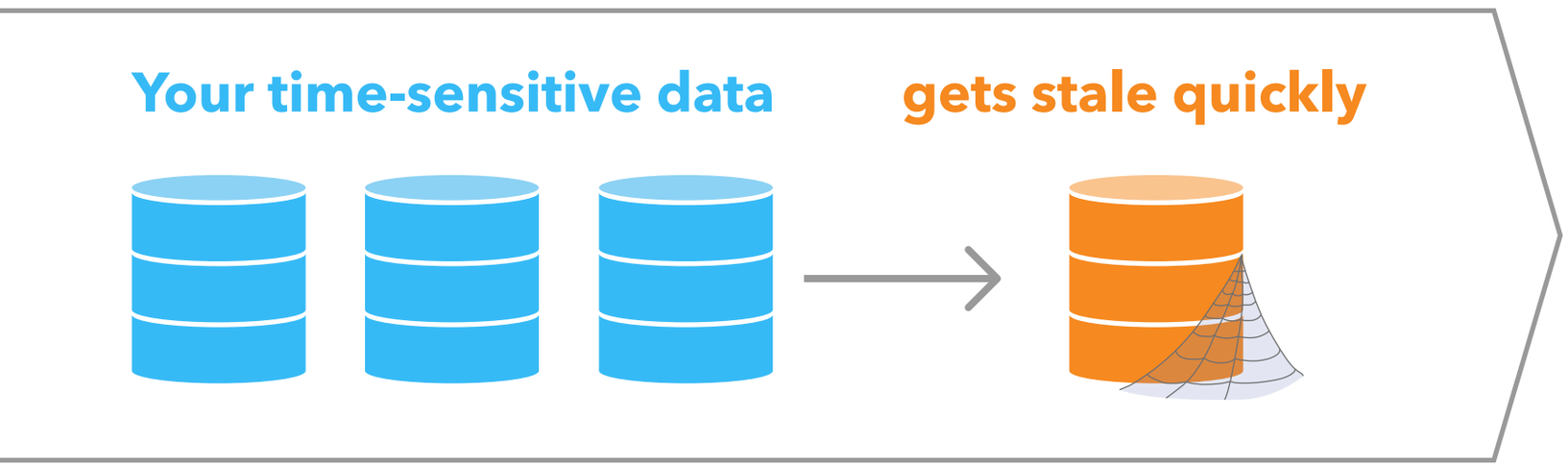

HBR, October 2012

# **USE IT** OR **LOSE IT**: THE IMPORTANCE OF **FRESH DATA**

If you can't analyze it fast enough, data can lose relevance and value, and opportunities can pass you by. In a real-time universe like the one we have today, no business wants to be stuck analyzing yesterday's news. But that is in fact what happens.

IDC reports that only 3 percent of potentially useful data is tagged as having potential at the time it is generated or received, and even less is actually analyzed. Some industries dispose of up to 90 percent of the data they generate, simply because they have no fast or reliable way to store, cleanse, sort, and analyze it.

So how does a company like Wal-Mart, which offers "everyday low prices" and backs it up with guarantees, manage to do this? By continuously monitoring price data from sources all over the world, and matching or lowering its prices accordingly. To do this, data must not only be analyzed, but captured and stored in the right place, in as close to real time as possible.



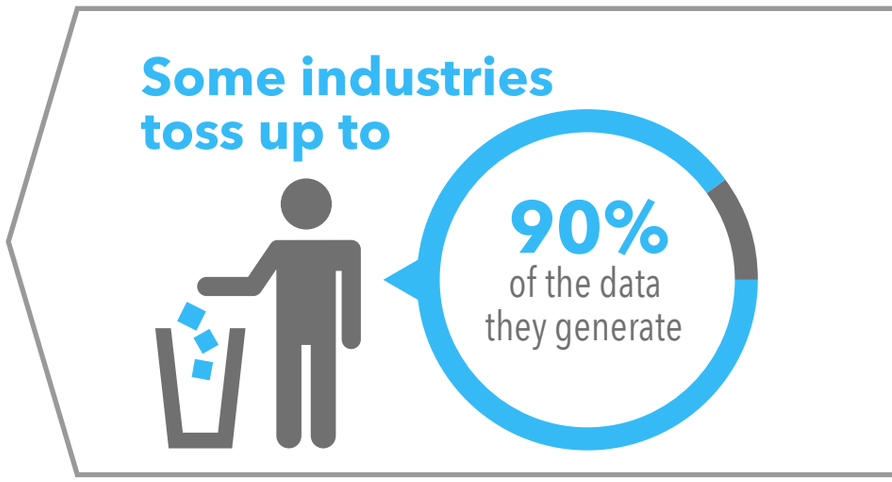
With the proliferating number of sources and targets for storage and analysis, many companies are still using Extract, Transform and Load (ETL) platforms, which is a 20-plus-year-old concept that is time- and resource-intensive to set up, manage, and monitor. The batch processes inherent in this technology result in much data becoming stale by the time it gets to its destination, rendering it close to useless in most competitive situations. Minimizing ETL when possible and maximizing the ability to quickly move only the changed data directly where and when it is needed are keys to ensuring fresh, real-time data, higher business efficiency, and increased productivity (see "Making ETL More Efficient" for more information).

To make fresh data available to the right stakeholders, many businesses are using a hybrid data infrastructure, with some data kept on premises and some in the cloud, which offers greater scalability and affordability. But to take advantage of the cloud, you must move your data there.

Read the next section to find out how slow cloud-seeding and moving data in the cloud can be, as well as a way to accelerate the process.

Only **3%** of potentially **useful data** is **tagged;** even **less is analyzed**

IDC, December 2012



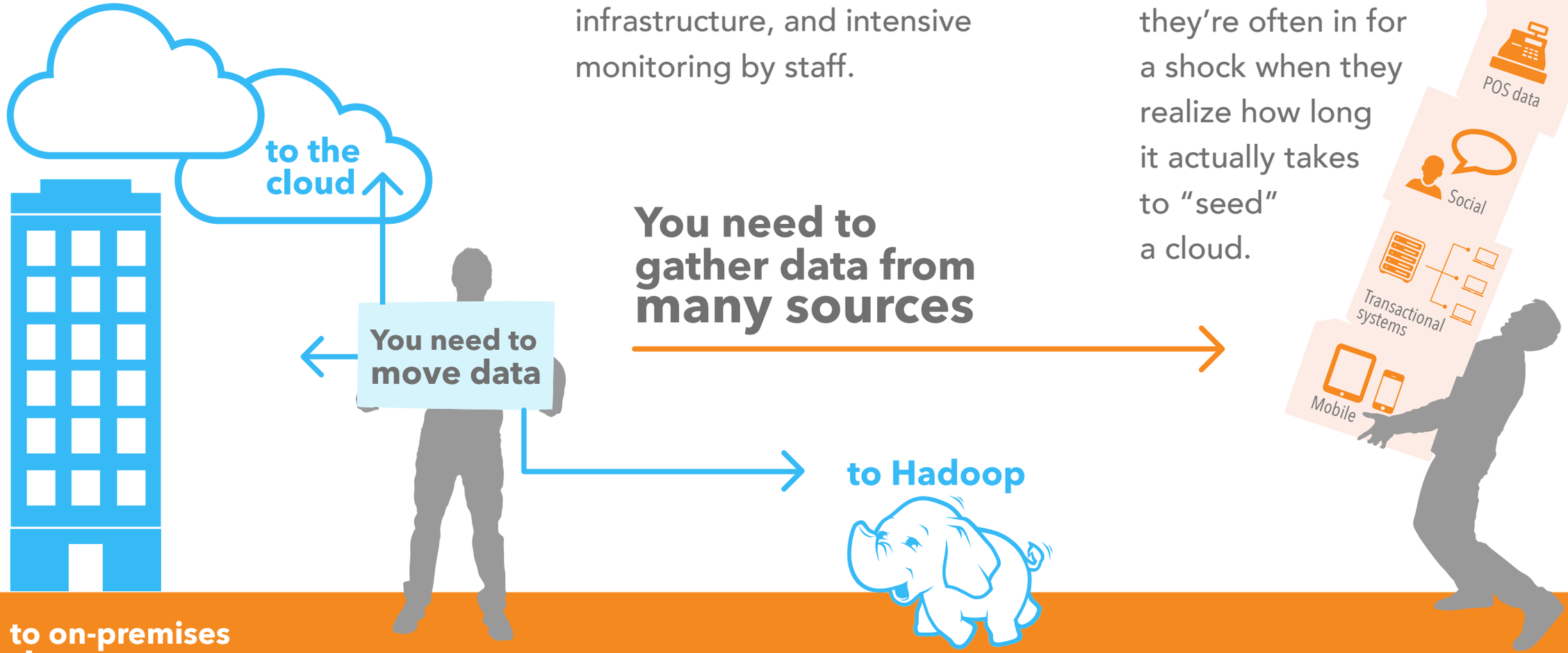

McKinsey, June 2011

# MOVING BIG DATA **OLD SCHOOL**

Moving data to the cloud is oftentimes a surprisingly complex and slow process.

Likewise, bulk transfers on-premises require a large investment of time, purchase and maintenance of infrastructure, and intensive monitoring by staff.

However, the versatility of the cloud—flexible pricing, scalable capacity, ubiquity—appeals to many businesses. But they're often in for a shock when they realize how long it actually takes to "seed" a cloud.

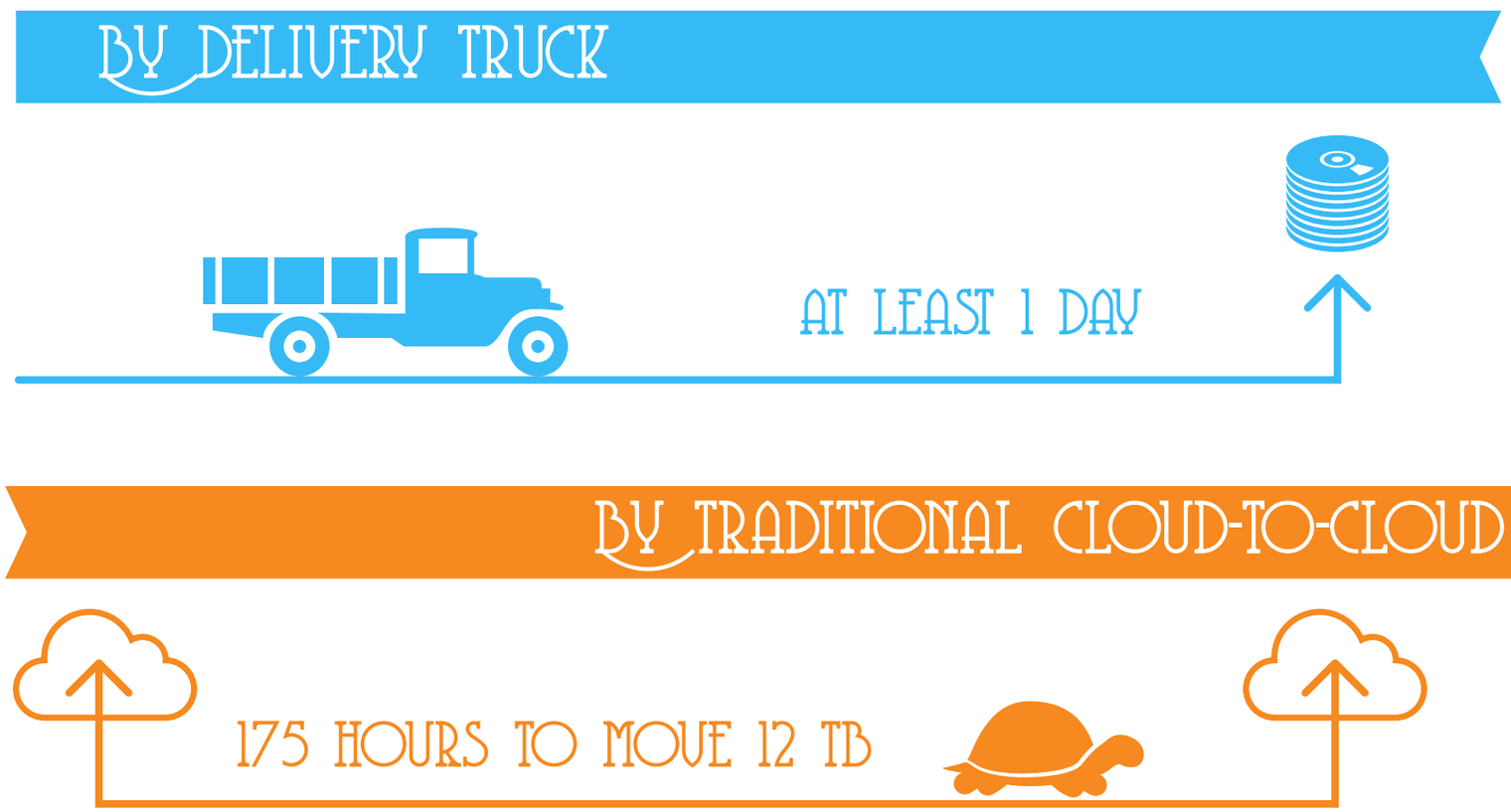


Traditionally, the best method available for loading a database to the cloud has been to send physical media—discs and tapes—from one location to another by delivery truck. That 'SneakerNet' approach takes at least a day, which is simply not sustainable for today's leading companies.

Traditional cloud-to-cloud transfers are often worse than the "truck-it" approach. It can take as much as 175 hours to move 12 TB of data from one cloud to another, according to Nasuni, an enterprise cloud storage company.

Not all data is moving to the cloud. Some of it goes to on-premises data centers, and some goes to distributed file systems like Hadoop. That's a full-on air-traffic-control operation. Worse, even when the data gets to the right place, cleansing and sorting it for analysis using traditional ETL methods is time consuming and requires expert resources to initiate, manage, and monitor that effort.

Nasuni, December 2012



## Moving Big Data New School

So that's the bad news about moving big data "old-school" style. There is another way—call it "new school" for convenience.

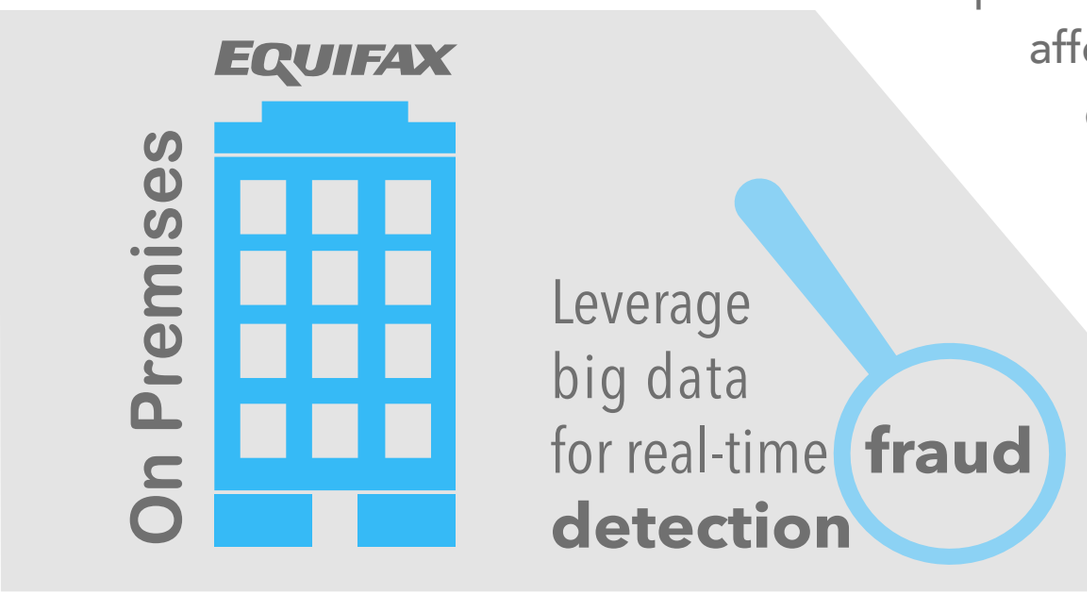
Etix, the largest independent ticketing company in North America, needed to move data from on-premises Oracle databases to the cloud-based Amazon Redshift data warehouse. Initial estimates were that this operation would take three months and $80,000 in labor. Instead, Etix used Attunity CloudBeam and was able to load its data in just a few minutes.

Because Etix can now maintain an up-to-date data warehouse for real-time analytics and identify actionable marketing data in minutes, the company was able to realize competitive advantage in just a few clicks.


Welcome to the speed and ease of moving data "new school." Next we'll look at some real world implications of this approach: the ability to synchronize data fast across multiple stores, providing a virtual way to "be everywhere at the same time."

# PERVASIVENESS 101 - HOW TO **BE EVERYWHERE** AT THE **SAME TIME**

You might hear people joke about needing a clone so they can be in two places at once. But there are cases where businesses really do need the latest data almost everywhere at the same time, and that's a real challenge.




Nowhere is this truer than at a credit agency like Equifax, which is bombarded with data on millions of people and their associated transactions all day long. Its employees constantly monitor data feeds for indicators of fraud or changes in purchasing behavior that would affect credit. It's absolutely critical that fraud be detected and requests for credit scores be serviced in as close to real time as possible. In this highly competitive market, the financial security of Equifax's customers, and the company's business, depend on it. A credit agency must have real-time access to mortgage, banking, credit card, and other financial records, all of which are operated by different partners using different formats, housed in databases thousands of miles apart.

In order to keep up, Equifax had been performing nightly manual batch loads from a multitude of sources into a data warehouse, where all of

a customer's transaction data could be compared. As one might imagine, this is a labor-intensive process that is always a race against time. Using Attunity, Equifax has established a real-time data flow that enables real-time credit decisions, as well as the ability to quickly extract key insights from big data, provide accurate and timely information to stakeholders, and increase its competitive advantage.

Such leading-edge approaches hold great promise, but they don't eliminate the need for ETL. The question then becomes to how to make ETL more efficient, described next.

# MAKING **ETL** MORE **EFFICIENT**

In data-intensive businesses, there has always been a necessary evil—moving and syncing the data. Formatting issues, duplicates, and errors need to be weeded out before any real analysis can begin. The traditional approach to this is through ETL.

ETL takes about 70% of data warehouse development time. It's not a real-time process by any stretch of the imagination, and it requires expert staff to monitor that process. While there is no way to avoid the need for properly formatted data, it turns out that far less of it has to go through ETL than has traditionally been thought. Not every application process requires complex structures, multi-dimensional analysis, and other aspects of data cleansing that traditionally fall under the umbrella of ETL.


ETL CONSUMES 70% OF DATA WAREHOUSE DEVELOPMENT TIME

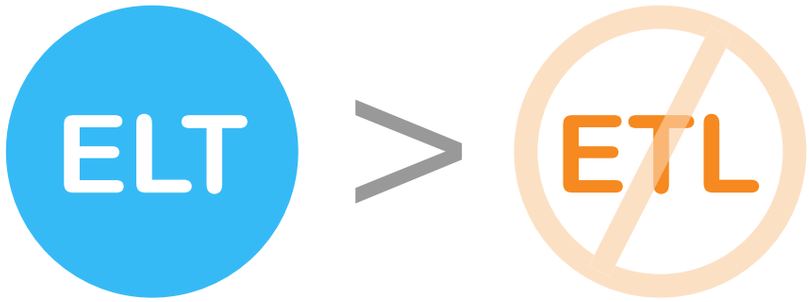
Understanding Cryptic Schemata in Large Extract-Transform-Load Systems, 2012

ETL REQUIRES EXPERTS

## Minimize ETL; Adopt ELT

Attunity and its customers have been able to identify that for the majority of use cases, most data does not have to be run through the ETL process. They have instead adopted the ELT process, an approach which focuses on getting the data to the target as quickly as possible and processing any required transformations using the powerful engine of today's multiple parallel-processing (MPP) data warehouses. This model significantly compresses preparation time and costs. With their massive memories and MPP capabilities, data warehouses can actually perform a lot of the tasks, including transformations, that had typically been run through ETL platforms. The advantage is that ELT is much faster and more efficient, and it enables businesses to better leverage their initial data warehouse investment.

The roadblock for organizations had always been to better understand which data can skip ETL, then moving the data fast enough into these data warehouses so that it can be analyzed effectively. By using Attunity's massive data-transfer capabilities and its ability to handle minor, SQL-lite transformations, companies can identify data that can skip the ETL process as it arrives, which enables them to greatly reduce the amount of ETL needed. They can avoid the opportunity cost of detouring data through the ETL process and wasting the expensive and powerful data-warehouse computing power that many of them already have.

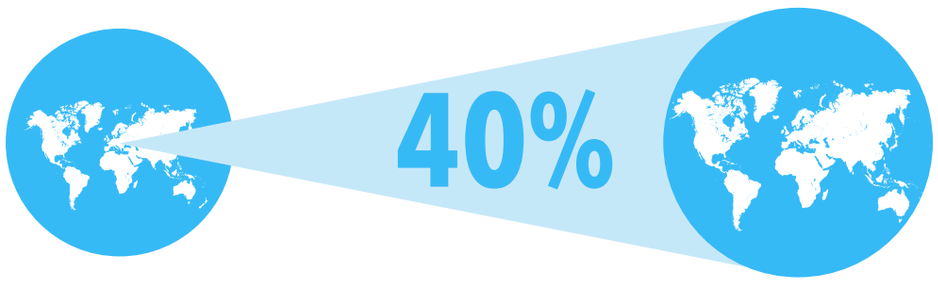


# CONCLUSION: THE REALITY OF **REAL TIME DATA**

In this series, we've talked all about how you should use your data, not lose it.

The growth in big data is awe-inspiring, and the number of data sources that companies are integrating is multiplying as well. But to get the value from big data, you need to be able to analyze it and analyze it fast, because freshness is critical across many businesses and use cases. If you can't move data where you need it in a timely manner, it quickly becomes irrelevant.

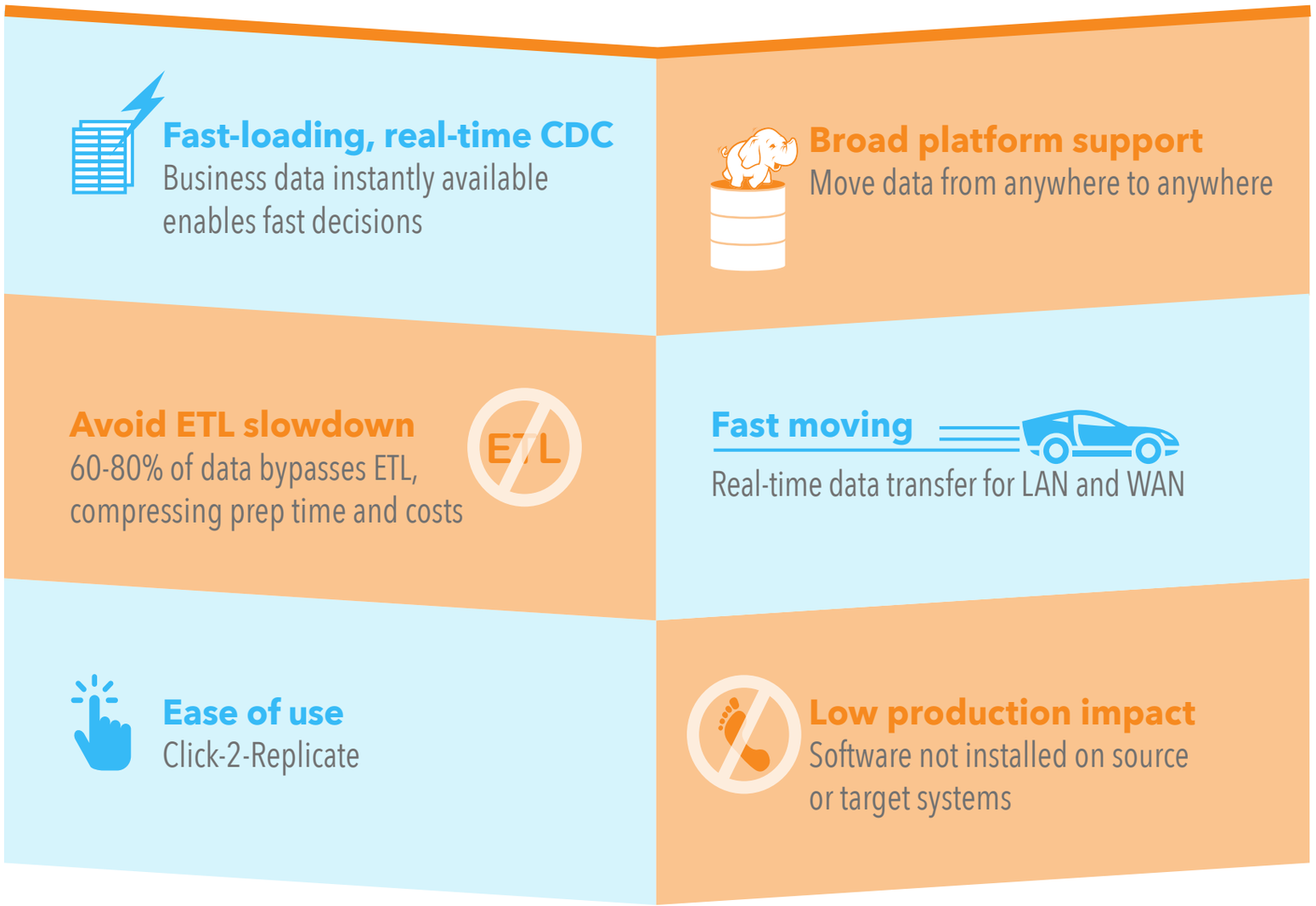
In a world where the press constantly talks about data speed and growth, little is said about how difficult and time-consuming it has been to move and replicate large datasets. It's surprising to realize that cloud seeding, a process that sounds electronic, is actually achieved using an overnight delivery truck laden with data stored on media. The efforts, costs, and expertise involved in loading data via traditional ETL are well known, which is driving the move to ELT, saving the transformation for the data warehouse to achieve higher efficiencies and faster time to value.




Projected **growth** in **global data** generated per year

McKinsey, June 2011

Attunity is in the business of making sure that enterprises have all the data needed where it's needed in near real time. Moving data 'new school' means:




**Fast loading, real-time change data capture (CDC)**, so the latest data is just where you need it, even if that means at thousands of locations distributed across the globe.

**Broad platform support and one-click replication.** Attunity supports myriad databases as well as nonrelational data stores. Users can set it and forget it using an easy Click-2-Replicate design.

**Real time data transfer for LANs and WANs**, with no software to install on source or target systems.

To highlight just a few customer use cases and successes:

### Visibility around the globe

Kongsberg Maritime wanted real-time visibility into ships at sea around the world; the question was how to achieve it. By partnering with Attunity, Kongsberg was able to create a "global platform designed to provide the full picture of an ongoing, seafaring operation in real time, which is essential for safety, troubleshooting and decision making," according to Jon Fredrik Lehn-Pedersen, Vice President Drilling and Advanced Vessels at Kongsberg.

### Moving to the cloud is a breeze

Attunity helped Etix, the largest North American web-based ticketing service provider, load its data warehouse in the cloud in four days instead of the three months it would have taken using other methods.

### Being everywhere at once

Equifax must essentially gather data from everywhere around the world at once to synthesize the latest transactional data on consumers in near real-time. Attunity helped Equifax move from a manual, time-consuming nightly batch process to creating a real-time data store that it leverages to build accurate, up-to-the-minute credit profiles and prevent fraud.

Given the ability to move data in real-time and replicate databases or changes with a single click, what new business ideas and use cases can you think of? We look forward to discussing your needs and hearing your feedback about this series.

# FOR **MORE** INFO


866.288.8648

sales@attunity.com

www.attunity.com


Follow Attunity on

Read the Blog ⟩⟩

ATTUNITY