Rethinking Data Integration in the Cloud: A Revolutionary Approach

December 2010

The Browser Has Evolved into Something New: The Best Way to Manage and Integrate Your Data in and Out of the Cloud

The World is Moving to the Cloud

By all accounts, cloud computing adoption is exceeding high expectations. In Gartner's 2009 CIO survey, cloud computing ranked #16 as a business priority. In one year, according to Gartner's latest survey, cloud computing jumped 14 spots up the priority list to #2, behind virtualization. "Gartner predicts that within two years 80% of Fortune 1000 enterprises will use the cloud at some level."¹ For many companies, cloud use will be at significant levels: Gartner has estimated that 30–35% of the IT workload would move to the cloud over the next five years, but the actual rate is outpacing that estimate.²

Government is behind, but not by far. For example, the U.S. Federal CIO, Vivek Kundra, launched the Federal Government’s Cloud Computing Initiative in September 2009. The U.S. Government is the world’s biggest IT buyer, spending over $76 billion annually on more than 1,100 data centers, 10,000 systems, 24,000 web sites, and 272,000 data sets on data.gov alone (up from just 47 data sets a year ago). Clearly, there’s gravitational pull to the U.S. initiative, which will accelerate the general move to the cloud.

Compelling Benefits are a Powerful Draw

Businesses and governments around the world are moving to the cloud because of its clear and compelling benefits—flexibility, speed, savings. But what’s not quite as clear about the cloud is … exactly what it is. Definitions abound, none widely accepted. Maybe this is because it’s tough to define something that’s changing before your eyes. To paraphrase John Lennon, the cloud is what’s happening while we’re busy making plans. We’re building it, but not to a blueprint. The major cloud providers began offering cloud services to exploit the large-scale infrastructure they’d invested in prior to the cloud; hence, there’s much heterogeneity among cloud offerings but few agreed-upon interfaces. Meanwhile, the “cloud ecosystem” continues to evolve and expand, encompassing private, public, hybrid,
and federated clouds, all in various stages of evolution, along with various types of cloud services including infrastructure-, platform-, and software-as-a-service.

**But There are Concerns as Well—About Data Portability and Integration, for One**

Even as organizations are drawn to the early cloud by its benefits, there are important concerns about data security and portability. Concerning portability, for example: What if you want to change service providers? Once your data is in the cloud, how difficult will it be to get it back out? What if you want to move data hosted in a cloud in Europe, say, to another region? Closely related to data portability is the challenge of keeping your data integrated across multiple applications inside and outside your firewall. Will there be “drift” between your on-premises data and your data in the cloud because of the barriers to integration? Or will you lose control of your off-premises data entirely, since you no longer have direct access to or control over the database behind those applications?

You need to maintain control of your own business data. That’s fundamental. The challenges involved, however, multiply in a hybrid world of private and public clouds, where your data lives both on-premises and in the cloud, possibly in multiple clouds. Some of this complexity is illustrated in figure 1.

Actually, figure 1 tells only part of the data complexity story. Besides being spread across enterprise and SaaS applications and other cloud services, your business-critical data is increasingly dispersed around the web and in databases; scattered across extranets, intranets, XML and FTP sites; and locked in Flash, Excel, PDF and Word files. How can you maintain control of all of your disaggregated data—and not only that but manage and integrate it with your business systems rapidly, in real-time, in the dynamic way that business actually happens?
Traditional Data Integration is No Match for Cloud Complexity

If you’re like many other businesses, your entry point to the cloud was a subscription to SaaS such as sales force automation (SFA), customer relationship management (CRM) or enterprise resource planning (ERP). Consider the task of migrating content from your on-premises system to a new SFA service. Or migrating your data from the SFA service back into one of your internal systems. Or integrating data across SFA and CRM services from different vendors. Basic, essential business decisions that should be simple enough to carry out, right?

You need to be able to do these things—and many more—to maintain control of your data. And you need to be able to do them rapidly in order to have full confidence in, and reap the full benefits of, the cloud.

There are three ways you can move and integrate data across applications, as shown in figure 2:

1. **At the back-end using data integration.** You can reverse-engineer the database and the business logic that ties the data together—or hire an army of consultants to do it for you. Either way, it’s extremely time-consuming and costly—and simply not a viable solution in view of the number and velocity of data integrations needed in a distributed, hybrid world. Furthermore, this approach typically does not work for SaaS applications, for which you have no way to access or control the underlying database.
At the application logic layer using an application program interface (API). This one sounds promising at first glance. After all, SFA, CRM, etc., vendors provide APIs, don’t they? Many do. That may account for some of your data integration needs; but beyond that, APIs simply can’t keep up with the complexity of the hybrid cloud. Moreover, APIs are increasingly rare in proportion to the exploding number of data sources on the web and in the cloud. Of more than 227 million active websites, for example, only about 2,200 have published APIs—fewer than one in 10,000. APIs might be a viable solution for the few data sources and services that have them, but for the vast majority of your data integration needs in the cloud, they are not a complete solution. Even for cloud services that have them, however, APIs may not be the best data integration tool. Naturally, SaaS vendors want to create customer “stickiness” and so their APIs are not designed to make it easy for customers to pack up and leave. APIs typically come with a steep learning curve for your engineers in dealing with the complexities of data migration and management between your in-house systems and the cloud-based services, which can stretch out to months and take a heavy toll on business agility. Case in point: One SFA vendor told a customer to call in two months when done migrating data from the internal system to the new service; as an alternative, the vendor offered to handle the migration on a consultant basis—for over $40,000. There is a better alternative than APIs that require two months or $40,000 of consulting to migrate data from one application to another, and that is:

At the presentation layer using browser-based data integration. Just about every application and data source you need to extract data from is moving to the web. Your HR tools, CRM apps, sales data, partner price lists, supplier inventory, and so on—they’re all accessible through a web browser. So are those 227 million active web sites and 24,000 U.S. government web sites we mentioned earlier. With browser-based data integration, your engineers will never again face the common problem of business users viewing data in their browser that the API was not designed to and cannot access—if an API exists, that is. Of course, we’re not talking about your standard browser here. Internet Explorer, Safari, Firefox and Chrome do a fine job of presenting data at the presentation layer. But when it comes to extracting, transforming, integrating and migrating data, they are simply not the right tool for the job.
A New Kind of Browser Platform Designed for Data Integration in the Cloud

The Kapow Extraction Browser is the first web browser purpose-built for secure, automated data extraction, transformation, integration, and delivery on a massive scale. It brings together the functionality of a presentation browser (so that you can browse the web as you normally do) and an extraction browser that automatically and precisely extracts and integrates data from virtually any source on the web, in the cloud or across the enterprise faster, cheaper and more accurately than any other method.

The Kapow Extraction Browser eliminates the need for APIs by leveraging the user interface of applications it encounters. In the cloud, web-based as it is, all data and business logic is available in the user interface—which means that you can access and extract any data you target. The Kapow Extraction Browser then automatically transforms the data into meaningful information using business rules, migrates the data where you want and integrates it with your existing business systems.

The Kapow Extraction Browser Automates Rapid Data Integration

The idea is to have a browser that can be tasked with performing the same workflow that a human can do, and therefore act as an agent for automated data extraction, transformation and integration.

How does this work? Even though a web page may look unstructured—information in the middle, a few related pieces on the right, a menu on top, a disclaimer at the bottom—it really is quite structured, but much of the structure is hidden away from the user. Web pages are a mix of logic and presentation, and the browser is simply presenting the page the way the web developer planned for it to be viewed by a human. The logical structure behind the web page is known as the Document Object Model (DOM), sometimes referred to by developers as the DOM tree because it’s shaped something like a tree. To integrate successfully with a web page, the automated agent that is performing the workflow must be able to operate on the DOM tree to trigger the same events as if a human were clicking on the mouse or typing in text. Moreover, the agent must be able to leverage “hidden” semantic information in the DOM, e.g., about the location of a login field or a title field.

The Kapow Extraction Browser is automated by instructions provided by programmatic robots—powerful scripted agents that automatically collect data and/or repeat data entry, data transformation, and business rules processing in real time. The Kapow robots mimic a user interaction step-by-step, automating the processes required to navigate and act on the DOM tree and extract the relevant data with all the metadata and context of input values and selections required to transform the data from a variety of formats, for example, to a single, standard format. After extracting the data, the Kapow browser can integrate and deliver that data into any format, such as an API/web service, SQL database, or back into enterprise or cloud applications via their web front-end.
Conceptually, all of these steps are the same steps that a human would go through to cut and paste data from one web application to another in a normal web browser. Usually, it’s not a single human doing this task in a particular company, however, but a number of employees performing the repetitive drudgery on a regular basis. By automating these steps, Kapow delivers data integration projects as much as 90% faster and 80% more cost-effectively than traditional methods and, by eliminating human error, improves data accuracy up to 100%.

As an illustration, The Wall Street Journal recently reported on Audi’s tremendous acceleration and cost-savings in data integration using Kapow: “Audi’s product managers used to check each of the 20 [data] sources individually and copy the results into spreadsheets. While the task was time consuming and inconvenient, Audi’s information-technology department didn’t want to pay as much as $500,000 to combine all of the data into one application using traditional software-integration techniques…. [But with Kapow] one developer was able to build the new application in just four days for a fraction of that cost.”

Some have raised a concern that, given the dynamic environment of the web, automated agents designed to integrate with particular web applications might quickly become obsolete when the particular web apps changed. However, when an agent is designed to interact with the logic and hidden semantic information, rather than the presentation of a web app, most changes in presentation do not affect the agent’s programmed ability to operate on that web app and integrate with its data. Only rare, major rewrites of web logic can disrupt the agent—for example, if new fields are introduced to a submit form.

Kapow is the Leading “Browser-Based” Enterprise Data Integration Platform for Rapid Data Delivery

The Kapow Extraction Browser is the engine of the Kapow Katalyst Platform, a revolutionary data integration solution for rapid data automation and delivery. Kapow Katalyst makes millions of potential data sources available to your business through browser-based data integration. Kapow also enables you to integrate with raw data at the database level, and with data logic at the application level using an API, for when these approaches are best for your particular data integration needs. As a complete data integration solution, Kapow uniquely integrates with any layer in the application stack—the presentation layer (JavaScript or HTML), the application layer (API, SOAP/REST/XML), or the database layer (SQL)—and automates the processes of extracting, transforming, integrating and migrating data from virtually any source in any of those layers to virtually any destination.

The Kapow Katalyst Platform includes a complete integrated development environment (IDE) that enables non-technical users to build robots with click-and-drag ease, no application coding required. With RoboServer, the highly scalable execution engine of Katalyst, you can deploy armies of automated robots across the web for massive automation of data integration or API-enablement processes. No longer will your employees be forced to cut and paste data repetitively from web pages, and no longer will you have to wait for the owner of the data source to supply a useful API, or for your engineers to reverse-engineer it over months. And, with the web-based Management Console, you can easily and collaboratively govern, monitor and manage the Kapow Katalyst RoboServer environment and data integration mechanisms.
Nearly 500 customers, including AT&T, Audi, ESPN, Intel, Morgan Stanley, Vodafone, and dozens of federal agencies rely on Kapow for data integration in the cloud as well as the complete range of data integration needs: web data syndication, mobile-enabling web sites, automated content migration, data for business intelligence, voice of the customer initiatives, comparison pricing—and much more.

Kapow’s revolutionary approach is proven to deliver data integration projects as much as 90% faster and 80% cheaper than traditional methods, with 100% data accuracy. For example: That data migration challenge mentioned earlier? The one that required two months, according to the SFA vendor, or more than $40,000 in consulting fees? Kapow completed the migration over a single weekend.