› White Paper

Data Integration Déjà Vu: Big Data Reinvigorates DI

sas
THE POWER TO KNOW®

Contents

About the Author

Olivier Penel is Principal Business Solutions Manager for Data Management at SAS where he leads the European and Asian Data Management and Governance Center of Excellence. For more than 15 years, Penel has consulted with Global 1000 companies across various industries in the areas of data governance, big data, master data management and data quality, helping them to drive value from their data. Penel has an engineering degree in human-computer interaction.

Data Integration, Again. Really?

Information technology evolves rapidly. But that doesn't always render existing technology extinct.

Think about communication channels, like radio, television and Internet, for example. How many people thought that TV would supplant radio and that the Internet would render both TV and radio useless or redundant?

That hasn't happened. Radio still educates and informs, but in a different way than in 1970. Television may have started with just three networks in the US, but now it's much more expansive. The Internet opened many new options for sharing information, but plenty of people still listen or watch through radio or TV. Think about streaming movies, satellite radio and set top boxes. Instead of dying out, old technologies often manage to coexist.

That's what has happened with data integration. Rather than being done primarily on a batch basis with internal data, data integration (DI) now needs to be implicit in everyday business operations. It needs to work with both indigenous and exogenous sources, while operating at different latencies, from real-time to streaming. Let's take a look at how data integration has gotten to this point, how it's continuing to evolve and what organizations must do to keep their approach to DI relevant.

Data Integration Adapts to Change

Data integration started way back when organizations realized they needed more than one system or data source to manage the business. With data integration, organizations could combine multiple data sources together. And data warehouses frequently used data integration techniques to consolidate operational system data and to support reporting or analytical needs.

But things kept getting more complex. When it became clear that the huge number of applications, systems and data warehouses had created a smorgasbord of data that was challenging to maintain, enterprise architects started to create smarter frameworks to integrate data. They created canonical models, batch-oriented ETL/ELT (extract-transform-load, extract-load-transform), service oriented architecture, the enterprise service bus, message queues, real-time web services, semantic integration using ontologies, master data management and more.

After all this time and with all these mature technologies in place, why would we still need new data integration paradigms? Why do organizations keep investing in this software?

It comes down to these three trends:

- Increasing numbers of indigenous and exogenous data sources that organizations use for competitive advantage, including social media, unstructured text and sensor data from smart meters and other devices.
- Unprecedented rate of growth in data volumes.
- Emerging technologies like Hadoop that expand beyond the reach of traditional data management software.

These trends have put tremendous pressure on existing infrastructures, pushing them to do things they were never intended to do. Bound by inflexible techniques in the face of big data, many organizations find it nearly impossible to make full use of all their data. On top of that, they need to keep an eye on the emergence of logical data warehousing, the necessary cohabitation of integration patterns, and the new capabilities required to support those requirements – such as Hadoop, NoSQL, in-memory computing and data virtualization.

Why Big Data Adds Big Complexity to Data Integration

Among all the trends that have affected data integration, the biggest game changer is big data. Big data is swiftly escalating data integration challenges. Why?

- With big data, differences among various data structures become much more significant.
- The consolidation of external data sources means organizations have little to no control over data standards at the source.
- Volumes and velocities are increasing exponentially, pushing both systems and processes to their limits.

We must **rethink** how organizations manage data. And we must redesign our information management strategy to match. Let's see how this is playing out today.

Internet of Things

According to Gartner, there will be nearly 25 billion devices on the Internet of Things (IoT) by 2020.[1] But those devices are already creating enormous amounts of continuously flowing data.

Think of remote patient monitoring, predictive asset maintenance, smart energy grid, location-based promotions and smart cities (building and traffic management). These are just a few previews of how IoT will change the world we live in.

The most pressing challenge now is to find economically viable ways to store all of this streaming data. Cloud and Hadoop platforms are some of the more promising answers. Another challenge is the ability to process (through analytics) this data in real time, to capture near-instant insight from the data. Here, new techniques like event stream processing can analyze data before it even hits the data store, identifying patterns of interest as the data is being created.

New Generation of Customer Intelligence

Customer care has been a longtime business focus, for obvious reasons. With customer relationship management (CRM) applications, businesses can improve the customer experience across channels and propose products and services customers are likely to buy. CRM, along with master data management solutions, are attempts to build a single view of customer data. That single view can improve marketing campaign efficiency, drive better retention, generate new cross-sell and up-sell opportunities, and shed more light on things like customer lifetime value.

What changes with big data is that organizations now have an opportunity to build a more complete and accurate view of customers by incorporating totally new data sources. Consider social media or web forums, or existing data the organizations already had but couldn't handle very well – such as emails and phone recordings.

With new data sources, organizations can:

- Do sentiment analysis for customer retention or product development, based on customer feedback.
- Conduct real-time marketing, which enables them to quickly identify customers who matter the most.
- Make next-best offers at the point of interaction or send tailored recommendations to mobile devices based on owners' locations.

Due to the volumes involved, the cost required to store all this additional data and the unstructured nature of the data, traditional enterprise data warehouses are not suitable for handling this new complexity. To use these new data sources for advanced customer intelligence, we clearly need new data integration techniques.

New Regulatory Requirements for Preventing Fraud and Reporting Risk

Financial institutions are being pressured like never before to tighten up their fraud prevention and risk management frameworks. And new risk reporting regulations are created every day.

What regulators are asking banks to do presents many data integration challenges:

- Risk reporting must now be done in a way that banks are not really prepared to do. Risk data aggregation must be performed at the enterprise level, integrating risk data across all departments, lines of business and countries.
- Banks must be able to recalculate entire risk portfolios in minutes rather than weeks. Regulatory reports, as well as third-party risk assessments, must be generated in real time based on live data. This requires a level of flexibility beyond the scope of current data infrastructures.
- Finally, banks are required to measure reports' trustworthiness based on the quality of the underlying data. This implies that they can establish the lineage of the aggregation process and measure data quality according to predefined standards.

In terms of fraud detection and prevention, financial institutions must be able to identify fraudulent behavior patterns based on transactional data in real time. They need to be able to detect fraud networks. And, of course, they need to stop fraudulent transactions on the fly.

Handling this highly volatile data in real time – so they can take immediate action – requires new data integration techniques.

Data Monetization

Partly fueled by the IoT, data monetization is now a concrete way to use valuable data assets to create new revenue channels. This is true for telecommunication and media companies, retailers, financial institutions, communication service providers and other industries as well. The main question for these companies is how to conform to privacy issues and regulations while making money with all that data.

[1] gartner.com/newsroom/id/2905717

The usual challenges still exist – sharing data between different organizations and consolidating internal and external data. But data integration applied to data monetization initiatives raises a whole new set of issues:

- How to share data while keeping control of it.
- How to ensure that security and privacy requirements are clearly defined and followed.
- How to manage the appropriately granular level of access rights.
- How to ensure that the governance framework and tools can effectively define what is and is not acceptable, control how data is shared and monitor how it is used.
- How to speed up data integration to allow near-real-time decision making.

Those challenges call for us to rethink our existing data integration paradigms and toolsets.

Additional Cost Optimization and Process Efficiency Pressures on IT

As always, the pressure is on IT and lines of business alike to reduce the cost of operations. Big data brings new potential to this area. Here are a few examples.

Price and stock optimization. Data plays a critical role in generating growth through price and sales effectiveness. Incorporating big data brings the potential for much deeper insight.

Delivery optimization. Route optimization is nothing new for big players in the logistics or shipping industries, but GPS data – as well as sensor data – offers new ways to optimize all sorts of things. Consider vehicle maintenance, mileage cost, self-improving route optimization, customer service and more. Fleet telematics and advanced analytics may take route optimization to the next level. But being able to effectively integrate and prepare the massive amounts of data being generated today is the underlying condition for success.

Predictive asset maintenance. This capability unleashes a massive opportunity for cutting costs in industries like oil and gas, manufacturing, logistics and telecommunications. But it poses serious hurdles for data integration. That's because it entails proactively gathering and analyzing vast amounts of data from sensors, aggregating this data with historical data, and being able to identify patterns to give early warnings and take preventive actions.

What are some of your greatest data management challenges?

In a 2014 CIO MarketPulse Survey[1]:

47% of companies cited complex and numerous data sources.

37% cited the lack of a unified view of corporate data.

[1] sas.com/resources/asset/131157_DM_Infographic_Final.pdf

IT infrastructure. When it comes to reducing IT costs, there are now big opportunities to store data inexpensively and to reduce the workload on technical resources by empowering nontechnical users. Big data ecosystems such as Hadoop provide a cost-effective way to store data compared to traditional data warehouse appliance servers. This is especially true as data volumes get larger. Hadoop can also be deployed on cheap commodity hardware for data processing and storage – and the software is cheaper than traditional database software. Hadoop also opens the door for business users or data scientists to manipulate and extract insight from big data without IT intervention. So technical resources won't need to handle ad hoc reporting or data-related queries.

New Challenges for Your Data Management Strategy

The disruptive effect of big data from a data integration perspective is clear. At this point, IT departments are back to the drawing board, trying to figure out how the promises of big data can be realized and what it means for their data management strategy. Three areas are especially critical to information strategies: data access and storage, metadata management and big data governance.

Data Access and Storage Plus Real-Time Access and Delivery

The vast amount of data involved in big data initiatives means organizations must find cheaper ways to store data so they can complement the existing data warehousing infrastructure. Traditional relational database management systems (RDBMS) are not necessarily an economically viable option.

Organizations are pushed to find ways to avoid the cost and complexity associated with traditional data integration techniques when dealing with large varieties of data sources and formats. For example, they must accommodate sources like operational applications, web and social media, sensors and smart meters, along with formats including file-based, voice recordings, relational databases and event stream data.

The ability of Hadoop to handle schema on-read (as opposed to schema on-write) provides the needed agility to quickly bring new data sources on board without having to shoehorn inadequate formats into a predefined data model. Hadoop can serve as:

- The next-generation data warehouse, to augment or supplement the traditional RDBMS.
- A new data store for new data types – particularly the unstructured data that an RDBMS cannot handle – and for new data sources, such as web, social network and sensor data.
- A data lake, to stage all the organization's available data in a minimally processed state.

Data access is traditionally conditioned to predefined data models, predefined data integration flows and predefined reporting models. Any change requires IT involvement, which often means long turnarounds on design, implementation and testing. But to keep pace with competitors, businesses need access to data in real time. Only then will they have the flexibility to extract valuable insight from the data, when it's needed. Techniques like data virtualization and self-service reporting make this possible.

Organizations need to be able to use data as soon as it is produced (or available), so employees can make decisions in real time and take action as soon as an event occurs. To do it, they must be able to analyze data streams on the fly, before the data even hits a data store.

Metadata Management

Traditional metadata management consists of developing a logical data model to describe how databases are related. This resolves inherent inconsistencies associated with data silos, and enables data sharing for reporting or analytical purposes.

But with increasing numbers of data sources, including sources that are not in the control of the consuming organization, it's increasingly difficult to manage metadata proactively. Besides, with the schema "on-read" principle in Hadoop, the format of data being loaded may be unknown at the point of entry. Eventually, that metadata has to be defined so the data can be shared and understood.

> Data professionals are used to thinking of metadata as having technical value but little worth outside of the IT organization. However, sometimes **metadata may be as valuable as the content** itself and can be used in the same way as big data. For instance, in the case of the PRISM privacy controversy in the US, the NSA went after call detail records instead of the calls themselves.[2]

[2] washingtonpost.com/news/wonkblog/wp/2013/06/12/heres-everything-we-know-about-prism-to-date/

With big data, it's not practical to map and try to make sense of every bit of data. Instead, organizations need to focus on:

- Data source mapping, meaning and relevance rather than data models.
- Semantic metadata applied to a selected number of business-critical data elements.
- Defining business terms and owners and relating that to technical metadata.

In turn, those who use the data will become responsible for providing useful business definitions of what the data is and does.

Big Data Governance

One of the major challenges for data integration in the context of big data is to establish and sustain the right level of governance. And it's not all about technology. Key issues like data quality, data privacy and security, relevance and meaningfulness must be considered at the enterprise level.

Energy Leader Transforms Data Into Customer Intelligence

Enerjisa, which serves 9 million customers, kept customer data on different systems and in different formats for data cleansing and analytics purposes. To enable various business areas to operate more efficiently, Enerjisa needed to create a single source of customer data that would be easy to access for all departments for advanced analytics purposes. Using SAS Data Management and SAS Data Quality, Enerjisa:

- Increased the rate of successful customer contact.
- Reduced total records by 25 percent.
- Increased completeness of records by 30 percent.
- Established a foundation for customer intelligence and analytics.

Let's look at this a little deeper. Linking to new data sources, especially for external sources and unstructured data, will put data out of reach for typical data governance programs. In other words, standards and data quality will no longer be controlled at the source.

All the same, trying to enforce traditional levels of quality for big data might annihilate the anticipated benefits of big data initiatives related to rapid data integration and handling data streams in real time. There is clearly a balance to be found between the data quality imperative and the benefits of big data velocity.

Bringing vast amounts of data into a data lake will raise questions around privacy regulations and security. Do we have the right to store this data? For how long? For example, on Facebook, if a user likes your page or friends you, you can store their information. But as soon as they unfriend you, you're required to remove that information from your repository.

Who should have access to data? And how are we allowed to use it? The data governance body must address those questions by defining the rules and monitoring their application across the organization. Metadata management and data lineage are great techniques to help organizations comply with privacy and security requirements.

Business glossaries are another method that can be used to store business terms like "profit" or "customer" and relate them to technical metadata like fields or reports. In this way, users can see how changing a field in a table will affect other data sources, targets, analytical models or reports downstream.

Even when there is no requirement to retire data from a storage perspective, we still need to manage the data life cycle to keep the focus on relevant data. That will avoid extraneous noise and prevent data lakes from becoming data swamps.

New Data Architecture Paradigms: How SAS® Can Help

When it comes to choosing ways to decouple data from consuming applications and processes, there is no silver bullet. Each organization has to adopt the integration paradigms and techniques that are best suited for it. There are several options:

- Data virtualization and logical data warehousing offer flexibility and fast deployment, augmentation of traditional integration architectures, and more.
- Data can be delivered with different capabilities (Hadoop, NoSQL, in-memory, etc.).

- High-volume data streams can be processed in real time.
- Data services can be delivered via the cloud – for example, integration platform as a service.

Hadoop, but Not Hadoop Alone

Although Hadoop is one of the key components of modern, big data-enabled infrastructures, it clearly falls short on its own when it comes to data management.

Hadoop brings a lot of value in terms of cheap data storage and distributed data processing. It's also fault-tolerant and scalable. But it's not mature enough to effectively manipulate data without specialized (and rare) skills, and without requiring a huge amount of custom development in MapReduce, Pig or HiveQL.

SAS® Data Management and Hadoop

The data management platform from SAS includes several components that are fully enabled for the Hadoop ecosystem:

- **SAS Data Management**, which provides an intuitive GUI, templates and a data transformation library for writing Hadoop programs in Pig, Hive and MapReduce. It also has a GUI for loading data into the SAS® LASR™ Analytic Server for visualization.
- **SAS/ACCESS® Interface to Hadoop and Impala**, which includes out-of-the-box connectivity between SAS and Hadoop using Hive and supporting other Hadoop languages, such as MapReduce and Pig.
- **SAS Data Loader for Hadoop**, which gives business users self-service data access, transformation, cleansing and profiling capabilities; allows for self-service data preparation and reporting; and pushes code processing down to the cluster for better performance and governance.

The answer is to have a modern data management platform that can abstract complexity. This type of platform can also reuse existing skills and data integration assets (such as data quality validations and data transformation flows) across Hadoop and traditional data warehouse systems.

To make sure Hadoop does not become another data silo that's isolated from the broader enterprise data management infrastructure, it's important to establish metadata lineage. Organizations also need to ensure consistency of data security rules across the whole enterprise data landscape, including Hadoop. The data management platform should work seamlessly across Hadoop and traditional RDBMS, and should provide:

- Access to Hadoop Distributed File System for loading from/to Hadoop.
- Embedded data governance (including business glossary, metadata management and granular security management).
- Embedded data quality (including profiling, monitoring and data quality transformation such as parsing, standardization, matching, etc.).
- Data preparation for analytics (aggregate, pivot, transpose, etc.).

Data Virtualization and the Logical Data Warehouse

We've known the limits of traditional data warehouses for years. For one, they take a lot of time and money to build and maintain. And in the era of big data, it's no longer practical to replicate data and structure each data mart to answer predefined queries. The notion of the warehouse as the single, monolithic "version of the truth" for reporting and analytics has proven ill-equipped to deal with today's massive variety and volume of data. And business users are simply dissatisfied with traditional data warehouses. They often provide the wrong level of data granularity and timeliness, and they aren't flexible enough to accommodate ever-changing business requirements.

With big data, new technologies came into play, such as Hadoop clusters and NoSQL databases. Now it's clear that these new ways of storing data are not going to replace the traditional RDBMS. Instead, they will extend or complement the RDBMS for cheap data storage and parallel processing.

In response to the notion that Hadoop could possibly become yet another silo of data, a few years ago Gartner came up with the concept of a logical data warehouse (LDW), one way to achieve data virtualization. The idea was to provide an enterprise data layer that presents a unified view of multistructured and unstructured data assets across organizational silos.

This shift moves from the concept of central repositories and data models to the concept of data services, data processing and access engines. An LDW provides a virtual data layer from both traditional and emerging data sources.

SAS Federation Server simplifies data access, administration, security and performance by creating a virtual data layer without physically moving data. This frees business users from Hadoop environment complexities. So now they can view data in Hadoop and virtually blend it with other database systems like SAP HANA, IBM DB2, Oracle or Teradata. Improved security and governance features ensure that the right users have access to the right data.

Streaming Analytics = Real-Time Data Analysis

Many big data scenarios are based on being able to analyze in real time vast amounts of streaming data coming from transactional systems, sensors, web navigation logs and other sources. In these cases, the traditional method of collecting, storing and analyzing data no longer works. Now we need to be able to monitor a confluence of high-volume data streams in real time, as they happen, to identify patterns and sequences of events, and to generate insight so we can take immediate action.

Instead of running queries against stored data, SAS Event Stream Processing stores queries and streams massive amounts of data through them, filtering, aggregating and detecting patterns in real time. This process occurs before the data is stored, reducing latency of the information being analyzed.

SAS Event Stream Processing can also differentiate between information that's relevant to the business versus information that doesn't matter – storing important information while discarding the rest. In turn, organizations can considerably cut the costs of storage and processing, and alleviate the load on traditional data integration frameworks. Finally, data quality routines like standardization can be applied in-stream, while the data is in motion.

In-Memory and In-Database Computing

In-memory and in-database computing accelerate time to value from analytics. But they also represent a way to reduce data movement and simplify data integration requirements.

By moving the logic to the data (in database), or by loading the data in memory for real-time analysis (in memory), there is no need to shoehorn various data sources into a canonical data model before data can be analyzed. SAS has a variety of solutions for in-memory and in-database computing, including SAS Visual Analytics and SAS Visual Statistics, SAS In-Memory Statistics for Hadoop, SAS High-Performance Analytics, SAS Scoring Accelerator and SAS In-Database technologies, and SAS Data Loader for Hadoop.

Data Integration, Again? Yes, Really.

The increasing number of data integration patterns, combined with high volumes and varieties of exogenous sources, make it imperative for organizations to have tools that help them glean value from their data. Hadoop is no longer seen as the "ETL killer" it was once thought to be. Today, new integration patterns – like event stream processing, data virtualization, and in-memory and in-database processing – have reinvigorated the data integration domain.

Just like the Internet didn't really kill TV, and TV didn't render radio useless, data integration is still alive and well today. In the media industry, multiple mediums now coexist as parallel channels for getting information, news and entertainment. With the exciting potential of things yet to come, there are plenty of reasons to keep data integration, and its numerous facets, near the top of your watch list for the foreseeable future.

Want to learn more?
Visit sas.com/data.

To contact your local SAS office, please visit: sas.com/offices

107865_S142353.0815