Setbacks Help Us March Onward: Unified Data Integration Platforms Power Business Innovation

An ENTERPRISE MANAGEMENT ASSOCIATES® (EMA™) White Paper
Prepared for Informatica
April 2013
Table of Contents

Setbacks Help Us March Onward............................................................................................................. 1
A Work of Art ............................................................................................................................................. 1
Almost There........................................................................................................................................... 1
Modern Production.................................................................................................................................... 2
Applying Modern Production to Data Integration....................................................................................... 2
Early Software Development: Hand Coding ................................................................................................. 2
Learning Curve and Productivity Issues ....................................................................................................... 3
Complexity and Resistance to Change .......................................................................................................... 4
Moving Toward Standardization .................................................................................................................. 4
Isolated Islands of Production ....................................................................................................................... 5
Proper Communication across Complex Tasks ............................................................................................. 5
Bringing Data Integration All Together ......................................................................................................... 5
Data Integration Operational Excellence ....................................................................................................... 6
Make the Best Quality of Goods Possible At the Lowest Cost Possible........................................................... 6
Hand Coding Attempts a Comeback ............................................................................................................. 9
The Only Constant is Change ....................................................................................................................... 9
Setbacks Help Us March Onward

When asked about lessons learned, Henry Ford, an American industrialist and founder of the Ford Motor Company, once said:

“Life is a series of experiences, each one of which makes us bigger, even though sometimes it is hard to realize this. For the world was built to develop character, and we must learn that the setbacks and griefs which we endure help us in our marching onward.”

This approach can be used to gain insight from experiences across a number of fields in history. The general patterns of change and development can be applied from one industry to another. Manufacturing and mass production, that Ford is most famous for, underwent massive change during the 1920s thanks largely to the innovations of the industrial revolution. Since then, manufacturing has continued to innovate their processes with inclusion of management practices, such as Operational Excellence\(^2\), and new technologies such as robotics.

The lessons from manufacturing, a decidedly bricks and mortar affair, have relevance to the data driven, and almost completely virtual, technology domains of the 21\(^{st}\) century. For example, the domain of data integration can learn from modern automobile manufacture and assembly. Key to both operational and analytical aspects of data driven enterprises, data integration should innovate along similar lines to auto manufacturing by using standard control practices, componentization, production automation and incorporating new data management technologies such as NoSQL and Apache Hadoop.

A Work of Art

The automotive industry can provide lessons for data integration projects by looking at the differences between customized builds and productized manufacturing. The Bentley Mulsanne is a handcrafted automobile from start to finish. Each component is hand built, or assembled, by some of the finest automotive builders in the world. The Mulsanne is the pinnacle of the 4-door sedan.

However, this level of quality comes at a price. The $300,000 price tag is just the beginning. A Mulsanne requires 12 weeks and 800 man-hours to build a Mulsanne. This is not to mention the amount of technical skill that it takes to custom build a Bentley. Bentley assembly team members have been trained as apprentices for years in the Bentley style and process before they are good enough to manufacture the Mulsanne, or any of the other models on the Bentley product line.

Almost There

Along the same lines, the General Motors Cadillac CTS is also a 4-door sedan. While not at the same quality level as the Mulsanne, the CTS is an equivalent vehicle with similar attributes. However, at $50,000 a piece, the CTS provides a significant price difference from the Mulsanne. For that price difference, Cadillac produces a vehicle faster, in greater quantities and requires less specialized fabrication skills.

Along the General Motors’ integrated robotic and human assembly line, the CTS can move from a painted chassis to a finished product in a single day. This is not forsaking quality for speed. The assembly teams responsible for the Cadillac are professionals and some of the best of their craft just as the Bentley teams. However, due to the application of existing and cutting edge fabrication technologies, the amount of time to train a member of the CTS production team is much shorter than for the Bentley production group.

\(^1\) [http://goo.gl/5EWwk](http://goo.gl/5EWwk)
Modern Production

The main difference between the Mulsanne and the CTS is the build techniques used to assemble the automobiles. Despite the level of technology in the vehicle, Bentley uses many of the same techniques as the automobile industry used in the early 1900s. In this early era of industrial manufacturing, specialized teams would use elaborate assembly techniques to build specialized autos. Prior to the 1910s, each automobile factory only built a few cars per day.

With the idea of creating a car to meet the needs of the masses, Henry Ford embarked on the development and “perfection” of the production assembly line. In the late 1920s, this technique was used to build a Model-T Ford approximately every 12 minutes. In just under 40 years from when Karl Benz invented the production automobile, auto manufacturing went from custom builds to mass production. Cadillac, with a few technological improvements, uses the same production techniques as the Ford Motor Company in the 1920s. With these techniques, the CTS is available for roughly one sixth of the cost of a Bentley and at six times the annual production.

Applying Modern Production to Data Integration

Data integration solutions share a similar history with automotive manufacturing. At the beginning of the “data integration era,” custom solutions were time-consuming and difficult. Highly skilled craftsmen in the form of data modelers, data (Extract, Transform and Load - ETL) developers and data stewards were required to design, construct and oversee solutions. In the end, results were a relatively small supply of expensive, long to build, custom implementations.

As with the development of the automobile, there were reasons for the time and effort required early in the data integration era. The technologies to move data between operational systems, or from operational platforms to early analytical environments, were developing and maturing. In addition, new business requirements demanded a progression of design disciplines (e.g., data warehouses and marts).

Early Software Development: Hand Coding

To look at the issues of these early custom data integration solutions, we need to look at lessons learned from previous technology periods. Starting with the widespread introduction of software development practices and programming languages in the 1960s, the productivity was similar to that of the early automobile builders. A slow, expensive process produced a small amount of results using highly skilled, and costly, technical resources. However, as the development processes and technology evolved, software development became more productive and more productionalized. Again, an approximately 40-year window existed from original rudimentary software development practices until we reached the productivity of the componentized; object oriented programming languages and the Eclipse integrated development environment (IDE).

http://en.wikipedia.org/wiki/Eclipse_(software)
Setbacks Help Us March Onward: Unified Data Integration Platforms Power Business Innovation

During this evolutionary path, we often wondered why software developers work best at night or exclusively by themselves. Recent notions suggest that custom code requires a special set of circumstances. The thought is that programmers, like many artisan automobile builders with their physical workshops, require a set of conditions to put themselves in a mental “happy place” of productivity. This happy place allows the developers to think about the problem and then start to build their custom code.

In this, it is similar to the workmanship associated with the Bentley. You need to have many of the right circumstances to build a handcrafted automobile – tools, supplies, etc. When developers are interrupted, productivity goes away and they have to start over.

This is why it is often an all-night effort to solve complex problems using the hand coding techniques. In addition, due to the isolated nature of this process, the resulting work product is inadequately documented. Both fellow coders and the management team poorly comprehend the placement of those components in the greater whole of an organization’s implementation strategy. These attributes of uneven production and poor documentation lengthened the software development process and again placed a premium on the work of the individual contributor in the form of the developer.

Learning Curve and Productivity Issues

During the early years, data integration solutions felt their way through productivity issues with results similar to early software development efforts. Hand coding-based options, such as PERL and UNIX shell scripts, were popular avenues. However, both required robust knowledge of those programming languages and operating system tools to make effective use of the strengths of each. The learning curve for these languages could often times be lengthy for new team members. This adversely impacted the productivity and agility of ETL code development. Developers without prior experience with ETL languages had particularly low levels of productivity.

In addition, there were issues that came with the amount of time to maintain those hand coded components. Should a key team member leave the organization, often times it required pulling their data integration scripts apart and relearning the application logic. This comes from the fact that “self-documenting” code rarely accomplishes that self-stated objective. Just ask anyone who has attempted to understand spaghetti PERL scripts. This hand coding approach to ETL jobs lacked both the ability to scale production and the agility to match changing requirements.

**Complexity and Resistance to Change**

Changing data specifications have led to a high level of complexity for analytical environments and their associated data integration efforts. This level of variation has led many data professionals to describe data warehouse and data mart implementations as efforts that are “permanently under construction.” The shifting nature of both data source analysis and target design requirements means that ETL code must be written and re-written and/or assembled in such a way to allow it the agility to adapt quickly.

However, this complexity stands in the way. Complex, hand coded data integration components often suffers from some of the same issues as other software-development, quality issues. These factors include:

- **Tightly Coupled Components** that are less adaptable to new requirements.
- **Large Code Modules** that have multiple functional aspects.
- **Duplicate Code in Multiple locations** that take additional resources to maintain in the event of changing requirements.

Some of these issues can be overcome with standard development practices. Again, this solution relies on the ability and skill of the ETL developers associated with the effort and their adherence to those standards. Overall, these hand coding techniques are slow to adapt to enterprise-wide change. New data sources associated with constantly fluctuating requirements and the process to implement those changes created an inflexible environment that often focused more on the status quo than on innovation.

**Moving Toward Standardization**

The next step in the evolution of data integration solutions was to take a particular tool, or toolset, to replace the hand coded ETL jobs. These tools allowed for the sharing of metadata or code between developers in a repository. These toolsets might be facilities provided by database vendors as packages for data acquisition. Alternatively, these tools could be individual applications for ETL specific functionality. These solutions had the advantage of being able to standardize the management and development of the data integration jobs.

Introduction of toolsets represented an improvement over the custom development process of the artisan ETL developer. Following the automotive example, this would be similar to having separate assembly lines for drive trains, chassis, and interior components. These environments added a level of standardization to components and implementation frameworks.

---


that were not available when everything was put together in a custom fashion. This step of moving from custom builds to standard tools and environments was significant for both automotive assembly and data integration.

**Isolated Islands of Production**

As independent data integration tools are deployed, there is a failure to assimilate metadata and management practice across the complete process. This is similar to the early automobile assembly lines of the 1910s. If there was a breakdown between the availability of parts for one production line, downstream assembly would come to a halt. The use of these individual production lines was an improvement, but not fully integrated.

A similar situation arises with environments based on independent tools. There is a lack of a cross-process, data integration metadata repository and management coordination. Without understanding adjustments to the technical and business decisions made at the data governance level, development teams have difficulty avoiding rework in their ETL components. Without understanding the number and type of new data sources from operational system owners, administration teams have issues managing process workloads and job scheduling. Due to these and other issues, it is important to have a comprehensive source of business and technical metadata in the form of a standardized repository.

**Proper Communication across Complex Tasks**

It is important to lower the barriers to access to this repository. Each of those groups should be able to work with the repository information in the context of their position within the organization. ETL developers should work with metadata in terms of the data sources and targets. Business analysts need to work with the information from a business context standpoint. Data stewards must work with both business and developers to ensure data can be trusted and auditable. Operational teams need to work with the information in terms of tactical operational concerns of system availability and data load times.

With this proper level of metadata communication and project coordination, the level of effort and inherent costs associated with re-work and data quality can be avoided. Using a higher level of coordination and communication than simply “interchanging a written specification documents,” data integration process management improves dramatically. Teams can implement and maintain much more complex efforts than they could otherwise. Associated Information Technology (IT) management teams can also manage the data integration process more effectively. Avoiding isolated metadata and information silos from independent platforms make this more effective management possible.

**Bringing Data Integration All Together**

In the late 1920s, Ford Motor Company refined the assembly process to produce Model-Ts at a rate of one every 12 minutes. This represented the integration of all the components into a single production facility. Not only was the assembly of parts brought under a single roof, the supply chains were also brought together. For example, onsite iron works ensured that too much or too little contributing materials, such as steel, did not influence overall production.

Later, this assembly process improved by creating highly integrated networks of suppliers. These networks avoided a single point of failure or chokepoint in a particular plant. This spread the associated work, and risk, across multiple factories and systems of suppliers. From top to bottom, the entire

---

process of creating an automobile was integrated. Each section of the process has visibility into the inputs and outputs of other components. With this outlook, management receives information on each of the disparate areas to monitor and manage the group as a whole. Using these techniques in 2011, a Ford plant was able to produce a vehicle in less than two minutes. Not the improvement from days to minutes that Henry Ford originally pioneered, but a significant improvement nonetheless.

**Data Integration Operational Excellence**

The modern data integration environment should be combined in a similar fashion. This level of incorporation enables IT organizations to manage by improvement methodologies such as Operational Excellence. By having a combined ecosystem, monitoring and management information is shared across the data integration process. Overall data integration process quality and production can be managed much like how auto manufacturers oversee its supply chain to integrate across multiple suppliers to control margin and production for optimum results.

Data governance processes and data stewards can make decisions and take corresponding actions relating to data quality and metadata. ETL developers can make choices for the productization of re-usable code components that meet the metadata criteria of data stewards and business requirements of end-user business analysts. These components can be standardized for continued reuse. With that, economies of scale can be created not only for the current ETL tasks, but also for future tasks. New data sources can be introduced into the environment without rebuilding from scratch.

**Make the Best Quality of Goods Possible At the Lowest Cost Possible**

When Henry Ford said:

“There is one rule for the industrialist and that is: Make the best quality of goods possible at the lowest cost possible…”

Ford’s focus was serving his customers with the best value. IT departments responsible for the data integration programs of their organizations should have the same goal in mind – “the best quality implementation for the lowest cost possible.” Using the techniques above of integrated platforms and comprehensive metadata repositories, data integration teams can improve the time to implement new data sources and new adjustments to those “constantly under construction” projects. They can also lower the costs associated with ongoing maintenance and upkeep.

Using integrated metadata repositories and development environments that encapsulate the hand coding aspects of development empower development teams to be more productive. Studies have shown that these types of teams have a 2.5x to 3.5x advantage in implementation timeframes. When you consider an average team of three developers and a three-month implementation timeframe, a 3x improvement in productivity would yield $50,000 in just operational labor cost savings over alternative environments for that single project. Alternatively, contemplate the increase in efficiency for that same team across the same three months. For the same amount of man-hours, an additional six months of business requirements could be implemented with the same team members. This would represent $150,000 in additional value to the organization as well as reducing the backlog of data integration tasks using the same team members.

---

9. [http://goo.gl/MrmhD](http://goo.gl/MrmhD)
Considering the common occurrence of underestimating projects and the prevalence of “scope creep,” new projects depending on data integration can be risky efforts. Business stakeholders sponsoring a particular project, and the CFO who provides the funding, require mitigation strategies to limit the chances of failure on a given project. The ability to have a 3x improvement in implementation capacity can be the difference between a project’s success or failure.


Big Data technologies, such as Apache Hadoop and other NoSQL data management platforms, represent new challenges in the world of data integration. They have introduced some of the most interesting opportunities in the world of data management in quite some time. And they have done so in a relatively short amount of time. Unlike software programming and traditional data integration, the rate of change in Big Data technologies such as Apache Hadoop is operating at a quickening pace.

In a little under 10 years, we have gone from the introduction of MapReduce as a processing methodology and Apache Distributed File System (HDFS) to the introduction of a long list of tools and solutions. The impact of the changes from Big Data technologies has changed the way that we look at analytical environments and data integration’s role with and around those platforms.

ENTERPRISE MANAGEMENT ASSOCIATES* (EMA™) analysts have developed a concept called the Hybrid Data Ecosystem.11 In this new ecosystem of data management, business requirements, such as workload, speed of response and economics, are the driving force of integration between the platforms. There is no single solution that meets all of these use case requirements. Big Data sources with multi-structured (e.g. non-relational) formats represent 25–30% of the new data management ecosystem with multi-structured data playing a significant role in the Hadoop, NoSQL and Discovery Platform nodes, per recent EMA research.12

It will be important to use data integration to bring the right information together with the right processing platform. Some use cases will bring pre-processed or aggregated Web translation data from multi-structured environments to traditional analytical platforms for cross-sell/up-sell recommendation generation. Other use cases will pull order information from operational platforms to Apache Hadoop for combination with call center records to improve customer experience.

For example, financial organizations are interested in finding a competitive advantage by adding multi-structured data to their models. Automated financial traders bring in twitter and other social media information, to look for stock trends keyed by real-time customer feedback or influence. This requires the agility to mix both structured relational information from stock exchanges and internal operational platforms with consumer generated social responses, such as twitter. Most of these social platforms are based on JSON, a multi-structure standard, and located in either Cloud sources or in Hadoop and NoSQL nodes of the Hybrid Data Ecosystem.

Omni-channel retailers are another example of organizations needing to incorporate these multiple sources and structures of data. Omni-channel retailers use an integrated strategy for “brick and mortar” as well as online sales. They base their business model(s) on the ability to optimize their pricing and inventory strategies between Point Of Sale (POS) systems and Web transaction information. Again, this requires the mixing of multiple nodes of the EMA Hybrid Data Ecosystem between Operational systems with structured data and Hadoop platforms that have become the de facto standard repository for historical Web log information data.
Hand Coding Attempts a Comeback

The Bentley Mulsanne represents a technologically robust but custom-built automobile. Similarly, Hadoop, and its subprojects, represent new technology with old habits. Using frameworks such as Apache Hadoop to implement MapReduce for high-performance distributed computing might be using the cutting edge of open source data processing and integration technology. However, these new technologies are re-creating the same hand coding issues of PERL and UNIX shell scripts.

Data integration and IT management teams should learn the lessons of their previous efforts. Just as the Cadillac CTS production line has used a mix of standard human elements and automated robots, data integration environments that include Big Data tools need to use a similar approach. A combined environment, such as the EMA Hybrid Data Ecosystem, includes traditional data stores (e.g. relational/structured) along with Big Data (e.g. non-traditional/NoSQL) approaches should be used with a comprehensive and unified data integration platform to achieve similar productivity and cost reductions.

The Only Constant is Change

In the early part of the 20th century, industries focused on manufacturing innovated processes and practices to fuel the Industrial Revolution. Those most successful at modernization changed the nature of their companies. Now, at the early stages of the 21st century, there is a new revolution centered on how to leverage data from both traditional operational information and Big Data sources to power competitive advantage.

As we move into this new “Data Revolution,” the pace of innovation in business, and the importance of a flexible and agile approach to support that change, will increase at a rapid pace. Data integration organizations and IT departments will be at the leading edge of this change. They need to learn from those “series of experiences” Henry Ford spoke of to keep “marching onward.” The lessons from Ford’s innovations will have particular relevance in the Data Revolution. Business stakeholders, in Marketing, Customer Care, and Sales will expect more from their technology partners. Those stakeholders and Finance teams of the CFO’s office will push for faster solutions, reduced cost of ownership and lower implementation risks.

Just as the assembly line broke new ground in the 1920s, pioneering data integration environments for the Hybrid Data Ecosystem will be critical in the 2010s. Adoption of standard management and production practices from other industries will remove some of the “art” from data integration projects in favor of the “science” of modular design to speed design and implementation. Incorporation of new Big Data technologies will forge new data connections between both traditional relational databases and multi-structured data stores. Using hand coded solutions or independent tools to power this innovation will skew delivery times, operational costs and chance of project failure toward that of an earlier era in information technology. The lessons of the past will be lost. The tactical appeal of “no cost” software license NoSQL solutions such as Apache Hadoop will quickly be outpaced by the rate of change required in the Big Data technology era. Innovative and forward looking data integration teams and the associated IT departments will move toward comprehensive and unified data integration platforms to match the expectations of business stakeholders. This will power the revolutionary business model change in this century’s version of the Ford Motor Company.
About Enterprise Management Associates, Inc.

Founded in 1996, Enterprise Management Associates (EMA) is a leading industry analyst firm that provides deep insight across the full spectrum of IT and data management technologies. EMA analysts leverage a unique combination of practical experience, insight into industry best practices, and in-depth knowledge of current and planned vendor solutions to help its clients achieve their goals. Learn more about EMA research, analysis, and consulting services for enterprise line of business users, IT professionals and IT vendors at www.enterprisemanagement.com or blogs.enterprisemanagement.com. You can also follow EMA on Twitter or Facebook.

This report in whole or in part may not be duplicated, reproduced, stored in a retrieval system or retransmitted without prior written permission of Enterprise Management Associates, Inc. All opinions and estimates herein constitute our judgement as of this date and are subject to change without notice. Product names mentioned herein may be trademarks and/or registered trademarks of their respective companies. “EMA” and “Enterprise Management Associates” are trademarks of Enterprise Management Associates, Inc. in the United States and other countries. ©2013 Enterprise Management Associates, Inc. All Rights Reserved. EMA™, ENTERPRISE MANAGEMENT ASSOCIATES™, and the mobius symbol are registered trademarks or common-law trademarks of Enterprise Management Associates, Inc.

Corporate Headquarters:
1995 North 57th Court, Suite 120
Boulder, CO 80301
Phone: +1 303.543.9500
Fax: +1 303.543.7687
www.enterprisemanagement.com