TREXIN INSIGHT PAPER

August 2, 2019

SOME EASY WAYS TO FAIL WHEN BUILDING YOUR DATA PLATFORM

Ideas to help you succeed when addressing a slightly provocative topic.

It's So Easy

It's very easy to fail when building a data platform. Many legacy, monolithic database systems (enterprise data warehouses and the like) and even newly-developed ones take far too much time to build and deploy, cost too much money, and most importantly, fail to deliver significant value (quickly enough) to the business, which was, or should have been, the overriding goal. They can also be overly fragile. For example, a minor change in an incoming data feed can break the data ingestion process. A minor modification to a report can impact other reports. In this post, we'll discuss some of the easiest ways to fail and also propose a new approach and a set of options to consider that can minimize the risk of failure.

How We Got Here

One of the benefits of working on many Client software projects is that we get to see a lot of different approaches. We get to see firsthand what works, and probably more importantly, what doesn't. We have seen software evolve from on-premise, monolithic applications to cloud-based microservices. Unfortunately, in many cases, our data architectures have not kept up.

In addition, often architectures are designed and deployed divorced from actual business needs and the realities of data consumption. The pervasive idea that IT can build an EDW and business will feast on all of its insightful and actionable data is compelling, but in reality, there is a set of critical gaps between having data available on one hand and being able to extract ongoing business value from it on the other.

The Pareto principle states that, for many events, approximately 80% of the effects come from 20% of the causes. Can we find a solution that could address 80% of our business data needs? The answer may lie in leveraging a data analytics accelerator.

Easy Ways to Fail

ONE EASY WAY TO FAIL: BUILD A FRAGILE ETL PROCESS

We worked for a healthcare provider that received monthly claims data from dozens of payers. Each payer had a different format and the Client built custom SSIS scripts to ingest the data. Unfortunately, the payer data was not always consistent and the formats were often updated without notification. This caused havoc on a monthly basis. One or more of the scripts would break and developers would have to scramble to discover, diagnose, and fix the problem. This would delay all the downstream processes which not only included monthly reporting but also impacted billing and collections.

A SECOND WAY TO FAIL: WRITING REPORTING REQUIREMENTS WITHOUT WORKING REPORTS

We worked for a healthcare payer to define next generation business reports. It took us weeks to build a consensus around the first set of requirements. It took us a few more weeks to deliver the first set of working reports. When they were shared with the users, the reports were not what they really needed or wanted.

EVEN MORE EASY WAYS TO FAIL

Here are some additional reasons that analytics efforts typically fail in our experience:

- **You haven't articulated your business questions and issues well enough**. "We need analytics" is not a business question. A prioritized backlog of mission-critical business questions that analytics can help answer is essential.
- **You're not setting expectations clearly with your (analytics) customers**. Analytics is not like traditional software development, or frankly like any other endeavor other than pure statistical research, with which most companies have no experience.
- **You're not going the last mile**. Your analytics program churns out "insights", but the business doesn't know how to turn that into valuable action.
- **You don't have enough valuable data, or don't understand the value inherent in your data**. This is a tough one. Sometimes the truth is that legacy data just doesn't yield much in the way of insight.
- **You don't have the right people to get the most business value from your data**. Is your team ready to adopt an analytics mindset? Do you have the right skills on the appropriate teams?
- **You're missing data quality and/or reasonable privacy/governance programs**. Another tough one. Most companies don't have a good grasp on the importance of establishing a comprehensive set of policies and procedures relating to data quality and governance.

Some Processes That Have Proven Helpful in Mitigating the Risk of Failure

In our experience, there are ways to minimize the risk of failure. Here are some of the processes that have been most successful for us.

CONSUMPTION-DRIVEN DESIGN

Traditionally, design and development of analytical platforms follow an *ingestion to consumption* approach of first defining the architecture and tooling, then considering business needs. Our experience has shown that the design and subsequent development of the platform benefits dramatically from a *consumption led* approach, often making the difference between robust user/consumer adoption and subsequent value on one hand, and creation of yet another warehouse most business users can't meaningfully use, on the other.

With today's options for visualization, reporting, machine learning, and advanced predictive tooling, the effectiveness of the analytical environment and its usage must truly be driven primarily by consumers and their requirements, tooling, and access models, coupled with agile processes described below. Together, these provide a significant input into the architecture of the entire platform.

AGILE DATA MODELING

Reporting and analytics requirements should *iteratively* drive the data model and/or storage buildout. In this kind of macro iterative process, for example, enterprises can start with a data lake (or similar lightly modeled data store) and move to a full-blown data warehouse as reporting and analytics requirements clarify – if indeed a DW is ever needed.

AGILE ANALYTICS

As mentioned above, the needs of analytics consumers play a major role in determining final architecture and related tooling. Key to deriving value from this approach is a continuously-maintained set of (often highly granular) **high-value business questions** stored in a backlog and an agile, iterative approach to prioritizing key business questions and answering them in the minimal amount of time. If executed correctly, answers to those questions provide tangible business value at the end of every (typically 2-4 week) iteration.

This agile approach and the value it provides is in stark contrast to long-term projects which typically take months before providing any business value. In a fast-paced business environment, answers from long-term projects that arrive many months after project launch may no longer be relevant to the key business questions at that time.

Accelerators Can Help

The Pareto principle states that, for many events, approximately 80% of the effects come from 20% of the causes. Can we find a solution that could address 80% of our business data needs? The answer may lie in leveraging a data analytics accelerator.

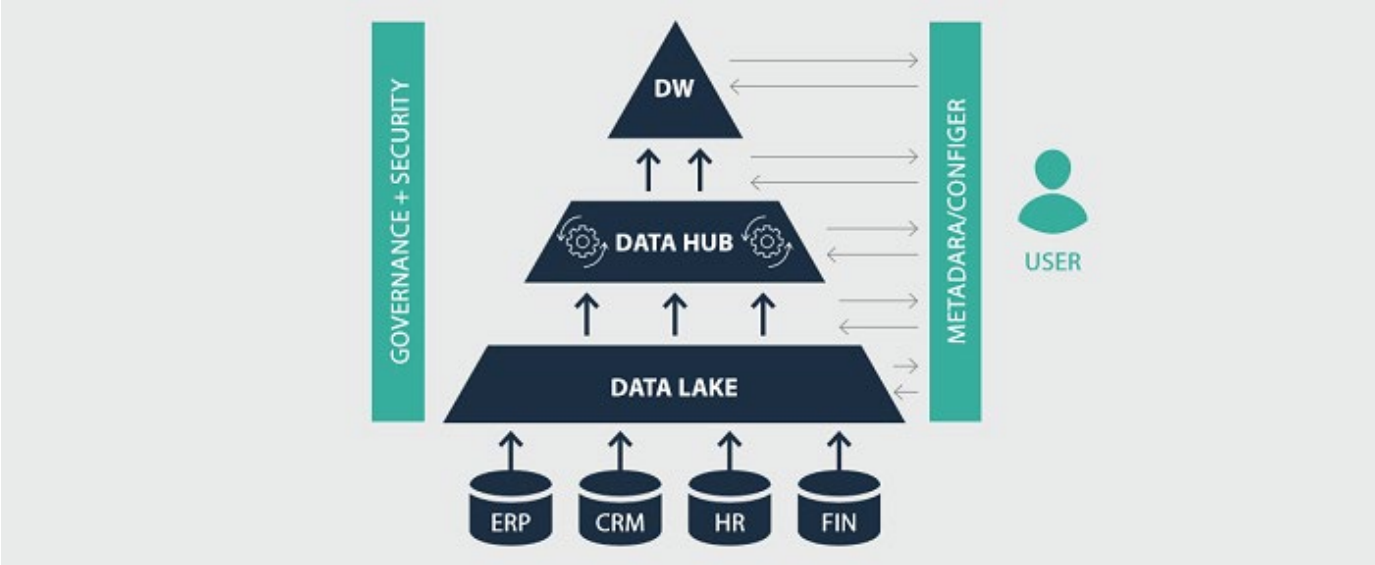
LARGE CLOUD PROVIDERS

Large providers of cloud services are well-known (think Amazon AWS, Microsoft Azure, Google and others). Each of the three named providers offers powerful, cloud-based, end-to-end Machine Learning capabilities that can be quickly leveraged by virtually anyone. Coupled with the highly-scalable infrastructure they offer, in the hands of someone with established data science skills, these providers can offer a nearly turnkey solution for cloud-based analytics.

Having said that, the power and comprehensive nature of the tools large cloud providers offer can make them daunting to people without a data science background. Another caution to consider is elaborated below: those without sufficient analytics and statistics knowledge can easily lead their organizations astray by making false conclusions from the analysis and modeling. This is a caution that warrants serious consideration and honest evaluation of data science skills in your organization.

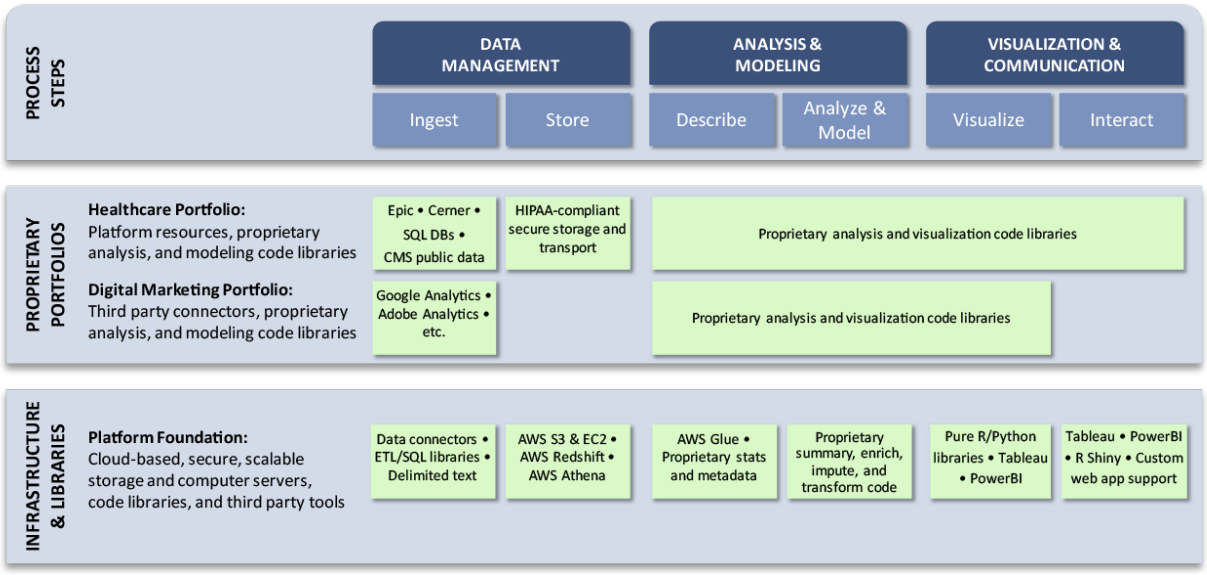
Besides those well-known providers, some other examples of analytics platform accelerators include our own (West Monroe Partners and Trexin) – though there are many more, and new ones created seemingly by the month. This is a trend we expect to continue.

West Monroe's Rapid Analytics Platform (RAP)



West Monroe's RAP is an accelerator consisting of service offering along with a mature toolset that allows Clients to forget about the mechanics of data ingestion and transformation. Users can quickly configure RAP to read in data from a number of sources and load it into a data lake where data scientists can access it. The data hub can be configured to transform the data in the data lake and deliver it to a data warehouse where BI tools can leverage it.

Trexin's Data Science Platform (TDSP)



This platform is designed to accelerate the end-to-end work of data scientists by providing a complete environment based on a secure, scalable infrastructure, including cloud-based storage and computer servers. This is coupled with a library of proprietary code that can facilitate all aspects of the data science process. Libraries support a variety of flavors of data ingestion, data prep, automated basic statistics, and metadata supporting rapid EDA, creation, and evaluation of predictive models and advanced visualization.

The platform is primarily available for Trexin data scientists, but Trexin also offers it to Client data scientists who need a powerful turnkey platform for analysis and modeling. We caution Clients against using this or any platform without sufficient statistics and analytical knowledge in order to avoid costly mistakes in evaluation of analysis and model results.

Final Thoughts

DELIVER QUICKLY AND OFTEN

Don't wait weeks or months before delivering to the business. Identify short-term MVPs, use a lightweight approach, and deliver quickly and often. You need to build confidence in your solutions and in your team. Consider data operationalization separately and get data reporting and analysis working first. It doesn't make sense to operationalize data if it can be leveraged to provide value to the business

USING AN ACCELERATOR: PROS AND CONS

The era of each data scientist and/or analytics group building their own custom infrastructure and analytic code sets is coming to an end. As more and more analysis and modeling is done, best practices patterns emerge and can be encapsulated in platforms such as those mentioned above. While there will be room for many years to come for creativity in coming up with analysis techniques, insightful interpretation of analysis results, beautiful/compelling data visualizations, and storytelling, there's also no need for individual data scientists or analytics groups to reinvent the wheel.

One caution in this arena: novice data scientists (i.e., those without much or any statistical knowledge, coding skills, or even business understanding, sometimes referred to as "citizen data scientists") are now able to easily and quickly construct their own predictive models using one of the platforms above, or many other similar products.

Without a solid understanding of statistics, however, it would be easy for those novice data scientists to draw conclusions not supported by the data, potentially causing an enterprise to invest millions of dollars based on a flawed interpretation and bad statistics. For this reason, we encourage those without significant statistical knowledge to collaborate with a statistician or senior data scientist to ensure interpretation of results is valid.

LOOKING FORWARD WHILE BEING CLEAR-EYED ABOUT THE PRESENT

While we encourage thinking strategically about a time horizon of 3-5 years, the reality for most enterprises is that they must "walk, perhaps crawl, before they can run", establishing fundamental analytics capabilities before jumping into more advanced realms such as Artificial Intelligence.

Having said that, however, there are truly many exciting and valuable opportunities to derive significant value from BI and reporting, and tried and true statistical analysis, enabled by the robustness and quality of many Clients' unique data sets. The future is bright.

This TIP was co-authored by John Crowell, Trexin's Analytics Capability Lead, along with friend and former colleague, Lorenzo De Leon. John welcomes comments and discussion on this topic and can be reached at john.crowell@trexin.com.