Balancing End User Needs with Flexible Data Architectures

By Jared Decker I President and Advisory Consultant, Expert Analytics

September 30, 2019 I Originally published in Enterprise Executive. Republished with permission.

To accommodate the massive influx and wide variety of data that organizations are capturing and storing for analytical and strategic value, approaches that accommodate flexible data structures and file types (such as data lakes) are increasingly common. These architectures require few accommodations when bringing data in but can involve significant processing prior to eventual analysis. In many cases, data from data lakes is combined with data warehouse data, creating an overall flexible data architecture. The goals for organizations on this path are quantity, speed, and flexibility. It is as if *garbage in garbage out*, which has influenced data warehouse design for decades with repercussions of accuracy when not controlling the format and quality of data, has shifted to *limited data limited opportunity* with repercussions of reduced competitiveness when strategic data gathering is limited or slowed. In this divide, end users have much at stake and their needs should be balanced with the organization's strategic data ambitions.

Depending on the type of data or industry in question, *limited data limited opportunity* can be a reality. But to be realistic, we should switch out the word *opportunity* for the word *prospecting* to more accurately describe the process of getting value from flexible data architectures. This is true

since organizations may seem better-positioned due to increased incoming data velocity and flexibility, but analysts and decision-makers are not necessarily seeing efficient analytical insights because they expend significant effort extracting, refining, reshaping, and enhancing data for it to be analytically useful. In this arrangement, data is easy come, but not so easy go.

Organizations that adopt flexible data architectures (such as Hadoop with its ability to store text, non-conformed columns, and heterogenous rows from different file types) are also accepting a prospecting challenge, with data science tools standing in for pickaxes. There are generally two ways to address the prospecting challenge: architectural approaches that maintain incoming data flexibility while enhancing downstream analysis with system processing bridging the gap, or specialized tools that are designed to allow end users to work effectively and efficiently with flexible data in place. Both are viable alternatives with various tradeoffs.

"Remember, before operations may be performed, data must be discovered, browsed, and comprehended."

Typical data operations fall into three categories: shaping, cleansing, and enriching. Shaping refers to the transformations that must be performed on data to make it useful to end users (and consumable by their analytic tools). Examples include pivoting columns to rows (and vice-versa), unpacking data structures (such as JSON, XML, or key/value pairs), and dataset operations like splits, joins and unions. Another form of shaping is aggregation, where data is summarized using aggregate functions. Cleanup refers to deleting unfit data, or changing values in place (such as converting invalid characters to blanks). Enriching refers to the creation of new data from existing data with user-defined logic such as if-then-else. Remember, before operations may be performed, data must be discovered, browsed, and comprehended. This creates a chicken and egg problem, or at minimum a very repetitive process where data is viewed, transformed, viewed, etc. (the view-transform cycle). With these operations in mind, how best to carry them out?

Architecture Approach

Architectural approaches have been proposed by industry thought leaders to improve end-user consumption of data from flexible data architectures, which typically amount to separating data into appropriate layers and locations, based on common characteristics, to undergo different kinds of processing in advance of query requests and analysis activities. The idea here is to allow for a great deal of flexibility for different types of data to be brought into the system, then perform the necessary processing downstream to make the data useful and the analysis process efficient prior to consumption. The main trade-off with an architectural approach is ongoing

reliance on development activities to prepare data for analysis via the processing components, along with any rigidity caused by ability to incorporate new data formats efficiently. The view-transform cycle may also be relatively slow given this dependency on fixed processing components. Flexible data by definition may also contain unexpected elements, which makes it a challenge to develop architectures that anticipate all necessary processing in advance. With an architectural approach, emphasis should be put on minimizing development time so as to not undermine the key points of a flexible data architecture, which are speed and agility.

Tools Approach

A relatively new category of software has emerged to work with flexible data architectures in a way that requires minimal up-front architectural changes in exchange for ad hoc data transformations, performed by end users, that can be single-use or scheduled as recurring jobs. Typically described as a data wrangling process, these tools combine aspects of data browsing, ETL, visual analytics, and data publishing. In some cases, powerful sampling techniques are incorporated to retrieve relevant and workable subsets of data to be used during the wrangling process, a welcome feature if querying huge data sources by allowing the user to work with big data at the speed of thought. The main tradeoff with a tools approach is up-front software costs vs. the long-term development costs in the architecture route. With a tools approach, extensive investigation should be done to select a platform that adequately addresses all necessary use cases for the long-term.

Flexible data architectures are becoming more common and practical, but when organizations adopt these architectures, they should heavily consider the efficiency of data consumption from the end user's perspective. When evaluating architecture vs. tools approaches, the key goal should be to focus equally on the end user as on the ability to rapidly and flexibly incorporate data. If these factors are balanced, a flexible data architecture can be a strategic asset.

Key Takeaways

- With big data, some organizations are storing more data with less up-front work to master/transform it. This approach is geared toward data lake architectures.
- If this strategy is applied, cleansing / mastering / transforming of the data will likely occur sooner or later, if the data is to be useful. This work can be done in advance by the IT

team, or it can be done when the data is ready to be consumed with high-powered tools that end users can leverage. In either case (custom work or high-powered tool) careful consideration and analysis must be done in order to establish the right long-term roadmap and adopt a lasting approach.

- It is not recommended that this be the sole approach to working with strategic data (data that will be used for decision-making purposes), but data lakes can supplement a data warehouse environment. The former stores the raw data which needs work prior to consumption, and the latter is ready to be consumed at any time.

About the Author

Jared Decker is the founder and president of Expert Analytics, and provides advisory consulting services to clientele regarding their strategic data platforms and initiatives. He has more than 16 years in BI and data analytics consulting, with 15+ years in data architecture roles, designing data warehouses and BI platforms for clients in verticals such as commercial real estate, capital management, credit card, and consumer retail. He is the co-author of several analytics books published by Wiley, is a frequent big data columnist for executive tech publications, and has provided onsite training engagements for numerous *Fortune 500* companies, including Halliburton, Humana, PepsiCo, PPG Industries, and consulting companies that service *Fortune 500* and government clientele.

About Expert Analytics

At **Expert Analytics** we provide data experts and expert teams to carry out your strategic data initiatives expeditiously and cost-effectively. Our experts are experienced in a wide range of data tools and platforms and are highly credentialed with decades of successful client deliverables, advanced degrees, and certifications. Our experts help you achieve success with your most strategic initiatives and improve your teams' performance across a range of projects through mentoring and advanced expertise.

Expert Analytics, based in Dallas, Texas, was created by data warehousing and analytics veterans and thought leaders. From the beginning, **expertise is our culture** and is deeply woven into every client delivery.

www.ExpertAnalytics.com
(800) XPRT137 | (800) 977-8137
Info@expertanalytics.com