ANALYTICS IS NOT THE SOLUTION

(AND BIG DATA IS NOT THE WHOLE ANSWER)

Michael Ralston, HP-Vertica, Inc. Sunnyvale, CA

ABSTRACT

Greater data availability leads to potential greater depth and subtlety of modeling, but building a model and gaining actionable business insight from analytic data is fundamentally a fixed process (there are no short cuts). There are different impacts however. Big Data Analytic Processing taxes the process in one way, while Analytic Exploration taxes it in another.

The process can certainly be optimized, streamlined, and supercharged – but only by understanding the strengths and interactions of the tools and techniques available, and their impact on the overall value. A great recipe, with better tools or more ingredients won't improve a meal that isn't served while it's still hot. Big Data Analytics is essentially mass production analysis – focus on optimizing the process and the value will emerge. Get it while it's hot!

ANALYTICS IS NOT THE SOLUTION

(AND BIG DATA IS NOT THE WHOLE ANSWER)

Michael Ralston, HP-Vertica, Inc. Sunnyvale, CA

KEYWORDS

Big Data, Advanced Analytics, Architecture, Haven, Columnar, HDFS, Hadoop, Vertica, HP, SAS, Modernization, In Database

INDUSTRY
Banking
Capital Markets
Communications
Education
Energy and Utilities
General
Government
Health Care
Insurance
Life Sciences
Manufacturing
Media & Gaming
Retail
Services

TARGET AUDIENCE JOB FUNCTION
Information Technology/Architect
IT Executive
Business Intelligence Executive

SKILL LEVEL
General

NOT SO FAST!

Often Analytic Solutions are compared based on, among other things, how "fast" they are. What kinds of analytics are you trying to do? Are you looking at performing analyses on petabytes of data? If you are considering actual Big Data analytics, with more involved processing than simple queries, then you must first examine the architecture of the storage. A Ferrari cannot get out of a traffic jam with more horsepower. Analytics and the ability to rapidly consume and serve mountains of data is now the prime services engine of most IT organizations, and time-to-value is the measure.

Big Data. Big Analytics.

Big Time!

INTRODUCTION

Why Analytics Value is tied to performance

- The biggest contributors to time variance in analytic insight creation are in domain awareness, model development and data cleansing – all of which require agile human interaction and expertise
- Reducing development time (process efficacy, domain awareness, modeling and data cleansing, etc.) will have greater incremental impact on the business value of the analytic solution than improving solution performance time because of the short lifespan of the model
- The opportunity window and corresponding business needs in which to provide market-leading analytic value, at scale, is finite, changing rapidly and increasingly brief
- Building a Big Data analytic environment to reduce data movement and increase processing performance on batch queries will provide definite value and should be done…but it is not the only improvement opportunity – the real choke point is in *analytic* processing time, especially during development

ADVANCED ANALYTICS CURRENT STATE

What are the Core Challenges?

The good news is that most enterprises have found very robust ways of culling the low hanging analytic fruit, and are reaping the rewards of this. Unfortunately, the next – and most valuable – increment of value is hidden in mountains of data. Understanding the best architecture, technology and analytic techniques to use *if* mountains of data are available becomes our next challenge. The analysis process itself involves many steps, very intimate knowledge of the data and a deep behavioral understanding of the continuously-changing domain.

All of this is subtext to the main challenge, which is *Time*. The concept of Time-to-Value has been around for decades, however in the pursuit of business insight, the difficulty is found in a shrinking opportunity window relative to the Time-to-Value performance. Insight, no matter how astute, has an expiration date. It is very easy to be distracted by the allure of Big Data – more data seems to enable better results…but if those results can't be produced in a timely manner, then the value drops significantly.

How to Address These Challenges?

Given the rapid and increasing pace of global business, the ephemeral goal of "finding nuggets of insight" in the tidal waves of data that are collected and available…is "quite" challenging. Unfortunately, it's not as simple as "asking an expert", however the process a data scientist physically and mentally applies is much the same. The objective is to streamline and scale this data-to-insight process as much

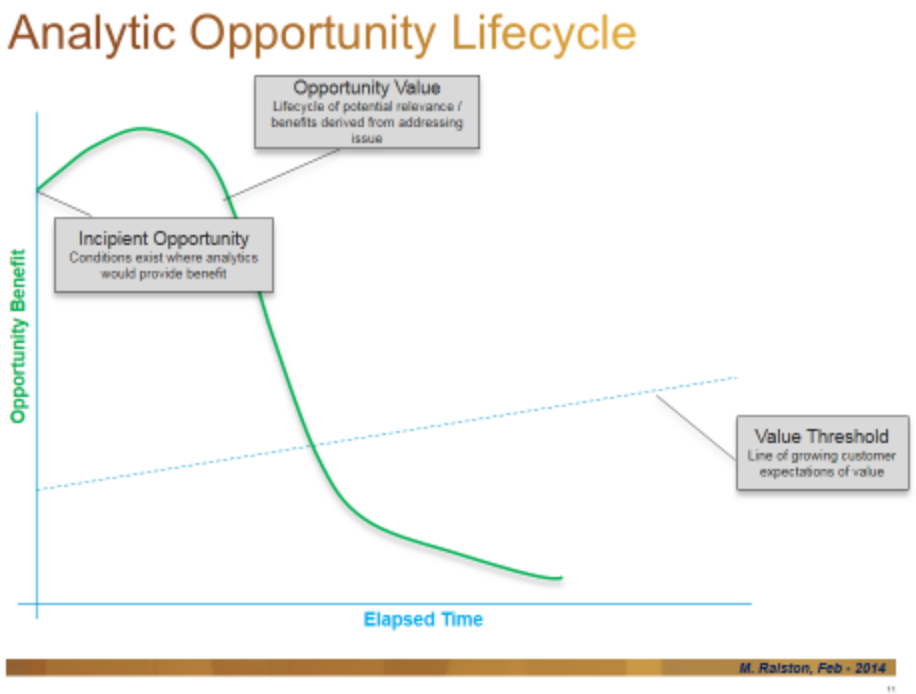
as possible. Creating an environment, purpose-built to process tabular data (i.e. a columnar format), allows massive scaling and performance improvements when doing analytics.

Where are the choke points in the process?

We build complex systems like spacecraft, musical masterpieces and behavioral models that all take a fraction of the time to *use* as they do to plan, develop and construct. Although there are very real and business impactful reasons to focus on performance time in production environments, we must also examine the data processing and analytics that take place during development – because this time is more sometimes actually more valuable.

Understanding the Analytics Opportunity Lifecycle

As an aside to the technical challenges, it is important to note that the reason for developing an analytic model is based on the presumption of an opportunity to gain business value from the insight provided. The so-called Analytic Opportunity Lifecycle is borne of the recognition of an opportunity, continues through the definition of the problem, identification and collection of the data and finally to the model development and deployment. All of this is in a race against time, as the business conditions that created the opportunity are of course evolving as well. The graphic below depicts the decaying value contribution of an analytic model over time.



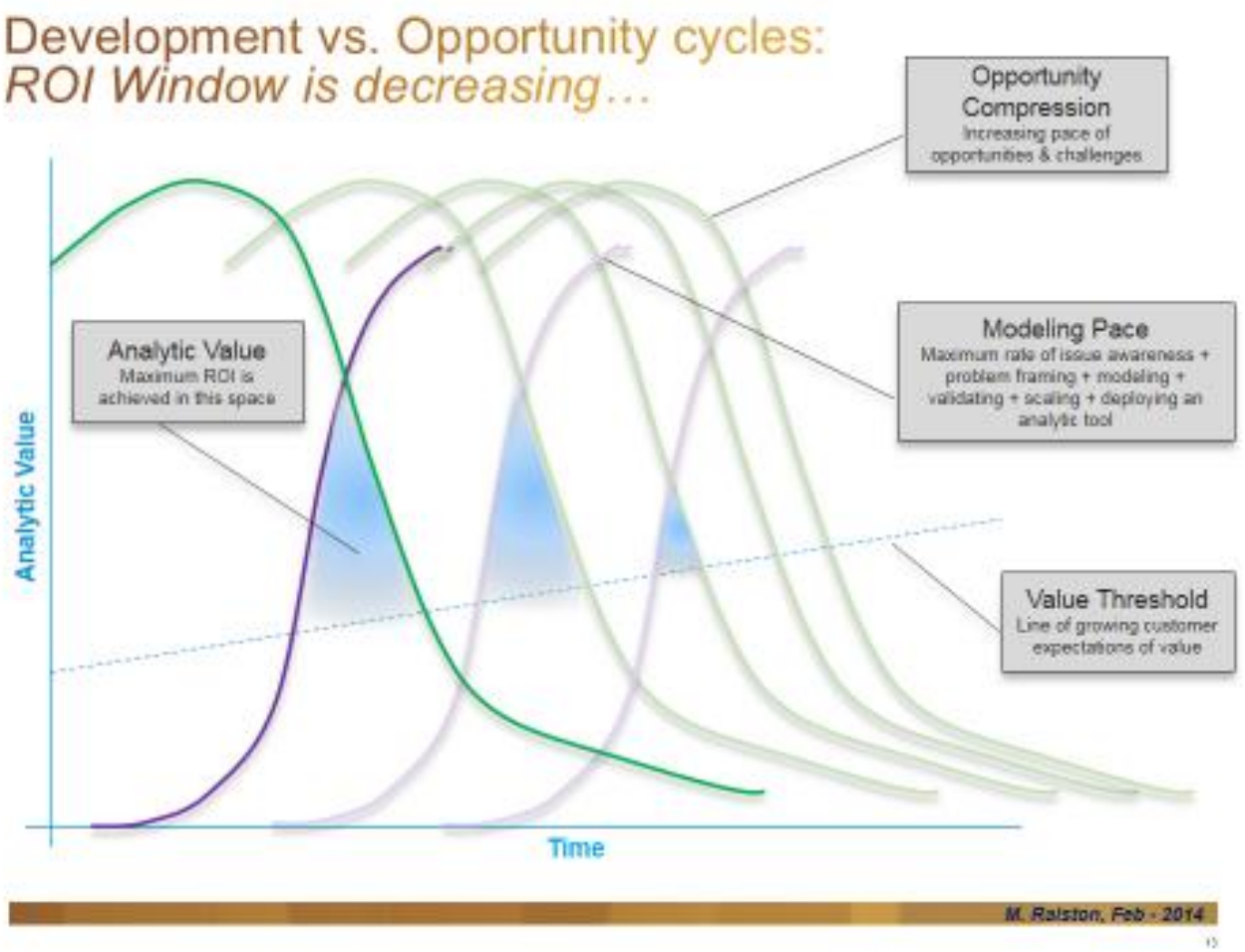
This concept is important because it helps focus our thinking about how and where to optimize the Big Data Analytics environment. It also gives a comparison context to the relative value of process improvement versus technology improvement versus model improvement.

The Development-Time Multiplier

Imagine an analytic model being deployed to predict the fraud risk in credit card transactions. Every minute in performance improvement for this model, when in production, may effectively save the enterprise thousands of dollars. They are understandably very keen to maximize the performance of their production systems. Now consider that this model is effective and relevant for about 2-3 months, after which time the environment has evolved enough where a re-tuned model would provide better lift.

Typically the enterprise is able to develop, test, produce and deploy new models at the rate of about 1 every 3-4 months. The development cycle is most impacted by moving & cleansing data and running and re-running the model during validation. The impact of slow performance here is delayed development and an effective multiplier effect on the time value of processing performance during the development phase.

This impact can be seen as a diminishing “potential” analytic value the model is able to produce, based on its time to deployment and how long it remains relevant. This value is further eroded by typical rising expectations from customers and ever quicker opportunity cycles.



ADVANCED ANALYTICS FUTURE STATE

Making Technology work for you: A Paradigm Shift Solution

The commonly held belief that better insight follows from more volume, more quality and more pace to the data, along with deeper analytic tools is fundamentally correct. The trick is to acknowledge the paradigm shift in architecture and process that is necessary to fully take advantage of all this data.

Historically, to perform advanced analytics, an architecture that supports much data preparation and allows for large datasets to be stored in memory while performing the analytics has been required. As the amount of data being stored and analysed has moved from gigabytes into petabytes (a.k.a. “Big Data”), the systems and architecture has begun to buckle. The first step – and still widely the only step taken so far – in this evolution has been the adoption of a distributed file system, making use of a Map-Reduce philosophy in distributed processing. This has opened the doors to great strides in data storage and in cost savings, but this Hadoop migration is merely the first step. The idea of moving the processing to the data is beneficial, but it simultaneously highlights to the opportunity to architect the database itself for the specific requirements of advanced analytics (versus simply query and reporting).

To take advantage of the inherent strengths of advanced analytic techniques, coupled with the opportunity to ingest truly colossal amounts of data, the enterprise needs to consider moving away from a functionally “layered” architecture and build with analytics in mind from the start.

Big Data plus Big Analytics

As noted, merely having more data does not guarantee faster or better analytic results. A highly distributed file system solves one problem, while opening the door to the next. Using advanced analytic techniques, the data scientist has need for extremely rapidly comparing huge amounts of data. The data scientist needs to perform hundreds of transformation, normalizing, exploratory and validating operations on the data during development. While operational in production, the analytics must be able to be processed on this huge data structure as its loading and often in an ad-hoc manner. A relational database or even an HDFS structure are dependent on reading entire records (all the columns) for these comparisons, whereas a database built with analytics in mind (i.e. columnar) will only read the columns being compared.

The details of how queries and SQL-based analytics are performed within a columnar database is beyond the scope of this paper, however the bottom line is that the advent of Big Data has led to the dawning of the age of Big Analytics, which in turn is driving the architectural revolution. Having schema-less databases and data stores optimized for high speed analytics will open the door for us to make even better use of the advanced analytic tools at our disposal.

Moving Forward

With an understanding of the business forces in play, namely the interacting analytic and opportunity lifecycles, the enterprise is in a position to craft an analytics – actually a Big Data Analytics – strategy

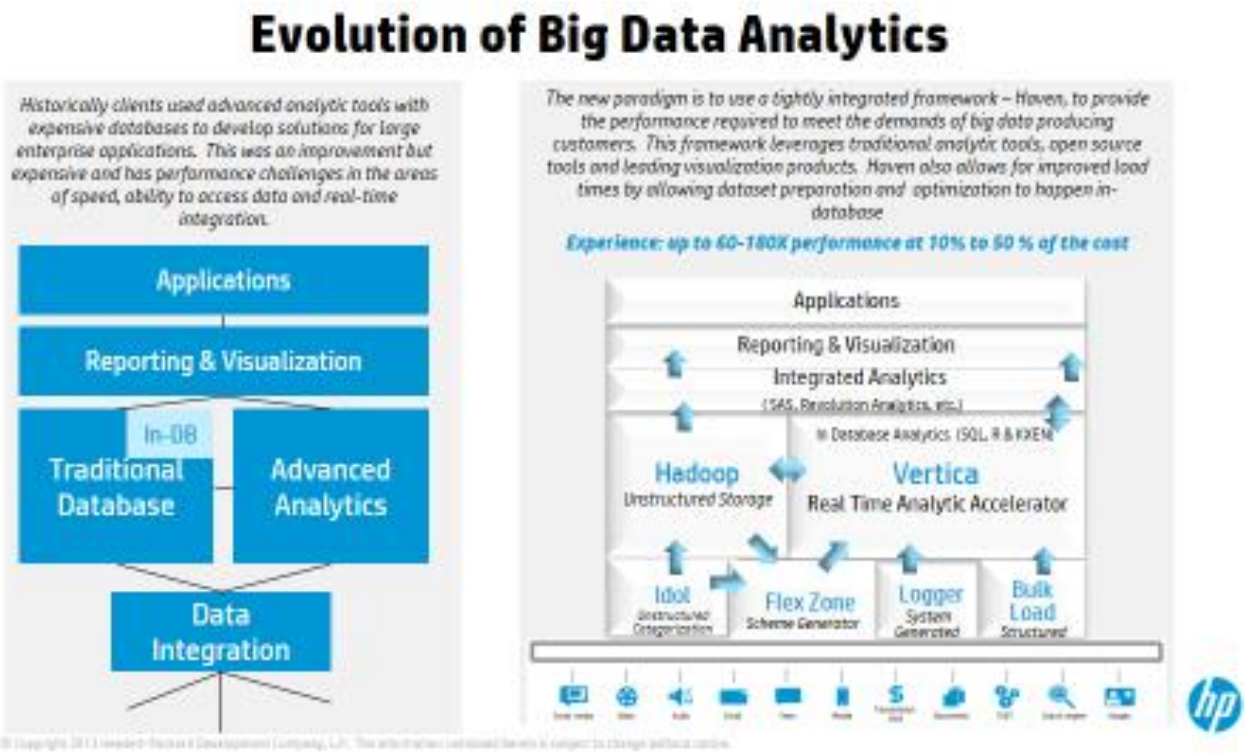
based on the full development cycle and the unique needs of analytic data processing. Fundamentally the idea of big data is not to solely stockpile data, but to analyse it. Low latency analysis drives the value proposition in production and multiplicatively in development. This low latency is the result of creating a storage architecture that is purpose built for performing analytic processes internally, rather than either pulling the data out (just lost all time gained from DFS), or by spending the effort to map complex analytics to distributed data (not practical for small or ad-hoc operations).

It is reasonable to assume that an ideal architecture will incorporate elements of both a columnar store and a distributed file system. The important knowledge for the architect, is gaining an understanding of both the strengths of the different options and an understanding of the business drivers of the analytic process.

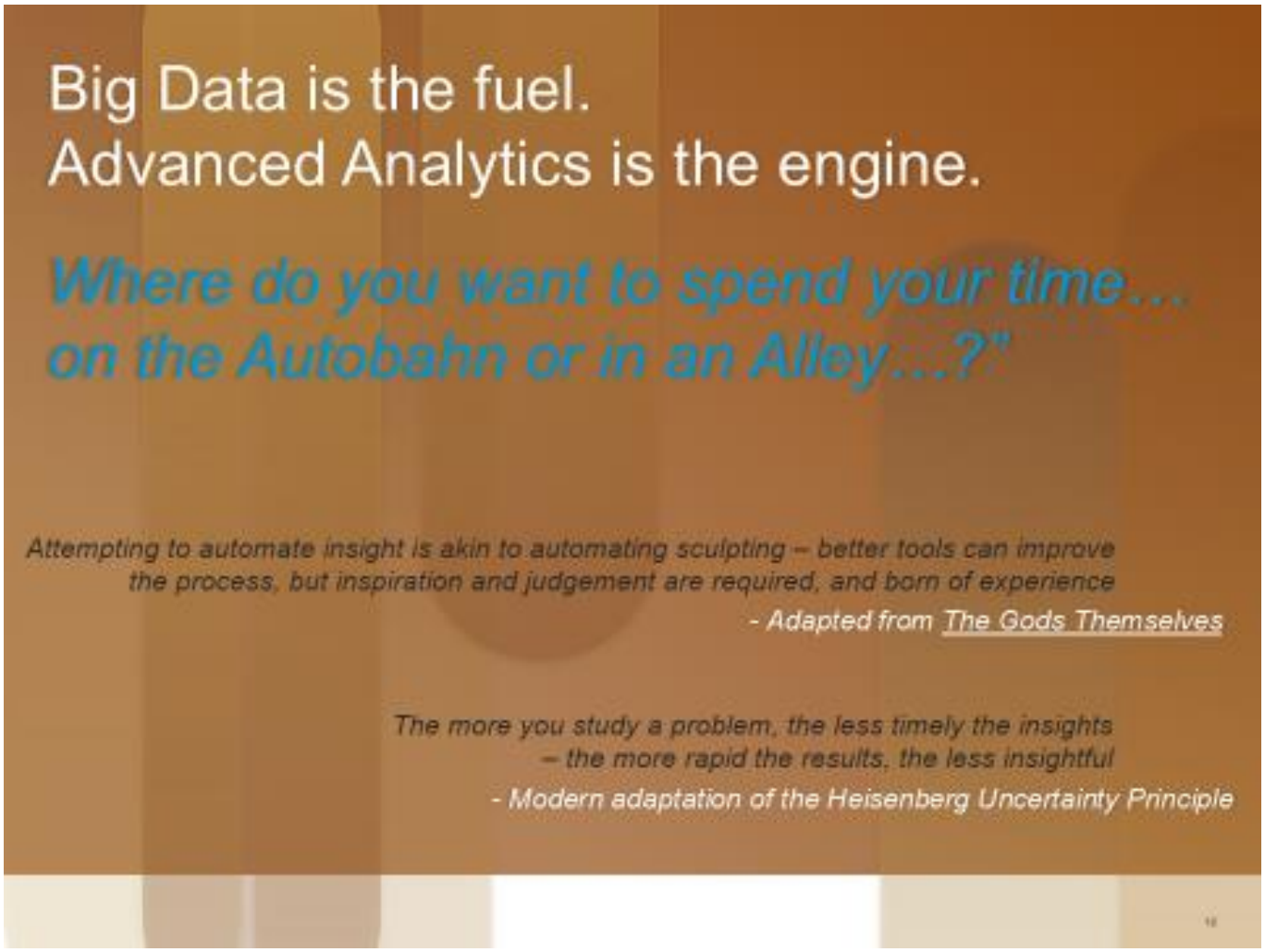
EXPERIENCES

An Example of an Integrated Architecture

Every enterprise is different, but for all looking toward a Big Analytics architecture, an instructive concept will be integration. Finding the right combination of tools, technology and applications will be driven by the adoption of an integrated platform that achieves the purpose-built goal of optimizing for analytics – particularly Big Data Analytics.



There is room and need for many of these technologies to coexist in this new paradigm, as long as it is built with purpose and understanding. The new shift will allow the real depth of analytic analysis to come to the forefront. With low cost, essentially limitless scalability, look for a renewed emphasis on trend detection and unstructured data analyses.



Michael Ralston, MS MBA
Analytics CTO
HP-Vertica Global Alliances

About the Author

With an innovation and thought leadership career spanning several decades, Michael has worked in all aspects of the analytics domain. In particular, optimization of complex systems, spacecraft design, six-sigma process improvement, Big Data solutions architecture and analytics strategy, are all areas of expertise. He now leads Analytic Offering Development for HP-Vertica as the Global Alliances Analytics CTO.

Responsible for the architecture and technology alignment of emerging strategic offerings, Michael works with HP-Vertica's Global Alliance Partners to bring analytics and database solutions together into scalable business offerings. Previously, Michael led the Analytics Center of Excellence within Cisco Systems Advanced Methodologies Group and helped launch the Solutions Innovation group at SAS.

AUTHOR CONTACT

Michael Ralston
Hewlett-Packard, Inc.
1140 Enterprise Way
Sunnyvale, CA 94089

+1.650.258.0522
michael.ralston@hp.com

ACKNOWLEDGEMENTS

Michael would like to acknowledge Mr. Michael Italiano and Mr. Rick Coughenhour for their instruction and support during his time at SAS. He would also like to thank Mr. Kevin McConnell for his support and guidance in turning this academic discussion into a paper and in his ongoing mentoring at HP.

REFERENCES

Davenport, Thomas H. and Jeanne G. Harris, 2007, Competing on Analytics, Harvard Business School Publishing Inc.

Davis, Stan and Christopher Meyer, 1998, Blur: The Speed of Change in the Connected Economy, Ernst & Young Center For Business Innovation.

McConnell, Kevin and Jeff Healey, 2013, "Big Data Analytics: What's Old is New Again", HP-Vertica Blog.

Owen, David, 2010, "The Efficiency Dilemma", The New Yorker Digital Edition, Conde' Nast Publishers.

Ralston, Michael, 2013, "Solving The Time to Market Dilemma", Cisco Systems, Big Data Summit (Internal).

TRADEMARK INFORMATION