BIG DATA ANALYTICS

Gain competitive advantage from the combination of big data and advanced analytics
# TABLE OF CONTENTS

**BACK TO THE FUTURE: THE ADVENT OF BIG DATA**  
1

**EXPLOITING THE “BIG ANALYTICS” BUSINESS OPPORTUNITY**  
1

**POINT #1: AGILE COMPUTING PROGRAM**  
2
   - ENABLE “AGILE” DATA WAREHOUSING  
   2
   - INTEGRATED DATA WAREHOUSE AND ANALYTICS  
   2

**POINT #2: LINEAR SCALABILITY**  
3
   - TRANSFORM ETL INTO DATA ENRICHMENT  
   3
   - SUPPORT EXTREME VARIABLE QUERY AND ANALYTIC WORKLOADS  
   3
   - ANALYZE MASSIVE, GRANULAR DATA SETS (BIG DATA)  
   4
   - ENABLE LOW-LATENCY DATA ACCESS AND DECISION MAKING  
   4

**POINT #3: PERVERSIVE, UBQUITOUS, AND COLLABORATIVE USER EXPERIENCE**  
5
   - ENABLE INTUITIVE AND PERVERSIVE USER EXPERIENCE  
   5
   - EXPLOIT THE COLLABORATIVE NATURE  
   5
   - ENABLE NEW BUSINESS APPLICATIONS  
   5
   - ATTRIBUTION-BASED APPLICATIONS  
   5
   - RECOMMENDATION-BASED APPLICATIONS  
   6
   - PREDICTIVE/FORECASTING-BASED APPLICATIONS  
   6
   - INSIGHT-BASED APPLICATIONS  
   7
   - BENCHMARK-BASED APPLICATIONS  
   7

**CLOSING THOUGHTS**  
7

**ABOUT THE AUTHOR**  
8
Throughout the 1970s and early 1980s, consumer package goods (CPG) manufacturers and retailers ran their businesses using AC Nielsen bimonthly market reports. These reports showed competitor and market data (e.g., revenue, units sold, average price, and market share) that CPG manufacturers used to set sales, marketing, advertising and promotional strategies, plans, and spend with their channel partners (e.g., distributors, wholesalers, and retailers).

Then in the mid-1980s, Information Resources Inc. (IRI) offered to install free point-of-sale (POS) scanners into retail locations in exchange for the residual scanner data. Retailers were more than happy to make this exchange, as labor was their largest cost component and there was limited understanding as to the value of the POS scanner data.

This POS scanner data, which was considered “big data” at the time, caused a game-changing, industry shift in power (between the manufacturers and retailers) and the way that both manufacturers and retailers ran their businesses. Data volumes jumped from megabytes to terabytes necessitating a new generation of storage and server platforms (e.g., Teradata, Red Brick, Sybase IQ, Sun, and Sequent) and analytic tools (e.g., Metaphor, Business Objects, Informatica, and MicroStrategy).

Leading-edge companies like Wal-Mart, Procter & Gamble, Tesco, and Frito Lay exploited this new “big data” and the new analytic platforms and tools for competitive advantage. These companies were at the forefront of developing new categories of big data, analytics-driven business applications to address business problems that previously were not addressable in a cost-effective manner (see chart to the left).

30 years later, it’s déjà vu all over again. There is an explosion in new, low-latency, fine-grained, diverse data sources (“big data”) that hold the potential to change the ways that organizations and industries operate. These new data sources are coming from a bevy of devices, customer interactions, and business activities that reveal new insights into the organizational and industry value chains.

The advent of these new, more detailed data sources will enable organizations to address business opportunities that they could not previously address, and lead to the creation of new families of business applications. However for that to happen, new platforms (infrastructure) and tools (analytics) must emerge. This white paper will outline the role that these new platforms can play and discuss the types of analytic-driven, big data-exploiting business applications that may emerge.

**EXPLOITING THE “BIG DATA ANALYTICS” BUSINESS OPPORTUNITY**

Big Data requires a new analytics platform from which both business and technology can gain competitive advantage. This requires a new technology infrastructure that is (a) massively scalable to petabytes of data, (b) supports low-latency data access and decision-making, and (c) has integrated analytics to accelerate the advanced analytics modeling and operationalization processes.

The ability to bring new scales of processing power to bear on massive data sets allows for the continuous identification of actionable insights buried within the big data, and enables the seamless integration of these actionable insights within the user’s work environment, wherever that might be. This new analytics platform can free organizations from the old ways of retrospective reporting by delivering forward-looking, predictive analytics to the masses and improved business decisions at all levels of the organization.
POINT #1: AGILE COMPUTING PLATFORM

Agility is enabled through highly flexible and re-configurable data warehousing and analytic architectures. Analytic resources can be quickly reconfigured and redeployed to meet the ever-changing demands of the business, enabling new levels of analytics flexibility and agility.

ENABLE "AGILE" DATA WAREHOUSING

The new analytics platform allows for the development of data warehouses that are free of the constraints found in today’s IT environments. Today, organizations are forced to use unnatural design techniques and unsophisticated reporting tools to mine insights from rapidly-growing, massive data sources using out-dated database technologies. As these data volumes continue to grow and new data sources come online, organizations are finding that today’s architectures, tools, and solutions are too expensive, too slow, and too rigid to support their strategic business initiatives.

Consider the impact of pre-building aggregates. Aggregates\(^1\) are commonly used to overcome the limited processing power of traditional relational database management systems (RDBMS) in handling multi-table joins and massive table scans. A database administrator (DBA) pre-calculates the most common aggregates during data preparation to speed ad hoc and reporting performance. The amount of data stored in these aggregate tables is growing to several times larger than the raw data itself. So much time is required to pre-build the aggregates that service level agreements (SLAs) suffer. Exploiting data “trickle feeds” to provide “realtime operational reporting” is a pipe dream because of the time required to rebuild the aggregate tables every time new data “trickles” into the data warehouse.

Eliminating these limitations enables an agile data warehouse environment that is as flexible and responsive as the business it serves by exploiting the following capabilities:

- **On-demand aggregation**—There is no need to pre-build aggregates to provide faster query and report response time. The power exists to create aggregates in real time, eliminating the debilitating need to constantly rebuild aggregates each time new data trickles into the data warehouse.

- **Index independence**—DBAs can eliminate the need for rigid indexing. DBAs won’t have to know ahead of time the questions that users want to ask so that they can build all the supporting indices. Users are free to ask the next level of detailed business questions without worrying about performance problems.

- **On-the-fly Key Performance Indicator (KPI) creation**—The business users are free to define, create, and test new derived (and composite) KPIs without having to involve the DBAs to pre-calculate them.

- **Flexible, ad hoc hierarchical structures**—The dimensional hierarchies do not need to be pre-defined when the data warehouse is built. For example, during a market intelligence analysis, organizations could have the flexibility to change the companies against which they are being benchmarked.

INTEGRATED DATA WAREHOUSE AND ANALYTICS

Traditionally, the data warehouse and analytics have resided in different environments. Moving data from the data warehouse into the analytics environment has required a separate ETL process where the data is selected, filtered, aggregated, pre-processed, re-formatted, and then transported to the analytic environment. Once in the analytic environment, the data analysts start building, testing, and refining the analytic models and algorithms. If during this process the data analysts realize that they need more granular data and/or different data, they have to repeat the entire data warehouse ETL process. This can add days, if not weeks, to the analytic process.

---

\(^1\) Aggregates are pre-calculated hierarchical or dimensional summaries of the facts (measures or metrics) usually defined by the “SQL Group By” phrase. For example, in the Geography dimension, one might create aggregates for all of the facts (e.g., sales, revenue, margin dollars, margin percent, and returns) by country, region, state, city, county, and zip code.
A single, integrated data warehouse and analytics environment—with in-database analytics—means that the data analysts don’t need to leave the data warehouse to perform analytics. Massive data sets can be transferred between the data warehouse and analytics environments at blazing transfer rates (5 TB to 10 TB/hour). This significantly speeds the analysis process and makes it easier to integrate the analytic results back into the data warehouse and business intelligence environment. For example, an integrated data warehouse and analytics environment supports the following types of analyses:

• Sub-segment and stream massive data sets between your data warehouse and your analytics environment to enable the creation of “analytics sandboxes” for analytic exploration and discovery.
• Interrogate massive data sets at the lowest level of granularity to flag “unusual” behaviors, trends, and activities to create actionable insights with corresponding recommendations.
• Speed the development and testing of different business scenarios to facilitate what-if analysis, sensitivity analysis, and risk analysis.

These types of benefits would be invaluable integrated into everyday tasks. Just ask a distribution manager who wants the system to flag potential vendor performance problems with respect to deteriorating service as measured by on-time-deliveries and the percentage of complete deliveries.

POINT #2: LINEAR SCALABILITY
Access to massive amounts of computing power means that business problems can be attacked in a completely different manner. Let’s look at a few examples where massive computational scalability can impact the business.

TRANSFORM ETL INTO DATA ENRICHMENT
ETL focuses on correcting the mistakes caused by the source systems; extracting, translating, cleaning, profiling, normalizing, and aligning all of the data to ensure that the users are comparing apples to apples in their analysis. With the processing power available for ETL (plus leveraging the availability of new computing languages like Hadoop), the traditional ETL process can be transformed into a data enrichment process. New and insightful metrics can be created including:

• Sequencing or ordering of activities—Identifying the sequence of activities that occurred before a specific event. For example, identifying that someone typically calls the call center twice after searching for support options on the website before successful problem resolution.
• Frequency counts—Counting how often a certain event occurs during a specified time period. For example, identifying the products that had x number of service calls during the first 90 days of use.
• N-tiles—Grouping items (e.g., products, events, customers, and partners) into buckets based upon a specific metric or set of metrics. For example, tracking the top tertile (top 10%) of customers based upon revenue or margin over a rolling three-month period.
• Behavioral “baskets”—Creating a “basket” of activities, including frequency and ordering, that precede a sales or “conversion” event in order to identify the most effective and profitable combinations of market treatments.

SUPPORT EXTREME VARIABLE QUERY AND ANALYTIC WORKLOADS
It’s very difficult to know in advance the types of queries and analytics that the business might want to perform based upon the most current business environment. A competitor’s pricing or promotional actions might require sudden analysis to better understand the financial and business impact upon your company. Most interesting analytics involve extremely variable workloads that are difficult to predict a priori.
Previously, organizations had to be content with performing cursory “after the fact” analysis and didn’t have the computational power to deep dive into the analysis as events were occurring or to contemplate all the different variables and permutations that might be driving the business. With the new platforms, these computationally-intensive, short-burst analytic needs can be supported. This capability manifests itself in the following ways for the business users:

- **Performance and scalability**—Agility to “drill into” the data to ask the second and third level of questions necessary to support decision-making. If business users want to look across all the fine minutia of the data to find those variables that are driving the business, they don’t need to worry about grinding the system to a halt by analyzing too much data.

- **Agility**—Supporting the rapid development, testing, and refining of analytic models that help to predict business performance. The Data Analysts are free to explore the different variables that might be driving business performance, learn from the results, and work those findings into the model’s next iteration. They are free to fail fast in their analysis without worrying about the system performance ramifications of their analysis.

**ANALYZE MASSIVE, GRANULAR DATA SETS (BIG DATA)**

One of the most significant advantages of the cloud is the amount and level of detailed data that can be brought to bear on analyzing and modeling business drivers. Not only does the cloud bring the potential for more efficient, on-demand processing power, but the cloud also delivers more efficient, cost-effective data storage capabilities. Instead of constraining the business to adapt to the data, the business is free to expand its analysis by taking advantage of all aspects of the data in the following ways:

- The ability to perform multi-dimensional analysis to the nth degree. The business is not constrained to contemplate only three or four dimensions, but can look at hundreds—if not thousands—of dimensions to fine-tune and localize business performance. With this level of multi-dimensional analysis, the business can find the business drivers by specific geography (e.g., city or zip code), product (e.g., SKU or UPC level), manufacturer, promotion, price, time of day or day of week, etc. With this level of granularity, local business performance can be dramatically improved.

- Find enough “small” diamonds across the volume of data to make a material difference to the business. The platform addresses two key challenges of localized analytics: the first is to find the business drivers at the local or specific level, and the second is to find enough of these local business drivers that can make a material difference to the business.

**ENABLE LOW-LATENCY DATA ACCESS AND DECISION MAKING**

Since data does not need to go through a heavy data preparation stage (with respect to pre-building aggregates and pre-calculating derived metrics), the latency between when the data is generated and when it is available to the business is greatly reduced. The ability to shrink the time between the data event and data availability means that the concept of operational analytics is a reality in the following ways:

- Exploiting continuous data feeds (trickle feeds) to provide low-latency operational reporting and analytics. The time between a business event (like a security trade) and the decision whether to buy or sell is dramatically reduced. We can see the impact of this low-latency decision-making clearly in the upswing of Wall Street algorithmic trading.

- Low-latency data access enables “timely,” in-flight decision-making. For example, campaign managers can reallocate online campaign budgets between best-performing and/or best-converting sites and keywords combinations, while the campaign is in-flight.
POINT #3: PERVERSIVE, UBIQUITOUS, AND COLLABORATIVE USER EXPERIENCE

Business users don’t want more data, charts, and reporting options no matter how elegantly they are presented. Business users want a solution that leverages analytics to identify and deliver material, actionable insights on their business.

ENABLE INTUITIVE AND PERVERSIVE USER EXPERIENCE

Coupling detailed data with massive analytic power has an interesting benefit—simpler, more intuitive interfaces. How is that possible? Think about the relationship between the iPod and iTunes. The iPod’s minimalist interface is one reason for its customer success (and dominant market share). Apple pushed most of the user complexity (e.g., managing playlists, adding new songs, and generating recommendations using the Genius feature) out of the iPod and into iTunes where it can be better managed. We can apply this same concept to improve the analytics user experience.

• The user experience can leverage analytics to do more of the heavy-lifting data analysis behind the scenes. Instead of presenting an ever-growing complexity of reports, charts, and spreadsheets, the interface can become more intuitive and present the users with the insights they need to know about the business.

• Based upon the insights teased out of the data, the user experience can lead with specific recommended actions (like the iTunes Genius feature). The complexity of identifying relevant and actionable recommendations is left to the analytics.

For example, envision a campaign manager interface that distills the myriad variables that impact campaign performance to only those variables that are material and actionable. Imagine that the user interface not only presents just those variables, but also presents recommendations to improve in-flight campaign performance. Now that’s a user experience that most users would relish.

EXPLOIT THE COLLABORATIVE NATURE

Collaboration is a natural part of the analysis and decision-making processes. Small communities of like users can quickly form to share learnings around specific subject areas.

For example, it would be very powerful for all the brand managers within a large consumer package goods company to create a community where data, information, and insights about brand management can be easily shared and discussed. Marketing campaigns that are working for one brand can be copied and expanded by other brands more quickly by sharing the resulting data and analysis.

ENABLE NEW BUSINESS APPLICATIONS

One way to appreciate the potential of this new “analytics” platform is to ask: what types of business problems does it—with its on-demand processing power, fine-grained data sets, low-latency data access, and tight data warehouse and analytics integration—enable the business to address what it could not address previously? Let’s review some business applications that analytics on these new platforms, especially when coupled with big data, can enable.

ATTRIBUTION-BASED APPLICATIONS

Attribution applications look to attribute “credit” for a particular event to a series of activities or transactions in a complex, multi-stage process. These applications need to retrieve, align, and analyze the series of activities, taking into account factors such as frequency, sequencing, recency, thresholds, and time decay between activities in order to accredit value to each activity. Sample Attribution applications include:

2 In electronic financial markets, algorithmic trading is the use of computer programs for entering trading orders with the computer algorithm deciding on aspects of the order such as the timing, price, or quantity of the order, or in many cases initiating the order without human intervention.
• Multi-channel marketing effectiveness applications where marketers are trying to attribute credit for a sale across multiple marketing channels. This is especially topical for online marketers trying to attribute credit for a conversion across multiple display ads, websites, and keyword searches.
• Partner attribution applications where sales organizations are trying to measure partner contributions in complex, multi-stage business transactions.
• Medical treatment attribution applications where health care organizations are trying to attribute the impact of different treatments and medications that resulted in some outcome.

RECOMMENDATION-BASED APPLICATIONS
Recommendation applications identify and create sets of “like” or similar users or products based upon behaviors, demographics, or some other discernable attributes. The applications analyze the transactions from these sets to create propensities that measure the strength of relationships between users and their behaviors and preferences. From these propensities, the applications are then able to make product (e.g., Amazon and Netflix) or people (e.g., LinkedIn and Facebook) recommendations. Sample recommendation applications include:

• Customer ad targeting applications that recommend “like” or similar target audience segments based upon behaviors and product purchase history (e.g., campaigns that are successful in targeting “Soccer Moms” have a high likelihood of also being successful targeting “New Grannies”).
• Product recommendation applications that recommend complementary products based upon what similar users have bought at what period of time (e.g., customers who bought a new home within certain zip codes are likely to purchase a new washer and dryer within three months of the home purchase).

PREDICTIVE/FORECASTING-BASED APPLICATIONS
Predictive and forecasting applications ingest a wide variety of variables, metrics, and dimensions to facilitate decision-making under different market scenarios. These applications leverage statistical and data mining techniques to distill a multitude of variables to identify those variables, and combination of variables, that are the best at predicting performance in certain situations. Given the time horizon of some of these decisions (like pricing), low latency access to the data and in-database analytics are critical to success.

Advanced predictive applications build in risk and sensitivity assessments so that the decision maker can understand which variables are the most important in making the decision. For example, if a certain variable is deemed to be critical in making a decision, then extra effort can be made to ensure the accuracy and completeness of that variable. Sample predictive/forecast applications include:

• Customer churn applications that predict the probability of customers’ attrition based upon factors such as usage activities, support requests, payment patterns, and the social impact of friends.
• Product maintenance applications that predict equipment failures based upon product usage information (especially information now being provided by embedded data devices), maintenance service records, and general product performance history.
• Employee performance applications that predict a potential employee’s performance based upon factors such as education, social-economy standing, previous job history, marital status, and certain psycho-behavioral responses.
• Clinical trial performance applications that model different drug outcomes based upon clinical trials so that a company can understand the effectiveness of certain treatments, and avoid catastrophic problems when drugs are used in certain combinations. This becomes even more important when trying to attribute outcomes across multiple treatments and medications (see attribution applications).
• Yield management, merchandising markdown management, and price optimization applications that build time-sensitive models that help the decision maker to understand when and how much to increase or decrease prices given current demand and supply conditions. These types of applications are most common with commodity products (e.g., perishable goods, airplane seats, hotel rooms, fashion clothes, and Cubs baseball tickets) with value that goes to zero at a certain point in time.

INSIGHT-BASED APPLICATIONS
Insight applications use statistical and data mining techniques to identify “unusual” behaviors or situations. Advanced Insight applications have the ability to perform complex analysis across hundreds to thousands of business dimensions. These applications are becoming more important as the volume of data minutia is growing from data sources such as web clicks, RFID sensors, and networked appliances. Sample Insight applications include:

• Product distribution and shrinkage applications that constantly monitor sensors and RFID data to identify discrepancies between where the product is supposed to be, and where it actually is.
• Fraud applications that are constantly monitoring financial transactions to spot “unusual” behaviors that may be indicative of fraudulent activities. These types of applications are being applied to credit cards, checking accounts, and insurance and Medicare claims.
• Money laundering applications that are constantly monitoring the flow of cash to identify “unusual” behaviors that may be indicative of money laundering such as an unusually high number of sequential, small, cash-only transactions.

BENCHMARK-BASED APPLICATIONS
Benchmark applications leverage analytics that compare one entity’s performance to some baseline. The comparison baseline could be an industry standard, a previous period, or a previous event (e.g., a marketing campaign). Sample benchmark applications include:

• Market share applications that provide share of market and share of wallet information. For example, big website companies can provide “share of voice” data and analysis that helps advertisers and agencies understand how their marketing spend compares to that of their competitors.
• Competitive benchmark applications that compare a company’s performance versus an aggregate of competitors, or an industry average. This provides a baseline against which companies can compare their financial or market performance.
• Campaign benchmark applications that compare the performance of a current marketing campaign to a previous and/or similar marketing campaign or event. For example, a company may want to compare the performance of their current “4th of July” campaign versus how the same campaign performed last year. The users may want to track percent of total campaign sales-to-date for each day of the campaign, and compare top and bottom performing geographies and products on a daily basis.

CLOSING THOUGHTS
These new, massively scalable platforms bring game-changing capabilities to the world of analytics. What are the advantages over today’s data warehouse and analytics platforms?

• The agility to provision and reassign massive computing resources on-demand as driven by business priorities.
• The ability to analyze more granular, more diverse, low-latency data sets (big data) while preserving the detailed nuances and relationships in the data that will yield the differentiated insights that enable optimized business performance.
• Cross-organizational collaboration around key business initiatives and rapid dissemination of best practices and organizational findings.
• Cost superiority to leverage commodity processing components to analyze big data to address and exploit business opportunities that previously could not be addressed in a cost-effective manner (if at all).
The ideal platform for analytics brings massively scalable processing power, the ability to exploit fine-grained data sets, low-latency data access, and tight data warehouse and analytics integration. If properly understood and deployed, it can be used to solve tough business problems that could not be addressed previously, and deliver material and actionable insights to the business.

CASE STUDY: HAVAS DIGITAL

The POWER OF INTEGRATED ANALYTICS TO DRIVE COMPLEX ATTRIBUTION MODELING

Havas Digital and EMC’s Data Computing Products Division are working jointly to enhance the Artemis Analytics Lab—a research and development initiative to embrace large-scale, big data analytics in the private cloud. This initiative greatly amplifies Havas Digital customers’ understanding of user behavior and their ability to optimize marketing campaigns accordingly.

The Artemis Analytics Lab is an initiative that combines Artemis’ expertise in data-driven marketing and EMC® Greenplum’s data computing foundation to provide industry-leading digital marketing analytics mining and in-database analytics.

“For many marketers, simple demographic information about users is no longer enough,” says Katrin Ribant, senior vice president of Artemis. “Our customers want to understand aspects of users’ behavior that can only be learned over time, and to infer conclusions that hide behind simple lists of transactions. Using advanced analytics, EMC Greenplum® provides analytics capabilities built directly into the database and executes on very large datasets, which enables us to generate rich new insights into user behavior and helps marketers predict how users will respond to new campaigns.”

A key differentiator of the Artemis system is its unique, state-of-the-art attribution mechanism, which more accurately computes the relative influence of advertising on purchase events. By moving modeling and other computations into the database and using EMC Greenplum, Havas Digital is now able to provide clients with a closer to real-time algorithmic attribution analytics framework.

ABOUT THE AUTHOR

Bill Schmarzo, Global Competency Lead at EMC Consulting, has over two decades of experience in data warehousing, BI, and analytic applications. Bill authored the Business Benefits Analysis methodology that links an organization’s strategic business initiatives with their supporting data and analytic requirements, and co-authored with Ralph Kimball a series of articles on analytic applications. Bill has served on The Data Warehouse Institute faculty as the head of the analytic applications curriculum.