ABS Group

DATA ANALYTICS FRAMEWORK

TABLE OF CONTENTS

Introduction 3

Set the Vision and Foundations for a Data-driven Business 4

Strategic Business Opportunities 6

Select the Best Data Analytics Approach 8

Secure the Data 11

Extract the Insights 12

Define and Execute the Responses 13

Continuous Improvement 14

AUTHORS

Henrique Paula
ABS Group Senior Vice President of Strategic Initiatives

Matt Mowrer
ABS Group Director, Government Programs – Safety, Risk and Compliance

Benjamin Roberts
ABS Group Consultant, Safety, Risk and Compliance

ACKNOWLEDGEMENT

The authors would like to thank Energective for providing guidance and assisting with this White Paper.

INTRODUCTION

Data Analytics and ‘Big Data’ have revolutionized many industries. However, these technologies have yet to manifest their full potential to deliver significant improvements in organizational performance and safety across a wide variety of applications. Early adoption has been driven by the recent, rapid decrease in the cost of advanced sensors, the expansion of wide area communication networks, the availability of high data storage capacity and ever-increasing computer processing power.

This document is structured around the seven (7) key factors required to succeed in Data Analytics, as shown in Figure 1. The first two (2) components are concerned with establishing the vision and capability for Data Analytics in the enterprise and putting in place a strategic planning process. The next four (4) components concern the execution of individual Data Analytics functional tasks. They focus on assessing the best Data Analytics approach to solving a problem, gathering the appropriate data, analyzing it and executing the appropriate response to the insights found. The final component is continuous improvement, a critical process that will help ensure the ongoing success of Data Analytics within an organization.

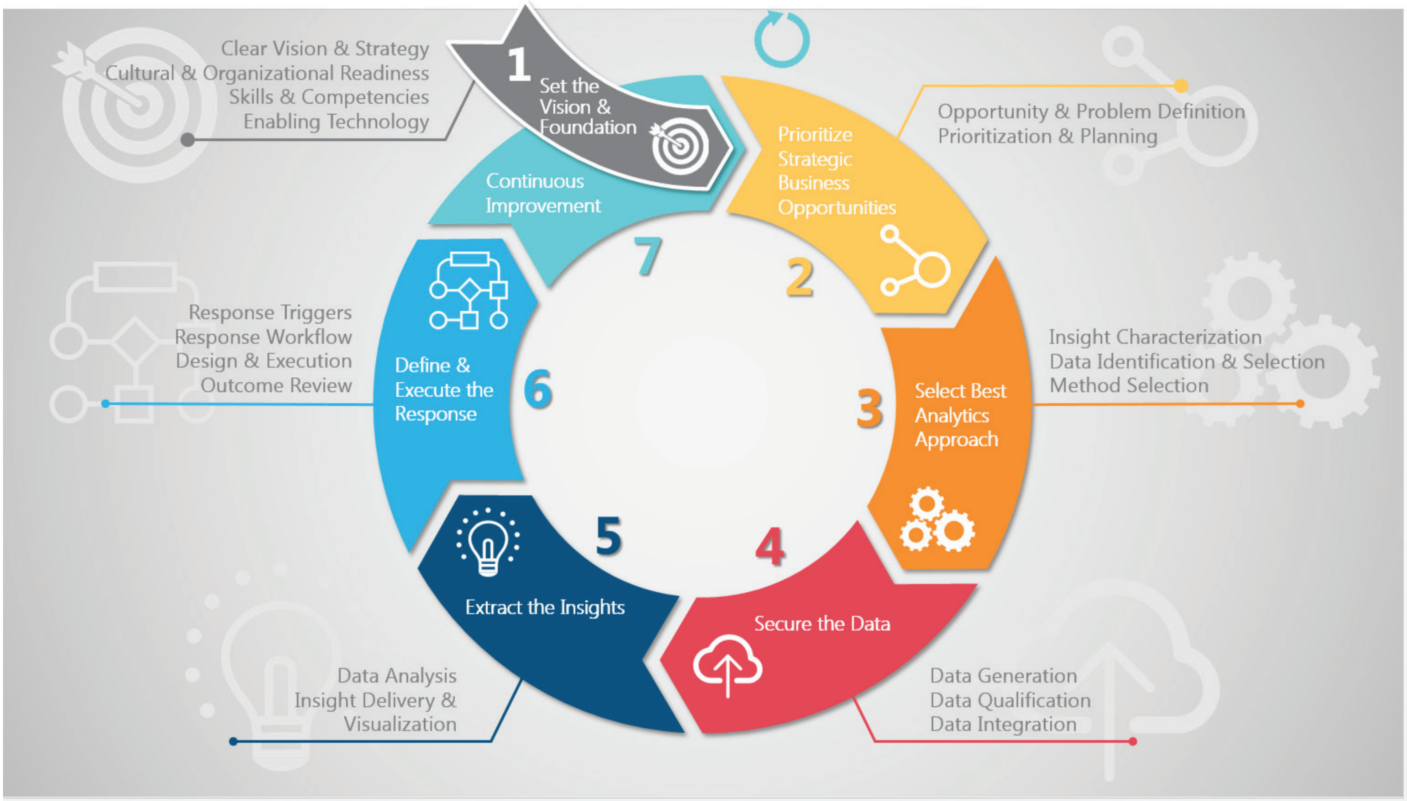


Figure 1. The 7 Steps to Successful Data Analytics

These success factors are presented as an iterative continuous improvement cycle, which together establish a robust strategic and tactical framework for managing Data Analytics across an enterprise. Each success factor will be described in more detail.

SET THE VISION AND FOUNDATIONS FOR A DATA-DRIVEN BUSINESS

Clear Vision and Strategic Directives

Setting a clear vision for Data Analytics and demonstrating executive support are key. The Vision describes what the business expects from Data Analytics. The associated strategic directives describe how it is going to achieve the Vision. Both set the tone and make clear that all are expected to play their role, as appropriate, to realize the strategic potential.

In defining the Vision, it is important to make clear that Data Analytics is not to be done for Data Analytics' sake. To ensure value creation, the first consideration has to be about the strategic priorities and objectives of the business. The role Data Analytics has to play to serve those priorities can be decided through consultation with the business, understanding strategic priorities and determining where the opportunity or challenge areas are. An appropriate investment budget can then be set, which will further underline the business's commitment. All of these factors should be considered as part of the normal business governance and strategic planning processes.

Cultural and Organizational Readiness

Value will only be realized from the insights that Data Analytics can bring if there is a willingness to accept and respond to them. A culture and organization has to be nurtured to allow this to happen, which recognizes the valuable knowledge and experience of employees and acknowledges how they can be helped in their roles and objectives through Data Analytics. Data Analytics often challenge conventional wisdom; so, creating a culture that consistently applies data demands changes to management systems and operating models. Culture change must be led from the top of the organization and may be supported by management consulting resources experienced at transforming companies. Senior executives and managers must continually stress the vision of a data-driven organization, where personnel at all levels make decisions based on objective evidence versus reacting based on gut feelings.

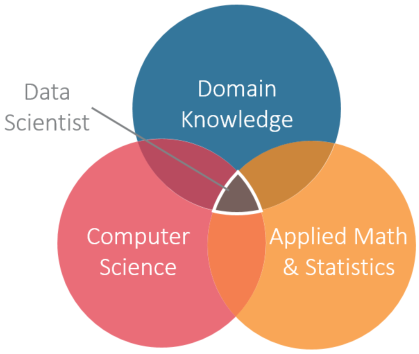


Figure 2. Key Skill Sets

Skills and Competencies

Undoubtedly, the skills and competencies required to perform Data Analytics successfully include deep domain expertise and experience in the industry itself. Invaluable are the people who can walk into a facility and know, based on what they see and hear, whether or not there is a problem. Augmenting this traditional expertise with new complementary talents and skill sets will bring the best value from Data Analytics. These skills include applied mathematics, statistics, computer programming, big data and visualization of insights. Such skills are in high demand, and the people with the required knowledge and talent – often referred to as Data Scientists – are scarce. Though the term “Data Scientist” is fluid and not universally defined, it is gaining popularity and prevalence due to the high demand for those with related skills. Not every person needs to have the full set of skills shown in Figure 2; organizations can create multidisciplinary teams that collectively have the requisite capabilities and experience.

Industry can benefit from working with academia and professional societies to help develop their people and capabilities in these areas. They could be young undergraduates taking particular industry-oriented Data Analytics courses, or they could be current employees who are intrigued – or even passionate – about this relatively new and potentially game-changing area and wish to be given the opportunity to take their careers in a new direction.

Awareness and Optimization of Technology Enablers

Awareness of the technologies available to perform Data Analytics is important. Broadly speaking the technology in question fits into the overall technical architecture, which comprises three (3) broad categories:

- Applications. The portfolio of software used by an enterprise
- Platform. Databases, Middleware, and Integration for sharing data between applications
- Infrastructure. Storage, computing, sensors and communication resources – the hardware

When defining the best technical architecture for Data Analytics for a particular enterprise, a strategic approach should be taken that is cost-effective, efficient, flexible and secure. For a particular individual Data Analytics solution, the technical architecture may need to be fine-tuned so that data flows are workable and viable, particularly if high bandwidth data streams are involved.

Not all of this technology needs to be run in-house. It is becoming increasingly common to outsource the technology or to put the technology on the so called “Cloud,” which is scalable and therefore cheaper than on-premises alternatives; however, security can be more of an issue on the Cloud.

A newer paradigm is also gaining popularity – the recently conceived so-called “Fog” or “Edge” computing – where the model is not to bring all data into a central storage and processing location, but to allow the processing to be performed where the data are being generated (i.e., with storage and processing capability built into a sensor). This local "intelligence" can identify trends and insights and determine when it is appropriate to communicate those back to a central monitoring location. This saves on communication costs and delays. In essence, Fog computing permits intelligence alerts, which spreads the load on storage and computing resources and helps optimize communication channels.

Open-source Data Analytics Enablers

Since the advent of "Big Data," innovative ways of integrating and analyzing large data sets have been conceived. Many of them are open-source, thereby helping to accelerate adoption. All of them have commercial, industrially scalable versions. There are many different solutions; five (5) of the most popular include the following:

Hadoop is an open-source software framework written in Java for distributed storage and processing of very large data sets on computer clusters built from commodity hardware. It does this by breaking the data down into small chunks and parceling them out between machines. It comprises a so-called "ecosystem" of optional software modules that enable advanced Data Analytics. It is tolerant to data of questionable quality, given the huge amount of data it typically processes, meaning the sample size is large and statistics can be used to whittle out spurious data values.

MapReduce is Google's alternative to Hadoop.

NoSQL refers to a type of database that is not based on traditional relational data models, which are tabular in nature. It has gained popularity because of its relative simplicity and, with the advent of "Big Data," it is easier to scale to hold large, heterogeneous data sets across computer clusters. NoSQL makes compromises to achieve simplicity and scalability. It accepts data of varying types and sizes and allows the data to be searched successfully; it is particularly useful since only 5% of all digital data are structured and fit neatly in a traditional database.

R is an open source programming language, popular among data scientists. It deviates from many traditional languages by focusing on raw data manipulation, statistical computation and visualization. Microsoft offers a commercialized version of the language as Revolution Analytics.

Python is a widely used general-purpose, high-level programming language that allows programmers to express concepts in fewer lines of code than would be possible in alternative common languages.

STRATEGIC BUSINESS OPPORTUNITIES

Opportunity and Problem Definition

A scope of work along the lines of "analyze several data sets and see what patterns can be found" is too loose, and yet seems to be how many Data Analytics projects begin. To ensure focused application of Data Analytics that serves the business, the scope and objectives should be clearly defined in terms of business priorities and associated key performance indicators (KPIs); for example, “reduce unplanned downtime by 5%” or “reduce maintenance costs by 3%.” Not only does this approach allow focus to be maintained on delivering a specific positive outcome, it also gives a clear indication of when that objective has been achieved. Table 1 lists several examples of opportunities where Data Analytics can improve organizational performance spanning a variety of departments.

Table 1. Potential Opportunities for Data Analytics

Engineering

- Improving process designs for availability, throughput and quality
- Minimizing life-cycle costs
- Improving operational safety
- Minimizing the potential for safety and environmental impacts

Operations

- Maximizing operational availability
- Maximizing throughput
- Minimizing off-spec products (product quality)
- Troubleshooting upsets and anomalies (downtime)
- Minimizing environmental impacts

Maintenance

- Optimizing spare parts inventory
- Minimizing unplanned downtime
- Maximizing maintenance effectiveness and efficiency activities
- Diagnosing equipment failures
- Optimizing maintenance planning
- Targeting inspection activity (risk and compliance)

Business (e.g., purchasing/logistics, marketing, human resources, finance)

- Minimizing capital and operating expenses
- Maximizing profits
- Optimizing production schedules
- Performing market research and competition analysis
- Rightsizing the workforce

Prioritizing and Planning Data Analytics Initiatives

The business value and cost of each candidate Data Analytics project should be estimated so they can be prioritized and planned accordingly. A high-level business case for each opportunity that identifies the anticipated benefits, costs, technical feasibility and risks will help decision makers choose the best candidates by addressing the initiatives with the highest return on investment. The choice should be informed by a cursory understanding of the available data and their quality to confirm the feasibility of the initiative.

SELECT THE BEST DATA ANALYTICS APPROACH

Insight Characterization

Having defined the opportunity or problem, the next step is to consider new insights that Data Analytics may provide to help address it or solve it. By “insight,” we mean a deep understanding of a topic – a very general term. To extract value from analytics, it is important to clearly define the type of insight being sought.

Note that it is imperative not to be overly prescriptive about target insights at this stage. There may well be insights from data that will be discovered and that cannot intuitively be anticipated. However, this is an important step in characterizing the type of insight required, which in turn helps target the right data.

For example, reducing downtime requires insights into the historical reliability of equipment, maintenance work history and root causes of unplanned outages. To reduce downtime, an analysis of the consistency – or lack thereof – of the work processes between more efficient and less efficient processes is foundational. An understanding of the maintenance workers’ and operators’ relative experience could also be instructive.

Correlation vs. Causation

In Data Analytics, two common terms are “correlation” and “causation.” Correlation discerns patterns and trends between different types of data that are not necessarily connected but appear to trend with each other, and so one can be used as an indicator of the other. Causation identifies root causes, which can then be addressed to prevent unwanted outcomes (e.g., not using appropriate lubricating oil leads to premature maintenance requirements). Understanding causation may have more impact and therefore more value than correlation.

Correlations can be found more quickly and cheaply than causations and therefore are often preferable. Applying equations and science that model the physical world to Data Analytics can be very powerful in this respect and can sometimes deliver more value.

Data Identification and Selection

Having considered the problem and/or opportunity sets and the initial insights that would be useful, it becomes possible to consider which data sets should be accessed and/or combined for the analysis. The mode of thinking should be “what would be useful in an ideal world,” and not be constrained by what data are known to exist. In this way, new data sources not currently in place or available could be conceived and implemented (e.g., by installing new or additional sensors).

To further illustrate this point, when considering condition monitoring of a pump, vibrational data are clearly useful. If the data are already being generated, then the source of the data must be identified and its integrity and validity for Data Analytics confirmed. If not, new sensors should be considered, of which there are many robust, lost-cost, intelligent examples available today.

Typically, to maximize the value of analytics for a particular problem or opportunity, multifarious data sources and data types will be in scope. For example, historical maintenance and failure records would be in scope of the data to be included. Pump performance data (e.g., flow rates, temperatures, pressures), particularly just prior to the occurrence of issues or failures, would be useful in developing an operating profile. Analysis of the certified skills and competencies of individuals who have undertaken the maintenance work could be useful, in which case permission to use potentially sensitive data might be sought from the Human Resources department, and so on.

Depending on the nature of the problem or opportunity being addressed, the number of candidate data sources for analysis may be large and prioritization will be required.

Internet of Things (IoT)

To date, the Internet has primarily served communications between humans, particularly in its application in the World Wide Web, email, social media, etc. The Internet-of-Things (IoT), sometimes referred to as the Industrial Internet of Things (IIoT), simply refers to Internet-enabled communications between nonhuman entities, such as devices and equipment, with data stores, applications and computers. The idea is that items as diverse as temperature sensors and washing machines can easily connect to IP networks and communicate data about their own condition or what they are measuring and also respond to external requests for data and execution of certain actions.

The IoT revolution has primarily been driven by the convergence of two factors. One of them is the proliferation of GPS technology. The second is the rapid increase of high-speed wireless Internet connectivity. Gartner has estimated that the number of connected devices will reach 25 billion by 2020 (Gartner Inc., 2014). In 2014 CISCO predicted that IoT will be a US $14.4 trillion industry in 10 years. (Nedeltchev, 2014).

In addition to selecting the data sources, the subsets of data from different selected sources should be clearly scoped out. This is where the subtleties of Big Data become apparent. It is not always necessary to look at all data, although in some cases it will be useful to do that, particularly when on a discovery mission to identify new patterns or correlations across large disparate data sets. Also, some applications will require analysis of real-time data, where all data generated are analyzed from a continuous stream. However, in many cases the data required can be honed in on and "right-sized," whether it be according to a particular time period, a particular equipment type, a geographic location, etc.

At the end of this exercise, a list of prioritized and well-defined data sources will have been created. Details of source systems or providers and data locations will have been gathered, as will relevant organizations and personnel from whom to get the data relevant to the objective.

In addition, a list of gaps in data will have been generated and prioritized and then addressed as appropriate; for example, through generation of additional data via new sensors.

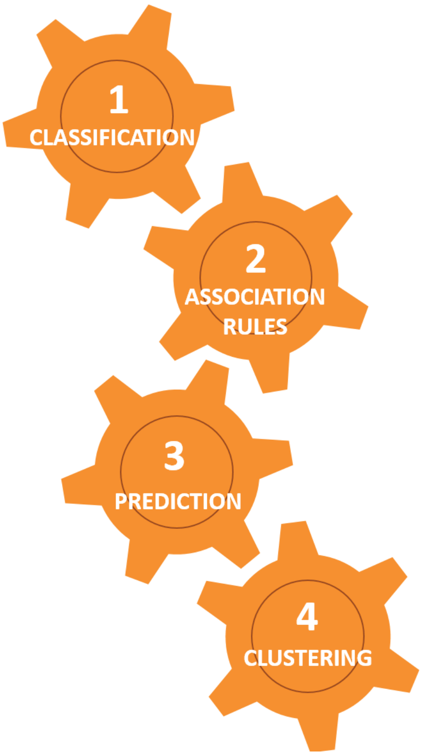
Businesses that take a strategic approach to Data Analytics across high-priority challenges or opportunity sets for a particular business will make investments across the enterprise in new data sources. A strategic approach to investment in new data sources will help advance the maturity of the enterprise with respect to data-driven performance management.

Both the Insight Definition step and the Data Identification and Selection step can be iterative, because considering the data available to solve a problem can give rise to ideas about additional, and perhaps better, insights not originally considered, which may in turn inform on the data sources required.

Method Selection

Data Analytics is largely about identifying trends and patterns in data and between data to create insight. Many different methods can be applied, and it is essential to select those that are most appropriate to realize a particular opportunity or solve a specific problem. There are four (4) main types of Data Analytics methods:

1. Classification. Classifying a new observation on the basis of historical observations that have already been categorized (e.g., assigning a given email into "spam" or "nonspam" classes)
2. Association Rules. Discovering interesting relations between variables in large data sets (e.g., if a large plane is leaving London Heathrow Airport in the afternoon and is heading East, it is likely destined for the Asia Pacific region)
3. Prediction. Exploiting patterns found in historical and transactional data to identify risks and opportunities associated with current data (e.g., gradually increasing vibrations of a certain frequency in a particular type of pump may show that it is about to fail)
4. Clustering. Grouping a set of objects in such a way that objects in the same group are more similar (in some sense or another) to each other than to those in other groups (e.g., all ball valves with mean time between failures greater than 5 years that have never been operated in the tropics)



SECURE THE DATA

Data Generation

Data generation is the act of creating and storing data from the identified sources. When it comes to asset-intensive industries, this may include access to data from the industrial control systems for equipment.

Role of Equipment Vendors

Equipment and control system vendors have a key role to play by sharing the data they generate and internalize with owners, operators and other industry stakeholders. These data can be combined with other data sets to potentially reveal performance interdependencies between different vendors' equipment and/or configurations.

Data Qualification

Due attention should be given to qualifying data for analysis, including data quality assessments. If the quality of the collected data (particularly from older, legacy data stores) is suspect, then data cleansing may be required. This step may not be required, particularly if there is already high confidence in the provenance of the data and especially if the data analysis will apply statistical methods to remove "noise" caused by exceptional or outlying values in the data. Data qualification will recognize faulty data, apply clever techniques to close gaps or correct data (such as linear regression), or eliminate it from the sample data set and make clear its absence.

Data Integration

Data integration is the act of securely integrating data from the source databases or data channels into a specially designated data store on which to perform the analytics. Over the last few years, new approaches to creating and managing large data sets have been developed.

Some companies will already have invested in a Data Warehouse (DW), which is a database that collects and stores integrated sets of transactional data from multiple operational systems. A DW is intended to enhance the flexibility of an organization to analyze and discover new information important to operations of the organization. It is designed for the purpose of query and analysis rather than transaction processing or shared data storage. A DW is a fundamental building block of a Business Intelligence (BI) system. The capabilities of a typical DW deployment are (1) an Extract, Transform and Load (ETL) solution, (2) an Online Analytical Processing (OLAP) engine and (3) client analysis tools and other applications that manage the process of gathering and delivering data to the business users for intelligent business decision making. There are three (3) common types of DWs:

- Enterprise Data Warehouse (EDW). Contains transaction data, summarized data, detailed data, semi-structured data and possibly voluminous unstructured data extracted from various systems such as Enterprise Resource Planning (ERP), Customer Relationship Management (CRM) and other external data sources
- Data Mart (DM). Collection of subject areas or subsets of the DW organized for decision support based on the specific needs of a given user group
- Operational Data Store (ODS). Contains current or near real-time, summarized structured data in a subject-oriented database extracted directly from operational transaction system data sources

Secure Environments

Due diligence must be applied to ensuring that sensitive data are secure – both the source data and the data generated by Data Analytics to derive insights. The data should be categorized and segregated according to the sensitivity of the source or of the potential ensuing decisions, and appropriately robust data security mechanisms should be put in place for each.

If operational data are being integrated from equipment and control systems, then particular attention will need to be given to avoid compromising the integrity of those operating environments when the data are being accessed and integrated. Typically, the SCADA and CONTROL systems environments are to be kept highly secure to avoid potentially catastrophic hacking. Recent examples of hacks into the control systems of operating automobiles demonstrate the importance of this security.

Data first need to be extracted from the secure environment into a Data Historian, which, as its name suggests, is where historical data (typically contiguous historical data) are stored for access. This is the original copy of the data generated in the SCADA and CONTROL systems and must be preserved as such. Therefore, tight controls must be in place as to how data are accessed and extracted from the Data Historian.

EXTRACT THE INSIGHTS

Data Analysis

Having obtained the data and selected the analytical methods, the next step is to apply the selected methods to the gathered and prepared data sets and execute the analysis of the data. The analysis could take the form of a one-off exercise on a huge historical data set or a continuous, real-time analysis on data as they are generated.

Insight Delivery and Visualization

The analysis will deliver insights in the form of trends, patterns, anomalies and forecasts. The notional "half-life" of a response, which indicates how long an insight will be valid, useful, or indeed valuable, must also be considered. In general, the longer the half-life of an insight, the more strategic its nature will be; conversely, the shorter, the more tactical.

Effective visualizations can be powerful tools in making complex information more accessible and easier to understand. Data visualization tools and techniques have evolved rapidly over the last several years, providing analysts with the capability to visualize multiple dimensions of an issue simultaneously to efficiently identify trends and correlations within the information. Examples of data visualization techniques include charts, graphs, reports, heat maps and dashboards, and tools can be used to generate static outputs or provide interactive exploration of the data set.

The way in which an insight is delivered will depend on the nature of the insight and the overall objective. If the objective is to illustrate a long-term business trend, then a report with supporting graphs will be appropriate. This is a long half-life insight example, which can afford discussion and deliberation among executives to decide whether and how to respond.

If the objective is to verify safety in response to a recognized anomaly in the process, then the insight will be an alert requiring an immediate response, with a graphical representation of supporting data to show the issue that triggered the alert.

DEFINE AND EXECUTE THE RESPONSES

Response Triggers

Some insights will be informational, but others will be prescriptive, involved actions that are referred to as Response Triggers. For short half-life responses, time is of the essence and an appropriate response will need to be invoked and executed quickly, potentially in seconds (e.g., actuate an emergency shutdown valve). Automatic responses are appropriate in cases such as these and need to be "triggered." Relevant experts must carefully prescribe the criteria for the triggered action(s).

Response Workflow Design and Implementation

In some cases, human intervention will be required as a response. The workflow for intervention needs to be sensibly thought out. For example, condition monitoring of equipment may indicate that a failure is imminent and require notifying maintenance of the possible failure modes and maintenance procedures. Maintenance processes must be updated to incorporate this insight and realize the benefits, which include more accurate troubleshooting and more efficient maintenance. Whatever the response, the people who execute it must be appropriately trained and ready for invocation at all times.

Outcome Review

Performance improvement driven by Data Analytics is not an exact science. The actual results and consequences of completed actions need to be compared with the expected results and any differences noted, so that the algorithms applied and/or the response actions invoked can be reviewed and refined, as appropriate. This will provide a feedback loop to the overall analytical process.

CONTINUOUS IMPROVEMENT

Data Analytics is an enabler of continuous improvement in an organization and, crucially, continuous improvement is critical to a highly effective and sustainable application of Data Analytics.

Continuous improvement should be applied to Data Analytics at two levels: (1) to each Data Analytics project and (2) to the way Data Analytics is being applied across the enterprise.

The approach to any Data Analytics initiative should be iterative, aiming to realize the key objectives according to the specified KPIs (see Step 2 above). A detailed review should be undertaken at the end of each cycle. The following questions are examples of what might be considered in the review:

- What data were used? Were the data the right data?
- How were the data analyzed? Can the analytics methods be improved?
- What types of insights were obtained? Were the right criteria used to trigger a response?
- How effective was the response? Was its execution efficient?
- What were the outcomes? How well did they meet the objective?
- How can the whole cycle be improved? What are the next iterations?

Regularly sharing information and insights about Data Analytics initiatives across business “silos,” and using collective feedback and reviews of what worked well and what could be done better, will maximize the value of Data Analytics to an enterprise and help ensure that the value is sustained. It will also allow successes to be recognized, celebrated and shared.

References

Gartner Inc. (2014, November 11). Gartner Says 4.9 Billion Connected "Things" Will Be in Use in 2015. Retrieved November 25, 2015, from Gartner: http://www.gartner.com/newsroom/id/2905717

Nedeltchev, P. (2014, January 15). The Internet of Everything is the New Economy. Retrieved November 25, 2015, from CISCO: http://www.cisco.com/c/en/us/solutions/collateral/enterprise/cisco-on-cisco/Cisco_IT_Trends_IoE_Is_the_New_Economy.html

ABOUT ABS GROUP

ABS Group of Companies, Inc. (www.abs-group.com), through its operating subsidiaries, provides a range of technical solutions to support the safety and reliability of high-performance assets and operations in the marine; offshore; oil, gas and chemical; power; and government sectors. Headquartered in Houston, Texas, ABS Group operates with more than 2,000 professionals in over 30 countries. ABS Group is a subsidiary of ABS, a leading marine and offshore classification society.

www.abs-group.com

ABS Group