From Data to Knowledge: The Process of Knowledge Discovery

Executive Summary

The proliferation of data in the last decades has reached astronomic figures and has led to new ideas about how to get the most out of the data. A clear border has been drawn between data, information, and knowledge. Raw data is considered as a new resource and a multitude of techniques have been developed to analyze it, break it down and refine it to extract the valuable knowledge it contains. That is known as Knowledge Discovery in Databases, also called KDD.

In the oil and gas industry, new challenges are pushing managers to optimize and improve operational efficiency. Using KDD to extract knowledge from historical data and applying it to maximize productivity is one of the most cost effective ways to operations' optimization. However, KDD involves a comprehensive process that encompasses a set of multidisciplinary activities such as collecting and cleansing data, choosing the data-mining algorithm that best fits a given task, and interpreting results. This process requires the right tools for data discovery and decision support. The trouble with these tools is that they can be hard to use. The friendliness of the user-interface is a challenge that needs to be addressed by a new generation of knowledge mining products. The objective is to offer a simple interface that enables almost everyone to process data.

KNet Analytics is an out of the box software application package designed to help engineers and operations in the manufacturing to transform their massive and unstructured data into knowledge and actionable decisions.

KNet Analytics empowers the casual users to use advanced analytics and machine learning algorithms to understand the depth of any process by generating expert rules, predictive models and fault propagation models in few clicks. KNet Analytics provides value immediately out of the box and does not require project implementation efforts. Users get instant gratification out of the box. KNet Analytics can be used for many different reasons such as behavioral patterns, diagnose abnormal conditions, understanding of the cause and effects of events, detection of different plant states and operating models, dynamic targeting, reverse engineering of your processes, connecting the dots together with cause and effect diagram, production optimization, plant behavior prediction, identification of suspicious sensor values and lab data.

Table of Contents

The Importance of Knowledge in Industry

The new digital landscape of the industry has led to data explosion at every stage of the business lifecycle. A report[1] from EMC[2] – was done in partnership with IDC[3] – tries to quantify the amount of data, its sources, and expected evolution:

- What the company calls "digital universe" currently contains 4.4 trillion gigabytes (4 400 000 000 000 GB) and it is doubling every two years. At this rate, it will reach 44 trillion gigabytes in 2020,
- 60% of data in the digital universe is currently attributed to mature markets such as Germany, Japan, and the United States, but by 2020, the percentage will flip, and emerging markets including Brazil, China, India, Mexico and Russia will account for the majority of data,
- In 2014, less than 20% of the data was "touched" by the cloud. In six years, that percentage will double to 40%;
- 18% of data in 2014 will be generated by mobile devices (include RFID tags, GPS devices, smart cards, cars, toys and dog collars). By 2020, this ratio will reach 27%,
- 66% of the digital universe bits are created or captured by consumers and workers, yet enterprises have liability or responsibility for 85% of the digital universe.

This new landscape has led to new ideas of how data can efficiently be used to maximize its value. The real imperative is the potential insight we can derive from this new, vast, and growing resource which is data. The amount of added value, this resource can offer, brings experts to claim: **"Data is the new oil"!** Michael Palmer, from the Association of National Advertisers, blogged[4] in 2006: *"Data is just like crude. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc., to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value."*

The oil and gas industry and the process industries are no exception to this trend. A set of new challenges emerged in the last decades with more visible impacts. The obvious challenge of managing a finite resource, the ever-changing environmental regulations and the consequent competitive pressures, makes it understandable why industry leaders are continually striving to find ways to optimize and to improve operational efficiency. One of the most effective ways to this improvement is plant historical data analysis, with a focus on optimization. In the oil and gas industry and process industries, "Data is the new oil" could not be more significant.

Data, information, and knowledge

As stated in Michael Palmer's quote, data cannot be used if not analyzed. Data is raw. It simply exists and has no significance beyond its existence. It can exist in any form, usable or not. In general, it consists of symbols that have no meaning if not linked to a specific context.

Information emerge when understanding relations takes place. It is an organized or structured data, which has been processed in such a way that the information now has relevance for a specific purpose or context, and is therefore meaningful, valuable, useful, and relevant. Example: If the temperature drops 15 degrees, it will rain.

Knowledge represents a pattern that connects and generally provides a high level of predictability similar to what is described or what will happen next. Example: If the humidity is very high and the temperature drops, substantially, the atmospheres is often unlikely to be able to hold the moisture so it will rain.

[1] http://idcdocserv.com/1678

[2] EMC is a multinational corporation considered as the world's largest provider of data storage systems

[3] IDC : International Data Corporation

[4] http://ana.blogs.com/maestros/2006/11/data_is_the_new.html

People and computer applications, both, depend on data and information, but effective decision-making requires more than merely data and information embedded in workflow processes. Rather, knowledge is the key ingredient. In industry, the ultimate goal is to obtain knowledge in order to assist in the decision-making process. A typical figure describing the essential associations between data, information, and knowledge is shown below.

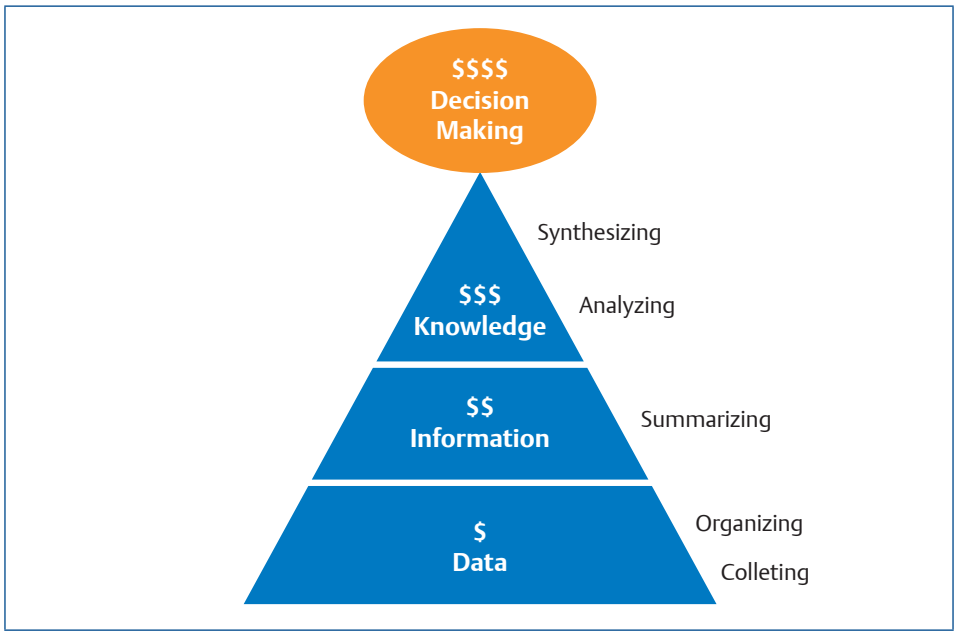

Figure 1: The links and process of converting data into vital knowledge

The importance of capturing operational knowledge

Departures of experts from the industry are adding to the stress endured by companies and leave a large vacuum of needed professionals where there are already shortages[5]. So how can the priceless knowledge be captured and made readily available during an abnormal situation or to the benefit of incoming generations of operators? What types of knowledge are required and how can it become operational intelligence?

Knowledge Discovery in Databases (or KDD) is a solution for the problem of retiring professionals. It provides a way to reach knowledge without needing the help of retiring professionals. In fact, KDD does not focus on capturing the knowledge from the professionals like expert systems do; it rather gets the knowledge directly from the plant data.

The Process of Knowledge Discovery in Databases (KDD)

KDD includes multidisciplinary activities. It encompasses data storage and access, scaling algorithms to massive data sets and interpreting results. Artificial intelligence also supports KDD by discovering empirical laws from experimentation and observations. Steps involved in the KDD process are the following:

1. Understand the requirements and set objectives

This includes understanding the background in which the data has been generated and identifying the goal of the KDD process from the customer's perspective.

[5] *Jodie Humphries: Oil and gas workforce - a shortage in skilled labor?; Oil & Gas Next Generation Magazine; February 2010*

2. Select and collect data

Select a target data set or subset of data samples on which discovery is to be performed. This is done through the choice of the potentially useful variables as per the objectives and the requirements. The collection can be done through extraction from existing databases or through new measures or surveys.

3. Clean collected data

This step includes:

- Handling missing fields, bad measures and outliers,
- Checking the validity of data and whether it satisfies data-type constraints, range constraints, mandatory constraints, etc.
- Checking the uniformity of data, for example the use of the same unit of measure in the entire dataset.

4. Pre-process and transform data

Pre-processing involves several techniques and strategies. Filtering, sampling, discretizing, normalizing, and decomposing are some of these techniques whose aim is to prepare the data for more advanced algorithms. The necessity of pre-processing is determined by the properties of the original dataset. Filtering for example is used with noisy signals to obtain a higher signal to noise ratio. Sampling helps extracting data subsets with smaller size when the original dataset is too big to be processed at once. Normalizing is applied to homogenize variables that differ by orders of magnitude.

5. Choose the data mining task

This step is the answer to the question: “What are we trying to mine?” It is the mathematical and statistical translation of the requirement and objectives. Examples of data mining tasks are:

- **Frequent patterns mining:** Frequent patterns are patterns that occur frequently in transactional data. For example a set of alarms that frequently appear together.
- **Clustering:** Cluster refers to a group of similar objects/values. Clustering refers to forming groups of objects/values that are very similar to each other but are highly different from the objects/values in other clusters.
- **Classification:** Classification requires finding a model that describes the data classes or concepts. The purpose is to use this model to best determine the class of an object whose class label is unknown, using attributes of that object.
- **Prediction:** It is used to predict values or events in the future. Regression analysis and time series models can be used for prediction when input variables change.

6. Choose the data mining algorithm

One data mining task can be accomplished through different algorithms. If we take clustering as an example, the most common algorithms are the k-Means algorithm, the Gaussian mixture models, and the OPTICS algorithm. Each algorithm has advantages and drawbacks which might make it difficult to choose the appropriate algorithm.

7. Execute the data mining algorithm

Executing an algorithm requires choosing the set of parameters needed to execute the algorithm. Continuing with the clustering example, the k-Means algorithm needs the number of clusters as an input parameter. OPTICS requires two parameters: the neighborhood distance and the minimum number of points required to form a dense region. Major attention should be given to the choice of the input parameters since it can considerably impact the output of the algorithm.

8. Evaluating results

The output of the data mining algorithm should be translated into knowledge after interpretation and contextualization. For example, clusters in a dataset can be interpreted as the result of different operating modes. In this case, the operating mode corresponding to each cluster should be identified. If the discovered knowledge does not meet expectations, a couple of steps back in the data mining procedure would be useful. One can re-execute a data mining algorithm with a different set of input parameters, use another data mining algorithm for the same task, substitute the data clean up method with an alternative one, or go back to the list of chosen variables and add/remove new variables to capture more accurate knowledge.

A common way to evaluate the results is to test them on data not used during model building - splitting data into “test” and “training” sets for choosing model parameters like the number of clusters. The recognized patterns must be valid on test data, and possess some degree of certainty. Various procedures for doing this are called cross-validation. In particular, users must avoid the problem of “overfitting”, where the data fits the limited training set very well, but gives poor results using other data such as test data (does not “generalize” well). As an example involving regression, a third-degree polynomial for just four data points will perfectly fit, but probably do poorly with any other input data. Once the parameters, like the number of clusters, are chosen using cross validation, then the entire data set can be used for final training.

9. Use the discovered knowledge

Use the newly discovered knowledge and incorporate it into another system for further action. The discovered knowledge must be represented in a form appropriate for the intended user. For human end-users, appropriate representations include natural language, formal logics, and visual depictions of information. Here, the friendliness of the knowledge mining tool has an important impact on the intelligibility of the results for human end-user.

But discoveries can also be intended to other computer programs; in this case, programming languages are more appropriate representations. Particular attention should be given to the maintainability of the code containing the extracted knowledge: the incorporation of the code should be easy each time the models are updated or fine-tuned.

10. Document the KDD process

The whole process ought to be documented and reported for interested parties. This includes making reports to describe the execution of the previous steps and the added value or the expected benefits of the knowledge discovery process.

The Need for Adequate Software Products

Choosing a knowledge mining tool that best fits your need is an important step in the knowledge discovery process. The trouble with statistical packages and quantitative analysis tools is that they require strong technical skills. Besides, making the best decision, when it comes to choosing a data mining algorithm and its parameters, might require a lot of learning or a strong theoretical background.

Some analytics software products are easier to learn and use than they used to be, and documentation is usually available on line from the vendor website or as part of the help system. However, the friendliness of the user-interface is still a challenge that needs to be addressed by a new generation of knowledge mining products. The interface needs to be simple to the extent that even people who have no knowledge of programming or statistics would be capable of processing data, provided they are familiar with software usage, can follow simple instructions, and understand the basics of their plant so they can interpret the results such as clusters and other discovered relationships.

Tools for KDD

Emerson offers a unique and an integrated solution for operational intelligence that covers the entire process from data extraction to online actionable decisions including modeling and knowledge discovery. The solution is based on the combination of two powerful and complementary software products:

- KNet Analytics
- KNet

These latest technologies empower oil & gas and other process industries with advanced analytics, modeling, data pre-processing, models deployment online, advanced expert systems applications, performance management and decision support.

Our solution consists of two main phases:

Phase 1: Knowledge discovery with KNet Analytics

KNet Analytics is an out of the box software application package designed to help engineers and operations in the manufacturing to transform their massive and unstructured data into knowledge and actionable decisions.

Knowledge Analytics empowers the casual users to use advanced analytics and machine learning algorithms to understand the depth of any process by generating expert rules, predictive models and fault propagation models in few clicks. KNet Analytics provides value immediately out of the box and does not require project implementation efforts. Users get instant gratification out of the box. KNet Analytics can be used for many different reasons such as behavioral patterns, diagnose abnormal conditions, understanding of the cause and effects of events, detection of different plant states and operating models, dynamic targeting, reverse engineering of your processes, connecting the dots together with cause and effect diagram, production optimization, plant behavior prediction, identification of suspicious sensor values and lab data.

This graphical environment also makes the KDD process easy for any engineer. End users can:

- **Load data** from several data sources including SQL Server, Oracle databases, PI historians, XML files, Excel files and text files.
- **Enhance the data pool** by using the data enrichment feature for replacing missing values by averages, minima, maxima or manually specified data,
- **Clean datasets** by deleting bad or incomplete data and detecting outliers for more accurate models,
- **Preprocess data:** Merge, duplicate, transpose, match, segregate, sort, filter and classify large datasets with few clicks,
- **Apply advanced algorithms** to determine correlations between variables and identify data clusters,
- **Create charts** to visualize data as well as the results of the models,
- **Validate models** automatically and expose the results in terms of errors, reliability and accuracy.

KNet Analytics offers pre-processing tools as well as advanced algorithms in order to allow end users, without the need of extensive domain expertise, to:

- Perform proactive decision-making to adjust their plans according to the data models
- Perform descriptive and predictive analytics using principal component analysis, neural networks, and linear and non-linear multivariate regressions such as rational regression, or polynomial regression
- Identify correlations and patterns using covariance analysis, correlations matrices, lagged-correlations, association rules, and classification algorithms
- Generate expert rules by building advanced decision trees

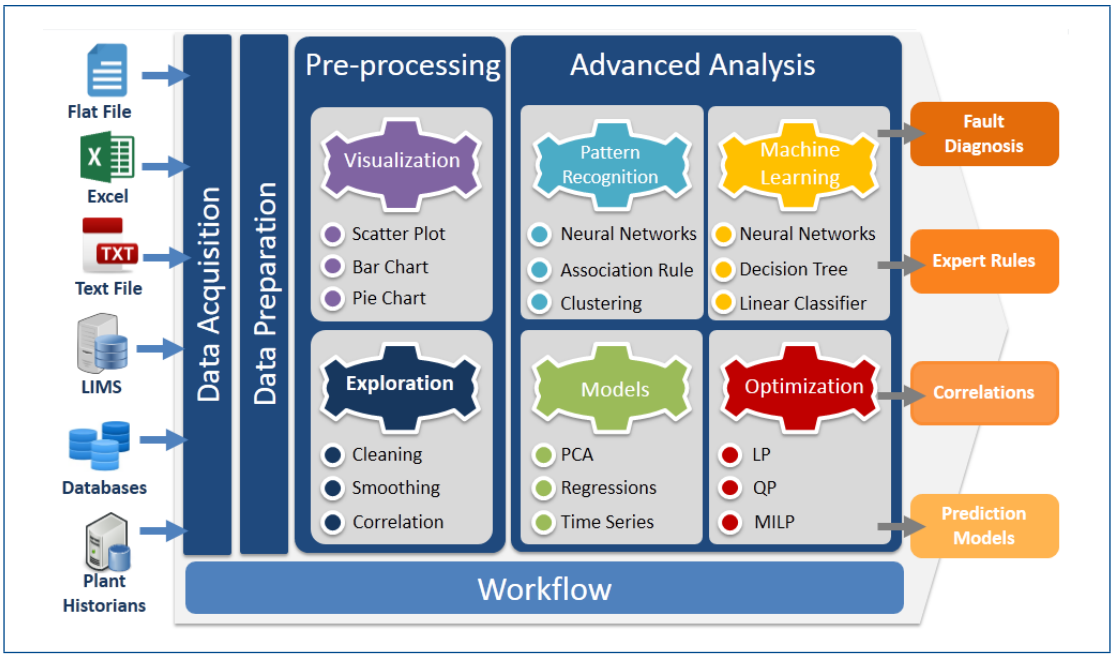

Figure 2: KNet Analytics features overview

The models built inside KNet Analytics can be exported online by automatically generating software plugins in dll format. These plugins can be easily called in the KNet framework and any environment supporting .NET. Any uncovered information and knowledge can also be configured as rules or fault trees using KRules and KRCA modules.

Phase 2: Online implementation with KNet

The online phase objective is to deliver real-time visibility and insight into operations for optimized performance management, maximized throughput, efficient management of abnormal conditions, and tight control of operating costs. To do so, KNet combines several modules and features to address the requirements of intelligent real-time expert systems applications such as:

- Exporting generated models from KAnalytics to the KNet framework for online execution,
- Online model-based reasoning using rules,
- Plant state and operating modes detection,
- Abnormal conditions detection, prediction, and diagnosis,
- Automating procedures through workflows,
- Data sources integration.

The figure below illustrates the entire process from data extraction to online actionable decisions.

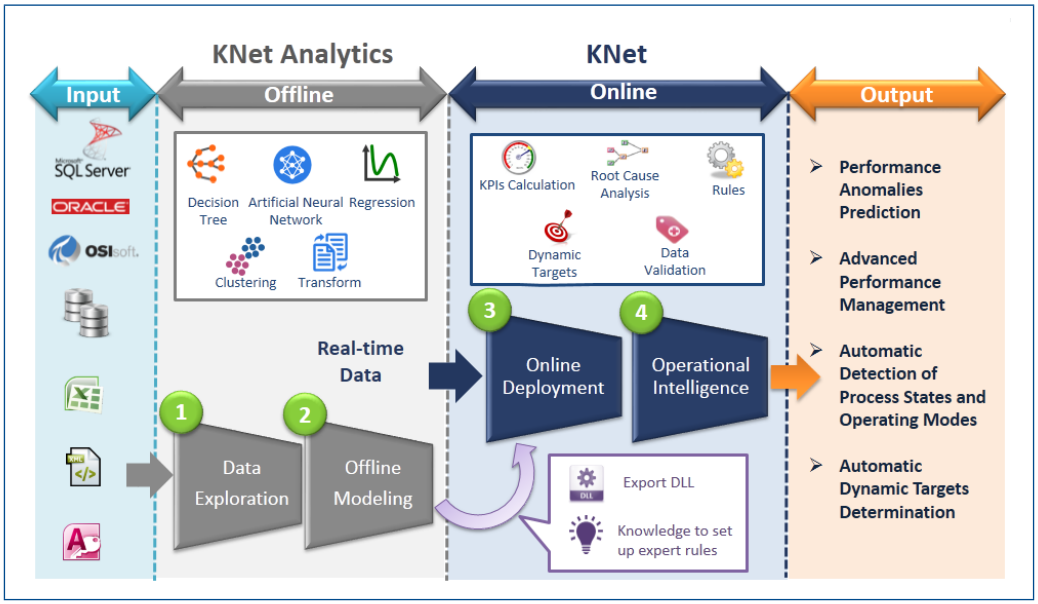

Figure 3: From historical data analysis to online operational intelligence

Keywords

- Data
- Information
- Knowledge
- Knowledge Discovery in Databases
- Advanced Analytics
- Industry

Emerson
Reliability Solutions
835 Innovation Drive
Knoxville, TN 37932 USA
+1 865 675 2400

www.emerson.com/ams