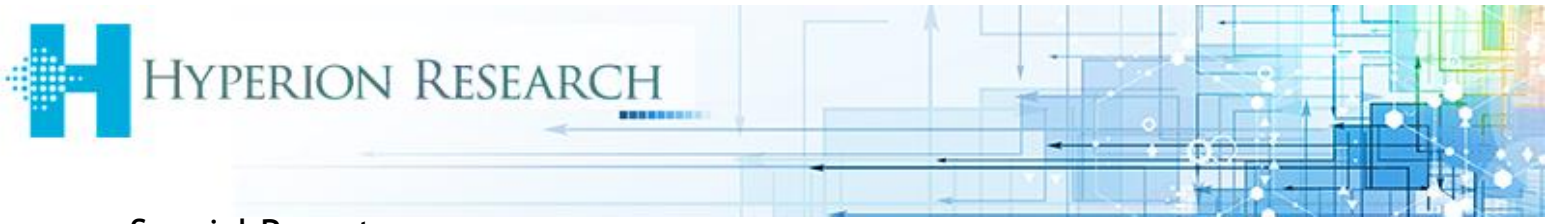


Special Report

Mapping Applications to Algorithms to Architectures: Data Ingestion, Machine Learning, Numeric Optimization and Data Mining

Alex Norton, Earl Joseph, Steve Conway, and Bob Sorensen
June 2018

HYPERION RESEARCH OPINION

This paper is the third of a four-part series based on a pioneering study of "big data" algorithms Hyperion Researched conducted in 2017.

Algorithms embodying mathematical models are the main intellectual capital and competitive weapons for advanced analytics in the commercial, academic, and government sectors. Hyperion Research defines advanced analytics problems as those that need high performance computing (HPC) resources to run effectively. We call this global market high performance data analysis, or HPDA. AI methodologies such as machine and deep learning are a subset of the HPDA market.

The variety of existing algorithmic types/subtypes is at least as large as the variety of HPDA problem types/subtypes the algorithms were designed to address. The study creates a comprehensive taxonomy of HPDA problem types and the preferred algorithm(s) for addressing each type. The study characterizes each problem type (application), along with its hardware and software requirements, and identifies the important attributes of each algorithm type associated with the problem.

The study presents this large collection of information in a series of Excel tables, with definitions provided for each row label (problem type/subtype) and column (problem attributes and hardware-software requirements). Hyperion believes that this layout will enable users to zero in on the data that is most relevant for addressing their own HPDA problems. An algorithm widely used for commercial fraud detection might, for example, have attributes closely matching those of a government problem that also involves analyzing faint signals. In that case, information about the commercial algorithm may be applicable to the government problem.

This final step of having users identify the nearest-equivalent algorithm is needed because precise algorithm matches are rare in the young, immature HPDA market. Even the algorithmic taxonomy used in this study represents a consensus of expert opinions Hyperion gathered, rather than a pre-existing standard classification on which everyone agrees.

One thing the experts surveyed for this study do agree on is the value of smart algorithms. As a general rule, smart algorithms need less data/fewer iterations to home in on useful solutions, and less-clever algorithms require more data/iterations. Hyperion estimates that demand for skilled algorithm developers will grow by 25-30% by 2020 while the supply of qualified job candidates remains fairly constant. Hence, competition for skilled algorithm developers will increase and they will be able to command higher salaries than today.

TABLE OF CONTENTS

LIST OF TABLES

IN THIS STUDY

Study Objectives

This document is part of a four-part series, which breaks out the larger, full study document into smaller, more specific segments of the full data set. The main objectives of the larger study were to:

- Consult experts to develop a taxonomy matrix that matches advanced analytics (HPDA) application (problem) types with the algorithms users prefer to employ for addressing these applications.
- Identify in detail the hardware-software requirements of the applications and the attributes of the algorithms that generate those requirements.
- Present these findings in a report designed to be employed as a reference tool for HPDA users (especially non-HPC specialists) from the broad spectrum of application domains investigated in the study. The report should especially help users to zero in on the data that is most relevant for addressing their own HPDA problems—with the understanding that users will often need to take the final step of translating the findings into the specific contexts the users operate in.

RESEARCH METHODOLOGY

The starting point for constructing an HPDA algorithm matrix was the taxonomy of HPC domains and sub-domains Hyperion closely tracks on a worldwide basis—some from as early as the 1980s and others added 4-5 years ago as new trackable HPDA use cases emerged in the commercial sector (e.g., fraud detection, affinity marketing, and advanced business intelligence). This study report includes a list of the domains and sub-domains, with a definition for each.

As with any first-of-its-kind study, our HPDA algorithm investigation next needed to map the uncharted landscape by consulting with users who are well-known experts in big data analytics that require HPC resources. Because there was no pre-existing, standard classification of HPDA algorithms (the purpose of the study was to create one), Hyperion asked each expert to submit a proposed taxonomy and then refereed an extended discussion among the experts to arrive at the consensus algorithmic taxonomy used in the study—i.e., a taxonomy that none of the experts objected to and that all said they could support.

The next step was to create a survey instrument that reflected the many possibilities inherent in the study, by including hundreds of questions but subjecting each respondent only to the subset of questions relevant for that respondent's area of expertise. The solution, not surprisingly, was a survey instrument that led respondents through branching questions. A bio-sciences expert, for example, would first be asked questions common to all respondents, then a set of bio-sciences questions, and finally (if appropriate) a set of genomics questions. A product design expert would be taken down an entirely different branching path after completing the initial set of common questions.

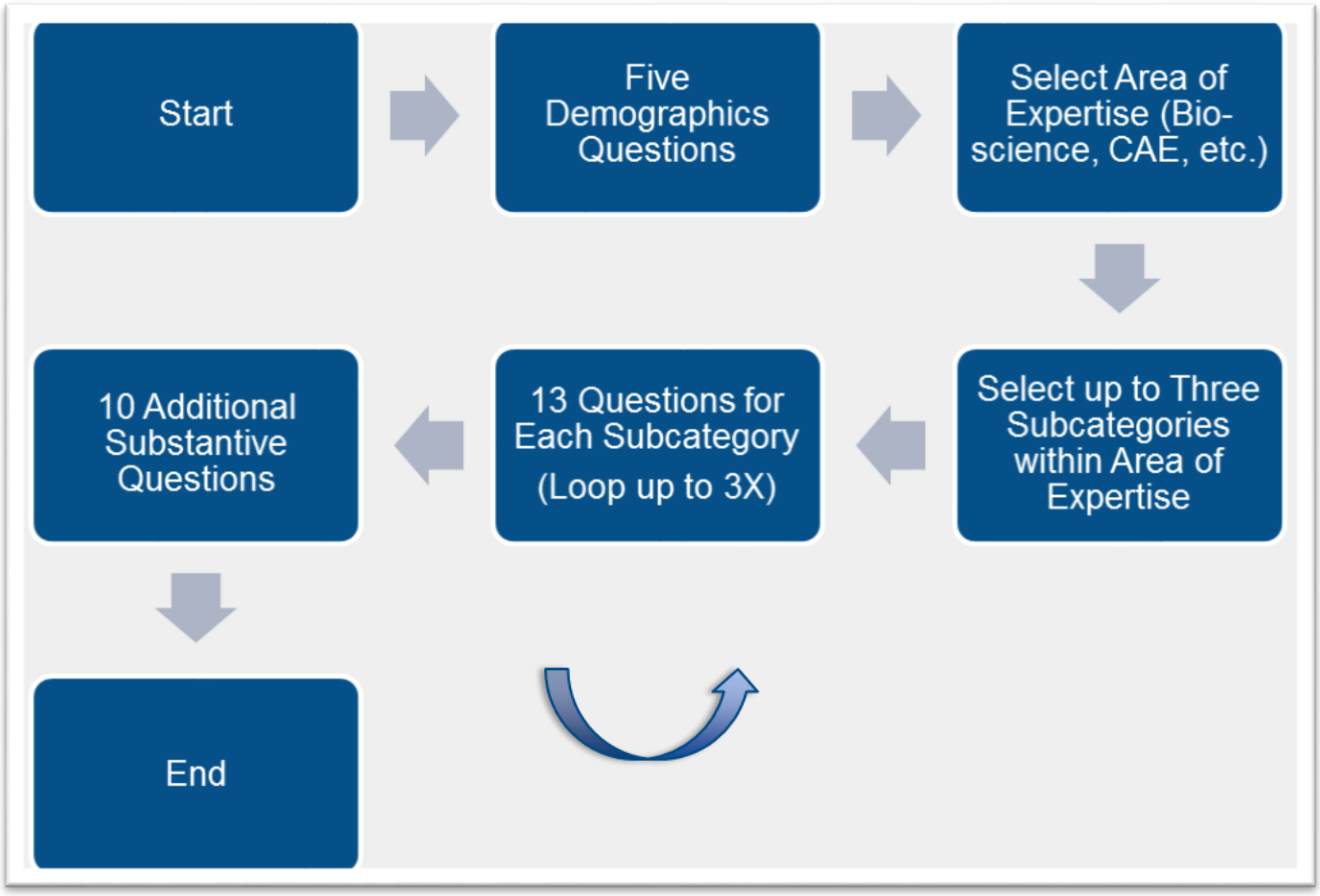
This methodology produced the exceptionally large collection of data that appears in the tables contained in this report.

Constructing a Comprehensive Survey Guide

As mentioned in the prior section of this report, the broad scope of this effort—and the wide range of questions asked of the surveyed experts—required that a relatively complex flow chart for the survey questions be used to adequately address the major study issues, keep the response rate high, and avoid the high dropout rate that would almost surely have occurred if every respondent had been subjected to the entire question set. To accomplish this, the survey used a branching/looping mechanism to allow each expert to drill down into their particular area(s) of interest but in a manner that reduced the overall number of questions each expert was ultimately asked to answer.

- The full survey consisted of 715 questions, but no individual expert ever saw the complete set. Depending on their willingness to provide answers and insights, each expert could choose to answer as few as 29 or as many as 55 questions.

The specific flow chart for the survey was as follows:



GENERAL REQUIREMENT DISTRIBUTIONS

Table 1 collapses the sub-segments (e.g., Genomics, Proteomics) into single rows representing the larger application domains (e.g., Bio-Sciences). In this collapsed view, the experts' preferences for data integration become clear. In five of the nine main categories, data integration was considered a critical (more than 75% of responses) data ingestion requirement, and three categories considered it their only (100% of responses) important data ingestion requirement.

- Commercial analytics stands out as the category most interested in data fusion requirements and least interested in data integration for those category experts expressing an interest in both.
- In what is an emerging trend in this study, Government labs continue to express some of the broadest and most diverse requirements.
- At this level of aggregation, the data ingestion requirements of Bio-Sciences and Economies/Financial categories are quite similar.

Table 1

Overall Distribution of Data Ingestion

Vertical	Data Fusion	Data Reduction	Data Integration
BIO-SCIENCES	11%	11%	78%
CAE: PRODUCT DESIGN			100%
CHEMICAL ENGINEERING			
COMMERCIAL ANALYTICS	54%	15%	31%
DEFENSE			100%
ECONOMICS/ FINANCIAL	13%	13%	75%
GOVERNMENT LAB	18%	41%	41%
MECHANICAL DESIGN			100%
WEATHER/ CLIMATE	100%		

Source: Hyperion Research, 2018

Definitions for the machine learning column labels in Table 2 are as follows:

- **Unsupervised:** in this form of machine learning, the computer is given the objective, trained in the rules for reaching the objective, and then turned loose on its own to pursue the objective without human help.

- **Semi-supervised:** here the computer is given the objective, trained in the rules for reaching the objective, and then turned loose on its own to pursue the objective, but with human intervention when needed.
- **Supervised:** in this mode the computer is given the objective, trained in the rules for reaching the objective, and continuously monitored and, when needed, guided by a human to pursue the objective.
- **Reinforcement Learning:** in this mode of machine learning, conceptually based on behavioral science, the computer is trained to take the best actions for maximizing performance.
- **Pattern Recognition:** in this very general mode of machine learning, the computer is trained to identify new visual patterns (e.g., images, human faces, or alphabetical characters) or sounds (e.g., natural language) based on the similarity to images or sounds with which the computer has been made familiar.

Table 2 is broad table that shows the preferences for machine learning applications. The experts' choices show that the categories of Bio-Sciences and Government labs have the most diverse requirements, although Bio-Sciences shows a strong preference for Pattern Recognition while Government labs cover the entire range of potential options.

- No category was wholly committed to a single machine learning requirement and only two of the entries here rise above 60 percent.
- Commercial analytics stands alone as the most interested in Reinforced Learning requirements. Even then, that interest is equally spilt across Unsupervised Learning, Reinforced Learning, and Pattern Recognition.
- CAE shows the strongest preference for Semi-Supervised Learning.

Table 2

Overall Distribution of Machine Learning

Vertical	Unsupervised	Semi- Supervised	Supervised	Reinforcement Learning	Pattern Recognition
BIO-SCIENCES		10%	10%	10%	70%
CAE: PRODUCT DESIGN		67%	33%		
CHEMICAL ENGINEERING					
COMMERCIAL ANALYTICS	33%			33%	33%
DEFENSE					

Table 2

Overall Distribution of Machine Learning

Vertical	Unsupervised	Semi- Supervised	Supervised	Reinforcement Learning	Pattern Recognition
ECONOMICS/ FINANCIAL	30%	10%			60%
GOVERNMENT LAB	30%	22%	26%	4%	17%
MECHANICAL DESIGN					
WEATHER/ CLIMATE					

Source: Hyperion Research, 2018

Definitions for the column labels in Table 3 are as follows:

Numeric Optimization: these algorithms are designed to find an ideal value for a function, i.e., a value for which maximization and minimization are equivalent. An ideal value can greatly reduce computer time-to-solution and memory use.

- *Continuous* – in this form of optimization, the variables chosen for a function are continuous, that is, there are no gaps between them.
- *Discrete* – here the variables chosen for a function are discontinuous, that is, they cover only a limited range of values, such as all integers.
- *Stochastic* – here the variables in the optimization procedure are random.

Table 3 is a summarized table that shows the preferences for numerical optimization in systems. This table shows that the overall preference was for continuous optimization requirements, followed by discrete, and then stochastic techniques. However, the wide diversity of preferences across categories is clear, as almost all options were selected less than 50% of the time within any major category.

- Commercial Analytics and Weather/ Climate are the only two exceptions, with both selecting continuous optimization as their majority selection.
- At this level of aggregation, the requirements of the Bio-Sciences and CAE product design categories appear similar.

Table 3

Overall Distribution of Numeric Optimization

Vertical	Continuous	Discrete	Stochastic
BIO-SCIENCES	40%	40%	20%
CAE: PRODUCT DESIGN	50%	25%	25%
CHEMICAL ENGINEERING	33%	33%	33%
COMMERCIAL ANALYTICS	100%		
DEFENSE	50%	50%	
ECONOMICS/ FINANCIAL	50%	14%	36%
GOVERNMENT LAB	26%	47%	26%
MECHANICAL DESIGN			
WEATHER/ CLIMATE	67%	33%	

Source: Hyperion Research, 2018

Definitions for the column labels in Table 4 are as follows:

Data Mining: is about extracting useful information from large volumes of data.

- *Query Processing* – involves extracting information from a database without changing the database.
- *Pattern Recognition* – as noted earlier, here the computer is trained to identify new visual patterns (e.g., images, human faces, or alphabetical characters) or sounds (e.g., natural language) based on the similarity to images or sounds with which the computer has been made familiar.
- *Network Analysis* – this involves identifying and evaluating the relationships within a group of entities, using tools such as graph analytics, semantic analysis, and others. The entities in question can range widely, from members of a social media network to people infected by a disease to cells within a human organ.

Table 4 shows the broad categories and their distribution in data mining preferences. When viewed at the main domain category level, all three options for data mining were deemed critical within some categories, such as Bio-Sciences, Commercial Analytics, Economic/ Finance and Government Labs, while five of the nine categories examined (CAE, Chemical Engineering, Defense, Mechanical Design and Weather) had no critical requirements for data mining.

Table 4

Overall Distribution of Data Mining

Vertical	Query Processing	Pattern Recognition	Network Analysis
BIO-SCIENCES	43%	38%	10%
CAE: PRODUCT DESIGN			
CHEMICAL ENGINEERING			
COMMERCIAL ANALYTICS	16%	26%	26%
DEFENSE		67%	33%
ECONOMICS/ FINANCIAL	53%	6%	
GOVERNMENT LAB	24%	35%	19%
MECHANICAL DESIGN			
WEATHER/ CLIMATE			

Source: Hyperion Research, 2018

SPECIFIC APPLICATION AREA DISTRIBUTION OF REQUIREMENTS

As Table 5 shows, when asked about important data ingestion requirements, the surveyed experts were generally spilt between two major preferences: 40% of those surveyed within each subcategory selected a single data ingestion requirement (four chose data fusion and six named data integration), while the remaining 60% of those having data ingestion requirements split their choices across two or even three options.

- Data integration was selected as a data ingestion requirement by fully 60% of all experts surveyed, albeit with varying degree of emphasis (from a high of 100% in nine different subcategories to a low of 30% in one subcategory).
- No expert cited data reduction as their sole critical data ingestion requirement. Although data reduction was seen as a critical ingestion requirement within six of the 24 categories, the emphasis ranged only from a high of 50% to a low of 17%.

U.S. Government scientific research and Commercial Analytics for affinity marketing and business intelligence were the only categories that cited requirements for all three data ingestion options.

Table 5

Specific Distribution of Data Ingestion

Application Area	Data Fusion	Data Reduction	Data Integration
BIO-SCIENCES			
Genomics			
Proteomics	100%		
Drug Discovery			100%
Bioinformatics			100%
Agricultural Research			
Epidemiology/Public Health		50%	50%
Precision Medicine			100%
CAE: PRODUCT DESIGN			
Structural Analysis			
Fluid-Structure Analysis			
Noise, Vibration, Harshness			
Crashworthiness			
Environmental Friendliness			
Materials Science			100%
CHEMICAL ENGINEERING			
Molecular Modeling			
COMMERCIAL ANALYTICS			
Fraud/Anomaly Detection	50%		50%
Affinity Marketing	50%	50%	
Business Intelligence	33%	17%	50%

Table 5

Specific Distribution of Data Ingestion

Application Area	Data Fusion	Data Reduction	Data Integration
Revenue Protection	100%		
Complex Pricing	100%		
DEFENSE			
Surveillance/Signal Processing			100%
Encryption			100%
Communications & Intelligence			
Anti-Terrorism			
ECONOMICS/FINANCIAL			
Portfolio Optimization			
Pricing Exotic Instruments	100%		
Global Risk Management		20%	80%
High Frequency Trading			100%
GOVERNMENT LAB			
Scientific Research	30%	40%	30%
Industrial Partnering		50%	50%
Law Enforcement			100%
MECHANICAL DESIGN			
3D Wireframe			100%
WEATHER/CLIMATE			
Climate Research	100%		

Source: Hyperion Research, 2018

As can be seen in Table 6, every potential choice for machine learning requirements was selected by at least someone, although there was little commonality among the various categories. Not surprisingly, the most frequently cited machine learning algorithm requirements were for the most common forms of machine learning today, Supervised Learning and Pattern Recognition (in many cases, Supervised Learning involves Pattern Recognition). The table shows, however, that more advanced forms of machine learning—Semi-Supervised and Unsupervised Learning—are starting to be used in some domains.

- The Bio-Sciences categories clearly had requirements for Pattern Recognition, as did some of the Economic/ Financial categories.
- Government labs continue to express one of the broadest and most diverse ranges of requirements, and this was the only category that expressed requirements in all of the options provided.
- In general, most subcategories relied exclusively on a single option. For example, precision medicine, structural analysis, and material science all cited requirements for Semi-Supervised Learning.

Table 6

Specific Distribution of Machine Learning

Application Area	Unsupervised	Semi-Supervised	Supervised	Reinforcement Learning	Pattern Recognition
BIO-SCIENCES					
Genomics			25%	25%	50%
Proteomics					100%
Drug Discovery					100%
Bioinformatics					100%
Agricultural Research					100%
Epidemiology/Public Health					
Precision Medicine		100%			
CAE: PRODUCT DESIGN					
Structural Analysis		100%			
Fluid-Structure Analysis					
Noise, Vibration, Harshness					

Table 6

Specific Distribution of Machine Learning

Application Area	Unsupervised	Semi-Supervised	Supervised	Reinforcement Learning	Pattern Recognition
Crashworthiness			100%		
Environmental Friendliness					
Materials Science		100%			
CHEMICAL ENGINEERING					
Molecular Modeling					
COMMERCIAL ANALYTICS					
Fraud/Anomaly Detection				50%	50%
Affinity Marketing					
Business Intelligence					
Revenue Protection	100%				
Complex Pricing					
DEFENSE					
Surveillance/Signal Processing					
Encryption					
Communications & Intelligence					
Anti-Terrorism					
ECONOMICS/FINANCIAL					
Portfolio Optimization	40%				60%
Pricing Exotic Instruments					
Global Risk Management	25%	25%			50%

Table 6

Specific Distribution of Machine Learning

Application Area	Unsupervised	Semi-Supervised	Supervised	Reinforcement Learning	Pattern Recognition
High Frequency Trading					100%
GOVERNMENT LAB					
Scientific Research	33%	17%	28%	6%	17%
Industrial Partnering	20%	40%	20%		20%
Law Enforcement					
MECHANICAL DESIGN					
3D Wireframe					
WEATHER/CLIMATE					
Climate Research					

Source: Hyperion Research, 2018

As can be seen in Table 7, experts were asked to select their most critical numeric optimizations, and in general the results were widely scattered across and within categories, with continuous, discrete, and stochastic methods selected in 15, 12 and 8 subcategories, respectively. It is interesting to note that in most cases, few respondents limited their preferences to only one optimization requirement.

- Indeed, in almost all cases except for business intelligence, experts within each subcategory selected two or even three of the provided choices.
- The Bio-Sciences category was quite divided. Sectors such as genomics, proteomics, and drug discovery used a mix of numerical optimizations, while others such as agricultural research and precision medicine selected none.
- Similarities across subcategories emerged as well. For example, molecular modeling and pricing exotic instruments in Finance/ Economics both expressed an equal reliance on all three optimization options. Likewise, climate research and high-frequency trading experts both opted for continuous and discrete optimization.

Table 7

Specific Distribution of Numeric Optimization

Application Area	Continuous	Discrete	Stochastic
BIO-SCIENCES			
Genomics	50%	50%	
Proteomics	50%	50%	
Drug Discovery	50%	50%	
Bioinformatics	25%	25%	50%
Agricultural Research			
Epidemiology/Public Health			
Precision Medicine			
CAE: PRODUCT DESIGN			
Structural Analysis	50%	25%	25%
Fluid-Structure Analysis			
Noise, Vibration, Harshness			
Crashworthiness			
Environmental Friendliness			
Materials Science			
CHEMICAL ENGINEERING			
Molecular Modeling	33%	33%	33%
COMMERCIAL ANALYTICS			
Fraud/Anomaly Detection			
Affinity Marketing			

Table 7

Specific Distribution of Numeric Optimization

Application Area	Continuous	Discrete	Stochastic
Business Intelligence	100%		
Revenue Protection			
Complex Pricing			
DEFENSE			
Surveillance/Signal Processing	50%	50%	
Encryption			
Communications & Intelligence			
Anti-Terrorism			
ECONOMICS/FINANCIAL			
Portfolio Optimization	25%		75%
Pricing Exotic Instruments	33%	33%	33%
Global Risk Management	75%		25%
High Frequency Trading	67%	33%	
GOVERNMENT LAB			
Scientific Research	25%	50%	25%
Industrial Partnering	29%	43%	29%
Law Enforcement			
MECHANICAL DESIGN			
3D Wireframe			
WEATHER/CLIMATE			
Climate Research	67%	33%	

Source: Hyperion Research, 2018

As can be seen in Table 8, survey respondents indicated that the two most important requirements for mining in their areas of HPC expertise were query processing and pattern recognition.

- In the main, the Bio-Sciences categories cited both data mining techniques as important and offered a similar preference pattern across many of the subcategories, while commercial analytics favored pattern recognition, and Economics/ Finance favored query processing.
- CAE showed almost no predilections for data mining requirements in general.
- As expected, the Government Lab category showed its typical broad and comprehensive preferences spanning all options.

Table 8

Specific Distribution of Data Mining

Application Area	Query Processing	Pattern Recognition	Network Analysis
BIO-SCIENCES			
Genomics	43%	43%	14%
Proteomics		100%	
Drug Discovery	50%	50%	
Bioinformatics	43%	29%	14%
Agricultural Research	100%		
Epidemiology/Public Health		50%	
Precision Medicine	100%		
CAE: PRODUCT DESIGN			
Structural Analysis			
Fluid-Structure Analysis			
Noise, Vibration, Harshness			
Crashworthiness			
Environmental Friendliness			
Materials Science			

Table 8

Specific Distribution of Data Mining

Application Area	Query Processing	Pattern Recognition	Network Analysis
CHEMICAL ENGINEERING			
Molecular Modeling			
COMMERCIAL ANALYTICS			
Fraud/Anomaly Detection	50%		50%
Affinity Marketing		100%	
Business Intelligence	33%	33%	
Revenue Protection		33%	67%
Complex Pricing			
DEFENSE			
Surveillance/Signal Processing		100%	
Encryption			
Communications & Intelligence			
Anti-Terrorism		50%	50%
ECONOMICS/FINANCIAL			
Portfolio Optimization	50%		
Pricing Exotic Instruments	100%		
Global Risk Management	57%	14%	
High Frequency Trading	100%		
GOVERNMENT LAB			
Scientific Research	35%	26%	26%
Industrial Partnering	17%	50%	

Table 8

Specific Distribution of Data Mining

Application Area	Query Processing	Pattern Recognition	Network Analysis
Law Enforcement		33%	33%
MECHANICAL DESIGN			
3D Wireframe			
WEATHER/CLIMATE			
Climate Research			

Source: Hyperion Research, 2018

SYNOPSIS

Hyperion's HPDA algorithm study is, to our knowledge, the first effort to create a comprehensive taxonomy of HPDA problem types and the preferred algorithm(s) for addressing each type. The study goes beyond this to characterize each problem type (application), along with its hardware and software requirements, and to identify the important attributes of each algorithm type associated with the problem.

What is this important? Because algorithms embodying mathematical models are the main intellectual capital and competitive weaponry for advanced analytics in the commercial, academic, and government sectors. Hyperion defines advanced analytics problems as those that need high performance computing (HPC) resources to run effectively. We call this global market high performance data analysis, or HPDA.

With the help of well-known experts, this special study created a notional taxonomy of HPDA algorithms and inserted them into a large matrix that pairs them with the applications users preferentially employ them for. The study captures detailed information on the hardware-software requirements of the applications and on the attributes of the algorithms. The study report aims to serve users, especially non-HPC specialists, as a kind of reference manual for matching their problems to relevant algorithm attributes and types. Hyperion believes that a reference tool of this kind is particularly valuable in the HPDA market, which is still formative and does not yet have standard definitions, approaches, and metrics.

This is the third of four papers we plan to publish based on the full study.

About Hyperion Research, LLC

Hyperion Research, consisting of the former IDC high performance computing (HPC) analyst team, provides HPC information, analysis, and recommendations based on technology and market trends. Research includes market sizing and forecasting, share tracking, segmentation, technology and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). We provide thought leadership and practical guidance for users, vendors and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com
