

Contrasting Xpiori Insight with Traditional Statistical Analysis, Data Mining and Online Analytical Processing

by Chris Brandin

Release 1.1

**Xpiori, LLC
2864 S. Circle Dr.
Ste. 1200
Colorado Springs, CO 80906
(719) 527-1315
www.xpiori.com**

© 2007 by Xpiori, LLC. All rights reserved.

Version 1.1

Copyright © Xpiori, LLC All Rights Reserved

Xpiori technology is protected by the following patents:

US Patent #5,742,611 (21 Apr 98)

US Patent #5,942,002 (8 Aug 99)

US Patent #6,157,617 (5 Dec 00)

US Patent #6,167,400 (26 Dec 00)

US Patent #6,324,636 (27 Nov 01)

US Patent #6,493,813 (10 Dec 02)

US Patent #6,792,428 (14 Sept 04)

Other U.S. and international patents pending.

The information in this white paper has been provided by Xpiori, LLC. To the best knowledge of Xpiori, it contains information concerning the current state of information processing technology. Xpiori, LLC disclaims any and all liabilities for and makes no warranties, expressed or implied, with respect to products described in this paper, including, without limitation, the implied warranties of merchantability and fitness for a particular purpose. No specific reliance should be made on the material provided herein without thorough investigation of the technology and its proposed application to specific circumstances. Product and technology information is subject to change without notice.

Executive Summary

Insight, an ad-hoc data exploration, analysis and discovery tool, enables rapid analysis and decision making with zero programming. The ad-hoc model of Insight breaks the traditional mold and creates a new paradigm in information analytics. When compared to traditional formalized statistical analysis, Insight provides significant value through reduction of cost and time to decision. Insight leverages the highly efficient pattern-matching core technology of the XMS native XML database to enable high-speed information analysis. The XMS database eliminates any need for costly star schemas, or data cubes. Because XMS processes XML information on the fly and without requiring schema definitions, Insight can immediately consume and analyze heterogeneous data stored in XMS.

Formal statistical analysis tends to be algorithmic. An algorithm is a complete set of mechanics for solving specific problems and implies a-priori concepts of the expected answers. Although it uses algorithms, Insight implements a heuristic analysis model. A heuristic is an incomplete set of guidelines with the potential for leading to greater learning and discovery. By allowing the analyst to work in a heuristic manner to perform ad-hoc exploration, Insight enables the discovery of previously unknown information that might otherwise remain hidden indefinitely.

Insight vs. Traditional Statistical Analysis

In information analysis, the goal of virtually any analytical model is to extract information from data by determining relationships between variables. Due to the analytical features of Insight, it becomes necessary to make a comparison with traditional analytics and describe the differences.

Formal Statistical Analysis

Traditional statistical analytics rely on formalized sets of algorithmic approaches. An algorithm is a complete set of mechanics for solving specific problems and implies a-priori concepts of the expected answers. By nature, an algorithm represents a pre-defined approach to solving specific problems, such as linear regression. In statistical analysis, the analyst commonly performs a series of steps, such as:

1. Create a hypothesis.
2. Compute descriptive statistics.
3. Identify outliers in the data.
4. Test for normal versus non-normal distribution.
5. Apply an appropriate algorithm.
6. Test the significance of the results.
7. Ensure that a type I or type II error did not occur.

At the end of the process, the analyst draws inferences from the statistical results. Formal statistical analytics apply rigid rules that data sets must conform to in order to apply a given algorithm and validate the results. For instance, algorithms based on analysis of variance, such as ANOVA, often require a normal distribution. Validation of results requires assurances that the data fit the chosen model, the results passed significance tests and no type I or type II errors occurred.

When applying formalized statistical analysis, a large portion of the effort involves hypothesis development and algorithm selection. The rigid rules imposed make ad-hoc drill-down and discovery very difficult or unfeasible. Application of formal statistical algorithms requires a-priori

knowledge, i.e. a pre-determined hypothesis. Hypotheses commonly influence data set selection or collection decisions and activities. Formal statistical algorithms assume a sampled population and, thus, often incorporate error term analysis. Outliers in the data set can skew the results and require decisions that determine whether to include a particular data point, usually a judgment call by the analyst.

Upon completion of the algorithmic process, the analyst can begin to make inferences from the results. Violations of the chosen model may require a restart from the beginning. Even worse, data sets collected to support a failed hypothesis may require additional design and implementation effort. Some analysis tools claim automation. However, an analyst must make choices and pre-program the analytical models based on a-priori conceptions of the answers, i.e. hypothesis.

Data sets stored in a relational database must be homogeneous and well formed in advance. Relational databases can be restrictive due to a reliance on primary and foreign key relationships among tables. Forming data sets for analysis often involves creation of a star schema, sometimes referred to as a data cube. The time and resources required to prepare data for analysis and program the analytics, even in so-called rapid development environments, can be costly. Because of this, unknown information can remain hidden indefinitely.

Xpiori Insight

Although it uses algorithms, Xpiori Insight implements a heuristic analysis model. A heuristic is an incomplete set of guidelines with the potential for leading to greater learning and discovery. Insight allows ad-hoc (heuristic) discovery and analysis without requiring a-priori knowledge of the data or preconceived notion of the answers. Its pattern-based empirical and probabilistic algorithms do not impose rigid conformity rules on data sets. The analytical model leverages the pattern-matching core technology of XMS. Highly efficient pattern-based set intersections performed by XMS enable the ad-hoc drill-down/drill-around and discovery capabilities of Insight. Because the analytical algorithms do not require conformance to rigid rule sets, Insight can compute correlations and display the results on the fly. The high-speed analytical capabilities of Insight provide for immediate inference and decision-making. By allowing the analyst to work in a heuristic manner to perform ad-hoc exploration, Insight enables the discovery of previously unknown information that might otherwise remain hidden indefinitely.

Insight Analytics Value Proposition

Insight provides significant value in many aspects due to its support for ad-hoc exploration and its ability to perform on-the-fly analysis of heterogeneous data sets. A word that illustrates the heuristic nature of analysis with Insight might be "eureka", having the same Greek root as the word, "heuristic". Archimedes made the exclamation famous after discovering his principle for identifying the composition of metals by water displacement while sitting in a bath. As the story goes, he ran naked through the streets shouting "Eureka!" (I have found it!)

Figure 1 - *Insight Use Scenario* shows how Insight fits into an information analysis and decision making scenario.

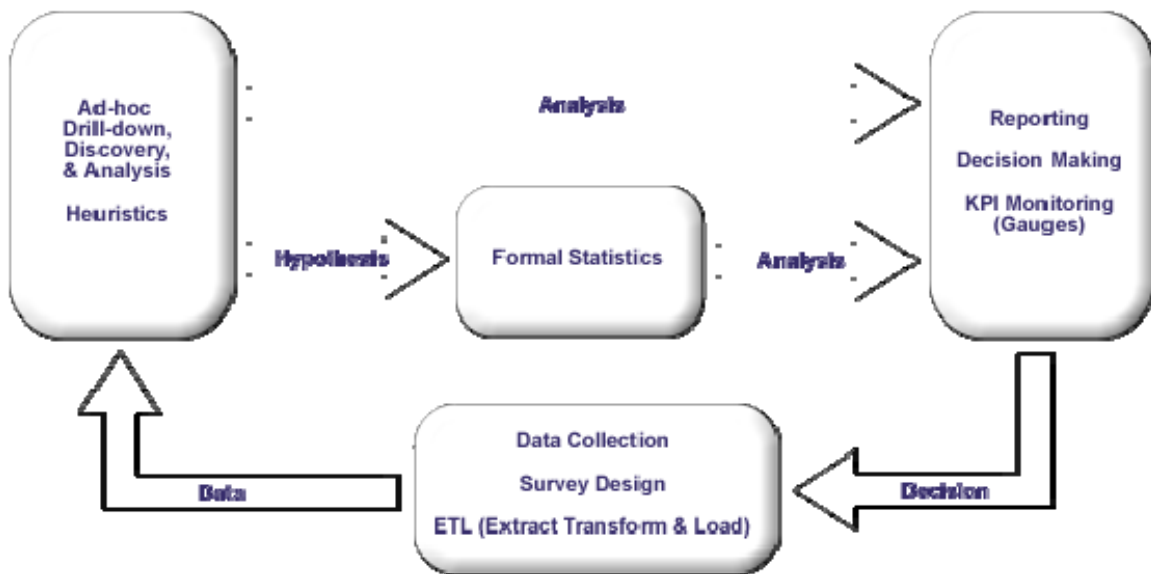


Figure 1 - Insight Use Scenario

Reduced Cost of Analysis

Insight can significantly reduce the cost of analysis. Compared with formalized statistical models, Insight virtually eliminates the costs associated with hypothesis development and rule conformity. The traditional methods of statistical analysis can have lengthy cycles and answering only one question can be costly, in terms of time and resources. Even in a rapid cycle environment, the turn-around time required using traditional methods may be measured in hours. Often, answering a single question can take days, weeks, or months. With Insight, users can perform multiple analyses within minutes, or even seconds, providing tremendous cost savings over traditional analytical methods.

Ad-hoc Exploration and Analysis Leads to Discovery

Due to the ad-hoc nature of its analytics model, the analyst can perform informal heuristic explorations of data sets and make discoveries without the need for a pre-developed hypothesis. The ability to perform ad-hoc exploration and analysis significantly increases the potential for new discoveries of previously hidden information.

Risk Mitigation and Early Stage Decision Making

Insight supports early stage decision making by reducing the analysis cycle and answering a greater number of questions in a shorter period-of-time. Insight provides value in data collection design and risk mitigation to refine data models and make decisions early in project life cycles.

No Star Schemas (Data Cubes) Required

XML, the data format used by the XMS database, contains both data and context. In contrast to a relational database, the XMS database does not rely on primary and foreign key relationships. Therefore, information stored in XMS will normally not require additional re-formatting prior to analysis, i.e. no star schema required. Additionally, Insight and XMS support heterogeneous data sets. Without requiring database design or migration, Insight can immediately consume previously non-existent data fields inserted into the database.

Accelerated Hypothesis Development

Insight does not eliminate the need for formal algorithms in all cases. When a formal algorithm is required, Insight helps to shorten the analysis cycle by accelerating hypothesis development. Hypotheses development using Insight reduces the probability of developing a failed hypothesis, leading to more accurate analytical models.

Business Intelligence Environments

The current reality is that an enormous amount of business enterprise data resides in warehouses or in distributed, unrelated databases. Business intelligence environments seek to make these data sources available on an enterprise-wide scale to enable faster and more accurate decision-making. When combined with an ETL solution and XMS as an operational data store, Insight provides a zero programming solution and more intimate engagement with enterprise information for business managers.

Insight vs. OLAP and Data Mining

In enterprise information analytics, the Insight analytics model bears some resemblance to traditional OLAP (on-line analytical processing) and DM (data mining). However, Insight provides significant advantages over the OLAP and DM paradigms. Data mining and OLAP are non-equivalent, complimentary technologies. Insight uniquely implements some necessary OLAP and DM features and renders others unnecessary. Industry efforts to integrate OLAP and DM indicate a trend in the information analysis industry toward the Xpiori Insight model.

OLAP

OLAP belongs to a sub-set of decision support tools used to discover patterns and relationships in a data set. The user forms a hypothesis about a relationship and verifies it with a series of queries against the data. In other words, the OLAP analyst generates a series of hypothetical patterns and relationships and uses queries against the database to verify them or disprove them. As such, OLAP analysis is a deductive process. An OLAP database is often synonymous with a data warehouse and typically constructed using star/snowflake schemas, or data cubes.

Data Mining

Data mining is an inductive process that uses the data, itself, to uncover patterns. Data mining relies heavily on the fields of artificial intelligence (AI) and statistics for pattern recognition and classification. Techniques include algorithms, such as neural networks, decision trees, and discriminant analysis. Traditional statistical techniques involve formal algorithmic modeling based on hypotheses. Data mining commonly includes predictive models using statistical techniques, such as linear regression. A data-mining tool might access an OLAP/relational database, or support various file formats, for data retrieval.

Feature Comparison

The following table presents a capability/feature comparison of Insight, OLAP, DM and traditional statistical analysis. Subsequent paragraphs provide further explanation.

Feature	Insight	OLAP	Data Mining	Traditional Statistical Analytics
Identify patterns natively, i.e. within DB context	X			
Enable real-time/dynamic changes to data sets	X	X		
Implement dynamic schema modifications	X			
Quickly analyze n-dimensional hierarchic data sets	X			
Rapidly discover inherent limitations of data sets	X			
Discover all existing correlations within a data set	X			
Quickly develop multiple hypothesis	X			
Extract insights without hypothetical prejudice	X			
Use to support traditional statistical analysis	X	X	X	X
Quickly deploy business intelligence environments	X			
Identify & monitor KPI's from live data sets	X			

Pattern Processing

Traditional OLAP and data mining methods use separate approaches to finding patterns and relationships within data sets. OLAP methods involve hypothetical query construction and data manipulation (data cubes). Data mining methods use database queries to retrieve data before applying a variety of statistical and artificial intelligence algorithms. Insight realizes a superior advantage over traditional methods by leveraging the core pattern processing technology of the XMS database. The native pattern-processing capability of the XMS database engine virtually eliminates any need for data cubes and external statistical analysis for identifying patterns and relationships.

Real-time/Dynamic Updates

While many OLAP databases support real-time and dynamic updates to data sets, doing so can be difficult and costly. An OLAP database may rely heavily on caching in order to improve query performance. In a real-time environment with low-latency updates, the performance of an OLAP database will suffer due to frequent cache invalidation.

Although the XMS database implements a cache, the performance of its core architecture is not affected in the same manner. In fact, the XMS cache remains valid during and after updates and does not require reload, thereby contributing to significantly better performance.

Dynamic Schema Modification

In data cube design, a star schema defines measures and dimensions that describe which data to present and how to present it. During the design process, the analyst must consider questions that ask why certain data should be included and why it requires a specific presentation. Resulting schemas are primarily incapable of supporting dynamic change.

The XMS database neither relies on, nor requires, explicit schemas to understand the structure of stored data. Instead, XMS recognizes implicit schemas from XML context information. XML

modifies context information on the fly without re-design. Storing a new document with additional XML tag hierarchy or inserting new context into an existing document automatically updates the implied schema.

Insight supports dynamic schema changes to data stored in XMS since it also does not rely on an explicit schema. Insight inspects the database and learns its structure through contextual information inherent in the stored XML. Therefore, Insight can adapt to frequent schema changes.

N-Dimensional Data Set Analysis

While OLAP databases enable analysis of n-dimensional data sets, the time associated with hypotheses development and data cube design results in a high cost compared to Insight. XMS supports dynamic and hierarchic heterogeneity without schema design, or re-design. Insight adapts to the XML structure based on its inherent context regardless of the existence of heterogeneous structures within a given data type. In other words, Insight and XMS view heterogeneous data sets of a given type in a manner similar to the abstraction concept of object oriented software design. This enables Insight to quickly analyze n-dimensional data sets without any need for data cube design and de-normalization, as with OLAP.

Inherent Limitation Discovery

When applying traditional analytical methods commonly used in OLAP and DM, the focus tends to be on answering questions formed from preconceived notions about the expected answers. Significant percentages of analysis resources may be spent making decisions about which questions to ask and creating a corresponding analysis model. Using Insight, analysis quickly converges toward the most significant correlations that exist within a data set. Insight correlations point to the important questions with relatively little effort from the user. Due to the high performance of XMS and rapid question/answer cycles, Insight analyses can quickly uncover deficiencies in data sets.

Analysis of a survey database might the data set chosen for the survey is incorrect or inadequate. Users realize the benefits of inherent limitation discovery at any stage in such a survey. For example, a behavioral health organization in Arizona used Insight to analyze the YRBS (Youth Risk Behavior Survey) raw data set from the CDC (Centers for Disease Control). Within minutes of the analysis, they realized that the data set was deficient because it did not include certain protective social factors that might help indicate causality of certain conditions. A previous attempt to analyze a summarized version of the same data set using traditional statistical analysis required nearly a man month of human resources and failed to uncover the deficiency.

Correlation Discovery

Insight effectively enables the discovery of virtually all correlations that exist within a data set in a relatively short period-of-time. In the OLAP and DM paradigm, correlation analysis moves at the pace of hypothesis development and data cube design.

Hypothesis Development and Prejudice

In traditional analytics, the requirement for hypotheses effectively contributes to delayed discovery and preclusion of potential correlations by virtue of choosing a hypothesis. A hypothesis provides an analytical focus and direction for the analyst. By nature, a hypothetical focus creates limitations that would preclude alternate views of the information under analysis. Often, in spite of subconscious or cognitive dissonance concerning a chosen hypothesis, the analyst must proceed to the proof, or rejection, of the hypothesis.

The absence of a requirement for pre-developed hypotheses allows the Insight user the freedom to explore a data set without limit. Focus and hypotheses emerge as output of the analysis instead of being requisite input. In instances where traditional statistical methods are needed,

Insight may be used to drive hypothetical development and significantly reduce the total time and cost of analysis.

Business Intelligence

Information analysis supports BI (business intelligence) through identification of important data elements and correlations. The results of such operations might include the data, itself, or documentation describing the data and analysis discoveries. In business intelligence environments, analytical discoveries feed forward into activities, such as decision support and KPI (key performance indicator) identification and monitoring.

Insight provides integrated BI support via gauges and executive dashboards. Gauges created from Insight correlations automate the BI process using a “push analytics” model, whereby analysis results are “pushed” to executive dashboards immediately upon discovery. Managers use these gauges for continuous monitoring. Once deployed, users in a BI environment can potentially progress from raw data to discovery and on to real-time monitoring in a matter of minutes.

#