DECISION MANAGEMENT SOLUTIONS

In-Database Analytics

Embedding Analytics in Decision Management Systems

James Taylor
CEO

In-database analytics offer a powerful tool for embedding advanced analytics in a critical component of IT infrastructure.

Organizations are adopting a new class of operational systems called Decision Management Systems to meet the demands of consumers, regulators and markets because traditional systems are too inflexible, fail to learn and adapt and crucially cannot apply analytics to take advantage of "Big Data." Decision Management Systems are agile, analytic and adaptive. They are agile so they can be rapidly changed to cope with new regulations or business conditions. They are analytic, putting an organization's data to work improving the quality and effectiveness of decisions. They are adaptive, learning from what works and what does not work to continuously improve over time.

Data mining and predictive analytics play a key role in Decision Management Systems. Decision Management Systems take advantage of the information available to an organization to improve the accuracy and effectiveness of each decision. This is achieved by embedding the results of mining historical data or by executing predictive analytic models derived from that data using mathematical techniques.

As the amount of data involved has grown and as data infrastructure has become more powerful, the adoption of in-database analytic technology has grown rapidly. Such products cover a wide range of capabilities:

- Support for reporting and OLAP.
- Data preparation and data quality tasks.
- Analytic model development including data discovery, variable selection, data mining and text analytics.
- Analytic model deployment and scoring.
- Ongoing analytic model management.

This paper will introduce in-database analytics and explore their role in embedding advanced analytics in Decision Management Systems.

CONTENTS

sas

Introducing In-Database Analytics

In-database analytics can mean exactly that—analytic capabilities embedded in a relational or columnar database. The phrase is also used to describe analytic capabilities embedded in data warehouse software, in data appliances and increasingly in Hadoop clusters.

Defining in-database

In-database analytic capability is delivered as a set of libraries, User Defined Functions, that deliver analytic or data mining functions such that they can:

- Access the data in the database, data warehouse, appliance or Hadoop file system in situ, without needing to extract it to some interim format.
- Directly use the memory, parallel processing capabilities and load balancing/processor management of the data infrastructure.
- Be accessed both from specialist analytic tools (for model creation or data quality tasks for instance) and from operational systems.

In-database analytic capabilities are specific to a particular database, data warehouse, data appliance or Hadoop distribution. Many vendors offer support for multiple data infrastructure platforms. Some capabilities are provided by the data infrastructure vendors, some by specialty analytic vendors, and some through partnerships between analytic and data infrastructure vendors.

When it comes to supporting Decision Management Systems, the core capabilities to look for today in an in-database analytic product are:

- In-database data preparation and quality.
- In-database modeling.
- In-database model deployment.

It is not that in-database analytic support for reporting or OLAP is unimportant, only that it is not the focus for Decision Management Systems. It should be noted that the same ROI often applies when using in-database analytics to improve manual decision making.

In the future, more extensive support for analytic model management and for wrapping analytics in business rules for in-database decision-making will become increasingly important.

In-database data preparation and quality

Data preparation, integration and cleaning often consumes 60-70% of the time and effort on an analytic project. In a traditional approach, data is extracted from the data infrastructure in which it is stored, processed through various preparation steps and then presented to the analytic modeling algorithms that need it.

With in-database capabilities, however, these steps all execute in-database. This means the original data is not extracted from the database but is processed in situ. The resulting cleaned and transformed data may be stored in the data infrastructure or passed out to a predictive analytic workbench for further processing. The net is that data required for analytic modeling is transformed in-database.

SQL Push Back is a common feature of analytic model tools. This involves generating complex SQL to handle a set of data extraction and preparation steps in a model workflow so that they can be performed in a single query. While this is similar to the use of an in-database analytic engine to handle data preparation and quality this is not the focus of this paper due to its more limited scope.

Sometimes in-database data preparation and quality capabilities support only single function calls and job/script/workflow management is handled outside the database. More complete offerings allow more whole scripts/jobs to be executed in-database.

In-database model development

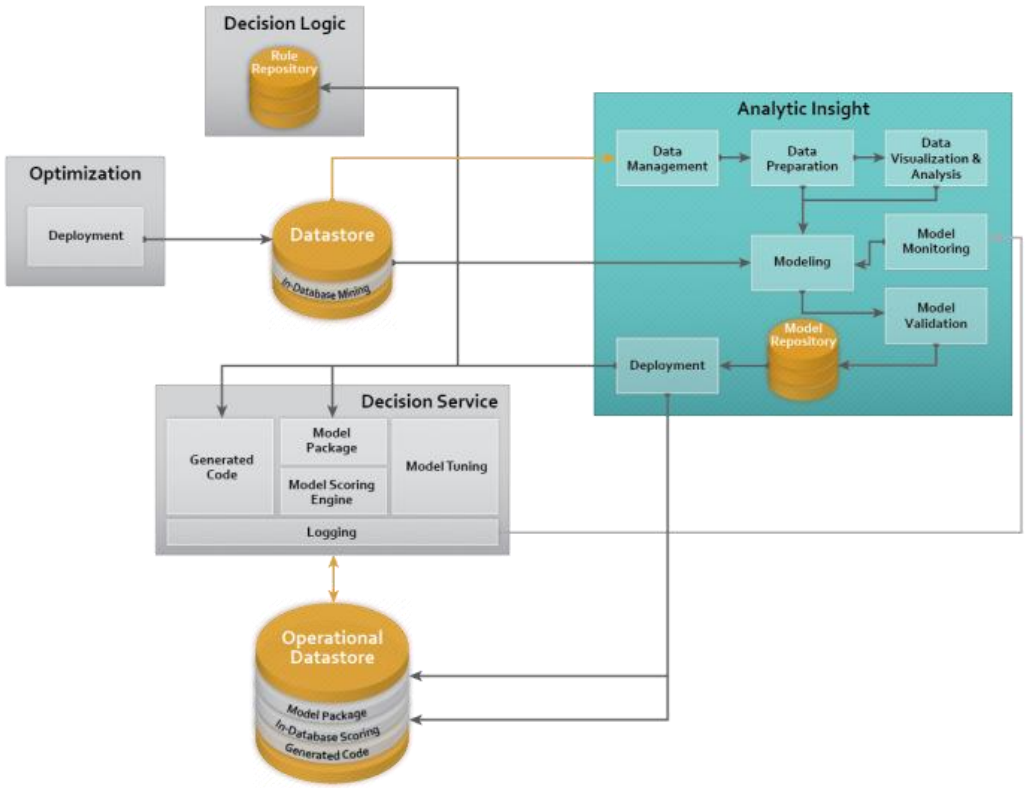
In-database model development allows predictive analytic models to be developed using algorithms embedded in the data infrastructure. These algorithms access tables and views directly to get the data they need, process the data using the data infrastructure's processing capabilities, and create a predictive analytic model. This model may be stored in the data infrastructure for in-database scoring or it may be passed out for use elsewhere.

These capabilities may be integrated with an external predictive analytic workbench so that they can be called as part of a modeling exercise. Some in-database modeling capabilities support the same wide range of modeling techniques as a predictive analytic workbench. Some support only a subset and are used by a predictive analytic workbench to offload some of the work of building predictive analytic models to the data infrastructure—essentially processing some of the steps defined in the workflow in-database.

In-database model deployment and scoring

In-database model deployment and scoring infrastructure takes models developed using some combination of in-database modeling infrastructure and a predictive analytic workbench and executes them in an operational datastore so they are available to operational systems accessing that datastore. This generally involves turning models into UDFs or stored procedures that can be called using SQL and that take database fields as input. These functions can be used in standard SQL statements and embedded into database views as though they are database columns, making the scores available widely. Another approach is to install an Embedded Process and run inside the database to read and write data from the database. The advantage of using the Embedded Process is that a single function or a stored procedure is used instead of multiple, user-defined functions.

Figure 1: In-Database Analytics in Context



The ROI of in-database analytics

As with any product, a return on investment can come from increased revenue or decreased costs. Predictive analytics often add top-line revenue by boosting sales or driving fraud out. These kinds of returns are due to the use of predictive analytics in general rather than the use of in-database analytics specifically. Nevertheless in-database analytics offer an ROI both by increasing value (though speed to market, improved accuracy and increased accessibility) and by decreasing costs.

Speed to Market

The key to deriving ROI from in-database analytics is a dramatic increase in speed to market. Using in-database analytics allows for value from analytics sooner. This increase in speed delivers analytic models more quickly by taking advantage of several characteristics of in-database analytics technology:

- Streamlining data preparation and other preparation activities though the elimination of data movement and replication.
- Executing analytic algorithms in the same memory space and using the high performance hardware common in data infrastructure.
- Massive parallelism of analytic algorithms so they can take advantage of MPP architectures to develop models across multiple nodes in the data infrastructure.
- Improved handling of very large datasets and large numbers of variables through execution of algorithms inside the database.
- Eliminating the need to re-code analytic models so they can be deployed.

This can result in a 10-100x overall reduction in time from when a team starts to when decisions are being made more analytically in a decision management system.

Improved Accuracy

Predictive analytic models developed using in-database analytic technology might be more accurate than those developed more traditionally:

- A faster cycle time can mean that more approaches can be considered, more iterations performed resulting in a superior approach.
- The ability to access data directly and not have to move it around can eliminate the need for sampling while simultaneously reducing the likelihood of errors being introduced in manual extraction/cleaning steps.
- Using the “main” data store rather than an analytical data mart may mean a more up to date and large dataset is available.
- Increased model development performance can be leveraged to develop more models using more algorithms for use in a more precise ensemble model.

Increased Accessibility

Lastly it is likely that the resulting analytics will be more accessible and so more likely to be used in more places, increasing their reach. This is particularly true in organizations where development tools are not analytically aware. For instance while most modern business rules management systems can easily execute an analytic model (and many import PMML), a system based on COBOL cannot easily be changed to execute a predictive analytic model. If the model is made available in the database, however, it may be possible to use it with far less change to the code.

The use of in-database analytics is a perfect vehicle to bring analytic and IT teams together. Too often the analytic and IT/data architecture teams can be at odds. The use of in-database analytics can bring the two groups onto the same side, using the approach to avoid data replication, improve governance and make analytic scores broadly available through the IT infrastructure.

Lower cost

The primary cost reduction is from less hardware and improved utilization. Because in-database analytic approaches reuse database appliances and other data infrastructure, less hardware has to be bought. At the margins there may also be reduced costs resulting from moving less data, though probably only in extreme cases or when data movement is billed for directly by an infrastructure provider. As data grows, the case for in-databases becomes stronger because the penalty for moving data out of the database becomes greater. Similarly the rapid growth in data volumes means that any analytical server must grow as fast as the database server does if analytic tasks are not performed in-database.

Many databases and data warehouse platforms have extensive support for integrating external data and making this data available through standard SQL. As such it is available to in-database analytic capabilities. Especially when such an approach is used to bring in data stored on Hadoop or commodity hardware it may lower the cost to store the increasingly large volumes of data being collected. With in-database analytics all this data is available to algorithms executing in the database.

Costs may also be lowered through better governance as analytic models are managed like metadata in a typical in-database analytic infrastructure. This may involve less cost and less risk than traditional approaches that involve widely spread scripts and workflows.

It should be noted that data infrastructure, especially high performance and more expensive varieties, tends to be very heavily utilized. Therefore, proper sizing of required capacity to take full advantage of in-database processing must be taken into consideration beforehand.

Applying in-database analytics

When developing analytics for use in a Decision Management System there's generally a simple sequence of steps:

- The team discovers and models the decisions that are to be supported. This identifies potential uses of analytics and establishes a decision-making context.
- The analytics group then proceeds through an interactive discovery process involving multiple model attempts, testing, and refinement until they have a suitable analytic model.
- This model is then deployed into a complete Decision Management System where it may be combined with business rules to make decisions.
- How well the decision is working is measured over time and the performance of the analytic models involved, changes in data distribution and more are monitored. Based on this analysis regular updates of the model may be needed.
- Over time changing business requirements or large shifts in performance will require the team to go back and repeat its original interactive discovery effort.

Over time this becomes a continuous sequence of steps, developing, embedding and improving predictive analytic models.

In-database analytic technologies add value throughout this sequence.

- Interactive modeling is faster thanks to more rapid turnaround of candidate models. Sampling is no longer required and models use 100% of the data.
- The preparation of data for modeling is improved by several multiples through in-database processing. This has a disproportionate impact because data preparation is such a large part of most analytic projects.
- Analytic model creation often involves processing a lot of data. For ensemble models, increasingly popular due to their increased accuracy, the data may need to be processed multiple times and through multiple steps. In-database analytics means that data does not have to be repeatedly moved out of the data infrastructure.
- Being able to deploy the model directly and support in-database scoring reduces the time to get access to the model and eliminates re-coding. The use of in-database scoring makes the analytic model available everywhere the data is.

These are all benefits of in-database analytic technology widely available today. Increasingly integrated model management allows for easier monitoring and managing of deployed models, adding further value. Longer term the possible deployment of a complete decision—business rules and predictive analytic models—in the database will increase this value significantly by making analytic decision-making pervasive throughout the data infrastructure.

Recommendations

Organizations developing Decision Management Systems should immediately consider in-database scoring as part of their deployment infrastructure. This might not be the only way the organization chooses to deploy analytic models—It may be easier to assemble the data you need to score a model using a business process rather than data infrastructure for instance—but it is generally highly worthwhile as part of the deployment approach.

Organizations should also evaluate their model development efforts to see how they could use their data infrastructure more effectively. This should be considered as part of an overall strategy to develop a higher performance, more industrialized analytic development approach. In-memory, in-database and other approaches may all be worth considering depending on the specifics of the environment. Obviously organization's whose data infrastructure is already being used at capacity may find other approaches more useful but it is a rare organization that will get no value out of in-database analytic model development.

Finally organizations should remember always that model development is not a one-time activity. Models must be monitored, updated and improved on a continuous basis. Ensuring that the model management being used is integrated with the in-database analytic technology in use will be important in creating an effective analytic environment going forward.

Works cited

Taylor, James. *Decision Management Systems Platform Technologies Report*. Palo Alto, CA: Decision Management Solutions, 2013.
Taylor, James. *Three steps to put Predictive Analytics to Work*. Palo Alto, CA: Decision Management Solutions, 2013.

Contact Us

If you have any questions about Decision Management Solutions or would like to discuss engaging us we would love to hear from you. Email works best but feel free to use any of the methods below.

Email : info@decisionmanagementsolutions.com

Phone : +1 650 400-3029

Fax : +1 650 352-9247