Аналитические платформы [M.007]

Даже самые мощные технологии извлечения закономерностей и машинного обучения, такие как KDD и Data Mining, не представляют особой ценности без инструментальной поддержки в виде соответствующего *программного обеспечения*. Рынок программных средств продолжает формироваться по сей день, однако в этой области уже можно выделить некоторые стандарты де-факто.

Программное обеспечение в области анализа данных

Рынок программного обеспечения KDD и Data Mining делится на несколько сегментов (рисунок 1).

Рисунок 1 – Классификация ПО в области Data Mining и KDD

Статистические пакеты с возможностями Data Mining и настольные Data Mining пакеты ориентированы в основном на профессиональных пользователей. Их отличительные особенности:

- слабая интеграция с промышленными источниками данных;
- бедные средства очистки, предобработки и трансформации данных;
- отсутствие гибких возможностей консолидации информации, например, в специализированном хранилище данных;
- конвейерная (поточная) обработка новых данных затруднительна или реализуется встроенными языками программирования и требует высокой квалификации;
- из-за использования пакетов на локальных рабочих станциях обработка больших объемов данных затруднена.

Плюсом статистических пакетов является их широкая распространенность. Настольные Data Mining пакеты могут быть ориентированы на решение всех классов задач Data Mining или какого-либо одного, например кластеризации или классификации. Вместе с тем эти пакеты предоставляют богатые возможности в плане алгоритмов, что достаточно для решения исследовательских задач. Существует немало свободно распространяемых настольных пакетов Data Mining с открытыми исходными кодами.

Однако создание эффективных прикладных решений промышленного уровня с помощью таких пакетов затруднено, поэтому в бизнес-аналитике, как правило, используются *СУБД с элементами Data Mining* и **аналитические платформы**.

Практически все крупные производители систем управления базами данных включают в состав своих продуктов средства для анализа данных и поддержку хранилищ данных. Эти

инструменты как бы встраиваются в СУБД. Отличительные особенности СУБД с элементами Data Mining:

- высокая производительность;
- алгоритмы анализа данных по максимуму используют преимущества СУБД;
- жесткая привязка всех технологий анализа к одной СУБД;
- сложность в создании прикладных решений, поскольку работа с СУБД ориентирована на программистов и администраторов баз данных.

Аналитические платформы

В отличие от СУБД с набором алгоритмов Data Mining, аналитические платформы изначально ориентированы на анализ данных и предназначены для создания готовых решений.

Определение

> *Аналитическая платформа — специализированное программное решение (или набор решений), которое содержит в себе все инструменты для извлечения закономерностей из «сырых» данных: средства консолидации информации в едином источнике (хранилище данных), извлечения, преобразования, трансформации данных, алгоритмы Data Mining, средства визуализации и распространения результатов среди пользователей, а также возможности «конвейерной» обработки новых данных.*

В аналитической платформе, как правило, всегда присутствуют гибкие и развитые средства консолидации, включающие богатые механизмы интеграции с промышленными источниками данных, инструменты очистки и преобразования структурированных данных и их последующее хранение в едином источнике в многомерном виде — в хранилище данных. Модели, описывающие выявленные закономерности, правила и прогнозы, также хранятся в специальном источнике данных — репозитарии моделей.

На рисунке 2 изображена типовая схема системы на базе аналитической платформы.

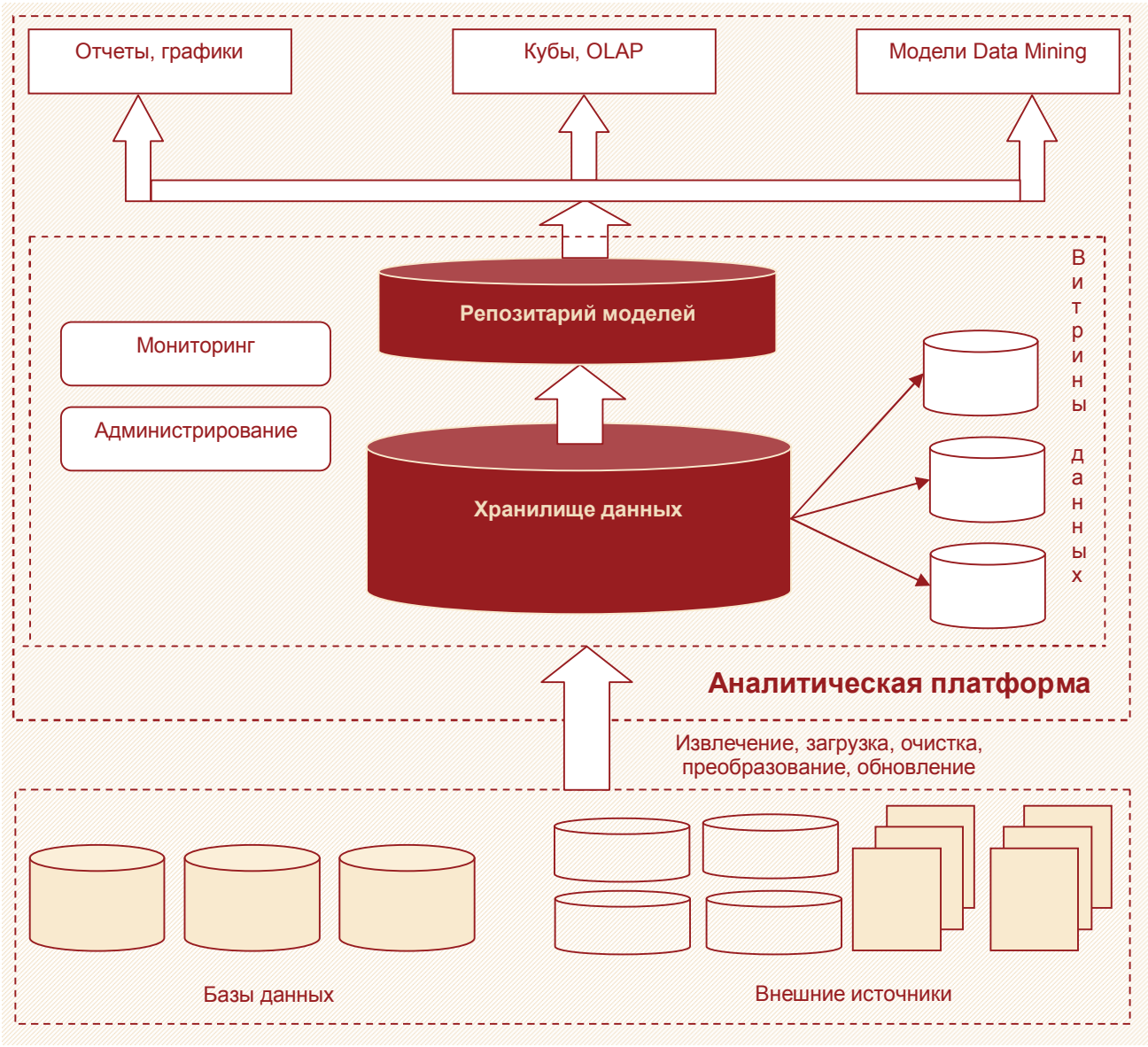

Рисунок 2 – Аналитическая платформа

Вообще говоря, приведенную на рисунке 2 систему можно построить с использованием нескольких программных решений, например, разделить функции извлечения/загрузки, OLAP-отчетности, хранилища данных, Data Mining между различным программным обеспечением. Но чтобы эти отдельные компоненты превратились в полноценную аналитическую систему, необходимо произвести интеграцию между ними на уровне обмена данными, а еще лучше — метаданными.

Языки визуального моделирования

Основным препятствием на пути все более широкого применения методов и программных средств анализа данных является сложность инструментов. Поэтому важно освободить аналитика от необходимости углубленного понимания сложных математических алгоритмов. Своеобразным ответом на это требование стало появление языков визуального моделирования. Сегодня их наличие является стандартом де-факто в полноценной аналитической платформе.

Язык моделирования позволяет аналитику в визуальной среде строить последовательности шагов по обработке данных от получения «сырых» данных до конечного результата. Шаги

представляют собой набор атомарных по отношению к данным операций, каждую из которых можно представить отдельным узлом. Примеры таких операций: выборка данных, фильтрация, сортировка, добавление нового столбца, построение Data Mining модели и т. п. Набор узлов образует графическую диаграмму. Такой способ представления очень близок к рассуждениям и действиям аналитика, которые он так или иначе проделывает. Существует две формы представления диаграмм: в виде *дерева* и в виде *графа* (рисунок 3 а, б).

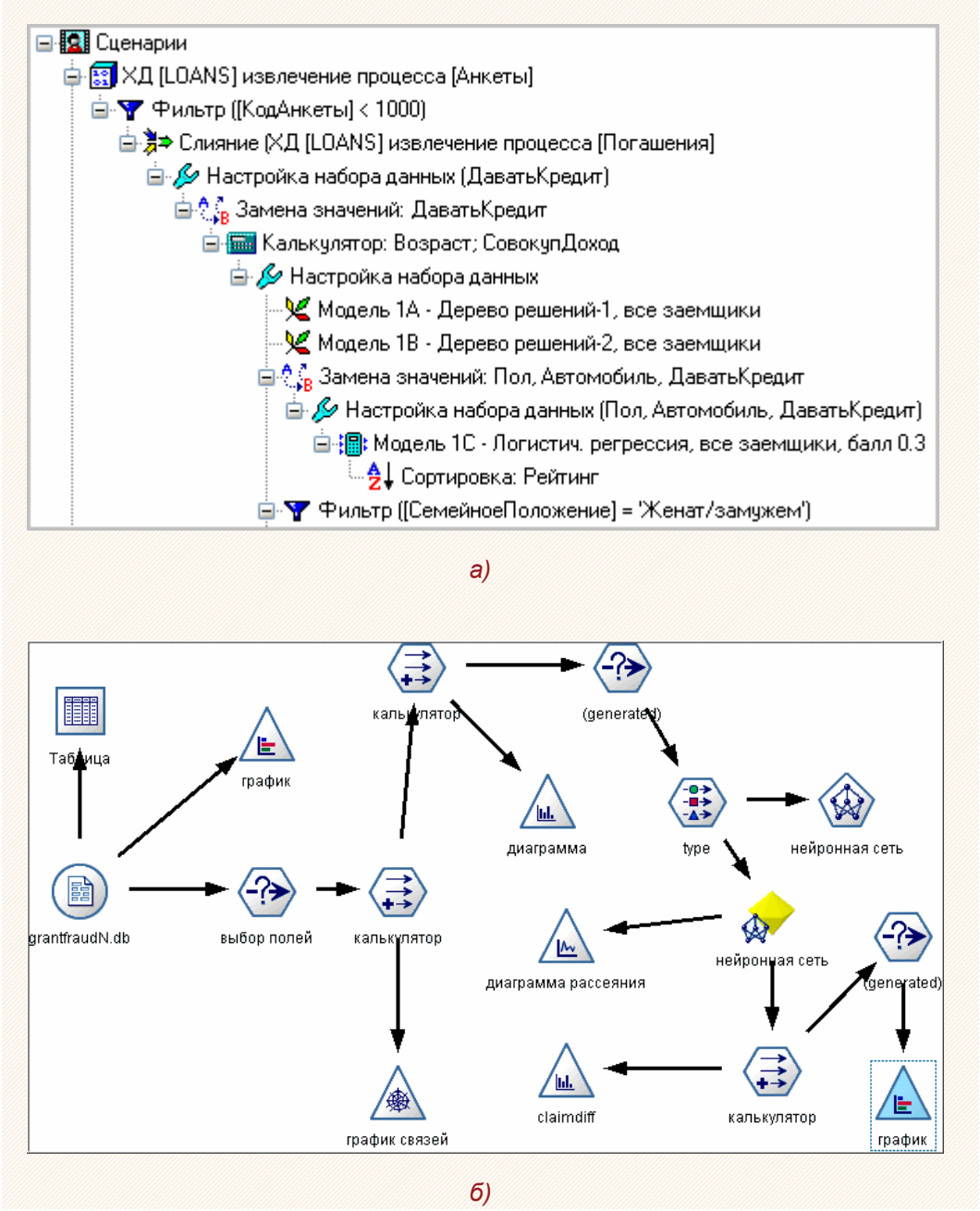

Рисунок 3 – Языки визуального моделирования: а – в виде дерева, б – в виде графа

Общие особенности языков моделирования в аналитических платформах следующие.

- Базовым узлом, с которого начинается диаграмма, является узел импорта, поскольку в аналитических платформах обычно отсутствуют средства для ручного ввода данных; предполагается, что данные уже имеются в каких-либо источниках.
- Графическое изображение, соответствующее какому-либо узлу, несет в себе большой семантический смысл. Оно помогает аналитику различать узлы по функциям и

определять их активность (часто еще не выполненный узел обозначается иконкой серого цвета, а выполненный — цветной).

- Диаграмма описывает формализованную последовательность действий над данными, и эти действия можно повторить на совершенно других данных, предварительно настроив соответствие колонок.

Каждая из форм представления имеет как достоинства, так и недостатки. У деревьев более жесткая структура по сравнению с графами, поэтому, к примеру, отображение двух узлов, сливающихся в один, затруднено. Вместе с тем дерево более компактно (в графе обязательно присутствуют стрелки, которые занимают место на диаграмме), что очень важно при большом количестве узлов, и позволяет выполнять множество интуитивно понятных операций, связанных с манипулированием ветвями(копирование, удаление, перенос и т. д.).