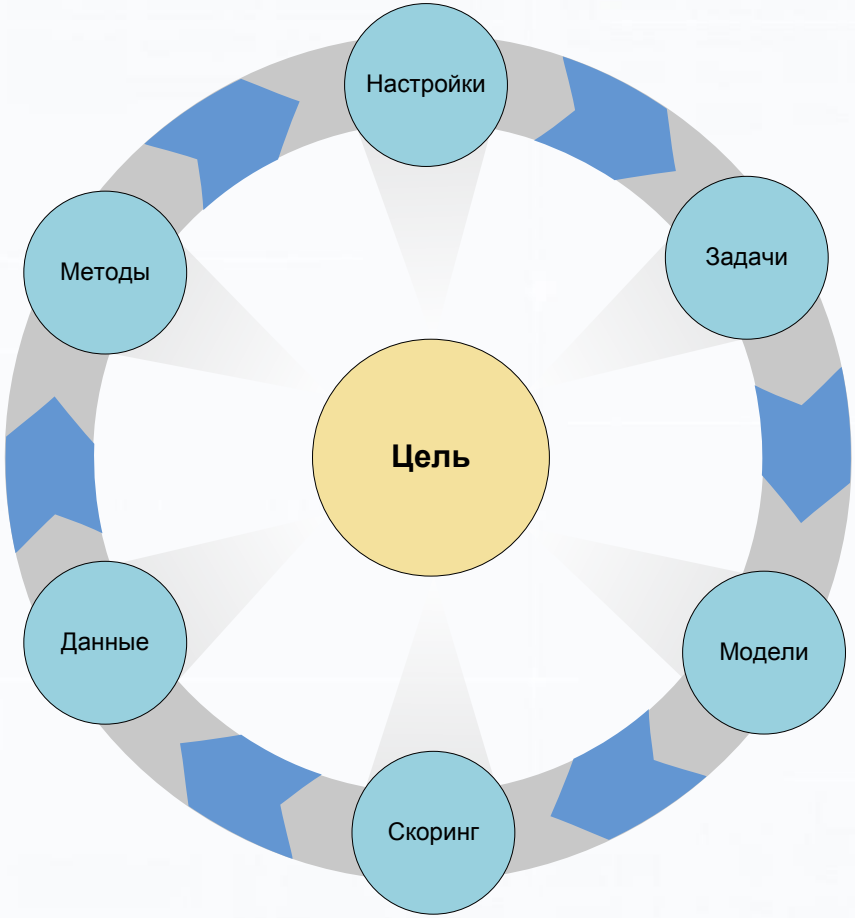
Аналитическая платформа с многоуровневой архитектурой

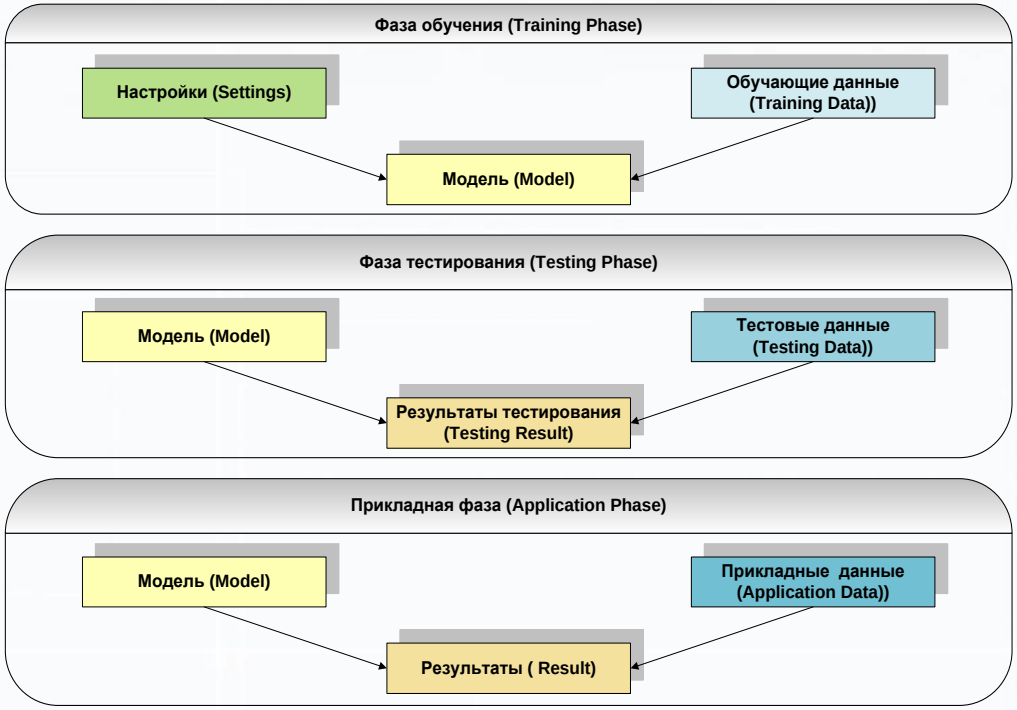
Дорогов А.Ю., Харьковский А.С., Абатуров В.С.
СПбГЭТУ «ЛЭТИ», ПАО «Интелтех»

2017

Технологический цикл предикативной аналитики



Аналитика в базе данных. Стандарт SQL/MM



Проблемы и вызовы

- рыночная адаптация
- эволюция версий
- проблемы алгоритмических расширений
- не регламентированная инфраструктура
- не регламентирован формат метаданных

История развития

В конце 1992 г. комитет по стандартизации SQL разработал дополнение для объектно-ориентированного SQL. Тогда же был принять стандарт, который описывал библиотеки классов объектных типов SQL. Предложенный стандарт стал известен как SQL/MM (MM расшифровывалось как MultiMedia)

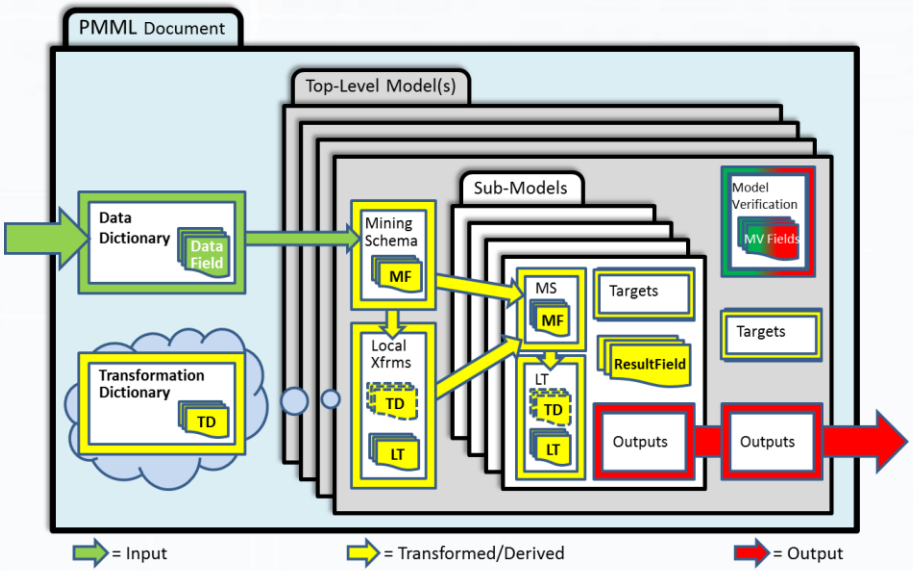
Возможности:

- обработка непосредственно в базе данных
- использование сервиса базы данных
- совместимость с языком SQL
- этапы аналитической обработки
- расширенные средства MultiMedia

Стандартизованные алгоритмы извлечения знаний

Ассоциативные правила
Последовательности
Классификация
Кластеризация
Регрессия

PMML – стандарт представления моделей Data Mining



История развития

PMML стартовал в 1998 году с версии 0.7. С тех пор стандарт пережил множество релизов вплоть до версии 4.2 в 2014 г.

Проблемы и вызовы

- рыночная адаптация
- эволюция версий
- проблемы алгоритмических расширений
- открытое ПО для восстановления алгоритма

Возможности:

- свободен к коммерческому использованию и имеет открытые коды
- гетерогенное окружение различных производителей
- пре- и пост- обработка встроены в стандарт
- поддержка мультимодельности
- встроенные механизмы расширения

Стандартизованные модели алгоритмов извлечения знаний

- Ассоциативные модели
- Базовые модели
- Деревья решений
- Кластеризация
- Регрессия
- Обобщенная регрессия
- Классификация по ближайшим соседям
- Нейронные сети
- Наивный байесовский классификатор
- Оценочные карты
- Последовательности
- Обработка текста
- Временные ряды
- Логические правила
- Машина опорных векторов

Поддерживающие компании

Alpine, SAND, SAS, TERADATA, TIBCO, ZEMENTIS, Angoss, open data, bigml, Experian, IBM, KNIME

Аналитические СУБД

ORACLE

- ***ODM*** (*Oracle Data Mining*) – расширение СУБД *Oracle Enterprise Edition* для выполнения аналитической обработки данных и бизнес аналитики;

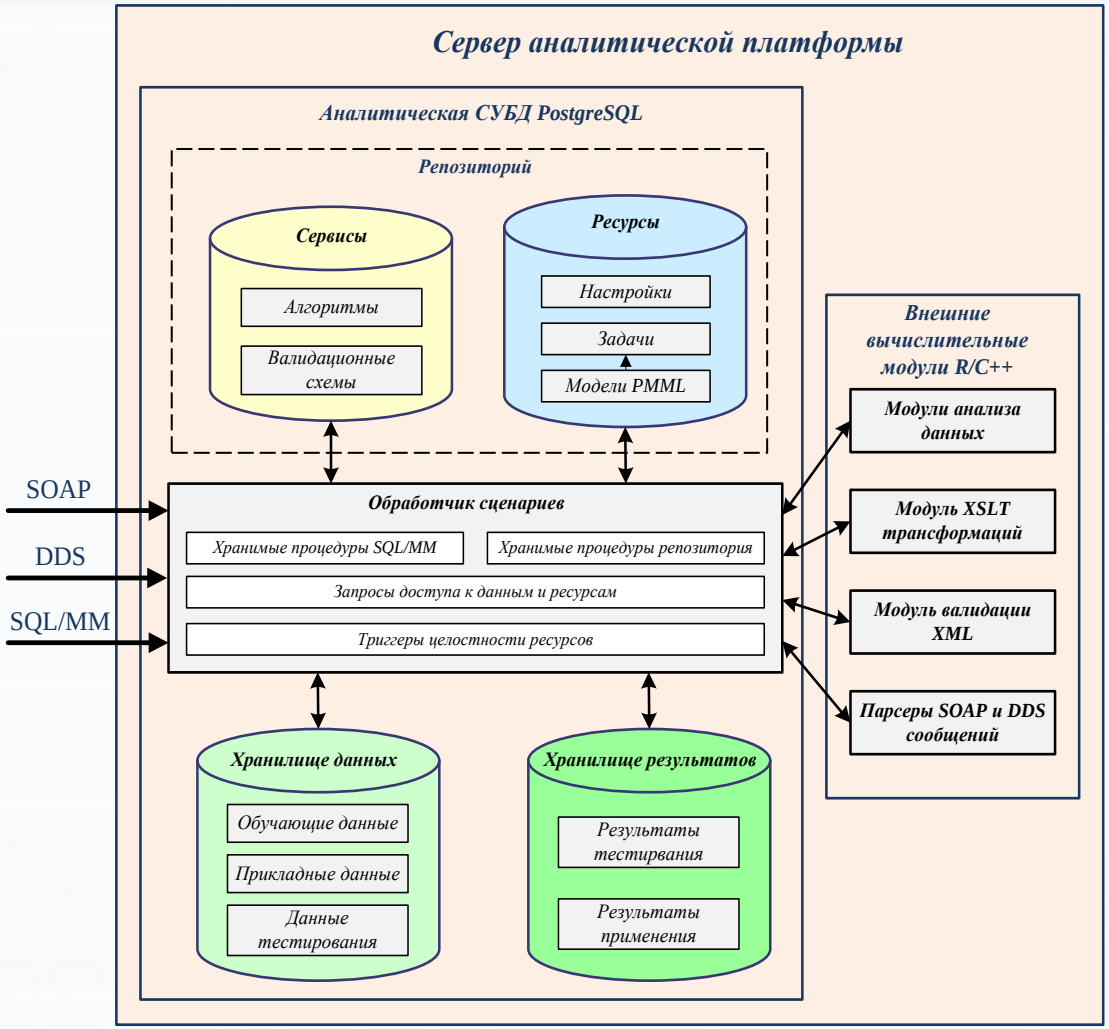
IBM DB2

- ***Data Warehousing Utilities*** (DB2) – средства бизнес-аналитики в СУБД DB2 Universal Database;

Microsoft SQL Server

- ***Analysis Services*** *(Microsoft SQL Server)* – служба СУБД *Microsoft SQL Server* для работы с *OLAP* и интеллектуального анализа данных.

Архитектура аналитической платформы Sorintella



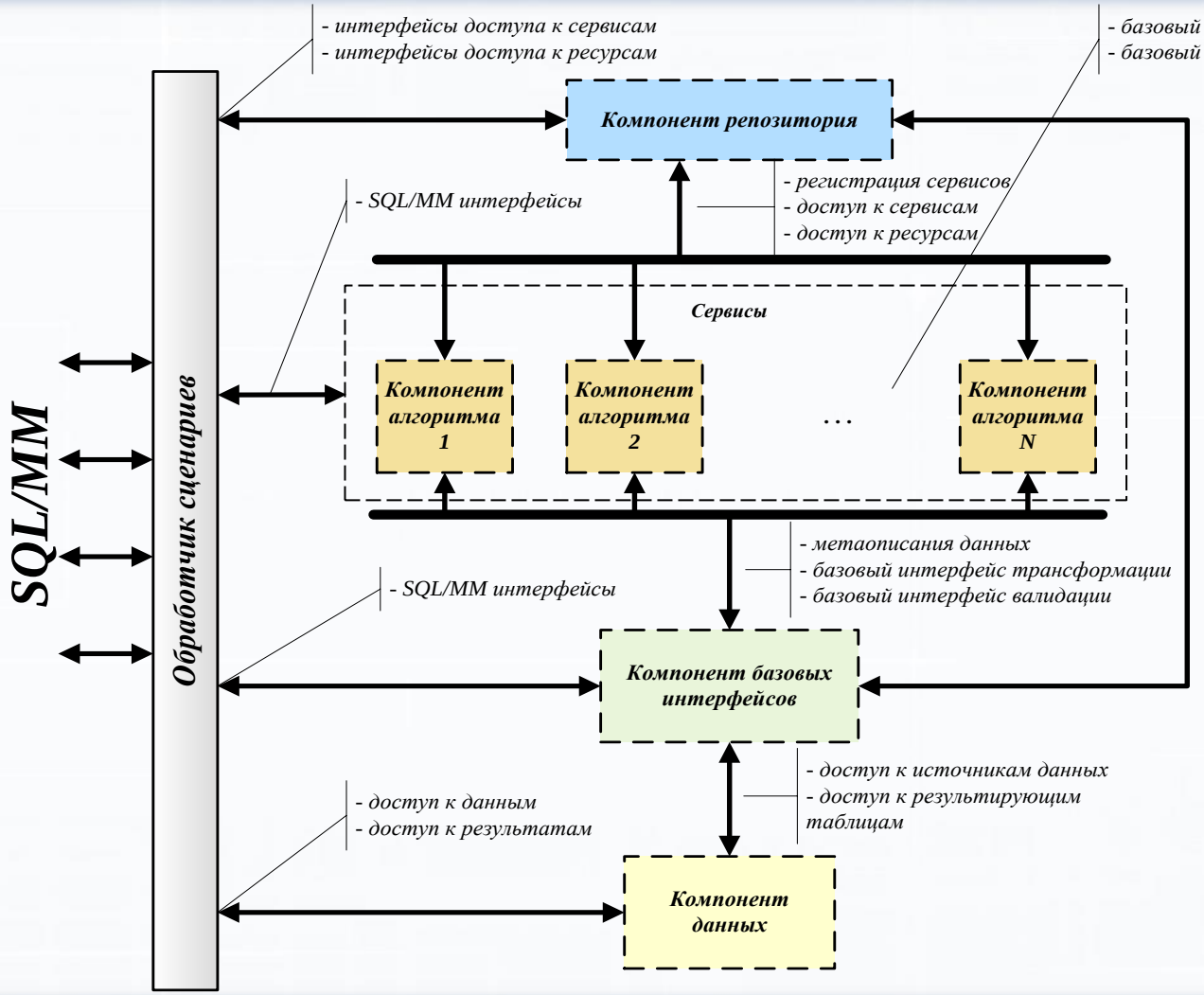
Тактико-технические характеристики

- обработка масштабных массивов разнородной информации (до 32 ТВ);
- расширяемость и масштабируемость аналитики;
- многоплатформенность;
- контроль целостности данных;
- интерфейсы доступа к языкам программирования высокого уровня;
- триггеры и правила для управления процессами;
- система управления правами доступа и авторизации;
- параллельная обработка пользовательских сессий;
- шифрование трафика;
- поддержка стандарта SQL/MM в части обработки пространственных и геоинформационных данных.

Инфраструктура обеспечивает:

- доступ и управление аналитическими ресурсами платформы (задачами, вычислительными моделями и настройками).
- спецификацию о реализованном наборе сервисов (описание алгоритмов).
- спецификации о допустимых настройках для реализованного набора алгоритмов.
- спецификацию по моделям представления знаний.
- спецификации интерфейсных функций стандарта SQL-MM для реализованного набора алгоритмов.
- поддержку целостности ресурсов и сервисов.

Компонентная архитектура аналитической платформы Sorintella



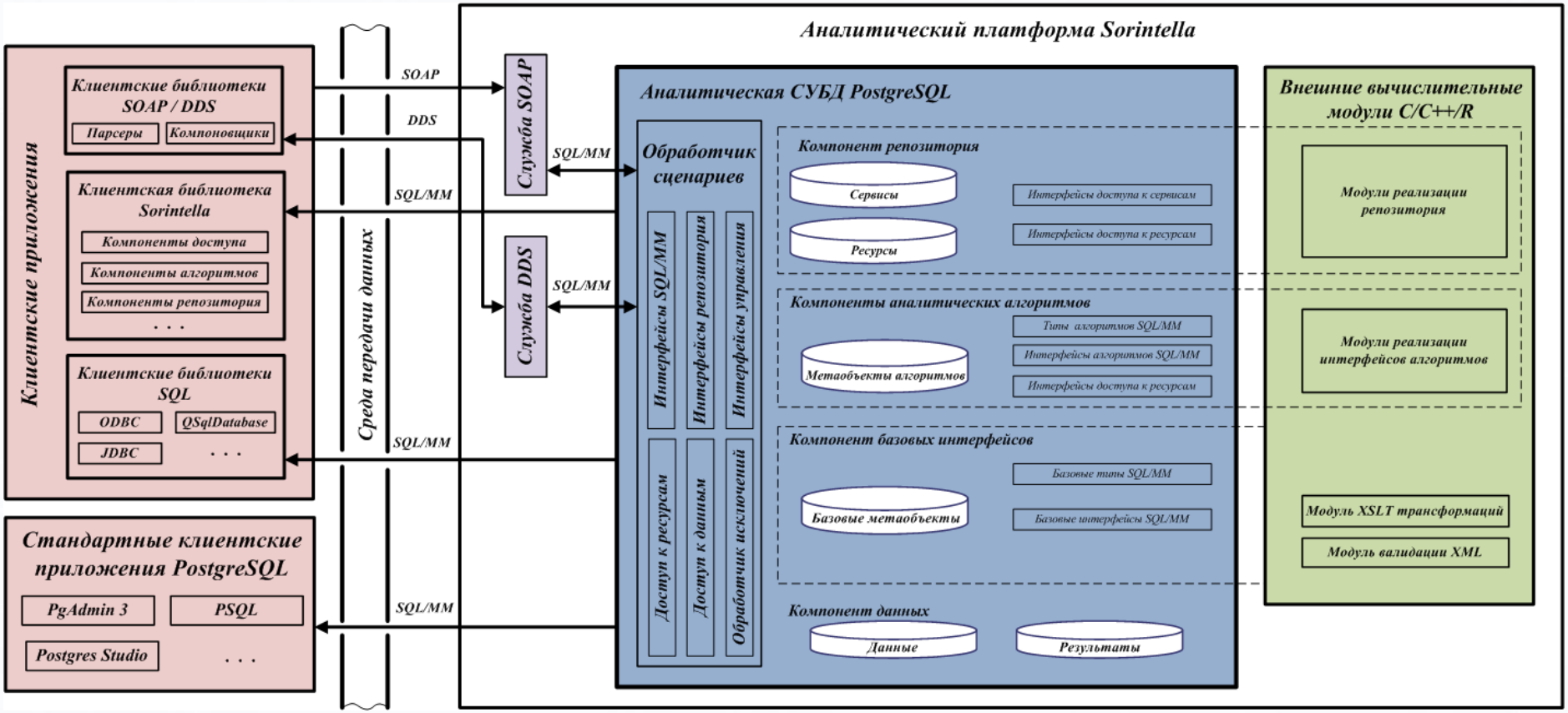
Области применения:

- Системы поддержки принятия решений
- Ситуационные центры
- Системы поддержки бизнеса и операционной деятельности
- Обработка архивных данных
- Распределенный интеллект
- Оперативно-техническое обеспечение средств связи
- Отказоустойчивые инфокоммуникационные системы

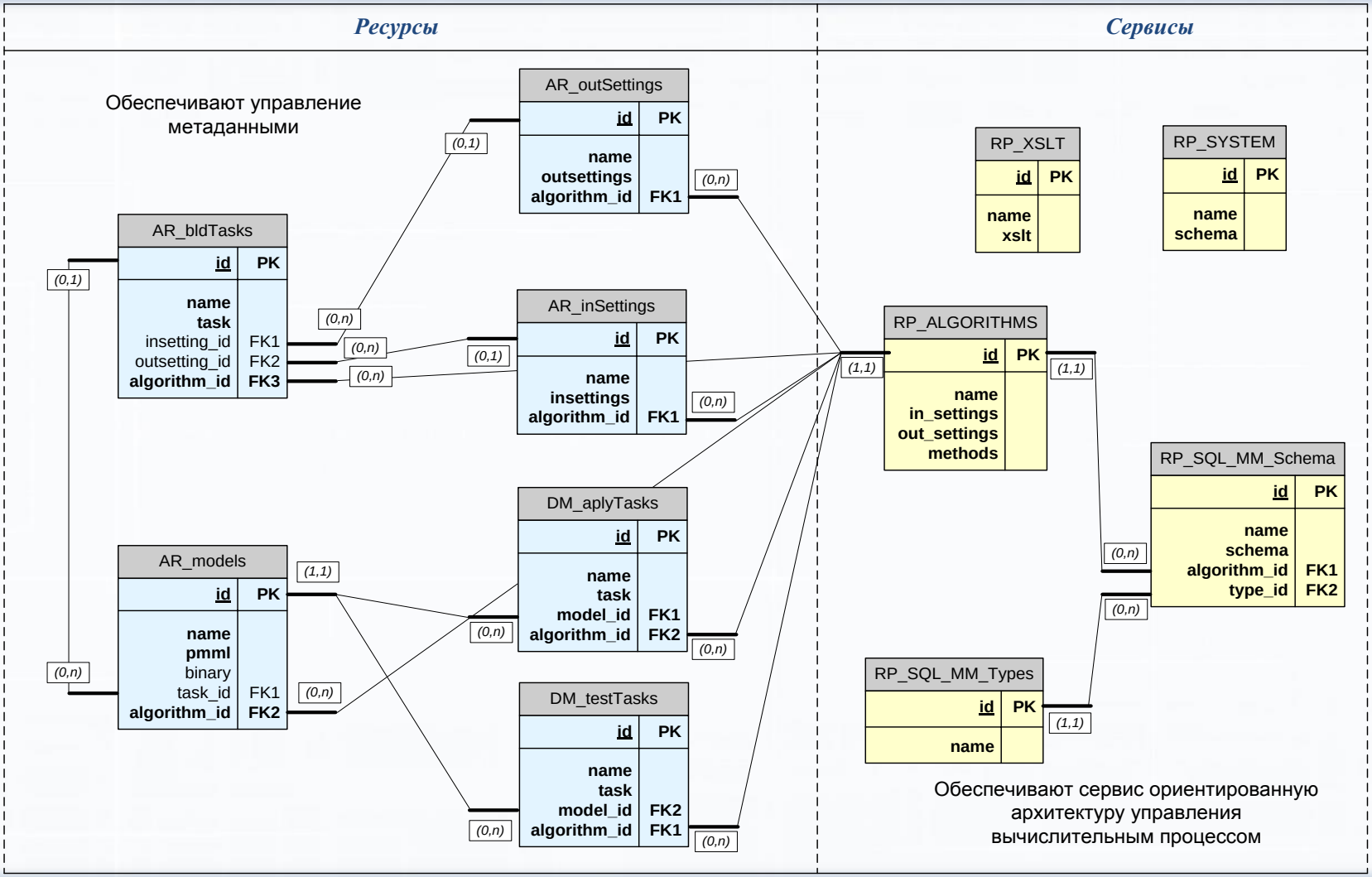
Компонентный подход позволяет:

- Обеспечить масштабируемость программных средств
- Определять состав компонентов в соответствии с решаемыми задачами

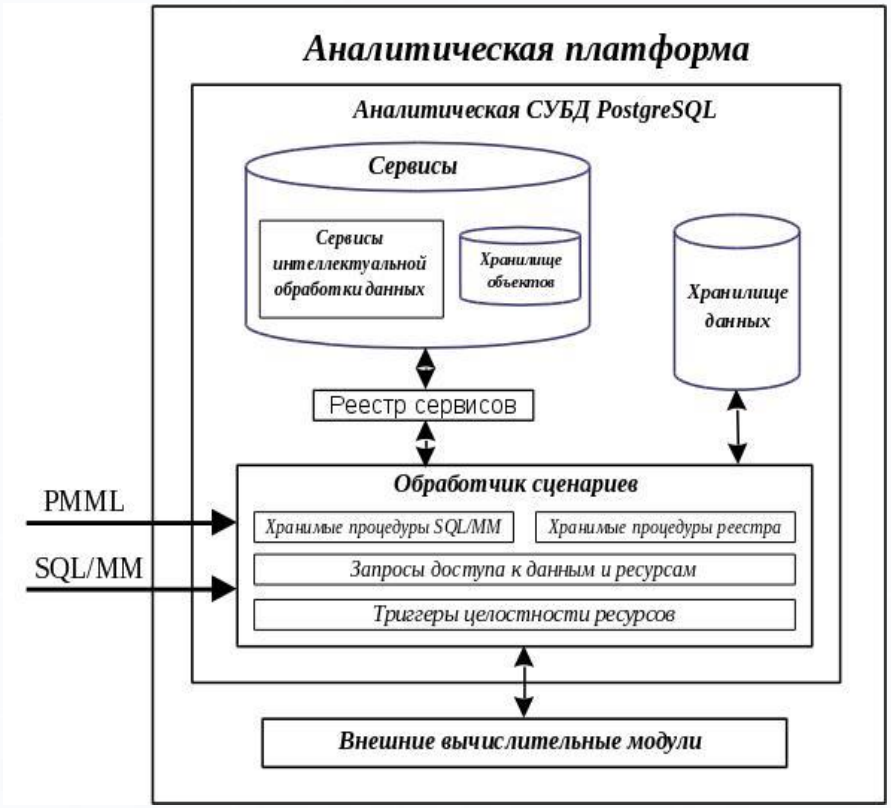
Компонентная архитектура аналитической платформы Sorintella

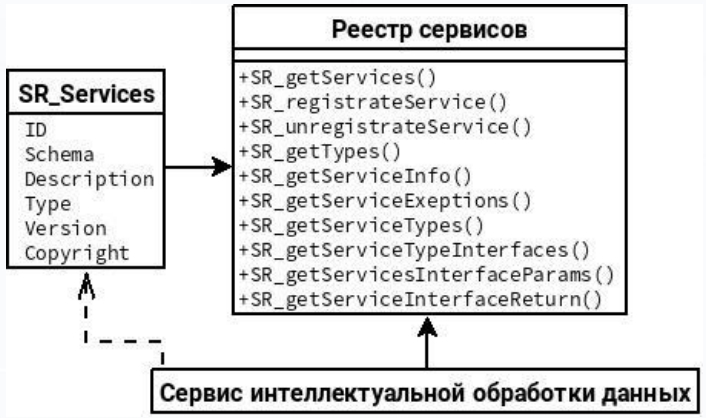


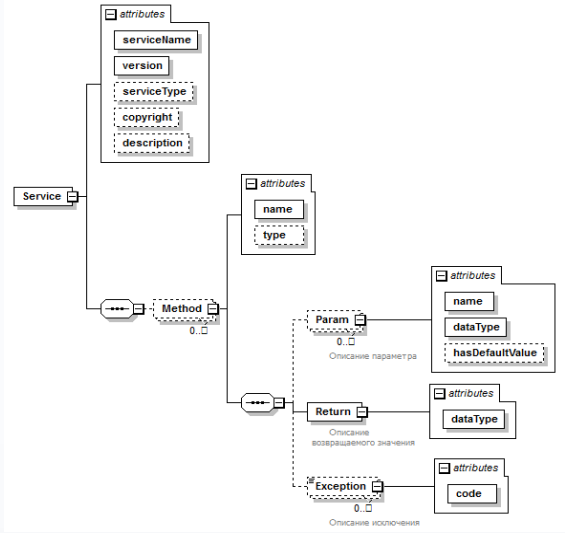
Архитектура репозитория Sorintella



Архитектура реестра сервисов Sorintella







Тактико-технические характеристики Sorinella

Сервер аналитической платформы предназначен для выполнения интеллектуального анализа данных в промышленных системах поддержки принятия решений

Методы интеллектуального анализа данных

Назначение	Метод
Классификация данных	Деревья решений
Предсказание событий	Ассоциативные правила
Распознавание состояний	Нейронные сети
Выявление зависимостей	Регрессионный анализ
Прогнозирование тренда	Анализ временных рядов
Построение когнитивных карт	Корреляционный анализ
Определение ключевых факторов	Структурный анализ
Фильтрация и сжатие данных	Спектральный анализ
Выявление аномальных состояний	Кластерный анализ
Выявление моделей поведения	Системный анализ
Логический вывод	Продукционные правила

Ключевые факторы производительности

- Серверные (хранимые) процедуры позволяют и позволяют существенно уменьшить трафик между клиентом и сервером
- Перенос функционала на RAM диск увеличивает скорость работы в разы
- Использование ускорителей на DSP-процессорах увеличивает скорость вычислений на порядок

Преимущества

- Обработка масштабных массивов разнородной информации (до 32 ТВ)
- Архитектура аналитической подсистемы в соответствии со стандартом SQL/MM Part 6
- Расширяемость аналитики
- Поддержка стандарта PMML для моделей аналитики
- Многоплатформенность
- Низкое требование к ресурсам
- Контроль целостности данных
- Высокая производительность
- Богатый набор типов данных, включая XML, ГИС, сетевые типы
- Наследование таблиц
- Масштабируемость к данным
- Интерфейсы доступа C,C++,C#,python,perl,ruby,php,Lisp и другие языки
- Методы доступа JDBC, ODBC
- Системные интерфейсы DDS, SOAP
- Процедурные языки PL/pgSQL, pl/Tcl, Pl/Perl и pl/Python. Кроме них, существует поддержка PHP, Java, Ruby, R, shell
- Серверные (хранимые) процедуры
- Триггеры и правила для управления процессами
- Система обмена сообщениями между процессами
- Развитая система прав доступа
- Аутентификация пользователя
- Шифрование трафика
- Аппаратное ускорение вычислений
- Возможность использования твердотельных и RAM дисков

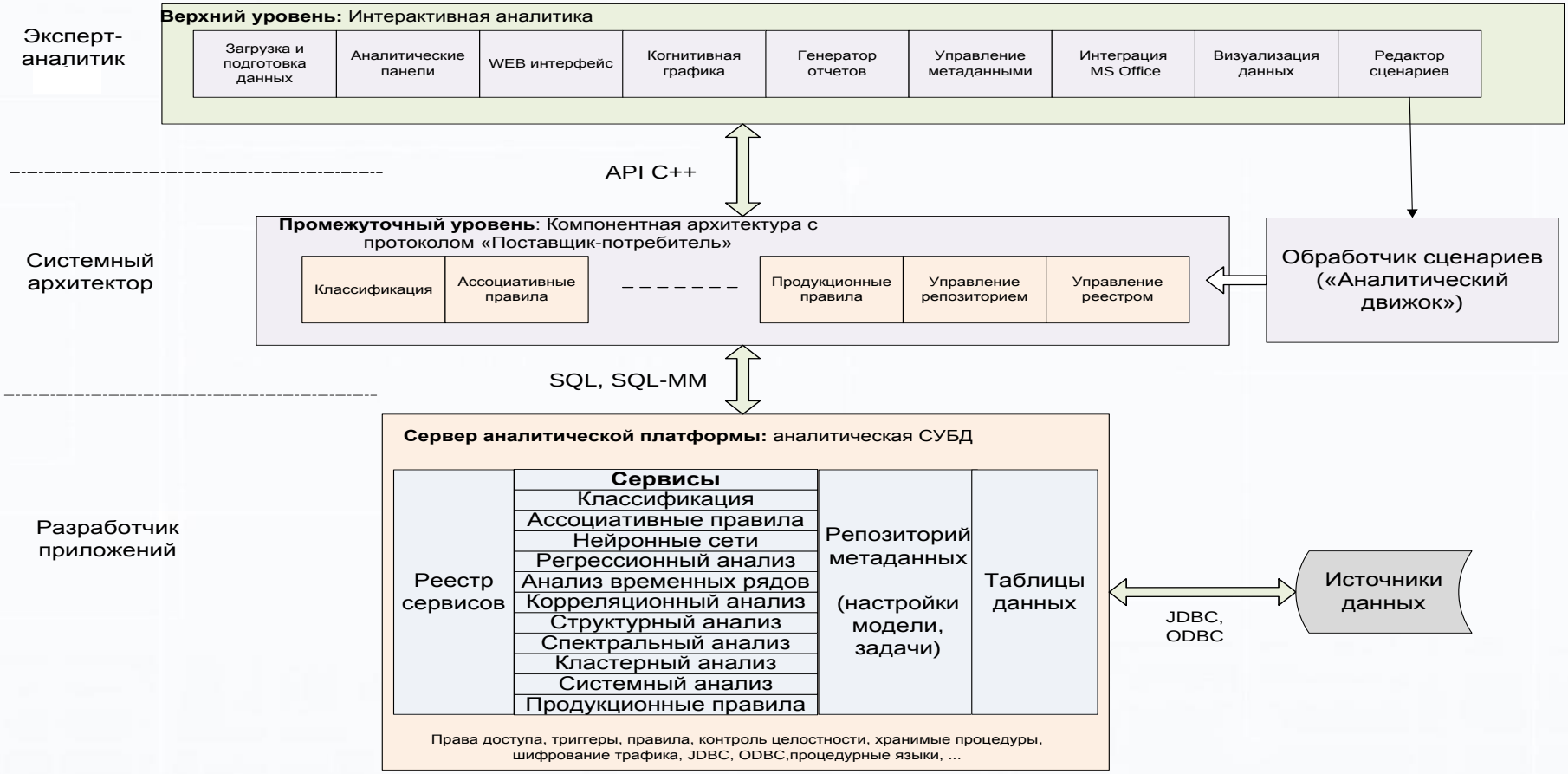
Уровни аналитической платформы

Нижний уровень платформы реализован в реляционной объектно-ориентированной СУБД. Данные обрабатываются там, где они собираются и хранятся. Методы интеллектуальной обработки реализованы в виде хранимых процедур. Для управления вычислительным процессом на этом уровне используется стандартное расширение языка SQL (международный стандарт ISO/IEC 13249-6-2006, SQL/MM Part 6). Аналитические сервисы полностью независимы, компонент реестра обеспечивает сервис-ориентированный доступ к алгоритмам. Формат реализуемых моделей знаний соответствует международному стандарту PMML.

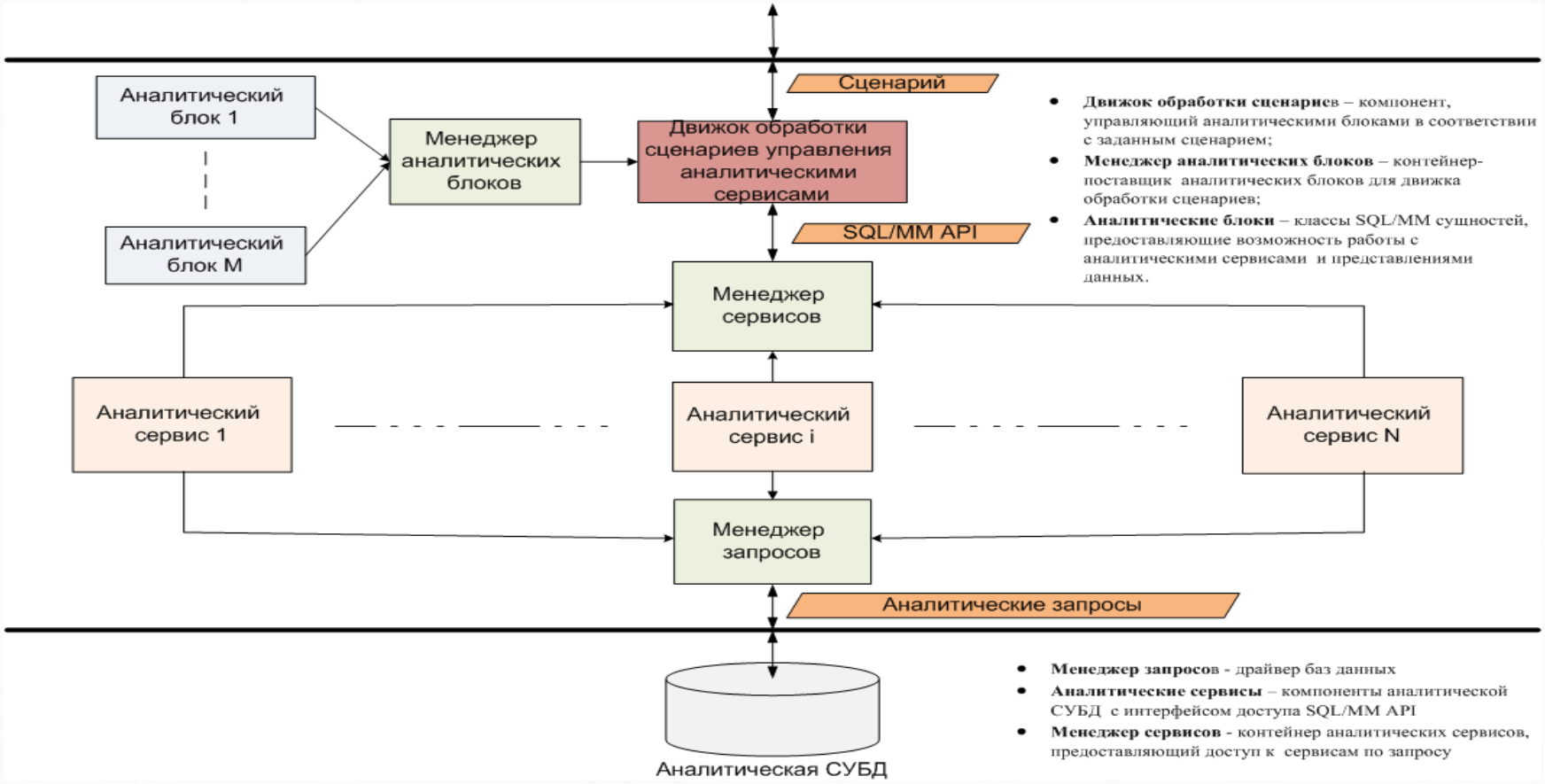
Второй уровень платформы представлен компонентно-ориентированной архитектурой на основе общей шины с протоколом поставщик-потребитель. Пользовательский интерфейс с аналитическими сервисами поддерживается через унифицированные наборы классов. Обработка сценариев реализуется встраиваемым программным движком.

Третий уровень представляет собой компонентную подсистему интерактивного взаимодействия пользователя с развитыми средствами загрузки и подготовки данных, визуального создания сценариев обработки и наглядного графического отображения результатов обработки.

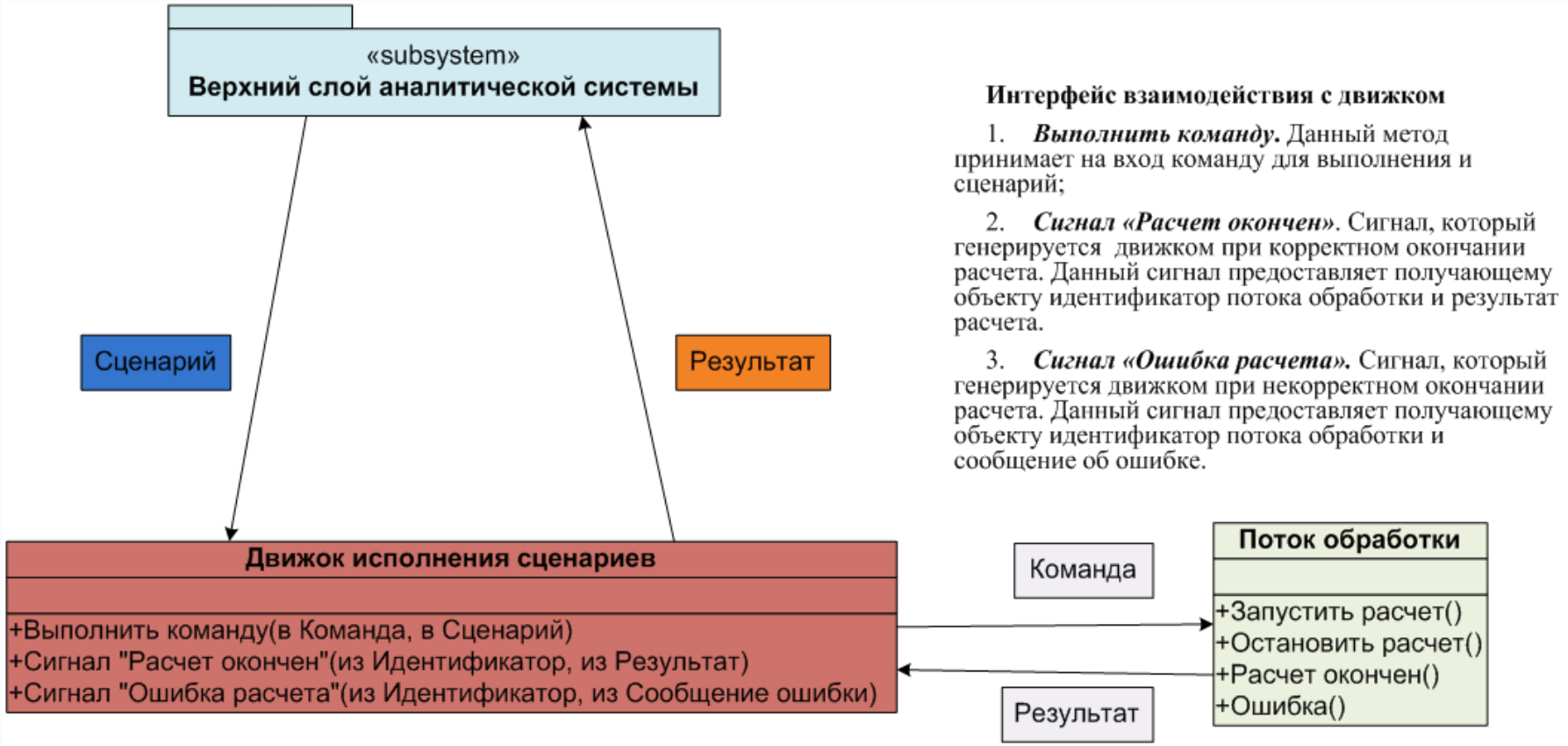
Многоуровневая архитектура аналитической платформы



Второй уровень . Компонентно-ориентированная архитектура



Обработчик сценариев. Аналитический движок



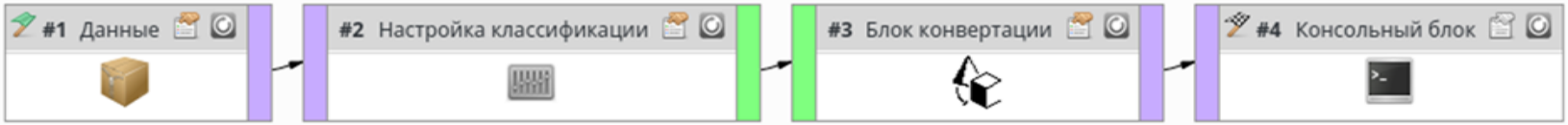
Интерфейс взаимодействия с движком

1. ***Выполнить команду.*** Данный метод принимает на вход команду для выполнения и сценарий;

2. ***Сигнал «Расчет окончен».*** Сигнал, который генерируется движком при корректном окончании расчета. Данный сигнал предоставляет получающему объекту идентификатор потока обработки и результат расчета.

3. ***Сигнал «Ошибка расчета».*** Сигнал, который генерируется движком при некорректном окончании расчета. Данный сигнал предоставляет получающему объекту идентификатор потока обработки и сообщение об ошибке.

Пример аналитического сценария

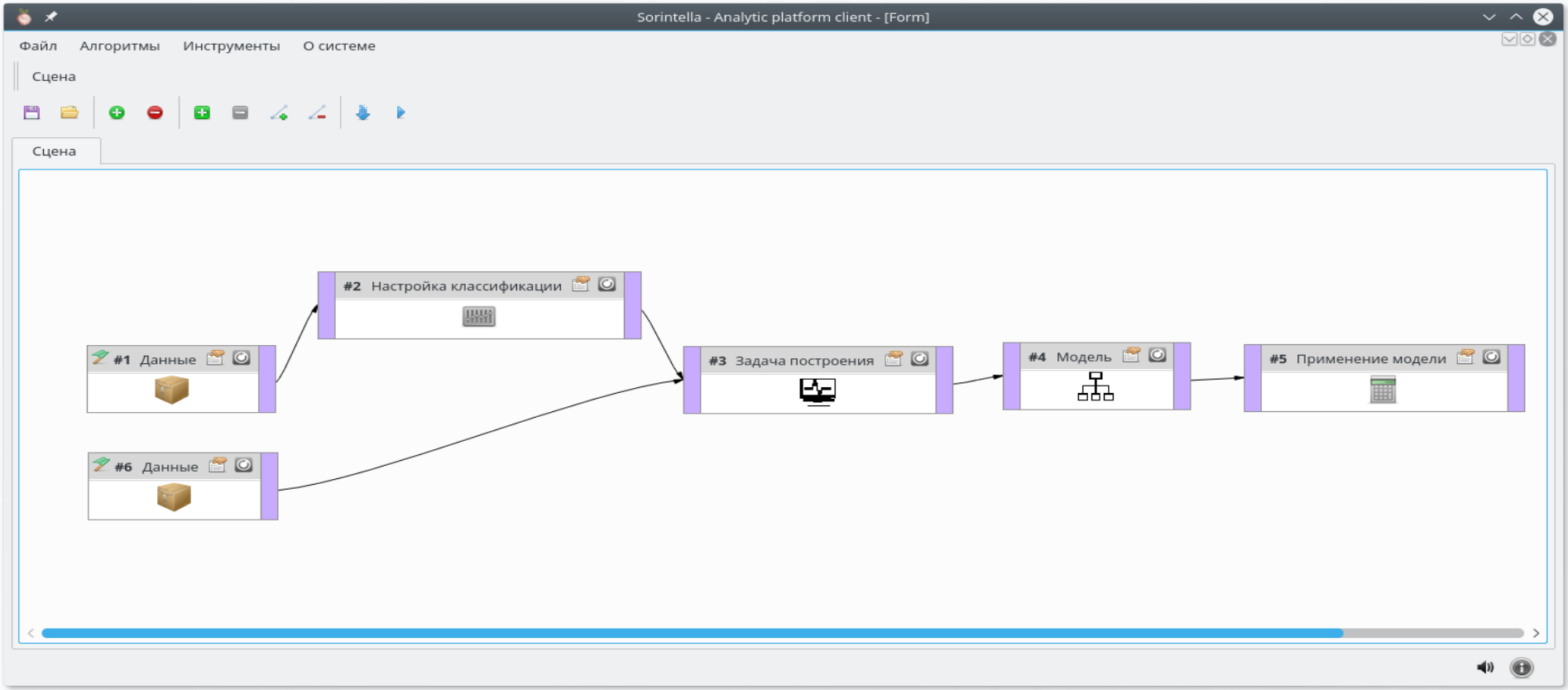


Сценарий обработки данных представляет собой ориентированный граф без петель. В начале графа располагаются узлы данных, которые связываются с узлами аналитической обработки данных. В конце графа размещены узлы представления результатов, полученных путем интеллектуальной обработки данных.

Сценарное описание представляет собой XML документ, содержащий два элемента – описание блока и описание связи.

```
<?xml version="1.0"?>
<Scena>
  <Block>
    <ID>d7e7e848-a604-4ed5-9bc6-ba37dcf6a933</ID>
    <Counter>1</Counter>
    <Property name="Данные">Data</Property>
  </Block>
  <Block>
    <ID>48fd18bf-f6d8-4833-928d-da13793e27bf</ID>
    <Property name="Настройки">Some properties</Property>
    <Counter>2</Counter>
  </Block>
  <Block>
    <ID>24c4ae76-267d-4631-a3ab-8642db26c4b0</ID>
    <Counter>3</Counter>
    <Property name="Тип входного параметра">Settings</Property>
    <Property name="Тип выходного параметра">String</Property>
  </Block>
  <Block>
    <ID>c3999cf7-e188-4db6-bb9e-33f93e0a49c7</ID>
    <Counter>4</Counter>
  </Block>
  <Connector>
    <Start>1</Start>
    <Finish param="input">2</Finish>
  </Connector>
  <Connector>
    <Start>2</Start>
    <Finish param="input">3</Finish>
  </Connector>
  <Connector>
    <Start>3</Start>
    <Finish param="input">4</Finish>
  </Connector>
</Scena>
```

Третий уровень. Графическая среда построения аналитических сценариев



Конкурентные преимущества

- Вся обработка данных выполняется непосредственно в базе данных, что исключает потери производительности связанные с обменами большими массивами данных.
- Клиент-серверная архитектура с унифицированным программным интерфейсом управления вычислительным процессом обеспечивает простое встраивание серверной части в приложения разработчика.
- В состав поставки могут входить только те аналитические методы, которые требуются разработчику приложения. Сервис-ориентированная архитектура обеспечивает пользователя информацией о составе аналитических методов и интерфейсах доступа к ним.
- Аналитический движок - встраиваемый компонент, реализующий выполнение сценариев, кардинально сокращает время разработки аналитических приложений.
- В качестве несущей базы данных используется СУБД с открытыми кодами, таким же требованиям удовлетворяют аналитическое программное обеспечение на языке R. В целом это обеспечивает возможность беспрепятственной сертификации аналитической платформы для специальных применений.
- Интерфейс управления вычислительным процессом соответствует международному стандарту SQL/MM.
- Построенные аналитические модели могут быть экспортированы в формате международного стандарта PMML, что позволяет выполнить их перенос на аналитические платформы других производителей с целью контроля и глубокого сравнительного тестирования.
- Пользовательские типы данных и поддерживающих их методы интерфейса реализованы в технологиях XML и XSLT, что обеспечивает их инсталляцию на любой объектно-ориентированной СУБД с минимальной адаптацией.
- Использование аналитических R-алгоритмов из международного репозитория *CRAN* гарантирует достоверность и надежность аналитических методов, наличие сопутствующей документации, тестовых баз данных и средств автономного тестирования алгоритмов.
- Снижение стоимости аналитической платформы для конечного пользователя достигается: 1) использованием несущей СУБД и алгоритмов аналитики из класса свободного программного обеспечения, 2) масштабированием состава аналитических методов под конкретного пользователя, 3) наличием открытой документации по аналитическим методам, стандартам интерфейса, и аналитическим моделям, 4) использованием готовых R-алгоритмов из международного репозитория *CRAN.*

Спасибо за внимание!