

Приложение № \_\_\_

к Тендерной документации по закупкам товаров –  
 Базовый комплект лицензий –  
 лицензии для аналитической платформы  
 способом открытого тендера  
 с применением торгов на понижение

Перечень закупаемых Товаров

№	Наименование Заказчика	Наименование Товара*	Ед. измерения	Кол-во	Срок поставки Товаров	Место поставки Товаров	Размер авансового платежа, %	Сумма, в тенге (без учета НДС)
1	2	3	4	5	6	7	8	9
1	АО «НАК «Казатомпром»	Базовый комплект лицензий - лицензии для аналитической платформы	Шт.	3	30 (тридцать) календарных дней со дня подписания Договора	Республика Казахстан, г. Астана, ул. Д. Кунаева, 10	0	350 000 000

\* Полное описание и характеристика Работ указывается в Технической спецификации (Приложение № \_\_\_ к Тендерной документации)

Председатель Тендерной комиссии  
 Управляющий директор по ИТ

Исмагамбетов Т.О.

Директор ДИТ

Игенберлин А.Е.

Главный менеджер ДИТ

Нурмаков К.К.

## Техническая спецификация

### 1 ТРЕБОВАНИЯ ПО ИНТЕРАКТИВНОМУ АНАЛИЗУ И ВИЗУАЛИЗАЦИИ

**М01.** Система должна обладать следующими служебными возможностями интерфейса:

- 1) Web-интерфейс, не требующий установки дополнительного ПО для конечных пользователей
- 2) Графический интерфейс, не требующий написания кода
- 3) Русифицированный интерфейс системы
- 4) Возможность настройки интерфейса в корпоративном стиле
- 5) Автомасштабирование под различные разрешения экрана

**М02.** ПО должно удовлетворять следующим требованиям к производительности:

- 1) Обработка данных по технологии in-memory – удержание витрины данных в оперативной памяти распределенного кластера в течение всего времени проведения анализа, т.е. в течение применения к витрине данных большого числа алгоритмов. По данному принципу должны работать все алгоритмы, присутствующие в данных требованиях
- 2) Возможность параллельной обработки больших объемов данных на распределенном кластере:
  - Для таблицы объемом в сотни гигабайт может быть подобран кластер, так чтобы отклик системы для типовых шагов анализа (например, построение гистограммы распределения) составлял до 5 секунд
  - Горизонтальная и вертикальная масштабируемость

**М03.** Система должна удовлетворять следующим требованиям администрирования и безопасности:

- 1) Единый инструмент администрирования
- 2) Поддержка интеграции с MS ActiveDirectory, LDAP
- 3) Поддержка Single Sign-On
- 4) Централизованное управление пользовательскими правами и полномочиями доступа, в том числе возможность авторизации и разграничения прав через MS Active Directory
- 5) Разграничение прав доступа
  - На уровне функций системы
  - На уровне объектов системы
  - На уровне данных

**М04.** Система должна обладать возможностью интеграции с другими системами:

- 1) Поддержка комментариев для визуального анализа и интерактивных отчетов

*Генерал*

- 2) Интеграция с Microsoft Office, как минимум с Excel, PowerPoint, Word, Outlook
- 3) Просмотр интерактивных отчетов с мобильных устройств
- 4) Создание почтовых сообщений со ссылками и скриншотами текущих объектов непосредственно из web-приложений при построении или просмотре отчета, проведении анализа
- 5) Настройка оповещений и почтовой рассылки по условию

**M05.** Система должна удовлетворять следующим требованиям по работе с данными:

- 1) Доступ к различным источникам данных
- 2) Настройка загрузки данных с помощью web-приложения
- 3) Возможность объединения различных источников данных в одном запросе
- 4) Возможность использования кода запроса
- 5) Возможность задание расписания для выполнения запроса
- 6) Простой процесс добавления своих данных
- 7) Поддержка одновременной работы с несколькими витринами данных

**M06.** Интерфейсы анализа и визуализации для конечных пользователей должны отвечать следующим минимальным требованиям:

- 1) Представление данных для пользователей в виде измерений и метрик
- 2) Создание вычисляемых показателей бизнес-пользователями. Должны быть доступны все основные математические, статистические, строковые, арифметические, условные операции, операции преобразования типов данных, функции работы с датой и временем
- 3) Создание вычисляемых категорий бизнес-пользователями
- 4) Создание пользовательских иерархий
- 5) Изменение способов агрегации показателей;
- 6) Создание расширенных фильтров на основании нескольких полей, содержащих несколько условий, несколько категорий и/или показателей, функции
- 7) Интерактивное создание фильтров выделением области на графике
- 8) Просмотр сводной статистики по всем показателям в одном окне
- 9) Возможность использования нескольких источников данных в проведении исследования

**M07.** Интерфейсы анализа и визуализации для конечных пользователей должны поддерживать использование аналитических методов в исследовании:

- 1) Обработка данных
  - Поиск и фильтрация выбросов
  - Создание вычисляемых полей
  - Динамическое исключение групп переменных
- 2) Исследование данных
  - Статистические средства анализа данных, в том числе построение распределений, вычисление описательных статистик
  - Анализ потоков данных с помощью Сенки-диаграмм во времени
  - Сетевые диаграммы
  - Ящичковая диаграмма
  - Облако слов
  - Расчет корреляции
- 3) Моделирование

*Александр*

- Прогнозные модели
  - Линейная регрессия
  - Логистическая регрессия
  - Общая линейная регрессия
  - Деревья решений
- Описательные модели
  - Кластерный анализ
- Возможность группового построения моделей для различных сегментов
- Автоматический перерасчет моделей при обновлении данных
- 4) Оценка и скоринг моделей
  - Сравнение моделей, используя lift и ROC кривые, misclassification и др.
  - Интерактивное определение порога отсечения для принятия решений
  - Скоринг данных в системе
  - Выгрузка кода для скоринга моделей

**M08.** В системе должны присутствовать инструменты создания интерактивной отчетности и информационные панели, соответствующие минимальным требованиям:

- 1) Поддержка анимированных диаграмм
- 2) Возможность формирования произвольных пользовательских иерархий
- 3) Возможность интерактивной детализации от верхнего уровня данных к исходным
- 4) Простая настройка интерактивности: зависимости объектов для фильтрации и выделения
- 5) Настройка детализации в другие разделы отчета/в другие отчеты/в другие системы путем формирования динамических ссылок
- 6) Пользовательский интерфейс для создания тем отчетов, содержащих заранее заданные цвето-шрифтовые схемы для различных видов объектов
- 7) Настройка порогового значения и рассылки оповещения об его достижении на почту или в форме sms-сообщения
- 8) Возможность настройки регламентированной рассылки отчетов на почту в виде прямой ссылки на отчет и вложения в формате pdf
- 9) Автоматическое обновление данных и оповещение пользователя в случае, если пользователь работает с данными в момент обновления
- 10) Возможность использования нескольких источников данных для построения отчета.

## 2 ТРЕБОВАНИЯ ПО ПРОГНОЗИРОВАНИЮ

**M09.** В системе должна быть реализована функциональность для моделирования и ситуационного анализа, содержащая следующие алгоритмы прогнозирования временных рядов:

- 1) Модели семейства ARIMAX;
- 2) Методы экспоненциального сглаживания;
- 3) Методы прерывистого спроса (Intermittent Demand Models);
- 4) Методы ненаблюдаемых компонент (Unobserved Components Models);
- 5) VARMAX модели для прогнозирования вектора временных рядов.

**M010.** Система должна поддерживать возможность прогнозирования иерархий:

- 1) Встроенный конструктор для создания иерархии временных рядов;

*Финанс*

- 2) Автоматическое формирование прогнозов на всех уровнях иерархии;
- 3) Автоматический подбор индивидуальной модели для каждого временного ряда;
- 4) Автоматическое согласование результатов прогнозирования (приведение к единому суммарному значению) на разных уровнях иерархии с возможностью изменения уровня согласования: сверху-вниз, снизу-вверх, из середины.

**M011.** Система должна поддерживать выбор и настройку композиции моделей прогнозирования (ансамблевые модели).

**M012.** В системе должна присутствовать возможность выделения валидационного периода для автоматического выбора наилучшей модели прогнозирования.

**M013.** Процесс подготовки данных и построения прогнозов должен быть автоматизирован:

- 1) Для обеспечения необходимого регламента оперативная работа алгоритмов отчуждаема от процессов анализа и настройки работы алгоритмов. Таким образом, система работает в трех режимах:
  - a. Первичного анализа и настройки
  - b. Тестовой эксплуатации алгоритмов анализа данных
  - c. Регламентного промышленного выполнения
- 2) Требования к процессам первичного анализа и настройки:
  - a. Процесс аналитической предподготовки данных выстраивается в виде пошаговой диаграммы в GUI
  - b. Результаты каждого этапа обработки данных выстраиваются в виде таблиц/графиков в GUI
  - c. Процесс аналитической предподготовки данных может быть:
    - i. Сохранен
    - ii. Отредактирован
    - iii. Передан другому пользователю
    - iv. Интегрирован с прогностическими моделями на уровне данных и метаданных

**M014.** Построение прогнозов должно осуществляться автоматически с помощью GUI:

- 1) Для построения прогнозов достаточно указать системе с помощью point-and-click интерфейса следующие параметры:
  - a. Целевая переменная
  - b. Переменная временного идентификатора
  - c. Группирующие переменные
  - d. Внешние факторы, влияющие на целевую переменную
  - e. Горизонт прогнозирования
- 2) Система выбирает наилучшую модель из следующего семейства:
  - a. Модели семейства ARIMAX;
  - b. Методы экспоненциального сглаживания;
  - c. Методы прерывистого спроса (Intermittent Demand Models);
  - d. Методы ненаблюдаемых компонент (Unobserved Components Models);
- 3) Из каждого семейства модель подбирает наилучшую модель. Например, для семейства моделей ARIMAX:
  - a. Автоматический подбор параметров (p, d, q);
  - b. Автоматический выбор наиболее значимых из указанных внешних факторов;

*Лужина*

- c. Автоматический выбор способа влияния внешнего фактора на целевую переменную (подбор передаточной функции);
  - d. Автоматический выбор оптимальной по критерию MAPE модели на валидационном периоде;
  - e. Автоматический расчет коэффициентов при параметрах модели.
- 4) Для построения прогнозов навыки программирования не требуются
  - 5) Пользователь, не обладающий математическими знаниями в области прогнозирования, имеет возможность построить базовый прогноз
  - 6) Пользователь, обладающий математическими знаниями в области прогнозирования, имеет возможность донастроить уже существующие модели и построить собственные с помощью GUI

**M015.** Требования к процессу тестовой эксплуатации алгоритмов анализа данных:

- 1) Процесс может быть поставлен на регламент в том же графическом интерфейсе, в котором проводилась подготовка данных
- 2) На обновленных данных система строит прогнозы в следующих режимах:
  - a. Диагностирование: Система автоматически строит наилучшую модель для каждого ряда
  - b. Выбор: Производится переоценка коэффициентов при параметрах моделей из уже существующего репозитория, затем из этих моделей выбирается наилучшая
  - c. Настройка: Модели для каждого ряда остаются те же, производится переоценка коэффициентов при параметрах модели
  - d. Прогноз: Модели и коэффициенты при параметрах модели остаются те же, обновленные данные подставляются в уже существующие уравнения

**M016.** Требования к регламентному промышленному выполнению:

- 1) Алгоритм построения прогнозов оформляется аналитиком в виде специальной сущности системы и она передается через внутренний слой метаданных (не через файловую систему, почту, физически и т.д.) для обеспечения надежности данного процесса
- 2) Постановка на регламент на сервере осуществляется с помощью GUI.
- 3) Постановка на регламент на сервере осуществляется с помощью интерфейса – отдельного от интерфейса первичной настройки и анализа, чтобы при необходимости этот процесс мог выполняться и сопровождаться отдельными ИТ-специалистами.

**M017.** Сценарный анализ («what-if», «что, если» анализ) с использованием GUI - интерактивный анализ влияния изменения независимых переменных на прогнозную переменную.

### 3 ТРЕБОВАНИЯ ПО ОПТИМИЗАЦИИ

**M018.** Система должна включать в себя следующие методы и алгоритмы для решения линейных (в том числе целочисленных и смешанно-целочисленных), квадратичных и общих нелинейных задач математической оптимизации (включая оптимизацию плана производства, составление оптимального расписания и интегрированное планирование):

- 1) прямой, двойственный и сетевой симплекс-методы;

*Лущина*

- 2) метод ветвей и границ с секущими плоскостями, поиском конфликтов и эвристикой;
- 3) метод внутренней точки с доверенной областью и линейным поиском;
- 4) метод активного множества с доверенной областью и линейным поиском;
- 5) алгоритмы для решения стандартных задач сетевой оптимизации: решение задачи коммивояжера, поиск потока минимальной стоимости, нахождение кратчайшего пути, поиск компонент связности и циклов;
- 6) методы локального поиска для оптимизационных задач с целевой функцией произвольного вида (в том числе нелинейной, недифференцируемой или разрывной);
- 7) алгоритмы многоцелевой оптимизации;
- 8) алгоритмы предварительной оптимизации;
- 9) метод декомпозиции для линейных задач, имеющих блочную структуру;
- 10) метод мультистарта для повышения вероятности нахождения глобального экстремума в нелинейных задачах со многими локальными экстремумами;
- 11) генетические алгоритмы.

**M019.** Система должна включать в себя методы и алгоритмы для дискретного моделирования событий и процессов с поддержкой:

- 1) генерации данных методом Монте-Карло;
- 2) планирования экспериментов;
- 3) иерархического моделирования;
- 4) выполнения моделей в режиме реального времени;
- 5) управления статическими и мобильными ресурсами;
- 6) анализа входных параметров и результатов;
- 7) создания, сохранения и редактирования собственных моделей и компонент.

**M020.** Система должна включать в себя методы и алгоритмы для планирования проектов и распределения ресурсов с поддержкой:

- 1) определения доступности ограниченных ресурсов;
- 2) использования календарей и графиков работы;
- 3) использования отношений приоритетности;
- 4) управления освоенными объемами;
- 5) анализа решений;
- 6) вычисления значения идеальной информации (VPI);
- 7) вычисления значения идеального контроля (VPC).

**M021.** Система должна включать в себя методы и алгоритмы для управления спецификациями материалов с поддержкой многоуровневых и вложенных спецификаций.

**M022.** Система должна предоставлять масштабируемую интегрированную программную среду, включающую в себя инструменты:

- 1) для чтения данных в форматах TXT, CSV, XML, DOC, RTF, XLS, XLSX, MDB;
- 2) для осуществления преобразований и обработки данных;
- 3) для формулирования задач на языке оптимизационного моделирования;
- 4) для настройки параметров алгоритмов решения задач оптимизации;
- 5) для создания, редактирования и выполнения хранимых процессов;
- 6) для графического моделирования дискретных событий и процессов;
- 7) для визуализации данных в форматах ActiveX, Java Applet, JPG, PNG, GIF;
- 8) для создания отчетов в форматах HTML, PDF, RTF;

*Линия*

- 9) для создания отчетов, настраиваемых диаграмм Ганта и сетевых графиков при планировании проектов;
- 10) для экспорта результатов в форматах TXT, CSV, XML, DOC, RTF, XLS, XLSX, MDB.

**M023.** Система должна предоставлять масштабируемую аналитическую инфраструктуру для осуществления сложных вычислений на больших объемах данных с поддержкой:

- а) многопоточного режима (с распараллеливанием вычислений по ядрам процессора) для решения оптимизационных задач и моделирования дискретных событий;
- б) распределенного режима (с распараллеливанием вычислений по элементам кластера) для решения оптимизационных задач с помощью основных методов математической оптимизации (как минимум, локального поиска, декомпозиции и мултистарта);
- в) обработки данных по технологии in-memory (разовой загрузки всей анализируемой витрины данных в оперативную память распределенного кластера и применения к ней аналитических алгоритмов) для решения оптимизационных задач с помощью основных методов математической оптимизации (как минимум, локального поиска, декомпозиции и мултистарта).

#### **4 ТРЕБОВАНИЯ К ИНТЕЛЛЕКТУАЛЬНОМУ АНАЛИЗУ ДАННЫХ**

**M024.** Система должна иметь графический интерфейс пользователя для построения процессов анализа данных в виде диаграмм, состоящих из последовательностей конфигурируемых узлов обработки данных, без необходимости написания кода.

**M025.** Система должна обеспечивать создание шаблонов процессов обработки и анализа данных и построения отчетности: частично или полностью процесс может применяться к другим данным без перенастройки параметров.

**M026.** Система должна поддерживать полный цикл разработки аналитических моделей:

- 1) доступ к данным
- 2) исследование данных
- 3) подготовка данных для моделирования
- 4) выбор и настройка аналитических моделей
- 5) оценка качества моделей
- 6) ввод в промышленную эксплуатацию

**M027.** Система должна иметь инструменты расширения функциональности:

- 1) возможность разработки собственных процедур обработки данных и аналитических моделей на внутреннем языке Системы
- 2) интеграция с R
- 3) поддержка PMML для обмена моделями

**M028.** Система должна иметь возможность импортировать и экспортировать источники данных в процессе разработки моделей:

- 1) таблицы данных промышленных СУБД
- 2) текстовые файлы с разделителем

*Гудин*



3) файлы MS Office

4) XML-файлы

**M029.** Система должна поддерживать одновременную работы с несколькими источниками данных в рамках сессии.

**M030.** Система должна обеспечивать пользователям простой процесс подготовки витрин данных без написания кода:

- 1) объединение таблиц данных, загруженных из разных источников
- 2) фильтрация и выбор объектов для анализа
- 3) транспонирование и агрегация данных
- 4) преобразования исходных характеристик объектов

**M031.** Система должна обеспечивать создание вычисляемых полей, в том числе с использованием статистических и строковых преобразований, преобразований дат и времени, макропеременных.

**M032.** Система должна содержать инструменты восстановления пропущенных (неизвестных) характеристик объектов:

- 1) без учета известных аналогов в данных - медианой, средним, модой, M-оценкой
- 2) с учетом профилей и аналогов в данных - с помощью деревьев решений

**M033.** Система должна обеспечивать преобразование типов данных и трансформацию исходных характеристик объектов, в том числе оптимальную бинизацию количественных полей, т.е. превращение диапазонов значений количественных полей в категории.

**M034.** Система должна позволять проводить сегментацию объектов с использованием методов автоматического поиска групп схожести объектов по количественным и качественным характеристикам (K-Means и SOM)

**M035.** Система должна предоставлять возможность интерпретировать сформированные группы схожести путем их профилирования

**M036.** Система должна обеспечивать формирование прецедентной выборки для выбранного объекта - нахождение заданного числа ближайших аналогов методом k-ближайших соседей.

**M037.** Система должна позволять проводить анализ последовательностей событий на основании транзакционных данных - метод поиска ассоциаций с поддержкой расчета значимости правил.

**M038.** Система должна обеспечивать анализ значимости характеристик объекта в связи с целевым событием/показателем, на основе чего производить отбор характеристик:

- 1) формирование независимых статистических характеристик - метод главных компонент
- 2) определение взаимосвязанных и исключение малоинформативных характеристик объектов - корреляционный анализ, кластеризация характеристик
- 3) выбор характеристик объектов, определяющих целевое событие - регрессионный анализ, деревья решений, алгоритм Random Forest.

**M039.** Система должна обеспечивать выявление аномальных объектов и выбросов:

- 1) определение стандартных коридоров и выбросов для индивидуальных характеристик объектов анализа - методы выявления простых однофакторных аномалий на основе анализа распределений

*Душман*

- 2) определение связанных характеристик объектов, стандартных коридоров их совместного изменения и выбросов - методы выявления многофакторных аномалий на основе корреляционного, регрессионного и кластерного анализа.

**M040.** Система должна позволять формировать логические правила (профили) выявления целевых событий при помощи деревьев решений:

- 1) в интерактивном режиме, при котором формирование правил производится пользователем итеративно с учетом рекомендации Подсистемы по отбору значимых характеристик и оптимальному разбиению объектов
- 2) с возможностью выбора информационного критерия разбиения объектов (Gini, Chi-square, Entropy)
- 3) с возможностью задания минимального охвата правил разбиения

**M041.** Система должна иметь режим автоматической генерации и выбора лучшей аналитической модели без использования расширенных настроек подсистемы из готовых шаблонов моделей.

**M042.** Система должна включать продвинутые методы углубленного анализа данных:

- 1) регрессионные модели, в том числе обобщенные линейные модели (glm) с выбором функций распределения и связи
- 2) нейронные сети (в т.ч. с автоподбором архитектуры)
- 3) алгоритм SVM
- 4) ансамблевые модели (boosting, bagging)
- 5) алгоритм Random Forest.

**M043.** Система должна иметь возможность прогнозировать дополнительные параметры целевого события

- 1) время до наступления события на основе алгоритмов анализа выживаемости;
- 2) величину потенциального ущерба одновременно с вероятностью возникновения события

**M044.** Система должна позволять создавать тренировочны валидационную и тестовую выборки для оптимизации сложности и оценки точности моделей

**M045.** Система должна предоставлять возможность анализа точности и стабильности полученных моделей на основе показателей Gini, KS (Kolmogorov-Smirnov), C-statistic, ROC curve, Lift Curve, Cumulative Lift Curve

**M046.** Система должна быть функционально масштабируема и иметь возможность без нарушения технологических и методологических процессов анализа данных и скоринга дополнена функциональностью автоматической перенастройки параметров моделей

- 1) при обновлении и поступлении новых данных
- 2) при наступлении определенных пользователем событий

**M047.** Система должна быть функционально масштабируема и иметь возможность без нарушения технологических и методологических процессов анализа данных и скоринга дополнена функциональностью автоматической перенастройки параметров моделей

**M048.** Система должна иметь широкие возможности применения и запуска в промышленную эксплуатацию скоринговых моделей

- 1) в сторонних приложениях – генерация скорингового кода SAS, C, JAVA, PMML
- 2) на стороне СУБД - перенос процедур обработки данных на сторону СУБД, возможности расширения функционала для переноса процедур скоринга на сторону СУБД

*Будина*

**M049.** Система должна обеспечивать регламентное применение моделей к новым данным - постановка моделей на скоринг по расписанию/по триггеру

**M050.** Система должна обеспечивать возможности мониторинга качества промышленных моделей во времени.

## **5 ОБЩИЕ ТРЕБОВАНИЯ К ПОСТАВЩИКУ**

---

**M051.** Потенциальный Поставщик в течение 15 календарных дней с даты поставки Товара должен установить и настроить программное обеспечение на оборудовании Заказчика.

**M052.** Потенциальный Поставщик обязан предоставить ссылку на официальный сайт с технической документацией подтверждающий функционал программного обеспечения соответствующий технической спецификации.

**M053.** Система должна обладать централизованными метаданными для всех компонент системы, отвечающих за интерактивный анализ и визуализацию, прогнозирование, оптимизацию, интеллектуальный анализ данных.

**M054.** Система должна обладать единым интерфейсом для администрирования прав пользователей и системных настроек. Интерфейс должен позволять:

- 1) Определять настройки логических серверов (служб) и других элементов инфраструктур (в т.ч. параметры их производительности – балансировку, кэширование и др.)
- 2) Задавать параметры библиотек данных, а также параметры подключения и доступа к внешним источникам/базам данных
- 3) Централизованно управлять всеми пользователями системы и определять роли/права доступа для всех объектов (включая метаданные, данные, приложения, службы и др.)
- 4) Осуществить сквозной «анализ влияния», а именно:
  - а. Оценить, на каких структурах данных, витринах, отчетах и приложениях скажется изменение содержимого таблицы или отдельного поля в системе-источнике (прямой анализ влияния);
  - б. Выявить, изменения в каких процессах преобразования данных, а также таблицах и полях систем-источников могли повлечь изменение содержимого показателя в рамках витрины данных или ячейки отчета (обратный анализ влияния).

**M055.** Система должна обладать веб-интерфейсом для мониторинга работоспособности системы, состояния запущенных сервисов, анализа логов, уведомления администратора в случае возникновения внештатных ситуаций.

Председатель Тендерной комиссии  
Управляющий директор по ИТ

Директор ДИТ

Главный менеджер ДИТ

Исмагамбетов Т.О.

Игенберлин А.Е.

Нурмаков К.К.