АГЕНТСТВО СТРАТЕГИЧЕСКИХ ИНИЦИАТИВ

Методические рекомендации по внедрению управления на основе данных

Введение

Данный документ, основывается на международной методологии исследования данных CRISP-DM, содержит методические рекомендации по работе с проектами, связанными с анализом данных, поддерживает внедрение управления основанного на данных и предназначен для:

- Заказчиков выполнения проектов и других заинтересованных в их результатах лиц.
- Постановщиков задач.
- Руководителей проектных команд.
- Участников проектных команд.
- Поставщиков и иных контрагентов, задействованных в проектах.

Данные роли распределяются согласно организационной структуре в субъекте Российской Федерации, а также с учетом отрасли в которой решается задача.

Первый раздел содержит глоссарий и предметный указатель. Второй раздел представляет пошаговые инструкции по выполнению этапов работ. Во третьем разделе раскрывается основная методология CRISP-DM, в нём же даны методические рекомендации по выделению крупных этапов работ, основанные на опыте проектов, связанных с анализом данных с 2009 года. Детальные методические рекомендации приведены в четвертом разделе.

В приложениях приведены:

- Приложение № 1 “Описания наборов данных с примерами”
- Приложение № 2 “Примеры наборов данных”
- …

Команда авторов:

- Алексей Драль
- Андрей Петров
- Вера Адаева
- Иван Бегтин
- Яна Коваленко
- ...

DISCLAIMER: если Вы читаете эту версию документа, то скорее всего являетесь экспертом в области анализа данных. Написать хороший документ без обсуждения разных точек зрения не представляется возможным. Поэтому мы будем премного благодарны, если Вы поделитесь своими комментариями, опытом и рекомендациями.

1. Глоссарий

Машинное обучение, МО (англ. machine learning, ML) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач. Для построения таких методов используются средства математической статистики, численных методов, методов оптимизации, теории вероятностей, теории графов, различные техники работы с данными в цифровой форме.

Программа и методика испытаний, ПМИ (англ. program and methods of testing, PMT) — раздел единой системы программной документации (семейство стандартов ГОСТ 19), определяющий требования, подлежащие проверке при испытании программного обеспечения, а также порядок и методы их контроля.

ETL (от англ. Extract, Transform, Load — дословно «извлечение, преобразование, загрузка») — один из основных процессов в управлении хранилищами данных, который включает в себя: извлечение данных из внешних источников; их трансформацию и очистку, чтобы они соответствовали предъявленным требованиям; и их загрузку в хранилище данных.

DevOps (акроним от англ. development и operations; по-русски обычно произносится как «девóпс») — технология (методология) активного взаимодействия специалистов по разработке со специалистами по информационно-технологическому обслуживанию и взаимную интеграцию их рабочих процессов друг в друга для обеспечения качества продукта. Предназначена для эффективной организации создания и обновления программных продуктов и услуг. Основана на идее тесной взаимозависимости разработки и эксплуатации программного обеспечения.

DataOps (акроним от англ. data и operations; по-русски обычно произносится как «дейтаóпс») — автоматизированная, процессно-ориентированная методология, используемая аналитическими командами, для улучшения качества и сокращения времени цикла аналитики данных.

Неструктурированные или слабо структурированные данные - данные имеющие слабо выраженную структуру, не имеет её вовсе, либо неопределённую структуру. Представлены, как правило, виде текстов, которые могут включать сведения о датах, значениях, интервалах и прочих фактах, подчинённых контексту описания.

СУБД - система управления базами данных - совокупность программных и лингвистических средств, обеспечивающих управление созданием и использованием баз данных.

NoSQL - подход к хранению данных, который опирается на принципах атомарности (изменений) и согласованности (данных). Является альтернативой традиционным реляционным способам хранения данных в СУБД. Ярким представителем здесь является MongoDB.

UX - User experience, опыт пользователя, опыт взаимодействия: восприятие и ответные действия пользователя, возникающие в результате использования (или предстоящего использования) продукта, системы или услуги. Сокращённое название термина “человеко-ориентированное проектирование” (ISO 9241-210 / ГОСТ Р ИСО 9241-210).

UI - User interface, пользовательский интерфейс - интерфейс, обеспечивающий человеко-машинное взаимодействие.

HTML - HyperText Markup Language, язык гипертекстовой разметки - стандартизированный язык разметки текстов, который используется для формирования представления и интерфейсов во Всемирной паутине.

HTML-тег - это элемент разметки гипертекста, как правило, не имеющий текстового выражения, но определяющий внешний вид заключённого в себя текста. Примеры: </br> - перевод строки, <p> - абзац, <h> - заголовок, <table> - таблица и так далее.

Ассессмент - оценка, как метод всестороннего исследования объектов и/или субъектов, целью которого является формирование рекомендаций. В контексте настоящей методологии:

- ассессмент персонала - методика поддержки внедрения изменений
- ассессмент данных - исследование и оценка качества данных

QA - [Software] Quality Assurance - набор методик, технологий и процедур мониторинга цикла разработки программного обеспечения предназначенный для обеспечения гарантий его качества.

CI - Continuous Integration, непрерывная интеграция - практика разработки программного обеспечения, заключающаяся в обеспечении автоматического слияния рабочих копий в основной ствол разработки, сборки проектов и их проверки для скорейшего выявления ошибок и интеграционных проблем. Является одним из базисных элементов экстремального программирования.

HR - human resources, человеческие ресурсы - область знаний, методики и дисциплины, связанные с изучением, организацией и управлением человеческим капиталом.

ГОСТ - Межгосударственный стандарт - региональный стандарт, пришедший на смену государственным стандартам СССР, принятый межгосударственным советом по стандартизации, метрологии и сертификации СНГ. Применяется, как правило, добровольно.

ISO - International Organization for Standardization, международная организация по стандартизации - международная организация, занимающаяся выпуском стандартов.

CSV - Comma Separated Values - текстовый формат, применяемый для передачи табличных данных, подчиняющийся RFC 4180, являющегося стандартом де-факто.

A/B-тестирование - метод маркетингового исследования, принцип которого заключается в сравнении контрольных групп с набором тестовых групп, в которых один или несколько атрибутов изменяются для того, чтобы определить, какие изменения приводят к улучшению целевых показателей.

ГЧП - Государственно-частное партнёрство - взаимодействие государства и бизнеса для решения общественно значимых задач на взаимовыгодных условиях.

2. Описание методологии анализа данных и разделения на этапы

2.1. Методические рекомендации по выделению этапов работы

Для работы над проектами по анализу данных, рекомендуется на начальной фазе проекта выделить следующие этапы:

1. Постановка задачи, как результат анализа собранных потребностей и формулировок разработанных требований.
2. Оценка социально-экономического эффекта от внедрения и сроков окупаемости.
3. Макетирование и прототипирование решения, не рассматривая само решение в отрыве от данных, над которыми оно должно работать.
4. Оценка результатов и, что не менее важно, ресурсов на получение более точной модели решения в случае повторного прохождения этапов №№1, 2..

При этом, помнить, что корректно поставленная задача - это модель предметной области, исходные данные, над которыми она действует, цель постановки и критерии её решения. Цель постановки и критерии решения здесь - ключевой фокус, та как всё остальное в постановке подчиняется им.

Макетирование и прототипирование решения, которое происходит на этапе № 2 после того, как задача сформулирована, позволяет предложить несколько решений, которые могут рассматриваться, как по отдельности, так и совместно в итоговом решении. Переход на третий этап возможен, когда появилась уверенность в том, что предлагаемый макет удовлетворит критериям успешности решения.

Здесь особое место занимает вопрос существующего потребительского опыта: слишком новаторские решения могут быть весьма результативны, но могут встречать серьёзное сопротивление на местах. Поэтому, в критерии результативности при оценке решений на Этапе № 3 необходимо принимать во внимание этот фактор, так как бывают постановки с очень короткими сроками реализации, иначе они будут не актуальны.

Если оценка возможности внедрения после этапа 3 будет положительной, то следующим шагом будет:

1. Планирование и согласование опытной эксплуатации разработанного решения, цель которого не только апробация разработанной модели, но и проверка действий участников внедрения в новых условиях. Такой подход позволяет получать дополнительные гарантии успешности внедрения.
2. Если учёт человеческого фактора приводит к тому, что оценка решения близка к удовлетворительной или хуже, следует рассмотреть возможность дополнительной проработки постановки задачи: снова зайти на Этап № 1 и сформулировать комплект частных технических заданий, относительно которых будет проведёт Этап № 2.

Оценка возможности внедрения должна включать в себя учёт вычислительной мощности, доступной заказчику, так как это может существенным образом повлиять на характере решения, а это неизбежно приведёт к уточнению постановки задачи из-за изменения эксплуатационных условий решения.

Такой подход позволяет выйти на Этап № 4 с хорошо проявленными границами проектов (подпроектов) и утвердить образ технического решения, так как в рамках этапа внедрения решения основной фокус будет смещён в сторону автоматики и развитию потребительского опыта людей (UX). Возникающие “болезни” решения будут не только оттягивать ресурс, но и создавать барьеры для внедрение изменений в создание людей. Поэтому в спектре компетенций четвёртого этапа присутствуют консультанты, специалисты HR и эксперты в UX.

В этой связи, если, например, решение будет охватывать большое количество людей, 4-ый этап целесообразно продублировать: в первую очередь выполнить ввод системы в опытную эксплуатацию и, при успехе, внедрить решение в промышленную. Для оценки качества перехода из опытной в промышленную эксплуатацию рекомендуем пройти сценарий этапа № 3 повторно, рассматривая полученный опыт, как моделирование (Этап 2). Такой подход позволит взглянуть на результат в комплексе.

Именно поэтому необходимо вести журнал проекта на всех этапах его выполнения, чтобы получить возможность объективной оценки результатов и осмысленного выбора шагов и их последовательности, когда субъектами изменений станут большее количество человек.

На каждом из этапов могут появиться условия или возникнуть события, благодаря которым проявят себя новые возможности их следует фиксировать в отдельный журнал (соответствующие практики существуют и в ГОСТах, и в стандартах ISO, и в CRISP-DM), так как они могут стать точками роста. Ввиду того, что проекты, основанные на аналитике данных,

фокусируются вокруг извлечения и преобразования информации, точки роста имеют шансы стать воронками роста, в которых все предыдущие шаги преобразования информации создают кумулятивный эффект, преодолеть который очень сложно. Ярким примером здесь является переход сервисов Яндекс от уже привычной навигации к цифровому сервису срочной аренды автотранспорта. Закрепление этих точек в долгосрочной перспективе формирует, таким образом, фундаментальные конкурентные преимущества.

2.2. Методология CRISP-DM

Данный документ использует нотацию, представленную в международной методологии по исследованию данных Cross-Industry Standard Process for Data Mining (CRISP-DM).

На диаграмме выделены стадии:

1. *“Определение целей и задач”* проекта (Business Understanding) | Понимание целей и задач бизнеса или государства, именуемого в дальнейшем “Заказчик” исследований. Результатом данной стадии является выбор и согласование метрики качества с Заказчиком. Выбор метрик качества должен учитывать балансировку классов, возможность проведения A/B-тестирования и т.п.
2. *“Исследование данных”* (Data Understanding) | Стадия “Исследования данных” включается в себя получение информации про формат и объем доступных датасетов (что может выявить дополнительные требования на инфраструктуру анализа данных), особенности инструментов для выгрузки данных, доступная разметка и ее формат. Доступные данные и инфраструктура для их анализа могут сильно повлиять на выбранные метрики качества стадии “Определения целей и задач” (Business Understanding). Именно поэтому на диаграмме выделены стрелки для перехода между этими стадиями в обе стороны.
3. *“Подготовки данных”* (Data Preparation) | Для обучения моделей машинного обучения данные должны быть представлены в специализированном формате (наиболее часто - в табличном виде, см. csv формат). В дополнение к этому, по опыту взаимодействия с Заказчиками по проектам анализа данных, часто необходима очистка данных: работа с опечатками, неправильным вводом данных, обновление разметки (исправление ошибок или самого формата разметки для удобства обработки на компьютере).
4. *“Моделирование”* (Modeling) | Результатом стадии “Моделирование” является разработка решения, основанного на алгоритмах машинного обучения. Исполнитель производит разработку решения на очищенных наборах данных и оптимизирует метрики качества, зафиксированные на предыдущих стадиях. Разные модели могут требовать

специализированную предобработку для обучения, поэтому диаграмме представлена стрелка от стадии “Моделирования” в стадию “Подготовка данных”.

5. *“Оценка решения”* (Evaluation) | Проведение оценки обобщающей способности выбранной модели на новых данных, а также пилотные запуски в формате A/B-тестирования производятся на стадии “Оценка решения”.
6. *“Внедрение”* (Deployment) | Стадия “Внедрение” используется для введения в продуктив выбранных моделей, которые решают поставленные задачи Заказчика исследований и приносят большую ценность, чем требуют затрат на использование, развитие и поддержку модели / решения.

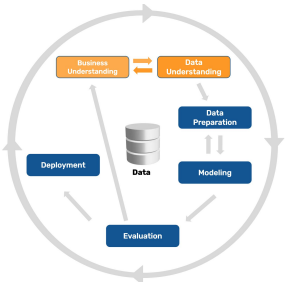
Внутренними стрелками на диаграмме выделены наиболее важные и частые зависимости и переходы между стадиями. Внешний цикл на диаграмме обозначает итеративную процедуру по ведению проекта. На каждом новом цикле происходит уточнение постановки задачи за счет более глубокого понимания данных и результатов экспериментов (моделирования, пилотного тестирования и опыта внедрения).

3. Чек-листы этапов проекта

Этап №1 “Постановка задачи и формирование представлений о доступных данных”

Требования к результату:

- ▢ составлена оценка социально-экономического эффекта проекта (или план её расчета), которая подтверждает целесообразность его запуска;
- ▢ представлено описание данных, процедура и сроки их выгрузки для проведения моделирования (см. Этап №2) и ко всем источникам данных имеется доступ;
- ▢ определены критерии успешного решение задачи;
- ▢ выбраны метрики оценки качества для задач машинного обучения, релевантные поставленной бизнес-задаче;
- ▢ описана процедура валидации или применения алгоритмов машинного обучения.
- ▢ заданы бюджетные и ресурсные рамки;
- ▢ составлена первичная карта рисков и ограничений и способов их преодоления.



Стадии[1]:

- Определение целей и задач
- Исследование данных

Требования к компетенциям:

- отраслевой эксперт (обычно представитель Заказчика или владелец продукта)
- специалист по анализу данных с опытом постановок задач и внедрения разработанных решений
- специалист по проектному управлению
- специалист по информационной безопасности

Описание процессов:

- **Постановка задачи:** отраслевые эксперты совместно со специалистами по анализу данных формулируют постановку задачи и способы её решения на основе доступных данных.
- **Поиск и исследование данных:** отраслевые эксперты совместно со специалистами по анализу данных и информационной безопасности определяют возможные источники данных для решения задачи, далее аналитик данных проводит качественные исследования выгрузок и следом они совместно принимают решения об их использовании.

[1] См. Раздел 3. Описание методологии анализа данных и разделения на этапы

- **Составление плана проекта:** специалист по проектному управлению совместно с отраслевыми экспертами и аналитиками данных определяют ресурсные, технологические и финансовые потребности, составляют перечень работ и сопоставляют их со сроками выполнения.

Требования к инфраструктуре / ПО:

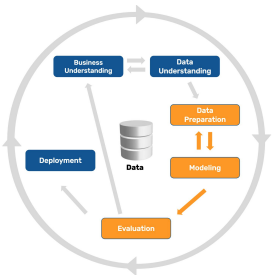
- Инструментарий для анализа и очистки данных.
- Средства обеспечения доступа команды к проектным материалам.
- Средства хранения источников данных и выгрузок из этих источников.
- Средства обеспечения командной коммуникации.
- Средства управления и хранения персональных данных[2] и закрытой информации[3].
- Инструменты для доступа к данным и их эксплуатационные инструкции.

[2] Для проектов, предусматривающих обработку персональных данных.
[3] Для проектов, предусматривающих обработку данных, представляющих собой коммерческую, служебную или государственную тайну.

Этап №2 “Подготовка данных и моделирование”

Требования к результату:

- ▢ результат моделирования соответствует критериям успешного решения задачи и целям Заказчика;
- ▢ получено подтверждение об уверенности Заказчика в том, что предложенная(ые) модель(и) подходит(ят) для решения поставленной задачи;
- ▢ получено подтверждение о возможности внедрения моделируемого решения;
- ▢ принято решение о выпуске модели(ей) на этап оценки.



Стадии:

- Подготовка данных
- Моделирование

Требования к компетенциям:

- специалист по анализу данных с опытом разработки моделей машинного обучения
- специалист по обработке данных с опытом работы со слабо структурированными данными
- отраслевой эксперт (обычно представитель Заказчика или владелец продукта)

Описание процессов:

- **Обработка, очистка и интеграция данных (при необходимости, разработка необходимых инструментов):** специалист по анализу данных ставит задачи по очистке и корректировке данных; специалист по обработке данных выполняет инструментальную обработку и реконструкцию данных; после чего аналитик данных готовит наборы для машинного обучения и отчёты о состоянии и подготовке данных.
- **Определение технологического стека, техник моделирования решения задачи и разработка макета решения:** специалист по анализу данных определяет методы моделирования решения, на основании которых специалист по обработке данных осуществляет выбор инструментов и разрабатывает макет решения.
- **Проведение исследования выбранной модели, результатов её применения и принятие решения о переходе на этап оценки:** на основании постановки задачи отраслевой эксперт принимает у аналитика данных результаты моделирования решения и, сверяя с критериями результативности, принимает решение о возможности внедрения.

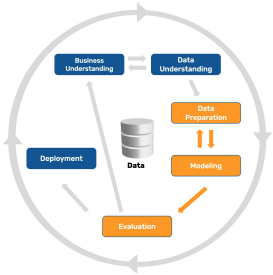
Требования к инфраструктуре / ПО:

- Удобный в использовании инструмент анализа данных с понятной визуализацией (Orange, Weka, Pentaho, …)
- Подходящий (для выбранных типов моделей) программный каркас для машинного обучения (Mahout, TensorFlow, …)
- Подходящие вычислительные мощности и (это очень важно) средства отображения (Монитор от 27’’, разрешением выше 1080p или аналогичное проекционное оборудование)

Этап №3 “Оценка модели и принятие решения о внедрении”

Требования к результату (из позиции заказчика):

- выполнена оценка результатов этапа подготовки данных и моделирования по критериям постановки задачи (см. Этап № 1);
- составлен отчёт о результатах моделирования, включающий в себя ведомость о сделанных выводах и сформулированных гипотезах;
- составлено представление о том, насколько хорошо сделанные выводы и гипотезы соответствуют поставленной задаче;
- принято решение о внедрении решения в административную практику (если проект предусматривает этот этап) с учётом рекомендаций аналитиков данных;
- уточнён и принят бюджет и ресурсы проекта.



Стадии:

- Оценка решения
- Принятие решения:
  - О внедрении
  - Об уточнении постановки задачи

Требования к компетенциям:

- отраслевой эксперт (обычно представитель Заказчика или владелец продукта);
- специалист в анализе данных;
- специалист в проектном управлении;
- специалист в управлении и внедрении изменений;
- специалист по инфраструктуре / DevOps;
- специалист в информационной безопасности.

Описание процессов:

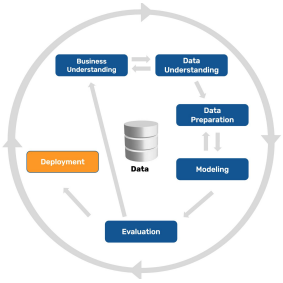
- **Составление отчётов о результатах моделирования:** специалист в анализе данных составляет отчёт о результатах моделирования, куда включает качественные показатели (в соответствии с постановкой), заключение о качестве данных и выявленных в них аномалиях, а также представление о возможных способах преодоления.
- **Формирование мотивированной оценки каждой разработанной модели:** отраслевой эксперт совместно со специалистом в проектном управлении выполняет оценку каждой разработанной модели с позиций:
  - организационной и технической возможности внедрения;
  - результатов сопоставления планируемого эффекта и инвестируемых средств.

- **Контроль информационной безопасности:** специалист по информационной безопасности осуществляет контроль результирующей постановки задачи, плана и ресурсного обеспечения проекта на соответствие принятым политикам.
- **Контроль технологической готовности к внедрению (в случае подготовки решения о внедрении)**: специалист по инфраструктуре осуществляет контроль обеспеченности предлагаемого решения вычислительными ресурсами и инфраструктурой и предоставляет соответствующий отчёт.
- **Осуществляется уточнение границ и принятие бюджета проекта (в случае подготовки решения о внедрении):** специалист в проектном управлении уточняет план, границы и бюджет проекта, включая изменение организационных аспектов и управление персоналом, определяет возможные источники финансирования и прочих ресурсов, после чего отраслевой эксперт принимает решение о его запуске.

Требования к инфраструктуре / ПО:

- Проекционное оборудование.
- Средства конференцсвязи, если над проектом работает распределённая команда.

Этап №4 “Ввод в опытную эксплуатацию”

<table>
<tr>
<td>Требования к результату:
определён компактный набор подразделений, достаточный для опытной проверки качеств решения;
составлен пошаговый план ввода в опытную эксплуатацию и интеграции для каждой модели, успешно прошедшей Этап №3;
определены рамки применимости моделей и выработаны организационно-технические меры поддержки этих рамок;
составлена ведомость потенциальных проблем внедрения их причин и способов разрешения;
назначены ответственные со стороны Заказчика, обеспечивающие интеграцию решения и данных для него (отдельно на вход и выход);
разработаны инструкции по интеграции в действующие административные процессы и адаптации действующей номенклатуры дел;</td>
<td> 

Стадии:
Внедрение
Требования к компетенциям:
специалист в проектном управлении;
отраслевой эксперт (обычно представитель Заказчика или владелец продукта);
специалист по внедрению с опытом разработки UX-решений и внедрению изменений административных процессов;
системный архитектор Заказчика;
системный инженер (DevOps);
инженер программист;
инженер данных (DataOps);
специалист в информационной безопасности.</td>
</tr>
<tr>
<td colspan="2">Описание процессов:
Запуск и наладка разработанной модели, интеграция её с информационными системами Заказчика: системный инженер, совместно со специалистами DataOps, DevOps и разработчиками решения осуществляют подготовку и запуск модели в составе информационной системы заказчика; сервисный специалист и специалист по контролю качества ПО осуществляют тестирование запущенного решения.</td>
</tr>
</table>

- **Планирование изменений административных процессов Заказчика:** специалисты по внедрению совместно с командой разработчиков и специалистов по управлению персоналом осуществляют планирование адаптации действующих процессов и изменения в номенклатуре дел.
- **Составление требований к процессам переходного периода[4]:** системный архитектор Заказчика совместно с отраслевым экспертом и специалстами UX разрабатывают требования к поддержке на период перехода от существующей системы к системе с набором новых качеств.
- **Обеспечение требований информационной безопасности:** специалист по информационной безопасности совместно с системным архитектором, инженерами по операциям и данным, а также аналитиком данных осуществляют диагностику решения на соответствие требованиям политики информационной безопасности и пишут диагностические сценарии.
- **Запуск внедряемой системы в опытную эксплуатацию:** команда внедрения под управлением руководителя проекта и отраслевого эксперта принимают решение о запуске системы в опытную эксплуатацию.
- **Устранение недочётов, оптимизация алгоритмов решения с учётом границ решения[5]:** по результатам проведения опытной эксплуатации команды разработчиков и внедрения осуществляют исправление выявленных ошибок и недочётов.

Требования к инфраструктуре / ПО:

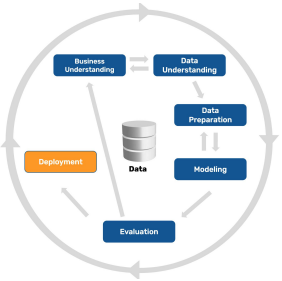
- Средства виртуализации вычислений, в зависимости от выбранного способа: платформы или операционной среды.
- Средства автоматического тестирования программных / аппаратных систем.
- Средства документирования и групповой работы.
- Средства версионного хранения данных: исходных кодов ПО, сценариев автоматизации, документации, чертежей / моделей и так далее.
- Экосистемные средства разработки (в зависимости от выбранного технологического стека), средства автоматизации сборки и использования CI[6].
- Экосистемные средства хранения данных (в зависимости от технологического стека).
- Экосистемные средства машинного обучения (в зависимости от технологического стека).

[4] Когда действуют два вида систем: создаваемая и развиваемая / замещаемая

[5] Улучшение качества на доли процента хорошей, работающей модели, может потребовать в несколько раз большей инвестиций (времени людей, машинного времени) по сравнению с разработкой референсного решения

[6] Анг. аббр. - непрерывная интеграция.

Этап №5 “Внедрение выбранного решения”

<table>
<tr>
<td>Требования к результату:
составлен пошаговый план внедрения и интеграции для каждой модели, успешно прошедшей Этап №3;
назначены ответственные со стороны Заказчика, обеспечивающие интеграцию решения и данных для него (отдельно на вход и выход);
разработан и утверждён план диагностики и обслуживания внедрённого решения;
разработаны и утверждены дополнения к соглашениям об уровне сервиса для каждого участка организационной / функциональной структуры, на которых осуществлена интеграция решения;
проведено обучение персонала Заказчика;
выполнена интеграция новых решений в действующие административные процессы;
выполнена адаптация действующей номенклатуры дел;
разработаны и утверждены регламенты действий в случае возможных простоев или продолжительного регресса, если это актуально для постановки задачи;
составлен финальный отчёт по внедрению, включающий в себя сведения о достигнутых результатах, использованных бюджетах, ресурсах и внешних компетенций.
</td>
<td> 

Стадии:
Внедрение
Требования к компетенциям:
специалист в проектном управлении;
отраслевой эксперт (обычно представитель Заказчика или владелец продукта);
специалист по внедрению с опытом разработки UX-решений и внедрению изменений административных процессов;
специалист HR с опытом в ассессменте персонала и внедренческом консалтинге;
системный архитектор Заказчика;
системный инженер (DevOps);
инженер программист;
инженер данных (DataOps);
инженер по качеству ПО с опытом автоматизации испытаний программного обеспечения;
специалист в информационной безопасности.
</td>
</tr>
<tr>
<td colspan="2">Описание процессов:</td>
</tr>
</table>

- **Масштабирование разработанного решения до рамок системы, обозначенной в постановке задачи:** системный инженер, совместно со специалистами DataOps, DevOps и разработчиками решения осуществляют масштабирование решения в составе информационной системы заказчика; сервисный специалист и специалист по контролю качества ПО осуществляют тестирование запущенного решения.
- **Внедрение изменений административных процессов Заказчика:** специалисты по внедрению совместно с командой разработчиков и специалистов по управлению персоналом осуществляют обучение персонала и адаптацию действующих процессов и изменения в номенклатуре дел.
- **Обеспечение процессов переходного периода (когда действуют два вида систем: создаваемая и развиваемая / замещаемая):** системный архитектор Заказчика совместно с системными инженерами Заказчика обеспечивают техническую поддержку на период перехода от существующей системы к системе с набором новых качеств.
- **Обеспечение отказоустойчивости и самодиагностики внедряемого решения:** системный архитектор Заказчика совместно с системными инженерами Заказчика обеспечивают запуск и наладку испытательных стендов для автоматической диагностики решения, а также совместно с аналитиком данных настраивают систему мониторинга.
- **Обеспечение требований информационной безопасности:** специалист по информационной безопасности совместно с системным архитектором, инженерами по операциям и данным, а также аналитиком данных осуществляют диагностику решения на соответствие требованиям политики информационной безопасности и пишут диагностические сценарии.
- **Составление документации (эксплуатационной, процессной, сервисной):** инженеры по операциям и данным составляют эксплуатационную документацию внедряемого решения, отраслевой эксперт и системный архитектор принимают её.

Требования к инфраструктуре / ПО:

- Средства автоматической диагностики программных / аппаратных систем.
- Средства агрегации и анализа эксплуатационных журналов внедряемых систем.
- Средства виртуализации вычислений, в зависимости от выбранного способа: платформы или операционной среды.
- Средства автоматического тестирования программных / аппаратных систем.
- Средства документирования и групповой работы.
- Средства версионного хранения данных: исходных кодов ПО, сценариев автоматизации, документации, чертежей / моделей и так далее.
- Экосистемные средства разработки (в зависимости от выбранного технологического стека), средства автоматизации сборки и использования CI[7].

[7] Анг. аббр. - непрерывная интеграция.

- Экосистемные средства хранения данных (в зависимости от технологического стека).
- Экосистемные средства машинного обучения (в зависимости от технологического стека).

4. Методические рекомендации

4.1. Этап №1 “Постановка задачи и формирование представлений о доступных данных”

4.1.1. Методические рекомендации по постановке задачи

Как уже отмечалось выше, корректно поставленная задача содержит 4 компоненты:

- Модель предметной области
- Исходные данные
- Цель
- Критерии решения

Поэтому на каждую задачу следует смотреть с учётом этих четырёх аспектов и не двигаться дальше в проект, пока по каждому из них не будет выработано удовлетворительной формулировки.

Для проверки качества постановки задачи можно использовать следующие вопросы:

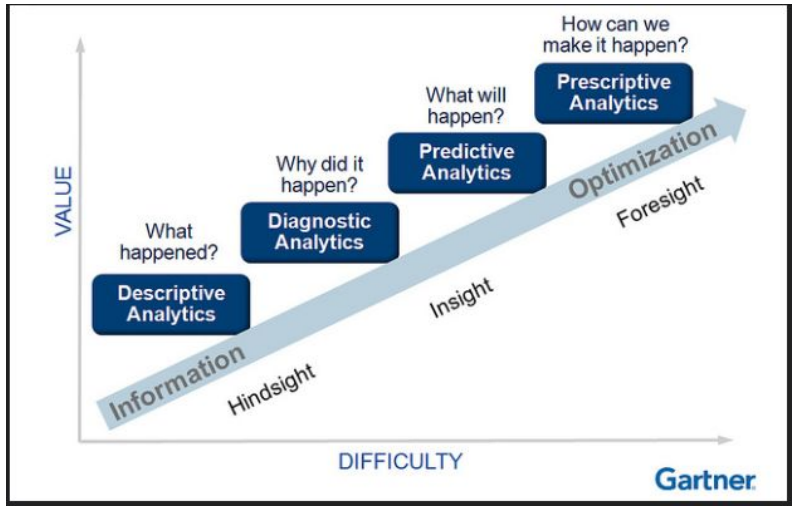
- Каков язык задачи? Все ли его понимают одинаково?
- Почему возникла задача? Зачем она требует решения?
- Из каких подзадач она состоит? Какие задачи предваряют её?
- Какова точная формулировка задачи?
- Каковы критерии приемлемости решения задачи?
- Какие данные требуются для решения задачи?
- Какие ресурсы требуются для её решения?
- Находится ли задача в рамках этических и правовых норм?

При этом, постановка задачи на аналитику на основе данных может относится к одному из следующих типов[8]:

1. Описательная аналитика (Descriptive Analytics), отвечает на вопрос “что произошло?”.

[8] Модель аналитической зрелости по Gartner (Data Analytics Maturity Model)

2. Диагностическая аналитика (Diagnostic Analytics), отвечает на вопрос “почему это произошло?”.
3. Предсказательная аналитика (Predictive Analytics), отвечает на вопрос “что произойдет?”.
4. Предписывающая аналитика (Prescriptive Analytics), отвечает на вопрос “что нужно сделать, чтобы это произошло?”.



Для получения обоснованных ответов 4-го типа (“Как можно добиться желаемого?”), требуется возможность получения обоснованных ответов 3-го типа (“Что произойдёт?”). Таким образом, внедрение систем 4-го типа, не имея внедрённой системы 3-го типа, будет контрпродуктивно. И так далее. Поэтому при формулировании постановки задачи так важно уделять внимание исходным данным - не должно быть сомнений в их полноте, достоверности и актуальности. Поэтому так важно развивать культуру обработки данных, их хранения и защиты.

Поэтому создаваемые в рамках проектов технические решения не могут являются решением задачи. Они являются средствами достижения поставленных целей и инструментарием, культуру использования которыми следует прививать на всех уровнях управления. Поэтому проект по внедрению управления, основанного на данных, - комплексный инновационный проект, вовлекающий, как видно из структуры Этапа 4, большое количество компетенций. Поэтому крупный масштаб решения / внедрения, присущий таким проектам, может сформировать стандарт де-факто на решеня подобного рода.

Следует также искать способы выстраивать кооперацию между государственными, социальными и бизнес-решениями, сохраняя и расширяя возможности доступа к данным и решениям, усиливая эффект от воронок развития. Так, например, государство может брать на себя задачу выравнивания доступа участников рынков к цифровым сервисам массового доступа, на условиях ГЧП или при помощи крупных некоммерческих образований, как это принято в мире.

Основные процессы на этом этапе:

- Определение организационной и функциональной структуры объекта внедрения (в разрезе ключевых интересантов результатов).
- Определение проблемной области в которой будет построено решение на основе данных.
- Определение текущего решения(ий) с описанием его(их) плюсов и минусов.
- Определение проблемы, требующей решения при помощи анализа данных.
- Определение точных критериев успеха и их трассировка к проектным целям.
- Формулирование постановки задачи, определение прочих связанных требований и критериев успеха (настолько подробно, насколько это возможно).

+ групповая работа (данные от Коричина Дениса)+ постановка задачи от Титаева.

Требования к результату:

- получено обоснование, чем конкретно аналитика данных поможет достичь проектных целей;
- существует представление о том, какие техники анализа позволяют получить лучшие результаты;
- определены критерии, благодаря которым будет понятно, что результаты аналитики данных достаточно точны и эффективны;
- существует понимание, каким образом результаты аналитики будут использованы и будет ли их внедрение;
- все выявленные риски и зависимости включены план проекта;

Требования к компетенциям:

- компетенции в формулировании постановок задач;
- компетенции в проектном управлении (ресурсы, коммуникации, структурирование проекта);
- отраслевая экспертиза (обычно представитель Заказчика или владелец продукта);
- компетенции в системном анализе;
- компетенция в информационной безопасности;

- компетенции в анализе данных.

4.1.2. Методические рекомендации по исследованию данных

Для ускорения погружения в постановку задачи для специалиста по анализу данных рекомендуется подготовить описание доступных наборов данных, содержащих нижеследующую информацию. Идеально, для каждого набора данных иметь:

- понятное описание данных, источника их возникновения и даты публикации/регистрации;
- ответственное лицо, которое отвечает за доступность данных и их чистоту;
- инструменты для их выгрузки для последующего анализа;
- наборы ранее созданных правил трансформации с указанием целевых систем;
- допустимые зоны передачи на основании положения об информационной безопасности;
- схемы метаданных и проверки целостности;
- способы получения: потоковое извлечение, пакетная передача, агрегация и т.д.
- наборы полномочий и характер доступа к данным.

Кроме того, необходимо иметь возможность быстро собирать сведения о состоянии данных в наборе:

- Статистика по заполненности полей
- Перечень возможных значений полей
- Корректность их заполнения в соответствии с правилами, действующими для полей
- Сведения о степени целостности данных: анализ потерянных ссылок или ссылки на несуществующие данные.

Состав процессов на этом этапе следующий:

- Определение типов данных, средств их хранения и компетенций доступных для анализа.
- Определение базовых аппаратных и программных требований.
- Определение типов данных, которые будут приобретены, включение этих сведений в бюджет проекта.
- Определение типов ограничений на доступ к источникам данных и способов их устранения.
- Определение рисков и способов их смягчения в случаях возникновения.
- Определение целей аналитики данных: описание технической задачи и формулировка метрик для оценки успешности решения.

- Составление плана проекта: обсуждение задач плана со всеми участниками команды внедрения, определение сроков выполнения задач, определение ресурсов для каждой из задач, определение циклических задач.
- Определение атрибутов данных[9] [10], которые наиболее и наименее полезны для выполнения проекта.
- Определение объёмов обрабатываемых данных, начиная с которых аналитика позволяет делать достаточно точные прогнозы/оценки.
- Составление отчётов над данными: по сбору, по описанию, по обзору, по качеству.

Требования к результату:

- ▢ все источники данных точно определены и доступны;
- ▢ определены ключевые атрибуты в доступных данных;
- ▢ определены ключевые проектные гипотезы;
- ▢ определены объёмы данных и составлена статистика по каждому значащему атрибуту;
- ▢ существует представление о том, какие проблемы качества были выявлены в наборах данных;
- ▢ существует понимание всех шагов подготовки данных;
- ▢ составлен план выполнения проекта.

Требования к компетенциям:

- отраслевая экспертиза (обычно представитель Заказчика или владелец продукта);
- компетенции в анализе и архиитектуре данных;
- компетенции в области хранения и передачи данных / DataOps;
- компетенция в информационной безопасности.

[9] С возможными форматами можно на сайте https://www.infoculture.ru/2018/12/10/datamaps/ и других, им подобным.
[10] Дополнительными источниками могут также проекты, как ресурсы НСУД: https://ru.wikipedia.org/wiki/Национальная_система_управления_данными

4.2. Этап №2 “Подготовка данных и моделирование”

4.2.1. Методические рекомендации по подготовке данных

Подготовка данных к моделированию предусматривают несколько этапов:

1. Очистку и реконструирование данных, если это требуется по результатам их исследования (на предыдущем этапе).
2. Конвертацию и объединение данных, получаемых из разных источников.
3. Составление тестовых выгрузок и наборов данных для обучения моделей.

Требования к результату:

- определены критерии отбора записей/элементов данных в соответствии с постановкой задачи;
- определены атрибуты или характеристики записей/элементов, а качество данных в них удовлетворяют условиям задачи;
- для каждого набора приняты решения об очистке, реконструировании и исправлении данных, принятые решения включены в спецификацию по подготовке данных;
- определены спецификации правил объединения и дополнения данных;

Описание процессов:

- Разработка инструментария отбора, очистки и конструирования данных для целей моделирования решения.
- Очистка и интеграция наборов данных, подготовка тестовых наборов.
- Формулирование гипотез или вопросов к данным для целей моделирования, удовлетворяющих требованиям постановки задачи.

Требования к компетенциям:

- отраслевая экспертиза (обычно представитель Заказчика или владелец продукта);
- компетенции в анализе данных;
- компетенции в ETL-обработке данных (инженерия данных);
- компетенции в программировании;
- компетенция в информационной безопасности;
- компетенции в DevOps.

4.2.1.1. Типы атрибутов данных и требования к ним

4.2.1.1.1. Строковые и текстовые поля

Строковые или текстовые данные, в общем случае, представлены произвольными последовательностями символов, взятых из некоторого заданного алфавита, подчинённого одной из таблиц кодировки: CP-1251, UTF-8, KOI-8 и так далее. Могут использоваться для хранения сериализованных данных, таких, как JSON, XML и прочих.

К ключевым требованиям для строковых типов данных относятся следующие правила:

- Таблица кодировки в рамках одного поля набора данных (или всего набора данных) должна быть одна.
- Если таблица кодировки текстового поля отличается от UTF-8, его значение должно быть сконвертировано.
- Если в качестве формата передачи данных используется формат CSV, значение поля должно помещаться в кавычки.
- Должно быть установлено правило сравнения неопределённых и пустых строковых значений.
- Значения строковых типов не должны содержать коды символов, не принадлежащих кодовой странице поля, за исключением специальных управляющих символов:
    - перевода строки
    - табуляции
    - и прочих (согласно стандарту POSIX: https://ru.wikipedia.org/wiki/Переносимый_набор_символов)

4.2.1.1.2. Числовые поля

Числовой тип данных может быть представлен двумя подтипами: целое или действительное. Могут отличаться разрядностью в битах (8-64) и признаком “плавающая запятая”.

Вместе с тем, стандарт описания данных СУБД предполагает указание количество десятичных разрядов и точность числа (также в разрядах). Это обстоятельство не только позволяет сокращать объёмы хранения данных, но и задаёт допустимую погрешность для вещественных значений.

4.2.1.1.3. Дата и время

Поля типа даты и/или времени в своей основе в рамках СУБД часто используют целочисленные типы данных (обычно хранится значение прошедших миллисекунд с начала заданной временной эпохи).

Вместе с тем, существуют форматы представления даты и времени (в зависимости от задачи), с помощью которых значения могут быть сконвертированы во внутренний вид. Поэтому, как в случае с выбором таблицы кодировки для строковых типов данных, должен быть задан[11] формат даты и времени.

За исключением особых случаев, описываемых отдельно и явно, рекомендуем использовать форматы представления даты в текстовой форме, соответствующие стандарту ISO 8601 (https://ru.wikipedia.org/wiki/ISO_8601).

4.2.1.1.4. Перечисления и классификаторы

Перечисления - это набор строго заданных значений, семантика которых определяется конкретной предметной областью (или областями) и/или поставленной задачей. Значения перечислений, как правило, являются литералами (https://ru.wikipedia.org/wiki/Литерал_(информатика)).

К ключевым требованиям для перечислений относятся следующие правила:

- Следует избегать использование пробельных символов в значениях литералов: либо удалять их, либо использовать символ подчёркивания (“_”).
- В наборе значений перечисления должно быть явно заданный “пустой” или “неопределённый” литерал.
- Должно быть установлено правило сравнения неопределённых и пустых литералов.

Классификатор - общий случай перечисления, представленный двумя полями:

- **Код элемента классификатора**: заданный по правилам, описанным в пункте 4.3.1.5. “Идентификаторы и ключевые поля”
- **Наименование элемента классификатора**: заданное по правилам, описанным в пункте 4.3.1.1. “Строковые и текстовые поля”

[11] Заключено соглашение о формате конвертации данных типа дата/время.

Классификатор может расширяться дополнительными полями, в зависимости от задачи. В этой связи, классификатор - это всегда отдельная таблица, на коды элементов которой ссылаются[12] другие таблицы.

4.2.1.1.5. Идентификаторы и ключевые поля

Идентификаторы или ключевые поля используются для однозначной идентификации объектов данных. Поэтому, в зависимости от сложности решения и объёмов данных, идентификаторы состоят из следующих компонент:

- Область данных: уникальный идентификатор области или секции данных, к которым относится объект/запись
- Пространство имён: уникальный идентификатор, к которому принадлежит запись или объект данных
- Уникальный идентификатор записи: уникальность которого распространяется на пересечение области данных с пространством имён.

В наиболее простых случаях, идентификатор может быть представлен рядом положительных неповторяющихся целых чисел или букв заранее заданного алфавита.

Ключевые поля могут быть следующих типов:

- Первичный: определяет базовый уникальный идентификатор записи в наборе
- Альтернативный: определяет дополнительный уникальный идентификатор записи в наборе, используемый для наложения дополнительных семантических ограничений на данные
- Внешний: определяет область, пространство имён и идентификатор записи во внешнем наборе данных

Для слабоструктурированных данных сказанное выше также справедливо, несмотря на то, что реляционный характер структуры хранения может быть слабо выражен.

4.2.1.1.6. Интервальные значения

В общем случае, интервальные значения - это вычислимые поля данных, расчёт которых позволяет определить интервал значений из определённой области.

Наиболее распространёнными значениями интервального типа являются:

- Градации возрастов: до 18, от 18 до 25, от 25 ло 40, ...

[12] В зависимости от формата передачи или архитектуры хранения данных, ссылка может быть недействительна. В этом случае, следует говорить о повреждении данных типа нарушения ссылочной целостности.

- Градации доходов: до 10 000, от 10 000 до 25 000, ...
- Размеры/расстояния: до 100, от 100 до 1000, ...
- И так далее.

В связи с этим, по возможности, данные должны хранить исходные значения для того, чтобы можно было вводить удобные для использования градации. Тем не менее, в исходных данных могут отсутствовать конкретные значения. В этих случаях, поставщик данных может предоставлять агрегаты, сборка которых основана на правилах разделения данных на интервалы.

В таких случаях рекомендуем использовать классификаторы, с помощью которых, в частности, определить вычисляемое выражение для сравнения с другими полями соответствующих типов.

Например, есть такие таблицы:

- таблица сроков полезного использования инвентаря: в которой на заданную дату рассчитано поле “возраст” в годах
- таблица инвентаря из внешнего источника: в которой поле “возраст” представлено в виде интервалов (до 3 лет, от 3 до 5 лет, более 5 лет)
- классификатор интервалов:
    - До 3 лет: val < 3
    - От 3 до 5 лет: val >= 3 and val < 5
    - Более 5 лет: val >= 5

Такой подход позволит однозначно привести первую таблицу ко второй и выполнить их объединение, пересечение или сопоставление. Вторую таблицу, таким образом, можно тоже привести к первой, для чего в классификатор ввести и определить атрибут “профиль распределения”, который может быть вычислен на основе исходных данных.

Предложенный способ позволяет задавать более сложные интервалы значений, например, состоящие из двух и более промежутков, если это требуется.

Тем не менее, последняя таблица может быть заменена простым перечислением, где вычисление может находиться на принимающей данные стороне.

Альтернативным способом указания простых интервалов может быть введение дополнительного атрибута в данные. Для рассматриваемого выше примера поля могут быть такими:

- age_start: число
- age_end: число

Где интерпретация значений (включает, не включает, ноль, плюс/минус бесконечность) лежит на принимающей данные стороне, а также требуется проверка непротиворечивости значений, например, в случае пересечения интервалов.

4.2.2. Методические рекомендации по машинному обучению и оценки моделей

Описание процессов:

- Формулирование гипотез или вопросов к данным для целей моделирования, удовлетворяющих требованиям постановки задачи.
- Определение технологического стека и техник моделирования решения задачи.
- Моделирование решения: разработка программного кода, выполнение процедур машинного обучения и отладка моделей.
- Определение аппаратных требований для развёртывания среды моделирования базового решения.
- Обеспечение безопасности данных.
- В случае работы с NoSQL-хранилищами данных произвести денормализацию данных.
- Очистка и реконструирование данных:
    - для неполных/отсутствующих данных: очищать, заполнять или исключать записи/атрибуты;
    - для ошибочных данных: исправлять, восстанавливать или исключать записи/атрибуты;
    - при нарушении целостности: конвертировать, замещать или исключать записи/атрибуты;
    - при ошибках в метаданных: идентифицировать и исправить или найти замену;
    - для требуемых записей/атрибутов (согласно требованиям задачи): определить правила нормализации, трансформации и вычисления;

Требования к результату:

- ▢ развёрнуты вычислительные среды для разработки и обкатки моделей решения;
- ▢ выбрана техника(и) моделирования данных и обоснована набором гипотез;
- ▢ определены наборы данных для проверки модели(ей), оценены ошибки измерения и обработки данных[13], выработаны критерии успешности моделирования, включая такой критерий, как точность модели;
- ▢ для каждой построенной модели получены подтверждение о том, возможно ли в результате её вычисления сделать осмысленное заключение, насколько устойчив этот результат, приемлема ли скорость вычислений;
- ▢ составлен отчёт о выявленных аномалиях в данных при моделировании, недочётах в исходных данных;
- ▢ предоставлен отчет о качестве моделировании (включая метрики качества предсказания и скорость обучения), который содержит выявленные препятствия по внедрению разработанных алгоритмов.

[13] Ошибки измерения не включают в себя ошибки в отношении пропущенных значений, а также не включают в себя проблемы недостатка или избытка охвата предметной области.

Требования к компетенциям:

- отраслевой эксперт (обычно представитель Заказчика или владелец продукта);
- компетенции в проектном управлении;
- компетенции в системном анализе;
- компетенции в анализе данных;
- компетенции в программировании;
- компетенции в DataOps;
- компетенции в инфраструктуре и системной инженерии;
- компетенция в информационной безопасности.

4.3. Этап №3 “Оценка модели и принятие решения о внедрении”

4.3.1. Методические рекомендации по оценке решений

Описание процессов:

- Формирование мотивированной оценки каждой разработанной на предыдущем этапе модели.
- Составление отчёта о результатах моделирования.
- Составление отчёта о тупиковых моделях и рекомендации по развитию настоящей методологии, если это требуется.
- Составление рекомендаций об использовании полученных моделей и средств их вычисления.
- Анализ инфраструктурной готовности и технических возможностей для внедрения (если оно предусмотрено проектом).
- Анализ и составление рекомендаций о возможности и способах внедрения (если оно предусмотрено проектом).
- Анализ сопоставления планируемых к достижению эффектов с инвестициями и прочими проектными расходами.
- Анализ источников финансирования и составление рекомендаций по их привлечению.
- Обеспечение безопасности данных.
- Принятие решений по следующим шагам.

Требования к результату:

- убедиться, что результаты моделирования могут быть легко представлены заинтересованным лицам;
- убедиться, что сделанные выводы и гипотезы ранжированы по степени убывания влияния на объект аналитики;
- для составленного списка ошибок моделирования выработаны способы устранения / обхода;
- для моделей, которые не принесли практической пользы, проведён анализ на предмет соответствия процесса их получения настоящей методологии и по составленным к ней поправкам приняты решения о включении в текст на будущее;
- даны рекомендации о следующем этапе проекта: внедрение (Этап № 4) или корректировка постановки (Этап № 1).

Требования к компетенциям:

- отраслевой эксперт (обычно представитель Заказчика или владелец продукта);
- компетенции в системном анализе;
- компетенции во внедрении изменений в бизнес-процессы;
- компетенции в проектном управлении;
- компетенция в информационной безопасности;

- компетенции в вычислительной инфраструктуре;
- компетенции в аналитике данных.

4.4. Этап №4 “Ввод в опытную эксплуатацию”

4.4.1. Методические рекомендации по опытной эксплуатации

Описание процессов:

- Выбор организационных единиц для выполнения опытной эксплуатации решения
- Планирование и организация процессов встраивания опытного решения в текущие административные процессы.
- Составление требований к процессом переходного периода (когда действуют два вида систем: создаваемая и развиваемая / замещаемая).
- Обеспечение требований информационной безопасности.
- Разработка и отладка интеграционного инструментария для внедряемого решения.
- Разработка и отладка методов конвертации данных для использования выбранными моделями.
- Запуск внедряемой системы в опытную эксплуатацию.
- Выполнение этапа опытной эксплуатации системы.
- Обеспечение сбора обратной связи, фиксации дополнений, пожеланий, критики и прочих видов замечаний.
- Составление документации: эксплуатационной, процессной, сервисной.
- Устранение недочётов, оптимизация алгоритмов решения с учётом границ решения[14].

Требования к результату:

- развёрнуты вычислительные среды для выбранных моделей решения;
- разработаны и утверждены требования к вычислительным мощностям и программному обеспечению;
- компоненты вычислительных моделей размещены в рамках отведённых для них вычислительных сред;
- составлена программа и методика испытаний[15] (ПМИ): для пуско-наладочных работ и промышленной эксплуатации решения;
- подключены и настроены средства мониторинга и диагностики рабочих моделей в соответствии с ПМИ;
- подключены и настроены средства интеграции с внешними системами / источниками данных;
- подключены и настроены средства мониторинга и диагностики соединений с внешними истониками данных в соответствии с ПМИ;

[14] Улучшение качества на доли процента хорошей, работающей модели, может потребовать в несколько раз большей инвестиций (времени людей, машинного времени) по сравнению с разработкой референсного решения

[15] Согласно ГОСТ 34 (РД 50-34.698-90).

- выполнена с оценками не ниже “удовлетворительно” программа и методика испытаний по разделу “Пуско-наладочные работы”, составлен и представлен отчёт о качестве запуска вычислительного комплекса.

Требования к компетенциям:

- компетенции в проектном управлении;
- отраслевая экспертиза (обычно представитель Заказчика или владелец продукта);
- эксперт по инфраструктуре Заказчика;
- компетенции в проектировании вычислительной архитектуры;
- компетенции в UX;
- компетенции в проектировании изменений административных процессов;
- компетенции в ассессменте;
- компетенции во внедренческом консалтинге;
- компетенции в управлении персоналом;
- компетенции в организации делопроизводства;
- компетенции в программировании;
- компетенции в инженерии данных;
- компетенции в DevOps;
- компетенции в автоматизации испытаний программного/аппаратного обеспечения;
- компетенции в сервисной инженерии вычислительных систем;
- компетенции в системном администрировании;
- компетенции в организации закупок;
- компетенция в информационной безопасности;
- компетенции в интеграции эксплуатируемых систем.

4.5. Этап №5 “Внедрение выбранного решения”

4.5.1. Методические рекомендации по внедрению решений

Описание процессов:

- Проектирование и реализация плана изменений административных процессов.
- Проектирование и реализация плана изменений номенклатуры дел.
- Планирование и организация процессов по повышению квалификации персонала.
- Обеспечение процессов переходного периода (когда действуют два вида систем: создаваемая и развиваемая / замещаемая).
- Масштабирование разработанного решения до рамок системы, обозначенной в постановке задачи.
- Развёртывание и настройка средств автоматического мониторинга и самодиагностики, интеграция их с элементами автоматических средств управления.
- Обеспечение требований информационной безопасности.
- Обеспечение сбора обратной связи, фиксации дополнений, пожеланий, критики и прочих видов замечаний.
- Составление документации: эксплуатационной, процессной, сервисной.
- Устранение недочётов, оптимизация алгоритмов решения с учётом границ решения[16].
- Запуск промышленной эксплуатации решения.

Требования к результату:

- развёрнуты вычислительные среды для выбранных моделей решения;
- разработаны и утверждены требования к вычислительным мощностям и программному обеспечению;
- компоненты вычислительных моделей размещены в рамках отведённых для них вычислительных сред;
- выполнена программа и методика испытаний[17] (ПМИ): для пуско-наладочных работ и промышленной эксплуатации решения;
- подключены и настроены средства мониторинга и диагностики рабочих моделей в соответствии с ПМИ;
- подключены и настроены средства интеграции с внешними системами / источниками данных;
- подключены и настроены средства мониторинга и диагностики соединений с внешними истониками данных в соответствии с ПМИ;

[16] Улучшение качества на доли процента хорошей, работающей модели, может потребовать в несколько раз большей инвестиций (времени людей, машинного времени) по сравнению с разработкой референсного решения

[17] Согласно ГОСТ 34 (РД 50-34.698-90).

- выполнена с оценками не ниже “хорошо” программа и методика испытаний по разделу “Пуско-наладочные работы”, составлен и представлен отчёт о качестве запуска вычислительного комплекса;
- составлена и передана эксплуатационным службам документация[18] по использованию программно-аппаратного комплекса, созданного в результате выполнения проекта.

Требования к компетенциям:

- компетенции в проектном управлении;
- отраслевая экспертиза (обычно представитель Заказчика или владелец продукта);
- эксперт по инфраструктуре Заказчика;
- компетенции в проектировании вычислительной архитектуры;
- компетенции в UX;
- компетенции в проектировании изменений административных процессов;
- компетенции в ассессменте;
- компетенции во внедренческом консалтинге;
- компетенции в управлении персоналом;
- компетенции в организации делопроизводства;
- компетенции в программировании;
- компетенции в инженерии данных;
- компетенции в DevOps;
- компетенции в автоматизации испытаний программного/аппаратного обеспечения;
- компетенции в сервисной инженерии вычислительных систем;
- компетенции в системном администрировании;
- компетенции в организации закупок;
- компетенция в информационной безопасности;
- компетенции в интеграции эксплуатируемых систем.

4.5.1.1. Пример описания используемых технологий

[18] Согласно стадии “Рабочая документация” ГОСТ 34.

Программное обеспечение	Порядок лицензирования	Сертификация ФСТЭК[19]	Ссылки на документацию	Комментарии
Java 8+	Открытая лицензия	Нет	https://www.java.com/ru/download/help/	
Scala 2.12.8	Открытая лицензия	**Нет**		
Spark 2.4.3	Открытая лицензия	**Нет**		
Библиотека MLib	Открытая лицензия	**Нет**		
Sbt 1.2.8	Открытая лицензия	Нет		Для запуска в консоли
Сервер почтовых служб MS Exchange	Проприетарная лицензия	Нет		Необязательный компонент
Клиент для создания API (например, Postman)				

[19] Для проектов, где такая сертификация требуется.

4.6. Доступный инструментарий

4.6.1. Talend Data Preparation

Инструмент предназначен для решения целого спектра разных задач, связанных с подготовкой, очисткой и обогащением данных, в частности отлично справляется с задачей анонимизации персональных данных.

Обладает широким набором функций визуализации обрабатываемых данных, позволяет записывать сценарии для последующего воспроизведения. Предоставляется на основании коммерческой и свободной лицензии. Существует разница в функционале между типами версий.

4.6.2. Metabase

Простая система класса BI, предназначенная для визуализации и анализа слабоструктурированных данных с открытым исходным кодом. Кроме свободной лицензии, предоставляется лицензия класса enterprise.

Для работы использует так называемую in-memory базу данных H2, которую не рекомендуется использовать для промышленной эксплуатации, для которой можно использовать MySQL (MariaDB) или PostgreSQL.

Методические рекомендации по работе с продуктом:

- https://docs.google.com/document/d/1gSiGvUNxmUHgPs6WjvEtjotV5bhmBkVqIkblNxQOn10

4.6.3. Docker

Средство контейнерной виртуализации вычислений Docker будет полезным для упрощения задачи развёртывания сред прототипирования под различные задачи моделирования данных. В частности, некоторые из приводимых в настоящем параграфе инструментов великолепно упаковываются с его помощью (для некоторых из них существуют официальные образы, поставляемые разработчиками продуктов):

- Metabase
- PostgreSQL
- GitLab

- И многие другие

Существует официальный репозиторий контейнеров для использования в составе ваших решений Docker Hub (https://hub.docker.com/).

4.6.4. Flask

Flask - программный каркас для разработки веб-приложений на языке Python, использующий набор инструментов Werkzeug и Jinja2. Предоставляет только базовые возможности для разработки.

Сайт среды: https://flask.palletsprojects.com/

4.6.5. Jupyter Notebook (JupyterLab)

Интерактивное средство разработки для Python и некоторых других языков, позволяющее создавать интерактивные приложения и комбинировать их с презентационными текстами, диаграммами и данными. Можно воспользоваться как предоставляемым разработчиком среды сервисом, так и развернуть его на своих площадях. Запуск проектов Jupyter Notebook обеспечивает среда JupyterLab, к которому можно подключать все необходимые в работе библиотеки и сервисы.

Сайт среды: https://jupyter.org/

4.6.6. TensorFlow

Программный каркас для решения задач машинного обучения от Google на языке программирования Python, имеющий широкое распространение в мире. Содержит в себе широкий набор инструментов моделирования и визуализации данных. Предназначен для использования профессионалами в области разработки и машинного обучения.

Сайт среды: https://www.tensorflow.org/

4.6.7. Яндекс.Подбор слов

Сервис Яндекс, позволяющий для заданного слова или словосочетания находить релевантные этому словосочетанию запросы к поисковому сервису Яндекс. Сервис позволяет эффективно решать задачи микро-таргетинга в маркетинге, когда существует необходимость более чётко определить образ целевой аудитории.

Сайт сервиса: https://wordstat.yandex.ru/

4.6.8. SQLite

Очень компактный SQL-сервер, используемый, как правило, для внедрения в более крупные прикладные разработки. В частности, может оказать неоценимую помощь, когда существует необходимость развернуть компактную базу данных в рамках мобильного приложения. Часто применяется для хранения журналов приложений или подготовки тестовых данных для последующего обучения модели.

4.6.9. Microsoft Excel / LibreOffice Calc

Широко известные табличные процессоры, предназначенные для хранения, обработки, вычислений и визуализации табличных данных. Могут, за счёт встроенных средств разработки сценариев, использоваться для решения таких задач, как:

- преобразование данных из одного формата в другой: например, из CSV[20] в XLS или во внешнюю базу данных (или наоборот);
- исследование данных: например, выяснение пропусков в данных, максимальных, минимальных значений в колонках и так далее;
- очистка данных: за счёт фильтров и встроенных инструментов автоматизации;
- объединение данных: за счёт встроенных инструментов автоматизации.

Обладают встроенной справкой на русском языке.

Сайты:

[20] Comma Separated Values - система записи строк, в которых значения разделены символом-разделителем (как правило, запятой). Имеет статус запроса на обсуждение (RFC) за номером 4180. Носит статус стандарта де-факто.

- https://products.office.com/
- http://www.libreoffice.org/

4.6.10. RusVectōrēs: семантические модели для русского языка

Сервис, предоставляющий доступ к уже подготовленным моделям русского языка в форматах word2vec, позволяющих ускорить подготовку моделей машинного обучения для обработки текстов на живом русском языке.

Сайт сервиса: https://rusvectores.org/ru/models/

4.6.11. Webscraper.io

Автоматизированный инструмент для сбора web-данных. Справляется с данными динамических форм, способен выгружать информацию в разных форматах.

Сайт сервиса: https://webscraper.io

4.6.12. Церебро Таргет

Сервис для поиска целевых аудиторий VK.

Сайт сервиса: https://церебро.рф

4.6.13. Data Wrangler

Инструментальный набор для очистки и преобразования данных от Стендфордского университета. Функционал близок к инструменту Talend Data Preparation.

Сайт сервиса: http://vis.stanford.edu/wrangler

4.6.14. Trifacta

Программное обеспечение для обработки и подготовки данных к анализу.

Сайт сервиса: https://www.trifacta.com

4.6.15. Open Refine

Инструмент для работы с сырыми данными: преобразование в разные форматы, очистка, расширение.

Сайт сервиса: http://openrefine.org

4.6.16. Infogram

Веб-сервис для создания инфографики, онлайн-карт и интерактивных схем.

Сайт сервиса: https://infogram.com

4.6.17. Tableau

Инструмент для визуального анализа, позволяющий осуществлять динамическую фильтрацию данных, выделять тренды или проводить глубинный когортный анализ.

Сайт сервиса: http://tableau.com

4.6.18. Mapbox

Онлайн-сервис, предназначенный для создания, редактирования и публикации карт.

Сайт сервиса: https://www.mapbox.com

4.6.19. Open Street Map

Некоммерческий веб-картографический проект по созданию силами сообщества участников - пользователей Интернета подробной свободной и бесплатной географической карты мира.

Сайт сервиса: https://www.openstreetmap.org

4.6.20. Data Studio

Инструмент, который даёт маркетологам простые средства для визуализации данных, полученных из разных источников.

Сайт сервиса: https://datastudio.google.com

4.6.21. D3

JavaScript-библиотека для создания статичных и интерактивных визуализаций сложных данных.

Сайт сервиса: https://d3js.org

4.6.22. Gephi

Программное обеспечение с открытым кодом для анализа и визуализации графов.

Сайт сервиса: https://gephi.org

4.6.23. Power BI

Комплексное программное обеспечение бизнес-анализа компании Microsoft, объединяющее несколько программных продуктов.

Сайт сервиса: https://powerbi.microsoft.com

4.6.24. Qlik

Платформа разработки аналитики, построенная на базе ассоциативного движка и библиотек.

Сайт сервиса: https://www.qlik.com

Приложение № 1: Описания наборов данных с примерами

1. Описание набора данных

1.1. Структура описания

Краткое название датасета	**Размерность датасета (объекты × атрибуты)**	**Формат предоставления**	**Описание датасета**

1.2. Пример заполнения

Краткое название датасета	**Размерность датасета (объекты × атрибуты)**	**Формат предоставления данных**	**Описание датасета**
График уборки	$2 \cdot 10^4 \times 5$	csv	график уборки урожая

Себестоимость культуры	678 × 3	csv	Затраты на сбор, обработку и транспортировку культуры
Характеристика склада	45 × 6	xls	Параметры склада с ограничениями по общей емкости и суточной производительностью

2. Детализация набора данных

2.1. Структура описания

[краткое название датасета 1]

Название переменной	Тип переменной	Единица измерения	Описание переменной

[краткое название датасета 2]

Название переменной	Тип переменной	Единица измерения	Описание переменной

...

[краткое название датасета n]

Название переменной	**Тип переменной**	**Единица измерения**	**Описание переменной**

2.2. Пример заполнения

[График уборки]

Название переменной	**Тип переменной**	**Единица измерения**	**Описание переменной**
Идентификатор поля	integer	-	номер поля
Культура	string		собираемая культура
Валовый сбор	float	тонна	Планируемый объем урожая по культуре
Дата начала	date (dd.mm.yyyy)		дата начала сбора культуры на поле

Дата окончания	date (dd.mm.yyyy)		дата окончания сбора культуры на поле

[Себестоимость культуры]

Название переменной	Тип переменной	Единица измерения	Описание переменной
Валовый сбор	float	руб/тонну	затраты, понесенные на поле для сбора и погрузки культуры
Тариф на транспортировку	float	руб/тонну/км	стоимость перевозки от поля до разных складов
Тариф на обработку	float	руб/тонну	стоимость приемки партии и подготовительных работ для приведения культуры в товарный вид

[Характеристики склада]

Название переменной	Тип переменной	Единица измерения	Описание переменной
Название склада	string		Название склада
Принадлежность	boolean		флаг принадлежности склада компании 0 - арендуемый 1 - собственный
Регион	string		Регион размещения склада
Емкость	float	тонна	Общая емкость склада
Дневная мощность	float	тонна/сутки	Максимальный возможный объем приема культуры на склад
Тариф	float	руб/тонну/сутки	Стоимость хранения культуры на складе

Приложение № 2: Примеры наборов данных

1. Проект “Анализ профиля бедности”

1.1. Структура данных

Краткое название датасета	**Размерность датасета** **(объекты × атрибуты)**	**Формат предоставления**	**Описание датасета**
Билайн, Белгородская область	2,6 млн. X 9	csv	Плотность населения Белгородской области, соцдем в разрезе сведений о доходах. Билайн
Билайн, Липецкая область	2,3 млн. X 9	csv	Плотность населения Липецкой области, соцдем в разрезе сведений о доходах. Билайн
Билайн, Пермский край	3,2 млн. X 9	csv	Плотность населения Пермского края, соцдем в разрезе сведений о доходах. Билайн
Билайн, Республика Саха	1,3 млн. X 9	csv	Плотность населения Республики Саха, соцдем в разрезе сведений о доходах. Билайн
Билайн, Республика Татарстан	6,3 млн. X 9	csv	Плотность населения Республики Татарстан, соцдем в разрезе сведений о доходах. Билайн
Билайн, Ростовская область	7,0 млн. X 9	csv	Плотность населения Ростовской области, соцдем в разрезе сведений о доходах. Билайн
Билайн, Новгородская область	0,9 млн. X 9	csv	Плотность населения Новгородской области, соцдем в разрезе сведений о доходах. Билайн

МТС, Белгородская область	1,5 млн. X 12	csv	Плотность населения Белгородской области, соцдем в разрезе сведений о доходах. МТС
МТС, Челябинская область	2,6 млн. X 12	csv	Плотность населения Челябинской области, соцдем в разрезе сведений о доходах. МТС
МТС, Пермский край	4,0 млн. X 12	csv	Плотность населения Пермского края, соцдем в разрезе сведений о доходах. МТС
МТС, Ростовская область	3,0 млн. X 12	csv	Плотность населения Ростовской области, соцдем в разрезе сведений о доходах. МТС
МТС, Республика Саха	1,4 млн. X 12	csv	Плотность населения Республики Саха, соцдем в разрезе сведений о доходах. МТС
МТС, Республика Татарстан	4,5 млн. X 12	csv	Плотность населения Республики Татарстан, соцдем в разрезе сведений о доходах. МТС

1.2. Детализация набора данных

Данные Билайн: Белгородская область, Липецкая область, Пермский край, Республика Саха, Республика Татарстан, Ростовская область, Новгородская область

Название переменной	Тип переменной	Единица измерения	Описание переменной
ccx	double	град.	Восточная долгота
ccy	double	град.	Северная широта
gender	enum	пол	Пол
age	enum	Градация возраста	Градация возраста

rev	enum	Градация дохода	Градация дохода
kids	enum	да/нет	Наличие детей
period	enum	Период	Период времени
region	enum	Наименование региона	Наименование региона
ctn_dst	int	человек	Количество человек

Данные МТС: Белгородская область, Челябинская область, Пермский край, Ростовская область, Республика Саха, Республика Татарстан

Название переменной	Тип переменной	Единица измерения	Описание переменной
square_size	int	метры.	Размер квадрата сетки
period_start	datetime	дата/время	Начало периода
period_end	datetime	дата/время	Конец периода
longitude	double	град.	Восточная долгота
latitude	double	град.	Северная широта
gender	enum	да/нет	Пол
age_start	int	лет	Возраст (от)
age_end	int	лет	Возраст (до)
income_start	int	рублей	Доход (от)
income_end	int	рублей	Доход (до)

kids	enum	да/нет	Наличие детей
cnt_dst	int	человек	Количество человек