**Анализ продаж молочной продукции и оптимизация управления запасами на молочных фермах**

***Цель проекта:***

Изучить особенности анализа данных по продажам молочной продукции и оптимизация складских запасов. Способы формирования интерактивных дашбордов в целях автоматизации процессов на производстве.

***Задачи:***

Изучить литературу, материалы и статьи касающиеся темы анализа продаж, общая эффективность производства, маркетинговые исследования.

Рассмотреть несколько вариантов имеющихся платформ для создания дашбордов, обработки данных и выбрать ПО соответствующее задачи автоматизации.

Произвести анализ эффективности молочных ферм с учетом местоположения, площади земель и поголовья коров.

Сформировать модель продаж и распределения различных молочных продуктов по различным брендам и регионам.

Оптимизировать управление запасами путем отслеживания количества запасов, минимальных пороговых значений и количества повторных заказов.

Разработать прогнозную модель для прогнозирования спроса и стратегий ценообразования.

Разработать дашборд, отражающий результаты анализа данных по продажам и запасам, позволяющий вручную оптимизировать рекомендуемые объемы и ассортимент заказа производителя.

***Инструменты:***

Git, VScode, Power BI (или аналог), Power Query, Python, Excel, draw i.

***Структура дипломного проекта:***

**Оглавление**

**Введение (2-3 стр)**

**Глава 1.** **Основы автоматизации процесса, связанных с аналитикой данных**

1.1. Что такое автоматизация бизнес-процесса

1.2. Какой бизнес нужно автоматизировать

1.3. Какие требования существуют к автоматизации процесса

1.4. Способы сбора и анализа данных

1.5. Что такое дашборд. Как он может повысить эффективность аналитических процессов

**Глава 2. Подготовка к разработке дашборда**

2.1. Сбор требований к разработке дашборда.

2.2. Сбор и обработка данных для анализа бизнес-процесса продажаи молочных товаров.

2.3. Сформировать список отслеживаемых метрик, необходимых для автоматизации, разработки дашборда

2.4. Составить предварительную визуализацию дашборда

2.5. Создать модуль обработки и анализа данных, а также протестировать его на данных компаний по производству молочной продукции

2.6. Изучить влияние условий хранения и сроков годности на качество и доступность молочной продукции

2.7. Произвести анализ предпочтений и покупательского поведения клиентов в зависимости от местоположения и каналов продаж

2.8. Провести маркетинговые исследования и анализ тенденций в молочной отрасли

**Глава 3. Описание процесса создания системы**

3.1. Разработка модуля сбора и обработки данных

3.2. Разработка модуля анализа данных

3.3. Разработка дашборда

3.4. Подведение итогов. Разработка предложений по улучшению процесса продаж и оптимизации производства.

Заключение (~ 4 стр)

Список используемой литературы и Приложения

**Введение**

На рынке прогнозируется уменьшение количества поставщиков молочной продукции, падение их производительности в связи со сложной экономической ситуацией, сезонным снижением цен в летний период и неутешительными для поставщиков прогнозами падения цен на сырье осенью. Небольшие поставщики, вероятно, не смогут обеспечивать поставку сырья на прежних условиях. Как сохранить прежних поставщиков, а в случае с новыми, наладить сотрудничество и предложить оптимальные условия работы и оплаты за сырье?

Чтобы гибко реагировать на изменение покупательского спроса, предприятиям важно отслеживать потребности рынка и периодически менять ассортимент. В сложных экономических условиях это становится как никогда актуальным, поскольку повышается риск того, что выпущенная продукция не будет продаваться, что приведет к затовариванию складов и потере денег — ведь молочная продукция в основном имеет короткий срок годности и, следовательно, необходимо своевременно ее реализовывать.

Проанализировать покупательский спрос и вовремя пересмотреть стратегию продаж поможет аналитика с классификацией по продажам продукции как по наименованиям (видам), так и по регионам поставки и сегментам. Руководителям отделов продаж (сбыта) и менеджерам по продажам в настоящее время важно анализировать потребности покупателей именно своего региона, поскольку на полках магазина, чаще всего и представлена продукция местных производителей.

Таким образом, используя аналитические отчеты, руководители предприятия могут получать необходимую для анализа информацию и своевременно принимать решение о выпуске наиболее востребованной и приносящей прибыль продукции.

На фермах и молокоперерабатывающих предприятиях выпуск продукции осуществляется под заказы клиентов. Как правило, это постоянные клиенты и спрос у них более-менее стабильный, поэтому производству понятно: что производить и в каком количестве. Но, даже у постоянных клиентов могут изменяться планы по количеству и ассортименту закупаемой продукции и эти колебания необходимо отслеживать, чтобы своевременно скорректировать планы производства. Для решения этой задачи начальник производства и сотрудники планового отдела могут использовать инструменты планирования производства и расчета потребностей в материалах с учетом отраслевой специфики. Программа поможет рассчитать, когда и сколько сырья и материалов (в зависимости от их качества) потребуется на выпуск продукции в соответствии с планами производства, а также запланировать выпуск полуфабрикатов.

Немаловажная задача — это обеспечение полной загрузки производственных мощностей. Ведь потери от простоев оборудования для заводов могут быть критичными.

В период, когда предприятия стремятся оптимизировать расходы, становится буквально обязательным не только своевременно отгружать продукцию, но и выстраивать маршруты автомобилей, развозящих продукцию клиентам так, чтобы максимально сократить затраты на доставку.

**Цифровизация всех бизнес-процессов — неизбежное будущее ферм молокоперерабатывающих предприятиях, ферм, а своевременная, корректная аналитика поступающих данных, снижает риски в части принятия неверных действий, шагов оптимизации бизнеса.**

**Цифровая трансформация –** глубокая реорганизация бизнес-процессов с широким применением цифровых инструментов для их исполнения, которая приводит к существенному улучшению их характеристик (сокращение времени выполнения, исчезновению целых групп подпроцессов, сокращению ресурсов, затрачиваемых на выполнение процессов) и/или появлению принципиально новых их качеств и свойств.

Результаты автоматизации и цифровизации в виде внутренних автоматизированных и прошедших реинжиниринг процессов, внедренных автоматизированных систем и надежных данных, во многих случаях служат основой для **цифровой трансформации.**

**Глава 1. Основы автоматизации процесса, связанных с аналитикой данных**

* 1. Что такое автоматизация бизнес-процесса

Помните старый мультфильм про Нехочуху? Ленивый мальчишка попал в страну великого Нехочухи, где большинство действий выполняли роботы. Надо сказать, мультипликаторы изобразили безрадостную картину автоматизации: человек внутри роботизированных процессов распадается как личность и теряет навыки социальной адаптации. Это идёт явно вразрез с любимым нами «Электроником», который как раз проповедует ценность соединения автоматизации и человеческого начала. А вот современный бизнес об автоматизации думает как о вынужденной, как о неоправданной трате, как о единственной волшебной таблетке от всех проблем, и в целом относится к ней с подозрением. А это ошибочный подход. Управленческая концепция BPMN рассматривает бизнес-процессы как важные ресурсы предприятия, и предполагает управление ими как одну из ключевых организационных систем. Ну а автоматизация — это как раз процесс внедрения и эксплуатации программных средств для оптимизации и ускорения протекания бизнес-процессов. Автоматизация упорядочивает, налаживает оперативную работу и управление задачами. Все сотрудники (причём совершенно не важно, удалённые или офисные) работают в едином информационном пространстве АСУ (PM, ERP, CRM…), координируют свои действия, получают и делегируют задачи, видят свои задания внутри проектов и бизнес-процессов, распределяют свою нагрузку во времени. Руководителю больше не нужно вызывать каждого к себе и разжёвывать задачи — достаточно поставить их в системе и проконтролировать выполнение. Вся оперативная работа ведётся в единой системе и любой сотрудник может получить доступ к данным в рамках своих компетенций, а также запросить доступ или выгрузку данных, если в этом есть необходимость. Вне автоматизации всё это представляет собой сплошную стихийную коммуникацию, в которой несложно забыть, ошибиться или просто «незаметно» не выполнить задачу.

Человеческий фактор — это, пожалуй, самое первое, с чем борется автоматизация в компании. Вот примерный минимум корпоративных проблем, с которыми сталкивается малый и средний бизнес: упущенные возможности контакта с клиентами, забытые и нигде не зафиксированные данные, растянутые во времени элементарные задачи, офисная и удалённая прокрастинация, работа «налево» по профилю компании, «свои» приватные клиенты менеджеров, хищение коммерческой информации и проч. Когда в руках, а точнее на серверах бизнеса появляется программное обеспечение для автоматизации бизнес-процессов всех уровней (управляющие, операционные, поддерживающие), большинство этих проблем легко и относительно безболезненно решаются.

**Автоматизация — это не лекарство от всех болезней бизнеса и не волшебная кнопка «Сделать все правильно».**

На основе данных по аудиту ферм специалистами сервиса Dairy Production Analytics партнёрами в разных странах, видим следующую проблематику в данном вопросе:

* Полное или частичное отсутствие в системах автоматизации достоверных данных для контроля производственных процессов
* Отсутствие квалифицированного персонала, который вносит неавтоматизированные данные в ручном режиме в системы автоматизации
* Отсутствие регламентов по внесению этих данных
* Отсутствие единого формата внесения данных, при котором одним оператором могут вноситься одни значения, другим – совершенно другие, но подразумевающие в итоге одинаковую информацию (как пример, один оператор в систему управления стадом вносит заболевание животного «Мастит», другой «Маст», «Маст.», «М» и т.п., по факту подразумевая мастит, но для системы автоматизации это, тем не менее, разные значения показателя)
* Отсутствие анализа этих данных на уровне систем автоматизации (основной аргумент при этом – там недостоверные данные, нам нечего там анализировать!)
* И самое главное – отсутствие понимания, для чего необходимы все эти данные.

В свою очередь, отсутствие достоверных данных порождает отсутствие контроля за ключевыми показателями на ферме. Приведем набор таких показателей:

* Количество поголовья (тёлки, нетели, бычки, коровы первой, второй, третьей и более лактаций, дойное поголовье, фуражное поголовье)
* Объёмы и показатели производства молока (надой на дойную корову, на фуражную корову, в целом по стаду за день, месяц, год, относительные показатели по надою и так далее)
* Показатели по выбытию животных, заболеваниях и их причинах
* Лактационные кривые по разным лактациям и по разным группам дойного поголовья
* Данные по оплодотворяемости коров и тёлок
* Данные по отёлам и изменению поголовья стада после отёлов
* Данные по кормлению, стоимости и структуре кормов
* И так далее…

Но есть и конкретные критические показатели, контроль за которыми особенно важен:

* Ежегодный процент выбраковки (количество выбывших коров, делённое на среднегодовое поголовье фуражных коров)
* Средний день доения
* Средний сервис-период
* Средний межотёльный интервал
* Процент стельности по стаду (количество стельных коров, делённое на количество голов в стаде)
* Средний расход доз семени на одну голову
* Средний выход тёлок
* PR-коэффициент
* Процент выбывших животных из новотельной группы
* Возраст выбывших животных до 60-го дня доения
* Процент животных в группе «Мастит»
* Количество повторных заболеваний маститом
* Процент животных в группе «Карантин»
* Процент хромых животных

И это ещё не полный список ключевых показателей. Очевидно, что без должного уровня автоматизации и качества данных, мы не можем контролировать эти показатели, особенно в рамках ферм с большим (более 500 голов) поголовьем.

Все вышеперечисленные метрики, а также управление складскими запасами, напрямую зависят от службы сбыта готовой продукции. Проводя полный и своевременный анализ продаж (воронки продаж), гарантируем корректное управление производственной системой.

* 1. **Какой бизнес нужно автоматизировать**

Любой бизнес необходимо автоматизировать. У компаний существуют разные метрики, бизнес-процессы, но есть группа целей автоматизации, которая универсальна для большинства организаций.

* **Экономия** — одна из ключевых целей автоматизации. Вложившись в программное обеспечение, компании используют его либо для увеличения производительности труда, либо для сокращения затрат. Таким образом, компания как минимум экономит время и деньги на изменении структуры сделок и продаж, увеличивая при этом оборот и прибыль. Чтобы автоматизация оказалась откровенно убыточной, нужно просто забросить купленную программу и забыть о потраченных деньгах.
* **Упорядочить и ускорить бизнес-процессы** — ключевая задача автоматизации. Сокращение времени на каждую операцию, определение чётких этапов, ответственных и сроков делает процесс эффективным, прозрачным и отлаженным. Чёткое распределение задач между сотрудниками минимизирует человеческий фактор и сокращает связанные с ним риски.
* **Увеличение прозрачности бизнеса в целом** — важная цель. Автоматизация помогает сотрудникам работать с клиентами и друг с другом в единой информационной системе и не конфликтовать «на переделе» клиентов и следовать индивидуальным KPI; клиентам — получать своевременное и качественное обслуживание; руководителю — видеть реальные результаты работы сотрудников. Такая прозрачность не только формирует позиционирование компании, но и ведёт к росту выручки благодаря исключению конфликтов и лояльности клиентов, а также стандартизации процессов (стандартный, фиксированный бизнес-процесс — признак надёжной компании).
  1. **Какие требования существуют к автоматизации процесса**

Прежде всего нужно выбрать правильные процессы для автоматизации:

• Большие процессы — ветвящиеся, сложные, затрагивающие несколько сотрудников;

• Повторяющиеся с каким-либо периодом (запуск акций, отгрузки, стандартные продажи и т.д.);

• Влияющие на все остальные процессы и активности в организации (обучение новых сотрудников, выпуск релиза продукта);

• Требовательные к срокам исполнения (запуск акций и промо, поддержка клиентов, некоторые внутренние процессы управления инфраструктурой и т.д.);

• Рутинные процессы — которые отнимают много времени, часто повторяются и при этом не требуют особо интеллектуальных действий (закупки, отгрузки, формирование пакетов документов и проч.)

Обратимся к теории бережливого производства. Необходимо смотреть на приоритет автоматизации через основные потери. Если изобразить потери на диаграмме, то сможем увидеть, что в любом процессе ценность занимает самую малую часть, все остальное это разного вида потери, которые бывают первого и второго рода (график 1).

Потери 1 рода — это действия, не создающие ценность, но без которых невозможно обойтись. Например транспортировка, оформление документов. Их невозможно удалить из процесса, но необходимо стремиться сокращать.

А вот потери 2 рода — это действия, не создающие ценности вообще, и их можно и нужно исключать из процесса полностью. Например ожидание, запасы, брак и т.д.

Потери можно найти в любом процессе, будь то производство, оказание услуг различного характера или здравоохранение. Для того что бы устранить данные потери, необходимо уметь их распознавать и знать способы борьбы с ними.

В классической теории бережливого производство были выделены 7 видов потерь:

• Перепроизводство

• Ожидание

• Запасы

• Излишняя транспортировка

• Излишнее перемещение людей

• Брак

• Излишняя обработка

В дальнейшем, совершенствуя теорию, исследователи дополнили первоначальный список еще одним видом потерь:

• Неиспользованный человеческий потенциал

* 1. **Способы сбора и анализа данных.**

Есть несколько методов сбора, необходимых для анализа данных. Они перечислены в порядке увеличения стоимости. Выгрузка из учетных систем. Обычно в учетных системах имеются механизмы экспорта данных или существует API для доступа к ним. Поэтому извлечение нужной информации из систем учета чаще всего относительно несложная операция. Получение сведений из косвенных данных. О многих показателях можно судить по косвенным признакам. Например, можно оценить реальное финансовое положение жителей определенного региона. В большинстве случаев имеется несколько товаров, предназначенных для выполнения одной и той же функции, но отличающихся по цене: товары для бедных, среднеобеспеченных и состоятельных. При наличии данных о продажах по регионам, можно проанализировать пропорции, в которых продаются товары для каждой из категорий клиентов: чем больше доля дорогой продукции, тем более состоятельны в среднем жители данного региона. Использование открытых источников. Большое количество данных присутствует в открытых источниках, таких как статистические сборники, отчеты корпораций, опубликованные результаты маркетинговых исследований и прочее. Покупка сведений у соцсетей, мобильных операторов и дата-брокеров. На рынке работает много компаний, которые занимаются сбором и продажей данных. Они предоставляют посредством API систематизированную информацию различного плана: кредитоспособность, клиентские предпочтения, цены на продукцию, геолокация и т.д. Проведение собственных маркетинговых исследований и аналогичных мероприятий по сбору данных. Это может быть достаточно дорогостоящим мероприятием, но, в любом случае, такой вариант сбора данных возможен. Ввод данных «вручную», когда данные вводится по различного рода экспертным оценкам сотрудниками организации. Данный метод достаточно трудоемкий и требует постоянного выделения ресурсов для обеспечения актуальности сведений. Очевидно, что если эксперт указал некоторый фактор как важный, то исключать его неразумно. Мы рискуем провести анализ, ориентируясь на второстепенные малозначащие факторы, и, следовательно, получить модель, которая будет давать плохие и нестабильные результаты. Такая модель не представляет практической ценности.

Собранные данные нужно преобразовать к единому формату. Идеальный случай — загрузка в базу или витрину данных. Но можно использовать и более легковесные форматы, например, Excel или текстовой файл с разделителями. Данные обязательно необходимо стандартизировать, т.е. одна и та же информация везде должна описываться одинаково. Обычно проблемы с единообразным представлением возникают при сборе информации из разнородных источников.

* 1. **Что такое дашборд. Как он может повысить эффективность аналитических процессов**

Термин business intelligence впервые появился в бизнес-литературе 1865 года. Его упомянул Ричард Миллар Девенс в своей книге «Энциклопедия бизнес-анекдотов». Он писал, что банкир Сэр Генри Фюрненс преуспел благодаря своему Business Intelligence — пониманию политики, нестабильности рынков и рыночных взаимоотношений.

Такое определение термина business intelligence ближе к бизнес-анализу, чем к техническому описанию и так как в 1865 году технологии были неразвиты, можно сделать вывод, что истоки термина уходят к бизнес-аналитике.

В 1958 году сотрудник IBM по имени Ганс Питер Луи написал научную статью A Business Intelligence System, полностью посвящённую business intelligence. Её можно считать точкой зарождения BI в его современном понимании, так как в статье речь шла о работе с данными, а сам термин BI определялся как «автоматизированная система, разработанная с целью распределения информации различных направлений деятельности любой индустриальной, научной или государственной организаций».

BI или Business Intelligence, если объяснять простыми словами, то BI - это Excel + Power Point в одном инструменте, в котором технических возможностей больше, чем в самих Excel и Power Point, например в BI можно настроить автопреобразование и автообновление данных, а также делать не статичные презентации как в Power Point, а интерактивные дашборды, в которых можно менять контекст данных в онлайн режиме.

Если определять что такое BI более научно, то BI - это автоматизация аналитических процессов для презентации данных в красивой форме. В данном определении акцент делается на 2 слова: автоматизация и презентация. Автоматизация помогает высвободить время аналитиков от рутинных задач, на которые аналитики тратят до 80% своего времени, а презентационный слой помогает делать интерактивные дашборды, которые в онлайн режиме показывают данные.

Если говорить о миссии BI, какую ценность BI несет для компаний и предприятий, миссия заключается в том, чтобы обеспечить организацию оперативными данными для помощи в принятии управленческих решение. BI делает решения более очевидными, понятными, легкими для объяснения, а также ускоряет их принятие, что крайне важно в условиях быстроменяющейся реальности. Например, без внедренной BI системы вы чаще всего будете решать проблемы по мере их обнаружения, а при помощи BI вы можете либо реагировать до их появления, либо быстрее принимать необходимые действия. Допустим, делаем аналитику прогнозирования амортизации оборудования - при помощи BI возможно настроить отправку уведомления сотруднику, когда необходимо начать заниматься обслуживанием оборудование, чтобы оно не вышло из строя.

Теперь давайте разберемся с тем, какие задачи BI решает, а для каких задач есть более профильные инструменты.

Профильные задачи для BI инструментов:

1. Автоматизация аналитических процессов. Очень часто внутри компаний одна и та же аналитика нужна к одним и тем же встречам. Раньше отчёты для них каждый раз готовил аналитик. Сейчас можно один раз настроить автоматическую аналитику при помощи BI-систем. Кейс: аналитик две недели готовит отчёт к встрече. После внедрения BI его время освобождается. В течение двух недель аналитик забирает данные из одной и той же системы, производит одни и те же преобразования (формат столбцов, расчеты, сводные таблицы) и производит много рутинной работы, BI позволяет 1 раз настроить этот процесс и высвободить время для более креативных задач.

2. Создание автоматических презентаций. Часто на встречах могут показывать презентации в PowerPoint с одинаковым набором графиков, таблиц и показателей. Меняются только выводы, основанные на данных. Создание таких презентаций тоже можно автоматизировать в BI. Кейс: на постоянной встрече показывают одну и туже

3. Alarm-отчёты. Многим руководителям не хочется погружаться в аналитику детально, либо у них нет на это времени. Часто они просят сообщать о чём-либо, когда возникают проблемы. Отчёты для них можно создать, когда известны границы для какого-то значения и подсвечивать, когда показатели выходит за пределы нормы. Кейс: руководителю приходит уведомление о том, что что-то идёт не так. Допустим у нас есть необходимость отслеживать план факт продаж ритейл магазина раз в час и если факт ниже плана на 20%, то система отправляет профильному сотруднику уведомление о том, что что-то идет не так и необходимо идти разбираться. Таким образом компания реагирует на проблему сразу, а не через какое-то время и если проблему можно решить быстро, то количество упущенной прибыли будет минимизировано.

**Для каких задач есть более профильные инструменты?**

1. Ad-hoc аналитика. Команду аналитики могут просить посчитать что-то «прямо сейчас» для конкретной непостоянной цели. Здесь может потребоваться очень узкая аналитика. В классическом варианте эту задачу решает бизнес-аналитики, а не BI-команда. Кейс: если нужна новая, уникальная аналитика на 1 раз, то делать это должен скорее бизнес-аналитик.

2. Дизайн и инфографика. Часто у заказчиков завышенные требования к визуальной части отчётов. Особенно эта проблема обостряется после посещения конференций и презентаций, где рассказывают кейсы. Если данные нужно визуализировать для продающий мероприятий (конференций, выставок) или для сайта, лучше обратиться к дизайнерам. BI — это, в первую очередь, рабочий инструмент для операционной деятельности компании. Он обеспечивает красивые презентации и отчёты, но это не основная его задача. Кейс: если нужен презентационный материал для выставки или конференции, это работа для дизайнера.

3. Хранение данных. Зачастую заказчики не разделяют тех, кто занимается хранением и обработкой данных и тех, кто готовит аналитику. Это особенно актуально, если BI относится к IT-департаменту. В идеале BI начинает работу на этапе создания витрины данных. Кейс: если нужно забирать данные из разных систем, объединять данные и организовывать их хранение, то это работа для Data Engineer.

**Каким организациям нужен BI?**

Общее правило: чем больше данных внутри компании, тем острее потребность в качественном BI. Обычно чем крупнее организация, тем больше потребность в BI, но стоит отдельно отметить, что в стартапах или недавно созданных компаниях (особенно в IT сфере) также потребность в BI может быть высока.

**Глава 2. Подготовка к разработке дашборда**

**2.1. Сбор требований к разработке дашборда.**

Можно выделить два основных типа дашбордов:

**Операционный** дашборд нацелен на предоставление пользователям критически важной информации для оперативного принятия решений. Актуальность информации в таких дашбордах должна стремиться к online. Это наборы ключевых метрик и информации об отклонениях, статусах и т.п. Обычно пользователями таких дашбордов являются менеджеры нижнего звена ответственные за текущую деятельность в организации (управление грузоперевозками, складами, закупками, продажами). Такой дашборд может содержать большое количество элементов, охватывающих все возможные метрики, которые пользователь хочет наблюдать.

**Аналитический** дашборд не несет таких требований ко времени как операционный и призван дать пользователю информацию о трендах и прогнозах, исследование взаимосвязи различных факторов. Обычно их используют аналитики и руководители организации для принятия стратегических решений, основанных на данных. В отличии от операционного дашборда, аналитический не стоит перегружать большим количеством элементов, а сконцентрироваться на основных показателях нужных для принятия решений и тем более лаконичным должен быть дашборд чем выше уровень руководителя, который работает с ним.

После того как определены пользователи дашборда, необходимо определить на какие вопросы пользователя должен ответить дашборд. Для этого требуется изучить логику принятия решений, по которой пользователь ищет ответ на свой вопрос. Должны быть определены метрики, на которые опирается пользователь в решении задачи и их взаимосвязь. Исходя из этого можно начинать сбор данных и формирование витрины, содержащей всю необходимую информацию для ответа на поставленные вопросы. Также следует получить требования по доступу к данным в зависимости от пользователя. Например, это могут быть ограничения по региону или департаменту/отделу пользователь, которого просматривает дашборд.

Ни один проект, связанный с данными, не обходится без стадии Data Profiling (Профайлинг). Что это такое?

Data Profiling (Профайлинг) - это процесс исследования данных, понимания структуры, взаимосвязей, контекста и возможного применения этих данных в дальнейшем анализе.

Необходимо понять значение данных с точки зрения бизнеса и текущих процессов, а также описать как эти данные могут быть использованы в анализе.

Профайлинг данных помогает:

В проектах связанных с исследованием систем-источников данных. В процессе data profiling (профайлинг) могут быть подчеркнуты риски низкого качества данных, что повлияет на дальнейшее использование данных в BI инструментах и Data-продуктах.

При миграции данных из одной системы в другую. В процессе переезда важно понимать структуру, взаимосвязи и бизнес-контекст данных, также важно проверить качество данных и при необходимости повысить его с помощью интеграционных и ETL/ELT процессов.

В процессе построение BI или Data продукта, которым пользуются конечные бизнес-пользователи. Data Profiling (профайлинг) позволит подчеркнуть все риски качества данных, а также увеличит шансы применения данных в нужном ключе.

**Важное правило:** Data Profiling (профайлинг) необходимо планировать как можно раньше (на первых шагах) работы над проектом и продуктом. Это упростит понимание данных, а также даст возможность управлять поставленными требованиями с большей гибкостью.

Data Profiling должен идти сразу после анализа и формулирования бизнес-требований. Вы должны понять, какие данные у вас есть и какие требования с помощью этих данных вы можете покрыть

Важно оценить качество данных до этапа разработки. Может случится так, что качество одного источника данных очень низкое и придется искать другой источник. Это лучше делать в начале, чем потом мигрировать все решение на другой источник

Необходимо проверять данные систематически. Нельзя слепо верить документации или описанной модели данных, они могут быть устаревшими или не совпадать с реальной структурой данных

Правильные вопросы при профайлинге данных:

Какие таблицы у нас есть и какие атрибуты они в себя включают?

Какие атрибуты включены в первичный ключ? Какой тип ключа используется (простой, составной, суррогатный)?

Какой тип историчности и обновления данных используется в таблице?

Как часто обновляются данные в таблице?

Сколько пустых (blank или null) значений содержит каждый атрибут таблицы? Какой процент от общего количество строк содержит пустые значения?

Сколько уникальных значений содержит каждый атрибут таблицы?

Какие самые популярные значения атрибутов? Если числовые значения, то какое минимальное, максимальное и среднее значение?

Какие корреляции есть между значениями атрибутов? Есть ли в этом зависимость от описываемого бизнес-процесса?

Как итогом после успешного профайлинга данных:

* Понимание данных, атрибутов, взаимосвязей
* Понимание бизнес-области применения данных
* Описанное качество данных и возможное улучшение этого качества при помощи дальнейшей обработки данных
* Сокращение процесса разработки путем проведения профайлинга и выявления всех зависимостей перед самой разработкой
* Большее покрытие бизнес-требований

После того как все требования собраны, ннеобходимо подготовить макет дашборда, в виде схемы расположения всех его элементов и используемый тип визуализаций и согласовать его с Заказчиком. Макет может быть подготовлен как в BI инструменте, так и с использованием графических редакторов, таблиц Excel или любым другим удобным способом.

**2.2. Сбор и обработка данных для анализа бизнес-процесса продажи молочных товаров.**

Рост потребления молока и молочных продуктов является в настоящее время общемировой тенденцией. Для обеспечения устойчивого роста производства молока необходимо повысить эффективность деятельности молочных ферм. Это требует, прежде всего, формализации и глубокого анализа протекающих на фермах бизнес-процессов. Производство и служба сбыта напрямую зависят друг от друга. Чем ближе находишься к пересечению двух процессов сбыта и производства, тем больше экономический эффект у компании, гармоничнее работают бизнес-процессы.

Рассмотрим набор данных (dataset) о продаже молочных товаров, комплексный набор данных о молочных фермах, продуктах, продажах и отслеживании запасов на складе. На схеме рис.1. представлена структура (точка) пересечения отдела продаж и производства. Жизненный цикл формирования коммерческого предложения на заявку потребителя.

Набор данных о продажах молочных продуктов сформирован подробный и информативный, касается молочных ферм, молочных продуктов, молокоперерабатывающих предприятий, продаж и управления запасами. Этот набор данных включает в себя широкий спектр информации, включая местоположение фермы, земельную площадь, поголовье коров, размер фермы, даты производства, сведения о продукте, информацию о бренде, количества, цены, срок годности, условия хранения, сроки годности, информацию о продажах, местонахождение клиентов и т. д. каналы продаж, объемы запасов, пороговые значения запасов и объемы повторного заказа.

Функции (Реквизиты):

* Местоположение: Географическое положение молочной фермы.
* Общая площадь земли (акры): общая площадь земли, занимаемая молочной фермой.
* Количество коров: количество коров на молочной ферме.
* Размер фермы: Размер молочной фермы (в кв.км).
* Дата: дата записи данных.
* Идентификатор продукта: уникальный идентификатор каждого молочного продукта.
* Название продукта: Название молочного продукта.
* Бренд: бренд, связанный с молочным продуктом.
* Количество (л/кг): Количество доступного молочного продукта.
* Цена за единицу: Цена за единицу молочного продукта.
* Общая стоимость: общая стоимость доступного количества молочного продукта.
* Срок годности (дни): Срок годности молочного продукта в днях.
* Условия хранения: Рекомендуемые условия хранения молочного продукта.
* Дата производства: дата производства молочного продукта.
* Дата истечения срока годности: дата истечения срока годности молочного продукта.
* Проданное количество (литры/кг): количество проданного молочного продукта.
* Цена за единицу (проданного): Цена за единицу, по которой был продан молочный продукт.
* Прибл. Общий доход (INR): приблизительный общий доход, полученный от продажи молочного продукта.
* Местоположение покупателя: Местоположение покупателя, купившего молочный продукт.
* Канал продаж: Канал, через который продавался молочный продукт (Розничная, Оптовая, Интернет).
* Количество на складе (литры/кг): количество молочного продукта, оставшегося на складе.
* Минимальный порог запаса (литры/кг): Минимальный порог запаса молочного продукта.
* Количество для повторного заказа (литры/кг): Рекомендуемое количество молочного продукта для повторного заказа.

**2.3.** **Сформировать список отслеживаемых метрик, необходимых для автоматизации, разработки дашборда.**

Список литературы:

1. <https://v8.1c.ru/metod/article/molokozavod-kak-obespechit-effektivnoe-proizvodstvo-i-sokhranit-sbyt-v-period-krizisa.htm>?
2. <https://habr.com/ru/companies/regionsoft/articles/520242/>
3. <https://dzen.ru/a/YK5JhhfRcSg4337H?experiment=919795&experiment=919795>

2

Приложение:

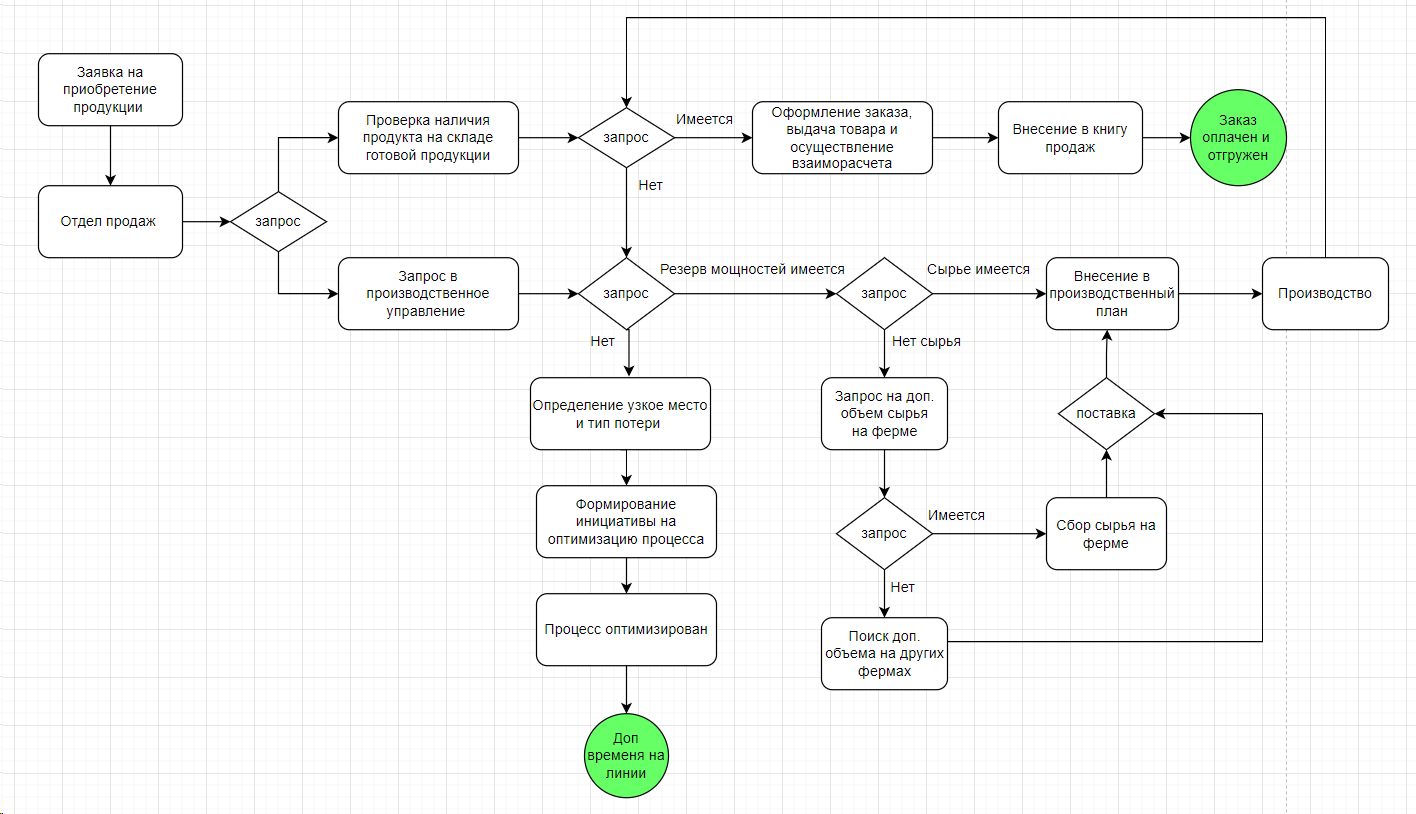


Рис. 1. Жизненный цикл формирования коммерческого предложения на заявку