# Определение требований

## Постановка задачи

Разработать систему автоматизации анализа данных о продажах и построения отчетов с учетом специфики розничной торговли. Основная задача системы – оценка эффективности работы предприятия и помощь в оптимизации продаж.

На основе данных о продажах/транзакциях (количеству и стоимости проданных покупателю товаров), получаемых от систем учета предприятия, POS-систем в точках продаж, либо вручную введенных данных, система должна предоставлять аналитическую информацию, которая может помочь предприятиям в принятии более эффективных решений и повышении уровня продаж.

Система должна поддерживать обработку данных, получаемых из точек продаж, в реальном времени, тем самым предоставляя актуальную информацию, недоступную при ручной обработке данных. Эта информация позволит руководителям предприятий быстрее принимать необходимые решения. Система также должна быть способна анализировать большие объемы информации со многих точек продаж одновременно, избавляя предприятия от необходимости проводить объемные вычисления самостоятельно, а также предоставляя возможность анализировать показатели в глобальном масштабе.

Система должна позволять расчитывать типовые показатели, такие как:

* Общая выручка за месяц;
* Средняя выручка с покупателя;
* Популярные товары;
* Продавцы, реализовавшие больше всего товаров
* Товары, приносящие наибольшую выручку;
* Покупатели, приносящие наибольшую выручку;

Кроме расчета типовых показателей, система должна предоставлять возможность расчитывать сравнительные и агрегированные показатели для двух и более наборов данных:

* Общая выручка с нескольких точек продаж;
* Сравнительная выручка по точкам продаж;
* Изменение уровня продаж;
* Популярные пары товаров;
* Структура продаж;

Необходимо предусмотреть возможность вывода данных в текстовом (табличном) и графическом (в виде графиков) виде.

## Требования к функционалу

* Загрузка и выгрузка данных в формате CSV;
* Генерация отчетов по заданным промежуткам времени;
* Визуализация данных;
* Поддержка аккаунтов пользователей;
* Хранение данных в обезличенном виде;
* Рассчет стоимости использования в зависимости от нагрузки.

### Загрузка и выгрузка данных в формате CSV

Должна быть предусмотрена возможность загружать исходные данные и получать готовые отчеты в формате CSV. Ввод и вывод данных должны быть реализованы как в виде графического интерфейса, так и в виде программного интерфейса (API) для возможности интеграции с другими системами и автоматизации процесса загрущки данных.

### Генерация отчетов по заданным промежуткам времени

Пользователь должен иметь возможность просматривать отчеты за выбранный промежуток времени.

### Визуализация данных

Графический интерфейс системы должен предоставлять возможность визуализации данных в виде графиков и диаграм. Визуализация должна происходить на стороне клиентского приложения, без обращения к серверу.

### Поддержка аккаунтов пользователей

Доступ к системе должен осуществляться только после авторизации пользователя по логину и паролю. Пользователи могут иметь доступ только к отчетам, относящимся к их организации. Пользователи должны иметь возможность изменить пароль. Также необходимо предоставить возможность регистрации новых пользователей для выбранной организации.

Кроме того, пользователь должен иметь возможность управления ключами доступа в программному интерфейсу на случай, если используемых ключ был скомпроментирован.

### Хранение данных в обезличенном виде

Необходимо предусмотреть поддержку обработки данных в обезличенном виде и возможность дальнейшей расшифровки таких данных клиентом. Данные для расшифровки должны храниться на стороне клиентского приложения и не отправляться на сервер.

Пользователь должен иметь возможность загружать список пар ключ-значение в формате CSV, с помощью которых разыменованные наименования товаров, названия точек продаж и имена продавцов и покупателей могли бы быть отображены в исходном виде.

### Рассчет стоимости использования в зависимости от нагрузки

Система должна автоматически измерять объем загруженных пользователем данных и на его основе рассчитывать стоимость использования системы за каждый месяц работы.

## Требования к реализации

* Платформонезависимость;
* Возможность масштабирования системы в зависимости от текущей нагрузки;
* Система не должна требовать от клиента проведения работ по интеграции, покупки оборудования и других предварительных расходов;
* Возможность интеграции с другими системами;
* Клиентское приложение должно работать без установки дополнительного программного обеспечения.

### Платформонезависимость

Система должна работать на любой современной операционной системе, включая Windows Server 2014 и Linux (Debian, Fedora, Arch Linux, Ubuntu Server и др. ), а также на облчаных платформах Microsoft Azure и Amazon Web Services. Также необходимо предусмотреть возможность работы системы на всех популярных веб-серверах (IIS, Apache, Nginx).

### Возможность масштабирования системы в зависимости от текущей нагрузки

Архитектура системы должна быть построена так, чтобы нагрузка могла быть распределена между несколькими серверами. При этом добавление новых серверов должно увеличивать мощность системы линейно и выход из строя любого из серверов не должен сказываться на работоспособности системы в целом. Подключение новых серверов не должно требовать остановки системы. Система должна автоматически подключать и освобождать используемые вычислительные мощности в зависимости от текущей загрузки.

### Система не должна требовать от клиента проведения работ по интеграции, покупки оборудования и других предварительных расходов

Для начала пользования системой не должно требоваться каких-либо работ и предварительных расходов со стороны клиента, за исключением случаев, когда требуется произвести инеграцию системы в существующую инфраструктуру компании.

### Возможность интеграции с другими системами

Система должна иметь возможность интеграции с другими системами при помощи программных интерфейсов. Программный интерфейс должен работать по протоколу HTTP для удаленного доступа к системе, а также поддерживать формат передач данных JSON. Необходимо предусмотреть меры по защите конфиденциальности передаваемых данных и защите от неправомерного доступа к программному интерфейсу. Программный интерфейс системы должен предоставлять все возможности, доступные через графический интерфейс.

### Клиентское приложение должно работать без установки дополнительного программного обеспечения

Пользователь должен иметь доступ к системе с любого устройства, имеющего выход в интернет, без установки какого-либо программного обеспечения. Минимальный требуемый набор операционных систем включает в себя Windows версии XP и старше, Mac OS X любой версии, iOS и Android.

# Архитектура разрабатываемой системы

Разрабатываемую систему было решено разделить на три изолированных друг от друга компонента:

1. Программный интерфейс для ввода данных
2. Модуль построения отчетов
3. Приложение для просмотра отчетов

Преимущества изоляции компонентов системы:

* **Отказоустойчивость**. Выход из строя компонента системы не сказывается на работе остальных компонентов.
* **Распределение нагрузки**. Загруженность одного из компонентов не влияет на скорость работы других частей системы.
* **Масштабируемость.** Компоненты могут быть масштабируемы независимо друг от друга, с учетом текуших потребностей системы.

Полный цикл от получения данных из точек продаж до просмотра готовых отчетов состоит из следующих этапов:

1. Данные из точек продаж отправляются на программный интерфейс.
2. Программный интерфейс размещает данные во временном хранилище.
3. Модуль для построения отчетов загружает данные из временного хранилища для обработки.
4. Готовые отчеты сохраняются в базу данных с отчетами.
5. Приложение для просмотра отчетов запрашивает необходимые отчеты из базы данных.



## Программный интерфейс для ввода данных

Данные со всех точек продаж в автоматическом режиме поступают на программный интерфейс. В задачи программного интерфейса входит:

* Аутентификация пользователя (приложения)
* Балансировка нагрузки между обработчиками
* Поддержка достаточной пропускной способности
* Трансляция полученных данных в необходимый формат для дальнейшей обработки
* Сохранение данных в промежуточное хранилище
* Сохранение статистики использованных ресурсов для биллинга

### Аутентификация пользователя

Все поступающие на программный интерфейс запросы должны проходить аутентификацию. Программный интерфейс должен самостоятельно определять, к какой именно организации относится клиент, отправивший запрос.

### Балансировка нагрузки между обработчиками

Программный интерфейс должен автоматически распределять нагрузку между имеющимеся обработчиками запросов с целью недопущения их перегрузки. Если какой-либо из обработчиков превысил допустимую нагрузку, последующие запросы должны быть распределены между остальными обработчиками до тех пор, пока перегруженный обработчик не закончит обработку данных.

### Поддержка достаточной пропускной способности

Программный интерфейс должен обеспечивать достаточную пропускную способность приема данных. Специфика розничной торговли предполагает неравномерность нагрузки в разное время суток и время года. Число запросов может изменяться на порядки в течении одного дня или в периоды праздников и распродаж. Приоритетной задачей интерфейса является получение данных от клиентов. Эту задачу допустимо решать в ущерб быстроты их обработки.

### Трансляция полученных данных в необходимый формат для дальнейшей обработки

Формат передачи данных между клиентом и программным интерфейсом может отличаться от формата, который требуется модулю построения отчетов. Программный интерфейс должен проводить трансляцию данных в необходимый формат.

### Сохранение данных в промежуточное хранилище

После трансляции, данные для построения отчетов должны быть сохранены в промежуточное хранилище, к которому будет обращаться модуль построения отчетов. Промежуточное хранилище должно отвечать требованиям быстродействия и одновременного доступа большого числа одновременно работающих с ним обработчиков программного интерфейса и модуля построения отчетов. Кроме количественных показателей производительности, это также требует от хранилища данных поддержку идемпотентности операций, то есть одни и те же данные, отправленные в хранилище несколькими обработчиками, не должны быть сохранены более, чем один раз.

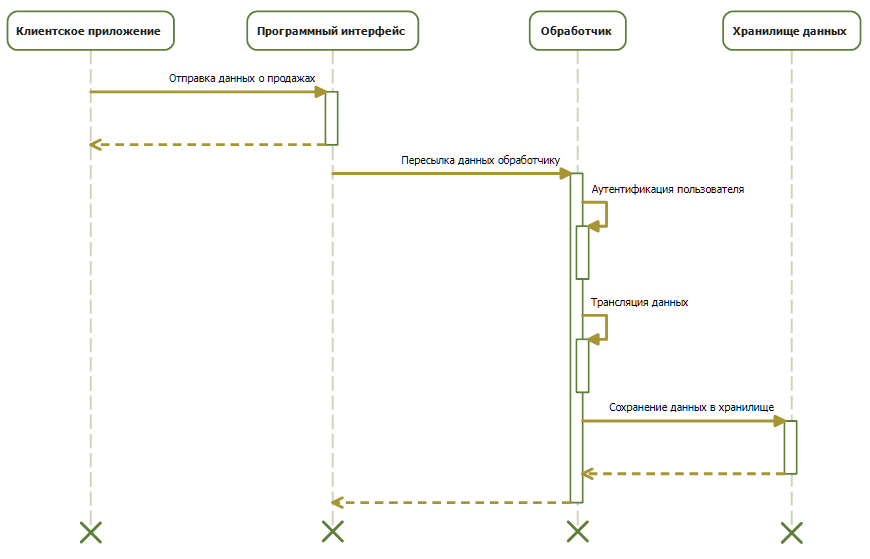
### Сохранение статистики использованных ресурсов для биллинга

Количественные показатели ресурсов, задействованных в обработке запросов, поступающих на программный интерфейс, должны быть сохранены для дальнейшей обработки системой биллинга.

## Взаимодейтсвие с программным интерфейсом

На диаграмме последовательности ниже показан процесс получения данных для обработки программным интерфейсом. Так как наибольший приоритет в работе программного интерфейса отдается пропускной способности, аутентификация пользователя и обработка полученных данных происходят уже после того, как завершается запрос от клиентского приложения. Это позволяет минимизировать время подключения пользователя к интерфейсу, тем самым повысив количество обрабатываемых запросов на единицу времени.

В случае, если аутентификация пользователя не удалась, обработчик завершает работу и данные не отправляются на дальнейшую обработку.



## Модуль построения отчетов

После того, как данные из точек продаж поступили в систему, они отправляются в модуль построения отчетов для обработки. Функции модуля:

* Проверка с заданной периодичностью наличия в хранилище новых данных
* Загрузка всех необходимых для построения отчетов данных
* Обработка полученных данных и рассчет заданных показателей
* Сохранение полученных результатов во временное хранилище
* Загрузка результатов в базу данных готовых отчетов

### Проверка с заданной периодичностью наличия в хранилище новых данных

С заданной периодичностью модуль построения отчетов проверяет, не поступили ли в хранилище новые данные. Если новые данные обнаружены, происходит их загрузка.

### Загрузка всех необходимых для построения отчетов данных

Модуль загружает из временного хранилища все необходимые для построения отчетов за заданный период данные и отправляет их на обработку. При наличии в модуле нескольких обработчиков, нагрузка распределяет между ними.

### Обработка полученных данных и рассчет заданных показателей

Выполняется обработка полученных данных и построение всех необходимых отчетов.

### Сохранение полученных результатов во временное хранилище

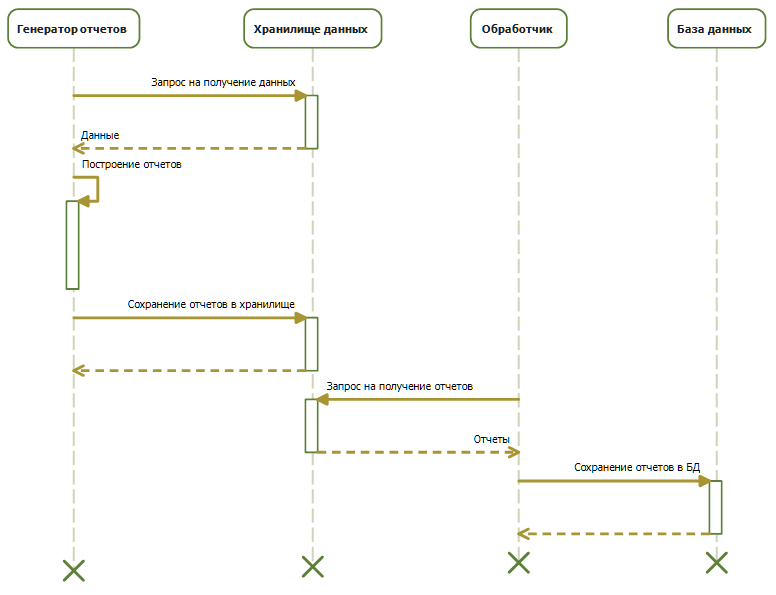
Результаты рассчетов сохраняются обработчиками во временное хранилище.

### Загрузка результатов в базу данных готовых отчетов

После того, как все обработчики закончили проводить рассчеты, полученные данные отправляются из временного хранилища в базу данных, с которой работает приложение для просмотра отчетов.

## Процесс генерации отчетов

На диаграмме последовательностей показан процесс обработки полученных через программный интерфейс данных, их обработка и сохранение результатов во временное хранилище, из которого их позже переносит в базу данных отчетов обработчик.



Данная архитектура позволяет запускать одновременно несколько генераторов отчетов, работающих в параллельном режиме. Использование хранилища данных для временного размещения в нем результатов работы генераторов отчетов позволяет снять лишнюю нагрузку на базу данных и изолировать компоненты системы друг от друга. В случае, если база данных по какой-либо причине оказывается недоступна, обработчик просто отложит задачу переноса отчетов из хранилища в базу на более позднее время, тем самым не прерывая работу генераторов отчетов.

## Приложение для просмотра отчетов

После того, как отчеты попадают в базу данных, они становятся доступны для просмотра через веб-приложение. Основные функции приложения:

* Авторизация пользователей
* Вывод отчетов за выбранный месяц
* Загрузка данных для обработки
* Управление данными для расшифровки

### Авторизация пользователей

Перед тем, как получить доступ к системе, все пользователи должны авторизироваться с помощью логина и пароля. Авторизированному пользователю доступны отчеты только той компании, на которую был зарегистрирован его аккаунт.

### Вывод отчетов

Пользователь имеет возможность вывести отчеты за текущий или любой прошедший месяц и год.

### Загрузка данных для обработки

Кроме загрузки данных напрямую с использованием программного интерфейса, система должна позволять загружать данные вручную. Эта функция реализуется приложением для просмотра отчетов, которое должно поддерживать загрузку CSV-файлов, их трансформацию в необходимый формат и отправку на программный интерфейс для дальнейшей обработки.

### Управление данными для расшифровки

Так как все данные должны храниться в системе в обезличенном виде, в приложении для просмотра отчетов необходимо реализовать функционал, позволяющий расшифровывать данные для удобства просмотра. Вся необходимая для этого информация должна храниться на стороне клиента и не отправляться на сервер.

# Проектное решение

## Требования к аппаратной части

Одним из ключевых требований к системе является платформонезависимость и переносимость программного обеспечения. Так как нельзя однозначно сказать, какие объемы данных должна будет обрабатывать система в долгосрочной перспективе, программное обеспечение должно поддерживать масштабирование. Немаловажной является также является стоимость эксплуатации системы.

Исходя из этих требований, мы получаем следующие критерии выбора аппаратной части:

* Поддержка широкого спектра операционных систем
* Максимальная дешевизна компонентов системы
* Возможность наращивания вычислительных мощностей системы
* Низкая стоимость эксплуатации

### Поддержка широкого спектра операционных систем

Долгий жизненный срок программного обеспечения предполагает смену программных платформ, на которых работает это ПО. Это может быть продиктовано сменой требований организации к набору используемого ПО, наличием в штате сотрудников с необходимыми для обслуживания систем навыками, финансовыми вопросами и т.д.

При выборе аппаратной платформы для системы необходимо учитывать возможность быстрой смены любого программного обеспечения и поддержку как можно более широкого спектра поддерживаемых операционных систем как на текущий момент, так и в долгосрочной перспективе.

### Максимальная дешевизна компонентов системы

Невозможность точно предсказать, насколько финансово успешным будет проект, выдвигает требование к максимальной дешевизне аппаратного решения, достаточного для вывода системы на рынок и последующей эксплуатации.

Аппаратное решение должно быть максимально дешевым. Стоимость замены компонентов платформы, как в случае выхода из строя, так и в случае необходимости повышения производительности системы, должны быть минимальны.

### Возможность наращивания вычислительных мощностей системы

Аппаратное решение должно позволять проводить замену компонентов с целью наращивания вычислительных мощностей системы. Кроме того, решение должно поддерживать возможность объединения нескольких независимых узлов системы в кластеры, тем самым позволяя увеличивать производительность, даже когда будет достигнут предел увеличения производительности отдельных узлов.

## Выбор аппаратного решения

Исходя из заданных требований, наиболее разумным видится использование в качестве аппаратного решения платформы облачных вычислений.

Облачные вычисления - концепция, подразумевающая обеспечение доступа по требованию к пулу вычислительных ресурсов, которые могут быть оперативно предоставлены и освобождены с минимальными затратами.

## Преимущества перед другими решениями

### Отсутствие предварительных расходов

Использование облачной платформы позволяет провести запуск системы без каких-либо начальных расходов. В отличие от использования собственных серверов, при использовании облачных серверов нет необходимости в покупке или аренде оборудования. Оплата облачных вычислительных мощностей производится по факту их использования, а не по предоплате.

Кроме отсутствия предварительных расходов на аппаратную часть системы, в случае использования платного программного обеспечения, облачные решения также не требуют покупки лицензий. Вместо этого оплата лицензионных отчислений включена в стоимость использования облачных ресурсов и также оплачивается почасово, по факту использования. При наличии собственных лицензий на используемое ПО, провайдеры облачных услуг позволяют использовать эти лицензии на облачных серверах, вычитая из стоимости аренды ресурсов стоимость лицензионных отчислений.

### Простота масштабирования системы

Облачные платформы позволяют легко и быстро масштабировать систему, как в сторону увеличения вычислительных мощностей, так и в сторону их уменьшения. Масштабирование может производиться в автоматическом режиме, в зависимости от текущей нагрузки.

Таким образом отпадает необходимость проводить рассчеты и делать прогнозы вычислительных мощностей, которые понадобятся для стабильной работы системы. Возможность масштабирования “вниз” также может снизить расходы в случае, если рассчет необходимых вычислительных мощностей оказался завышенным. В случае покупки или аренды классических серверов, вернуть деньги за уже приобретенные мощности будет невозможно. А в случае заниженного прогноза требуемых мощностей, либо резко возросшей нагрузки на систему, оперативное увеличение доступных мощностей может быть затруднено или вовсе невозможно, тем самым снижая стабильность работы системы.

### Широкая поддержка программного обеспечения

Использование технологий виртуализации позволяет провайдерам облачных услуг предлагать на выбор практически все существующие на сегодняшний день операционные системы.

Кроме того, большинство провайдеров облачных услуг предоставляют предустановленные конфигурации программного обеспечения, заточенные под конкретные задачи, что снимает необходимость проводить настройку ПО самостоятельно.

### Низкая стоимость эксплуатации и обслуживания

Управление и использование облачных мощностей требует гораздо меньших усилий, так как отсутствует необходимость в обслуживании аппаратной части системы. Все работы по обслуживанию, замене и починке оборудования производятся самим провайдером облачных услуг и не требуют наличия в штате соответствующего персонала.

Настройка, масштабирование и обслуживание системы в облачной среде не требует специальных знаний и может проводиться специалистами, имеющими лишь базовые знания в области системного администрирования.

Возможность масштабирования как “вверх”, так и “вниз” позволяет оплачивать только то количество вычислительных мощностей, которое требуется для работы системы в данный момент. Масштабирование системы может выполняться в полностью автоматическом режиме. Это особенно ценно в случае автоматизации работы розничной торговли, когда основная нагрузка системы приходится на светлое время суток, а ночью практически отсутствует. Кроме того, в периоды праздников и распродаж, когда нагрузка на систему резко возрастает, автоматическое масштабирование гарантирует, что система будет доступна и сможет обслуживать клиентов.

## Недостатки перед другими решениями

### Высокая стоимость эксплуатации в долгосрочной перспективе

Как правило, в долгосрочной перспективе стоимость аренды вычислительных мощностей может превышать стоимость покупки и обслуживания собственных аппаратных решений. При использовании платного программного обеспечения, стоимость его аренды через несколько лет выходит дороже покупки лицензий.

Разница в цене между облачными и реальными вычислительными мощностями становится все более заметной по мере роста производительности системы. В этом случае имеет смысл заранее спроектировать систему так, чтобы возможно было осуществить перенос отдельных частей или всей системы целиком на другую аппаратную платформу, без необходимости внесения изменений в программное обеспечение системы.

### Отсутствие доступа к аппаратной части

В некоторых ситуациях, преимущество облачных платформ, заключающееся в отсутствии необходимости обслуживания аппаратной части системы, может быть недостатком. При использовании облачных серверов, выбор используемой аппаратной платформы как правило недоступен или сильно ограничен. В особенности это касается систем ввода-вывода и графических подсистем.

## Выбор платформы облачных услуг

Облачные вычисления предоставляются широким рядом компаний и отличаются по цене и предлагаемым услугам. Модели осблуживания провайдеров облачных услуг можно условно разделить на две категории: инфтраструктура как услуга и платформа как услуга.

### Инфраструктура как услуга

Инфраструктура как услуга (IaaS, Infrastructure-as-a-Service) – модель, при которой пользователю предоставляется возможность самостоятельного управления ресурсами. Пользователь может самостоятельно устанавливать и осблуживать программное обеспечение, включая операционные системы и прикладное ПО. Как правило, пользователю предоставляется ограниченный доступ к настройке аппаратной инфраструктуры – системам хранения данных, сетевым сервисам. Пользователь может развертывать и управлять кластерами виртуальных серверов, организовывать репликацию и резервное копирование данных, устанавливать и настраивать программное обеспечение.

Крупнейшим представителем данной категории является компания Amazon с ее облачной платформой Amazon Web Services (сокращенно AWS). Кроме непосредственно продажи виртуальных серверов, AWS предоставляет доступ ко множеству программных и инфраструктурных решений:

* Amazon Simple Storage Service (Amazon S3) – географически распределенная система хранения данных, имеющая высокую масштабируемость, надежность и скорость работы.
* Amazon CloudFront – система доставки данных (CDN, Content Delivery Network), позволяющая передавать данные конечным пользователям с минимальной задержкой и высокой скоростью, вне зависимости от их географического местоположения.
* Amazon Simple Queue Service (Amazon SQS) – сервис очередей сообщений, позволяющий с высокой надежностью передавать данные между распределенными компонентами системы.
* Amazon Kinesis – сервис буферизации данных, поддерживающий тысячи одновременных подключений, и позволяющий распределить нагрузку на компоненты системы.
* Amazon Elastic Beanstalk – утилита для простого управления кластерами серверов любого размера.
* Amazon Elastic Load Balancing (ELB) – балансировщик нагрузки, позволяющий автоматически добавлять и убирать вычислительные мощности, а также временно отключать от кластера перегруженные или неисправные узлы.
* Amazon Elastic MapReduce (EMR) – кластер серверов Apache Hadoop, интегрированный с другими сервисами платформы.
* AWS Data Pipeline – сервис для выполнения задач по перемещению данных между хранилищами и базами данных.

Как и у большинства других облачных провайдеров, оплата вычислительных и инфраструктурных мощностей производится по факту их использования. Тарификация виртуальных серверов является почасовой и зависит от выбранных аппаратных характеристик. Дополнительно, пользователь оплачивает входящий и исходящий трафик и стоимость хранения данных. В случае, если предполагаемые требования к объему вычислительных мощностей известны заранее, клиент может сделать предоплату по более низкой цене, чем при почасовой аренде.

### Платформа как услуга

Платформа как услуга (PaaS, Platform-as-a-Service) – модель, при которой клиенту предоставляется возможность использования готовой инфраструктуры для размещения программного обеспечения. Контроль и обслуживание инфраструктуры, в том числе сети, серверов, операционных систем, систем хранения данных и других компонентов осуществляется провайдером и как правило недоступен для конфигурации пользователем, за исключением выбора необходимой вычислительной мощности, пропускной способности каналов связи и объема систем хранения данных.

Многие провайдеры, работающие по этой модели, предоставляют доступ к дополнительным сервисам, таким как реляционные и нереляционные базы данных как услуга, серверам с различным предустановленным и неконфигурируемым программным обеспечением, автоматически масштабируемым системам хранения данных, системам анализа данных и другим программным комплексам.

Одним из крупнейших представителей данной категории является облачный сервис Microsoft Azure. Ключевыми компонентами платформы являются:

* Azure Web Sites – сервис для размещения веб-приложений, поддерживающий множество современных языков программирования и фреймворков, включая .NET, Node.js, PHP, Python, Java и другие.
* Azure Virtual Machines – инфраструктурный сервис, схожий по возможностям с IaaS-платформами, предоставляющий виртуальные сервера, работающие на Windows Server 2008/2012 или одном из нескольких дистрибутивов Linux.
* SQL Azure – сервис, основанный на Microsoft SQL Server, обслуживание которого лежит на провайдере, а оплата производится только по потребленным ресурсам.
* Azure CDN – географически распределенная сеть доставки данных.
* HDInsight Hadoop – кластер Apache Hadoop, обслуживанием которого занимается провайдер услуги, а клиенту предоставляется доступ лишь к выполнению прикладных задач.

Исходя из требований к аппаратному решению, наиболее подходящим вариантом является платформа Amazon Web Services. Преимуществами этой платформы для разрабатываемой системы являются:

* Наличие множества готовых сервисов, позволяющих не заниматься созданием и обслуживанием собственных инфраструктурных решений, таких как системы хранения и передачи данных.
* Возможность настройки инфраструктуры исходя из требований и особенностей разработки и эксплуатации системы.
* Доступность широкого рядя программного обеспечения и возможность его настройки под нужды проекта.

## Выбор программного решения

### Язык программирования

Исходя из требования, что клиентское приложение должно работать на компьютерах пользователей без установки какого-либо дополнительного программного обеспечения, а также то, что приложение должно работать как на персональных компьютерах, так и на мобильных устройствах, единственным доступным решением остается разработка веб-приложения.

Существует множество различных платформ и языков программирования, позволяющих разрабатывать веб-приложения. Все они предоставляют примерно одинаковые возможности, поэтому на выбор используемого инструмента в большей степени влияют экономические и организационные факторы, такие как доступность и навыки разработчиков, жизненный цикл и перспективность выбранной технологии, стоимость технической поддержки проекта.

В качестве основы разрабатываемой системы была выбрана платформа Node.js, а в качестве языка программирования – JavaScript. К достоинствам этого выбора можно отнести:

* **Открытость**: Node.js является полностью открытой и бесплатной платформой.
* **Мультиплатформенность**: Node.js работает на всех современных операционных системах, включая Windows, Linux и FreeBSD. Кроме того, платформа поддерживается всеми популярными веб-серверами – Apache, Nginx, IIS, а также может работать и без использования сторонних веб-серверов.
* **Универсальность**: использование языка JavaScript позволяет писать на одном и том же языке как серверную, так и клиентскую часть приложения. Кроме уменьшения числа разработчиков и требований к их навыкам, это также позволяет переиспользовать код между серверной и клиентской частями системы. Кроме того, плафторма Amazon Web Services позволяет использовать Node.js для написания модулей к системе AWS Lambda – вспомогательному сервису, упрощающему интеграцию между различными компонентами AWS.
* **Популярность**: по статистике крупнейшего репозитория открытых проектов GitHub, JavaScript является самым популярным языком разработки на сегодняшний день - около 15% всех проектов написаны на JavaScript. Пакетный менеджер Node.js – NPM является на сегодняшний день крупнейшим ресурсом открытого ПО для разработчиков, и его популярность продолжает расти.

Используемой версией JavaScript был выбран стандарт ECMAScript 6 (ES6). На текущий момент, стандарт ES6 все еще находится на стадии черновика и не был окончательно утвержден (планируемая дата выхода окончательной редакции – июнь 2015 года). Тем не менее, было решено использовать самую свежую версию стандарта с тем, чтобы обеспечить как можно более долгий жизненный цикл разрабатываемой системы. Затраты на доработку кода в случае каких-либо изменений в окончательной редакции стандарта не должны быть существенны, особенно на фоне остальных приемуществ.

Так как далеко не все веб-браузеры имеют поддержку стандарта ES6, исходный код приложения перед публикацей транслируется в наиболее распространенную версию стандарта - ECMAScript 5.

### Операционная система

Как было сказано выше, Node.js работает на всех современных операционных системах, а приложения, написанные на нем, являются платформонезависимыми (при условии, что не используются сторонние модули, работающие не на всех платформах). При таких условиях, при выборе операционной системы на первый план выходят требования к простоте конфигурации и сопровождения, а также надежность и стоимость эксплуатации.

Как будет показано далее, для повышения стабильности работы системы было решено использовать “горизонтальный” вариант масштабирования системы – одновременный запуск нескольких серверов с распределением нагрузки между ними. Использование коммерческих операционных систем при горизонтальном масштабировании может быть экономически затратно, поэтому выбор был сделан в пользу дистрибутива Linux, предоставляемого Amazon. К достоинствам этого варианта можно отнести:

* **Отсутствие необходимости в конфигурации и администрировании системы**: дистрибутив постявляется со всем необходимым предустановленным программным обеспечением и оптимальными настройками для работы Node.js. Дистрибутив активно поддерживается разработчиком, своевременно и в автоматическом режиме получает обновления.
* **Интеграция с другими сервисами**: дистрибутив тесно интегрирован со всеми сервисами AWS, включая ElasticBeanstalk – платформу для автоматического развертывания веб-приложений с поддержкой горизонтального масштабирования.

### Хранилище данных

В качестве хранилища данных было решено использовать сервис Amazon Simple Storage Service (Amazon S3). S3 – веб-служба, предоставляющая возможность для хранения и получения любого объёма данных, в любое время из любой точки сети. На такой выбор повлияли следующие характеристики этого сервиса:

* **Отказоустойчивость и высокая скорость работы**. Сервис поддерживает тысячи одновременных подключений и может сохранять необходимую производительность практически при любых нагрузках на систему.
* **Простая интеграция с другими сервисами**. S3 тесно интегрирован со всеми сервисами платформы, включая EMR, используемый для построения отчетов, и Data Pipeline, ответственный за перенос данных из хранилища в базу данных.
* **Простота администрирования**. Сервис не требует какой-либо работы по администрированию и поддержке.

### База данных

Для хранения готовых отчетов, данных компаний, пользователей и биллинга используется MySQL.

MySQL – реляционная система управления базами данных, разрабатываемая компанией Oracle. Выбор данной базы данных продиктован следующими факторами:

* **Простота администрирования**. База данных работает на основе сервиса Amazon RDS, позволяющего легко масштабировать сервер СУБД, автоматически выполнять резервное копирование данных и проводить обновление программного обеспечения.
* **Дешевизна эксплуатации**. MySQL является открытым проектом и не требует каких-либо финансовых затрат на лицензионные отчисления.
* **Простота интеграции**. Сервис Data Pipeline поддерживает работу с MySQL без необходимости написания собственных модулей для загрузки данных.

### Генератор отчетов

Для обработки полученных данных, подсчета показателей и генерации отчетов используется Apache Hadoop – платформа для выполнения распределенных программ, работающих на кластере серверов. Преимуществами такого подхода является:

* **Высокая масштабируемость**. Платформа может практически неограниченно масштабироваться и обрабатывать любые объемы данных.
* **Простота работы**. В качестве инструмента для обработки данных используется язык запросов, схожий с SQL. Это позволяет начать работу с системой и приступить к созданию отчетов без какой-либо специальной подготовки.

Apache Hadoop разработан в рамках парадигмы MapReduce, согласно которой приложение разделяется на большое количество элементарных заданий, выполняемых на узлах кластера и сводимых в конечный результат. Таким образом, процесс обработки данных состоит из двух этапов:

1. **Шаг Map**: предварительная обработка данных. Главный узел кластера разделяет входные данные на части и распределяет их между остальными узлами. Операции предварительно обработки работают независимо друг от друга и могут производиться параллельно.
2. **Шаг Reduce**: свертка обработанных данных. Главный узел кластера получает ответы от рабочих узлов и на их основе формирует результат вычисления. Аналогично операции map, reduce может производиться параллельно мнодеством узлов кластера.

Для простоты развертывания и управления кластером Hadoop используется сервис Amazon EMR. Сервис использует специальную сборку Hadoop, включающую модули для интеграции с другими сервисами платформы. Кроме всего прочего, это позволяет использовать хранилище S3 в качестве распределенной файловой системы кластера, без необходимости перемещать куда-либо данные, которые были загружены программным интерфейсом системы.

За работу шага Map ответственнен фреймворк Apache Pig, использующий высокоуровневый язык программирования Pig Latin. Использование Pig позволяет проводить предварительную обработку данных без необходимости написания низкоуровневого кода для управления работой кластера.

На шаге Reduce используется фреймфорк Apache Hive, который предоставляет язык запросов HiveSQL, схожий с SQL. Использование этого фреймворка также позволяет абстрагироваться от низкоуровневых деталей операции свертки и сосредоточиться непосредственно на обработке данных.

# Программная реализация

## Программный интерфейс

### Принцип работы

Внешним элементом программного интерфейса, с которым взаимодействует пользовательское приложение в точках продаж является веб-сервис Amazon Kinesis. Сервис служит буфером, позволяющим получать данные одновременно от сотен клиентов и распределять их между обработчиками.

В данной реализации в качестве обработчиков выступает приложение, работающее на сервисе AWS Lambda. Приложение представляет собой модуль Node.js, принимающий на вход данные в формате JSON, полученные программным интерфейсом.

Обработчик вызывается автоматически при поступлении данных в буфер. Каждый обработчик получает за раз не более 10 запросов на загрузку данных. Количество используемых обработчиков автоматически изменяется, чтобы оперативно обрабатывать поступающие данные. После того, как обработчик заканчивает работу, модуль отключается и тарификация используемых ресурсов останавливается, что позволяет не нести расходы за простои.

Тарификация пользователей системы производится по факту использованных вычислительных ресурсов. Для упрощения рассчетов, общий объем вычислений приравнивается к количеству загруженных через программный интерфейс записей и продажах товаров.

### Формат данных

Клиентские приложения взаимодействуют с программным интерфейсом по протоколу HTTP. Все данные передаются в сериализованном виде в формате JSON. Корректный формат входных данных имеет следующий вид:

{

"sales": [

{

"check": "check1",

"customer": "customer1",

"item": "item1",

"location": "location1",

"price": 120.5,

"staff": "staff1",

"time": "2015-04-27T15:19:07.294Z"

},

{

"check": "check1",

"customer": "customer1",

"item": "item2",

"location": "location1",

"price": 200,

"staff": "staff1",

"time": "2015-04-27T15:19:07.294Z"

},

{

"check": "check2",

"item": "item2",

"location": "location2",

"price": 200,

"staff": "staff2",

"time": "2015-04-27T16:20:05.344Z"

},

]

}

|  |  |
| --- | --- |
| sales | Массив записей о приобретенных товарах. |
| check | Уникальный идентификатор покупки. Каждая покупка может включать в себя несколько записей о приобретенных товарах. |
| customer | Уникальный идентификатор покупателя (может отсутствовать). |
| item | Уникальный идентификатор товара. |
| location | Уникальный идентификатор точки продажи. |
| price | Цена реализованного товара (с учетом скидок). |
| staff | Уникальный идентификатор продавца, реализовавшего данный товар (может отсутствовать). |
| time | Время совершения покупки. |

Обработчик проверяет корректность полученных данных и преобразовывает их в формат CSV, который затем сохраняется в хранилище данных S3. Хранилище настроено таким образом, чтобы данные, загруженные более недели назад, автоматически удалялись. Это позволяет не хранить данные, для которых уже были сгенерированы все необходимые отчеты.

Данные в хранилище имеют аналогичную структуру, но вместо формата JSON используется формат CSV, который лучше подходит для обработки генератором отчетов:

1,location1,"2015-04-27T15:19:07.294Z",check1,staff1,customer1,item1,120.5

1,location1,"2015-04-27T15:19:07.294Z",check1,staff1,customer1,item2,200

1,location2,"2015-04-27T16:20:05.344Z",check2,staff2,,item2,200

### Тарификация

После обработки полученных через программный интерфейс данных, обработчик подсчитывает количество записей о приобретенных товарах и отправляет эту информацию вместе с уникальным идентификатором компании, на которую зарегистрирован текущий пользователь, в очередь сообщений для дальнейшей обработки системой биллинга. Формат сообщений имеет следующий вид:

{

"uuid": "aa74268e-a51b-4ed0-bec5-5ffba2f6015a",

"company": 1,

"sales": 3,

"time": "2015-04-27T21:35:16.129Z"

}

|  |  |
| --- | --- |
| uuid | Уникальный идентификатор сообщения. |
| company | Уникальный идентификатор компании. |
| sales | Число полученных записей о продажах товаров. |
| time | Время обращения к программному интерфейсу. |

Система биллинга раз в сутки загружает сообщения, сохраненные в очереди и подсчитывает общее число записей, которое были загружены в систему каждой компанией. Эта информация затем сохраняется в базе данных приложения для просмотра отчетов и становится доступна для просмотра пользователями.

## Модуль генерации отчетов

### Принцип работы

С заданной периодичностью сервис Amazon EMR развертывает кластер виртуальных серверов под управлением Apache Hadoop. В задачи кластера входит:

1. Загрузка данных из хранилища
2. Распределение данных между узлами-обработчками
3. Рассчет показателей и сохранение данных

### Загрузка данных из хранилища

В качестве одного из параметров, кластер получает интервал времени, для которого требуется рассчитать необходимые показатели и сгенерировать отчет. Перед тем, как данные могут быть отправлены узламикластера, они должны быть загружены в распределенную файловую систему (Hadoop Distributed File System, HDFS). Благодаря тому, что данные уже находятся в хранилище S3, которая имеет интеграцию с Hadoop, нам не требуется предварительно перемещать их куда-либо для того чтобы главный узел кластера имел к ним доступ.

Формат CSV, в котором хранятся данные, также поддерживается платформой и не требует предварительной обработки. Для загрузки данных используется следующий Pig-скрипт:

sales =

LOAD '$INPUT/sales' USING CSVExcelStorage(',', 'NO\_MULTILINE', 'UNIX', 'SKIP\_INPUT\_HEADER', 'SKIP\_OUTPUT\_HEADER')

AS (

company: int,

location: chararray,

time: datetime,

check: chararray,

staff: chararray,

customer: chararray,

item: chararray,

price: float

);

### Распределение данных между узлами-обработчками

После предварительной загрузки данных, необходимо выделить показатель, исходя из которого они будут распределятся между узлами кластера. В качестве такого показателя был выбран уникальный идентификатор покупки (чека). Таким образом, все записи, относящиеся к одной покупке, всегда будут находиться только на одном узле кластера.

Для удобства обработки, данные дополнительно разбиваются по трем категориям:

* *checks\_by\_location*: Данные о компании, точки продажи, продавце и покупателе, относящиеся к каждой покупке.
* *checks\_totals*: Общая сумма каждой покупки.
* *items\_by\_check*: Данные о товарах, относящихся к каждой покупке.

Дальнейшее распределение данных между узлами происходит автоматически, под управлением Apache Pig.

### Рассчет показателей и сохранение данных

На этапе вычислений и свертки данных используется фреймворк Apache Hive. Перед тем, как данные, которые были подготовлены на предыдущем этапе, могут быть использованы, необходимо создать виртуальные таблицы Hive, основанные на данных, полученных от Pig:

DROP TABLE IF EXISTS checks\_by\_location;

CREATE EXTERNAL TABLE checks\_by\_location

(

check string, company int, location string, customer string, staff string

)

ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' LOCATION ${INPUT}/checks\_by\_location/';

После этого с данными можно работать так же, как в реляционных СУБД.

Для сохранения результатов вычислений также необходимо создать таблицы необходимого формата. При создании таблиц укажем, что данные должны быть сохранены в формате CSV:

DROP TABLE IF EXISTS report\_item\_pair;

CREATE TABLE report\_item\_pair

(

company int,

location string,

item1 string,

item2 string

)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE LOCATION '${OUTPUT}/report-item-pair';

Для построения запросов используется язык HiveSQL, очень похожий на SQL. К примеру, для рассчета общего числа продаж каждого из товаров можно использовать следующий запрос:

INSERT OVERWRITE TABLE report\_item\_popular

SELECT checks\_by\_location.company, checks\_by\_location.location, items\_by\_check.item, COUNT(items\_by\_check.item)

FROM items\_by\_check

JOIN checks\_by\_location ON items\_by\_check.check = checks\_by\_location.check

GROUP BY checks\_by\_location.company, checks\_by\_location.location, items\_by\_check.item;

В общей сложности, на данном этапе происходит рассчет следующих показателей:

* *report\_customer\_average*: Средняя выручка с каждой покупки.
* *report\_customer\_top*: Рейтинг покупателей, приносящих максимальный доход.
* *report\_item\_pair*: Популярные пары товаров.
* *report\_item\_popular*: Рейтинг популярности товаров.
* *report\_item\_top*: Товары, приносящие наибольший доход.
* *report\_staff\_top*: Рейтинг продавцов по общей выручке.
* *report\_sales\_total*: Общая выручка с разбивкой по точкам продаж.

По мере выполнения вычислений, полученные данные сохраняются в хранилище в формате CSV, откуда они затем будут перемещены в базу данных приложения для просмотра отчетов.

### Перенос отчетов из хранилища в базу данных

После завершения работы кластера и сохранения отчетов, запускается дополнительный обработчик, который считывает результаты вычислений из хранилища и выполняет SQL-скрипты для вставки данных в базу отчетов.

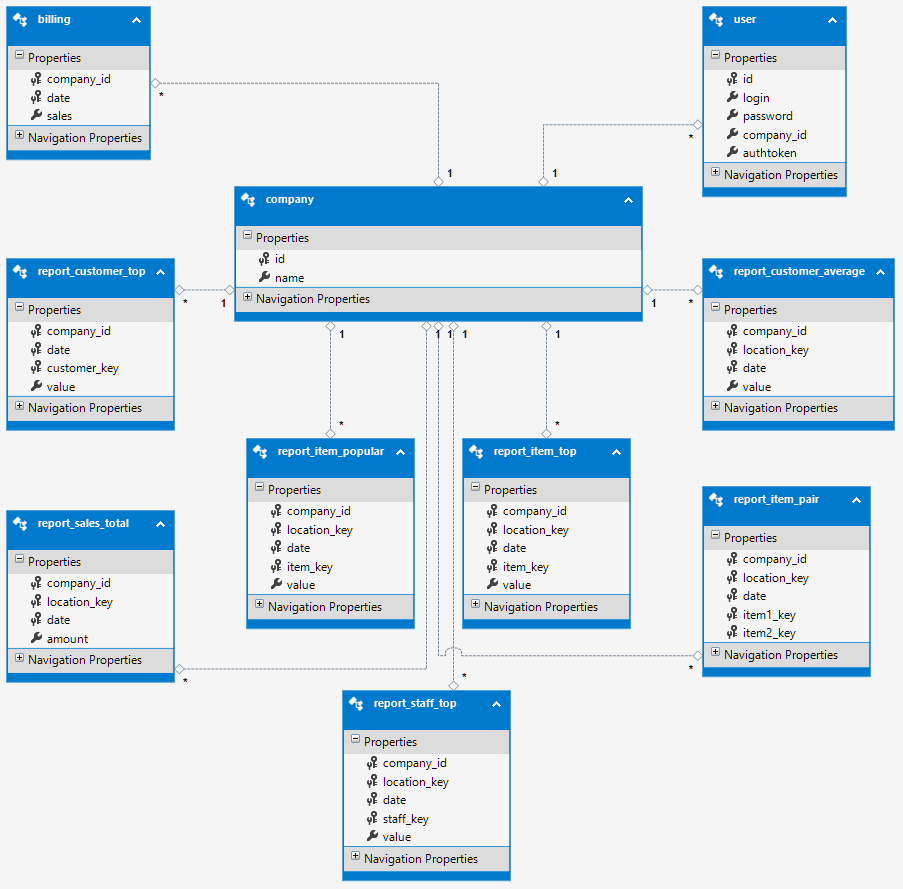
В случае, если база данных недоступна, обработчик откладывает свою работу на 10 минут до тех пор, пока операция вставки не завершится успешно.

При наличии в базе данных записей с ключевым значением, совпадающим с переносимыми данными, вместо вставки происходит обновление данных. Это необходимо для случаев, когда построение отчетов было выполнено дважды для одно и того же набора данных, к примеру при перессчете агрегированных показателей в течение одного дня.

## Приложение для просмотра отчетов

### Структура базы данных

База данных приложения имеет следующую структуру:



### Описание таблиц

Таблица **company** содержит записи о компаниях, зарегистрированных в системе.

|  |  |  |
| --- | --- | --- |
| Поле | Тип | Назвначение |
| id | int | Уникальный идентификатор |
| name | varchar(255) | Название компании |

В таблице **user** хранятся записи о пользователях системы.

|  |  |  |
| --- | --- | --- |
| Поле | Тип | Назвначение |
| id | int | Уникальный идентификатор |
| login | varchar(32) | Логин пользователя |
| password | varachar(255) | Захешированный пароль |
| company\_id | int | ID компании, на которую зарегистрирован пользователь |
| authtoken | varchar(255) | Уникальный ключ для авторизации на программном интерфейсе системы |

Таблица **billing** содержит записи о количестве запросов, обработанных системой для каждой компании, исходя из которых рассчитывается стоимость использования системы.

|  |  |  |
| --- | --- | --- |
| Поле | Тип | Назвначение |
| company\_id | int | ID компании |
| date | date | День |
| sales | int | Число запросов |

Таблица **report\_customer\_top** содержит отчеты о покупателях, приносящих наибольшую выручку.

|  |  |  |
| --- | --- | --- |
| Поле | Тип | Назвначение |
| company\_id | int | ID компании |
| date | datetime | Дата |
| customer\_key | varchar(255) | Уникальный идентификатор покупателя |
| value | decimal(10,4) | Общая выручка |

Таблица **report\_customer\_average** содержит отчеты о средней выручке с каждой покупки.

|  |  |  |
| --- | --- | --- |
| Поле | Тип | Назвначение |
| company\_id | int | ID компании |
| date | datetime | Дата |
| location\_key | varchar(255) | Уникальный идентификатор точки продажи |
| value | decimal(10,4) | Средняя выручка |

Таблица **report\_item\_pair** содержит отчеты о популярных парах товаров.

|  |  |  |
| --- | --- | --- |
| Поле | Тип | Назвначение |
| company\_id | int | ID компании |
| date | datetime | Дата |
| location\_key | varchar(255) | Уникальный идентификатор точки продажи |
| item1\_key | varchar (255) | Первый товар пары |
| item2\_key | varchar (255) | Второй товар пары |

Таблица **report\_item\_popular** содержит отчеты о количестве проданных товаров.

|  |  |  |
| --- | --- | --- |
| Поле | Тип | Назвначение |
| company\_id | int | ID компании |
| date | datetime | Дата |
| location\_key | varchar(255) | Уникальный идентификатор точки продажи |
| item\_key | varchar (255) | Уникальный идентификатор товара |
| value | int | Количество проданных единиц товара |

Таблица **report\_item\_top** содержит отчеты о выручке с каждого проданного товара.

|  |  |  |
| --- | --- | --- |
| Поле | Тип | Назвначение |
| company\_id | int | ID компании |
| date | datetime | Дата |
| location\_key | varchar(255) | Уникальный идентификатор точки продажи |
| item\_key | varchar (255) | Уникальный идентификатор товара |
| value | decimal(10,4) | Общая выручка с товара |

Таблица **report\_sales\_total** содержит отчеты об общей выручке с каждой из точек продаж.

|  |  |  |
| --- | --- | --- |
| Поле | Тип | Назвначение |
| company\_id | int | ID компании |
| date | datetime | Дата |
| location\_key | varchar(255) | Уникальный идентификатор точки продажи |
| amount | decimal(10,4) | Общая выручка с точки продажи |

Таблица **report\_staff\_top** содержит отчеты о выручке, которую принес каждый из продавцов.

|  |  |  |
| --- | --- | --- |
| Поле | Тип | Назвначение |
| company\_id | int | ID компании |
| date | datetime | Дата |
| location\_key | varchar(255) | Уникальный идентификатор точки продажи |
| staff\_key | varchar (255) | Уникальный идентификатор продавца |
| value | decimal(10,4) | Общая выручка |

### Принцип работы приложения

Приложение для просмотра отчетов является веб-приложением, построенным по одностраничному принципу: загрузка с сервера выполняется один раз, после чего все необходимые данные приложение получает без перезагрузки страницы. Это позволяет повысить скорость работы и снизить нагрузку на веб-сервер.

Серверная часть веб-приложения построена на основе фреймворка Express.js и имеет следующую структуру:

* Страница входа в систему. Перед использованием системы все пользователи должны пройти аутентификацию. В случае, если пользователь пытается открыть какую-либо страницу или получить данные через программный интерфейс, не будучи аутентифицированным, он автоматически перенаправляется на страницу входа в систему.
* Страница приложения для просмотра отчетов. Основной раздел сайта, на котором располагается клиентское приложение, с помощью которого пользователь взаимодействует с системой.
* Программный интерфейс. Веб-сервис, через который клиентское приложение получает данные отчетов, биллинга и управляет аккаунтом пользователя.

### Программный интерфейс

Взаимодействие с программным интерфейсом приложения происходит по протоколу HTTP, все данные передаются в формате JSON. Аутентификация пользователя производится с помощью cookie, которую пользователь получает после успешного входа в систему.

Программный интерфейс предоставляет следующие методы:

|  |  |  |
| --- | --- | --- |
| HTTP-метод | URL | Операция |
| GET | /api/profile | Получение данных об аккаунте пользователя. |
| POST | /api/profile/change-password | Запрос на смену пароля пользователя. |
| POST | /api/profile/sign-up | Запрос на регистрацию нового пользователя. |
| POST | /api/profile/create-company | Запрос на регистрацию новой организации. |
| POST | /api/profile/generate-auth-token | Запрос на создание нового ключа доступа к программному интерфейсу. |
| GET | /api/report/customer-average | Отчет о среднем чеке. |
| GET | /api/report/customer-top | Отчет о постоянных покупателях. |
| GET | /api/report/item-pair | Отчет о популярных парах товаров. |
| GET | /api/report/item-popular | Отчет о популярных товарах. |
| GET | /api/report/item-top | Отчет о прибыльных товарах. |
| GET | /api/report/sales-total | Отчет о продажах. |
| GET | /api/report/staff-top | Отчет о лучших продавцах. |
| POST | /api/report | Отправка данных о продажах. |

Пример выполнения запроса на получение отчета о покупателях, которые принесли наибольшую выручку с 8 по 9 мая 2015 года:

GET /api/report/customer-top?date=2015-05-08&days=2

[

{

"customer\_key": "customer1",

"date": "2015-05-08",

"value": 1764

},

{

"customer\_key": "customer9",

"date": "2015-05-08",

"value": 2866.25

},

{

"customer\_key": "customer8",

"date": "2015-05-08",

"value": 1085.75

},

{

"customer\_key": "customer5",

"date": "2015-05-09",

"value": 1479.5

},

{

"customer\_key": "customer6",

"date": "2015-05-09",

"value": 3308

},

{

"customer\_key": "customer7",

"date": "2015-05-09",

"value": 3845.5

}

]

|  |  |
| --- | --- |
| Поле | Назначение |
| customer\_key | Уникальный идентификатор покупателя. |
| date | День, к которому относится текущая запись. |
| value | Выручка с текущего покупателя за день. |

Наименования товаров, названия точек продаж и имена покупателей и продавцов хранятся в базе данных в обезличенном виде. Получив от программного интерфейса данные отчета, клиентское приложение самостоятельно сопоставляет идентификатор разыменованной записи и отображает ее пользователю в исходном виде. Таким образом сохраняется конфиденциальность данных. Схема работы этого алгоритма показана ниже:



### Клиентское приложение

ыы