Бюджетные расходы по уровням и месячным периодам

Дмитрий Донецков (ddonetskov@gmail.com)

18 июня 2019 г.

Аннотация

Описывается подход к подсчёту суммарной стоимости госконтрактов по №44-ФЗ[1] по уровням бюджета и периодам (месяцам).

Содержание

1. Цель

Необходимо подсчитать сумму госконтрактов по Федеральному закону №44-ФЗ в разрезах по уровням бюджета и месячным периодам.

2. Подход

Подход заключается в получение необходимых первичных данных по госконтрактам, их нормализации при необходимости, и последующей агрегации по заданным измерениям при помощи доступных средств.

Данные по госконтрактам, подпадающим под действие Федерального закона №44-ФЗ, размещаются в ЕИС в сфере закупок (далее - ЕИС)[2]. Существует механизм выгрузок информации по госконтрактам из ЕИС, который на ежедневной основе размещает данные в XML-формате на публичном FTP-сервере ftp://free:free@ftp.zakupki.gov.ru в каталоге FCS_regions. Разработчики ЕИС предоставляют документацию, которая описывает форматы данных. Это позволяет сторонним лицам работать с вышеуказанными выгрузками. Однако, самостоятельная работа с выгрузками может требовать значительных ресурсов, т.к. форматы данных меняются, необходимо работать с данными в иерархической структуре, уметь объединять нужные атрибуты в однотипные наборы, соединять разнотипные наборы и т.п.

[1] http://www.consultant.ru/document/cons_doc_LAW_144624/
[2] http://zakupki.gov.ru/

Для получения данных может использоваться и ИС «ГосЗатраты»[3], которая содержит данные из ЕИС. "Госзатраты"обладает публичным API. Документация на API доступна по ссылке https://github.com/clearspending/clearspending-examples/wiki). Данное API позволяет получать информацию по госконтрактам, хорошо подходит для "точечной"работы с ними, когда детализированы условия поиска, т.к. количество возвращаемых записей ограничено системой 500 записями. Для больших выборок требуется обращение к ИТ-специалистам, управляющей данной системой.

Изучение [1] и API «ГосЗатраты» указывает, что контракты могут иметь различия, существенное с точки зрения цели: разниться по количеству стадий исполнений, платежи по ним могут проводиться в несколько этапов, могут возникать неустойки, отнесённые как на заказчика, так и на поставщика, контракты могут отменяться. Вся эта информация представлена четырьмя типами, которые описывают различные аспекты жизненного цикла контракта и несут следующую информацию:

- **contract**, **contract2015** - о заключенном контракте (его изменении),
- **contractProcedure**, **contractProcedure2015** - об исполнении (о расторжении) контракта,
- **contractCancel**, **contractCancel2015** - об аннулировании контракта,
- **contractProcedureCancel**, **contractProcedureCancel2015** - об отмене исполнения (расторжения) контракта.

Типы без *2015* в конце их имени действовали до 01 января 2015, с *2015* - после 01 января 2015. Далее, для упрощения рассмотрения будем учитывать только типы, действующие после 01 января 2015; а их имена использовать без *2015* на конце, что согласуется с документом [1]. В нём тоже есть двойственное именование: на стр. 26 приведены типы с данным окончанием, но, далее, в документе они приводятся уже без окончания. Типы для выгрузок из ЕИС, как они определены в fcsExport.xsd, тоже не имеют *2015* на конце. Видимо, данное окончание было введено только на уровне документации для обозначения перехода.

Если следует учитывать типы, действующие до 01 января 2015, то необходимо провести дополнительный анализ по их отличию от текущих в части состава полей, их смыслового содержания.

Выгрузки из ЕИС по каждому из вышеприведённых типов в рамках одного контракта могут быть в схемах разных версий (поле **schemeVersion**), иметь несколько версий внутри схемы одной и той же версии (поле **versionNumber**). Во внимание необходимо принимать только те объекты, которые описываются наиболее поздней версией схемы и имеют максимальный номер версии. Остальные объекты являются историческими, хотя они могут представлять интерес для анализа того, как меняются параметры контрактов во времени. Такой анализ выходит за рамки текущего рассмотрения.

Данные типы имеют вложенную структуру. Так, например, в **contract** входит тип **stages**, внутрь которого входит тип **payments**. Такая иерархическая структура не очень удобна при аналитической обработке больших массивов данных, поэтому при выборке данных из "Госзатраты"следует привести результат к "плоскому де-нормализованному виду.

Для подсчёта сумм контрактов, которые были указаны на момент их заключения, достаточно воспользоваться информацией только из **contract**. Для подсчёта *фактически* потраченных сумм по контракту необходимо просуммировать все фактические выплаты, описанные типом **contractProcedure**, как в пользу заказчика, так и исполнителя. Таким образом, возникает, как минимум, два уровня подсчёта:

- планируемые траты по контракту (цена контракта при его заключении, запланированные этапы выполнения контракта с платежами по ним),
- фактические траты по контракту (фактичекие платежи по контрактус учётом неустоек и отмен).

В изначальной цели не конкретизируется, какой именно вариант подсчёта должен быть реали-

[3] https://clearspending.ru/

зован. Руководствуясь соображением, что для ответственного анализа нагрузки на бюджет следует учитывать именно фактические траты, необходимо взять данные из **contractProcedure**. При этом, полезно взять данные и из **contract**, что, например, позволит соотнести суммы контрактов и суммы фактических платежей по ним.

Следует отметить, что можно ввести и более содержательные уровни подсчёта. Например:

- траты по контракту, направленные на достижение заявленных результатов контракта,
- траты по контракту, направленные на достижение тех или иных целей более высокого порядка (национальные/федеральные/региональные программы и т.п.), в контексте которых был заключён контракт.

К сожалению, в ЕИС отсутствует явная увязка с подобными уровнями, а попытка провести её на основе существующих данных требует приложения серьёзных усилий, включая исследовательские, т.к., вероятно, требует разработки и применения методов продвинутой аналитики, машинного обучения. Данные уровни не входят в настоящее рассмотрение.

Согласно ст. 34 Федерального закона №44-ФЗ для некоторых контрактов может быть указана не твердая цена, а ориентировочная или максимальная (с формулой цены). Для таких контрактов особенно важно учитывать именно фактические платежи, т.к. сумма платежей может сильно отличаться от ориентировочных/максимальных цен.

Итак, имея корректные данные по типам **contract** и **contractProcedure** в "плоской"структуре, их можно подвергнуть требуемой агрегации при помощи программных средств: Excel (Pivot Tables), Python, R и т.п.

Но! практика подсказывает, что данные непосредственно из ЕИС не всегда являются корректными ("Прозрачность информации не означает её достоверности"[4]), поэтому следует рассмотреть и вопрос приведения первичных данных к таковым настолько, насколько это возможно (перед их анализом).

3. Искажения в данных

Данные по контрактам при поступлении в ЕИС проходят логические проверки: на это указывает документация [1, стр. 13], [2]. Детальное изучение сущности данных проверок требует значительного времени. Беглое же ознакомление с ними подсказывает, что осуществляется три типа проверок: а) на корректность типа значения и его макс. длины на уровне соответствия XSD-схеме, б) на ссылочную целостность, в) на соответствие логическим бизнес-правилам. При этом, не очень понятно проводятся ли указанные проверки только для интеграционного взаимодействия или же и для ручного ввода контрактов на сайте. Интересно отметить, что для многих базовых типов (**productName** и т.п.) в XSD-схеме [3, файл IntegrationTypes.xsd] не установлены ограничения на минимальную длину полей. Данные типы могут не иметь значения для рассматриваемой цели, но отсутствие некоторых минимальных ограничений по ним может косвенно пояснять, почему практика свидетельствует о том, что данные в ЕИС могут быть пропущены или искажены, несмотря на ряд логических проверок.

ЕИС существует и активно используется в масштабах всей страны с 2011 года. Естественно, что за всё это время сложилась разнообразная практика её применения. К сожалению, не всегда благонадёжная. Некоторая доля контрактов может неумышленно или умышленно иметь искажённые данные, тем более, что, как было пояснено выше, для этого разработчиками системы не были ещё закрыты даже некоторые простые "лазейки".

Искажения, скорее всего, носят разные виды. Полный (почти полный) их перечень можно выявить только после тщательного аудита данных и итерационного формирующегося опыта работы с данными вкупе с пониманием сути Федерального закона №44-ФЗ. Список некоторых известных

[4] https://clearspending.ru/page/about/faq/#Vsio-Zhe-Po-Zakonu

искажений приведён по ссылке https://clearspending.ru/in-control/.

Можно отметить, что некоторые существенные поля (например, signDate или budgetLevel) указаны в [1], как необязательные к заполнению, но при изучении комментариев к ним становится понятным, что, в конечном итоге, они заполняются ЕИС. Вероятно, какие-нибудь ошибки, "дыры"в процедурах могут приводить к тому, что эти поля так и остаются незаполненными. Для ответа на данный вопрос следует провести первичный (exploratory) анализ данных.

Умозрительно, для целей анализа, исходя их описания типов данных, предусмотрим следующие виды искажений вместе со стратегией по работе с ними:

Таблица 1: Возможные искажения данных

Характер искажения	Стратегия реагирования
Пропуск атрибутов, которые указаны в [1] необязательными, но являются существенными для целей анализа	Отброс таких контрактов с их пометкой, что по ним следует позже сформировать методику восстановления данных.
Указание дат в различных форматах (DD.MM.YYYY или MM.DD.YYYY)	Даты согласно XSD-схемам последних версий должны быть указаны в соответствии типом xs:date. Для более ранних версий, возможно, не было такой строгой типизации, поэтому может понадобится своя собственная валидация. Все контракты с датами, не удовлетворяющими xs:date отбрасываются с их маркировкой.
Для контрактов заключённых не в российских рублях, указан валютный курс сильно отличающийся от установленного Центробанком РФ	Для выявления таких контрактов следует соотнести установленный валютный курс с официальным, при существенном расхождении (более 1%?) - отбрасывать.

В целом, при отсутствии хорошего опыта работы с данными конкретной природы, записи с искажениями разумно пока отбросить, а при описании результатов анализа - указать долю тех данных (от общего количества), по которым данный результат был получен. По мере накопления опыта будет вырабатываться методика и устранения искажений (а не просто отброса записей с ними), чтобы повысить долю используемых данных.

4. Методика

Учитывая вышеизложенное, методика целевого анализа данных представляется следующим образом.

Необходимо выгрузить все объекты типов **contract** и **contractProcedure** за интересующий период в денормализованном ("плоском") виде. Для каждого контракта отобрать только актуальные объекты, отбросив исторические (признаки для этого были показаны выше).

После такой обработки для каждого контракта должны остаться: только один объект типа **contract** и только актуальные объекты типа **contractProcedure**. Последние могут отсутствовать, если контракт ещё не начал исполняться.

Далее, следует привести процедуры работы с возможными искажениями в данных, описанные в Таблице 1.

К этому моменту, данные должны представлять из себя "чистый непротиворечивый набор, готовый к целевому анализу. Тогда, инструментами анализа данных, возможно подсчитать как

суммарные цены контрактов по периодам (по объектам типа **contract**), так и фактические выплаты по контрактам по периодам (по объектам типа **contractProcedure**).

5. Техническое задание для формирования выгрузки

Необходимо предоставить данные по госконтрактам по Федеральному закон №44-ФЗ: значения отдельных элементов объектов типов **contract** и **contractProcedure** раздельными наборами (наборами файлов) за период с DD.MM.YYYY по DD.MM.YYYY (включительно):

1. Для типа **contract** выбираются контракты с датой заключения (./signDate) внутри указанного интервала.
2. Для типа **contractProcedure** выбираются контракты с датой платежного документа, т.е. значения следующих элементов находятся внутри указанного интервала: ./executions/payDoc/documentDate, ./penalties/penaltyAccrual/penaltyDocument/documentDate.
3. Для каждого контракта, по которому в выборку попадает тип **contractProcedure**, обязательно следует включить в выборку и тип **contract** для данного контракта, даже если дата его заключения (signDate) выходит за пределы указанного интервала. Это требуется для подсчёта сумм, указанных в тех контрактах, которые были заключены ранее указанного периода, но исполнение по которым началось уже после начала периода.
4. Для типа **contract** требуется набор полей, приведённый по ссылке: ссылка, вкладка «contract».
5. Для типа **contractProcedure** требуется набор полей, приведённый по ссылке: ссылка, вкладка «contractProcedure».
6. Выгрузка предоставляется файлами в одном из двух форматов: а) CSV с разделителем в виде символа табуляции в кодировке UTF-8, б) Apache Parquet.
7. На каждый тип и каждый месяц создаётся свой собственный файл с именем согласно маске: <имя типа>_<год><месяц>.<расширение>.
8. CSV-файлы должны быть заархивированы (zip/bz2), раздельно.

Ожидаемый размер выборки для указанного периода: 3-4 млн. записей типа **contract** и 6-8 млн. записей типа **contractProcedure** в расчёте на один год. Для более точной оценки количества записей в выгрузке, её размера в байтах можно сформировать её для нескольких произвольных месяцев и интерполировать пропорционально всему периоду.

6. Открытые вопросы

1. Следует ли приводить затраты разных периодов к уровню текущих цен, чтобы их можно было сравнивать в абсолютных значениях более точно?
2. Если задачу воспринимать, как оценку нагрузки на бюджет, то следует ли затраты по исполнению контрактов суммировать с выплатами по соглашениям по субсидиям?

Источники

1. Форматы информационного взаимодействия по 44-ФЗ: Альбом ТФФ 9.2 в режиме принятых изменений. — 30.05.2019. — URL: http://zakupki.gov.ru/epz/main/public/download/downloadDocument.html?id=31298.
2. Форматы информационного взаимодействия по 44-ФЗ: Интеграционные контроли и алгоритмы РК 9.1. — 13.03.2019. — URL: http://zakupki.gov.ru/epz/main/public/download/downloadDocument.html?id=30266.
3. Форматы информационного взаимодействия по 44-ФЗ: Схемы 9.2 итерация 5. — 11.06.2019. — URL: http://zakupki.gov.ru/epz/main/public/download/downloadDocument.html?id=31299.

A. Замечания по «ГосЗатраты»

A.1. Расхождение между ЕИС и ГосЗатраты по сумме выплат

На примере контракта regnum=1490908343515000003 ("Керченский мост") можно видеть, что сумма выплат по данным из "Госзатраты"составляет 8 596 938.09 рублей (см. пример подсчёта по сссылке), а по ЕИС - 222 177 201 692,82 рублей при цене контракта в 222 823 769 686,32 рублей (см. ссылку на контракт на ЕИС).

Это может быть связано с тем, что в "ГосЗатратах"учитываются не все объекты типа **contractProcedure**, а только самый последний, хотя таких объектов может быть несколько за время исполнения контракта (необходимо учитывать все).

A.2. Наименование товаров и услуг - указано

На сайте "ГосЗатраты"в разделе Аномалии и ошибки, Некорректное заполнение поля "Наименование товара/работ/услуг"приведены контракты, у которых предположительно неверно заполнено поле "Наименование товара/работ/услуг". Три произвольно выбранных контракта из списка в данном разделе (https://clearspending.ru/contract/01353000026110000053/, https://clearspending.ru/contract/03353000270110000001/, https://clearspending.ru/contract/03352000117110000018/) содержат наименования работ и товаров в данном поле. Для некоторых из них оно, возможно, сформулировано общими словами, но все из них не обладают описанными признаками аномалии:

> В ряде случаев при вводе информации о государственном контракте заказчики не заполняют поле "Наименование товара/работ/услуг"или указывают в нем код ОКДП/ОКПД, в результате чего невозможно получить конкретные данные по контракту. Данная проблема приводит к затруднениям в последующем анализе структуры заказов государственных и муниципальных заказчиков и поставщиков.

Возможно, следует дополнить или уточнить описание признаков?