Создание системы для OLAP – кубов

Миронов Дмитрий Сергеевич, студент магистратуры
Научный руководитель:
МИРЭА — Российский технологический университет (г. Москва)

В статье автор описывает построение системы позволяющую быстро внедрять аналитические методы и принимать решения на основе данных

Ключевые слова: *OLAP, куб, аналитика, база данных.*

При создании системы для OLAP – кубов нужно учитывать две основные проблемы, это потребляемая память при расчете каждой агрегации куба и скорость расчета всех агрегаций. Поскольку при каждом последующем расчете агрегаций, объем куба увеличивает в несколько раз.

На текущий момент есть два основных решения в области BI – Hyperion planning и Qlik sense. Оба решения предлагают ограниченный функционал и высокую стоимость владения. Детально прописывая план внедрения платформы, многие компании сходятся во мнении [1], что быстрее и дешевле создать свою платформу, используя более современные инструменты анализа данных, разработанные для DS (data science) и открытые библиотеки для визуализации для современных фреймворков JavaScript.

В данной статье описывается построение системы позволяющую быстро внедрять аналитические методы и принимать решения на основе данных. В виду того что в создаваемой системе основной функционал будет построен на создании OLAP – куба.

Пример функционального решения

Перед началом формирование OLAP - куба, необходимо создать его структуру (рис. 1), то из чего он будет состоять. Основой, конечно же являются данные и аналитики. Аналитик — это измерение куба, то что будет группироваться, при формировании куба.

Необходимо указать столбцы основного файла и их иерархию. Пример иерархии или же одной аналитики– «месяц – неделя – год». Данный аналитик будет называться в структуре, например, «Дата». Таких аналитиков в кубе может быть не ограниченное количество, но с каждым добавляемым аналитиком и глубины иерархии, увеличивается объем куба и сложность при его расчете. В качестве данных необходимо указать столбцы с числовыми значениями на основе которых буду проводиться расчеты.

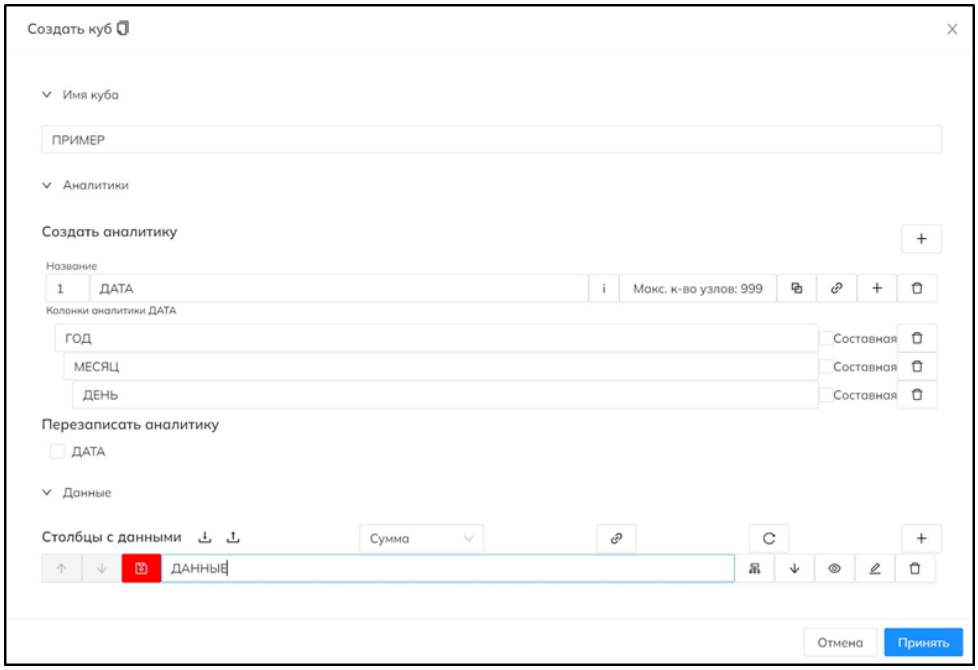

Рис. 1. Формирование структуры [разработано автором]

После описания аналитиков и данных куба, формируются параметры каждого аналитика и столбов с данными, для того чтобы эффективно хранить полученные значения и быстро выводить данные при запросе.

Для каждого аналитика в структуре должна находится следующая информация: название аналитики, названия столбцов иерархии аналитики, порядковый номер аналитики, длина индекса для данной аналитики.

Как формируется куб.

Основной идеей формирование куба, заключается в том, чтобы формировать каждый последующий разворот куба мы будет на основе

предыдущего, так идя он начальной таблице, мы постепенно будет ее увеличивать, добавляя сгруппированные значения и из полученной таблице будем рекурсивно группировать её по каждой иерархии аналитика (рис. 2).

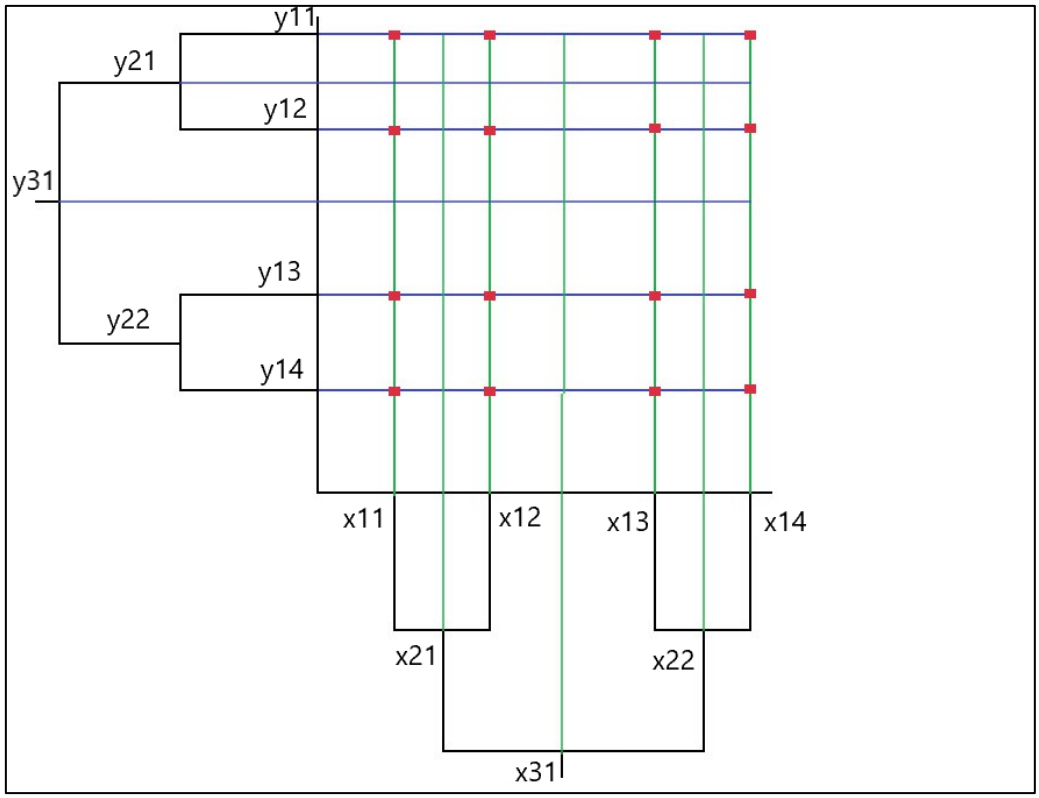

Рис. 2. Агрегация в двумерной плоскости [разработано автором]

На рис. 2 представлена первая агрегация, при которой формируются значения на листьях. Добавляя полученные значения к основным, мы можем делать группировку следующего уровня иерархии аналитика. Нам не приходится каждый раз рассчитывать куб до нужного уровня, поскольку при каждой агрегации уровня иерархии аналитика, сформированные данные уже будут, но используя больше памяти при каждой последующей агрегации над кубом. Данную проблему можно решить тем, что каждый столбец данных считать отдельно, последовательно. Таким образом если у нас в изначальных столбцах данных находятся 100 столбцов, мы соберем куб 100 раз для каждого столбца, следовательно, мы уменьшим объем потребляемой памяти при формировании куба, но скорость формировании всего куба увеличиться.

Как формируются индексы.

С точки зрения математики для того чтобы хранить плоскость на прямой необходимо каждой точки дать свой уникальный индекс (номер), таким образов если у нас многомерная плоскость, то для каждой из плоскостей нужно присвоить свой индекс. Сформированный куб это и есть многомерная пространство плоскостей. Каждый индекс будет состоять из индексов аналитика.

Для того чтобы сформировать индекс аналитика, нужно подготовить таблицу, в которой они будут храниться. Данная таблица будет состоять из уникальных значений каждого уровня иерархии аналитика, в которой в столбце «Имя» будут находится имена всех уникальных значений аналитики, а в столбце «Индекс» число для каждого уникального значения в столбце «Имя».

Перед подстановкой индексов, нужно отредактировать с агрегированные данные. Для этого нужно создать столбы с названиями всех аналитиков и подставить первое не пустое значение из названий столбцов иерархии аналитика. Таки образов в итоговой таблицу окажется количество столбцов равное количеству аналитиков.

Далее необходимо заменить значения в столбцах на значение в таблице с индексами и соединить каждую строку в одну ячейку. Таким образом получаться индексы для каждого значения OLAP-куба.

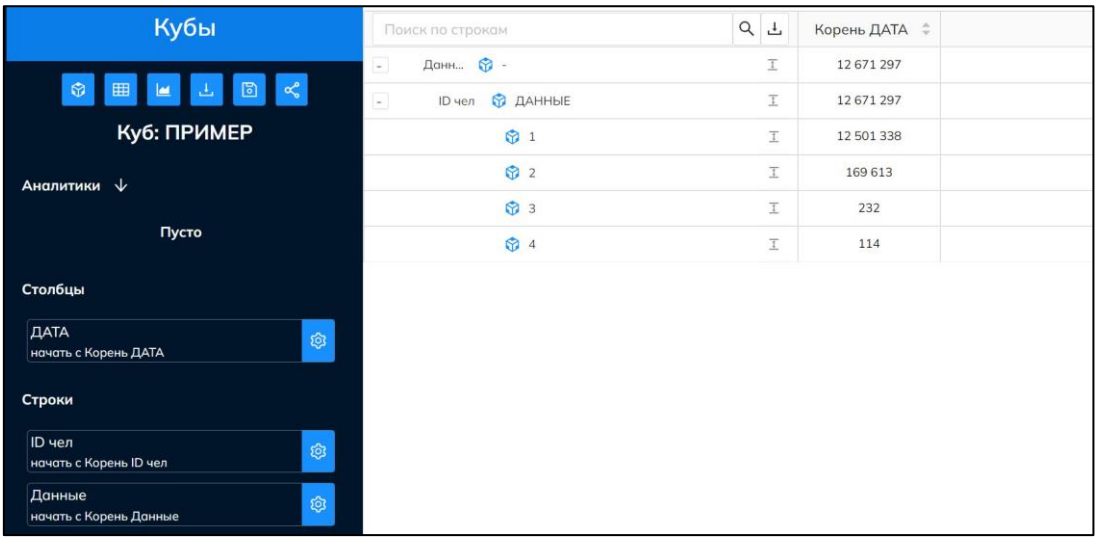

Рис. 3. Разворот куба [разработано автором]

Литература:

1. Маршак С. Я. Вот какой рассеянный. М. : АСТ, 2016.
2. Чуковский К. От двух до пяти. М. : Детская литература, 2017.