ПРИНЦИПЫ ПОСТРОЕНИЯ МАСШТАБИРУЕМЫХ СИСТЕМ УПРАВЛЕНИЯ И КОНТРОЛЯ ДАННЫХ НА ОСНОВЕ ПРОГРАММНЫХ КОМПОНЕНТ С ОТКРЫТЫМ КОДОМ[1]

Тимофеев Б. М., студент кафедры информационно-аналитических систем СПбГУ, timbog@mail.ru

Аннотация

В данной статье разобраны современные подходы к построению масштабируемых систем управления данными, их основные компоненты и характеристики, проанализированы основные требования к таким системам и на основании них предложены варианты решения для хранения данных и метаданных.

Введение

В связи со стремительным ростом объемов хранимой информации становятся все более актуальны системы, позволяющие хранить и оперировать большим количеством слабоструктурированных данных. Такие системы можно условно разделить на 2 категории: коммерческие и находящиеся в открытом доступе. Поскольку использование коммерческих систем зачастую связано со значительными финансовыми затратами, исследование посвящено построению систем на основе программных компонент с открытым исходным кодом.

К требованиям данной системы необходимо отнести надежность (потери данных могут быть критическими для пользователей, вся информация должна дублироваться), защищенность от несанкционированного доступа (в том числе и разграничение прав пользователей внутри самой системы), а также приемлемая стоимость, что и обуславливает использование компонентов со свободной лицензией. Отдельно следует выделить такое требование к системе, как сохранение данных в течение определенного периода времени (архив),

[1] Работа поддержана Центром разработок EMC в СПб и Исследовательским центром EMC в Сколково

поскольку информация может иметь актуальность даже спустя большой промежуток времени. Таким образом, данные, доступ к которым давно не осуществлялся, могут быть отправлены в архив системы хранения.

Системы для управления данными становятся все более и более востребованными, также многим их пользователям необходима возможность использования метаданных. В качестве примера области, где востребованы подобные платформы, может служить любое учреждение, в котором важна не только сама информация, но и ее описание.

Для иллюстрации подобного объединения в данной статье будет использоваться пример поликлиники с медицинской лабораторией. В таком случае существует две основных группы пользователей системы: врачи, которым необходима информация о здоровье человека (например его медкарта), и лаборанты, в обязанности которых входит исключительно проведение анализов, не имеющие полного доступа к информации о пациенте. Информация, которая хранится в электронной медкарте, может быть разделена на данные и метаданные следующим образом: перенесенные болезни, результаты анализов и т. п. будут являться непосредственно данными, а имя пациента, дата создания карты, имена лаборантов, производивших анализы для пациента, будут являться описанием данных, то есть метаданными. Немаловажную роль в подобного рода учреждениях играет разделение прав доступа к информации: каждый сотрудник может иметь доступ исключительно к той части данных, которая затрагивает его деятельность, таким образом лаборант не сможет просматривать сведения о заболеваниях пациента (он может только добавить результаты анализа).

Структура хранимой информации

Для решения вышепреведенной задачи может служить масштабируемая система управления и контроля данных, построению которой на основе компонентов с открытым исходным кодом посвящена данная статья.

Одним из наиболее важных факторов в построении подобных систем является разделение данных и их описания (метаданных), а также контроль пользовательского доступа к ним (механизм авторизации пользователей с разными правами на доступ к данным).

Системы управления данных, в силу их универсальности (независимости от предметной области данных), должны уметь оперировать самыми разными данными, однако даже в них присутствует

структуризация информации, которую можно условно разделить на 2 типа: данные и метаданные.

Данные – это та часть информации, которая имеет основную ценность для пользователя, то есть это информация в “сыром” виде, без описания.

Таким образом, объекты являются листьями дерева хранимых данных, датасеты представляют из себя нетерминальные вершины. Еще одной характерной особенностью датасетов является то, что для них определены права доступа.

Метаданные – это не информация в чистом виде, а ее описание. Одним из примеров описания данных являются такие их метрики, как дата создания, дата последнего изменения, создатель и т. д.. Наиболее подходящим способом представления метаданных является использование пар “ключ-значение” (то есть хэш-таблицы), с помощью которых можно абстрагироваться от предметной области данных. Для удобства пользователя, каждая единица данных может иметь служебные метаданные (такие как контакты создателя), заполняемые системой автоматически.

Модуль авторизации пользователей

Одним из основных принципов построения данных систем является возможность настройки доступа к информациии ее владельцами. Как было указано в примере, приведенном во введении, в некоторых случаях пользователи должны иметь полный доступ к части данных, в то время как другая часть полностью скрыта от них.

Для того, чтобы нагляднее демонстрировать права пользователей, их принято разбивать на группы, ниже будет продемонстрировано три основных из них:

Администратор данных (Data steward) – пользователь системы, контролирующий доступ к данным другими пользователями внутри какого-то датасета.

Обычный пользователь – пользователь системы, работающий с датасетом и обладающий правами доступа к нему, заданными администратором данных.

Системный администратор – пользователь системы, отвечающий за поддержание работоспособности системы, который имеет полные полномочия по ее управлению, в т.ч. наделению пользователей правами администратора данных, установке размера дискового пространства для каждого пользователя и.п.

Помимо вышеперечисленных ролей, существует также разделение на методы, с помощью которых пользователи могут оперировать данными.Такими методами являются:

1) чтение метаданных

2) запись метаданных

3) чтение данных

4) запись данных

Таким образом, администратор данных может выбрать пользователей и методы их доступа к какому-либо подконтрольному ему датасету. Например, врач может разрешить лаборантам записывать в датасет, содержащий результаты обследования пацентов. Также он может открыть доступ на чтение результатов этих анализов для врача, который тоже нуждается в этой информации.

Компоненты системы

Для построения вышеописанной системы необходимо наличие следующих компонентов:

- Система хранения данных
- Система хранения метаданных
- Модуль авторизации пользователей
- Балансировщик нагрузки на систему
- Система резервного копирования данных

Для каждого из этих компонентов платформы существуют решения, находящиеся в открытом доступе.

Заключение

Построение конкретной системы управления данными во многом зависит от требований к ней. В контексте данной работы рассматривались такие требования, как надежность, защищенность от несанкционированного доступа, а также низкая стоимость.

Востребованность подобного рода систем растет не только в области научных исследований, но также и среди обычных пользователей, которые могут пользоваться облачными сервисами (например социальными сетями или облачным хранилищем), использующими в качестве хранилища данных подобные платформы.

Литература

1. Сайт системы хранения данных iRODS [электронный ресурс] – Режим доступа: http://irods.org/, свободный.

2. Сайт платформы ATTIVO [электронный ресурс] – Режим доступа: http://www.attivio.com/platform, свободный.