Отображение модели данных NetCDF в реляционную модель для работы с коллекциями данных дистанционного зондирования

Д.Л. Чубаров, В.А. Кихтенко, Н.Н. Добрецов

Институт вычислительных технологий СО РАН
Новосибирск

Всероссийская конференция «Обработка пространственных данных в задачах мониторинга природных и антропогенных процессов»

Бердск, 2017

План доклада

1) Модели данных и тест времени
2) Модель файл-переменная-атрибут (модель netCDF)
3) Как перейти от файлов к коллекциям без потерь?
4) История одного подхода к решению проблемы работы с коллекциями геоданных — история hVault
5) О чём забыли при разработке hVault
6) Что было бы если … – современные инструменты для работы с коллекциями

Модели данных для архива спутниковых измерений

Особенности спутниковых измерений

1) **Данные имеют пространственные и временные атрибуты** – каждое измерение связано с объектом в пространстве и отметкой времени,
2) **Большой объём данных** – ограничение на возможность копирования,
3) **Поток данных** – данные поступают по мере регистрации как единичные измерения, или порциями с ограниченным временным и пространственным охватом,
4) **Множественность характеристик** – одновременно измеряется несколько связанных характеристик, их нужно различать при обработке
5) **Многомерные массивы** – для ускорения вычислений необходимо выполнение операций с массивами,
6) **Один запрос — много данных** – данные используются целиком, либо блоками, которые могут иметь сколь угодно большой пространственный или временной охват.

Тест времени

- 1985 – NASA CDF
- 1987 – NCSA HDF
- 1989 – UCAR netCDF

Основные характеристики:

- Хранение нескольких многомерных массивов в одном файле
- Хранение сопроводительной информации вместе с данными – модель распространения данных на CD или на ленте

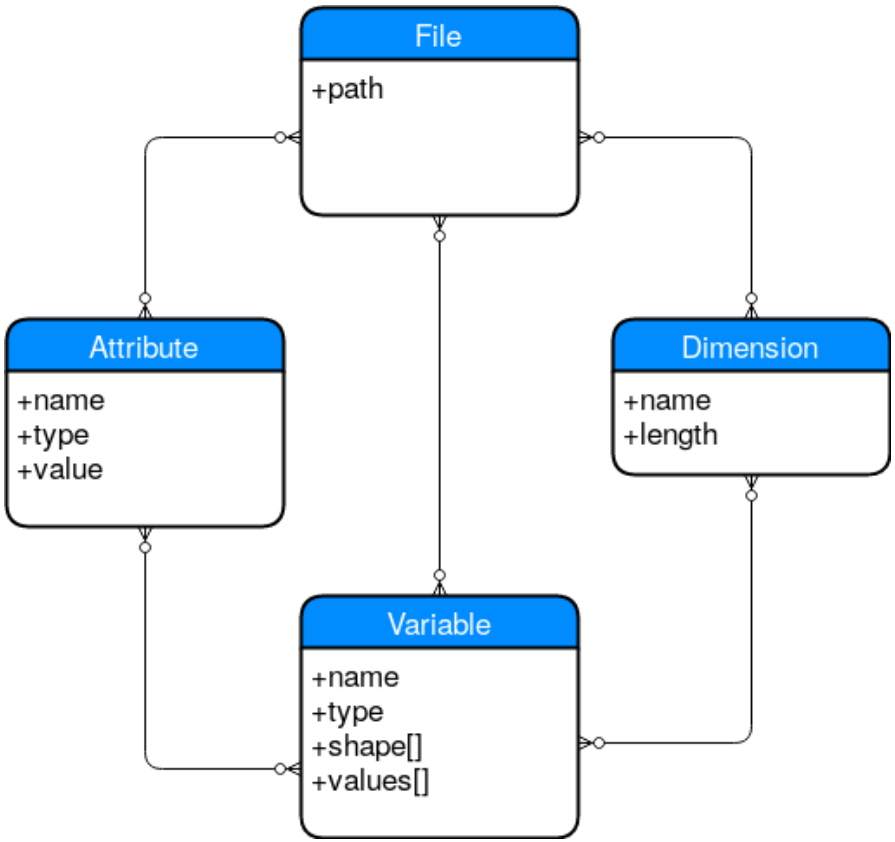
Проблемы:

- Представление чисел на различных архитектурах ЭВМ
- Кодировки текста
- Ограничения на размер массива
- ….

Результат:

- Единая модель хранения результатов измерений, основанная на HDF5

Модель файл–переменная–атрибут (netCDF)

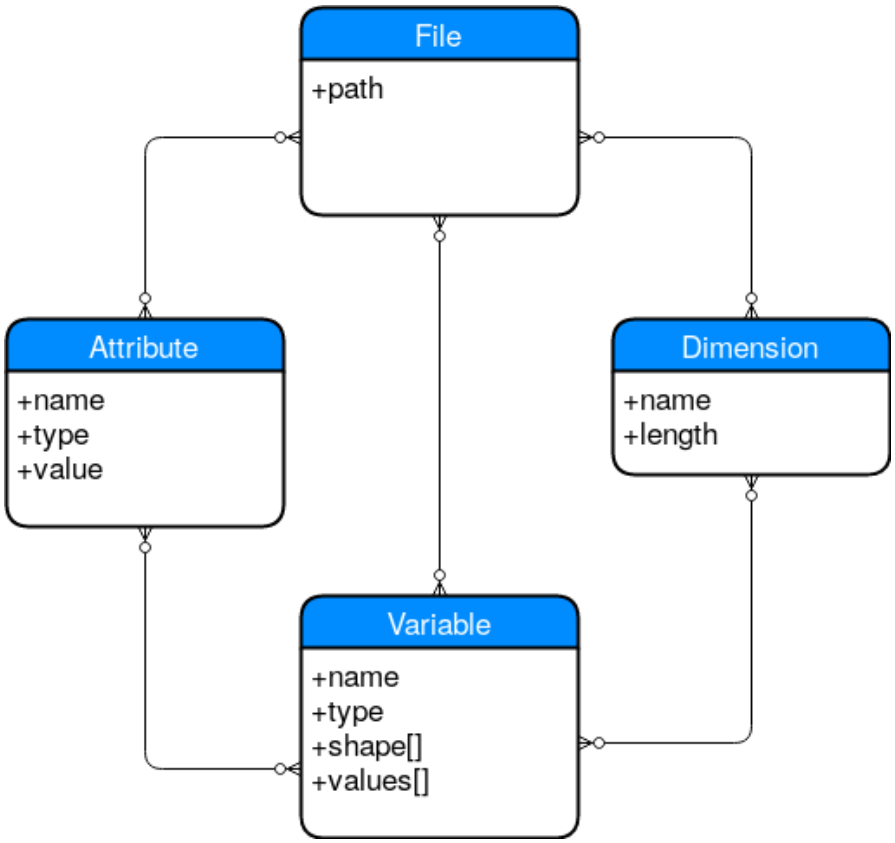


- Файл содержит одно или несколько измерений
- Все характеристики в одном файле
- Многомерные массивы
- С каждым файлом связаны пространственные и временные атрибуты
- С каждой переменной связаны атрибуты, постоянные для всех файлов коллекции

Работа с коллекцией

- Обработка данных там, где они есть — принцип *in-situ*
- Учёный получает только необходимые ему результаты
- Промежуточные результаты остаются в коллекции
- Сохраняется последовательность операций, ведущих к получению результата — может быть воспроизведена при пополнении коллекции или изменении исходных данных

Модель файл–переменная–атрибут (netCDF)



Проблема работы с коллекцией

Работа со всеми значениями одной переменной либо с блоком, выходящим за пределы одного файла, требует работы со списком файлов

Почему это проблема?

- С точки зрения программной инженерии – нарушение инкапсуляции
- С точки зрения учёного – дополнительные сложности

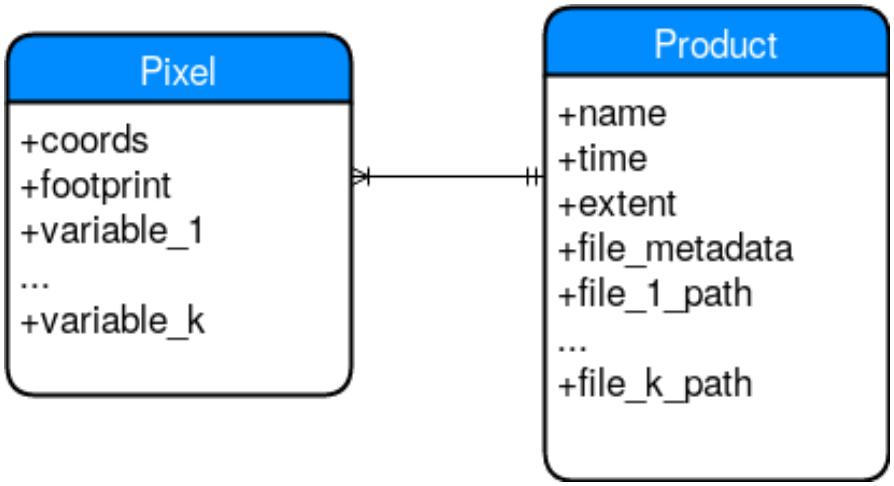
Проблема работы с коллекцией

Как устранить файл из модели
файл — переменная — атрибут?

История hVault

- Дано: постоянно пополняющаяся коллекция стандартных продуктов MODIS
- Надо: выбрать данные по температуре поверхности Земли, восстановленной с помощью двухканального алгоритма (Wan, Dozier), MOD11A1, для заданной точки и интервала времени

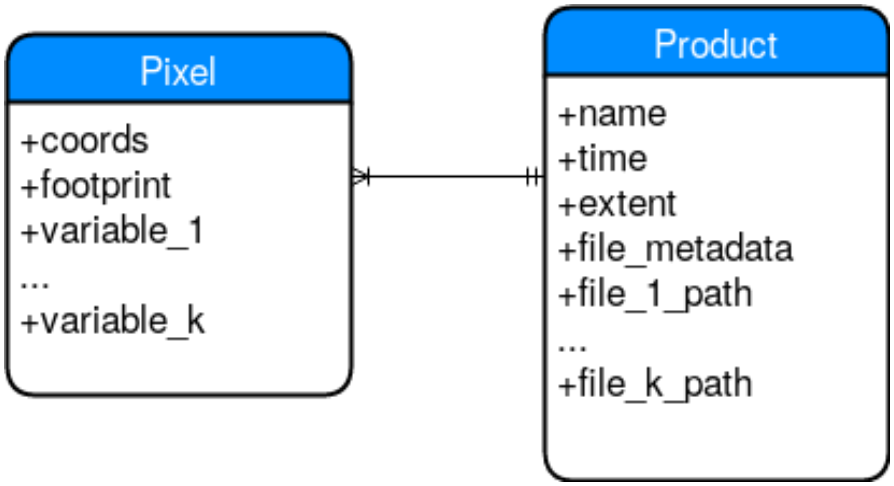
Отображение в реляционную модель



Основные принципы

- Коллекция файлов HDF как источник внешних данных для реляционной СУБД
- С каждым пикселем ассоциирован кортеж значений измеряемых характеристик

Отображение в реляционную модель



select pixel.lst_day
from pixel ⋈ product
where
ST_Contains(*Berdsk*,pixel.coords) **and** product.time between 28-08-2017 and 31-08-2017

Основные принципы

- Коллекция файлов HDF как источник внешних данных для реляционной СУБД,
- с каждым пикселем ассоциирован кортеж значений измеряемых характеристик,
- файлы не видны

Результат

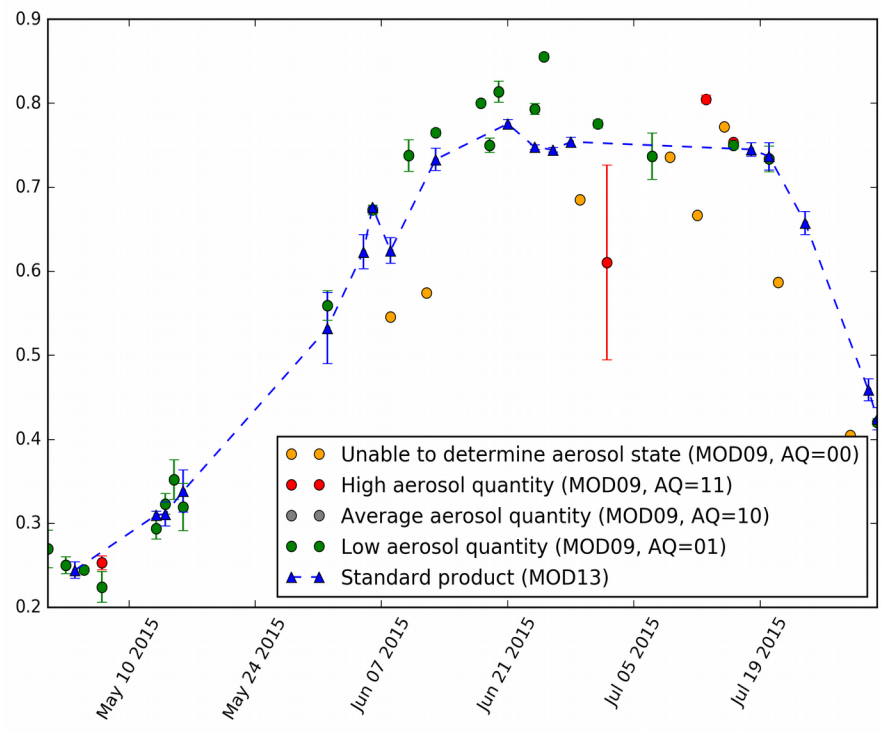
Особенности

- Возможность работы с продуктами L2 (атмосферные продукты MODIS)
- Возможность работы с большинством форматов пространственных данных посредством GDAL
- Обработка может полностью осуществляться средствами СУБД

Проблемы

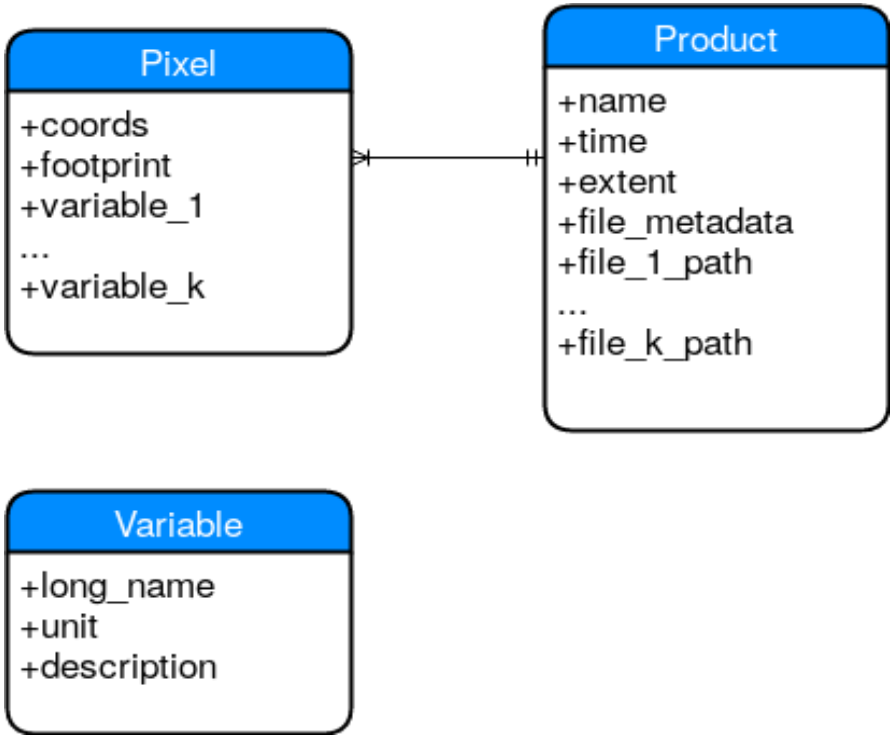
- Отсутствие поддержки работы с массивами
- Пропали атрибуты
- При добавлении новых характеристик нужно менять структуру таблицы *pixel*

Пример NDVI по данным MOD09



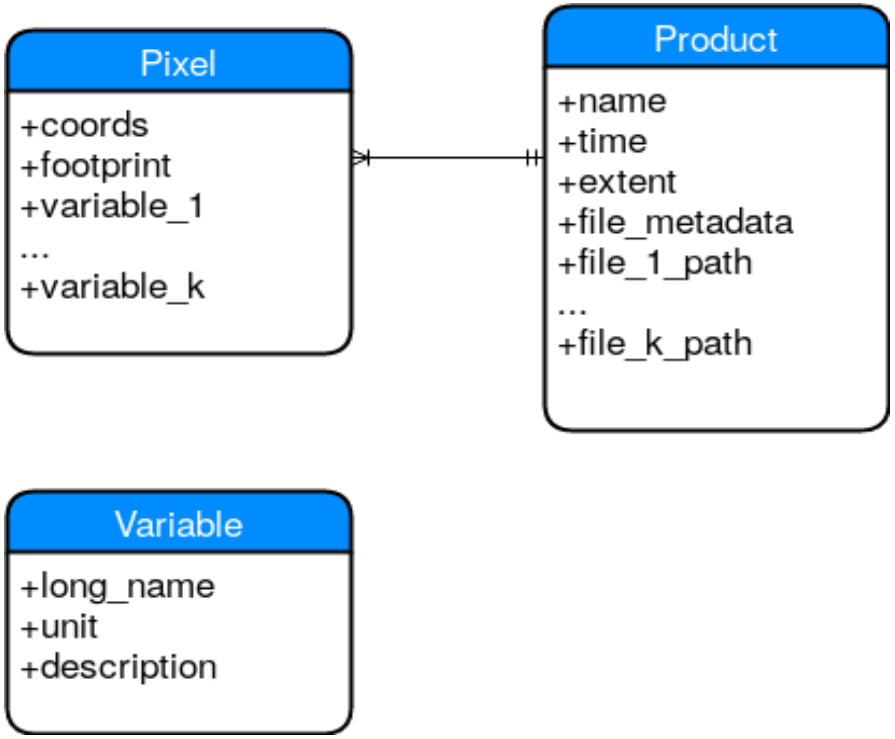
- Вычисление NDVI по данным MOD09 — фильтрация по состоянию атмосферы (атмосферный аэрозоль)
- Атрибуты переменной QA содержат описание битовой маски

С добавлением атрибутов



- Атрибуты переменных загружаются при добавлении нового продукта
- При добавлении каждого продукта MODIS меняется структура таблиц для Pixel и Product

С добавлением атрибутов



- Атрибуты переменных загружаются при добавлении нового продукта
- При добавлении каждого продукта MODIS меняется структура таблиц для Pixel и Product
- Как быть с продуктами **VIIRS/NPP ?**

что было бы если ...

- Задачи обработки климатических сеток — подсчёт числа снежных дней в году – Ophidia (data chunk extenstions)
- Задачи предоставления доступа к архиву Landsat/Sentinel – построение безоблачных композитов – Australian Geoscience Data Cube (AGDC)
- Наличие инфраструктуры для массивно-параллельных вычислений – Google Earth Engine
- Задачи отображения различных спутниковых покрытий для большого числа пользователей при ограниченных ресурсах – GeoSMIS

Вместо заключения

- При переходе от работы с отдельными файлами к работе с коллекцией в режиме in-situ необходимо обеспечить сохранение связей, содержащихся в исходной модели данных. И это возможно
- Для обеспечения эффективной работы с коллекцией необходимо преодолеть несколько неочевидных препятствий
- Сервисы?