Использование облачных ресурсов Google Cloud Storage и их интеграция с ресурсами ЦКП "ИКИ-Мониторинг"

Балашов И.В., Бурцев М.А., Прошин А.А., Лебединская Д.И.
Институт космических исследований РАН, Москва, Россия

Цель работы

- Исследование и разработка способов интеграции внешних облачных хранилищ стандарта S3 с архивами и информационными системами на базе ЦКП «ИКИ-Мониторинг»

Актуальность работы

- На сегодня облачные хранилища крупных компаний, таких как Google и Amazon, обеспечивают доступ к многопетабайтным архивам открытых данных ДЗЗ. Так, по ряду оценок, объём данных КА семейства Landsat и Sentinel, предоставляемый Google Cloud Storage, составляет около 50 Пб.
- В последние несколько лет появились API к хранилищам такого стандарта, а также специальные формы организации данных (например, Cloud-Based GeoTIFF), обеспечивающие возможность прямого доступа к требуемым фрагментам файлов данных без скачивания их целиком.
- Эти факторы сделали возможной прозрачную интеграцию хранилищ S3 с архивами ЦКП «ИКИ-Мониторинг» на базе технологии UNISAT для доступа к внешним данным.

Технология UNISAT

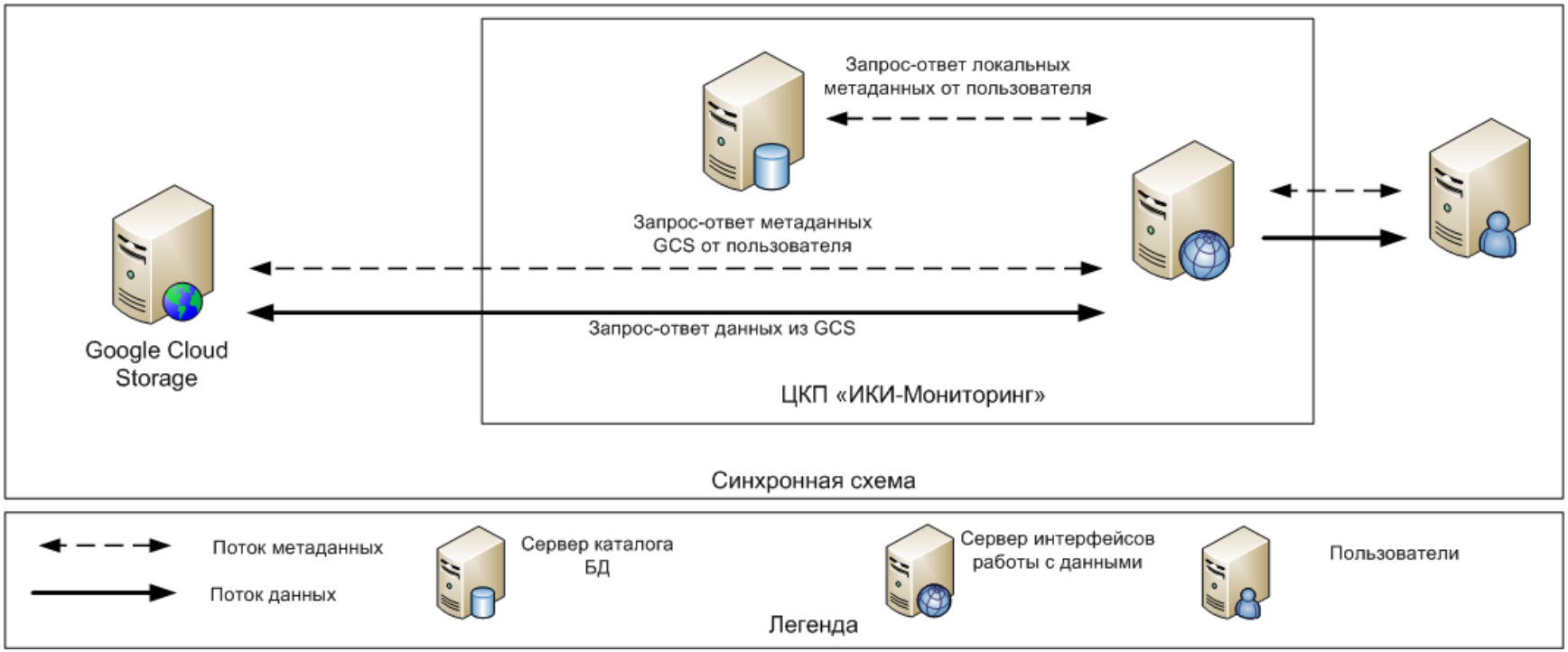
Технология UNISAT обеспечивает:

- Единообразное распределённое хранение разнотипных данных и каталогов развёрнутых описаний к ним;
- Предоставление пользователю достаточно сложных инструментов для работы с данными и их анализа, которые до недавнего времени были доступны только в рамках специализированного программного обеспечения;
- Предоставление доступа к «виртуальным» продуктам, т.е. продуктам, которые динамически формируются на момент запроса их пользователями на основе данных, имеющихся в архивах. Такой подход позволяет обеспечить доступ пользователей к интересующим их новым информационным продуктам, не прибегая при этом к переобработке исходных данных и не увеличивая размер имеющихся архивов.

Объект интеграции

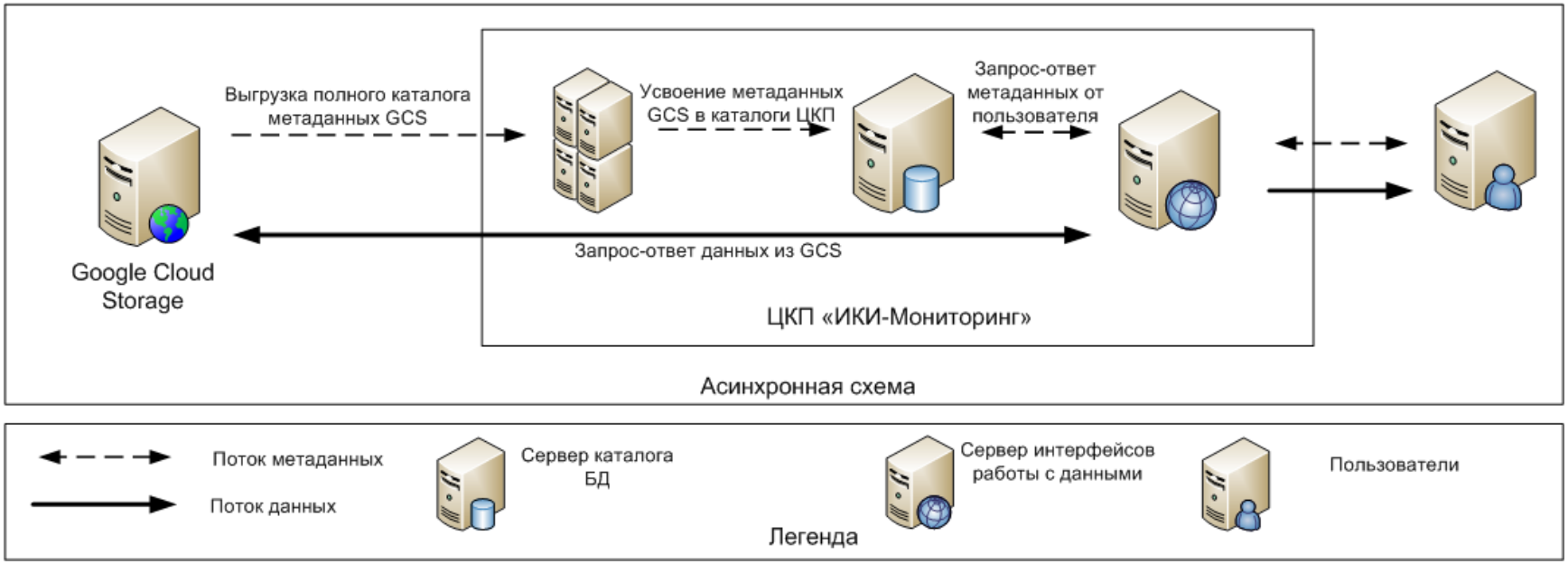
- Наиболее крупными и интересными представителями облачных систем, хранящих данные ДЗЗ, на сегодня являются Amazon (AWS) и Google Cloud Storage.
- Для пробной интеграции были выбраны ресурсы Google Cloud Storage, так как они предоставляют широкий спектр открытых, бесплатных данных, в том числе данные КА Landsat и Sentinel на весь мир с момента начала работы этих миссий.
- Подробнее с доступными данными можно ознакомиться по здесь: https://cloud.google.com/storage/docs/public-datasets

Возможные схемы интеграции - синхронная



Синхронная схема предполагает прямые запросы метаданных к каталогу GCS из интерфейсов доступа посредством API на базе SQL-подобной библиотеки BigQuery. Запросы на отображение данных также идут напрямую к GCS.

Возможные схемы интеграции - асинхронная



Асинхронная схема предполагает регулярную выгрузку каталога метаданных GCS с последующим усвоением в локальные каталоги Unisat. После этого все запросы метаданных работают только с локальными БД, запросы на отображение данных идут напрямую к GCS.

Выбор схемы интеграции

Синхронная схема	Асинхронная схема
Проще в реализации	Сложнее в реализации
Выше оперативность доступных данных	Неизбежны расхождения между реально доступными данными в GCS и известными нам
Не требует дополнительных программно-аппаратных ресурсов	Требуются дополнительные ресурсы для получения, конвертации и усвоения метаданных в локальные архивы
Хуже быстродействие получения метаданных по запросу в интерфейсе	Лучше быстродействие получения метаданных по запросу в интерфейсе

Наиболее критичным параметром в данном случае является быстродействие получения метаданных. При прямом запросе метаданных из GCS посредством BigQuery скорость выполнения запроса может составлять порядка 10-15 секунд, что становится неприемлемым.
Поэтому асинхронная схема выглядит предпочтительнее.

Текущее состояние

- Разработана и реализована схема и механика усвоения метаданных в рамках асинхронной схемы интеграции;
- Проверена возможность отображения данных напрямую их GCS в рамках систем на базе ЦКП «ИКИ-Мониторинг»;
- Ведется разработка модулей картографических интерфейсов для работы с данными GCS.

Технические решения

- Библиотека работы с растровыми изображениями GDAL с драйверами S3 (vsicurl);
- Python-обвязка для GDAL – pygdal;
- Конвертер метаданных из формата GCS в формат Unisat, реализованный на Python;
- Интерфейсные модули для работы с данными GCS – в процессе разработки.

Предварительные результаты

- Интеграция систем ЦКП «ИКИ-Мониторинг» с Google Cloud Storage технически реализуема и может обеспечить полную функциональность работы с данными, предоставляемую системами ЦКП;
- Оптимальным вариантом интеграции является асинхронный;
- Полноценная реализация интеграции ожидается весной следующего года.

Спасибо за внимание!

Работа выполнена в рамках темы "Большие данные в космических исследованиях: астрофизика, солнечная система, геосфера" (госрегистрация №0024-2019-0014)