Островок!

Моисеев Виталий

Как мы строили Lakehouse на Ozone

ostrovok.ru

Моисеев Виталий

В ИТ 20 лет

С “Большими Данными” более 10 лет

Ostrovok.ru
Ex.Aliexpress Russia , Ex.Ozon, Ex.Sber

Dev > DE > Platform Lead > Head of Data

За это время успел поработать с 5+ EB и очень большой платформой данных

@amnesiacus
itspec144@gmail.com

О компании

- Компания основана в 2010 году
- 700+ технических сотрудников

2,5 млрд

фотографий отелей храним и отображаем на своих серверах

100ТБ

данных обрабатывают наши - PostgreSQL сервера

30K

поисковых запросов в секунду обрабатывает наш поисковый кластер

37K

бронирований совершают на нашей платформе в день

400+

нагруженных железных серверов

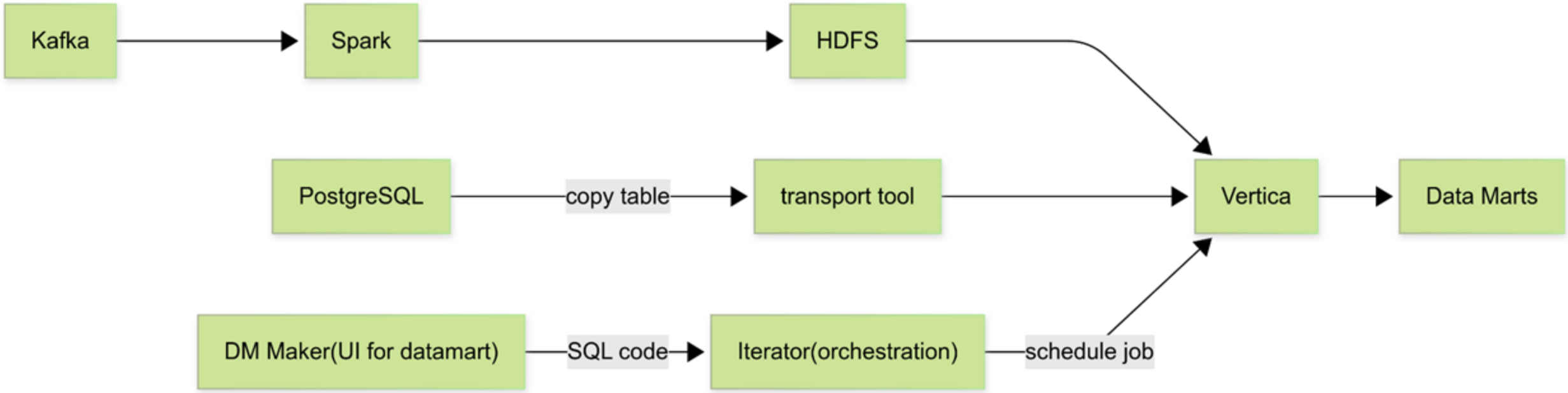
О чем поговорим?

платформа v.1 и выборы новой платформы

О технологии Apache Ozone и как мы ее применили для построения Data LakeHouse

Архитектура новой платформы и полученные инсайты

История начинается с платформы d.OTA v.1



320ТВ
HDFS кластер

3k таблиц
в хранилище

100ТБ
Vertica кластер

Предпосылки к появлению новой платформы

рост объемов данных и стоимости хранения и владения

Импортозамещение Vertica и Cloudera

Рост кол-ва пользователей

Legacy стек технологий

Ограничения масштабирования

Необходимость квотирования платформы по командам

Новые требования к работе с данными

Целевая нагрузка

Обработка

30+ ТВ Ежедневно

Хранение

700+ ТВ с ростом до РВ+

Пользователи

1000+

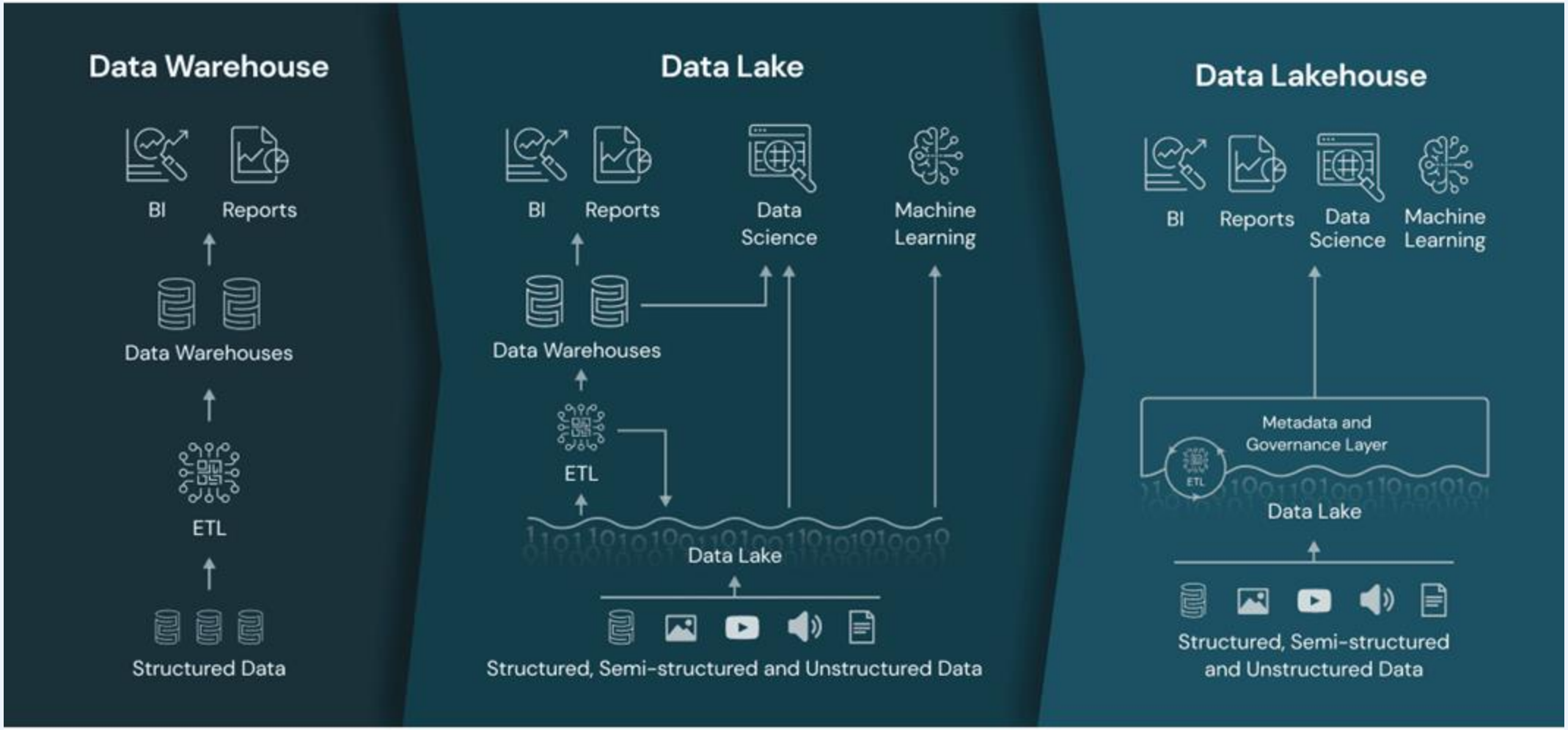
ETL dev.

100+ человек

Пожелания к новой платформе данных

- поддержка хранения разных форматов (parquet, JSON, картинки)
- kubernetes-ready
- разделение compute & storage
- поддержка schema evolution
- использование разных движков (возможность интерактивных запросов)
- эластичное масштабирование
- возможность хранить как маленькие так и большие файлы
- гибкость в управлении TCO по части стоимости хранения и обработки

Наши Пожелания близки к “Lakehouse”



https://www.databricks.com/glossary/data-lakehouse

Recap по “Data Lakehouse” от DataBricks

Основные Характеристики “Data LakeHouse” из публикации DataBricks

- Единое хранилище данных
- Разделение хранения и вычислений
- Поддержка BI и ML/AI в одном хранилище
- Использование открытых форматов данных
- Масштабируемость и высокая доступность

Полный Paper от DataBricks тут

как мы выбирали новый стек ?

популярная и часто встречающаяся архитектура аналитической платформы в РФ

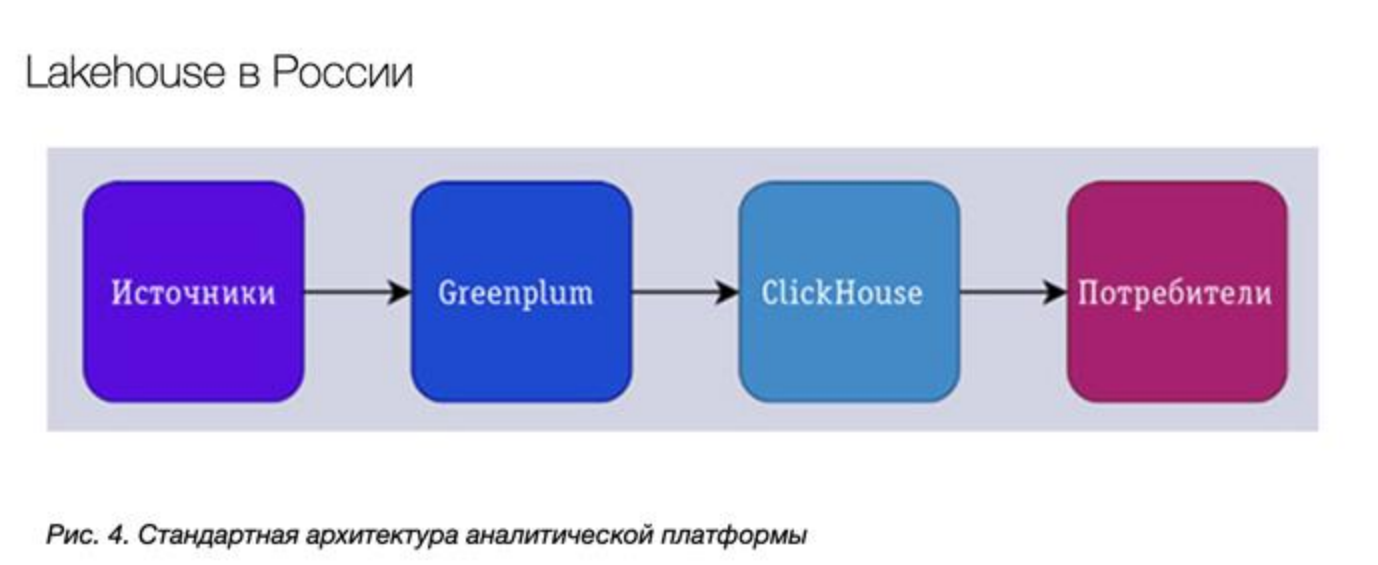


Рис. 4. Стандартная архитектура аналитической платформы

данная архитектура не покрывает существенной части наших пожеланий

https://osp.ru/os/2025/01/13059362

как мы выбирали?

	Hadoop + Trino + Spark	S3 + Trino + Spark	Комментарий
поддержка хранения разных форматов (parquet, JSON, картинки)	ДА	ДА	поддерживают обе технологии
разделение compute & storage	ЧАСТИЧНО ДА в случае worker нод и YARN (ДА при Spark в K8S)	ДА	поддерживают обе технологии, при переносе Spark в K8s
возможность хранить как маленькие так и большие файлы	НЕТ	ДА	в случае HDFS существует проблема хранения маленьких файлов
Эластичное масштабирование	НЕТ (NameNode имеет ограничения)	ДА	в случае HDFS могут быть сложности с масштабированием метаданных / NameNode
гибкого управления и оптимизации TCO в части стоимости хранения и обработки	НЕТ (так как платим за доп сервис)	ДА	в случае HDFS для маленьких файлов и картинок потребуется отдельный сервис

S3 + Trino хорошо подходит для задач построения “Lakehouse”

Выбрали основные компоненты

s3

HADOOP - > S3

trino

vertica - > trino

Iceberg

All -> Iceberg

k8s

BM - > k8s

Планируемые объемы для S3+Trino

Целевое и будущее состояние

s3 целевой

700+ TB

s3 в будущем

PB+

trino целевой

500+ CPU Cores / 6TB RAM

trino в будущем

Trino as a Service

Рассмотрели варианты реализации s3 он-прем

Ceph	minio	Apache Ozone
“Универсальный солдат”	“популярное решение”	“новая звезда”

Как мы выбирали s3 он-прем?

Решение	Тип	Консистентность	Масштабируемость	интеграции с Bigdata	Лицензия	Заметки
Ozone	Объект / Файл	Сильная	Поддержка Exabyte , миллиарды обьектов	Нативная	Apache 2.0	Hadoop-native мульти-протокол KV с HDFS API & S3 API
Ceph	Объект / Блок / Файл	Настраиваемая	Multi-PB, очень большие	Через Ceph RGW (RADOS Gateway)	LGPLv2.1 / Ceph Foundation	Общего назначения: RADOS (Блоки), RGW (Обьекты), CephFS (Файлы)
MinIO	Объект	Сильная	До PB	Через S3 коннекторы	AGPLv3 (SSPL-like)	Cloud-native S3 API, нет FS семантики

https://ozone.apache.org/docs/edge/concept/comparison.html

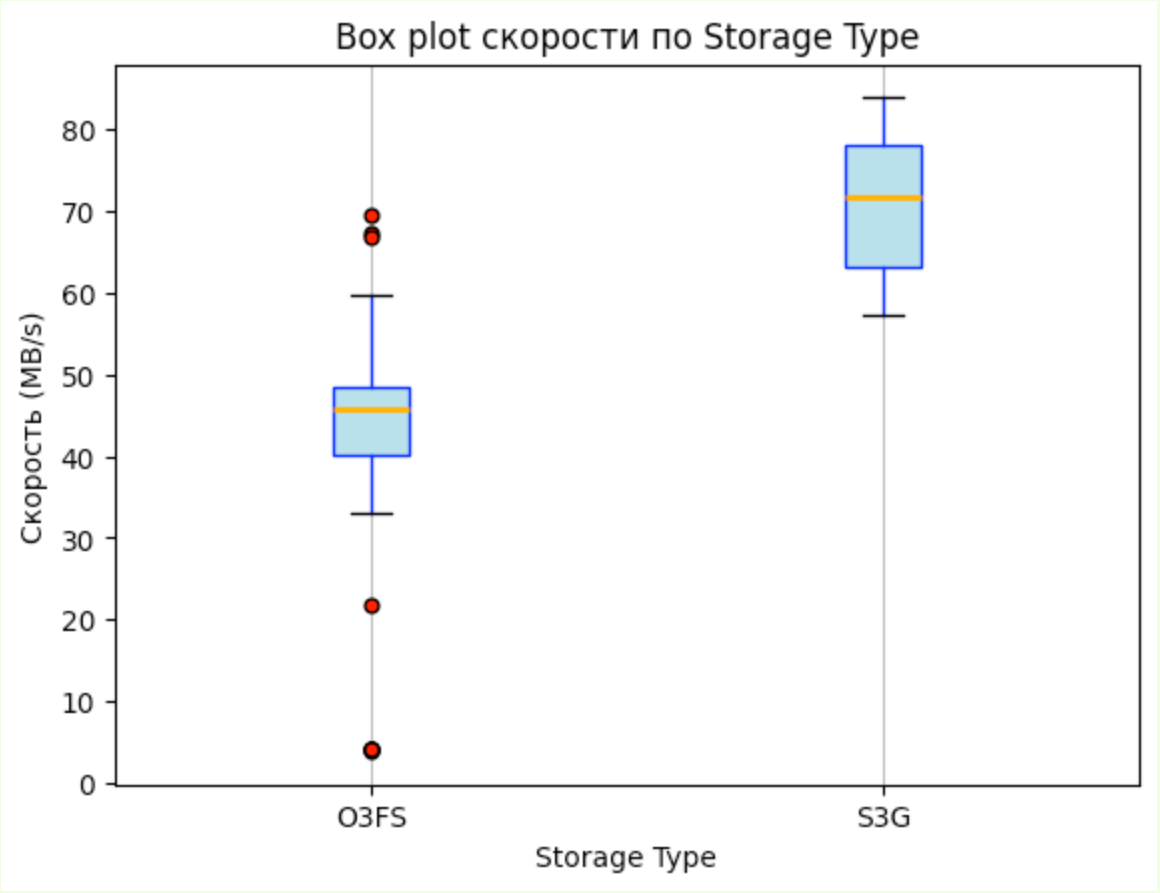
Как мы выбирали s3 решения ?

	CEPH	MINIO	Apache Ozone
Сложность освоения с нуля	Высокая	Низкая	Средняя (случае опыта с Hadoop снижается)
Сложность миграции с HDFS	Высокая	Средняя/Высокая	Средняя/Низкая (облегчается за счет поддержки Hadoop API)
сложность развертывания	Высокая	Низкая (может увеличиться в зависимости от объемов)	Средняя (случае опыта с Hadoop снижается)
сложность эксплуатации	Высокая	Низкая (может увеличиться в зависимости от объемов)	Средняя (случае опыта с Hadoop снижается)

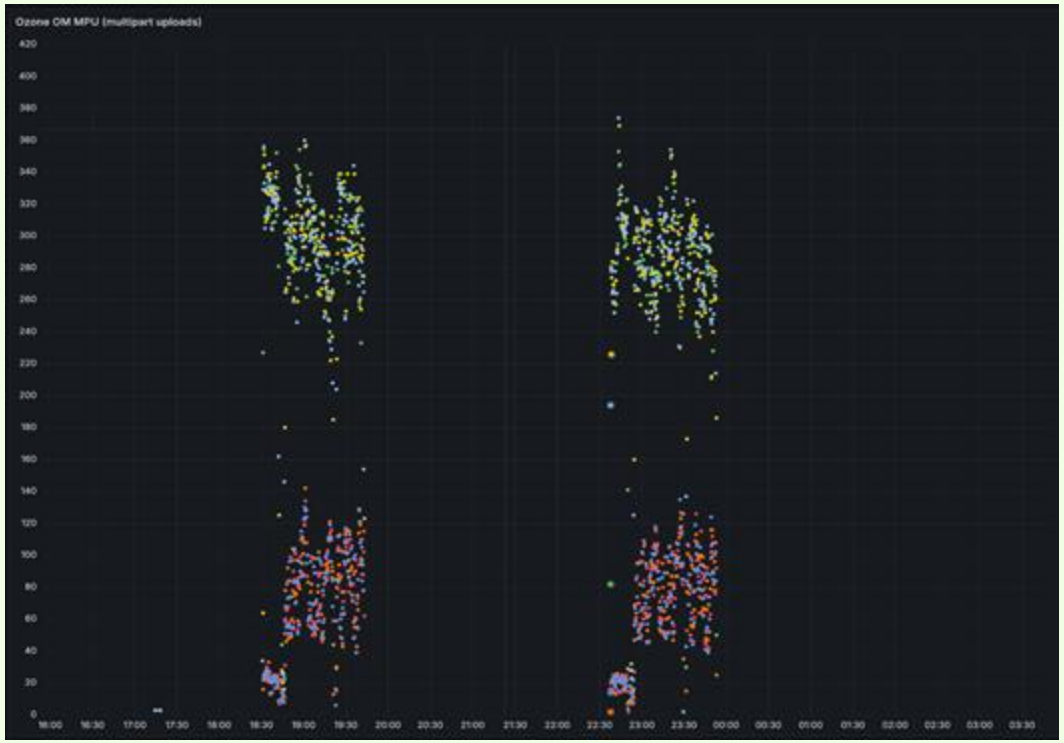
Выбрали Apache Ozone

Замеры скорости Apache Ozone

результаты теста HDFS->OzoneS3(O3FS) | AWS Cli -> OzoneS3(S3G):



скорость копирования около 80MB/s



Ozone v.1.4 (2024 год)

10Гбит/сек Общей сети

что такое Ozone?

- Масштабируемое объектное хранилище для крупномасштабных аналитических рабочих нагрузок:
- Один из Топовых проектов Apache
- Продукту 6+ лет
- Разработан с целью устранения всех недостатков и ограничений HDFS

мы будем говорить о версии v.2.0

https://docs.cloudera.com/cdp-private-cloud-base/7.3.1/ozone-overview/topics/ozone-introduction.html

компании использующие Ozone:

- Tencent
  - tencent Cloud
- Shopee
  - Practice ozone at Shopee
- DiDi
  - Using Ozone at DiDi
- Cisco
  - Cisco Data Intelligence Platform
- Cloudera
  - Cloudera Data Platform
- Hitachi
  - Hitachi Analytics Infrastructure

Элементы хранения есть в Apache Ozone

- VOLUME aka User Account)
- BUCKET aka AWS S3 Bucket
- KEY aka File on s3

какие протоколы поддерживает Ozone?

OFS (совместимо с Hadoop)
Глобальный уровень
представления ФС. ofs://

O3FS (совместимо с Hadoop)
Представление на уровне 1 bucket.
o3fs://

Протокол CSI
опция как PV для k8s

Шлюз HttpFS
реализация интерфейса,
совместимого с WebHDFS.

Протокол S3
клиенты и приложения на базе S3

Java API
API на основе RPC

Что по фичам в Ozone?

Разработчики Озон и документация Cloudera заявляют о следующем:

- поддержка высокоплотных ноды (384 вместо 100ТБ на дата ноду)
- масштабируемость метаданных
- можно хранить и картинки и разного размера файлы
- multi-tenancy s3 и bucket snapshot (PIT снимок для bucket)
- Поддержка kerberos и AWS access.keys/secret.key (для доступа к s3 не нужен keytab)
- квотирование по размеру Volume/Bucket/Key
- управление глубиной сканирования при s3 ListObject

Что в разработке для Apache Ozone?

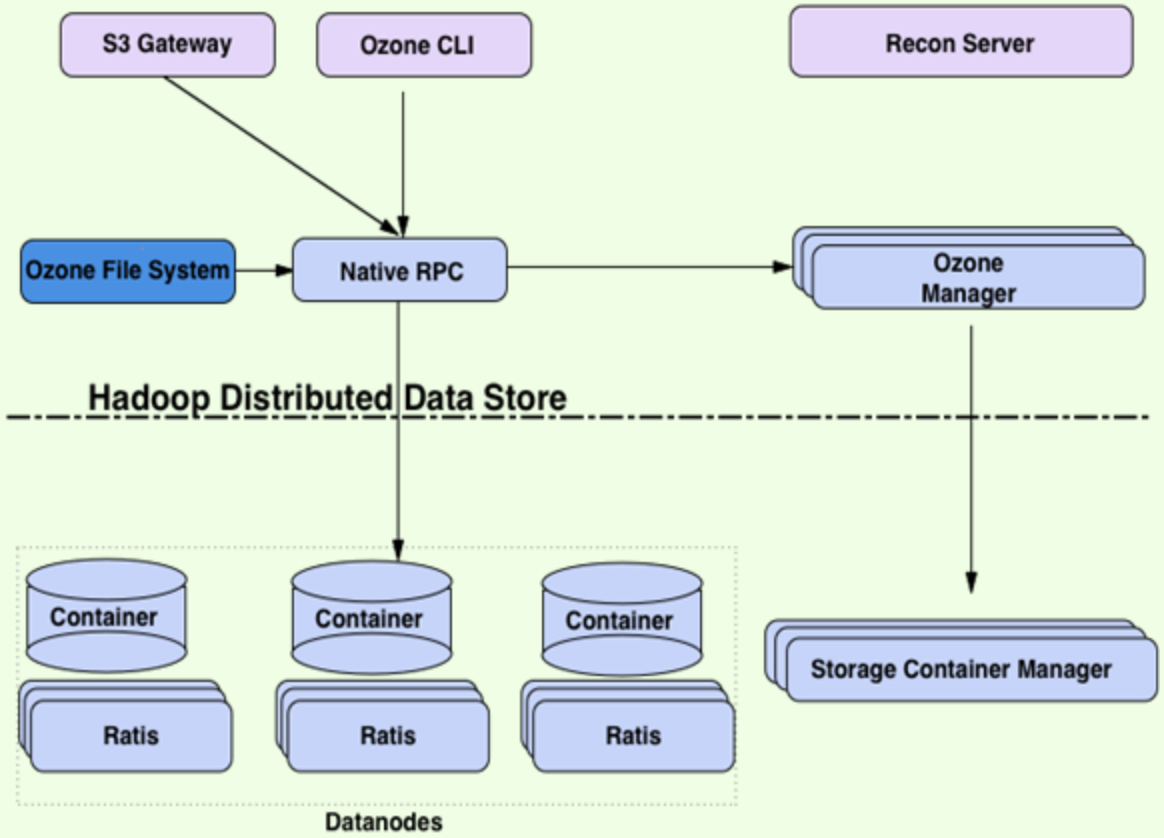
в ближайших релизах -> 2.1.0

- Disk balancer (HDDS-5713)
- S3 lifecycle management HDDS-8342

в планах

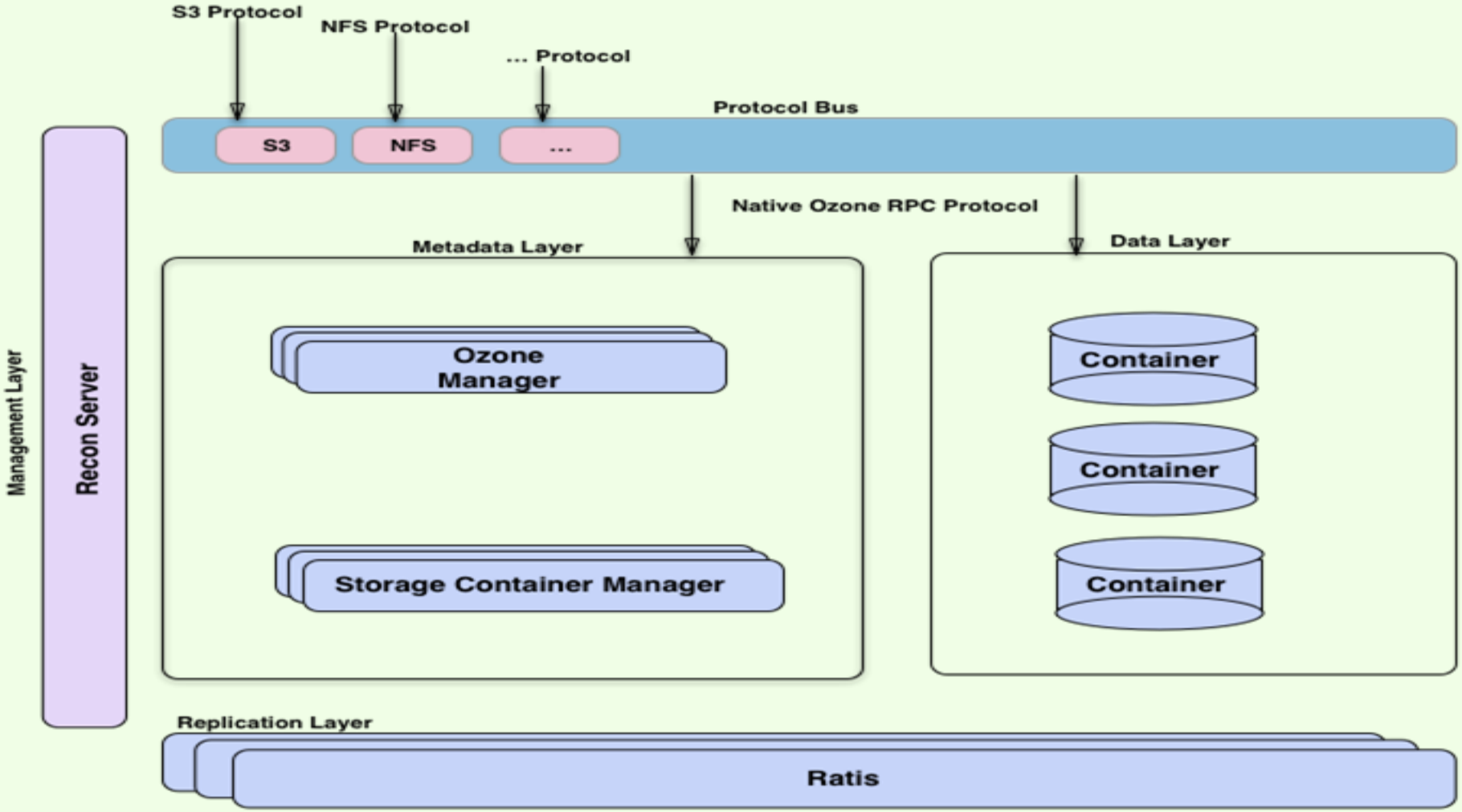
- федеративность
- Обновление на Apache Ozone с HDFS

Архитектура Ozone



https://ozone.apache.org/docs/current/concept/overview.html

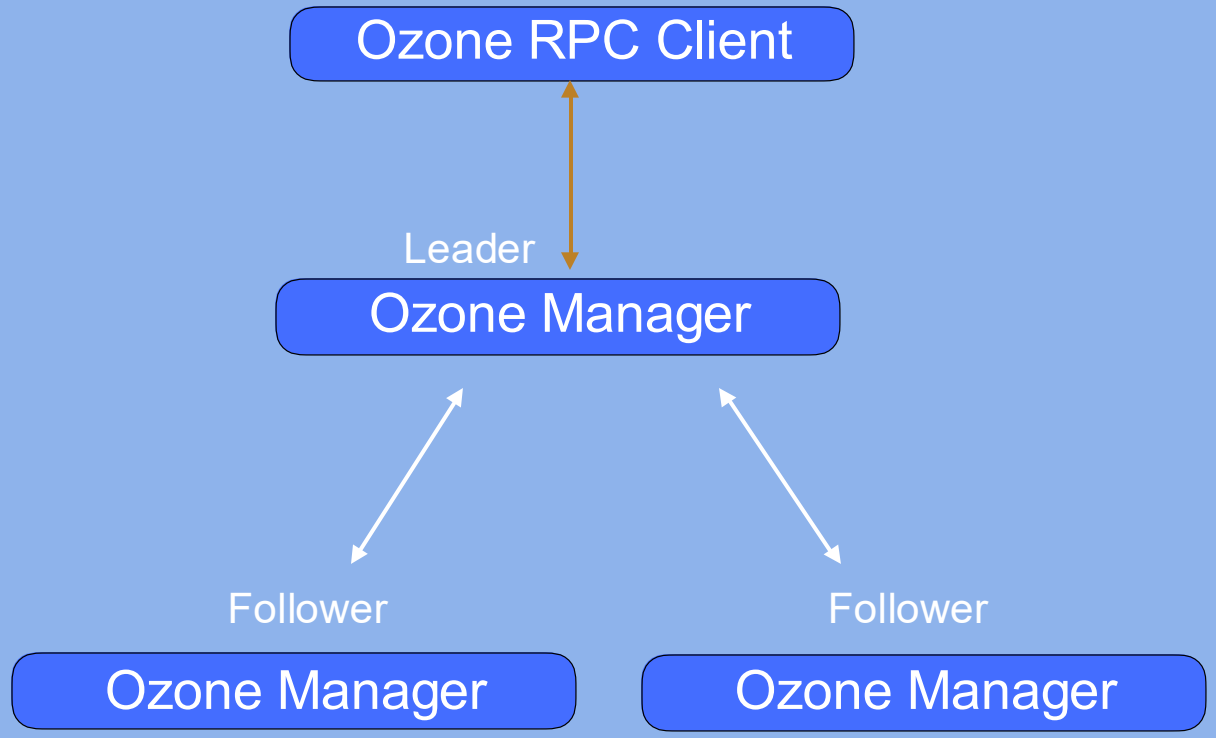
Архитектура Ozone



https://ozone.apache.org/docs/current/concept/overview.html

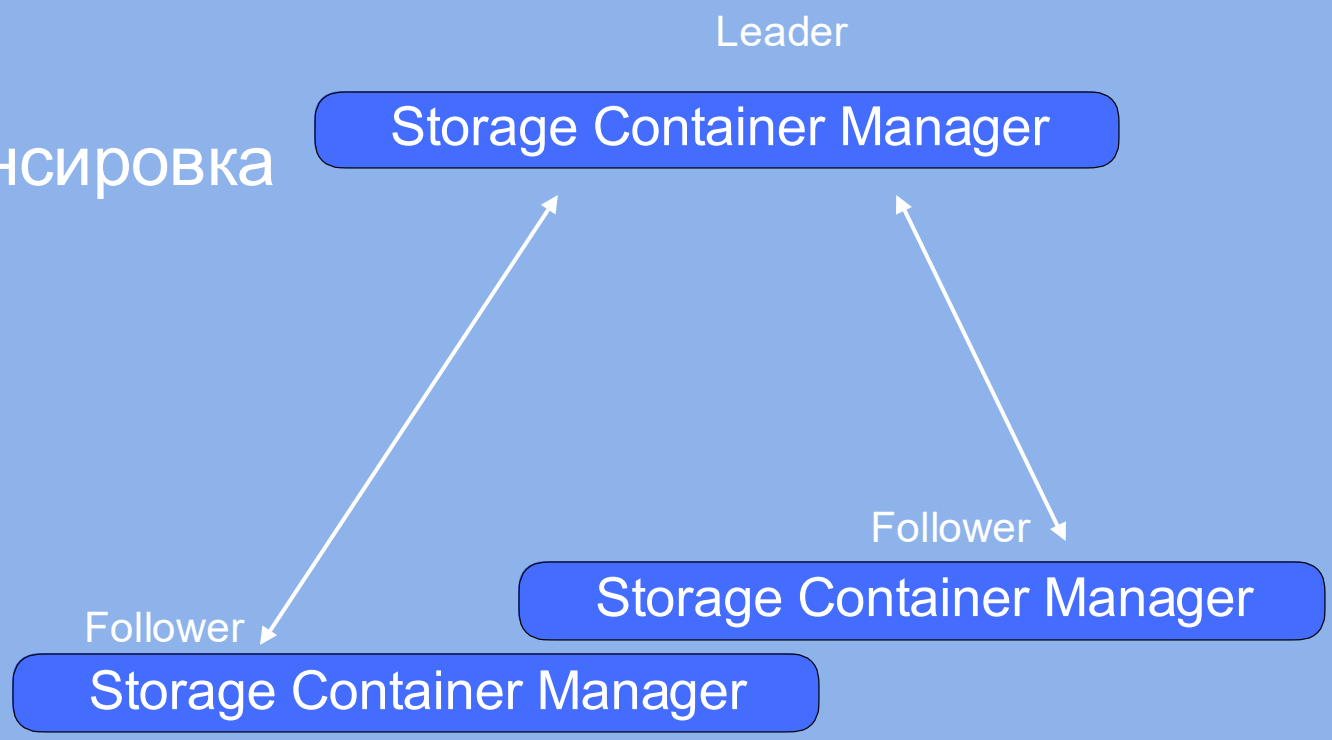
Ozone Manager (OM)

- Хранит только метаданные
- Томов, бакетов, ключей, ACL
- Данные реплицируются с помощью **Apache Ratis** (имплементация Raft протокола консенсуса)



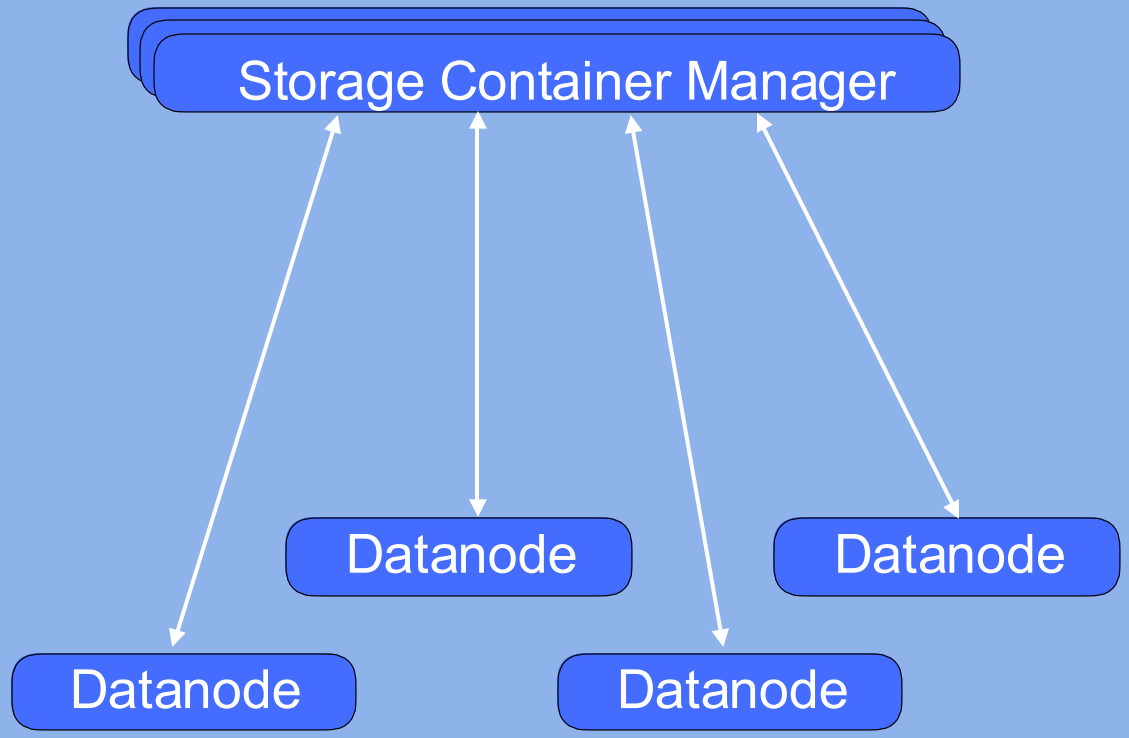
Storage Container Manager (SCM)

- Отслеживает метаданные контейнеров
- Запись, инфо о реплике, балансировка кластера, репликация, и т.д.
- Данные реплицируются с помощью **Apache Ratis**



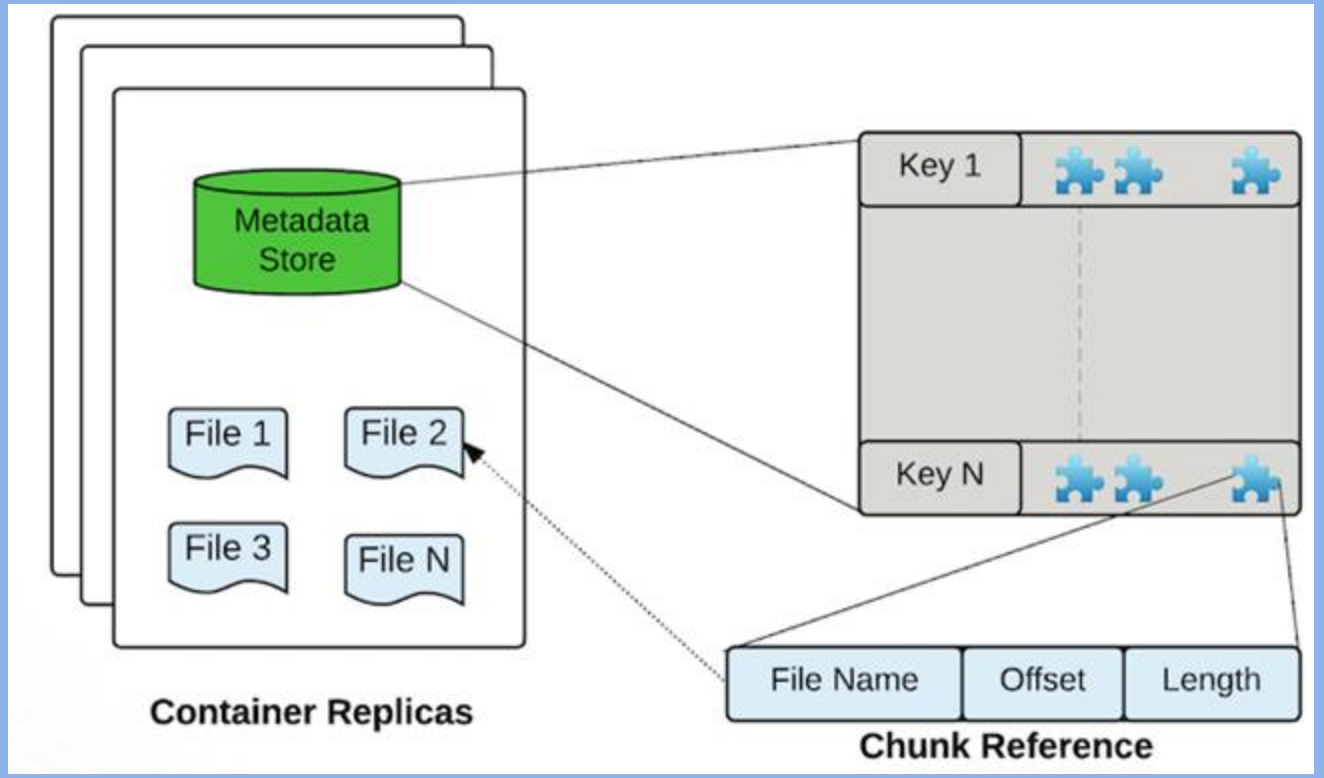
SCM + Дата Ноды

- Дата ноды периодически посылают heartbeat всем SCMs
- Лидер SCM принимает решение на основе здоровья дата ноды.
- Лидер SCM отвечает на heartbeat командами для дата нод
  - Replicate data, delete blocks, и т.д.



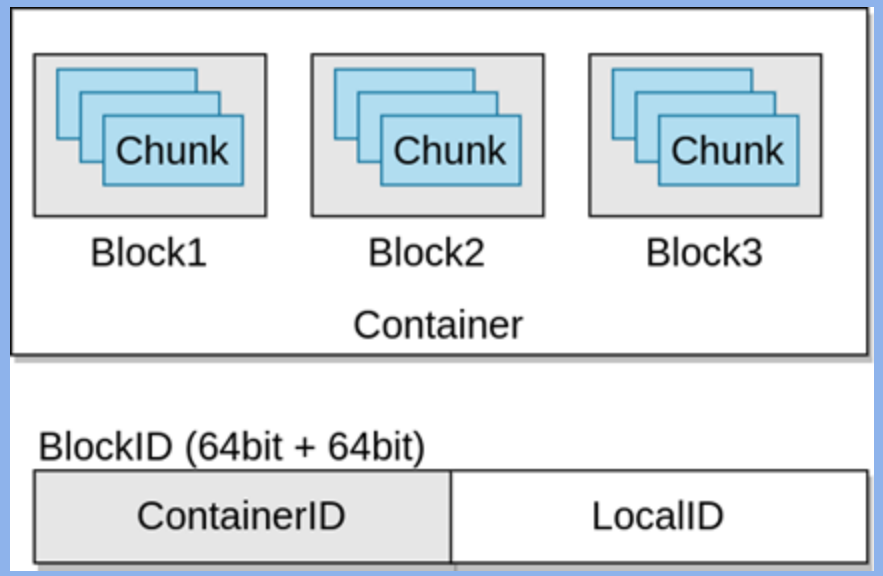
Datanodes

- Содержит несколько Томов (**volumes**)
- Каждый Том(volume) содержит **контейнеры**
- Каждый container это 5gb коллекция блоков
- RocksDB для каждого Тома(volume) хранит метаданные



Контейнеры

- это “самостоятельные суперблоки”
- являются единицей репликации и управляются через SCM
- состоит из блоков по 256 МБ
- блоки это локальная информацию и не управляется SCM. Блок содержит ID контейнера и LocalID
- контейнер бывает открытый и закрытый

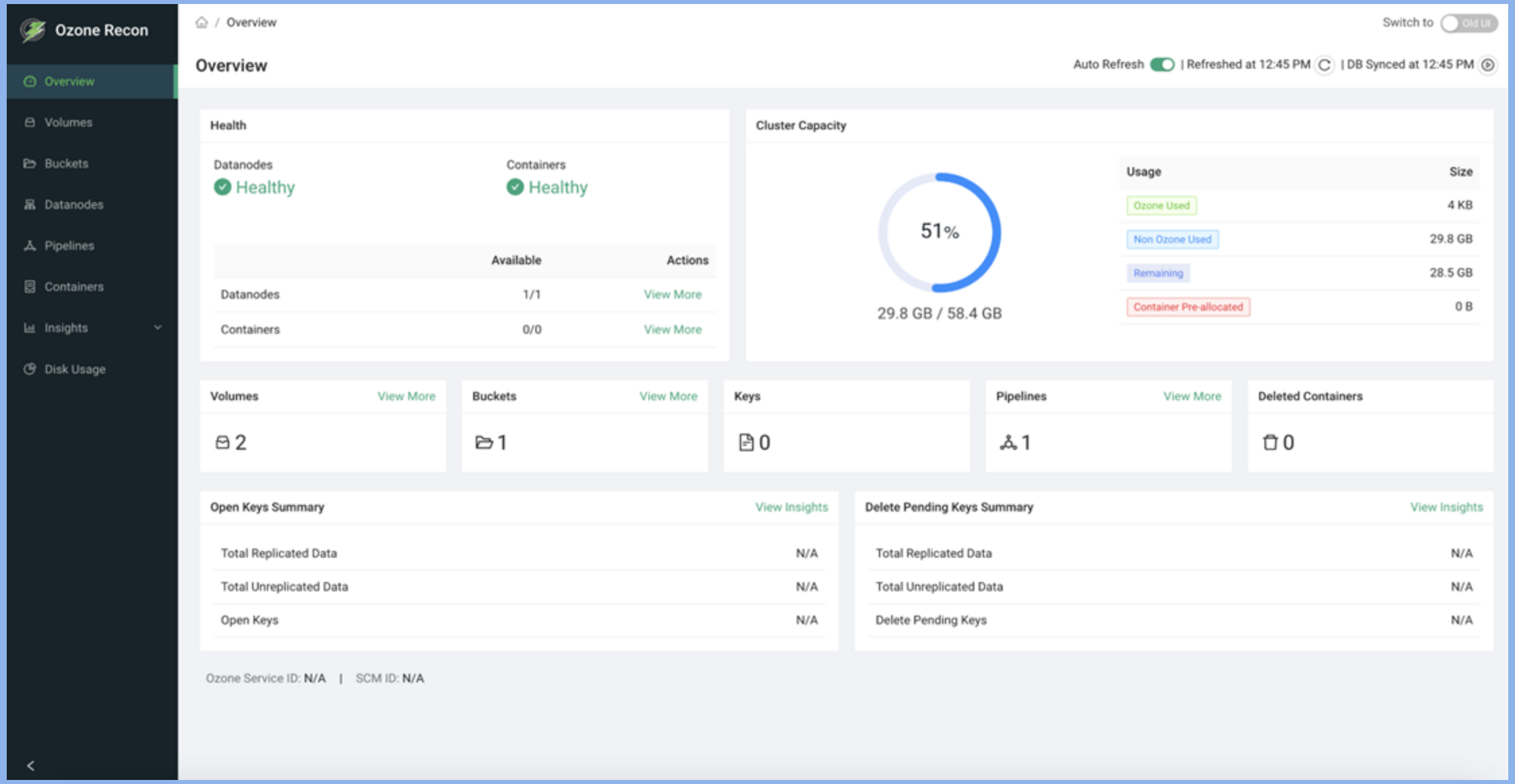


S3 Gateway

- принимаем вызовы S3 REST API от клиентов
- Транслирует их в вызовы Ozone и пересылает операции с метаданными в OM
- Передает данные напрямую в Дата ноды и обратно

https://docs.acceldata.io/odp/documentation/using-ozone-s3-gateway

Recon сервис UI



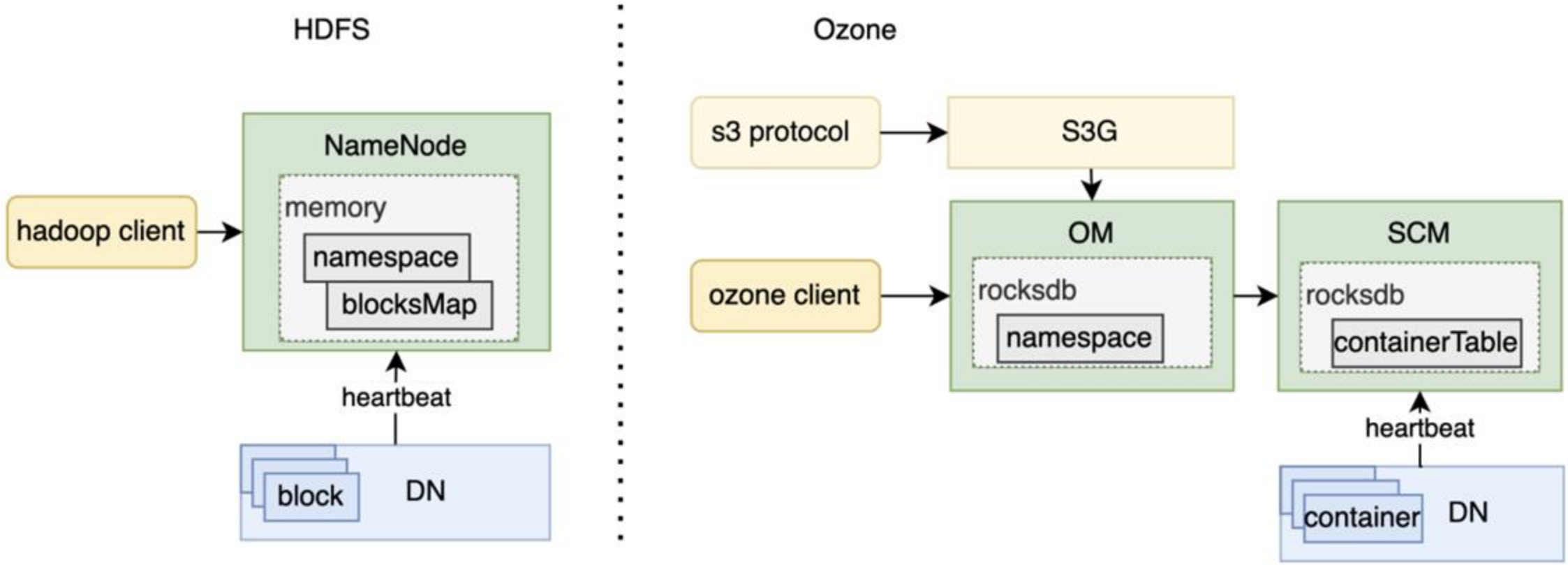
Строительные блоки Apache Ozone

- Ozone разделяет управление пространством имён и управление пространством блоков
- Пространство имён (namespace) управляется Ozone Manager (OM)
- Пространство блоков (blockspace) управляется Storage Container Manager (SCM)
- Не отслеживает отдельные блоки данных как HDFS. Вместо этого SCM отслеживает контейнеры*, в которые собираются блоки.

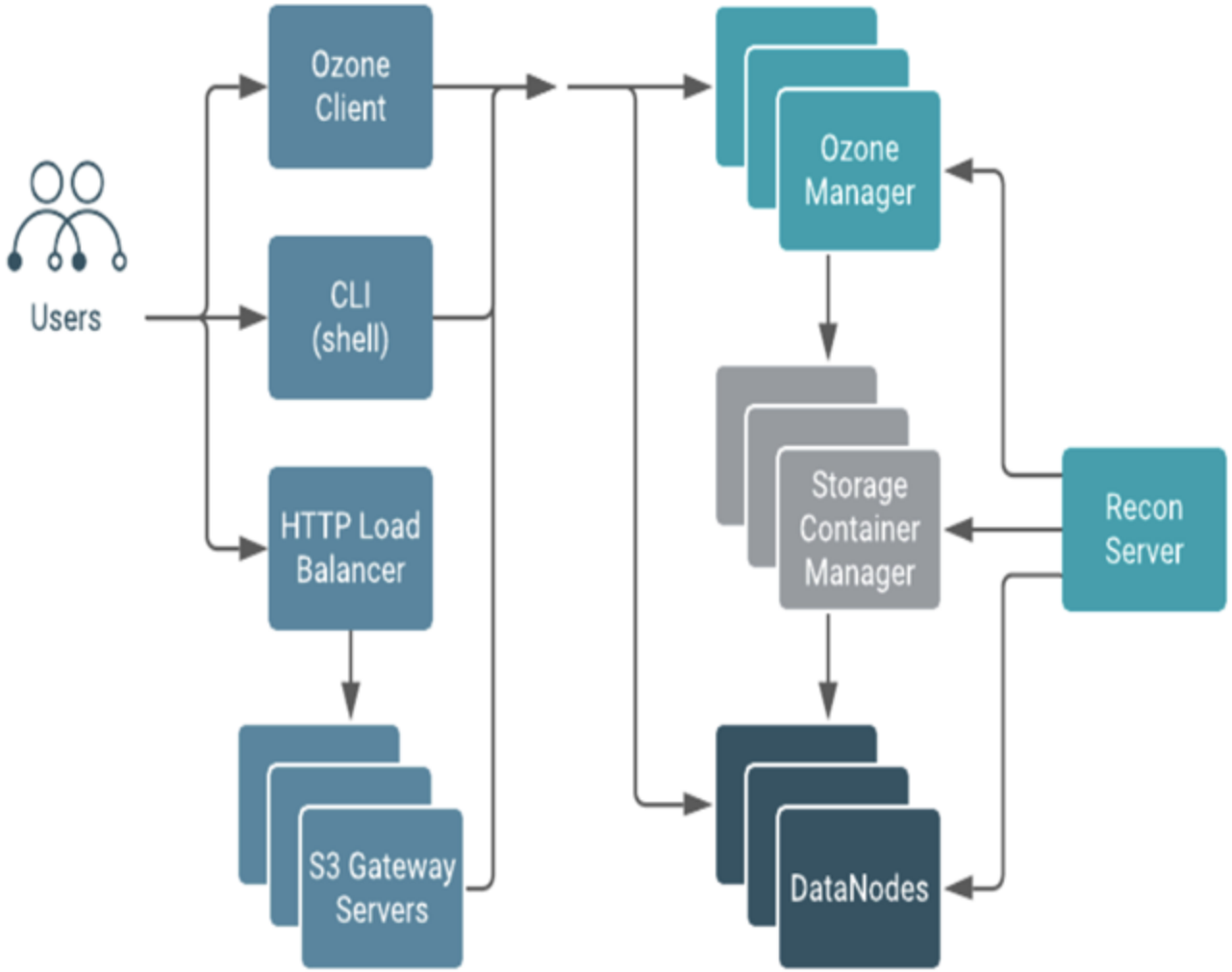
По умолчанию каждый контейнер* может иметь размер до 5 ГБ

- Благодаря независимому масштабированию пространства имён и управления блоками, Ozone может масштабироваться до сотен миллиардов файлов (ключей) в одном кластере.

Сравнение hdfs vs ozone взаимодействия



Наш Кластер Ozone сейчас



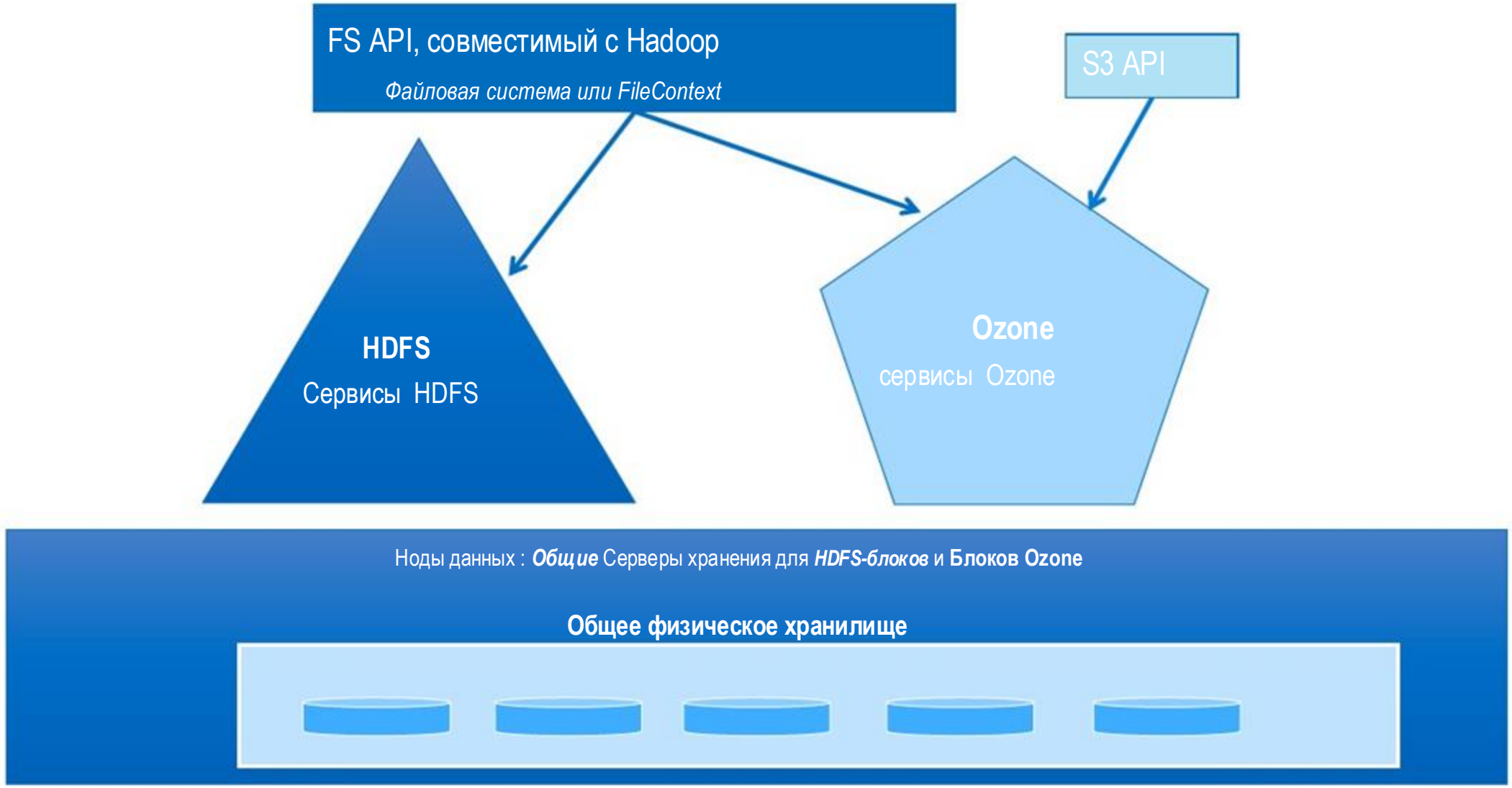
Версия Ozone: v.2.0

Сервисы:

- 3 шт OM в режиме HA
- 3 шт SCM в режиме HA
- 3 шт s3G
- 16 шт Дата Нод
- Кластер интегрирован с Ranger
- Кластер Керберизирован

Миграция данных

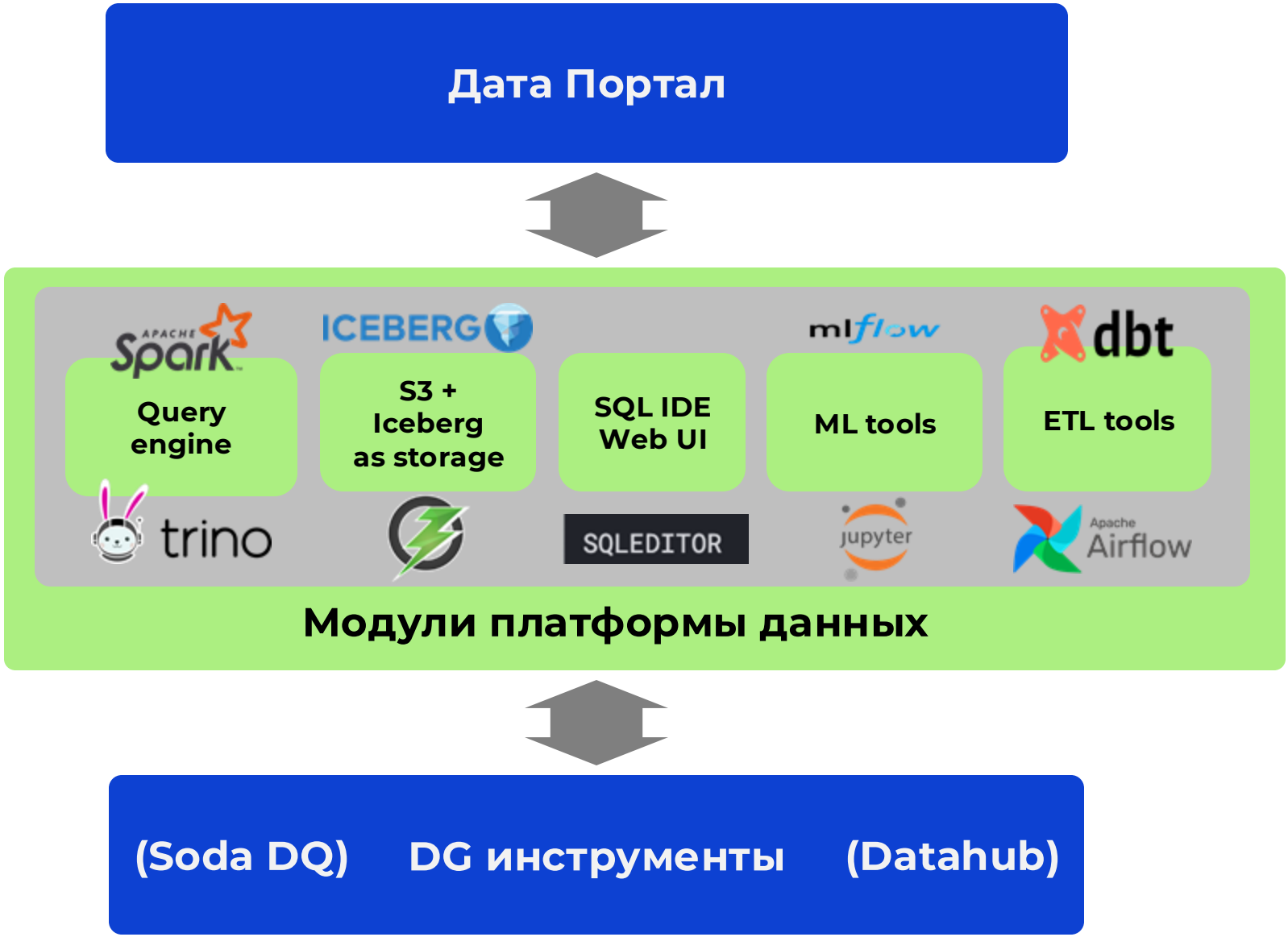
совместно используется дата ноды одновременно работающими сервисами hadoop и apache ozone



сложности встреченные на пути и выученные уроки

Ограничения керберизации уже развернутых кластеров Ozone(v.1.4 и v.2.0)	с включенным HA нет возможности керберизировать кластер, только без HA режима. перевести HA в режим без HA также нет возможности.	решение только ре-деплой кластера с миграцией данных дает эту возможность.
ресурсоемкий листинг объектов в buckets с большим кол-вом ключей	Могут Долго проходить операций типа ListObjects на buckets большим кол-вом ключей	workaround в применении ozone.s3g.list-keys.shallow.enabled=true, Ozone S3 Gateway (s3g) включает **"shallow listing"** режим для операций типа ListObjects. В этом режиме перечисление объектов возвращает только объекты, находящиеся непосредственно в текущем "каталоге", не заходя глубже (не делая глубокий обход всего дерева ключей/префиксов). S3 gateway (s3g), будет использовать выбранный режим (shallow или deep) list-операций.
ограничения при использовании Ozone s3 как PV в k8s (только Fuse over S3)	нет поддержки **sync, vsync, libaio** (Linux native asynchronous I/O), **posixaio.** так как CSI все еще в alpha версии	workaround не удалось найти , остановили исследование проблемы

платформа dota v.2: текущее состояние



Планы на будущее:

SSO на базе Keycloak с правилами Ranger для trino, spark, iceberg catalog и s3

- управление доступом к данным
- маскирование
- использование как хранилище политик

Планы на будущее:

Освоение Real-time ETL на базе Flink

- Предоставление возможности near real-time отчетности
- Повышение скорости доставки данных
- Сокращение времени подготовки витрин

Итоги

Не хадупом Единым

Вопросы

Моисеев Виталий

Ostrovok.ru