Опыт создания и миграции иерархического хранилища данных

Мирянов Сергей
ООО «РН-БашНИПИнефть»

UFADEVCONF 2023

ПЕРЕЧЕНЬ СОКРАЩЕНИЙ, ИСПОЛЬЗУЕМЫХ В ПРЕЗЕНТАЦИИ

Сокращение	Расшифровка
ACID	Atomicity Consistency Isolation Durability, Атомарность Согласованность Изоляция Надежность, гарантии предоставляемые системой управления БД
API	Application Programming Interface, интерфейс программирования приложений, программный интерфейс
ASCII	American Standard Code for Information Interchange, формат кодирования текста
B+ tree	Сбалансированное n-арное дерево поиска, с большим количеством потомков в узле
COW	Copy-on-Write, копирование при записи
GC	Garbage Collection, сборка мусора
HDF5	Hierarchical Data Format, версии 5, формат представления иерархических данных большого объема
JSON	JavaScript Object Notation, текстовый формат
Msgpack	Бинарный формат кодирования данных
MVCC	Multiversion Concurrency Control, управление параллельным доступом посредством многоверсионности
POSIX	Portable Operating System Interface, переносимый интерфейс операционных систем
SWMR	Single Writer Multiple Reader, Один Писатель, Много Читателей, режим работы с файлом HDF5
Undo/Redo	Отмена/повторение действий
UTF8	Стандарт кодирования символов, позволяющий более компактно хранить и передавать символы Юникода
WAL	Writer-ahead Log, журнал упреждающей записи
Zip	Формат архивации файлов и сжатия данных без потерь
ПО	Программное обеспечение

Список внешних программных библиотек
Blosc/Blosc-c
GZIP
h5py
libmdbx
LMDB
LZ4
LZF
Cython
numpy
OpenMP
PyTable
RocksDB
SQLite
SZIP
Xxhash/XXH3
ZStd

ПЛАТФОРМА REXLAB

- **11** лет разработки
- **200** разработчиков
- **1M** строк кода на Python, **300K** строк кода на C++
- **10** наукоемких программных продукта
- MS Windows, Linux
- Сотни пользователей
- **https://rn.digital**

- Много **кода**
- Мало **тестов**
- Мало **типов**
- Недостаточно **абстракций**
- Десктопное ПО

HDF5

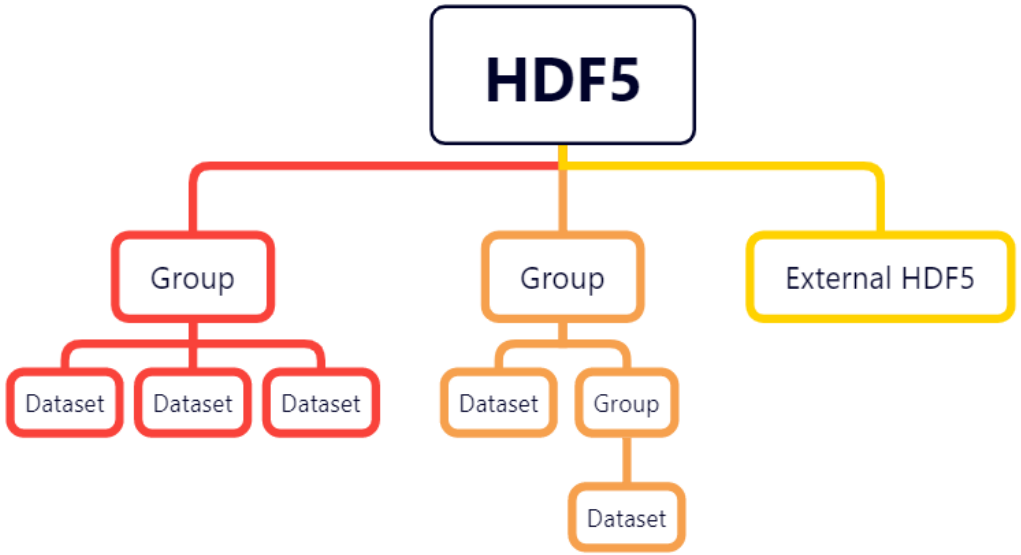
- Открытый иерархический формат хранения данных
- Гетерогенные большие данные
- Данные и мета-данные
- Срезы (Data-slicing)

- Многомерные массивы
  - Xarray, Zarr, Blosc2 NDim, …
- Датафреймы
  - Pandas, Arrow, Parquet, ORC, …
- Иерархическая файловая система

× Можно повредить при аварийном завершении

× Низкая эффективность сжатия данных

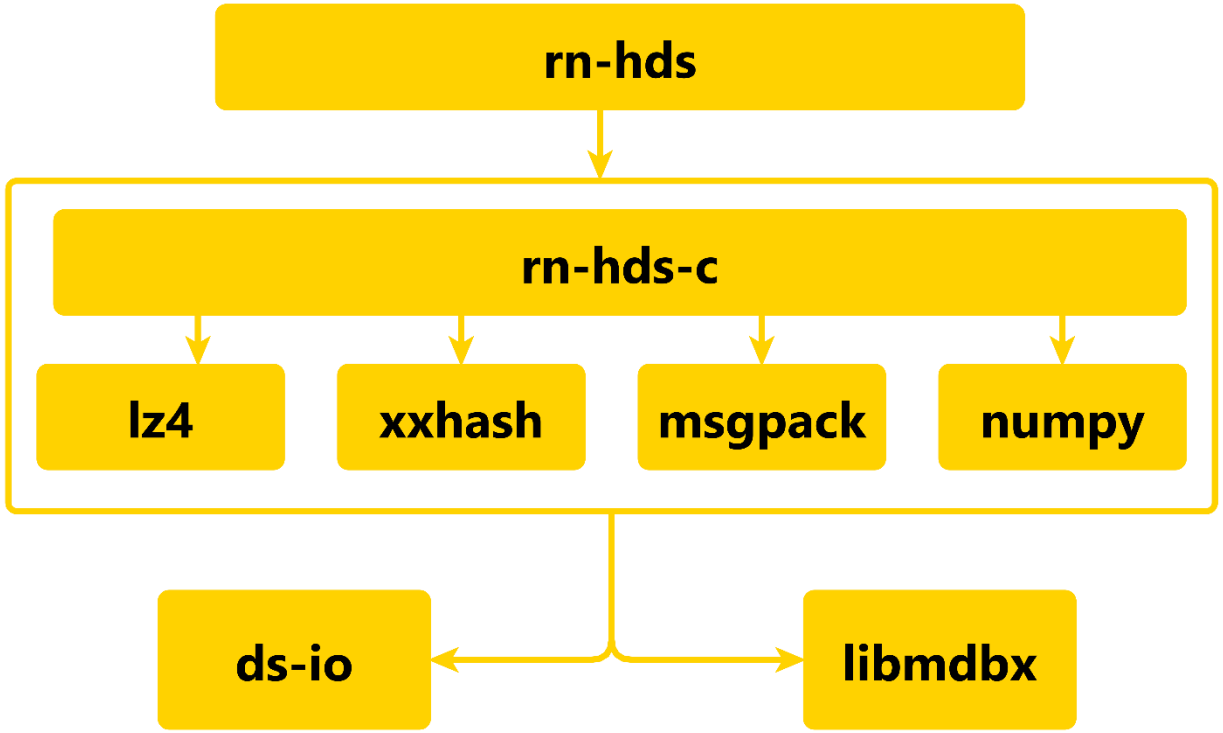
× Потеря совместимости



ПОСТАНОВКА ЗАДАЧИ

- Надежное хранение данных
- Эффективное хранение данных
- Многопроцессный доступ к данным
- Минимум изменений в платформенном коде
- Привычный интерфейс и ожидания

СТРУКТУРА РЕШЕНИЯ



Работа с многомерными массивами данных	**ds-io**
Низкоуровневое хранилище в формате «ключ-значение»	**libmdbx**
Управление деревом узлов, атрибутами, датасетами	**rn-hds-c**
H5py-like интерфейс	**rn-hds**

LIBMDBX

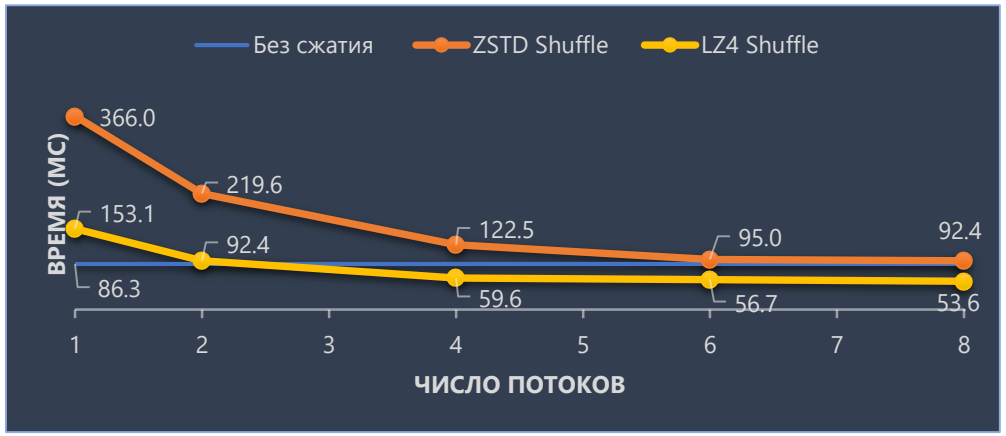
- Провели сравнение
  - ✓ **HDF5**
  - JSON
  - SQLite
  - ✓ **RocksDB**
  - Filesystem
  - Zip
  - ✓ **LMDB**
  - ➢ **libmdbx**

+ mmap
+ B+ tree
+ ACID, MVCC, **No WAL**
+ Изоляция read/write транзакций
+ **Многопроцессный доступ**
+ **Управление размером БД**
± COW + **Shadow Paging**
- Только одна пишущая транзакция
- Файловая блокировка при открытии пишущей транзакции

РАБОТА С ДАТАСЕТАМИ

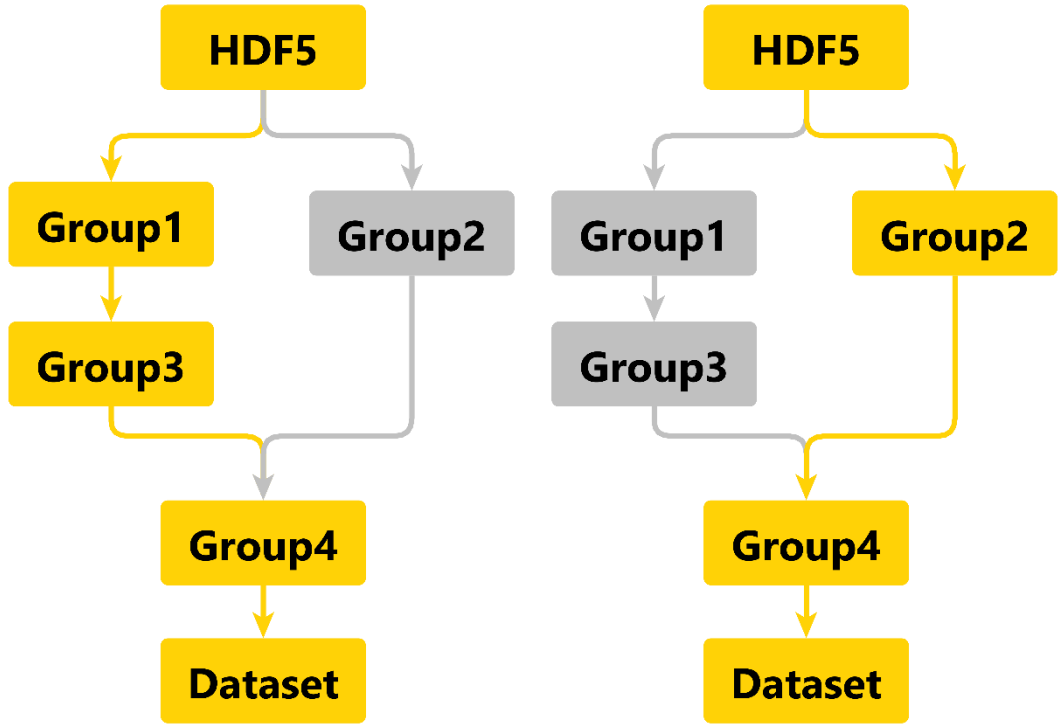
- rn-hds-c
  - COW
  - Дедупликация (XXH3)
- ds-io
  - OpenMP
  - LZ4
  - ZStd
  - Blosc (byte shuffle)

	HDF5 (comp=0)	HDF5 (comp=5)	HDF5 (comp=5, shuffle)	Blosc2 (zstd, shuffle)	ds-io (zstd)	ds-io (zstd, shuffle)
Время записи (мсек.)	141	5 000	2 283	344	124	**106**
Время чтения (мсек.)	88,3	695	471	105	**66,2**	74,4
Размер файлов (Кбайт)	202 190	105 763	59 561	58 159	60 025	**38 288**
Степень сжатия (раз)	1	3,14	4,62	4,71	5,96	**7,39**



РАБОТА С ДЕРЕВОМ

- Компактный формат
- COW
- Дедупликация (XXH3)
  - Ссылки на атрибуты
  - Ссылки на датасеты
- Msgpack
- Отложенная загрузка



КАКИЕ БЫЛИ СЛОЖНОСТИ

1. Прототип
   - Python + libmdbx + Json

- Нет актуальных биндингов
  - Обновили
  - Поправили баги
- Не поддерживает numpy
- На основе cython
  - Нет нормальной поддержки float16
  - Можно быстрее
- Json
  - Медленно

КАКИЕ БЫЛИ СЛОЖНОСТИ

1. Прототип
   - Python + libmdbx + Json
2. Прототип 3
   - Python + rn_libmdbx
     - libmdbx
     - msgpack

- Заменили Json на Msgpack
  - Быстрый
  - Простой API
- Работа с атрибутами за один вызов
- Стало сравнимо с h5py

КАКИЕ БЫЛИ СЛОЖНОСТИ

1. Прототип
   - Python + libmdbx + Json
2. Прототип 3
   - Python + rn_libmdbx
     - libmdbx
     - Msgpack
3. Работа с удаленными объектами

- HDF5 позволяет работать с объектами, у которых нет ни одного родителя
  - Отложенное удаление
  - Копится мусор

КАКИЕ БЫЛИ СЛОЖНОСТИ

1. Прототип
   - Python + libmdbx + Json
2. Прототип 3
   - Python + rn_libmdbx
     - libmdbx
     - Msgpack
3. Работа с удаленными объектами
4. Слабая типизация h5py

- H5py + numpy
  - В датасет можно положить почти что угодно
  - Своеобразная поддержка ASCII и UTF8
    - np.dtype metadata
- Много тестов
  - Мало тестов
- Hyrum's Law

КАКИЕ БЫЛИ СЛОЖНОСТИ

1. Прототип
   - Python + libmdbx + Json
2. Прототип 3
   - Python + rn_libmdbx
3. Работа с удаленными объектами
4. Слабая типизация h5py
5. Дорогие пишущие транзакции

- Пакетный режим
  - Ослабление гарантий
  - Снижение накладных расходов
    - Sync
    - GC

СРАВНЕНИЕ РЕЗУЛЬТАТОВ

Операция	Время, мсек		Разница, раз	Время, мсек		Разница, раз
	P1	P2		S1	S2	
Открытие проекта	127 824	27 249	4,69	175 483	114 677	1,53
Открытие файла	116 146	15 747	7,38	102 534	42 127	2,43
Сохранение файла	147 462	80	1 835,13	117 742	38	3 119,20
Закрытие проекта	2 801	1 265	2,21	4 599	2 649	1,74
Загрузка списка скважин	17 832	7 392	2,41	50 897	20 361	2,50
Загрузка списка каротажа	8 110	438	18,52	6 741	106	63,75
Статистика по каротажу	16 405	5 074	3,23	36 010	11 007	3,27
Расчет 3D каркаса	50 899	41 668	1,22	57 213	44 901	1,27
Размер	Мб		Разница, раз	Мб		Разница, раз
	41 861	7 220	5,77	14 422	3 473	4,15

ИТОГО

- Получили надежное хранилище
- С возможностью писать в него из нескольких процессов
- Экономию занимаемого пространства
- Выигрыш по времени

‼ Но

- Libmdbx сложная технология
- Python не подходит для интенсивной нагрузки
- Древовидный API медленный
- Сборка мусора
  - Libmdbx
  - Атрибуты и датасеты
  - Узлы
  - Python

БОНУС

- Журнал изменений
  - Undo/Redo
- Контрольные точки
  - Восстановление в случае сбоев
  - Фиксация логических изменений
  - Компактификация изменений

‼ Но
- Сложная логика на клиенте


РОСНЕФТЬ
СПАСИБО ЗА ВНИМАНИЕ!

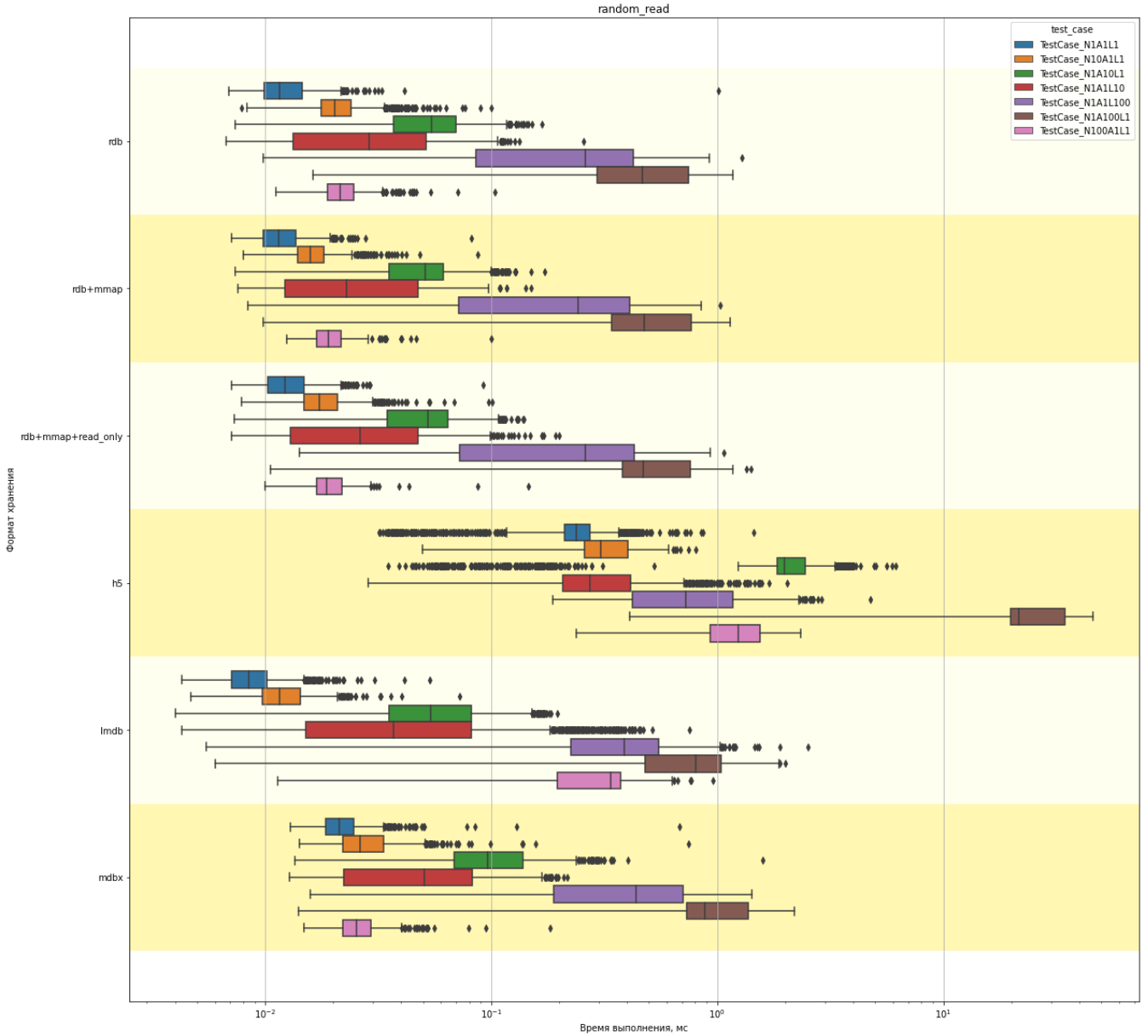
РОСНЕФТЬ

ООО «РН-БашНИПИнефть»

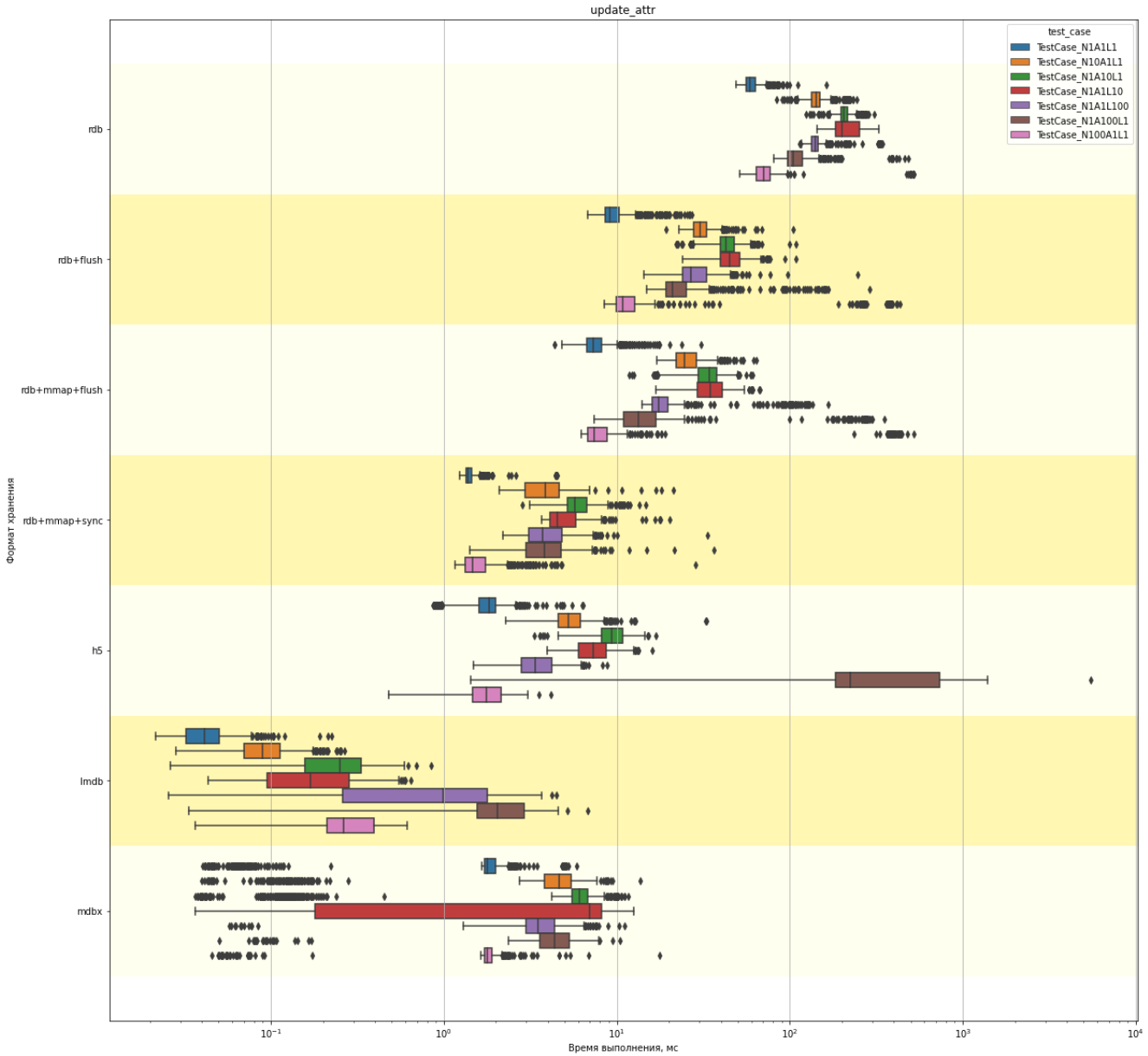
По всем возникающим вопросам просьба обращаться к
Мирянову Сергею Николаевичу

по адресу электронной почты: **SN_Mirianov@bnipi.rosneft.ru**

ЧТЕНИЕ АТРИБУТОВ



ЗАПИСЬ АТРИБУТОВ



СОЗДАНИЕ УЗЛОВ