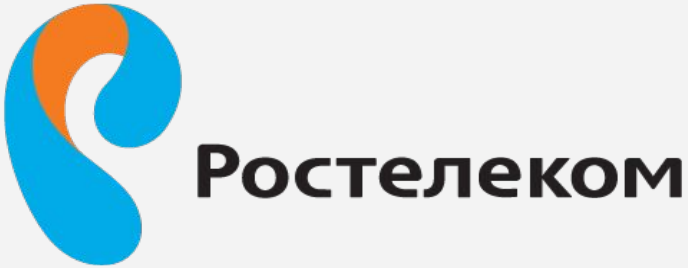
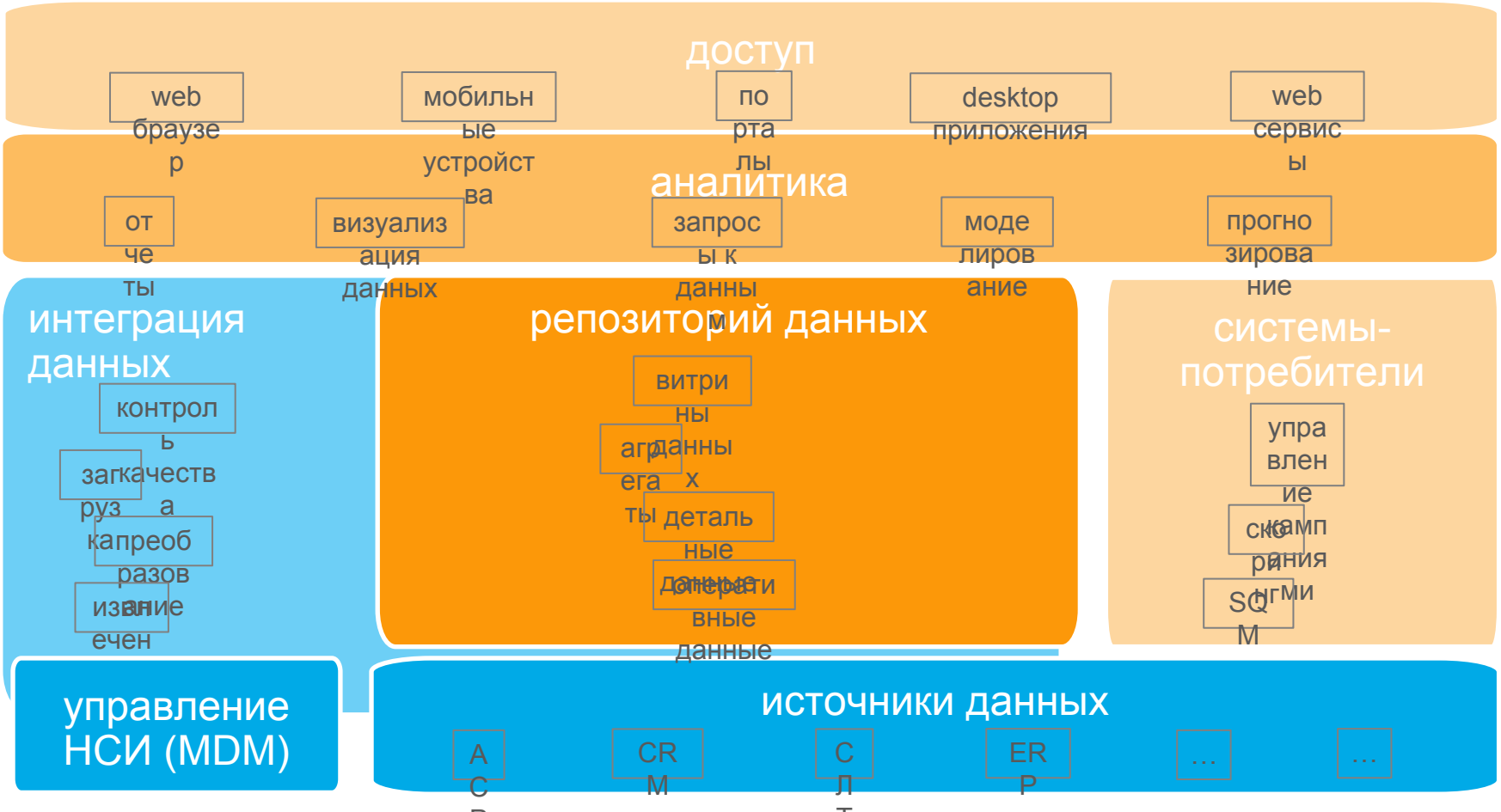
Борис Емельянов

PGDAY'17 RUSSIA
КОНФЕРЕНЦИЯ ПО БАЗАМ ДАННЫХ

Опыт использования GP в Ростелекоме



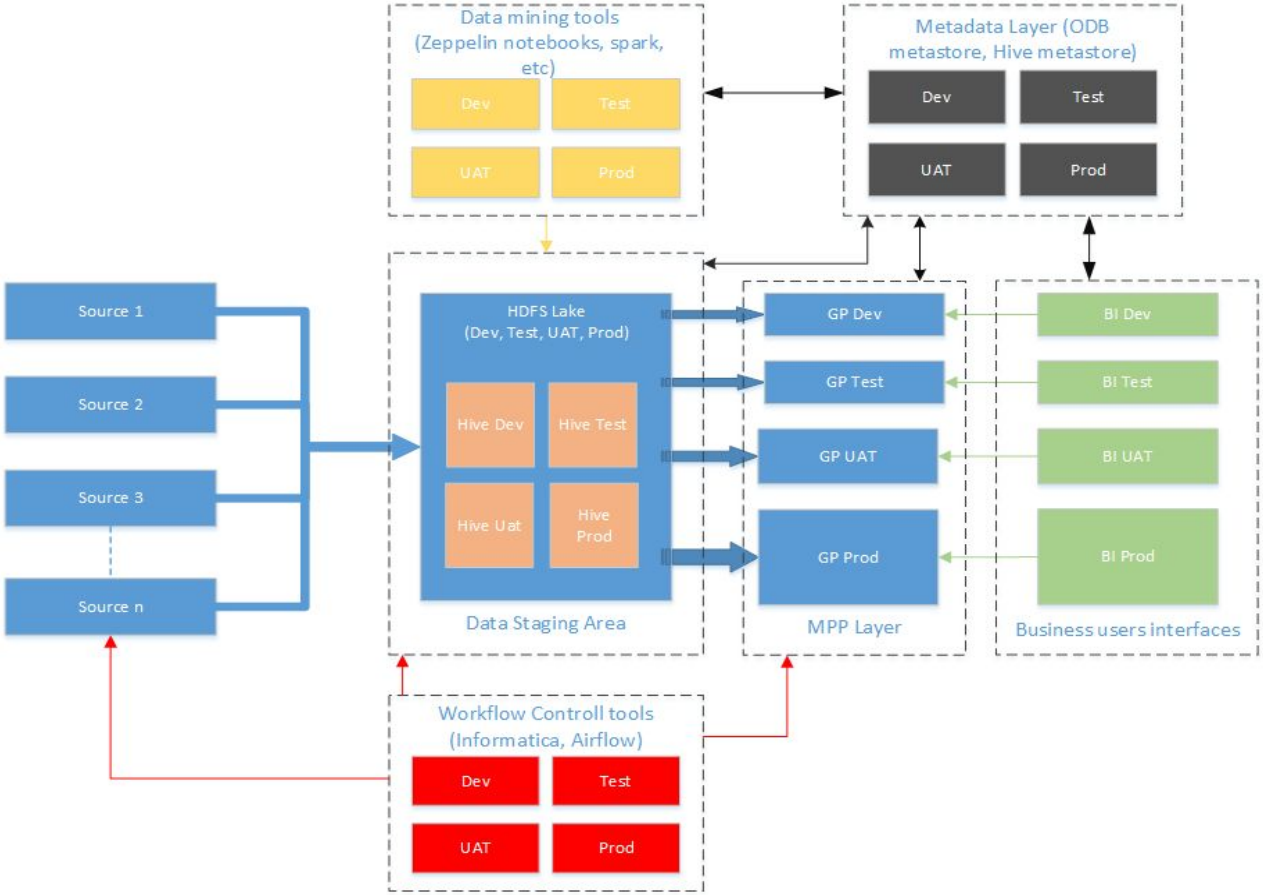
Основные компоненты



Выбор технологии

- Appliance (HW vendor lock, price, scaling)
  - Teradata
  - Exadata
  - Netezza
- Software
  - Greenplum (postgresql)
  - Vertica

Высокоуровневая архитектура



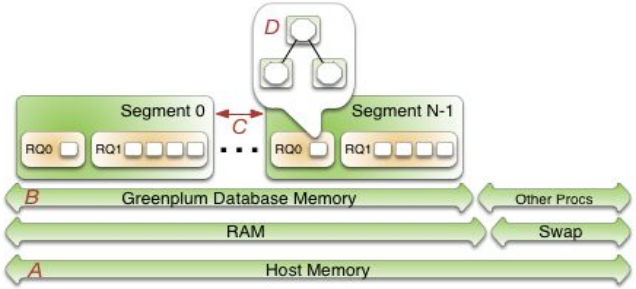
Размер инсталляции

- 60 систем источников
- более 6000 таблиц источников
- 250 ГБ ежедневной выгрузки
- 80 ТБ данных в GP
- 14 сегмент-серверов, 256 ГБ ОЗУ, 10 ТБ полезной емкости на сервер

Установка и настройка

- Автоматизирована с помощью rpm + puppet
- 3 primary segmets per host (преобладает большое количество относительно мелких запросов)
- nofiles → 524288
- max_connections → 400
- max_appendonly_tables → 20000
- max_locks_per_transaction → 256
- max_resource_queues → 18

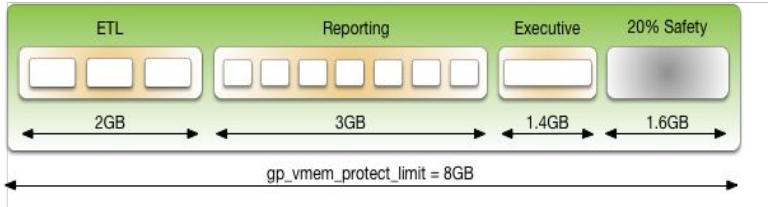
Управление нагрузкой



- http://greenplum.org/calc/
- ((SWAP + (RAM * vm.overcommit_ratio / 100)) / max_number_segments_per_server_with_mirror_failure
- gp_vmem_protect_limit → 65536

- Сложности разделяемой среды
- Без оверкоммита ресурсы, чаще всего, утилизированы не полностью
-
- Использование spread mirrors стратегии

Очереди



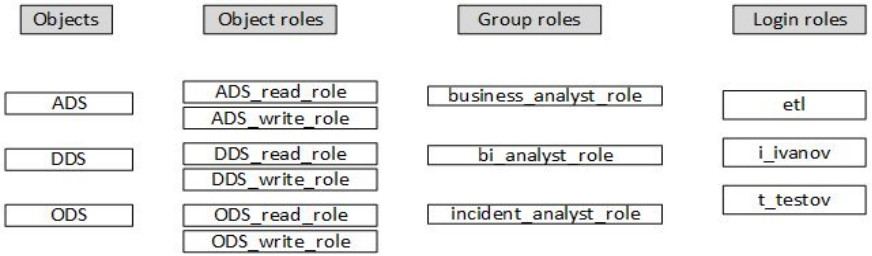
- ACTIVE_STATEMENTS
- MEMORY_LIMIT
- MAX_COST
- MIN_COST - я только спросить! Проблема с функциями.

- Принцип пользователь-очередь
- Стараться распределять запросы типу операций, а не согласно административной структуре
- gp_toolkit.gp_resqueue_status
- Очереди сейчас: ETL, BI, dev, Analitycs.
- В планах: динамически перестраивать очереди, выделение отдельных очередей для тяжелых и легких ETL процедур, Запуск дополнительных сред.

Борьба со спиллами

- Выяснить, почему это происходит
  - Читаем EXPLAIN, EXPLAIN ANALYZE
  - Используется ли партиционирование?
  - Перекошенные таблицы в запросе
- Увеличивать память на зарос: память очереди, statement_mem, max_statement_mem
- Ограничить размер и количество файлов (gp_workfile_limit_per_segment, gp_workfile_limit_files_per_query)
- Если от спиллов не избавиться
  - Компрессия (gp_workfile_compress_algorithm)
  - Запись промежуточных результатов во временную таблицу с высокой компрессией

Модель прав доступа



- Postgresql 8.2 – нужно выдавать права на каждый объект отдельно
- Минимальная единица – схема
- DDL скрипты + автогрантер

```
FOR rec IN
SELECT 'grant select on ' || nsp.nspname || '.' || cls.relname || ' to ' || replace(nsp.nspname, 'edw', current_database()) ||
'_read_role ; ' grt
FROM pg_catalog.pg_class cls
INNER JOIN pg_catalog.pg_namespace nsp ON nsp.oid = cls.relnamespace
LEFT JOIN pg_inherits inh ON inh.inhrelid = cls.oid
WHERE nsp.nspname LIKE 'edw%'
  AND cls.relkind IN ('r','v','S')
  AND inh.inhrelid IS NULL
  AND NOT EXISTS
    ( SELECT 1
     FROM pg_catalog.pg_class cls1
     INNER JOIN pg_catalog.pg_namespace nsp1 ON nsp1.oid = cls1.relnamespace
     WHERE nsp1.nspname LIKE 'edw%'
       AND array_to_string(cls1.relacl,',') LIKE '%' || replace(nsp1.nspname, 'edw', current_database()) || '_read_role' || '=r%'
       AND cls1.relkind IN ('r','v','S')
       AND cls.relname = cls1.relname
       AND nsp.nspname = nsp1.nspname )
LOOP
EXECUTE rec.grt ;
```

Heap vs AO

- AO – compression!
- AO – можно бэкапить инкрементально
- В нашем случае запросы по сжатым AO таблицам работают быстрее
- AO – создает дополнительные таблицы (и файлы), что увеличивает размер каталога:
  - pg_aoseg_<oid>
  - pg_aoseg_<oid>_index
  - pg_aovisimap_<oid>
  - pg_aovisimap_<oid>_index
-
- AO скорее подходит для больших фактовых таблиц (а не всех подряд!)

Мониторинг

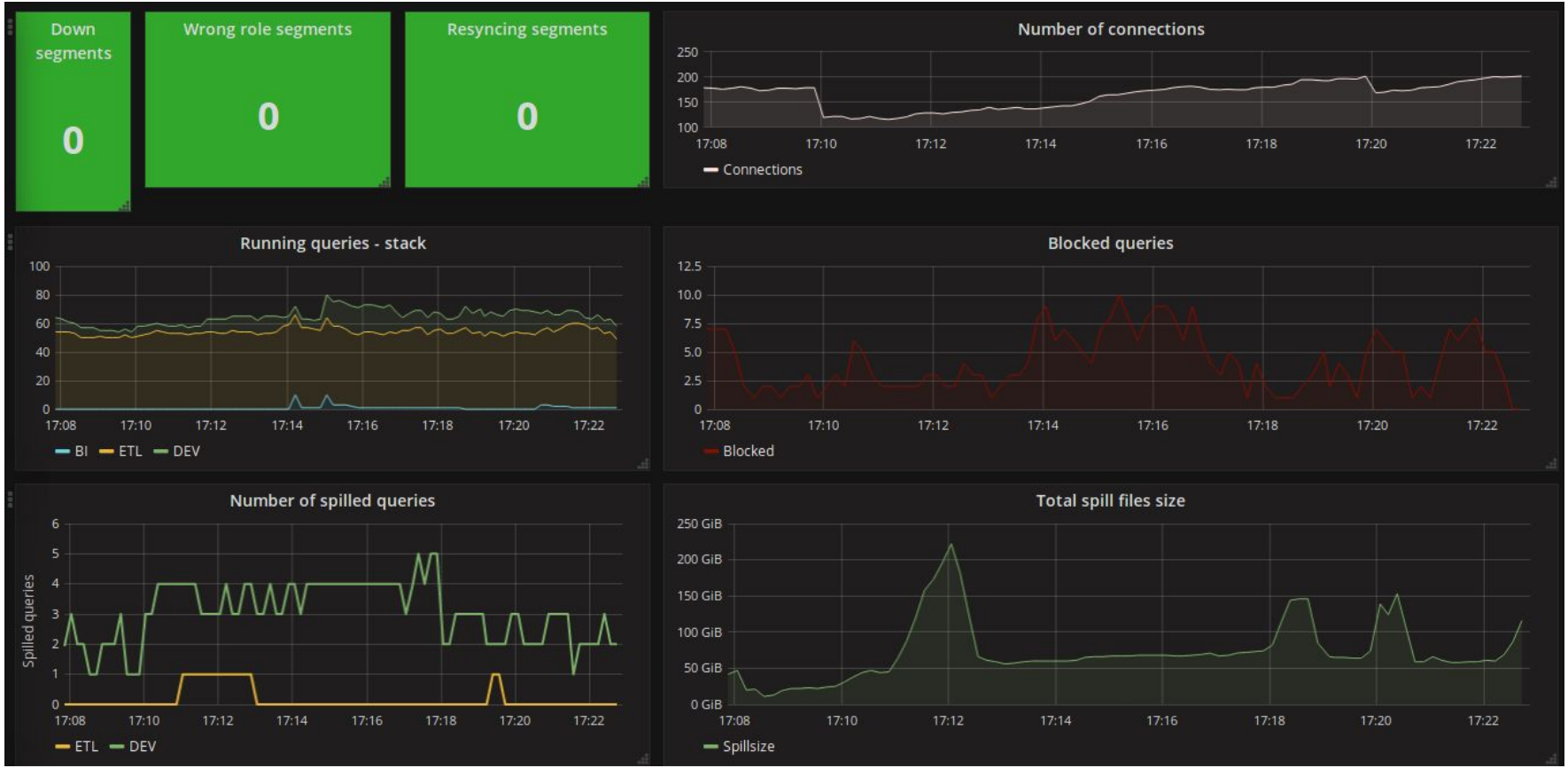
Что мониторим

- Общие метрики ОС
- Состояние сегментов
- Подключения, активные запросы
- Заблокированные запросы
- Состояние очередей
- Спилл-файлы: количество запросов, объем спилла
- Статистика по запросам

Инструментарий

- GP command center – из коробки
- Gpperfmon database
- Zabbix – alerting
- Ganglia – hadoop legacy
- Moving to Telegraf+InfluxDB + grafana
-

Мониторинг


Down segments
0
Wrong role segments
0
Resyncing segments
0
Number of connections
250
200
150
100
17:08
17:10
17:12
17:14
17:16
17:18
17:20
17:22
Connections
Running queries - stack
100
80
60
40
20
0
BI
ETL
DEV
Blocked queries
12.5
10.0
7.5
5.0
2.5
0
Blocked
Number of spilled queries
Spilled queries
6
5
4
3
2
1
0
ETL
DEV
Total spill files size
250 GiB
200 GiB
150 GiB
100 GiB
50 GiB
0 GiB
Spillsize

Регулярное обслуживание

- VACUUM – max_fsm_pages, max_fsm_relations
- gp_toolkit.gp_bloat_diag
- Время от времени проверять на скрытые поля AO таблиц
-

```
SELECT nspname,
       relname,
       count(DISTINCT (comp_inf).datafile) AS file_qty,
       round(sum((comp_inf).hidden_tupcount)/sum((comp_inf).total_tupcount),2) AS percent_hidden,
       sum((comp_inf).total_tupcount) AS total_records
FROM
  (SELECT relname,
          nspname,
          gp_toolkit.__gp_aovisimap_compaction_info(cls.oid) AS comp_inf
   FROM pg_class cls
   JOIN pg_appendonly ao ON ao.relid = cls.oid
   JOIN pg_namespace ns ON ns.oid = cls.relnamespace
   WHERE nspname LIKE '%edw%' ) t
GROUP BY nspname, relname
HAVING sum((comp_inf).total_tupcount) > 0
AND round(sum((comp_inf).hidden_tupcount)/sum((comp_inf).total_tupcount),2) > 0.1
ORDER BY total_records DESC;
```

Регулярное обслуживание

- Отключение старых сессий select pg_terminate_backend(procpid) from pg_stat_activity where current_query='<IDLE>' and usename not in (...) and cast(extract(epoch from now()) - extract(epoch from backend_start) as int) > ${TIME}
- ANALYZE – analyzedb многопоточный, учитывает состояние таблицы, для AO работает инкрементально
- Analyzedb хранит рабочие файлы на мастере, нужно чистить
- Проверка на наличие перекоса в таблицах
- Быстрый способ: по размеру сторфайлов таблицы на сегментах: https://discuss.pivotal.io/hc/en-us/articles/204407723(http://www.pivotalguru.com/?p=519)
- Удаление временных схем (иногда gpcheckcat)
- Частый VACUUM, ANALYZE, REINDEX системного каталога
-

Резервное копирование

- Инструмент из коробки – gpcrondump
- Создает локальные дамп-файлы параллельно на всех сегментах
- Можно использовать запись в named pipes, https://discuss.pivotal.io/hc/en-us/articles/203694696-How-to-use-gpcrondump-with-named-pipes
- https://discuss.pivotal.io/hc/en-us/articles/218815537-Greenplum-How-To-Troubleshoot-Long-Running-and-Hanging-Backups
- Возможные альтернативы:
  - Gptransfer
  - HDFS writable external tables
-

Детские ошибки

- Правильный выбор ключа распределения
- Неуказание ключа распределения ведет к distributed randomly = переезд всех данных по сети
- Неправильный выбор ключа распределения (либо правильный выбор, но неверное наполнение)
- Гибкая стратегия партиционирования
- Сбалансированное использование AO/Heap таблиц
- Мы прошли путь от “нигде” до “везде”
- Регулярное обслуживание
-

PGDAY'17 RUSSIA

Спасибо!

Борис Емельянов
b.emelyanov@rt.ru
b.emelyanov@yandex.ru