Igor Putyatin
Pivotal Data Engineering

PGDAY'17 RUSSIA
КОНФЕРЕНЦИЯ ПО БАЗАМ ДАННЫХ

Greenplum Best Practices

Содержание

- Обзор Greenplum
- Оптимизация физической модели данных
- Управление памятью
- Как ускорить Greenplum?
- Мониторинг использования ресурсов и отладка запросов

Обзор Pivotal Greenplum

Greenplum – **массивно параллельная** СУБД,
обладающая **линейной масштабируемостью**;
применяется в mission critical системам, работающих с большими данными.

Основное назначение:

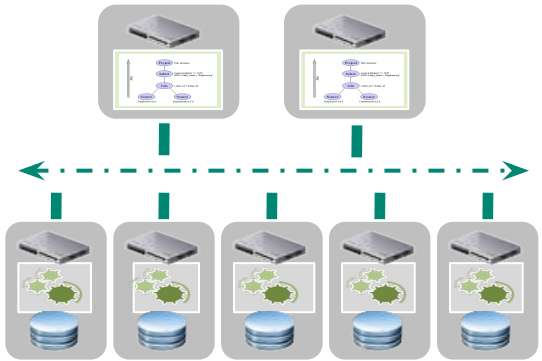
- анализ данных
- загрузка и хранение структурированных данных
- работа с Business Intelligence приложениями
- использование алгоритмов машинного обучения.

Pivotal Greenplum : Массивно-параллельная архитектура

PGDAY'17 RUSSIA

Массивно-параллельная обработка данных

Хранилище данных Greenplum



Аналитические инструменты СУБД

Встроенные библиотеки

MADlib, Apache Solr, GPText, PostGIS Geospatial Objects for PostgreSQL

Программирование

python, R, Perl, Java, THE C PROGRAMMING LANGUAGE

Pivotal Greenplum

Массивно-параллельная загрузка данных из внешних источников

Хранилища данных в экосистеме Hadoop

Hortonworks, cloudera, MAPR

Облачные хранилища данных

S3, Google Cloud Platform, Azure

Pivotal

Механизм индексации и поиска текстов

GPText

- Реализован на основе Apache Solr
- 5 лет успешной эксплуатации
- Интеграция с Madlib для реализации машинного обучения на текстовых данных

Применение

- Мониторинг корпоративных коммуникаций
- Исследование настроения клиентов
- Поиск по хранилищу документов
- Обработка данных из соцсетей, и многое другое

Apache Solr

Библиотека функций data science

Generalized Linear Models
- Linear Regression
- Logistic Regression
- Multinomial Logistic Regression
- Ordinal Regression
- Cox Proportional Hazards Regression
- Elastic Net Regularization
- Robust Variance (Huber-White), Clustered Variance, Marginal Effects

Utility Modules
Array and Matrix Operations
Sparse Vectors
Random Sampling
Probability Functions
Data Preparation
PMML Export
Conjugate Gradient
Stemming
Sessionization
Pivot
Path Functions
Encoding Categorical Variables

Other Machine Learning Algorithms
- Principal Component Analysis (PCA)
- Association Rules (Apriori)
- Topic Modeling (Parallel LDA)
- Decision Trees
- Random Forest
- Conditional Random Field (CRF)
- Clustering (K-means)
- Cross Validation
- Naïve Bayes
- Support Vector Machines (SVM)
- Prediction Metrics
- K-Nearest Neighbors

Matrix Factorization
- Singular Value Decomposition (SVD)
- Low Rank

Linear Systems
- Sparse and Dense Solvers
- Linear Algebra

Descriptive Statistics
Sketch-Based Estimators
- CountMin (Cormode-Muth.)
- FM (Flajolet-Martin)
- MFV (Most Frequent Values)

Correlation and Covariance
Summary

Inferential Statistics
Hypothesis Tests

Time Series
- ARIMA

Graph
- PageRank
- Single Source Shortest Path

Инсталляции Greenplum

Greenplum используется в крупных банках, финансовых компаниях, телекоммуникационных компаниях, страховых компаниях, госорганах и других организациях в качестве основного хранилища данных и аналитической платформы

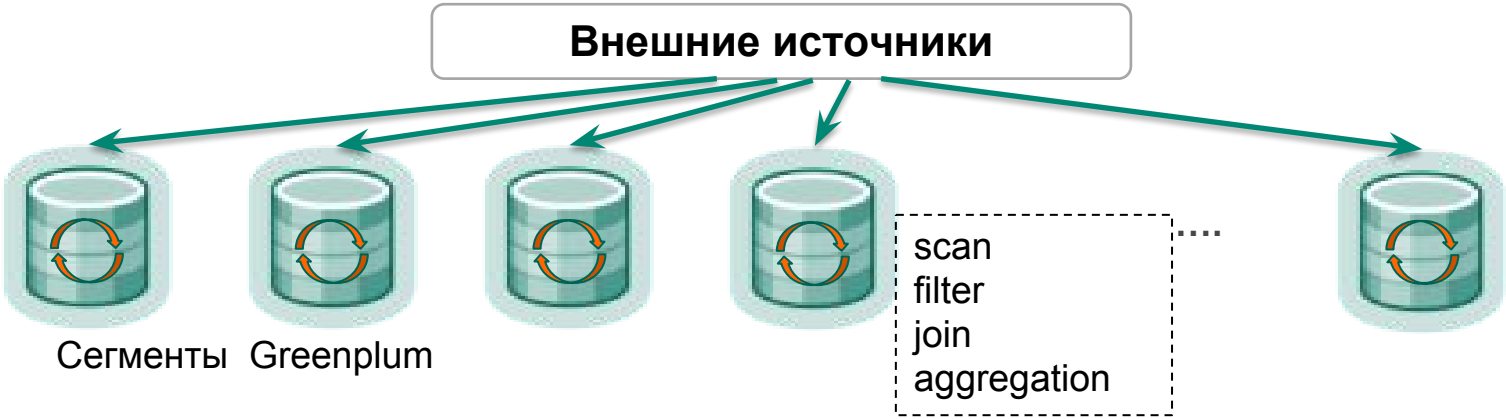
На данный момент самый крупный работающий в продуктиве кластер Greenplum состоит из **128** сегментных хостов.

- Суммарная скорость чтения данных с диска – **407 Гб/сек**
- Суммарная скорость записи – **386 Гб/сек**
- Максимальный объем данных без учёта сжатия – **1,5 петабайт**

Назначение Greenplum

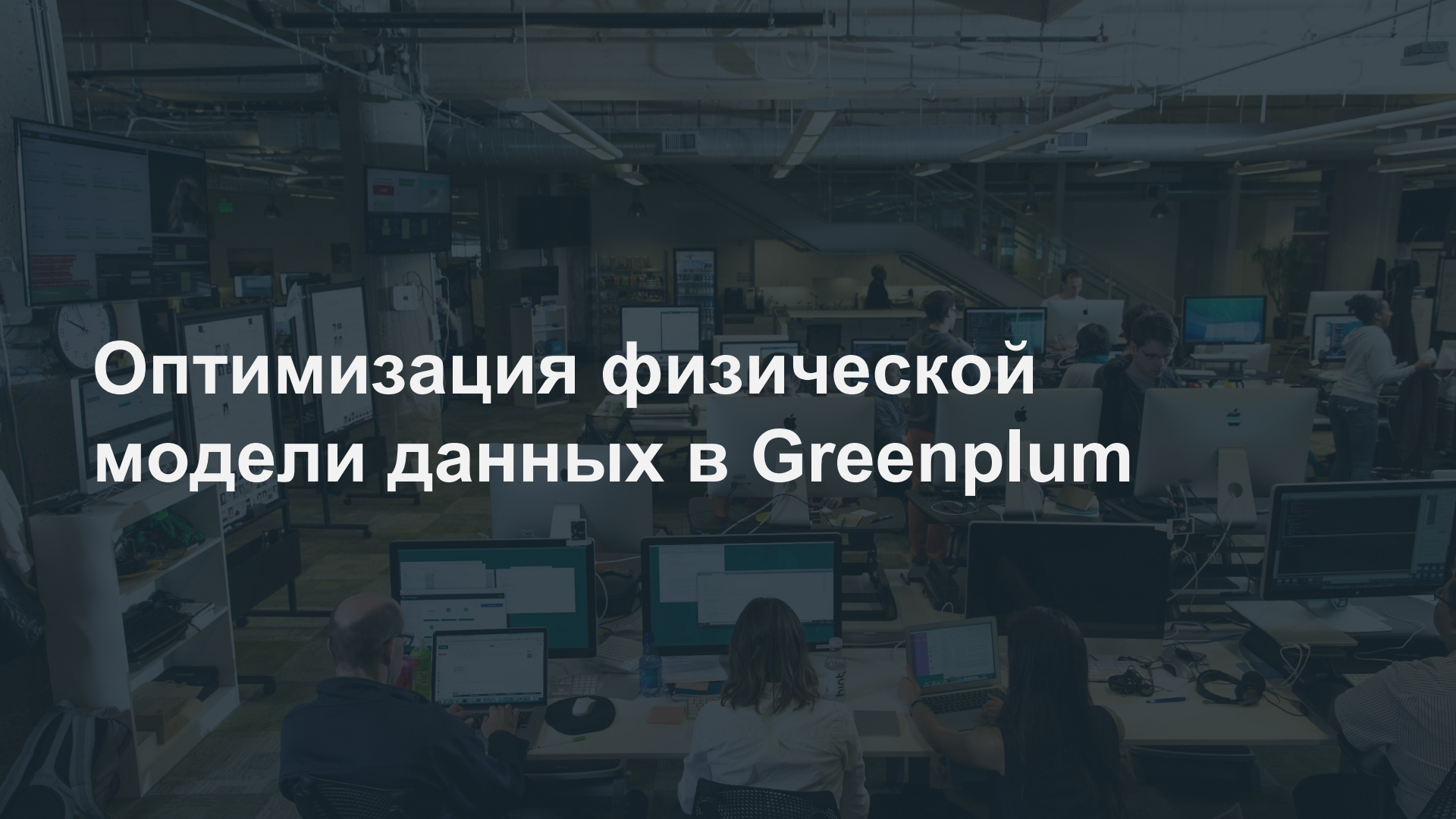
Greenplum работает максимально эффективно на тех операциях, которые можно распараллелить

- Full Scan таблицы
- Агрегация
- Соединение таблиц (join)
- Параллельная загрузка данных из внешних источников



Примеры неправильного использования Greenplum

- Подключение клиентских приложений, работающих в режиме реального времени и часто (несколько раз в секунду) запрашивающие данные из Greenplum. Данные для такого рода приложений должны готовиться в Greenplum и выгружаться в in-memory решение, например Pivotal Gemfire.
- Построчная загрузка данных из источников в реальном времени
- Реализация итеративных алгоритмов с построчной обработкой данных
- Реализация итеративных алгоритмов с DDL операциями



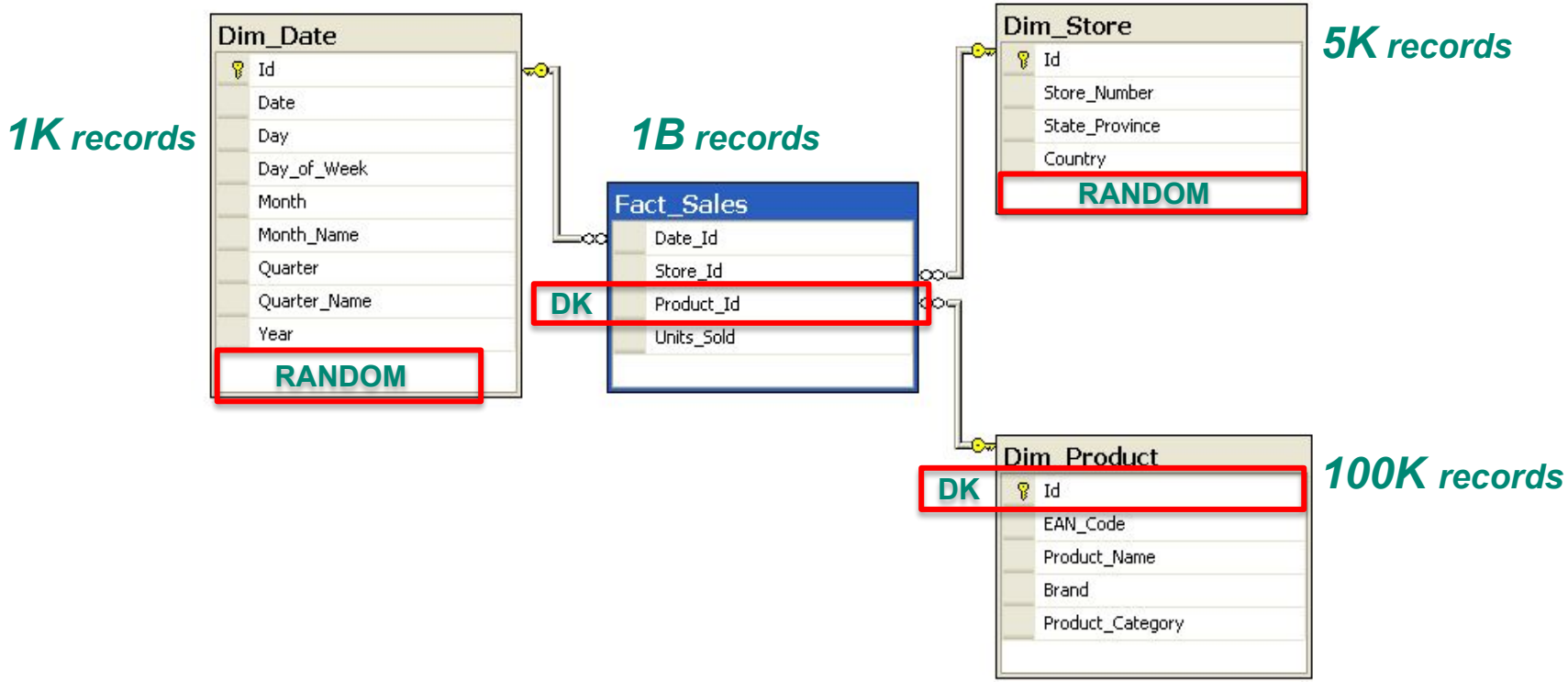
Оптимизация физической модели данных в Greenplum

Выбор атрибута распределения

Необходимо добиваться по возможности равномерного распределения данных по сегментам.

- Выбирать атрибут с хорошей селективностью (большое количество уникальных значений)
- Атрибуты ключа распределения должны использоваться в join
- Один, максимум два атрибута в ключе распределения
- В ключе не должно быть null или значений по умолчанию
- При отсутствии атрибутов с хорошим распределением, или для маленьких таблиц выбирать RANDOM распределение
- Всегда явно указывать способ распределения в DDL
- Не забывать указывать способ распределения для временных таблиц

Выбор атрибута распределения



Ориентация данных и сжатие

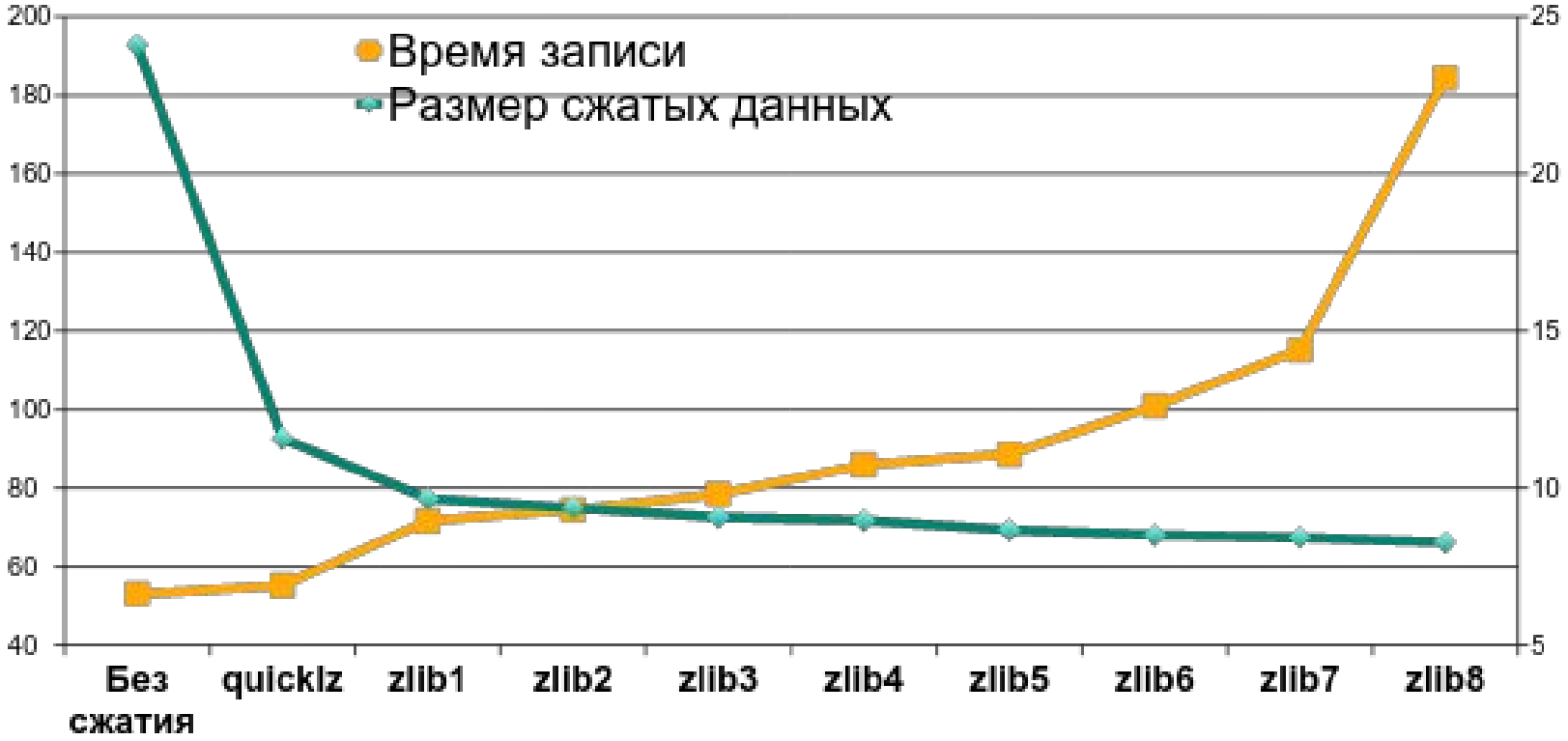
Строковая ориентация

- Применяется по умолчанию
- Доступны алгоритмы сжатия **quicklz, zlib**
- Оптимально для операций update
- Данные таблицы хранятся на сегменте в одном файле.
- При достижении размера более 1Гб таблица разбивается на файлы по 1Гб

Колоночная ориентация

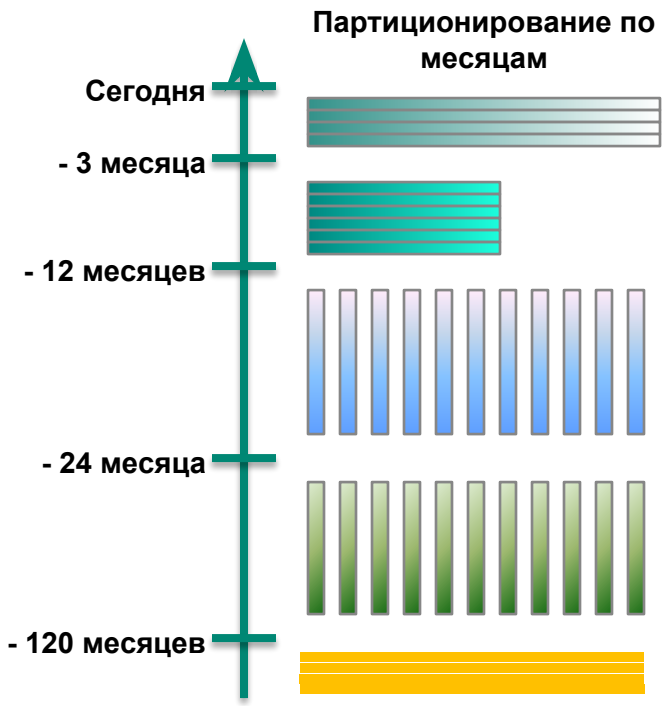
- Доступны алгоритмы сжатия **quicklz, zlib, RLE**
- Более эффективное сжатие
- Данные каждой колонки таблицы хранятся в отдельном файле.

Эффективность алгоритмов сжатия



Pivotal

Пример использования различных способов хранения данных



Способ сжатия	Влияние
Без сжатия	Быстрые операции вставки/обновления/удаления
Хранение по строкам + лёгкое сжатие (Quicklz или Zlib 1)	Лучше производительность / Меньше IO
Хранение по столбцам + лёгкое сжатие (Quicklz или Zlib 1)	Лучше производительность если не все столбцы участвуют в запросе
Хранение по столбцам + максимальное сжатие (RLE, Zlib 4+)	Возможно снижение производительности запроса из-за нагрузки на ЦП
Хранение данных в HDFS (GZIP)	Возможно снижение производительности запроса из-за необходимости считывания данных из HDFS

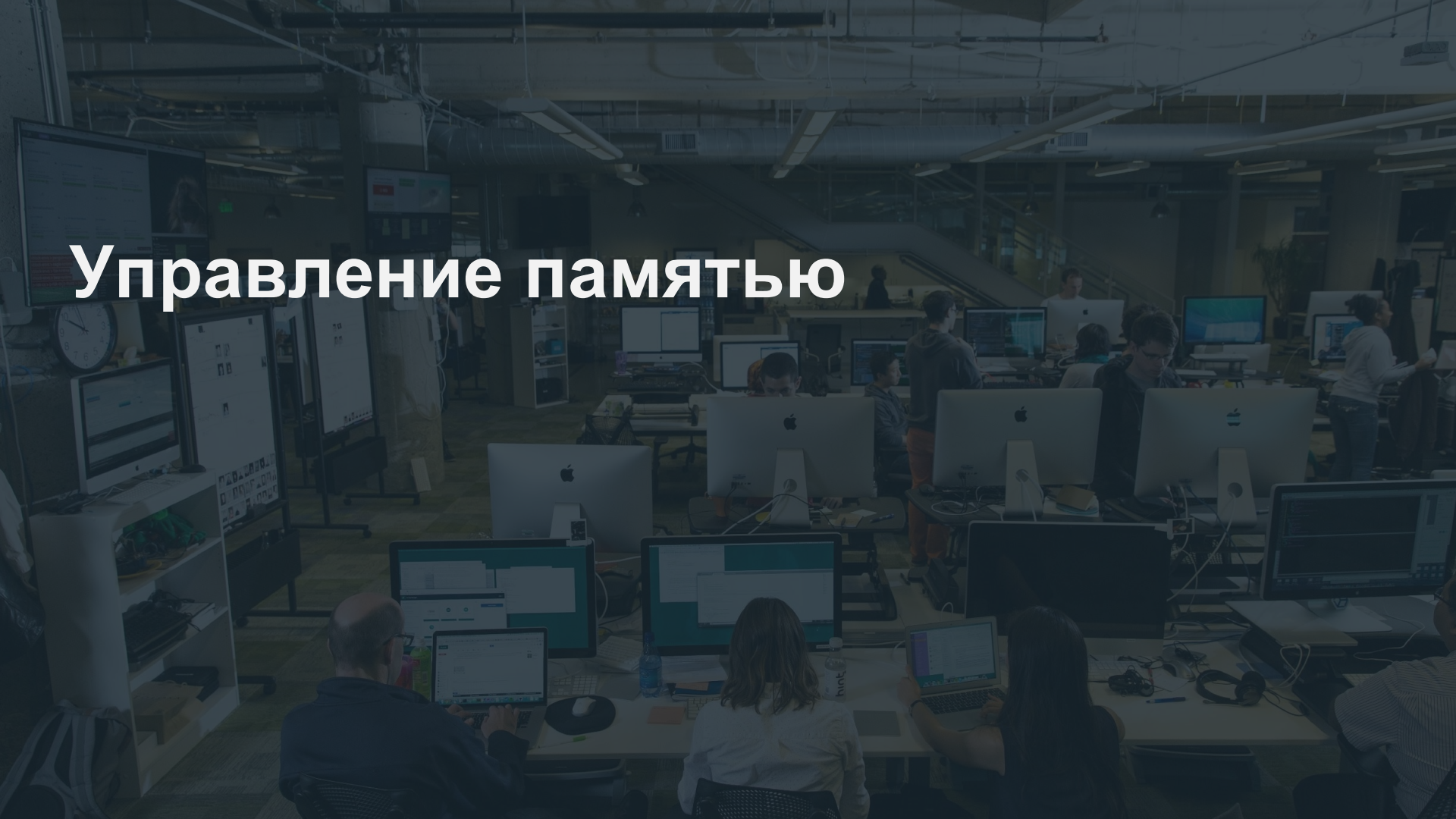
Паттерн доступа к данным меняется на разных этапах жизненного цикла данных

Pivotal

Партицирование

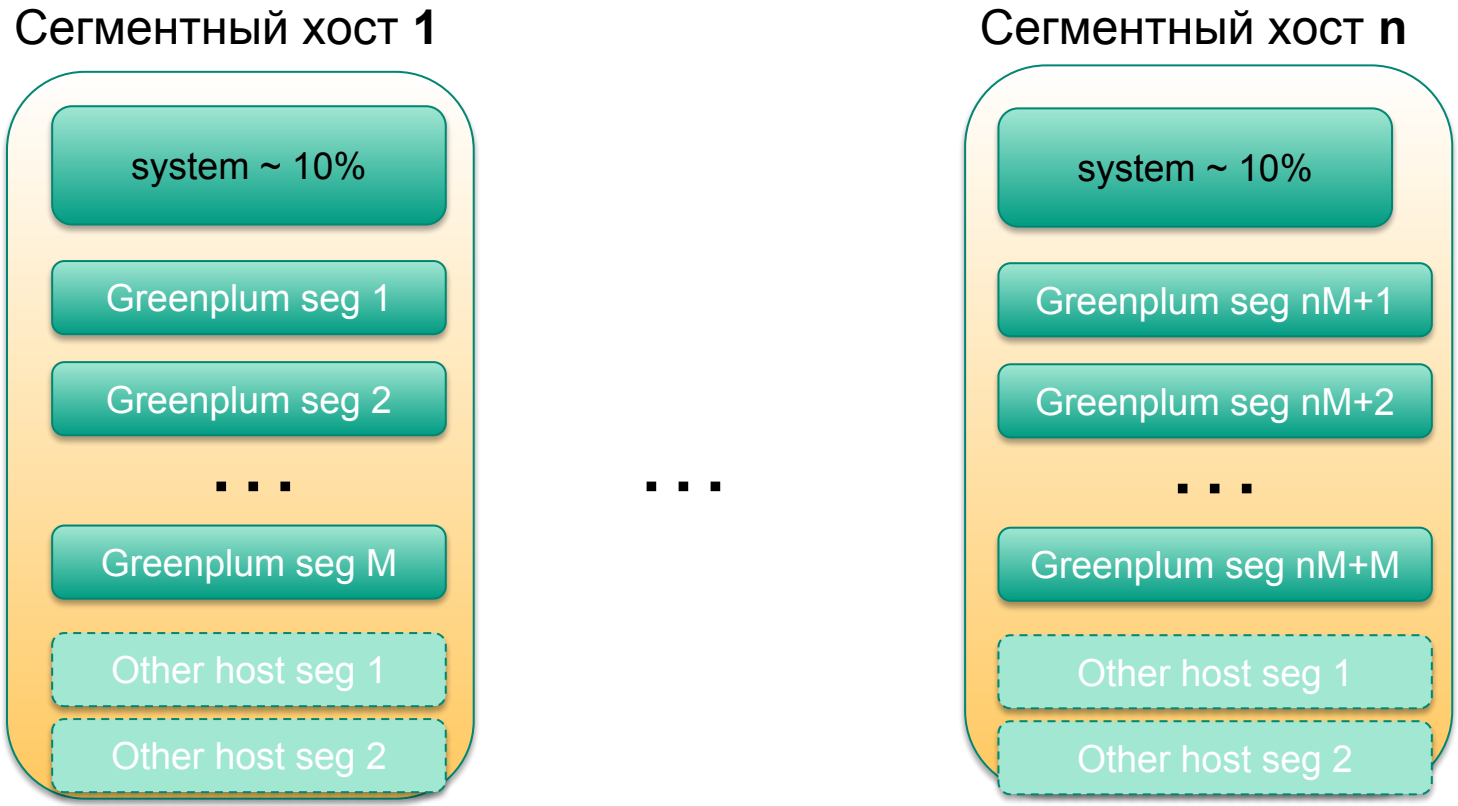
Благодаря MPP-архитектуре операция full scan в Greenplum выполняется наиболее эффективно, соответственно можно использовать партиции бóльшего размера по сравнению с не-MPP базами

- Следует избегать создания большого количества партиций
- Не использовать многоуровневое партицирование
- Размер партиции – от сотен мегабайт до десятков гигабайт (в зависимости от размера кластера)

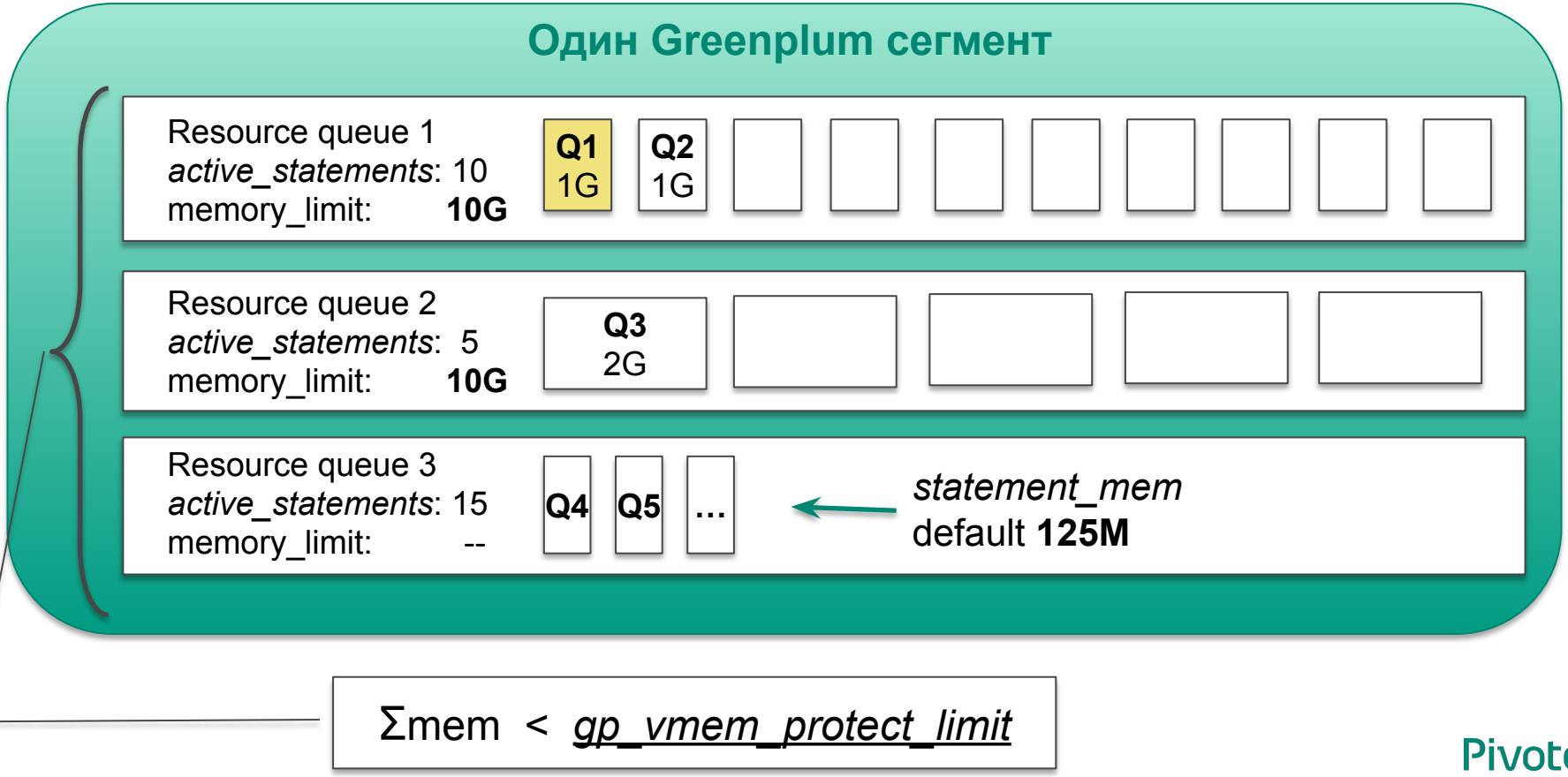


Управление памятью

Использование памяти в Greenplum



Использование памяти в Greenplum



Временные файлы

- При превышении запросом выделенной памяти данные вытесняются во временные файлы на диск
- Временные файлы хранятся в той же файловой системе, что и файлы пользовательских таблиц, и удаляются при завершении сессии
- Часто временные файлы кэшируются xfs
- Предельный размер временных файлов контролируется параметрами
  - gp_workfile_limit_per_query
  - gp_workfile_limit_per_segment

Как ускорить Greenplum?

Утилизация системных ресурсов

CPU

- Чтение сжатых данных.
- Большое количество вычислений

Диски

- Чтение несжатых данных
- Вытеснение временных данных на диск (spill) при нехватке памяти

Сеть

- Редистрибьюция
- Объемный результат запроса идёт на мастер-хост

Память

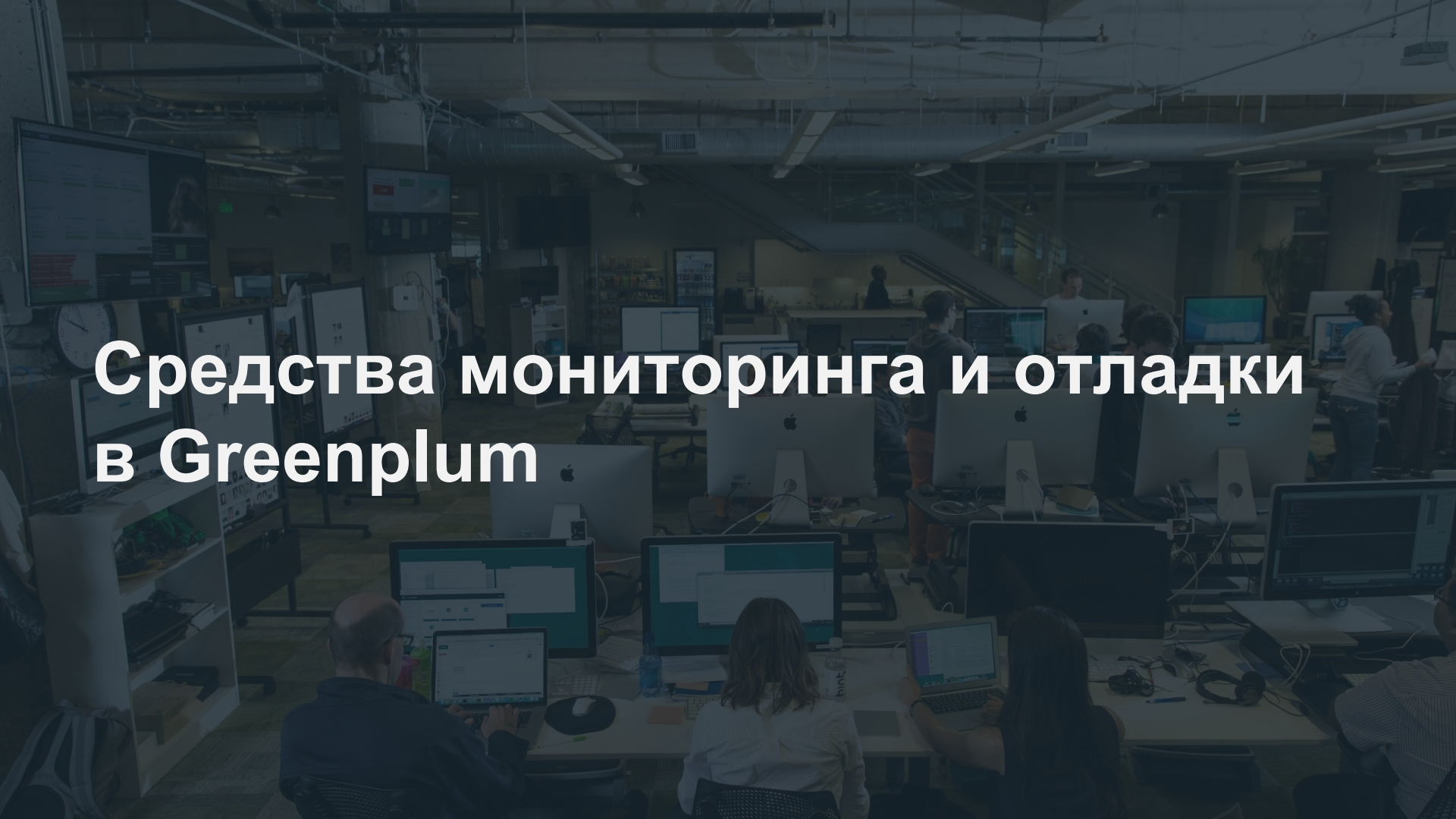
- Джойны и агрегации

Типичные причины общего снижения производительности Greenplum

- Часть сегментов Greenplum упало и работают зеркалированные сегменты
- Проблемы с диском на одном из сегментов
- Раздувание каталога
- Чрезмерная нагрузка на файловую систему и каталог из-за большого количества объектов БД
- Высокое потребление ресурсов кластера не-Greenplum процессами
- Неоптимальные настройки параметров ОС.

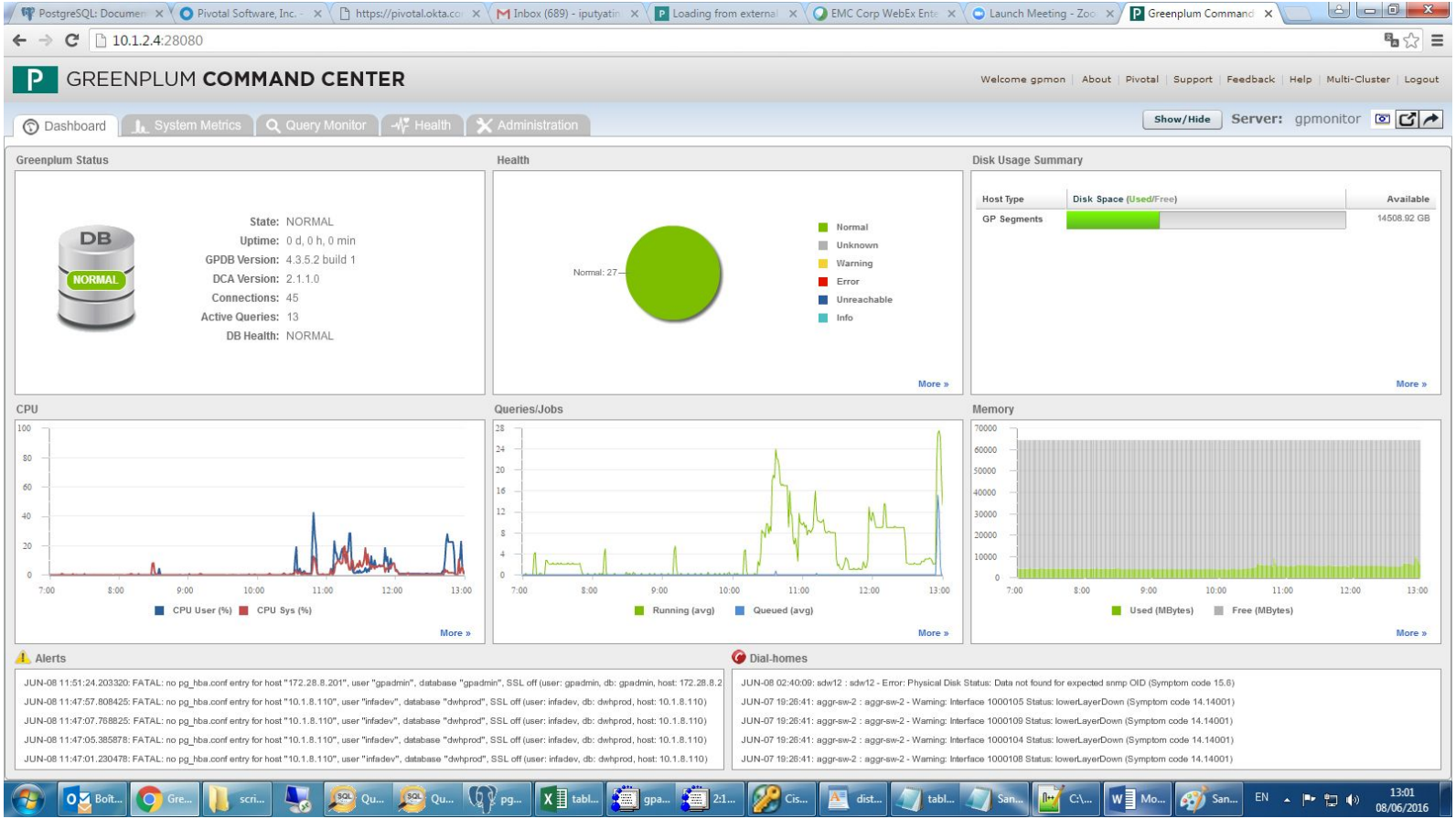
Причины снижения производительности отдельных процессов

- Неравномерное распределение данных по сегментам в таблицах
- Неоптимальный план запроса из-за отсутствия статистики
- Блокировка объектов БД разными запросами
- Ограничения очереди ресурсов по количеству активных запросов
- Ошибка в запросе или в данных (дубли, пропущенное условие джойна)



Средства мониторинга и отладки в Greenplum

Greenplum Command Center



Greenplum Command Center

GPCC – это веб-приложение для мониторинга всех показателей Greenplum в реальном времени и их истории

- Утилизация ресурсов (CPU, память, сеть, диски и др.)
- Выполняющиеся запросы
- Мониторинг «здоровья» аппаратного обеспечения: состояние жестких дисков, доступность сети, свободное дисковое пространство и прочее

Все показатели с историей хранятся в отдельной базе данных gpperfmon.

Greenplum Command Center

Вывод показателей утилизации ресурсов в разрезе узлов кластера

Host Metrics Realtime statistics by server

What is this?

Last Sync
2017-04-04 19:54:06

Hostname ▲	CPU Total/Sys/User (%)	Memory In Use (%)	Disk R (MB/s) | Skew	Disk W (MB/s) | Skew	Net R (MB/s) | Skew	Net W (MB/s) | Skew
ip-172-21-10-116.eu-central-1.compute.internal	49.55	5.68	186.1	1289	1.63	1.62
ip-172-21-3-71.eu-central-1.compute.internal	50.83	5.67	186.3	1259	1.66	2.39
ip-172-21-4-253.eu-central-1.compute.internal	49.92	5.65	192.9	1297	2.39	1.73
ip-172-21-8-224.eu-central-1.compute.internal	49.57	5.65	191.1	1274	1.69	1.63
ip-172-21-3-26.eu-central-1.compute.internal	0.01	1.50	0	0	0	0

Pivotal

Планы запросов

Планы запросов - основной инструмент отладки и оптимизации для разработчика запросов

- Explain – план запроса, который формируется планировщиком на основе запроса, DDL таблиц и статистики. Выводится оценочное количество записей в промежуточных результатах запроса.

- Explain Analyze – план запроса с фактическими значениями количества записей в промежуточных результатах запроса, а также фактическим потреблением ресурсов (память, временные файлы, время выполнения отдельных этапов). Для вывода этого плана Greenplum сначала выполняет сам запрос.

На что следует обращать внимание

В выводе explain

- Правильная оценка количества записей в таблицах
- Правильная оценка количества записей после фильтров
- Partition elimination
- Отсутствие лишних редистрибьюций

В выводе explain analyze

- Соотношение среднего/максимального количества записей по сегментам
- Значение Start offset by

Планы запросов

http://planchecker.cfapps.io

Инструмент для подсветки синтаксиса плана и выявления проблемных мест

```
-> Partition Selector for transaction (dynamic scan id: 1)
(cost=10.00..100.00 rows=50 width=4)
   Partitions selected:  240 (out of 240)
   Rows out:  0 rows (seg0) with 0.087 ms to end, start offset by 33
ms.
```

WARNING: Detected 240 partition scans | Check if partitions can be eliminated

WARNING: 100% (240 out of 240) partitions selected | Check if partitions can be eliminated

Процессы Greenplum

Процессы сегментов Greenplum (инстансов postgresql)

```
ps -ef | grep postgres | grep silent
```

Процессы пользовательских сессий

```
ps -ef | grep postgres | grep con
```

Процессы конкретной сессии

```
ps -ef | grep postgres | grep con<sess_id>
```

Процессы бэкапа

```
ps -ef | grep dump
```


Pivotal
Transforming How The World Builds Software