

Igor Putyatin  
Pivotal Data Engineering

PGDAY'  
RUSSIA 17

КОНФЕРЕНЦИЯ  
ПО БАЗАМ ДАННЫХ

# Greenplum Best Practices

Pivotal

# 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:

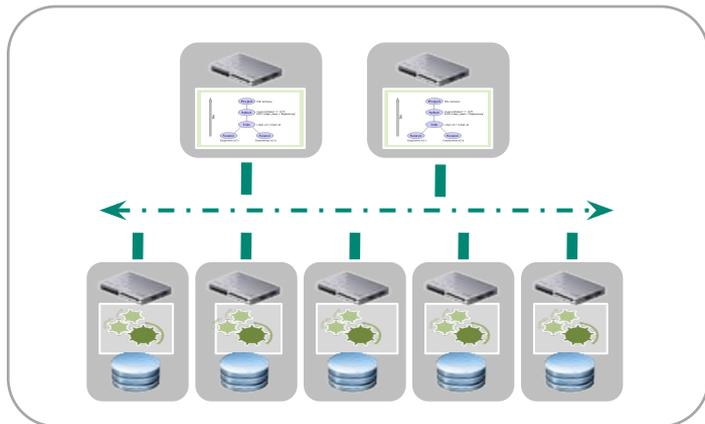
- анализ данных
- загрузка и хранение структурированных данных
- работа с Business Intelligence приложениями
- использование алгоритмов машинного обучения.

# Pivotal Greenplum :

## Массивно-параллельная архитектура

### Массивно-параллельная обработка данных

#### Хранилище данных Greenplum



#### Аналитические инструменты СУБД

##### Встроенные библиотеки



##### Программирование



### Массивно-параллельная загрузка данных из внешних источников

#### Хранилища данных в экосистеме Hadoop



cloudera

MAPR

#### Облачные хранилища данных



# 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



## 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

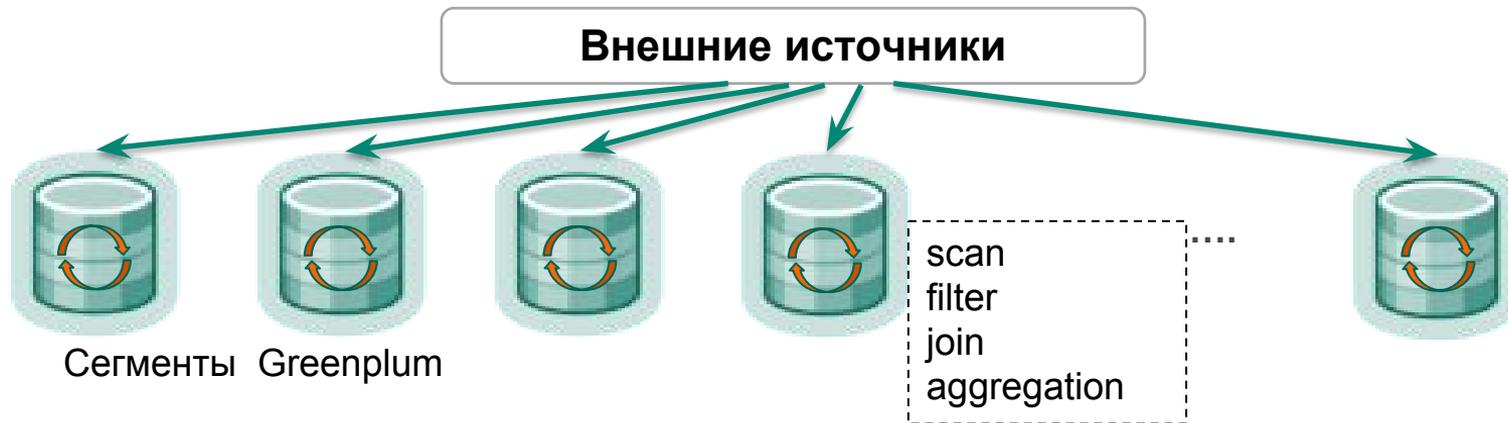
На данный момент самый крупный работающий в продуктиве кластер Greenplum состоит из **128** сегментных хостов.

- 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

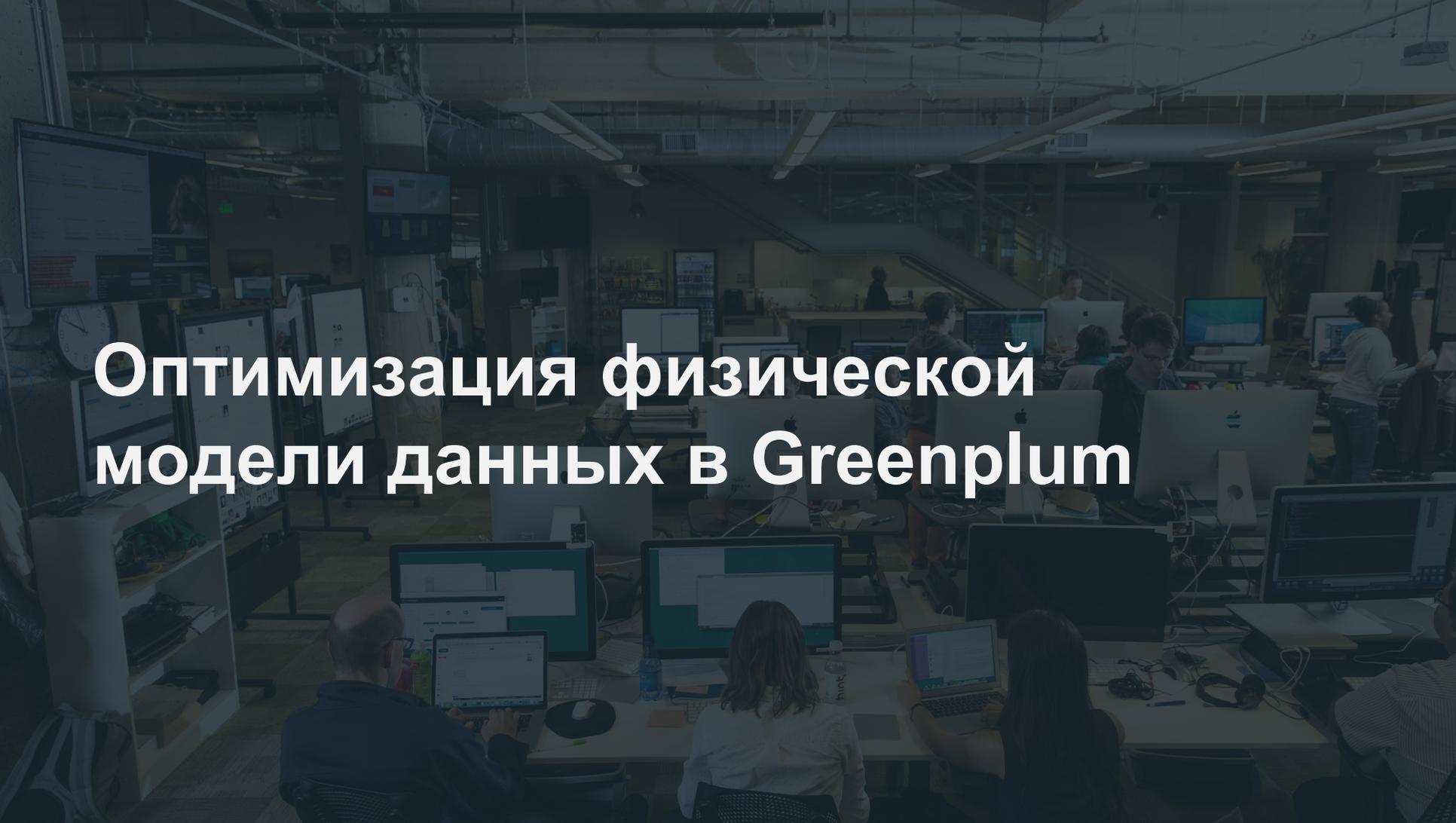
Greenplum работает максимально эффективно на тех операциях, которые можно распараллелить

- 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

1K records

Dim_Date	
Id	PK
Date	
Day	
Day_of_Week	
Month	
Month_Name	
Quarter	
Quarter_Name	
Year	

**RANDOM**

1B records

Fact_Sales	
Date_Id	
Store_Id	
Product_Id	DK
Units_Sold	

5K records

Dim_Store	
Id	PK
Store_Number	
State_Province	
Country	

**RANDOM**

100K records

Dim_Product	
Id	PK
EAN_Code	
Product_Name	
Brand	
Product_Category	

**DK**

# Ориентация данных и сжатие

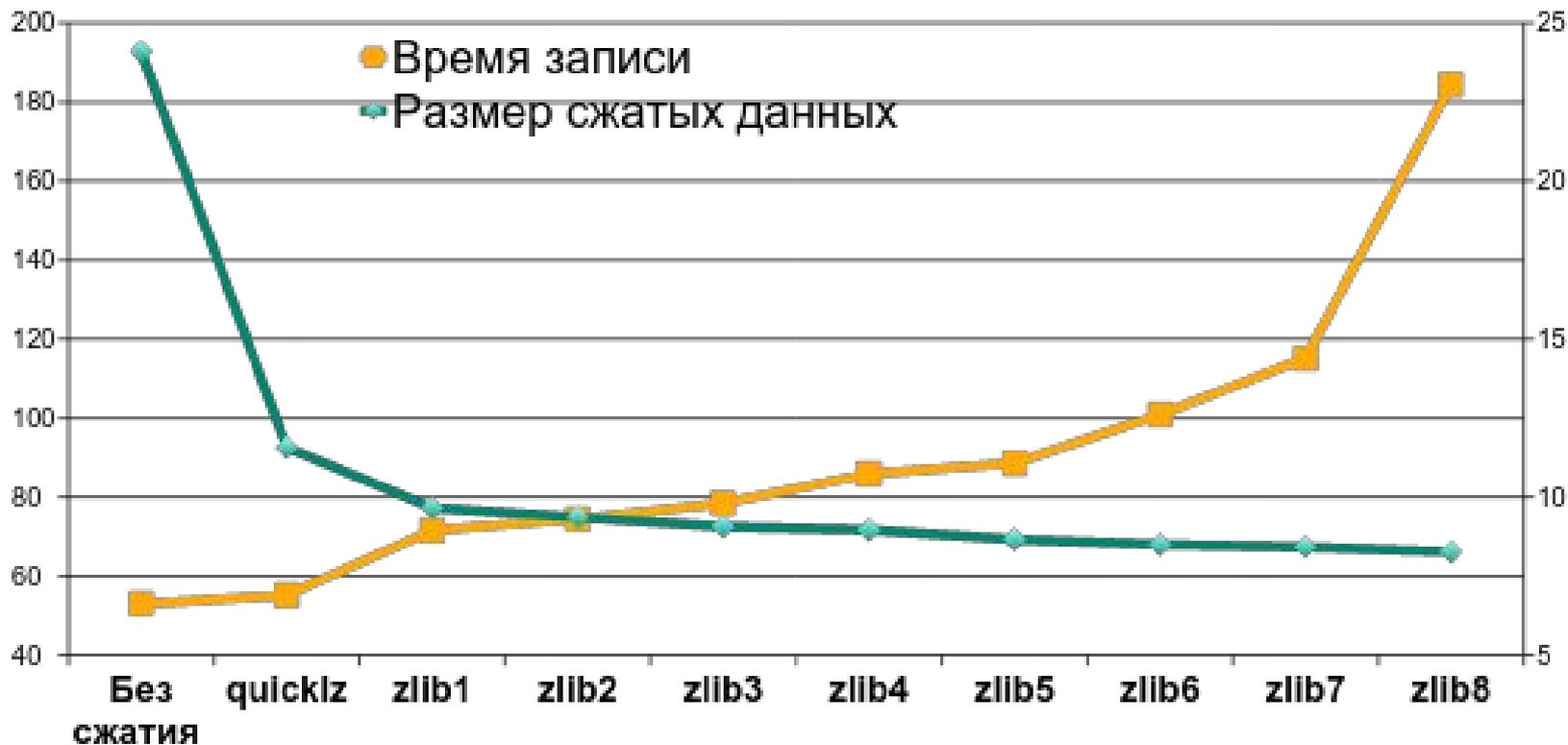
## Строковая ориентация

- 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

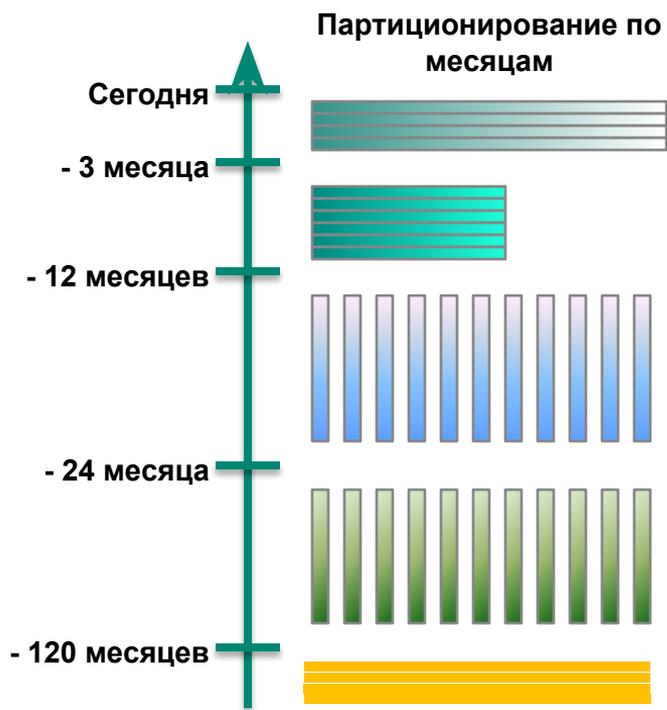
## Колоночная ориентация

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

# Эффективность алгоритмов сжатия



# 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

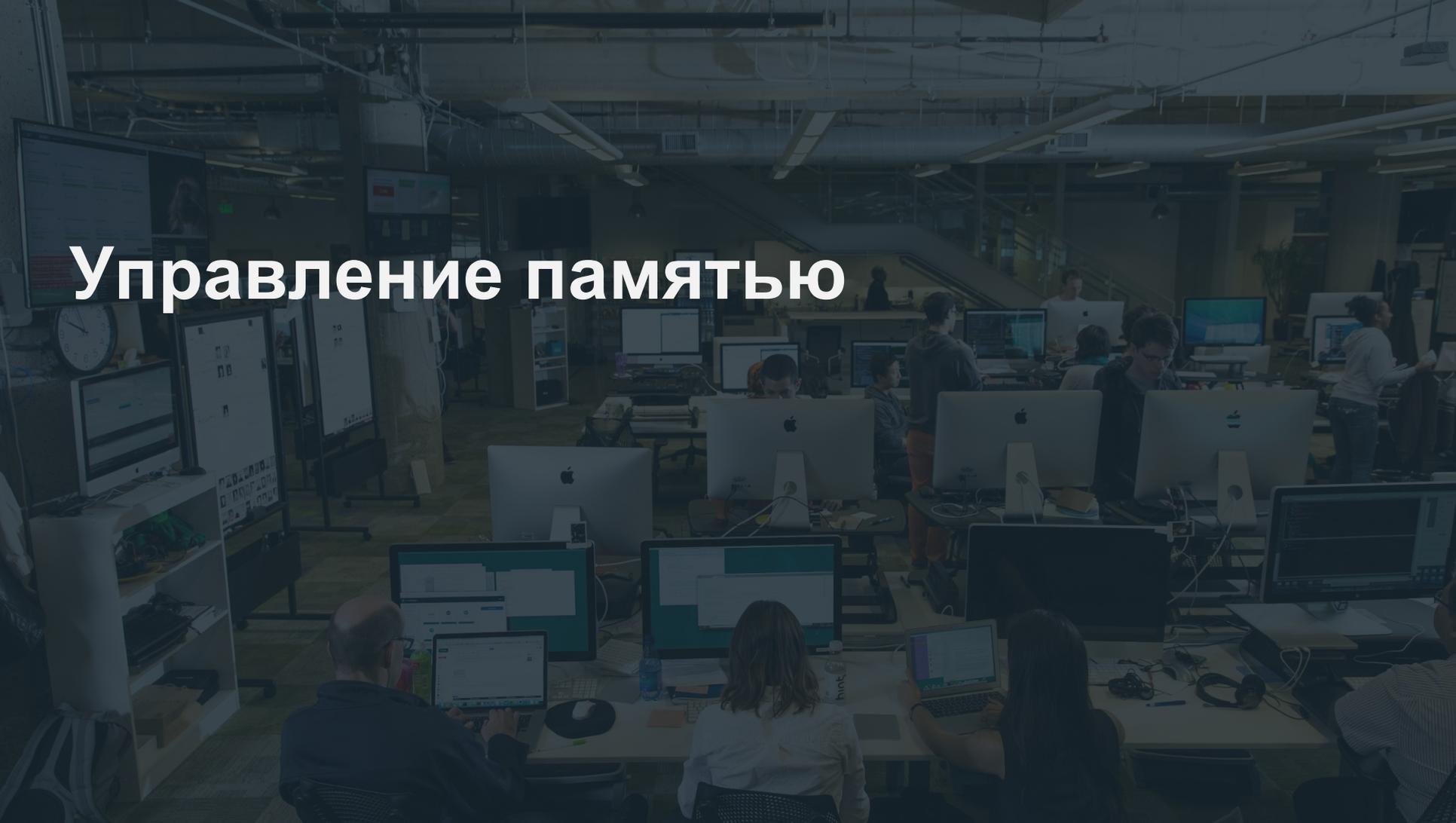
Паттерн доступа к данным меняется на разных этапах жизненного цикла данных

# 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

A dimly lit, modern office with many people working at desks with computers. The text "Управление памятью" is overlaid in white. The office has a high ceiling with exposed pipes and lights. There are several large monitors and laptops on the desks. People are seen from behind, focused on their work. The overall atmosphere is professional and busy.

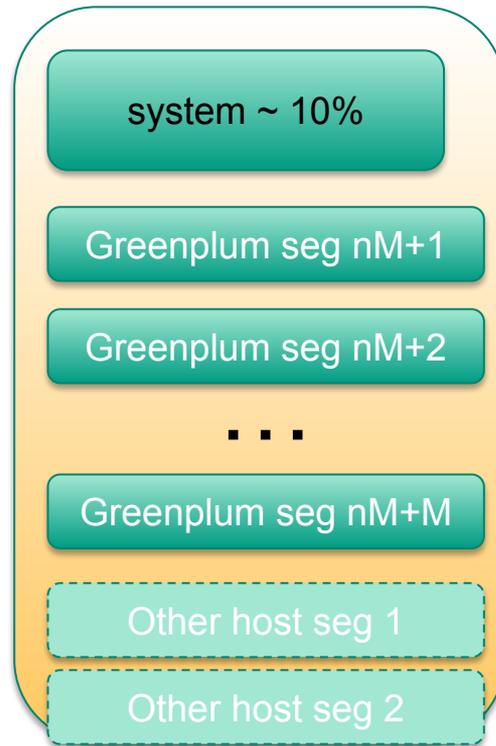
# Использование памяти в Greenplum

## Сегментный хост 1



...

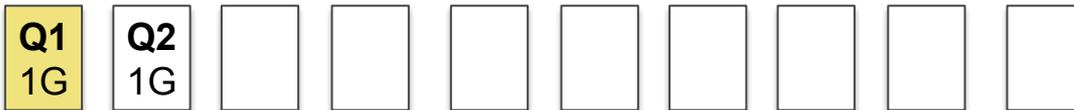
## Сегментный хост n



# Использование памяти в Greenplum

## Один Greenplum сегмент

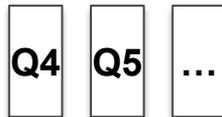
Resource queue 1  
*active\_statements*: 10  
*memory\_limit*: **10G**



Resource queue 2  
*active\_statements*: 5  
*memory\_limit*: **10G**



Resource queue 3  
*active\_statements*: 15  
*memory\_limit*: --



← *statement\_mem*  
default **125M**

$$\Sigma \text{mem} < \text{gp\_vmem\_protect\_limit}$$

# 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?

A dimly lit office with many people working at computers. The text 'Как ускорить Greenplum?' is overlaid in white. The office is filled with desks, each with multiple computer monitors. Some people are sitting at their desks, while others are standing and talking. The lighting is low, creating a focused and professional atmosphere. The text is centered in the upper half of the image.

# Утилизация системных ресурсов

## СРУ

- 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

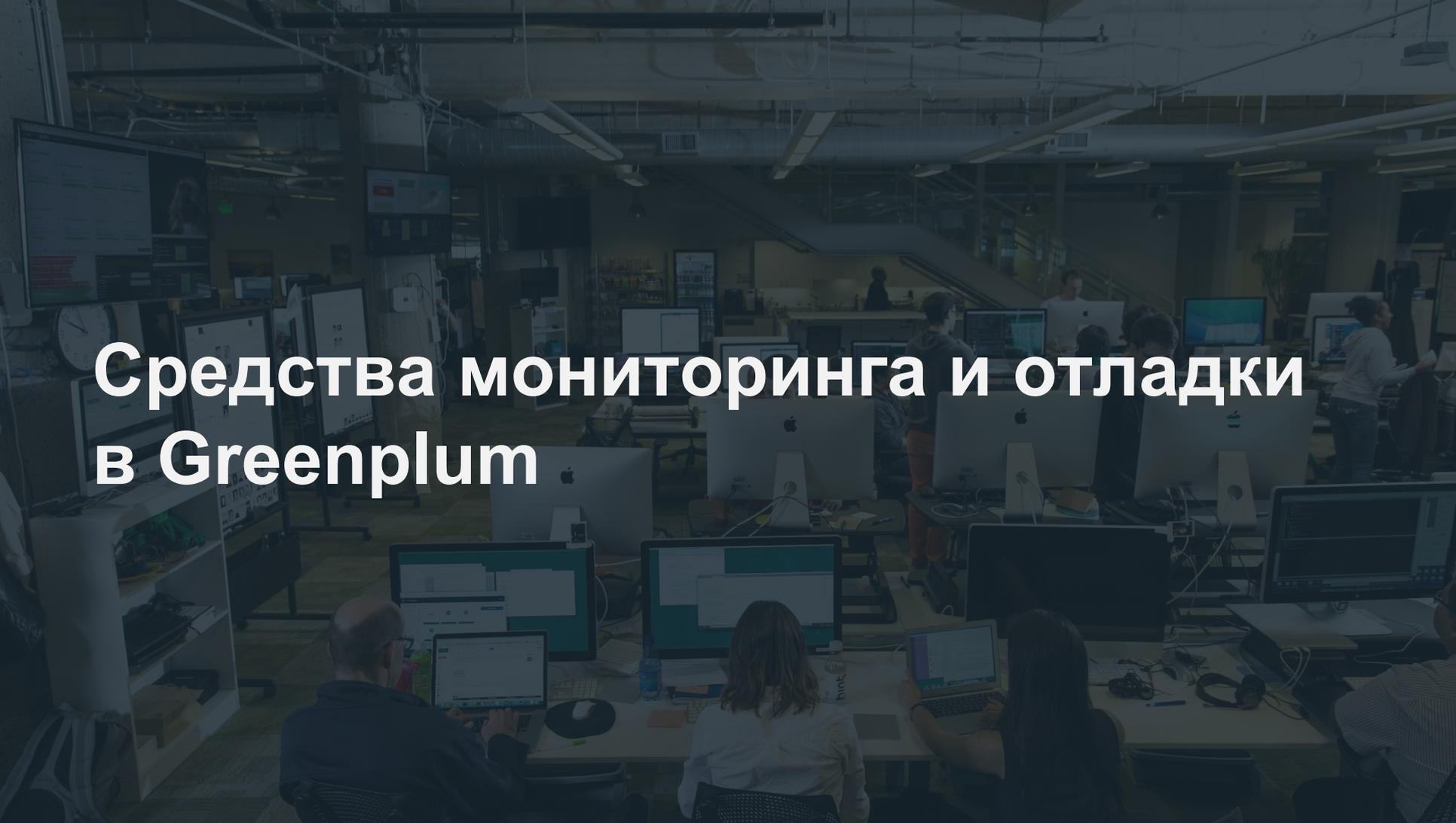
- Джойны и агрегации

# 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

PostgreSQL: Document... Pivotal Software, Inc. - https://pivotal.okta.co... Inbox (689) - iputyatin... Loading from external... EMC Corp WebEx Ent... Launch Meeting - Zoo... Greenplum Command

10.1.2.4:28080

**GREENPLUM COMMAND CENTER** Welcome gpmon | About | Pivotal | Support | Feedback | Help | Multi-Cluster | Logout

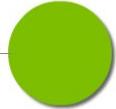
Dashboard | System Metrics | Query Monitor | Health | Administration Show/Hide Server: gpmonitor

### Greenplum Status



State: NORMAL  
 Uptime: 0 d, 0 h, 0 min  
 GPDB Version: 4.3.5.2 build 1  
 DCA Version: 2.1.1.0  
 Connections: 45  
 Active Queries: 13  
 DB Health: NORMAL

### Health

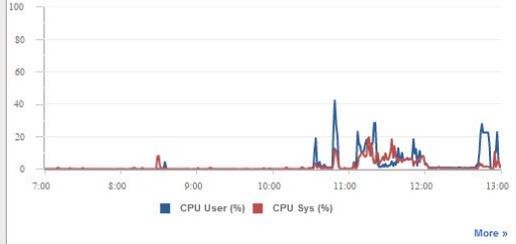


- Normal
- Unknown
- Warning
- Error
- Unreachable
- Info

### Disk Usage Summary

Host Type	Disk Space (Used/Free)	Available
GP Segments		14508.92 GB

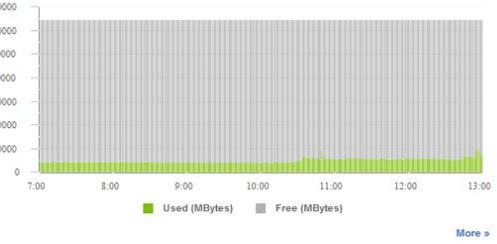
### CPU



### Queries/Jobs



### Memory



### Alerts

JUN-08 11:51:24.203320: FATAL: no pg\_hba.conf entry for host "172.28.8.201", user "gpadmin", database "gpadmin", SSL off (user: gpadmin, db: gpadmin, host: 172.28.8.201)

JUN-08 11:47:57.808425: FATAL: no pg\_hba.conf entry for host "10.1.8.110", user "infadev", database "dwhprod", SSL off (user: infadev, db: dwhprod, host: 10.1.8.110)

JUN-08 11:47:07.768825: FATAL: no pg\_hba.conf entry for host "10.1.8.110", user "infadev", database "dwhprod", SSL off (user: infadev, db: dwhprod, host: 10.1.8.110)

JUN-08 11:47:05.385878: FATAL: no pg\_hba.conf entry for host "10.1.8.110", user "infadev", database "dwhprod", SSL off (user: infadev, db: dwhprod, host: 10.1.8.110)

JUN-08 11:47:01.230478: FATAL: no pg\_hba.conf entry for host "10.1.8.110", user "infadev", database "dwhprod", SSL off (user: infadev, db: dwhprod, host: 10.1.8.110)

### Dial-homes

JUN-08 02:40:09: sdw12 : sdw12 - Error: Physical Disk Status: Data not found for expected snmp OID (Symptom code 15.8)

JUN-07 19:26:41: aggr-sw-2 : aggr-sw-2 - Warning: Interface 1000105 Status: lowerLayerDown (Symptom code 14.14001)

JUN-07 19:26:41: aggr-sw-2 : aggr-sw-2 - Warning: Interface 1000109 Status: lowerLayerDown (Symptom code 14.14001)

JUN-07 19:26:41: aggr-sw-2 : aggr-sw-2 - Warning: Interface 1000104 Status: lowerLayerDown (Symptom code 14.14001)

JUN-07 19:26:41: aggr-sw-2 : aggr-sw-2 - Warning: Interface 1000108 Status: lowerLayerDown (Symptom code 14.14001)

# 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

# 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 редистрибуций

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
```

A photograph of the Golden Gate Bridge in San Francisco, partially obscured by a thick layer of fog. The bridge's iconic towers and suspension cables are visible against a dark, overcast sky. The foreground shows a steep, rocky hillside with sparse vegetation.

# Pivotal®

Transforming How The World Builds Software