Задание для выполнения:
В этом пособие изучаем Главу 3. Частные вопросы управления ИТ-инфраструктурой, а именно:

Примеры инфраструктурных решений, применяющихся в крупных сетевых проектах.

- Пример реализации инфраструктуры в Google.
- Пример реализации инфраструктуры для проекта Flickr

Создать документ-отчет, в котором дать описание каждого примера реализации и управления ИТ инфраструктурой.

—

Пример реализации инфраструктуры в Google

Используемые технологии:

Google File System (GFS)

GFS — это распределенная файловая система, предназначенная для работы в условиях, когда компоненты системы часто выходят из строя. Основные особенности GFS включают высокую надежность, масштабируемость и доступность данных. Система не требует немедленного освобождения дискового пространства при удалении файлов, что упрощает управление данными.

Пример реализации:

1. Репликация данных:
   Мастер в GFS управляет репликациями чанков, распределяя их между серверами с учетом загруженности и других факторов, таких как наличие свободного места и размещение на разных стойках. Реплики распределяются так, чтобы обеспечить максимальную надежность и доступность данных, даже в случае выхода из строя целых стоек. В случае потери реплики система автоматически реплицирует чанк, чтобы восстановить его число.
2. Механизм блокировок:
   Для предотвращения конфликтов между операциями в GFS используется система блокировок на чтение и запись. Например, при создании файла система блокирует пути к директориям и файл, предотвращая изменения, которые могут привести к несоответствиям в данных.

3. Сборка мусора:
   При удалении файла система не освобождает место немедленно, а переименовывает файл, что позволяет в случае необходимости восстановить данные. Реальная очистка происходит через регулярную сборку мусора, когда скрытые файлы удаляются, и связанные с ними чанки отцепляются от файлов.

MapReduce

MapReduce — это модель для обработки и генерации больших наборов данных, используемая в Google для выполнения параллельных вычислений на кластере машин. Она состоит из двух этапов: map и reduce, что позволяет эффективно распределять задачи и обрабатывать данные в распределенной среде.

Пример реализации:

1. Распределение задач:
   В MapReduce задания распределяются между серверами типа Map, которые обрабатывают входные данные, и серверами типа Reduce, которые агрегируют промежуточные результаты. На каждом этапе MapReduce происходит сжатие данных для эффективной транспортировки между серверами, что снижает нагрузку на каналы передачи данных.
2. Применение в задачах анализа данных:
   Одним из примеров применения MapReduce является подсчет количества слов на всех страницах. В этом случае данные из GFS передаются в MapReduce, где на этапе map данные разбиваются на пары ключ/значение (слово и его количество), а на этапе reduce результаты агрегируются и записываются обратно в GFS.
3. Механизмы обработки сбоев:
   MapReduce автоматически восстанавливает выполнение задач в случае сбоев. Это достигается путем создания резервных копий промежуточных данных и использования нескольких серверов для обработки одних и тех же данных.

BigTable

BigTable — это распределенная система хранения данных, которая использует Google File System для хранения данных и обеспечивает масштабируемость и высокую доступность. BigTable не является реляционной базой данных и не поддерживает SQL-запросы, что

позволяет ей эффективно работать с большими объемами данных, которые требуют специфической обработки.

Пример реализации:

1. Механизм хранения данных:
   BigTable управляет данными с использованием таблиц, которые разбиваются на блоки. Каждый блок имеет размер 64 КБ и хранится в формате SSTable. Доступ к данным возможен по ключам строки, столбца или временной метке.
2. Типы серверов:
   В BigTable используется три типа серверов: Master: управляет распределением таблиц по серверам и следит за их состоянием. Tablet: обрабатывает запросы чтения/записи для таблиц, деля их на части при превышении размера. Lock: обеспечивает синхронизацию доступа к данным, предотвращая одновременные изменения данных.
3. Механизм устойчивости и отказоустойчивости:
   В случае выхода из строя одного сервера, другие серверы берут на себя его задачи. Это позволяет системе продолжать функционировать без сбоев, обеспечивая высокую доступность и устойчивость к отказам.

Пример реализации инфраструктуры для проекта Flickr

Инфраструктура и оборудование

Для обеспечения эффективной работы и масштабируемости Flickr использует специализированное оборудование и программные компоненты. В числе основных характеристик:

- Оборудование:
  На серверах используются процессоры EMT64, с операционной системой RHEL 4, 16 ГБ оперативной памяти и 6 жестких дисков с 15000rpm, объединенных в RAID-10. Размер пользовательских метаданных достигает 12 ТБ, что исключает хранение фотографий, и использованы 2U корпуса для серверов.
- Объем хранения:
  Flickr управляет двумя петабайтами дискового пространства, с ежедневным добавлением более 400,000 фотографий. Это требует

масштабируемых решений для хранения и быстрого доступа к данным.

Механизмы балансировки нагрузки и управления трафиком

Одним из ключевых аспектов эффективной работы Flickr является распределение нагрузки и управление входящими запросами.

- Коммутаторы приложений:
  Входящие запросы поступают на дублированные контроллеры приложений Brocade ServerIron ADX, которые обеспечивают балансировку трафика и коммутацию запросов на основе виртуальных ферм серверов. Используется интеллектуальное распределение нагрузки для максимального использования всех доступных серверных ресурсов.
- Использование виртуальных IP (VIP):
  Все запросы направляются на VIP, что позволяет скрывать реальные IP-адреса серверов, обеспечивая безопасность и надежность системы.
- Обслуживание сессий:
  Каждая сессия назначается определенному серверу, а все сообщения в рамках сессии обрабатываются этим сервером. Таблицы сессий синхронизируются между двумя коммутаторами, что обеспечивает отказоустойчивость.

Масштабируемость и репликация данных

Flickr использует несколько методов для обеспечения масштабируемости и высокой доступности данных.

- Масштабирование MySQL:
  Для работы с базами данных используется модификация MySQL Dual Tree, позволяющая масштабировать систему без использования кольцевой архитектуры. Добавление новых мастер-серверов позволяет эффективно распределять нагрузку и повышать производительность, при этом без значительных вложений в оборудование.
- Репликация данных:
  В центре базы данных находится таблица пользователей, и каждый сегмент системы хранит данные о более чем 400 тысячах пользователей. Активная репликация осуществляется по принципу мастер-мастер, что позволяет поддерживать систему в режиме

одновременной активности обоих серверов. Для оптимизации работы данные привязываются к случайным сегментам, а миграция пользователей проводится для балансировки нагрузки.

Хранение и обработка фотографий

Основная нагрузка на инфраструктуру Flickr приходится на хранение и обработку изображений.

- Система хранения данных:
  Все фотографии хранятся в системе хранения данных, в то время как метаданные и ссылки на местоположения файлов в файловых системах хранятся в базе данных. После загрузки изображения система генерирует различные размеры фотографий, и на этом процесс заканчивается. Данные фотографии обрабатываются с помощью ImageMagick.
- Кэширование и прокси-сервера:
  Для ускорения доступа к часто запрашиваемым изображениям используется кэширование с помощью Memcached, а также проксирование через серверы Squid для хранения HTML-страниц и изображений в кэше. Это позволяет обеспечить быстрое реагирование системы на запросы пользователей.

Управление сессиями и мониторинг

Flickr активно использует средства мониторинга и управления для обеспечения бесперебойной работы.

- Мониторинг и логирование:
  Для мониторинга распределенных систем используется Ganglia, что позволяет отслеживать состояние серверов и эффективно управлять производительностью системы. Subcon используется для хранения конфигурационных файлов в SVN-репозиториях, что упрощает развертывание на серверных кластерах.
- Процесс резервного копирования:
  Регулярные резервные копии данных выполняются с помощью процесса ibbackup с использованием cron. На каждом сегменте настроено выполнение резервного копирования в разное время для минимизации воздействия на производительность.

Обработка пиковой нагрузки

Flickr сталкивается с периодами пиковой нагрузки, когда количество запросов резко возрастает. Чтобы справиться с этим, используются следующие меры:

- Обработка пиковых нагрузок:
  В моменты пиковой нагрузки, когда система получает до 6-7 тысяч запросов в секунду, инфраструктура должна работать при полной загрузке. Однако система может поддерживать нормальную работу даже при увеличении нагрузки до 50% от максимальной мощности, обеспечивая стабильную производительность.
- Миграция и балансировка нагрузки:
  Миграция пользователей и сбалансированное распределение нагрузки между сегментами позволяет системе работать эффективно и выдерживать увеличение числа запросов.