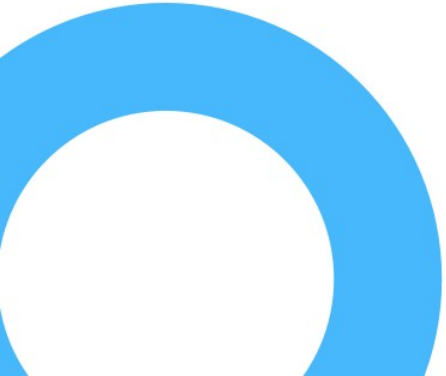
Hadoop High availability: опыт Badoo

Крашенинников Александр



badoo

О чем доклад?

badoo

- Узлы и компоненты Hadoop
- Зачем обеспечивать высокую доступность кластера
- Как постичь дзен Hadoop High Availability (HA)
- Наш опыт внедрения HA

Что будет в докладе

Что такое Hadoop?

Hadoop

badoo

- ✦Программный комплекс для распределённых вычислений
- ✦Open-source
- ✦Иногда - синоним “Big Data”

Использование Hadoop в Badoo

Hadoop в Badoo

- Инсталляция из 67 машин
- Объём файловой системы ~ 850 Tb, 90M файлов
- CPU: 2700 cores
- RAM: 10 Tb

История появления

- С 2013 года - долговременное хранилище BI данных
- Концепция - распределённая файловая система + средства обработки

Hadoop в Badoo: Deep storage

- Deep storage для хранения исторических данных
- “Backup” данных из аналитической базы
- По некоторым источникам есть история за несколько лет
- > 600 Tb занятого дискового пространства

Hadoop в Badoo: ETL

- ETL (Extract Transform Load) - важный процесс BI
- Данные из внешнего/внутреннего мира => информационная модель компании
- Вход может быть в 100500 разных видах, выход - набор таблиц в РСУБД стиле

Hadoop в Badoo: ETL

- Загрузили много данных из внешних источников, посчитали свёртки, выгрузили в аналитическую БД
- > 700 заданий в сутки
- Десятки Tb на чтение

Hadoop в Badoo: realtime

- Многие считают, что Hadoop и realtime - несовместимо
- Но при правильных инструментах это не так :)
- Keywords: Apache Spark, Storm, Samza

Hadoop в Badoo: realtime

- In-house система потоковой аналитики на базе Apache Spark
- Строим статистику в “rolling window” (10 минут, 1час, 1 день)

Hadoop в Badoo: realtime

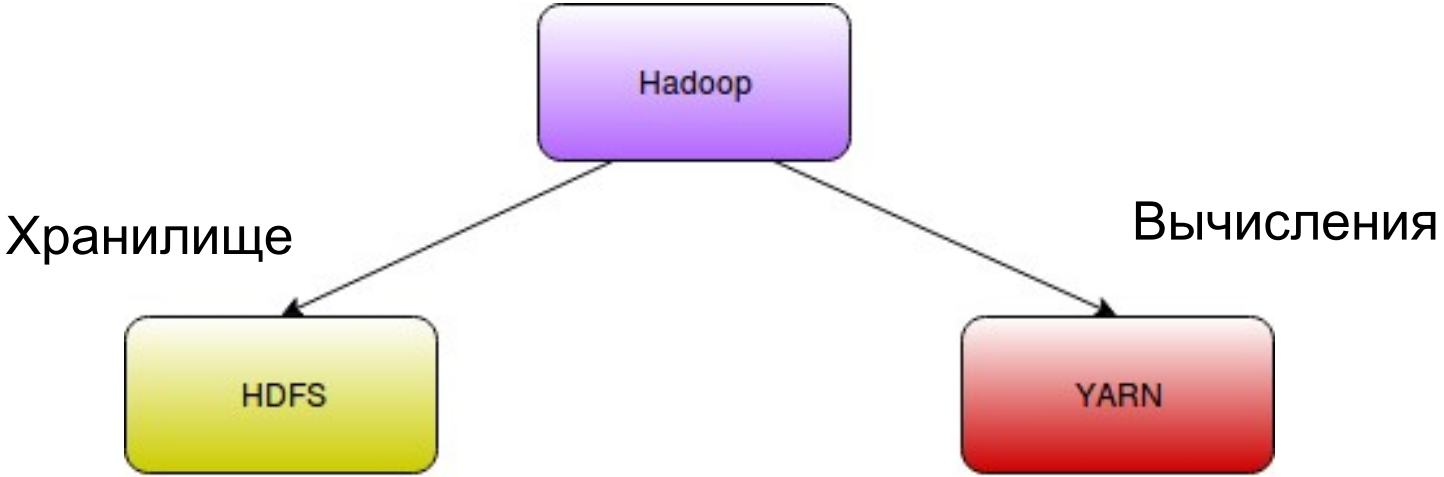
- > 800 типов статистических событий
- 450 - 600 K событий в секунду на вход
- 120 K метрик в секунду на выходе
- Anomaly Detection на выходные данные

Цена downtime

- ✦ Недоступность части данных для аналитиков к началу business time
- ✦ Деградация графиков для realtime
- ✦ Невозможность получения оперативной статистики
- ✦ Anomaly Detection “слепнет”

Из чего сделан Hadoop

Hadoop "на пальцах"



Hadoop
Distributed
File**S**ystem

Yet
Another
Resource
Negotiator

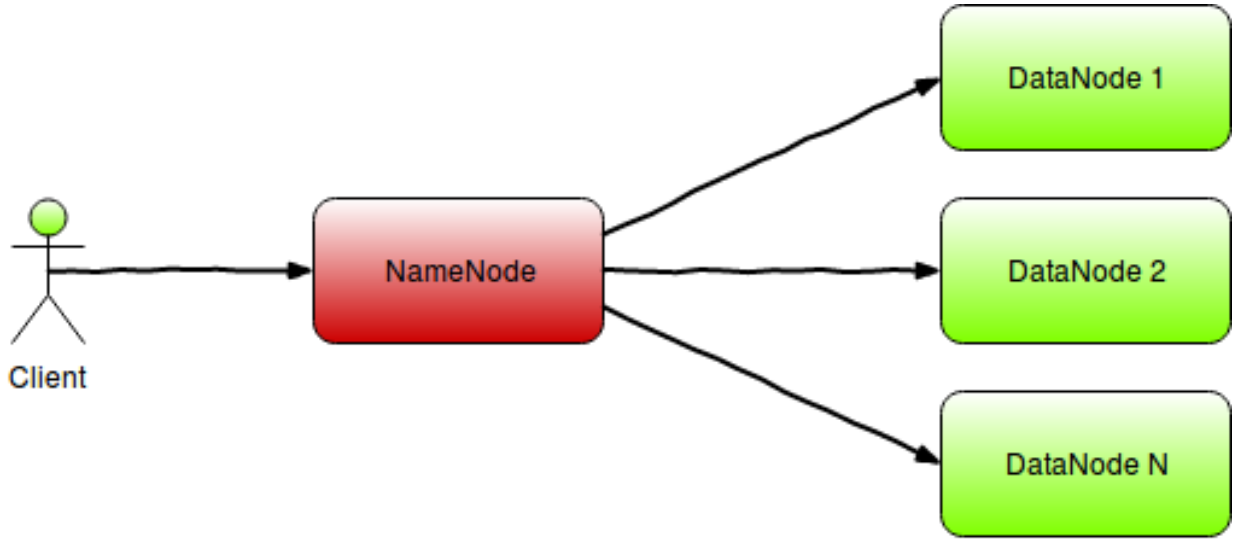
HDFS

badoo

- ✦Иерархическая файловая система
- ✦Распределена по узлам кластера
- ✦Поддерживает репликацию данных

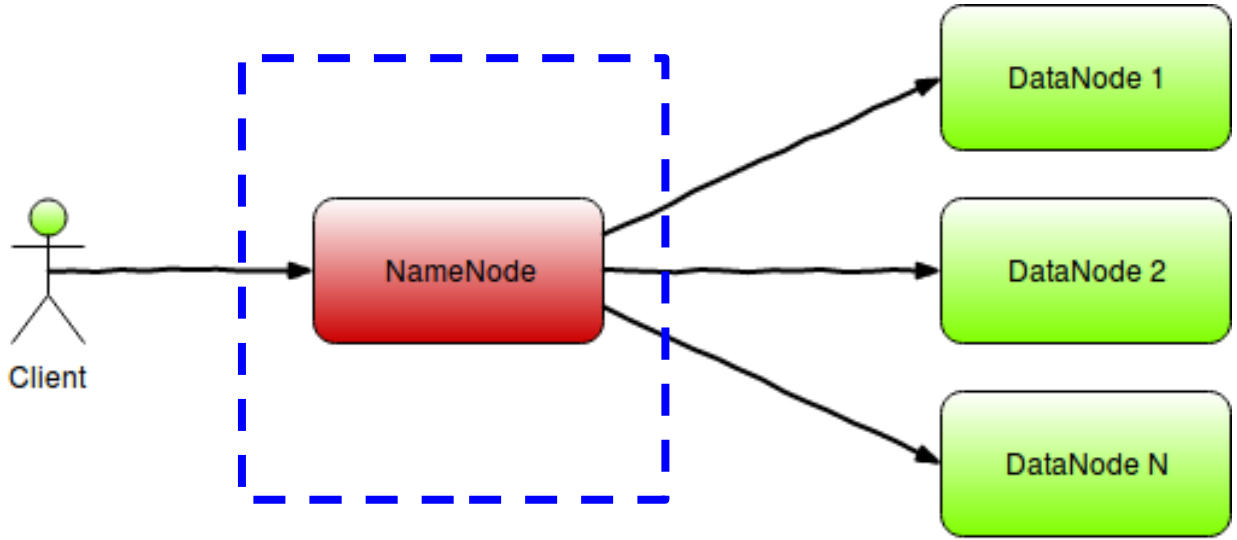
HDFS

badoo



HDFS

badoo

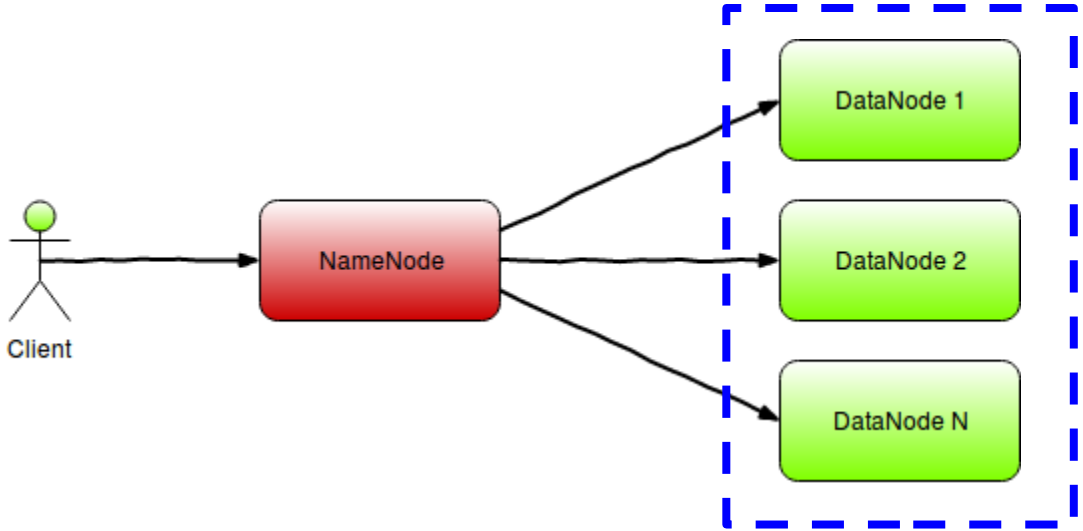


HDFS: NameNode

badoo

- Координатор файловой системы
- Хранилище индекса FS
- Входная точка для операций с данными

HDFS



HDFS: DataNode

badoo

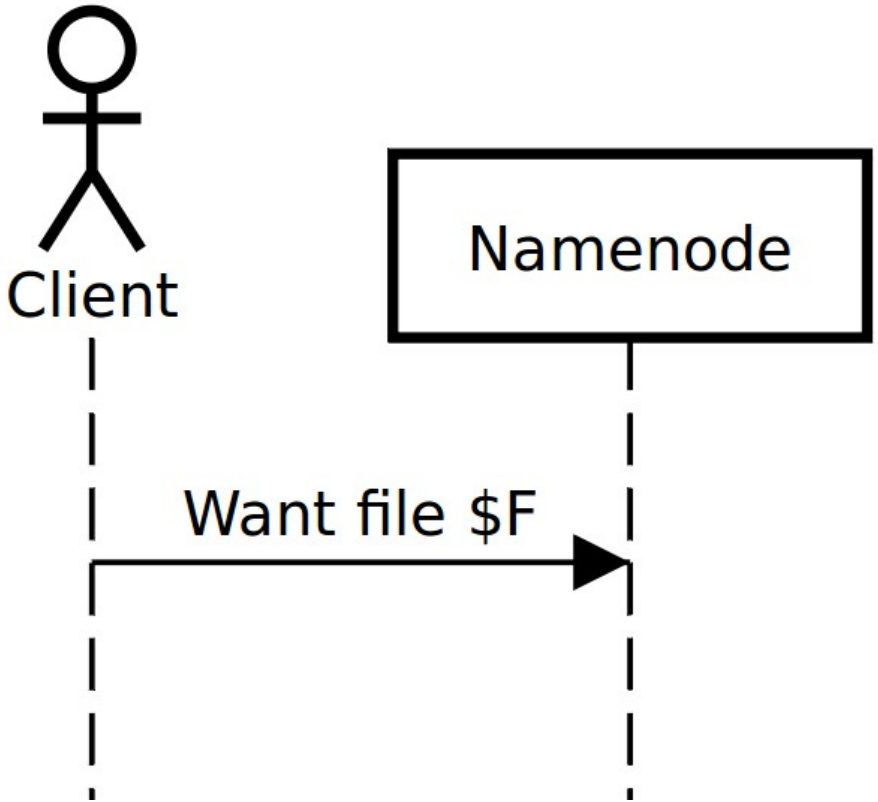
- Принимает read/write запросы **к содержимому** файлов HDFS
- Хранит содержимое файлов на своей дисковой подсистеме

HDFS: запрос на чтение

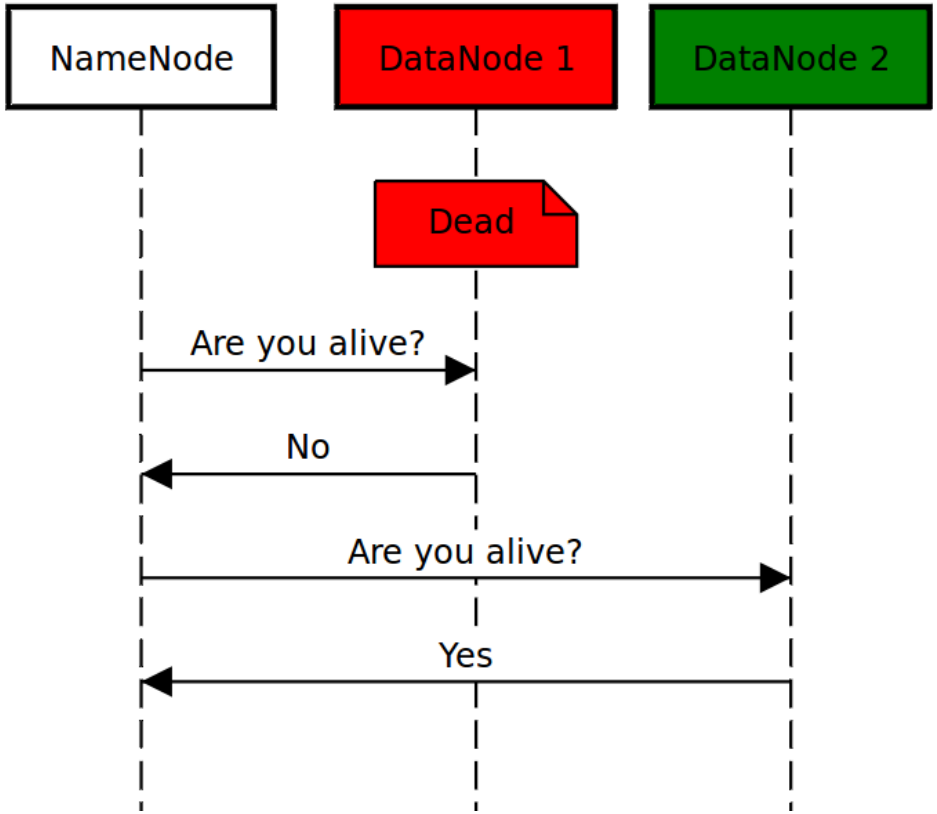
badoo

- Клиент обращается к NameNode с запросом getFile
- NameNode определяет **работающие** ноды, где хранится файл
- Сообщает клиенту адрес, куда обращаться

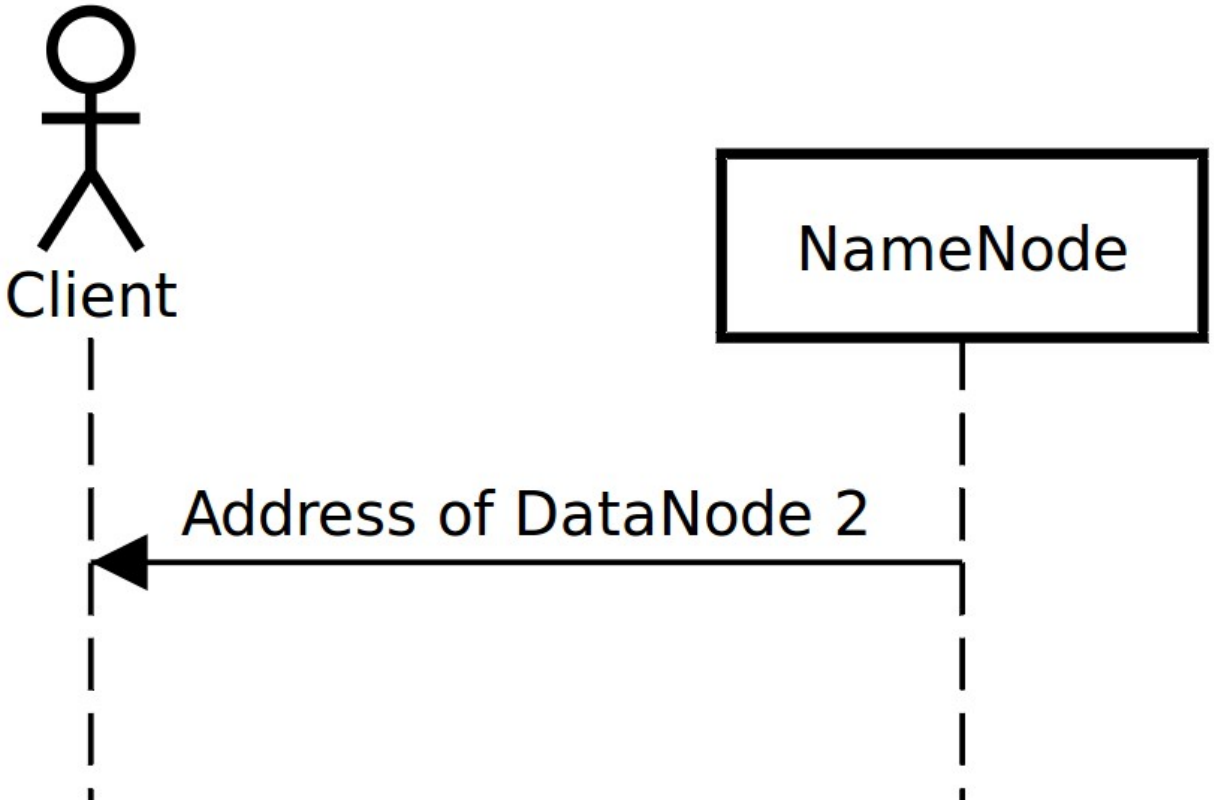
HDFS: чтение файла



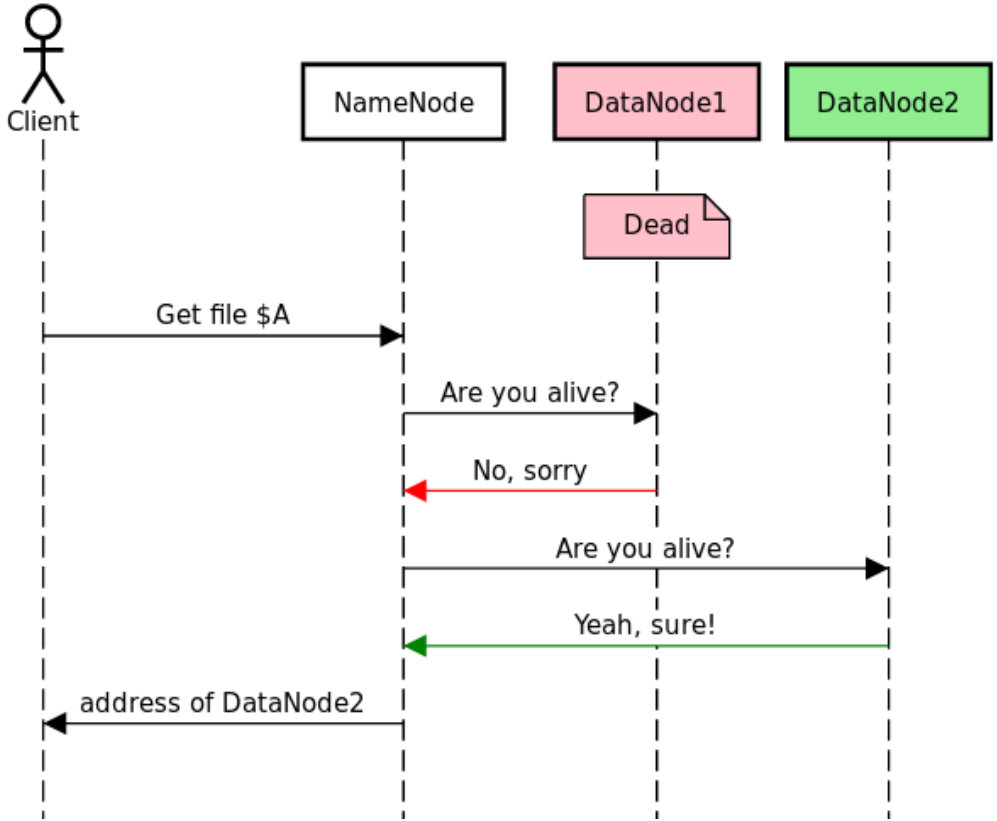
HDFS: чтение файла



HDFS: чтение файла



HDFS: запрос с отказом DataNode



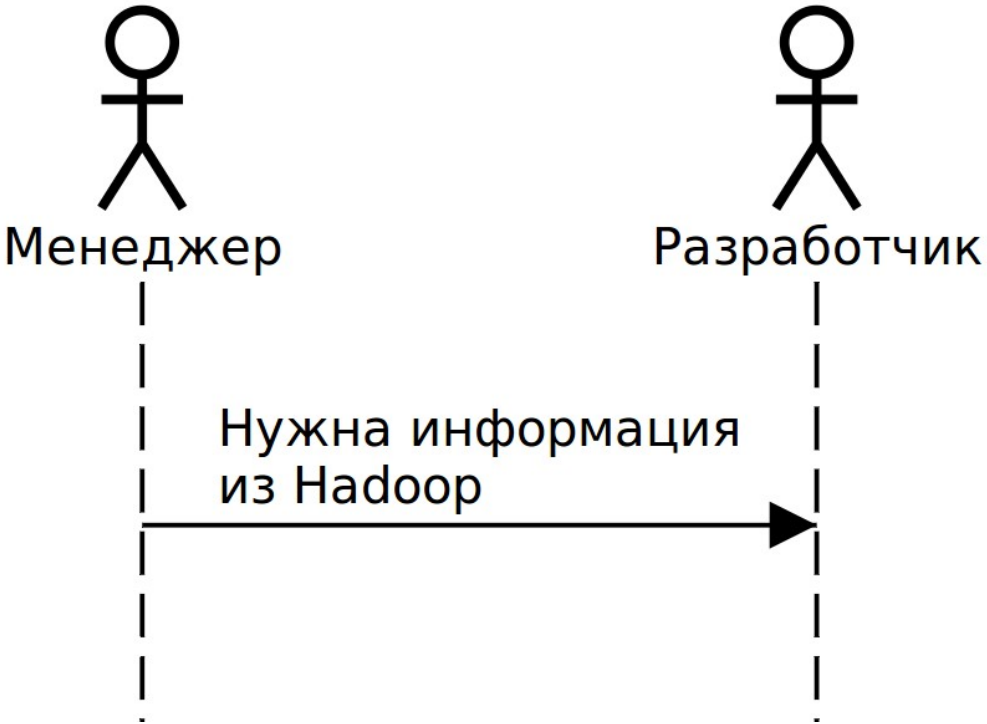
NameNode отправляет клиента на “живую” реплику файла

HDFS: реальность

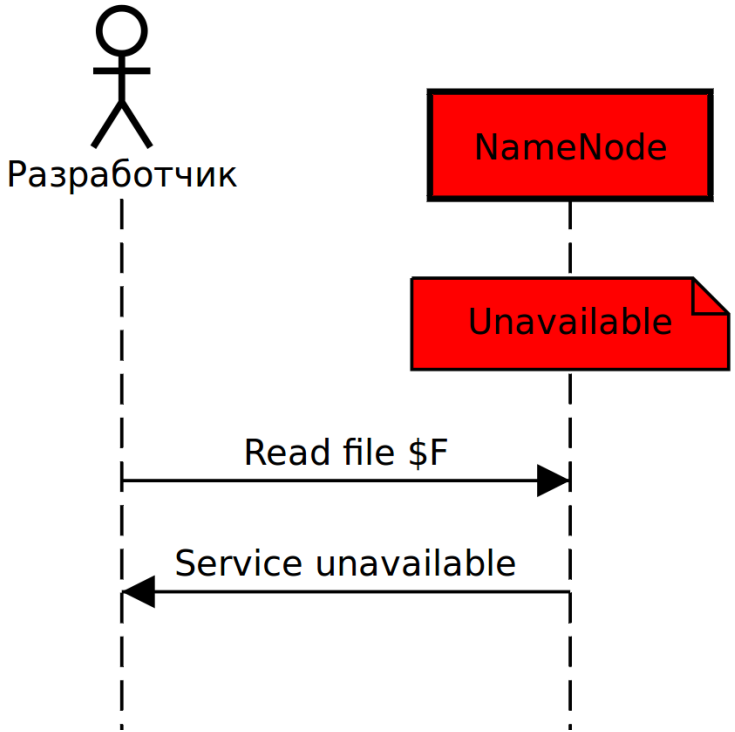
badoo

Что будет, если сервис NameNode недоступен для клиентов?

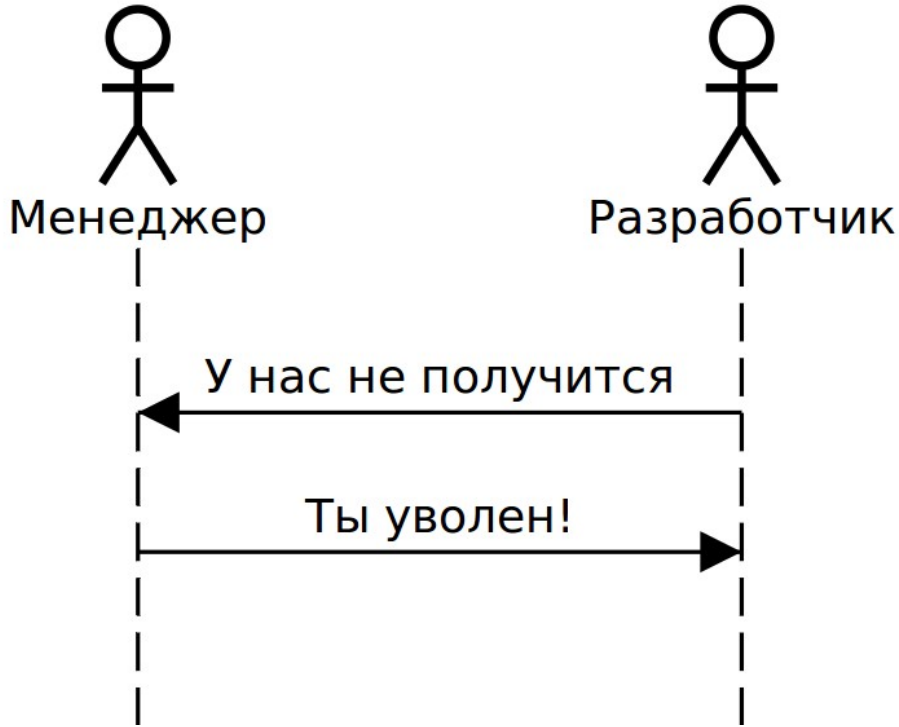
HDFS: запрос с отказом NameNode



HDFS: запрос с отказом NameNode



HDFS: запрос с отказом NameNode



HDFS: выводы

badoo

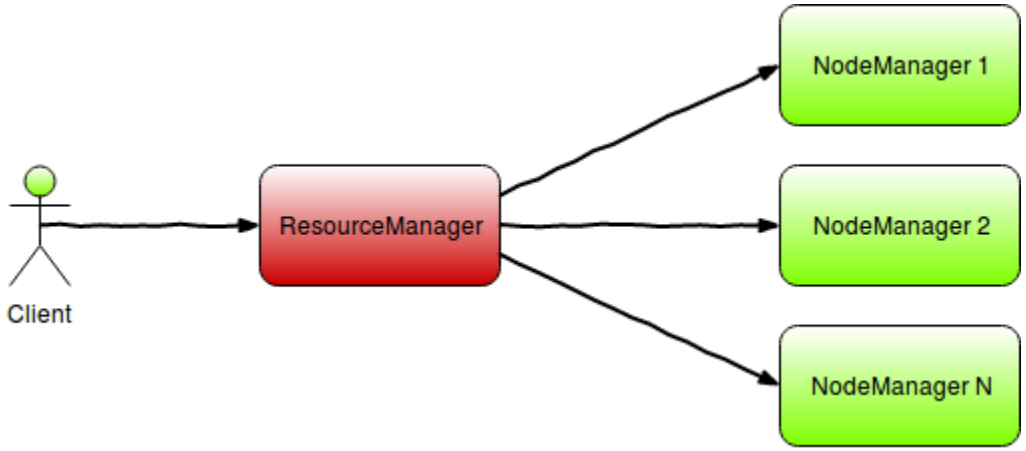
- Выход из строя DataNode не критичен, пока есть хотя бы одна, хранящая копию данных
- Выход из строя NameNode ведёт к полной недоступности HDFS

YARN

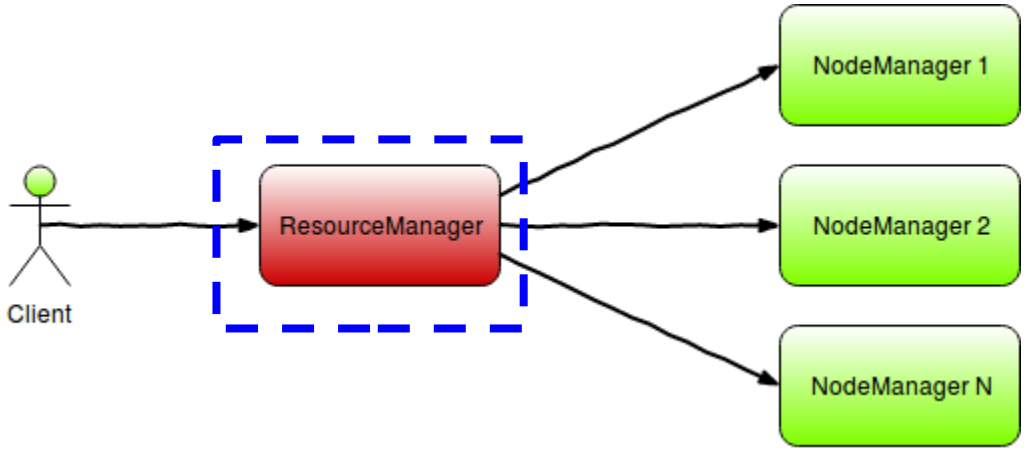
badoo

- Набор сервисов для выполнения вычислений на кластере
- API для запуска клиентских программ
- Менеджмент ресурсов

YARN


NodeManager 1
ResourceManager
NodeManager 2
Client
NodeManager N

YARN

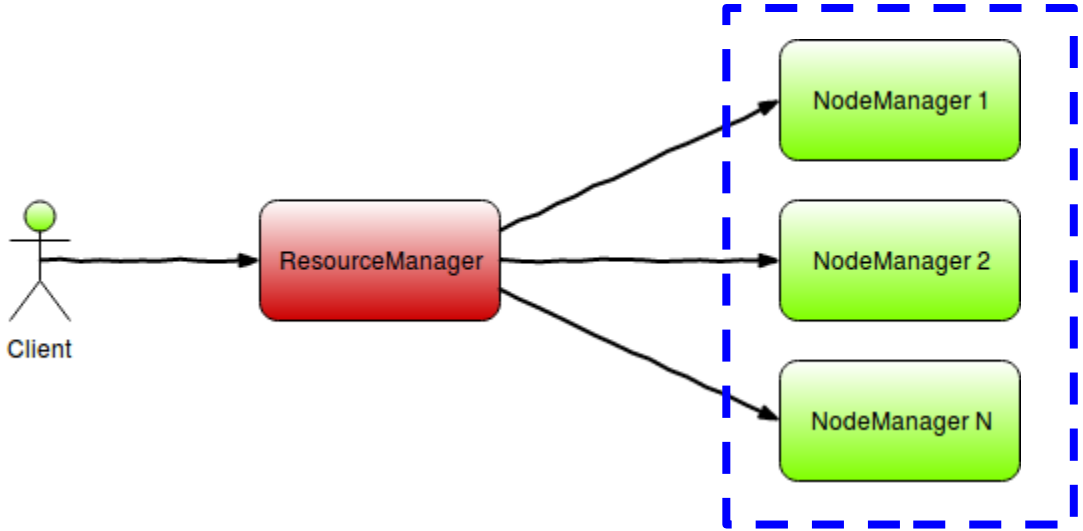

NodeManager 1
ResourceManager
NodeManager 2
Client
NodeManager N

YARN: ResourceManager

badoo

- Координатор вычислений на кластере
- API для запуска программ
- Менеджмент ресурсов кластера (CPU, RAM)

YARN


Client
ResourceManager
NodeManager 1
NodeManager 2
NodeManager N

YARN: NodeManager

badoo

- Сервис на “рабочих” машинах кластера
- Принимает запросы на создание JVM с заданным размером HEAP и резервированием CPU
- Исполняет пользовательский код в этих JVM

YARN: использование

badoo

- Клиент обращается к NodeManager за ресурсами
- NodeManager создаёт JVM (“контейнеры”) на машинах кластера
- Запуск клиентского кода в контейнерах

YARN: отказ NodeManager

badoo

Что происходит при недоступности ResourceManager?

YARN: отказ ResourceManager

Вспоминаем картинку про менеджера и разработчика

YARN: выводы

badoo

- Выход из строя NodeManager приводит к потере части ресурсов кластера
- Выход из строя ResourceManager’а приводит к невозможности запуска чего-либо на кластере

badoo

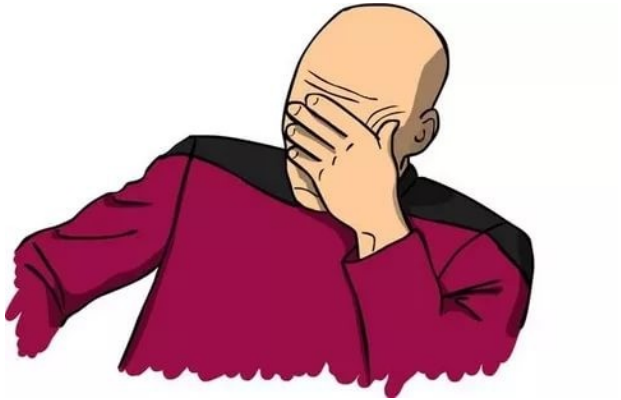
Опыт Badoo по эксплуатации Hadoop до HA

До HA: проблемы NameNode

- NameNode пишет все операции модификации данных в edit-log
- При старте NameNode, edit-log объединяются в образ файловой системы fsimage

До HA: проблемы NameNode

При высоком uptime перезагрузка могла достигать нескольких часов (sick!)



Варианты решения проблем NameNode

- перезапуск NN раз в месяц
- ручное поэтапное подсовывание edit-логов
- использовать свежий Hadoop :)
- правильно настроить edit-logs

Fuckup из жизни

1. Однажды мы не смогли подняться из edit-логов
2. Кластер лежит, пользователи бегают и судорожно машут руками
3. Админ пьёт валокордин
4. Члены команды BI изучают stack-trace ошибок

Реанимационные меры

1. Патч кода восстановления
2. Копирование edit-логов на отдельную машину
3. Восстановление с патченным кодом
4. Копирование fsimage обратно
5. PROFIT!
6. Downtime - почти сутки
7. Литры кофе и тонна нервов

Так зачем нужен HA?

При HA

- ✦ Fault-tolerance
- ✦ Низкий downtime
- ✦ Не теряем $$$
- ✦ Крепкий сон

Риски без HA

- ✦ HDFS => набор несвязных бинарей
- ✦ Вместо вычислительного кластера - куча несвязных машин


ЖЫЗНЬ С БОРОДОЙ
ЖЫЗНЬ БЕЗ БОРОДЫ
pikabu.ru

И что же делать?

High Availability в теории

badoo

- ✦Hot standby
- ✦Round-robin request balancing
- ✦Шардинг

High Availability в теории

badoo

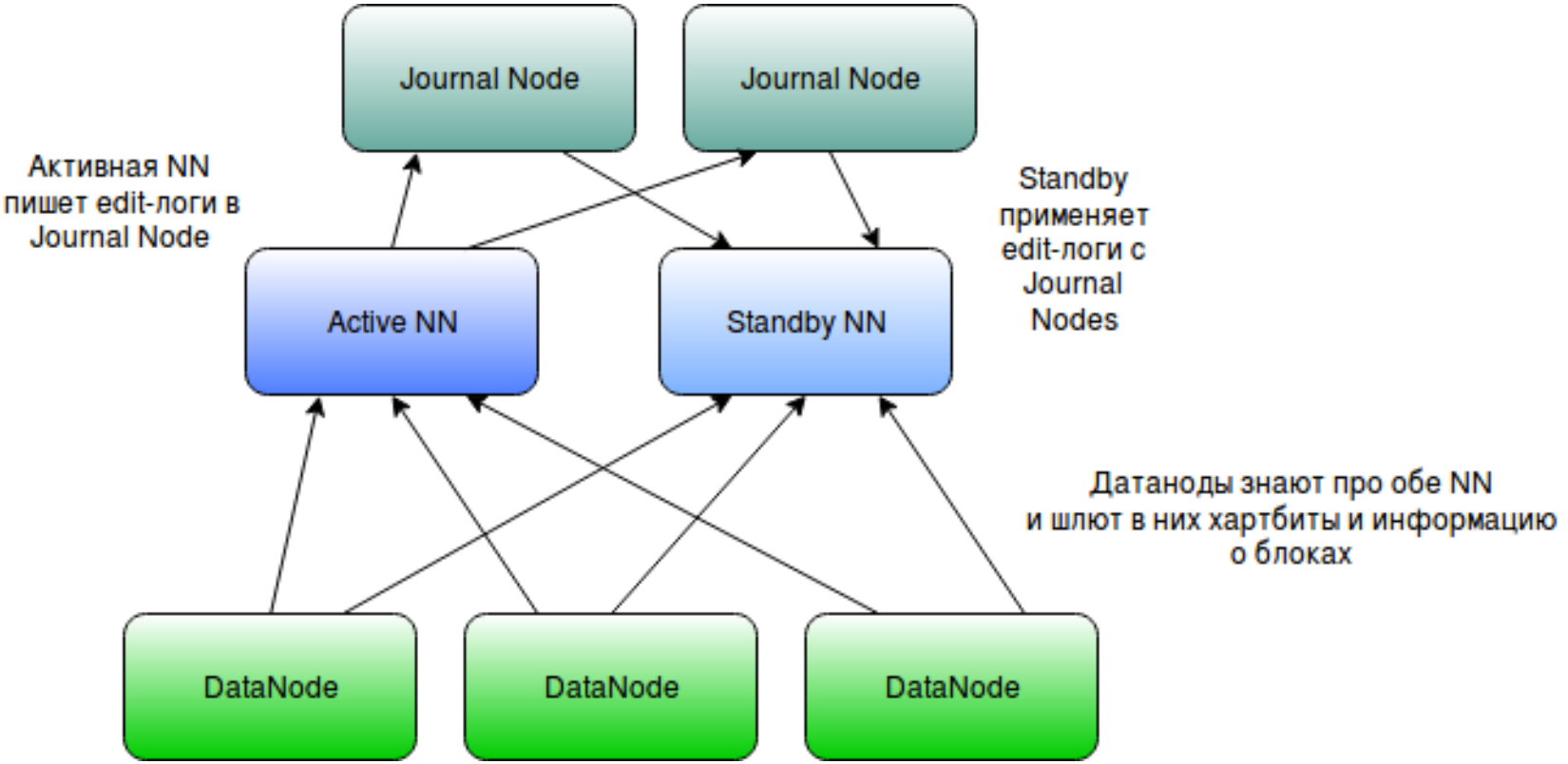
- Hot standby - используется в Hadoop для резервирования SPOF (NameNode, ResourceManager)
- Шардинг - Federated NameNode

HA HDFS: общая схема

badoo

- Требует развертывания дополнительных сервисов
- Предоставляет hot standby для NameNode

HA HDFS: общая схема



HA HDFS: общая схема

badoo

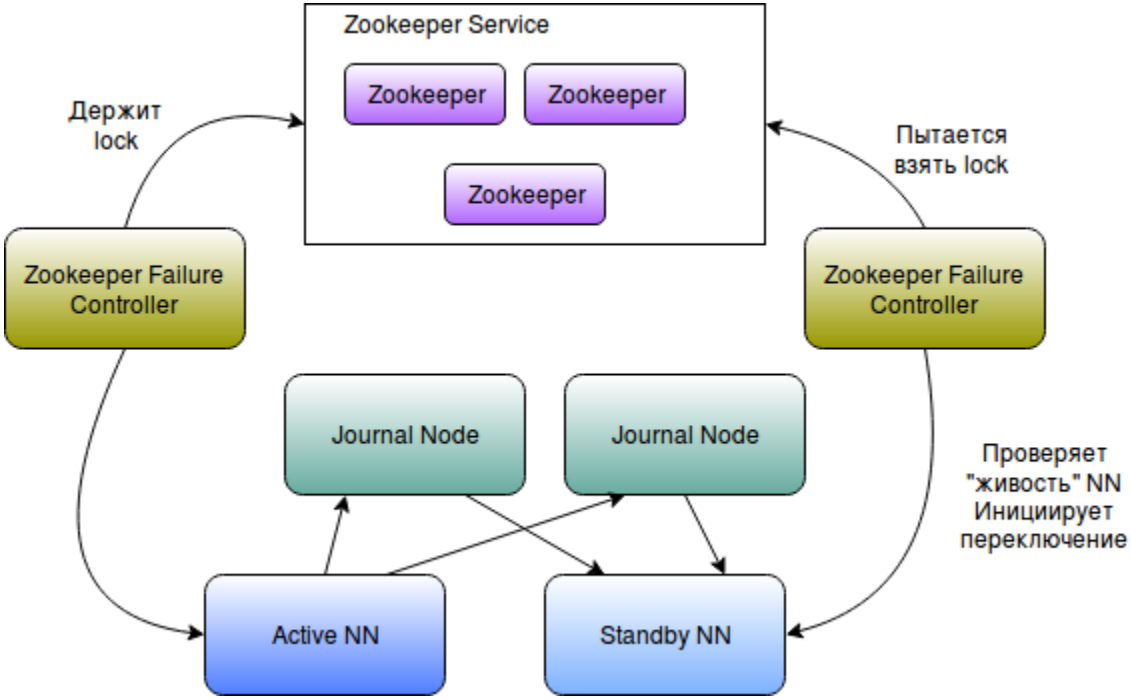
- Запросы обслуживает одна активная NN
- Есть реплика, читающая журнал с транзакциями

HA HDFS: сервисы

badoo

- Кластер Zookeeper
- Сервис ZKFC (failure controller)
- Сервис журналов (транзакций)
- Standby NameNode

HA HDFS: компоненты failover



HA HDFS: Zookeeper

Распределённая система с функционалом:

- lock-service
- leader election
- иерархическое хранение конфигураций

В рамках HDFS HA:

- выбирает активную NN
- оповещает клиентов об изменении статуса NN

HA HDFS: ZKFailoverController (ZKFC)

badoo

Сервис для автоматического переключения NN

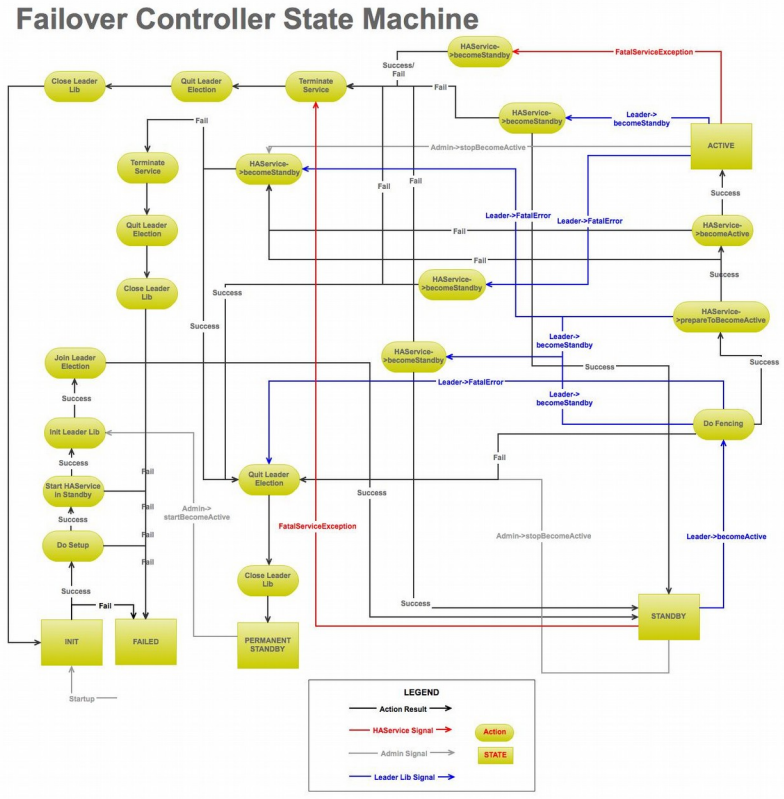
- запускается на машинах NN
- мониторит живость локальной NN
- подписан на уведомления от ZooKeeper
- производит failover

HA HDFS: JournalNode

badoo

- Распределённое хранилище транзакций NN
- Отсутствие SPOF - несколько экземпляров сервиса
- Запись в самую медленную реплику не тормозит общий процесс

Механизм failover



https://issues.apache.org/jira/browse/HDFS-2185

HDFS failover: summary

badoo

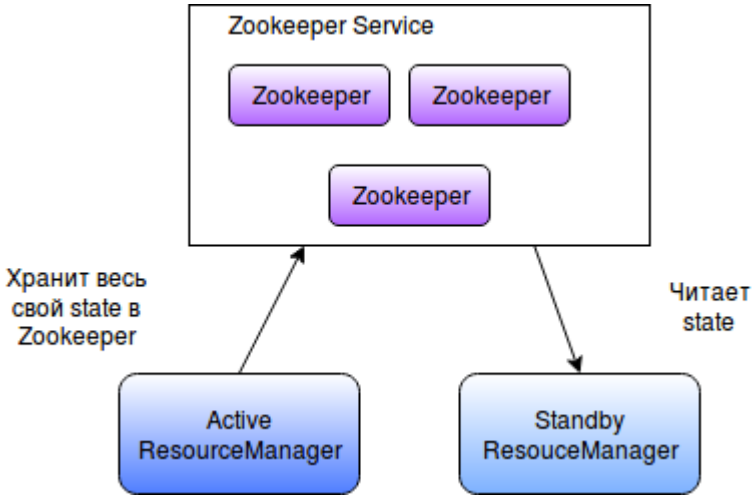
- Активная NN держит лок в ZK
- “Пропажа” лока - сигнал к переключению
- Произвольный ZKFC пытается сделать свою NN активной в ZK
- Если удалось - переводит остальные NN в standby, а локальную делает active

HA YARN: общая схема

badoo

- Hot standby для ResourceManager (входная точка кластера)
- Требует наличия ZooKeeper и второго экземпляра ResourceManager

HA YARN: общая схема



HA YARN: общая схема

badoo

- ResourceManager’ы используют механизм leader-election, предоставляемый Zookeeper
- Для репликации состояния мастера используется хранилище Zookeeper
- Контроллер failover’а встроен в сам демон RM, и не требует отдельного приложения

Адаптация клиентских приложений

HDFS: клиентский API

badoo

- Главная мысль - переход к HA потребует изменения клиентского кода
- Об этом не принято широко писать в документации

HDFS: клиентский API

badoo

- Native API

  клиенты - в основном - JAVA

  JAVA-клиенты знают про конфигурацию HA

- WebHDFS

  HTTP-based, для любого языка

  клиенты не знают, какая NN активная

WebHDFS: standby NN

```
curl http://stanbynn.domain:50070/webhdfs/v1/?op=LISTSTATUS

{
  "RemoteException":
    {
    "exception":"StandbyException",
    "message":"Operation category READ is not supported in
state standby"
    }
}
```

Standby-нода не обслуживает клиентские запросы

http://$host:50070/jmx?qry=Hadoop:service=NameNode,name=NameNodeStatus

WebHDFS: standby NN

badoo

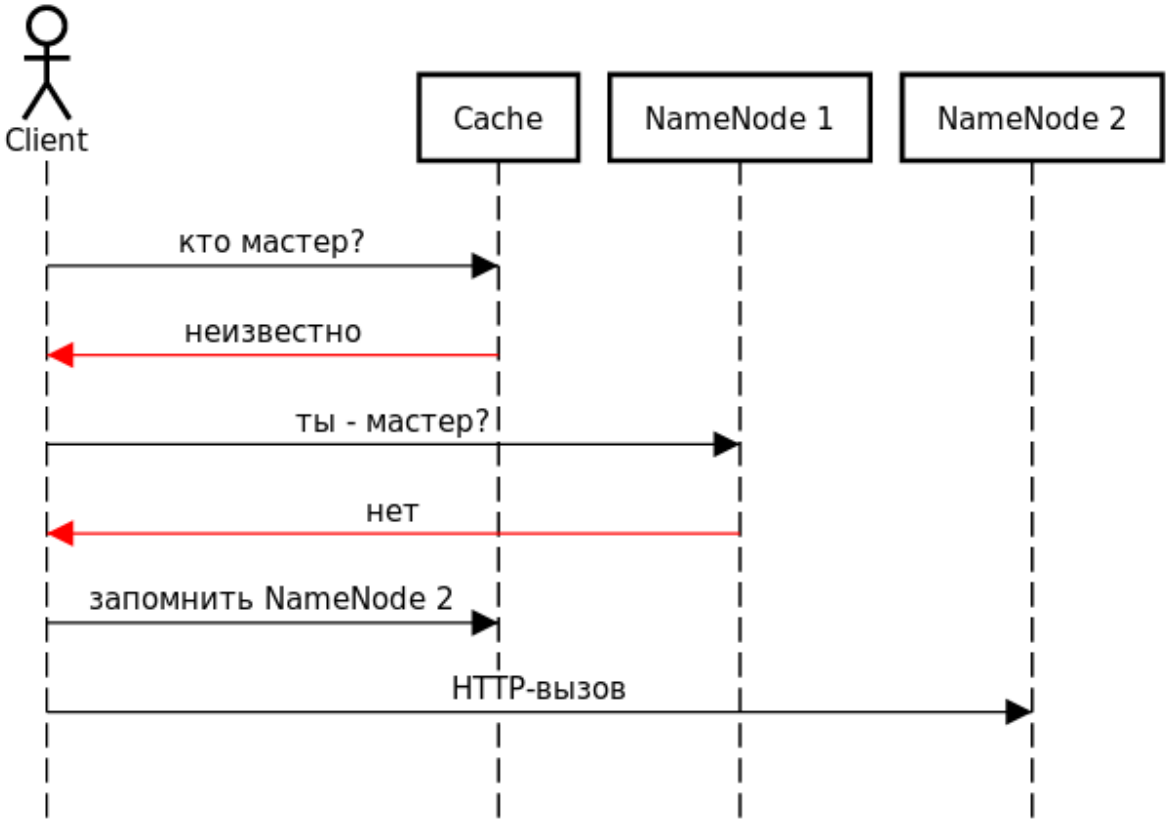
- Определённо, должен быть workaround!
- Мы придумали 3 варианта
- Надо выбрать один :)

Select master: cached metadata

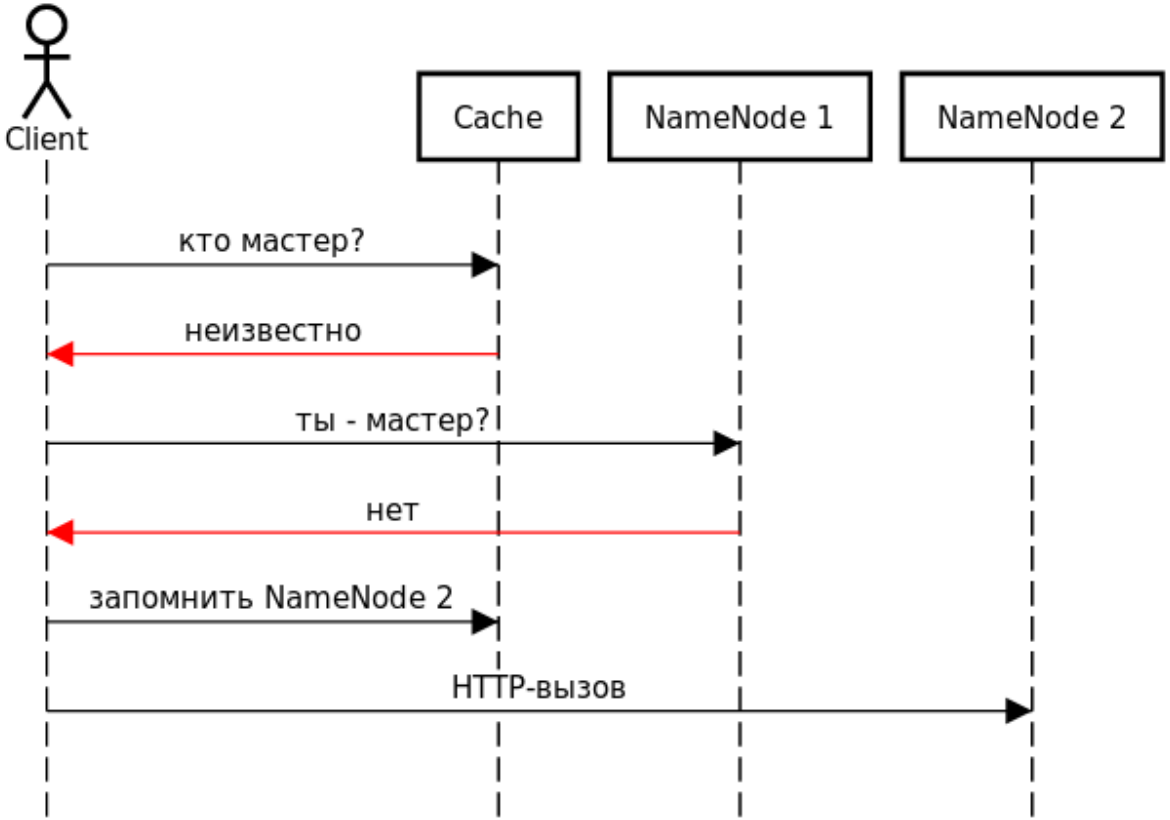
badoo

- Иметь на клиентах список всех NameNode
- Последовательно опросить их, вычислив мастера
- Положить в кеш адрес активного сервиса

Select master: cached metadata



Select master: cached metadata



Pro:

1. Просто, как валенок

Cons:

2. Лишний HTTP-вызов
3. Имплементация под каждый ЯП
   (у нас их много :)

WebHDFS: software balancer

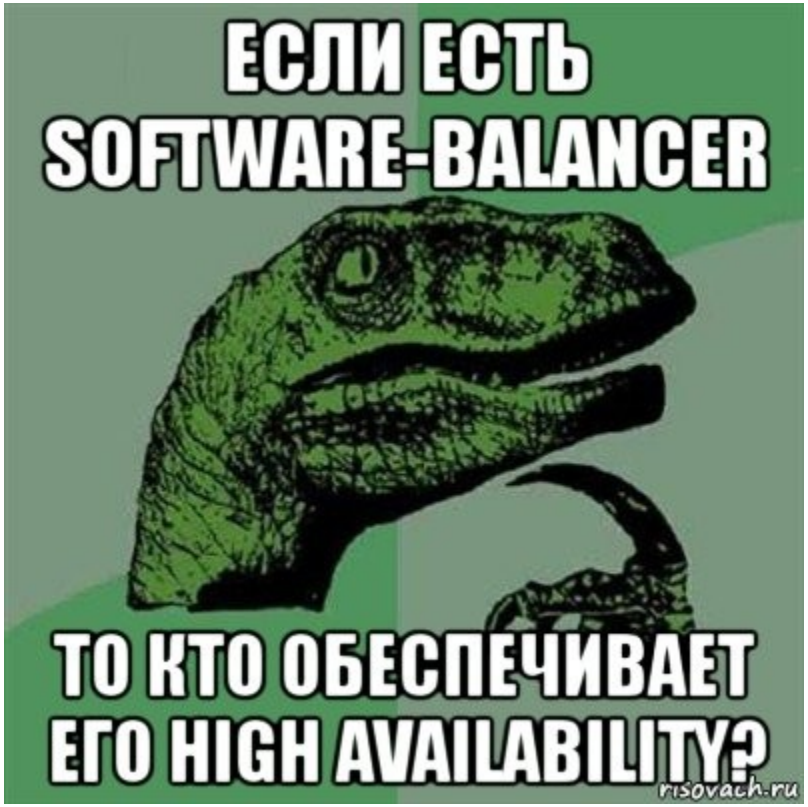
badoo

- Проксирование запроса через балансер
- Определить список upstream для NameNode
- Описать способ валидации (проверочный URL)

Преимущества software balancer

badoo

1. Простота конфигурации
2. Более одного решения
   Nginx
   HAProxy
   ваш любимый


ЕСЛИ ЕСТЬ
SOFTWARE-BALANCER
ТО КТО ОБЕСПЕЧИВАЕТ
ЕГО HIGH AVAILABILITY?
risovach.ru

Select master: hardware balancer

badoo

- Физическая железяка
- В Badoo в каждом ДЦ
- Маршрутизирует весь трафик
- Имеет функции и возможности тюнинга, превосходящие программный балансер
- Умеем его готовить

Советы из опыта эксплуатации (software)

HDFS: rock your NameNode

badoo

- Простое правило:
  1M файлов = 1 Gb heap size для NN
- В противном случае наблюдается деградация
- Не влезли в RAM одного хоста - добро пожаловать в Federated HDFS (мы пока не доросли до него)

HDFS: rock your NameNode

badoo

- Вдохновлялись статьей от Hortonworks
- Асинхронное логирование HDFS
- Достаточное число потоков для обработки API запросов
- Выключили atime
- Патч для socket listen queue - HADOOP-14560
- Заменили JSON либу WebHDFS

Тюнинг Zookeeper

badoo

- Обязательно!
  Настройка ротации snapshot
- Обязательно!
  -XX:OnOutOfMemoryError="kill -9 %p"
- Просто примите это за реальность

Советы из опыта эксплуатации (hardware)

HDFS: datanode

- Т.к. репликация программная, то нет большого смысла использовать RAID
- JBOD - наш друг: сила в шпинделях!

HDFS: выбор диска для блоков

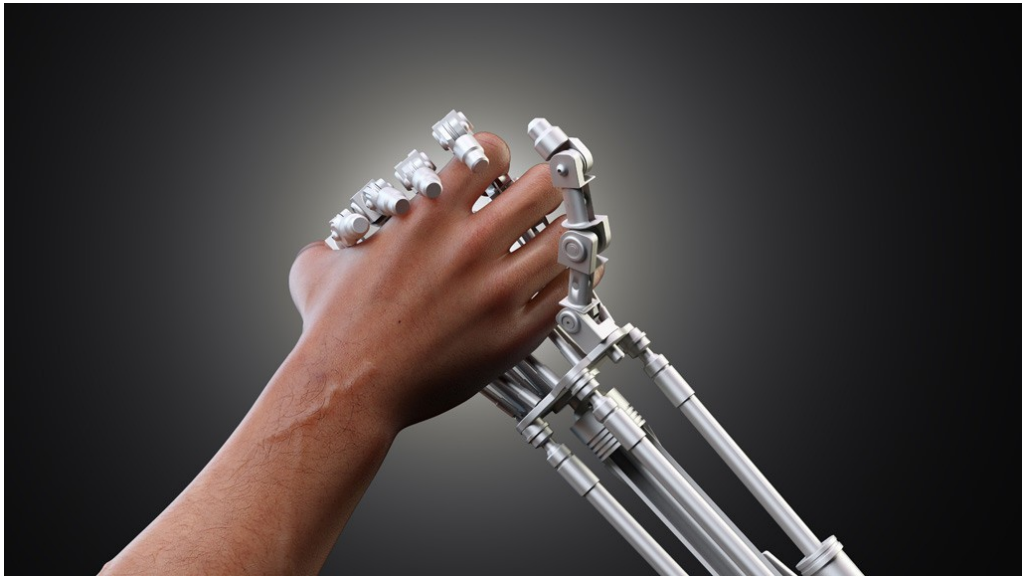
```
akrasheninnikov@bihadoop63.mlan:~> df -h
Filesystem                      Size  Used Avail Use% Mounted on
devtmpfs                         95G  8,0K   95G   1% /dev
tmpfs                            95G     0   95G   0% /dev/shm
tmpfs                            95G  9,3M   95G   1% /run
tmpfs                            95G     0   95G   0% /sys/fs/cgroup
/dev/mapper/root_vg-root_vol     20G  3,4G   16G  18% /
/dev/mapper/root_vg-local_vol    48G  4,7G   42G  11% /home
/dev/sda1                       486M   43M  418M  10% /boot
/dev/sda3                       1,6T  1,4T   82G  95% /local/hadoopdata/1
/dev/sdc1                       1,7T  1,4T  193G  88% /local/hadoopdata/3
/dev/sdi1                       1,7T  1,4T  152G  91% /local/hadoopdata/9
/dev/sdb1                       1,7T  1,4T  153G  91% /local/hadoopdata/2
/dev/sdg1                       1,7T  1,4T  155G  91% /local/hadoopdata/7
/dev/sdh1                       1,7T  954G  613G  61% /local/hadoopdata/8
/dev/sde1                       1,7T  955G  612G  61% /local/hadoopdata/5
/dev/sdf1                       1,7T  1,4T  150G  91% /local/hadoopdata/6
/dev/sdj1                       1,7T  947G  620G  61% /local/hadoopdata/10
/dev/sdd1                       1,7T  991G  575G  64% /local/hadoopdata/4
```

HDFS: выбор диска для блоков

- By default - Round Robin
- Но на первом диске - OS => он заведомо меньше
- Всегда образуется перекос
- Решение 1: настройка для датаноды AvailableSpaceVolumeChoosingPolicy
- Решение 2: Hadoop 3 + междисковый балансер

Человек против машины

badoo



Может ли один аналитик уложить кластер?

Нет ничего невозможного!

badoo

- Пишет 4-х этажный SQL
- Запускаем, ждём 5 минут
- Забиваем IN/OUT eth на датанодах
- Забиваем свитч
- Залезаем в iowait
- PROFIT!

How to avoid

badoo

- На датанодах - 10Gb eth

или:

- Стараемся уважать data-locality
  (работаем с данными по месту их физического хранения)

Что вы узнали

Что вы узнали

- Из каких базовых сервисов состоит Hadoop
- Какие там существуют SPOF
- С помощью каких средств они устраняются
- Какие действия предпринять для адаптации клиентов кластера к HA

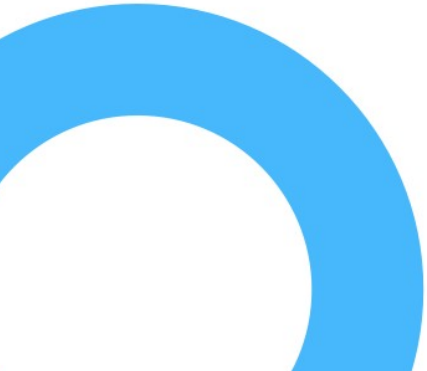
Полезные материалы

badoo

- https://tech.badoo.com
- Доклад Badoo про использование Hadoop
- Инструкции по настройке HA
- Тюнинг NameNode
- Near-realtime аналитика на Hadoop

Кажется, я обещал 2 картины?

Спасибо!



krash@corp.badoo.com
krash3@gmail.com

badoo