Введение в Big Data и распределенные файловые системы

Алексей Драль, aadral@bigdatateam.org
CEO at BigData Team, http://bigdatateam.org
https://www.facebook.com/bigdatateam

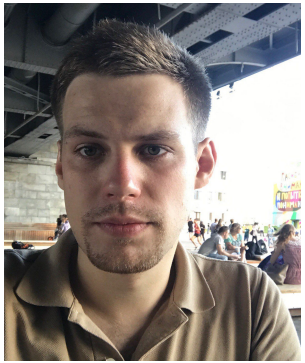
18.02.2021, Moscow, Russia

BIGDATA TEAM

Разработчики курса

Алексей Драль

BIGDATA TEAM



Артем Выборнов

RAMBLER GROUP

Андрей Титов

NVIDIA

Антон Горохов

Yandex

Павел Клеменков

NVIDIA

и многие другие!

Канал Telegram

Виктория Брынза

Обо мне

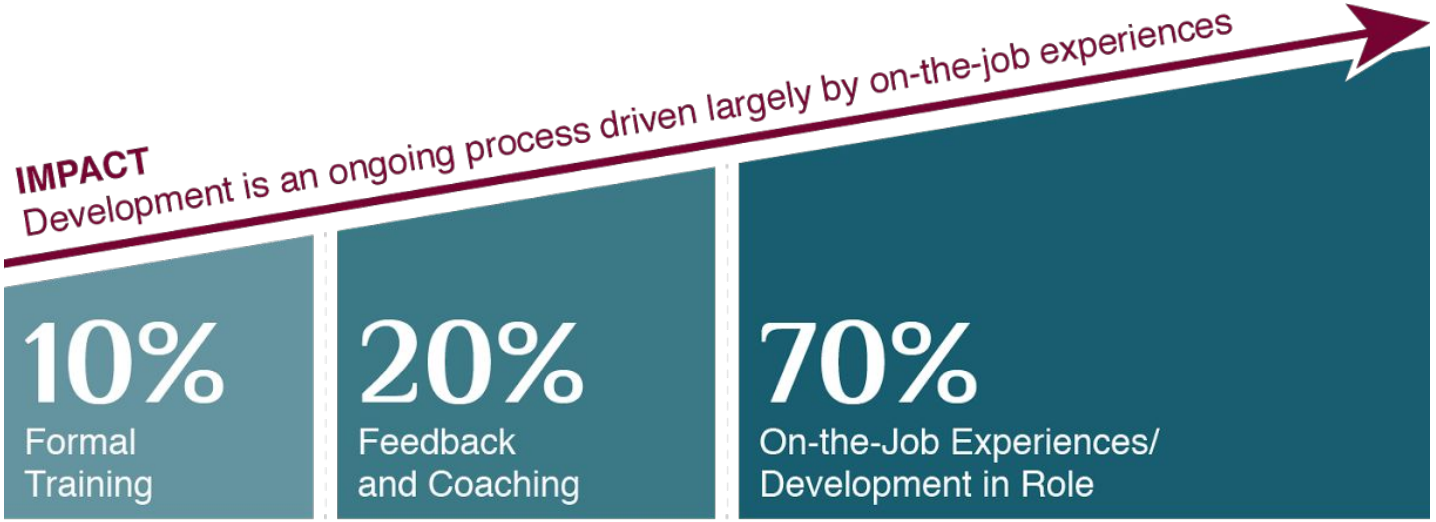
- СУНЦ МГУ
- МГУ
- ШАД Яндекса
- Рамблер
- Яндекс
- Amazon AWS
- МФТИ
- Сбербанк
- BigData Team

BIGDATA TEAM

План курса

1. Введение в Big Data. Распределенные файловые системы, HDFS
2. Hadoop MapReduce
3. Оптимизация MapReduce вычислений
4. Hive: SQL поверх больших данных
5. Модель вычислений Spark: RDD
6. Spark DataFrames, Spark SQL
7. Оптимизация Spark вычислений
8. Потоковая обработка данных (Kafka, Spark Streaming)
9. Cassandra + Spark
10. Оптимизация хранилища (Data Layout)

Артефакты обучения



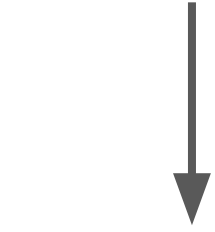
The 70-20-10 rule was developed by Morgan McCall, Robert W. Eichinger and Michael M. Lombardo at the Center for Creative Leadership.

- 10 домашних заданий
- 2 тестирования

BIGDATA TEAM

Метрики F2F обучения

до
20-30%



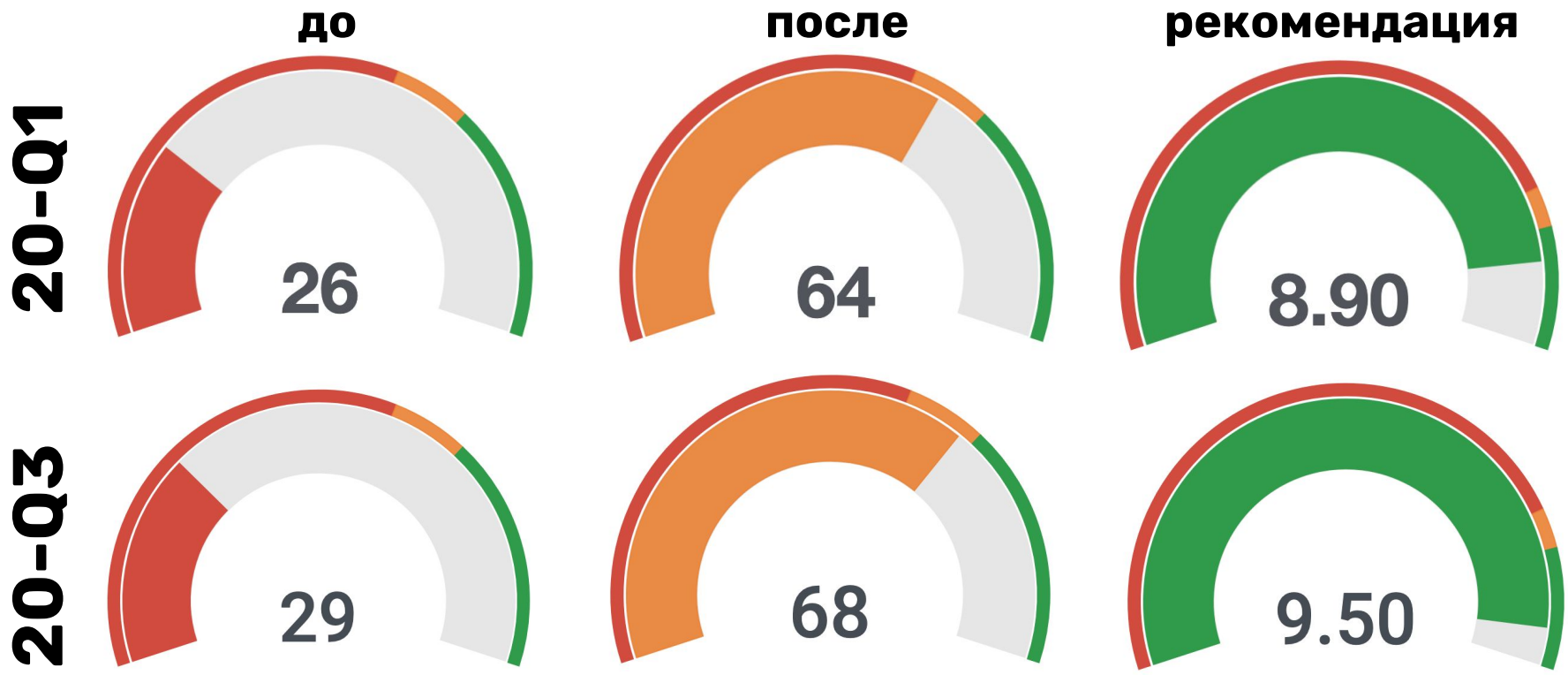
после
70-80%

HDFS	MR	MR	MR	MR	Hive	Hive	Hive	Hive	Spark	Spark	Spark	Spark	Spark	RT	RT	RT	RT	NoSQL	NoSQL
0	0	0.75	0	0	0	0	0	1	1	0	0	1	0	0	0.66	0	0	0	0
0.5	0	1	0	0.6	0	1	0	1	0.75	0	0	1	0	0	0.66	0.66	1	1	1
0	0	1	0	0	0	0	0	1	0.5	0	0	0	0.5	0	0	0	0	0	0
0	1	0.75	0	0.4	0	0	0	1	0.75	1	0	0	0	1	0	0	0	0	0
0.75	0	1	0	0.8	0	0	0	1	0.75	1	1	1	1	0	0	0	0	0	0
0.25	0	1	0	0	0	1	0	1	0.5	1	0	0	0.5	0	0.66	0.66	0	0	0
0.75	0	1	0	0.8	0	0	0	1	0.75	0	0	0	0	0	0.99	0.66	1	0	0
0.75	1	1	1	0.8	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0
0	0	0.75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.75	0	0.5	0	0	0	0	0	1	1	0	1	0	0.75	0	0.33	0.66	1	1	0
0.5	0	0.75	0	1	0	0	1	1	0.75	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	1	0	0	0	1	1	1	1	0	0.75	0	0.99	0.33	1	0	0
0.75	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0
0.25	0	0.75	1	0	0	0.75	0	1	0.75	1	1	1	1	1	0	0	0	0	1
1	0	0.5	0	0	0	0	0	0	0.75	0	0	0	0	0	0	0	0	0	1
0.75	0	1	0	0	0	0	0	1	0.75	0	0	0	1	1	0.66	0	0	0	0
0.5	0	1	0	1	1	0	1	1	0	1	1	0	1	0	0	0	0	0	0

HDFS	MR	MR	MR	MR	Hive	Hive	Hive	Hive	Spark	Spark	Spark	Spark	Spark	RT	RT	RT	RT	NoSQL	NoSQL
1	0	1	1	0.8	1	1	1	1	1	1	1	1	0.6	1	1.00	1.00	1	0.25	0.75
0.75	1	1	1	0.6	1	1	1	0.75	0.5	1	0.5	1	0.6	1	1.00	1.00	1	1	1
1	0	1	1	1	1	0	1	0.5	0.5	0	0	1	1	0	0.67	1.00	0	0.5	0.75
1	0	1	1	1	1	0	1	0.5	0.5	0	0	1	1	0	0.67	1.00	0	0.5	0.75
0.5	1	1	1	1	1	1	1	0.75	0.25	1	0.75	1	0.8	1	1.00	1.00	0	0.75	1
1	1	1	0	0.2	1	0	1	0.75	1	0	1	0	1	1	1.00	0.50	0	0.75	0.75
0.5	1	0	0	0.8	0	0	0	0.5	0.75	0	0.75	1	0.6	0	0.67	0.50	0	0.5	0.75
0.5	1	0	1	0.8	1	1	0	0.75	0.75	1	0.5	1	0.8	1	1.00	0.50	1	0.5	1
0.75	1	1	1	1	1	0	0	1	0.5	0	0.5	1	0.8	1	0.33	0.83	0	0.75	0.5
1	1	1	1	0.8	1	1	1	0.75	1	0	1	1	0.6	1	1.00	1.00	0	0.75	1
1	0	1	1	1	1	0	0	1	0.5	0	1	1	0.6	1	1.00	0.50	1	0.75	1
0.75	0	1	1	1	1	1	1	0.75	1	0	0.5	0	0.8	1	1.00	1.00	1	0.5	1
0.75	1	1	1	0.6	1	1	1	0.75	0	0	0.5	1	0	0	0.33	0.00	1	1	1
1	0	1	1	0.8	1	0	1	0.75	0.25	0	0.5	1	1	1	0.67	1.00	1	1	1
0.5	1	1	1	0.8	1	0	0	1	1	1	0.5	1	0.8	0	1.00	0.67	0	0.75	0.25

BIGDATA TEAM

Точка отсчета 21-Q1



Intro quiz: https://rebrand.ly/x5bd2021q1_quiz_intro

План занятия

- Мотивация работы с Big Data
- Многопроцессорные вычислительные системы
- Распределенные файловые системы (GFS / HDFS)
- Hadoop Sizing и архитектура Namenode
- HDFS Workshop

Мотивация работы с Big Data

Big Data: Motivation

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Q&A

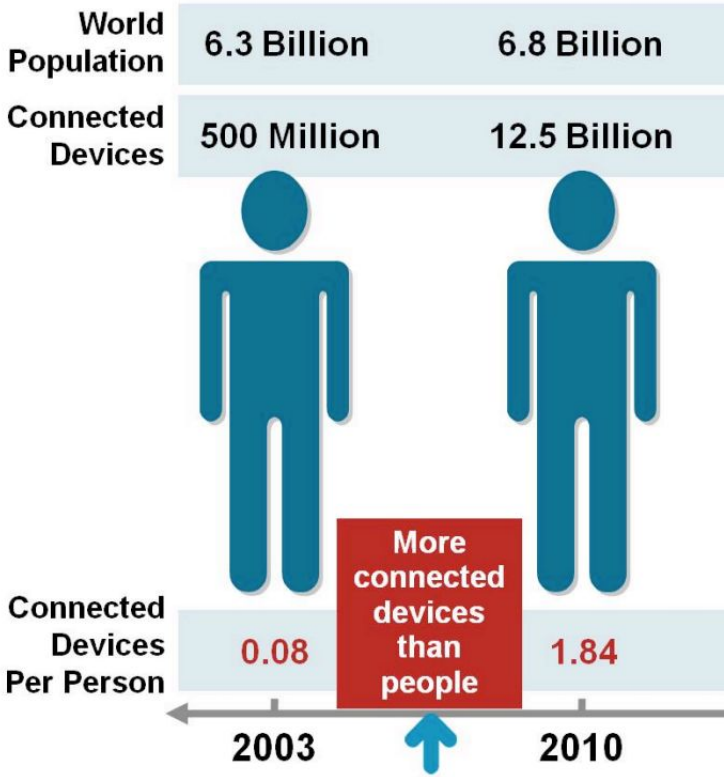
Откуда берутся (большие) **данные**?

BIGDATA TEAM

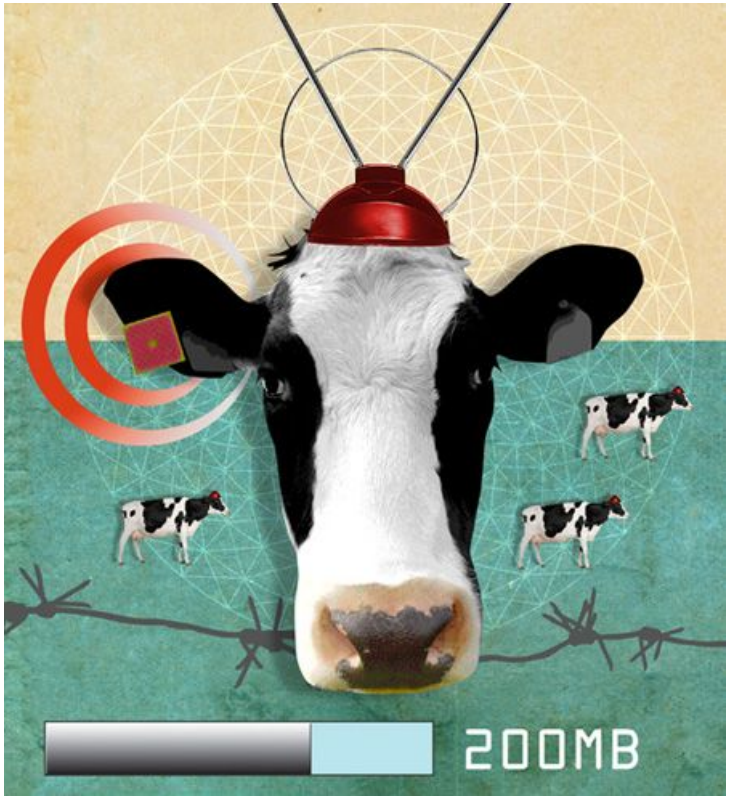
Internet of Things (IoT)

Q&A: IoT vs Human Beings?

Internet of Things (IoT)



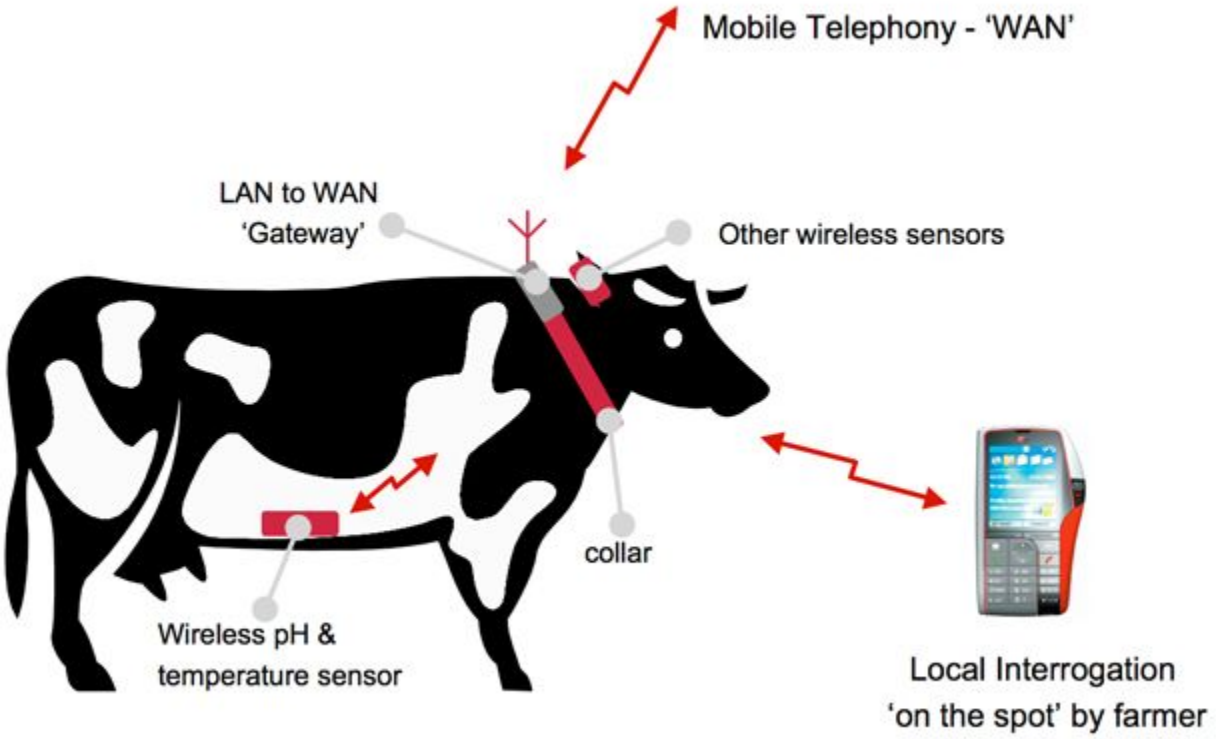
Source: Cisco IBSG, April 2011



Source: The Economist, 2010

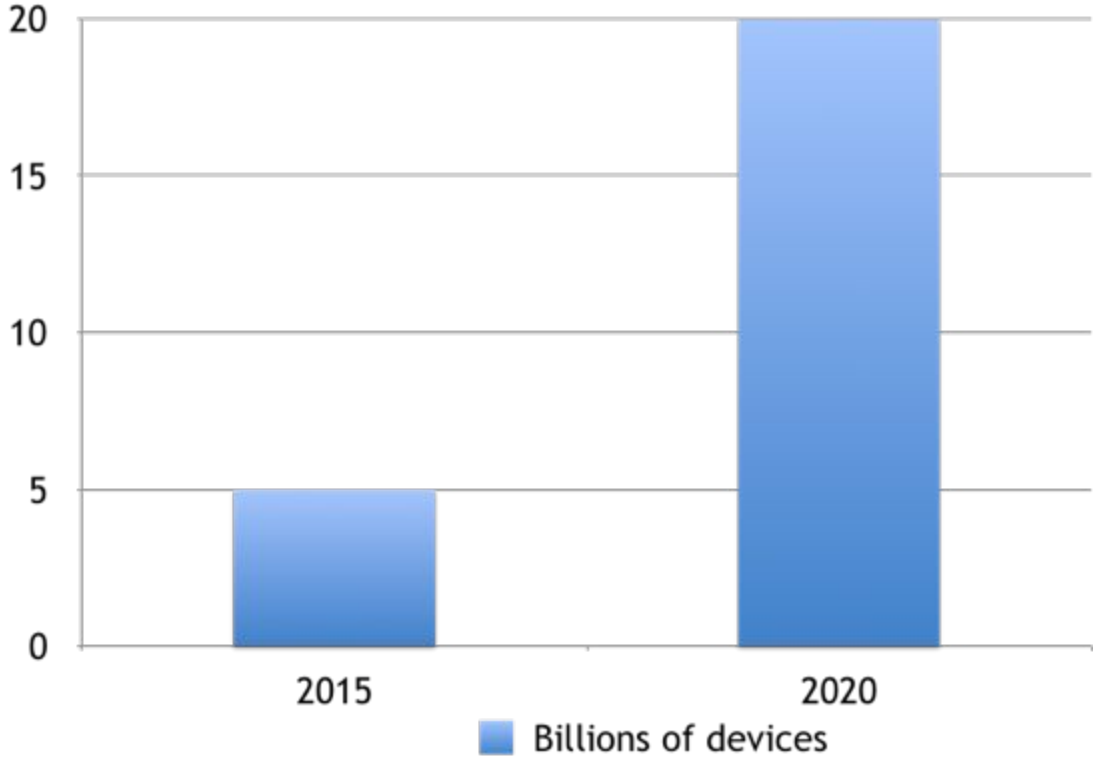
BIGDATA TEAM

Internet Cattle



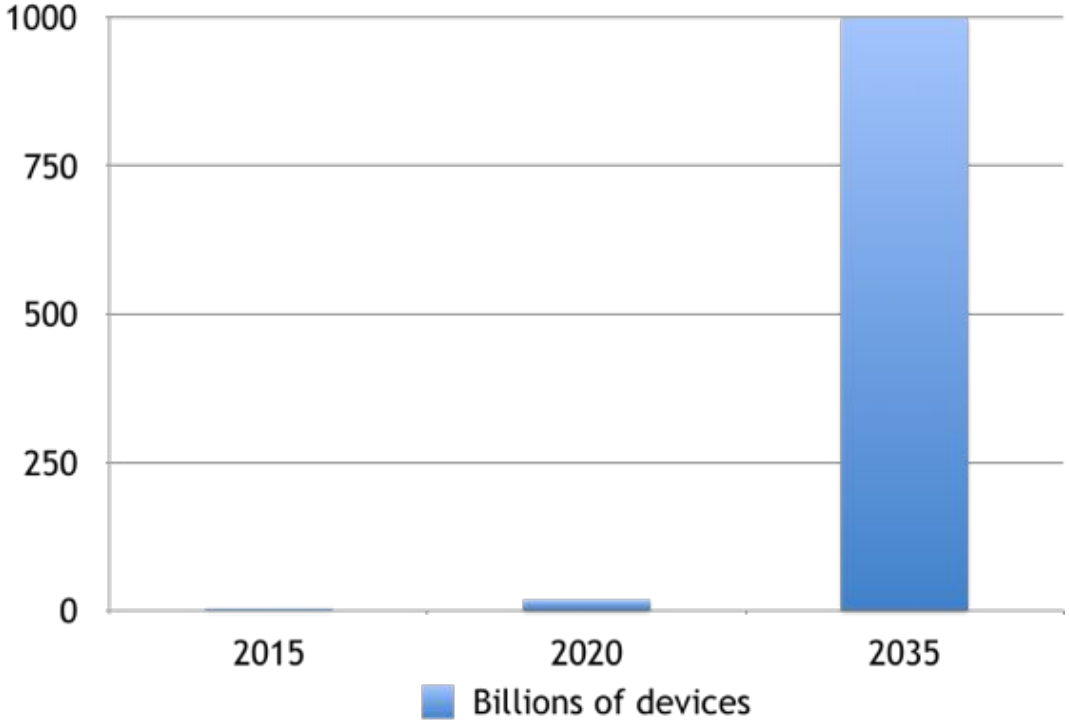
https://www.theengineer.co.uk/internet-cattle/ Scotland (16.08.2017)

Internet of Things



Gartner

Internet of Things

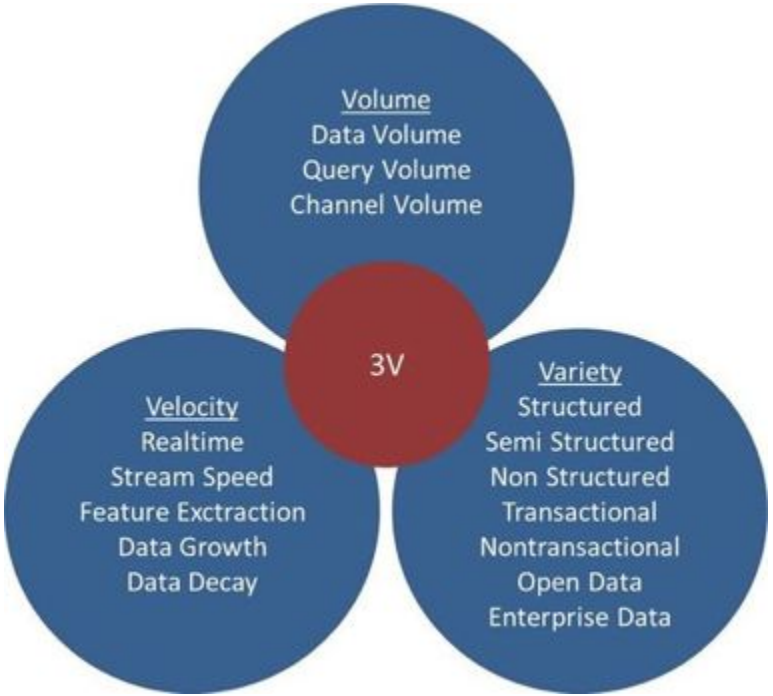


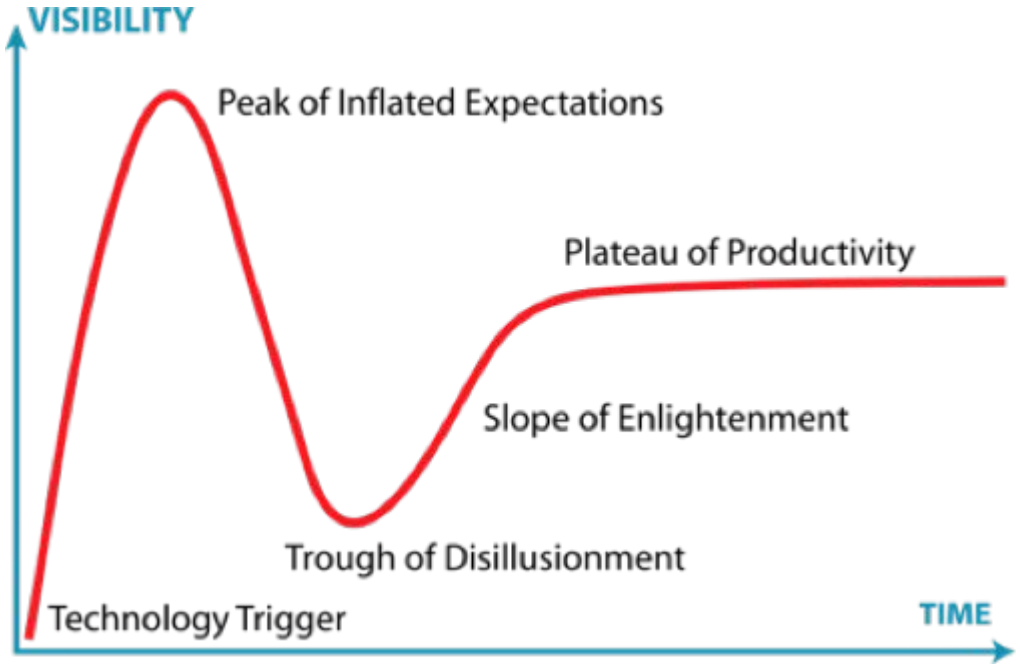
Gartner

Q&A

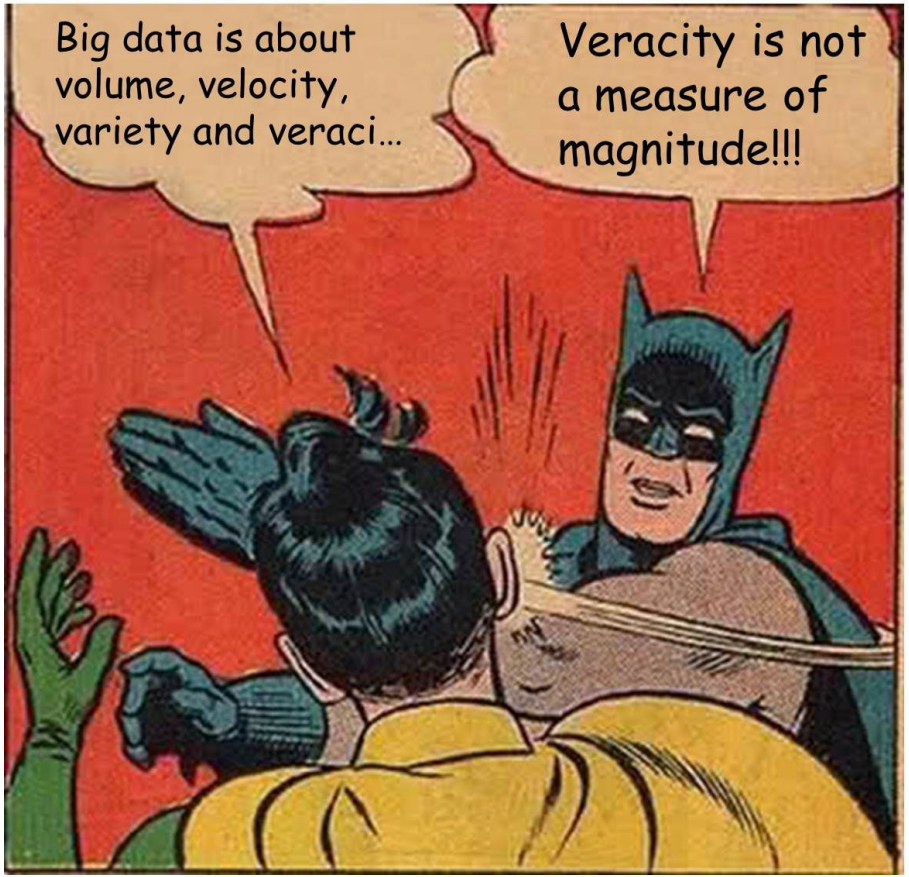
Что такое Big Data?

Big and Small Data





4+ Vs of Big Data



https://blogs.gartner.com/doug-laney/batman-on-big-data/

Q&A

Где применяется (data) **science**?

Q&A: Why Big Data?

Data Scientist? Data Engineer? Data …?

indeed

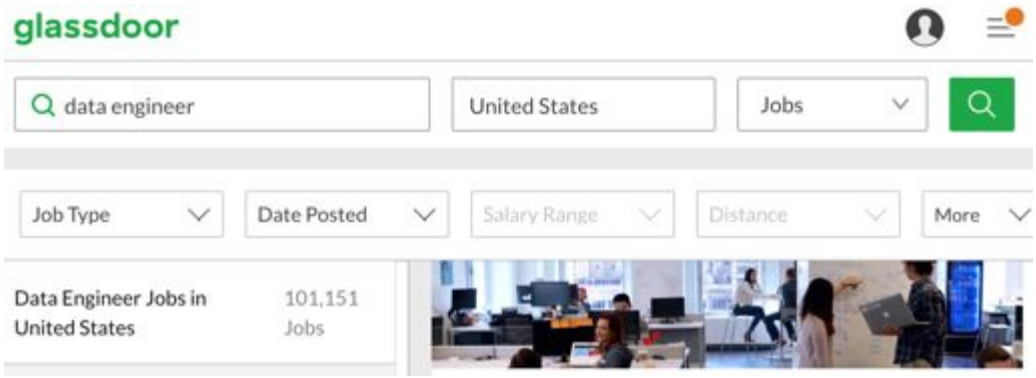
Index jobs

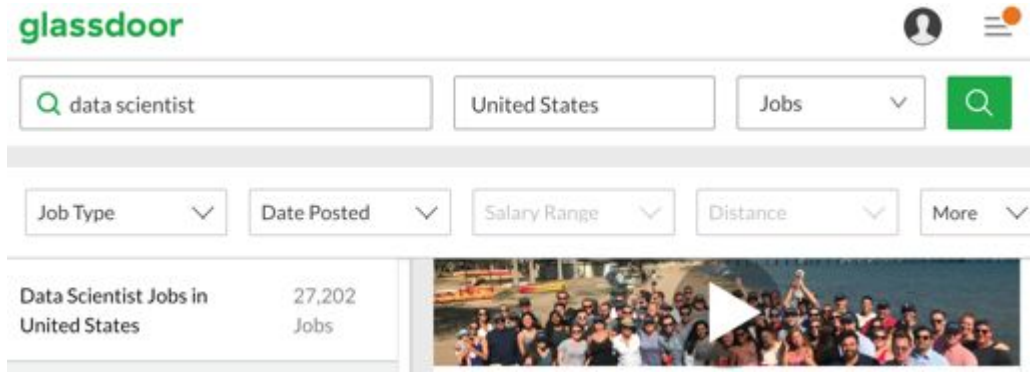
My recent searches

data engineer - 34,228 new

data scientist - 9,279 new

4x





BIGDATA TEAM

Coursera Specialization on Big Data

Capstone Project

Hadoop, Spark **Hive, Spark** **Spark ML** **Real Time**

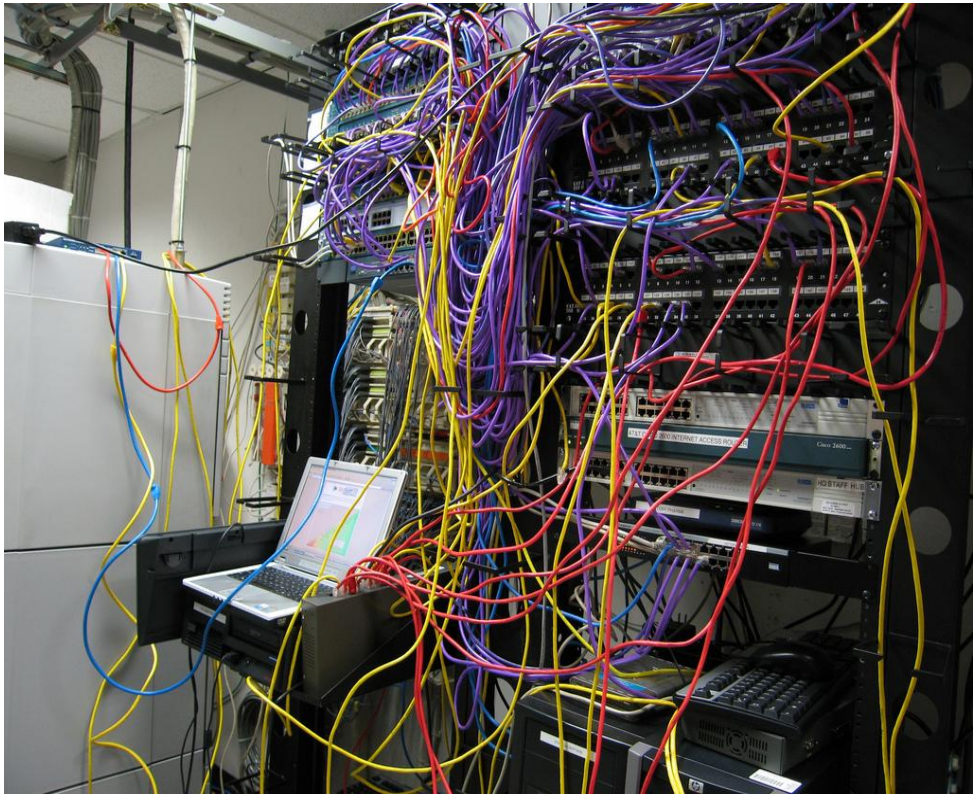
Yandex

https://bigdatateam.org/big-data-engineering

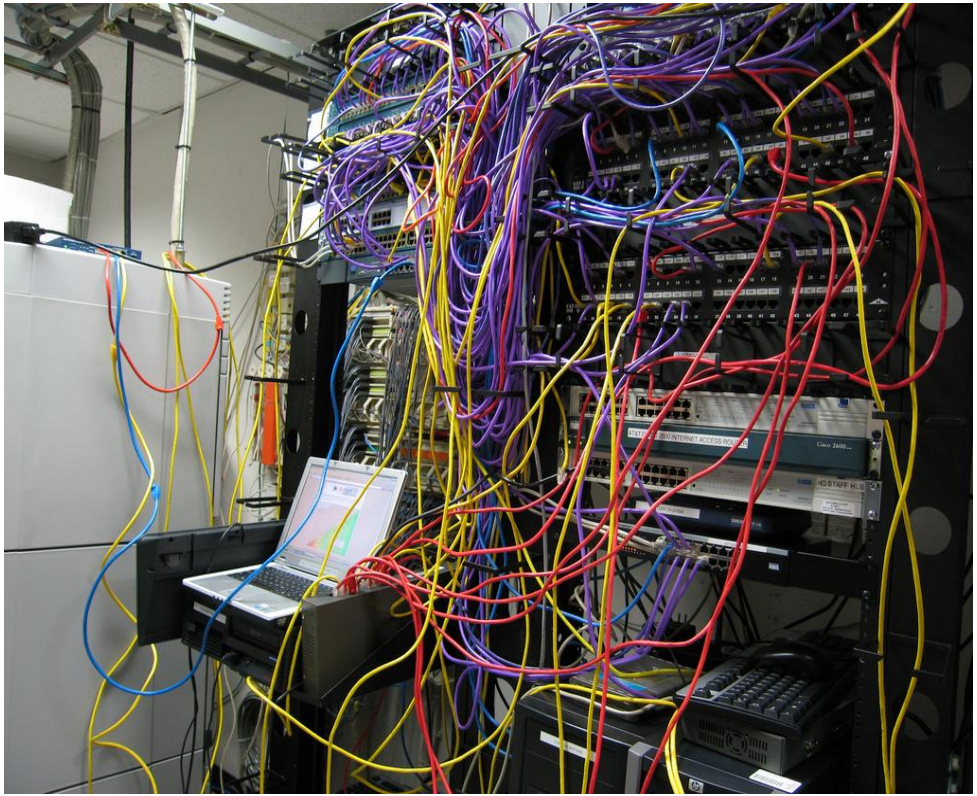
Многопроцессорные Вычислительные Системы (МВС)

BIGDATA TEAM

Q&A: Что здесь может сломаться?



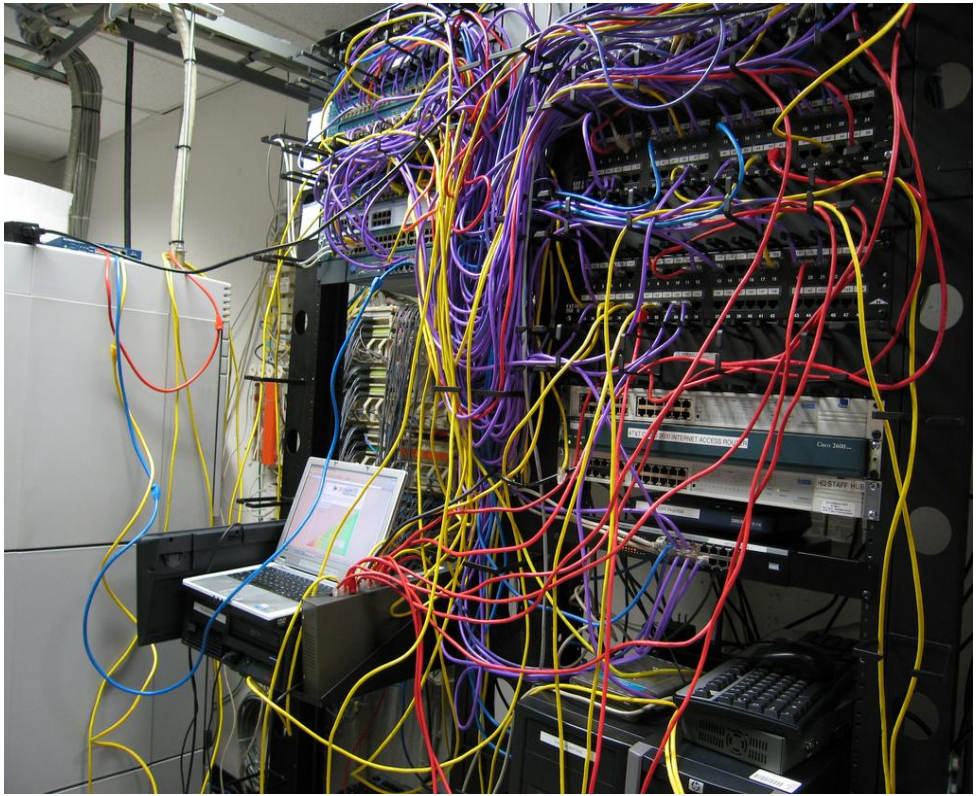
Компоненты МВС



- Сервера (узлы)
- Сеть (соединения)
- Модель синхронности

BIGDATA TEAM

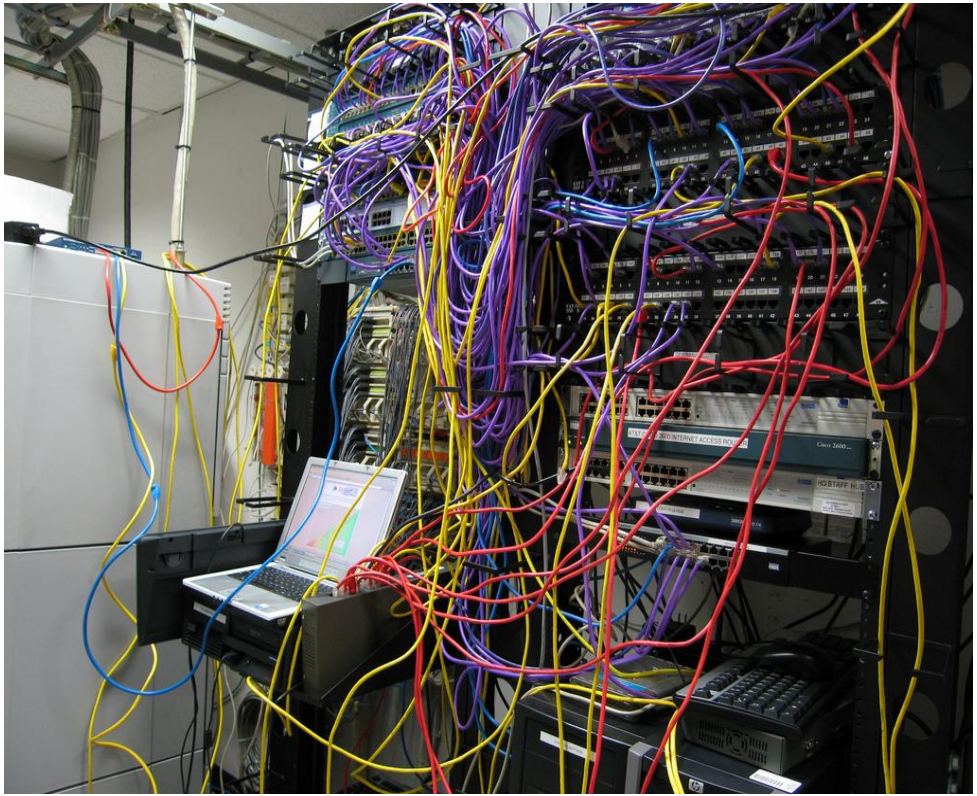
Гарантии относительно поломки узлов (Node Failures)



- Fail-Stop
- Fail-Recovery
- Byzantine

BIGDATA TEAM

Гарантии относительно поломок сети (Link Failures)



- Perfect Link
- Fair-Loss Link
- Byzantine

Модель синхронности

- Clock Skew
- Clock Drift

МВС

- Параллельные вычисления
- Распределенные вычисления
- Грид вычисления

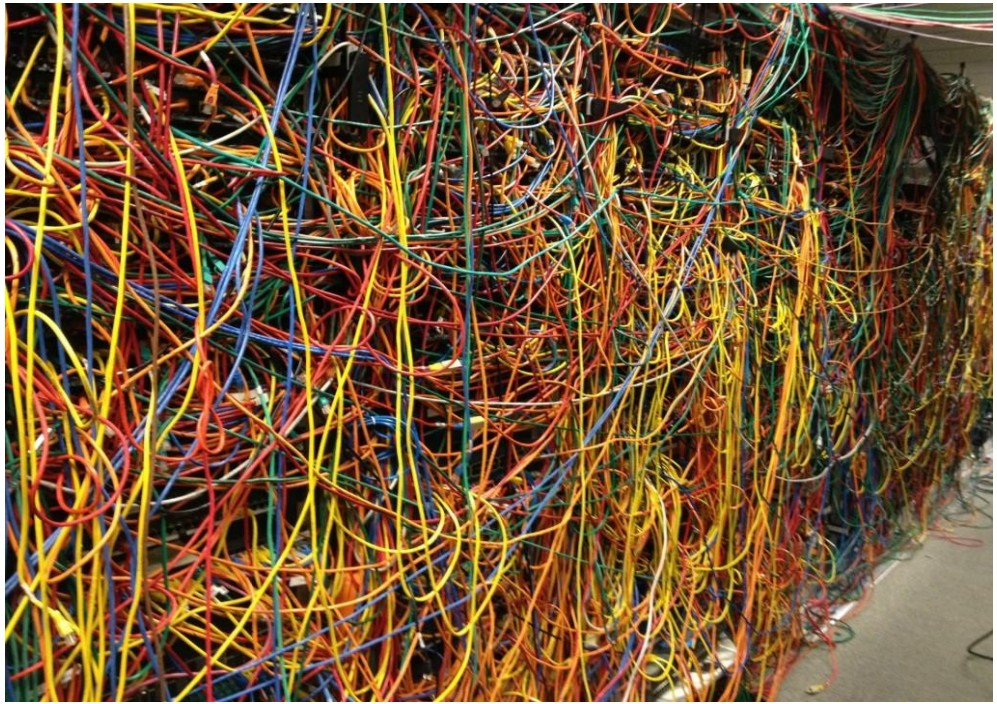
Параллельные вычисления

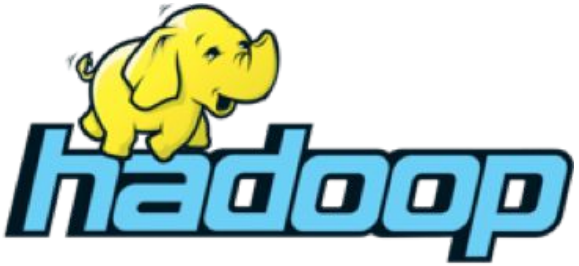
OpenMP

MPI

Fail-Stop + Perfect Link + Synchronous

Распределенные вычисления



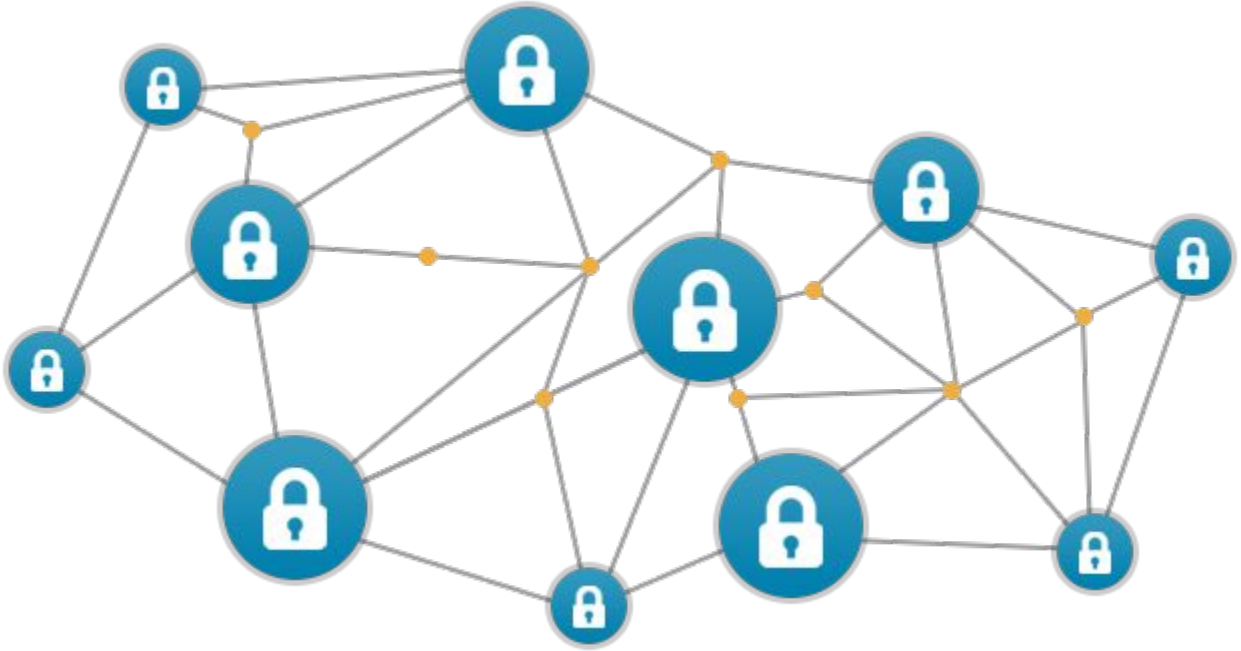


Fail-Recovery + Fair-Loss Link + Asynchronous

Грид вычисления

Byzantine-Failure + Byzantine Link + Asynchronous

Blockchain



Грид vs Распределенные системы

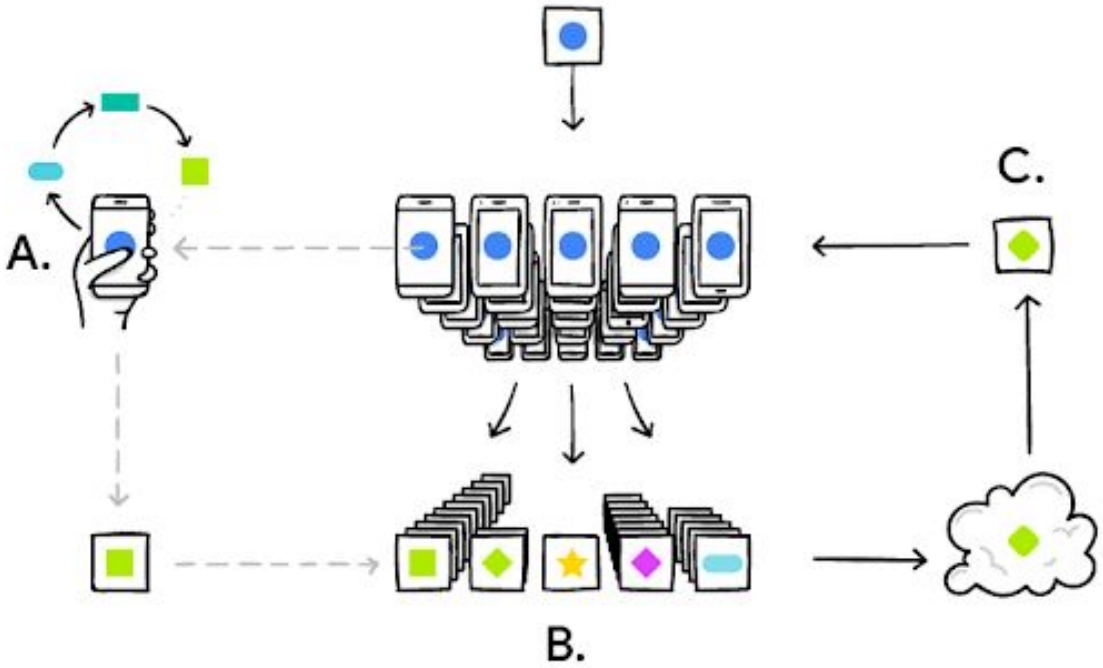
Visa:

- avg: 1,700 TPS
- Scale up to: 24,000 TPS

Bitcoin:

- avg: 5-6 TPS
- Scale up to: 150-160 TPS

Federated Machine Learning

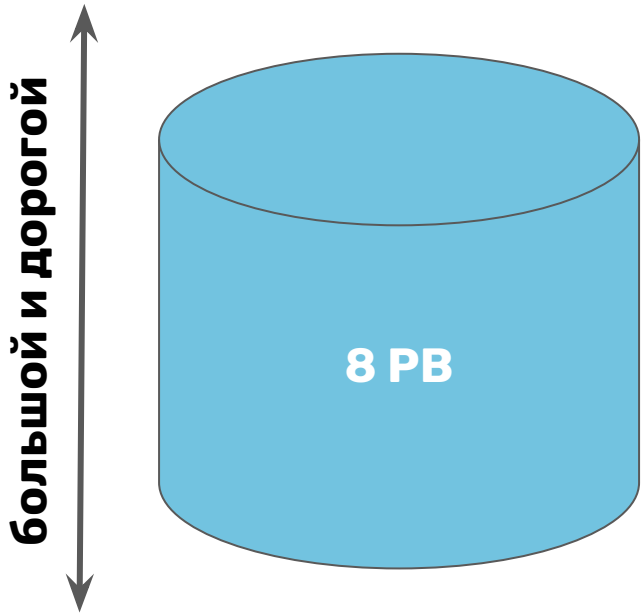


https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Распределенные файловые системы (GFS / HDFS)

BIGDATA TEAM

Распределенные файловые системы



вертикальное масштабирование (scale-up)

горизонтальное масштабирование (scale-out)

Distributed File Systems

The Google File System

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung

Google*

ABSTRACT

We have designed and implemented the Google File System, a scalable distributed file system for large distributed data-intensive applications. It provides fault tolerance while running on inexpensive commodity hardware, and it delivers high aggregate performance to a large number of clients.

While sharing many of the same goals as previous distributed file systems, our design has been driven by observations of our application workloads and technological environment, both current and anticipated, that reflect a marked departure from some earlier file system assumptions. This has led us to reexamine traditional choices and explore rad-

1. INTRODUCTION

We have designed and implemented the Google File System (GFS) to meet the rapidly growing demands of Google's data processing needs. GFS shares many of the same goals as previous distributed file systems such as performance, scalability, reliability, and availability. However, its design has been driven by key observations of our application workloads and technological environment, both current and anticipated, that reflect a marked departure from some earlier file system design assumptions. We have reexamined traditional choices and explored radically different points in the design space.

The Google File System, Symposium on Operating Systems Principles (SOSP, 2003)

Q&A

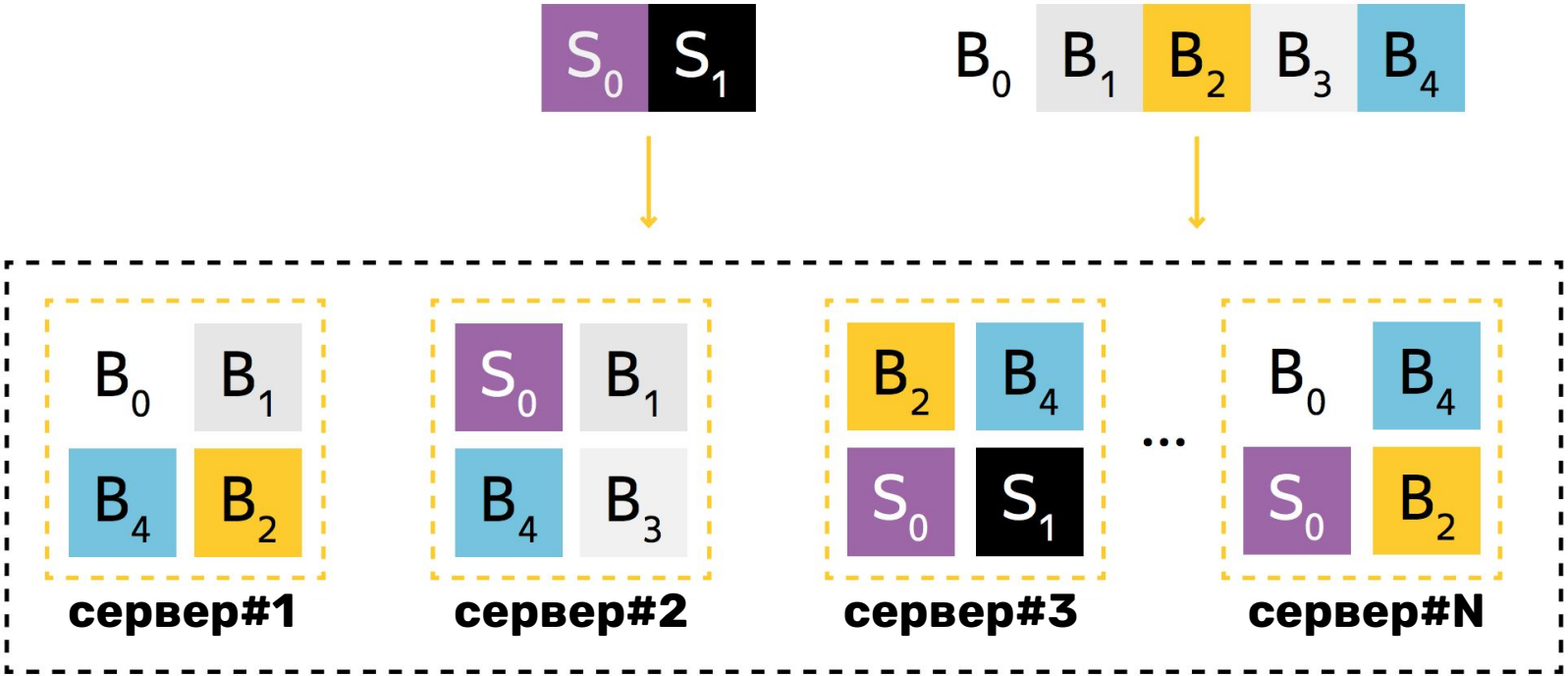
Как обеспечить гарантии хранения данных в рамках fail-recovery?

Q&A

Как храним файлы размером 2 TB и 10 GB в кластере?

BIGDATA TEAM

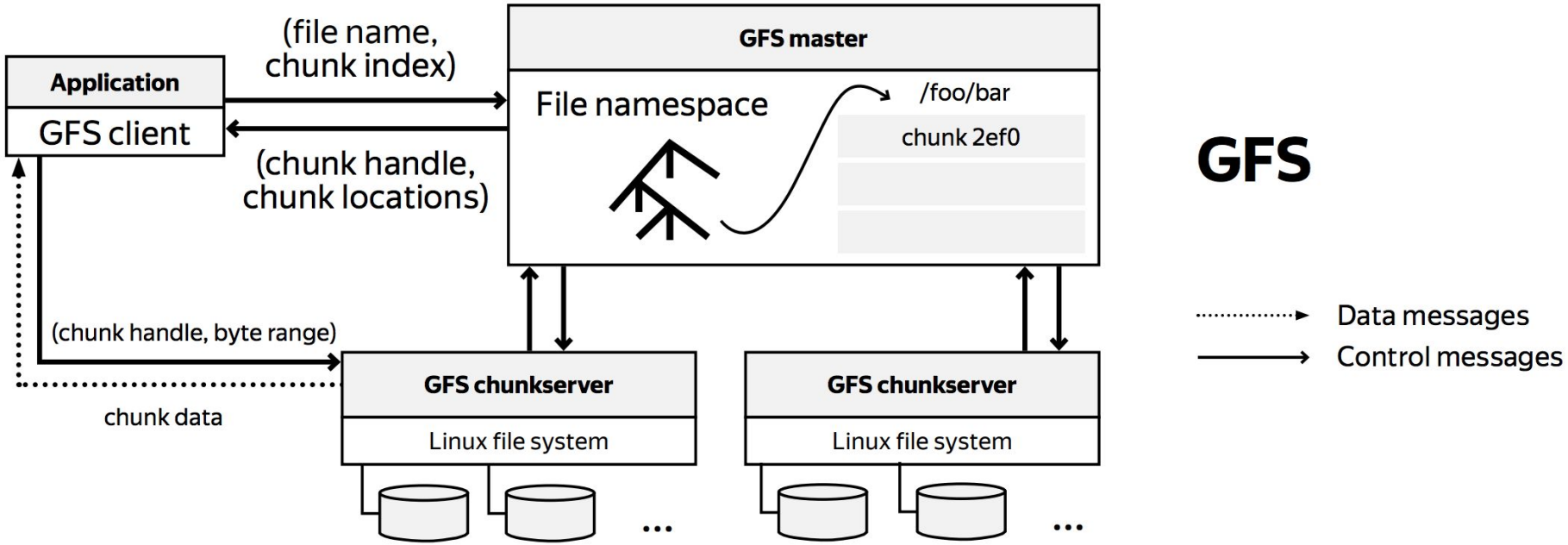
Распределенная ФС



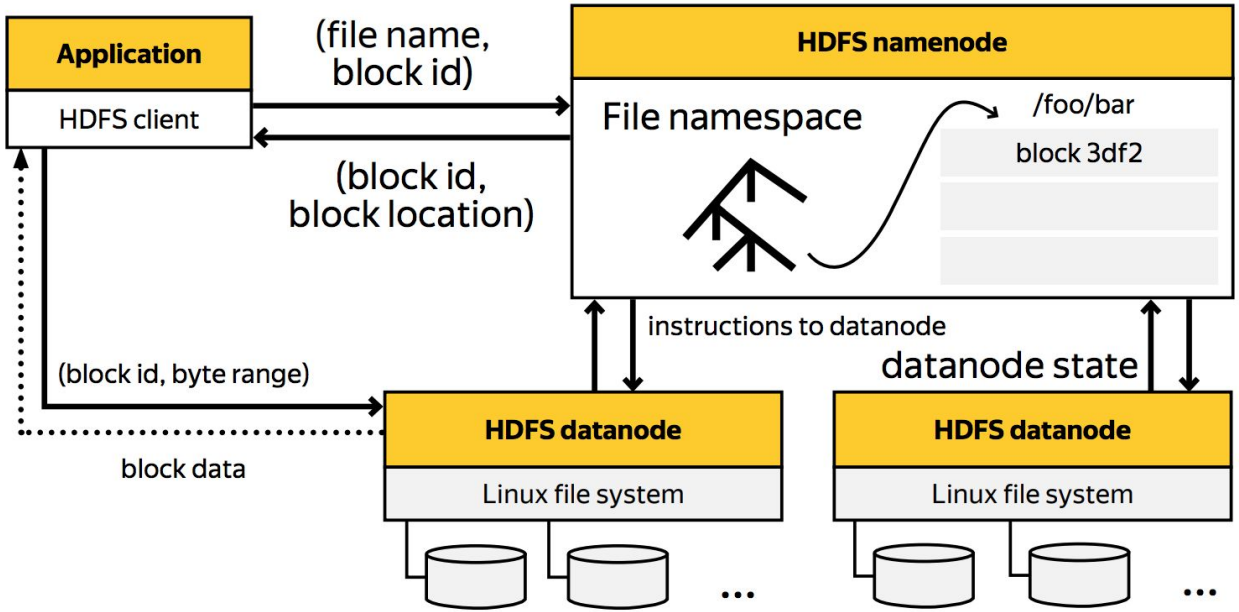
Принципы GFS

- Поломки компонент - это норма (используем репликацию)
- Равномерная утилизация компонент кластера
- Семантика write-once-read-many

GFS Data Flow

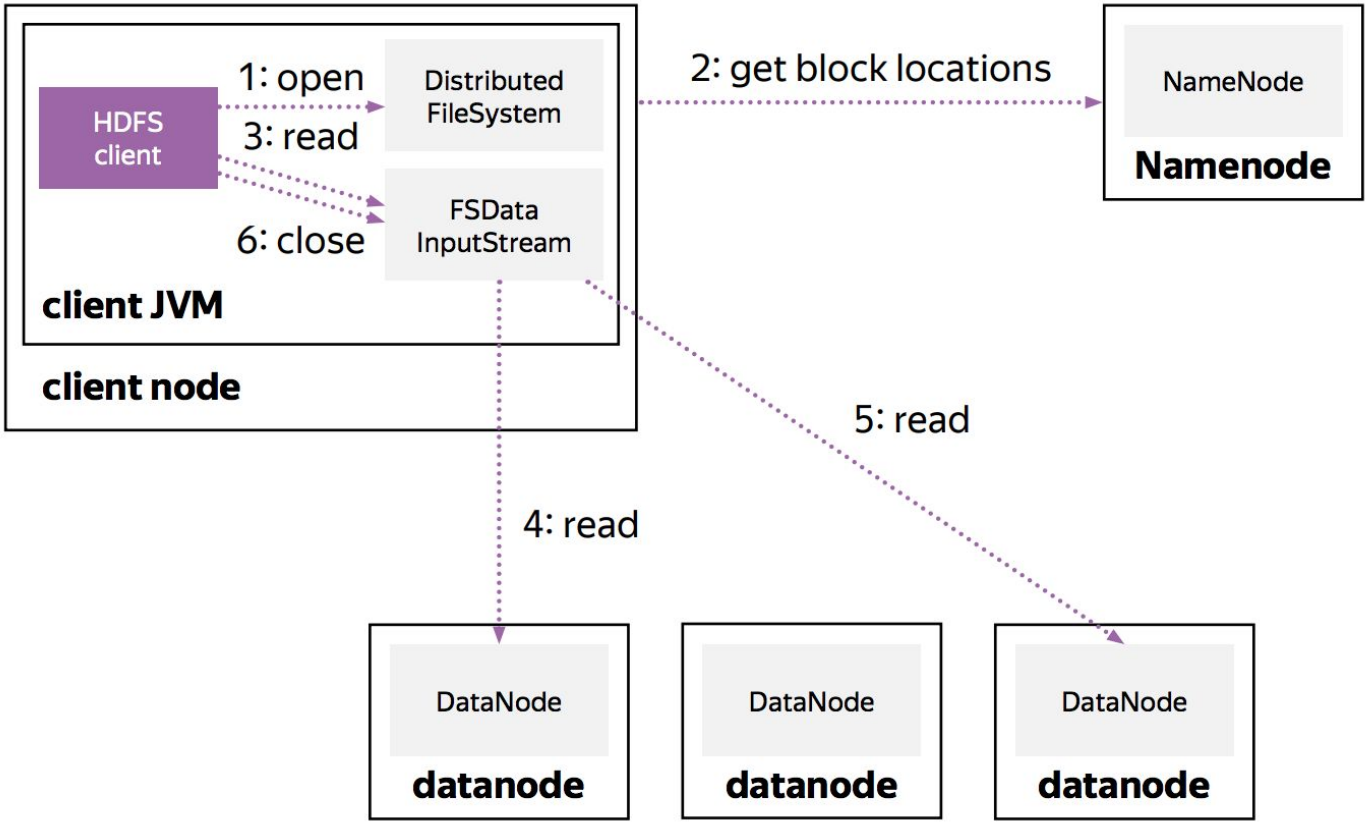


HDFS Data Flow

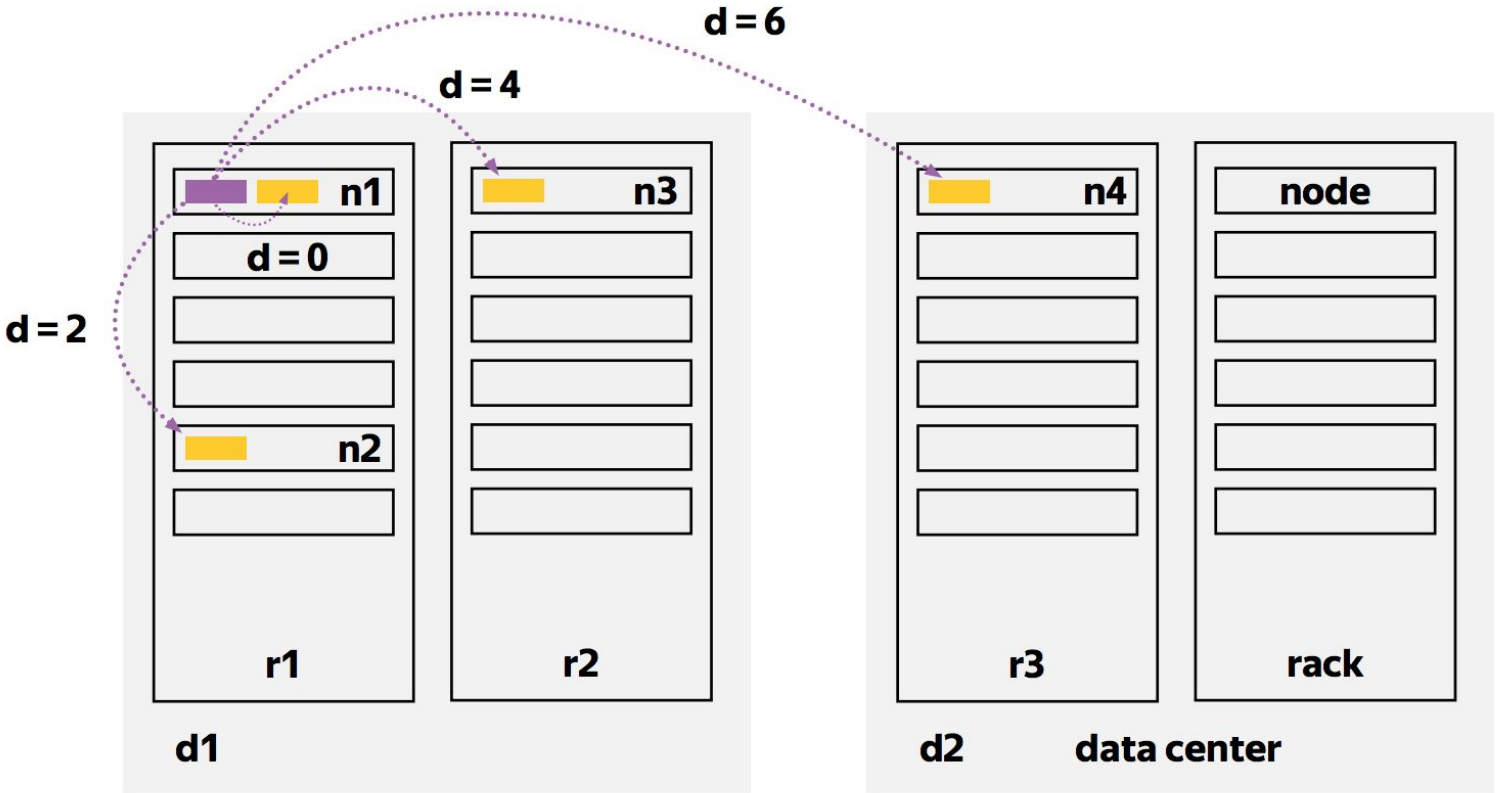


HDFS

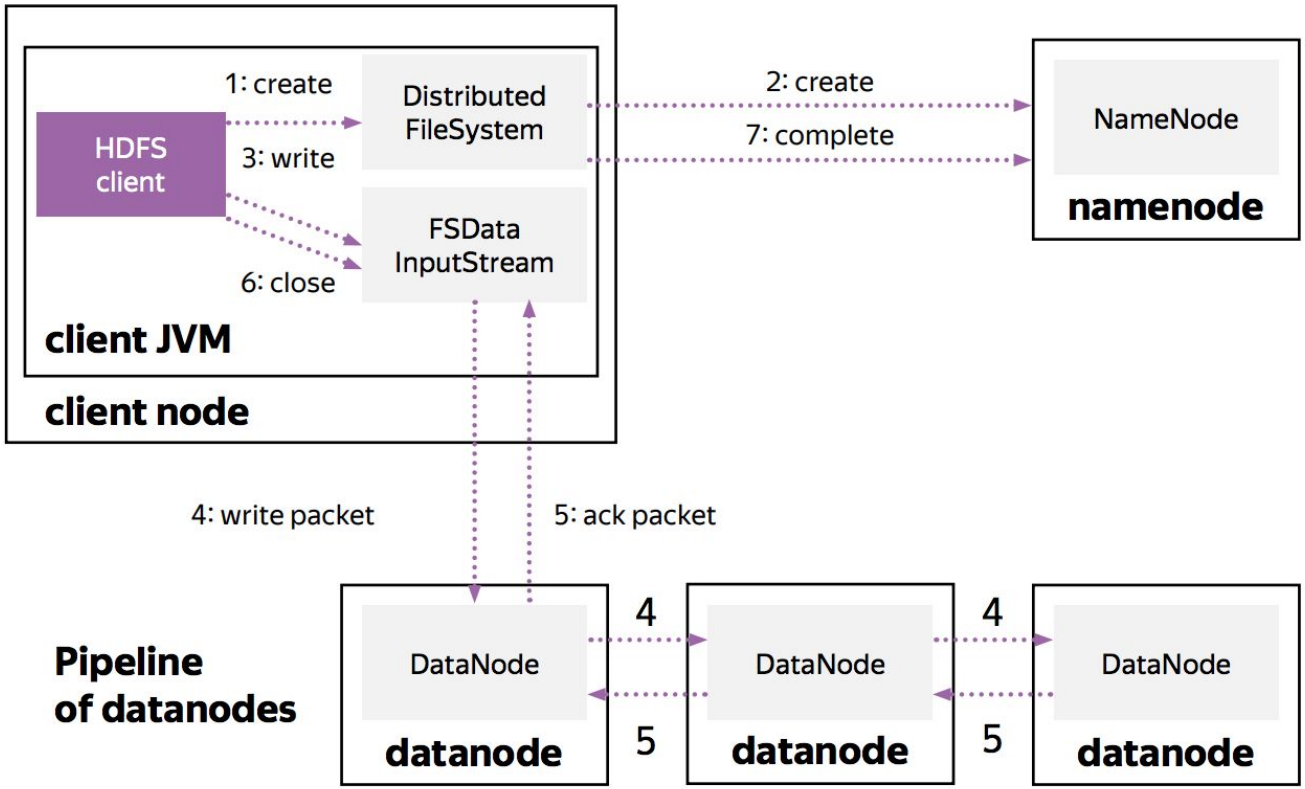
HDFS read


HDFS client
1: open
3: read
6: close
Distributed FileSystem
FSData InputStream
client JVM
client node
2: get block locations
NameNode
Namenode
4: read
5: read
DataNode
datanode
DataNode
datanode
DataNode
datanode

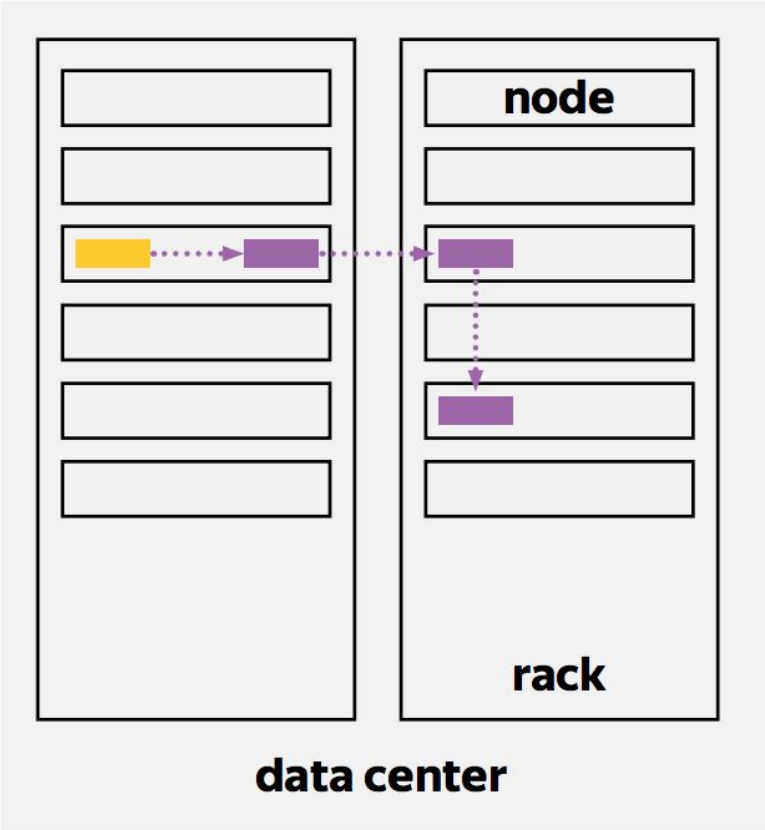
HDFS “distance”



HDFS write

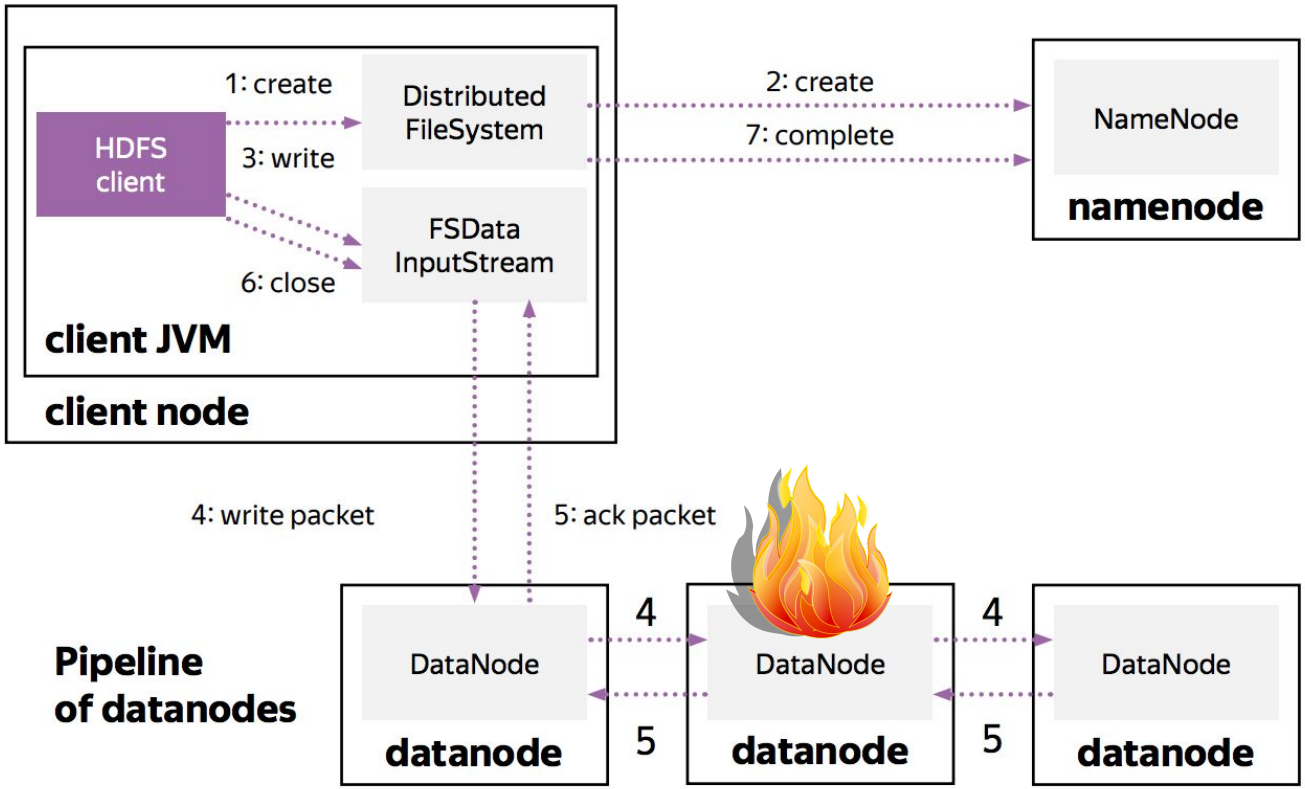

1: create
3: write
6: close
HDFS client
Distributed FileSystem
FSData InputStream
client JVM
client node
2: create
7: complete
NameNode
namenode
4: write packet
5: ack packet
Pipeline of datanodes
DataNode
datanode
4
5
DataNode
datanode
4
5
DataNode
datanode

HDFS Replica Placement



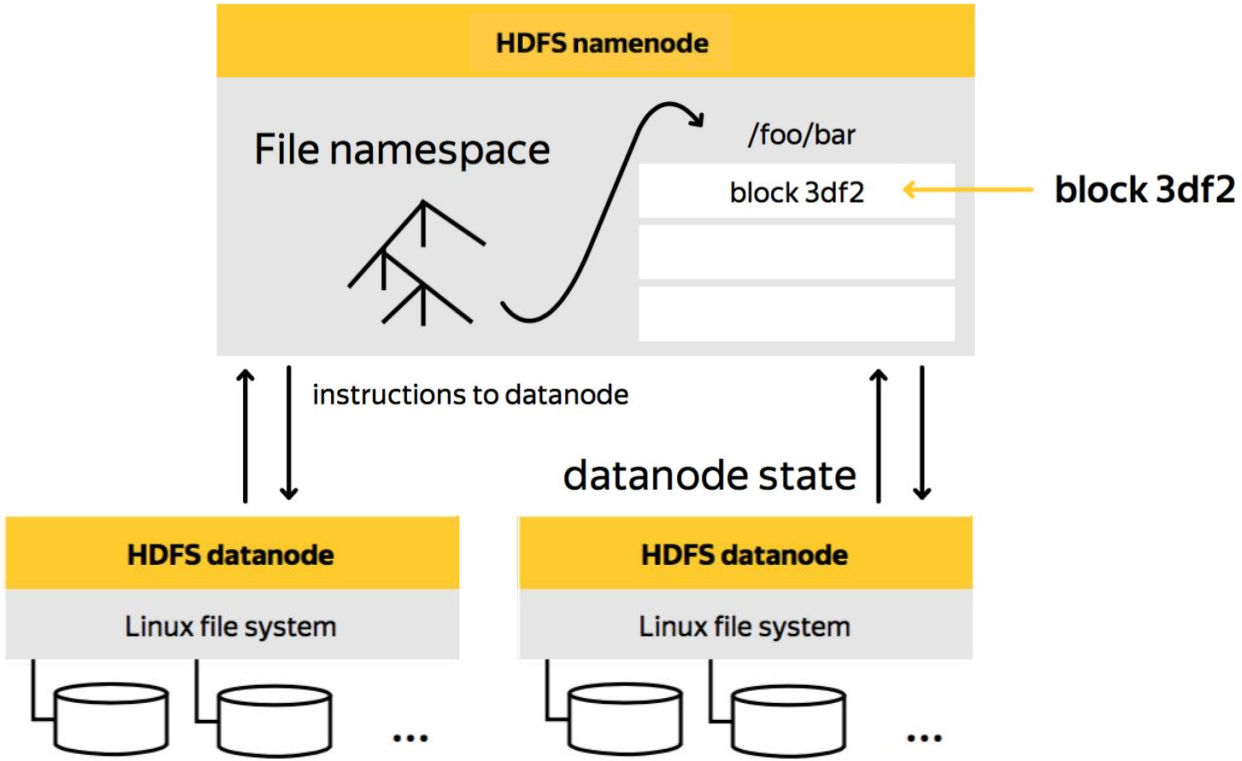
BIGDATA TEAM

Запись данных в HDFS v.2

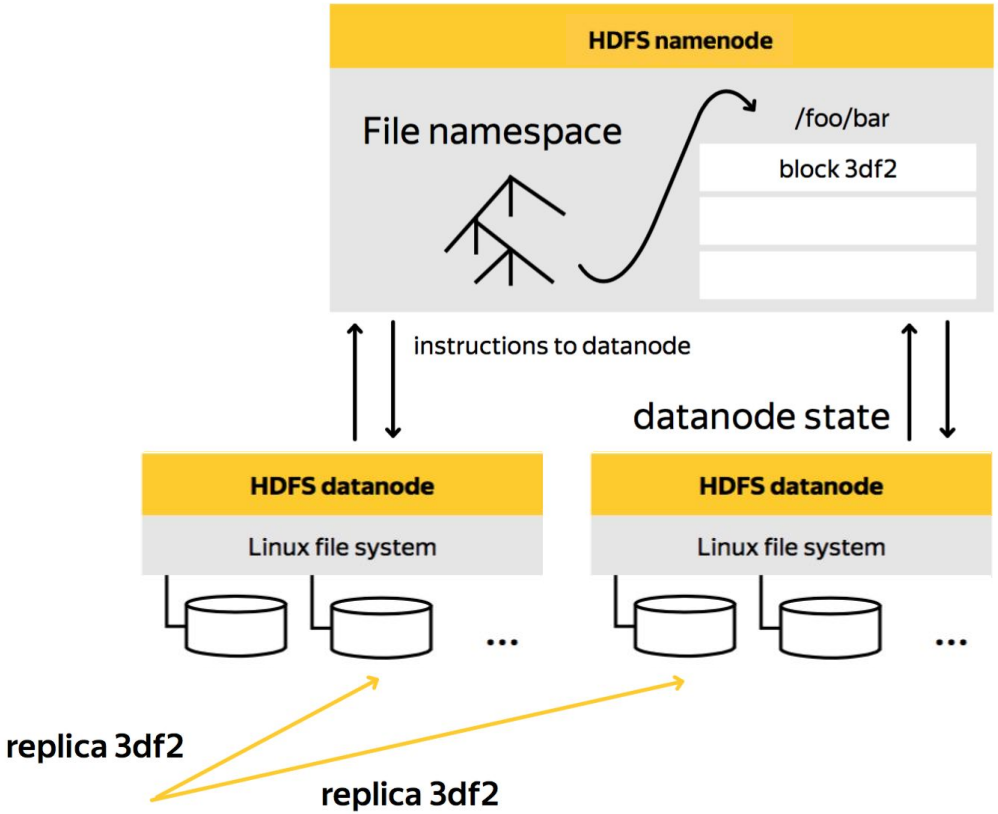


чанки, блоки и реплики

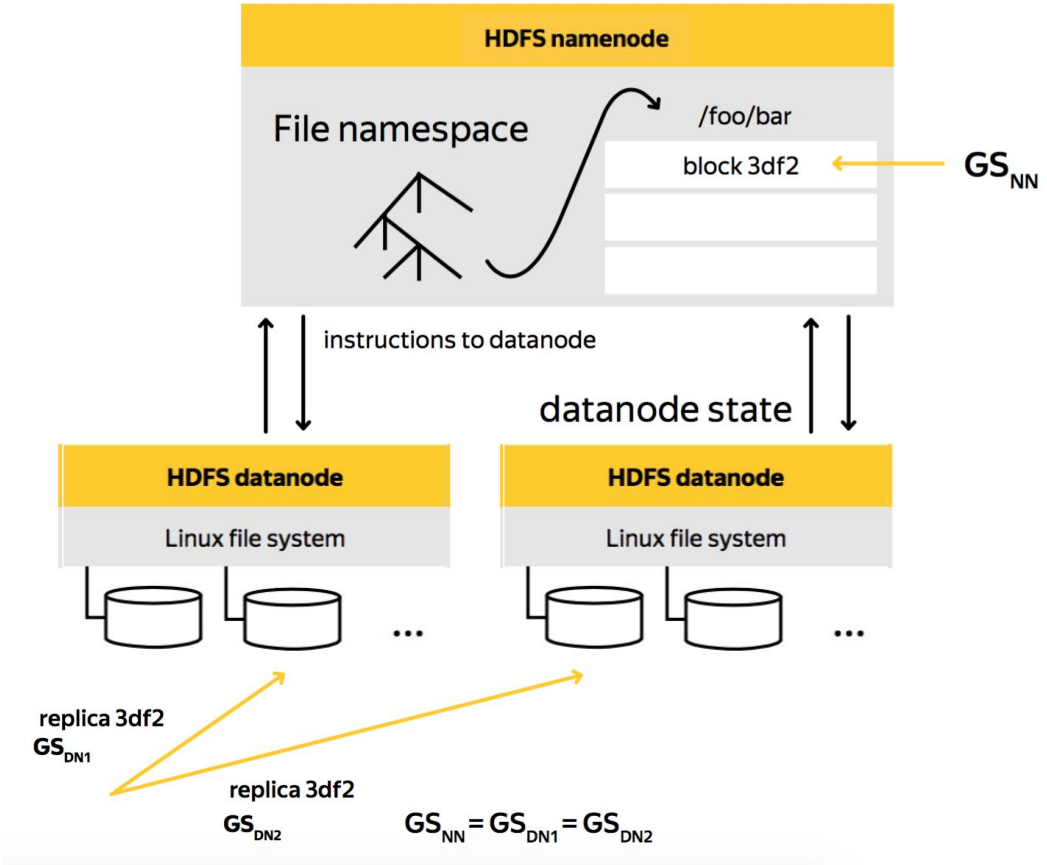
HDFS Block


HDFS namenode
File namespace
/foo/bar
block 3df2
block 3df2
instructions to datanode
datanode state
HDFS datanode
Linux file system
HDFS datanode
Linux file system

HDFS Replica



Finalized Replica



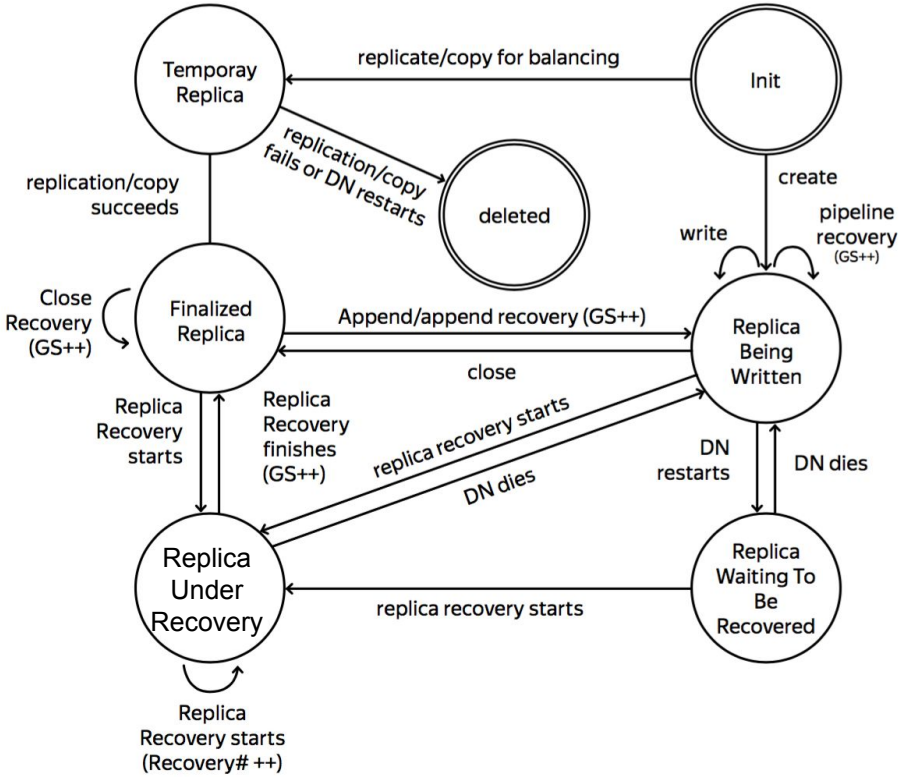
BIGDATA TEAM

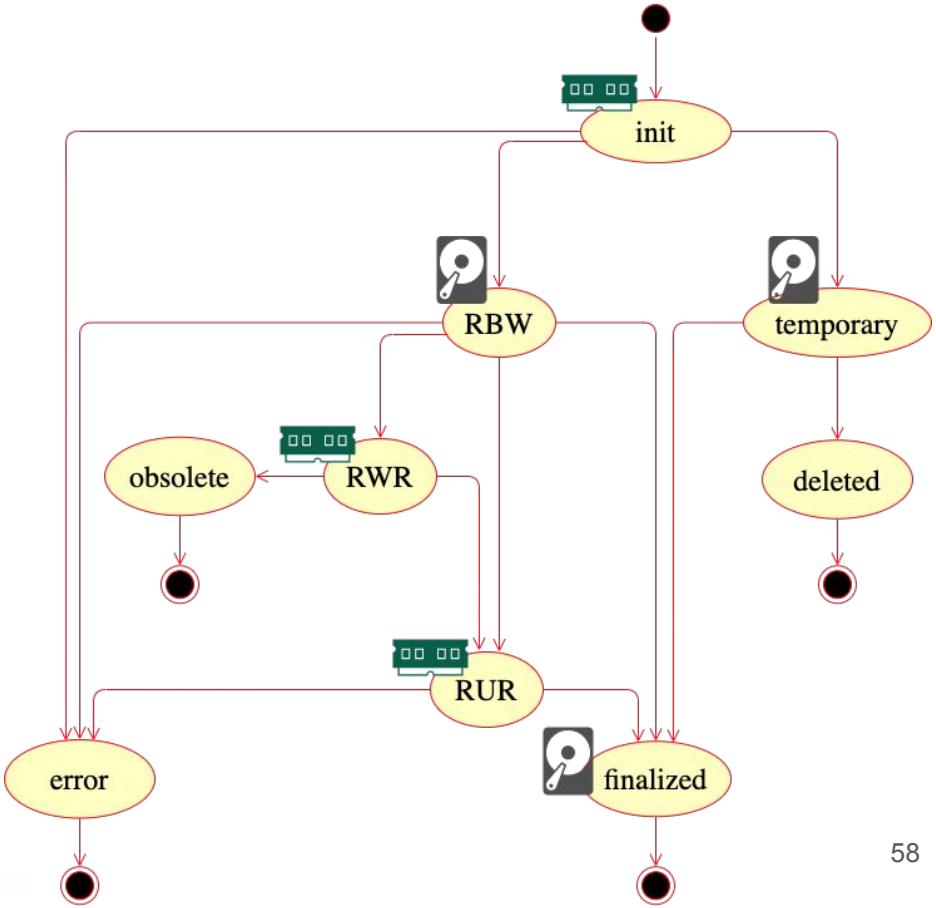
Процедуры восстановления HDFS

- Replica Recovery
- Block Recovery
- Lease Recovery
- Pipeline Recovery

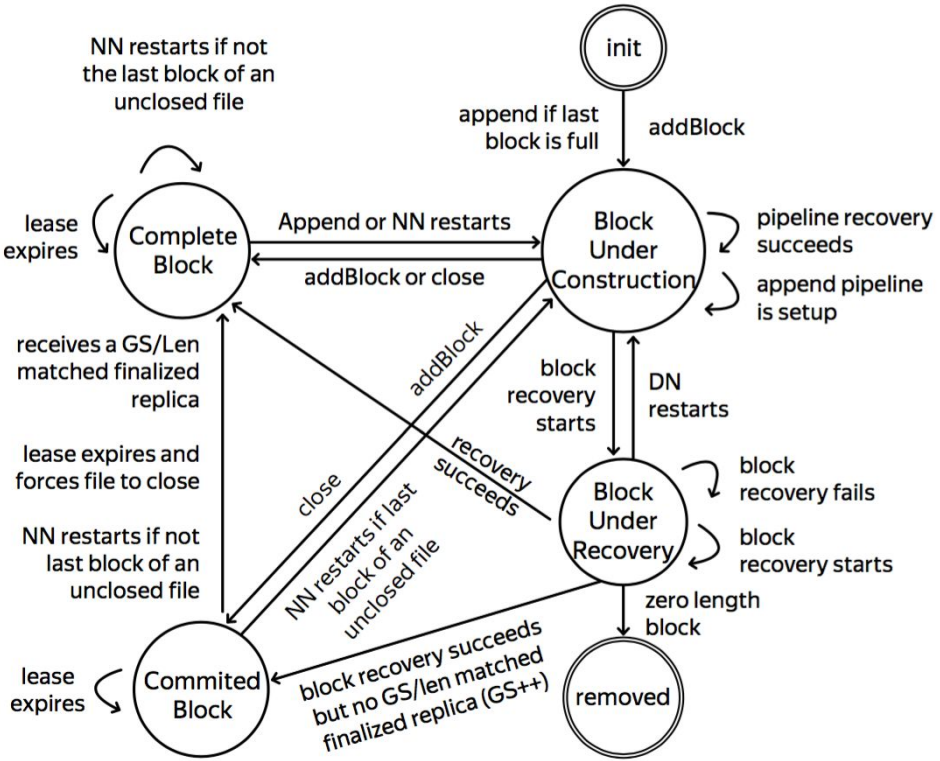
BIGDATA TEAM

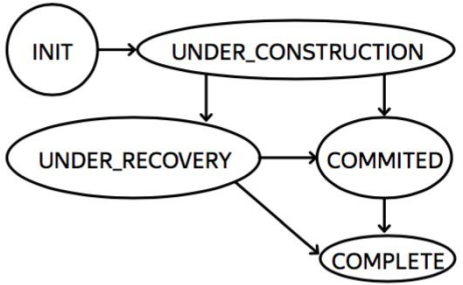
Replica State Transition v.2021





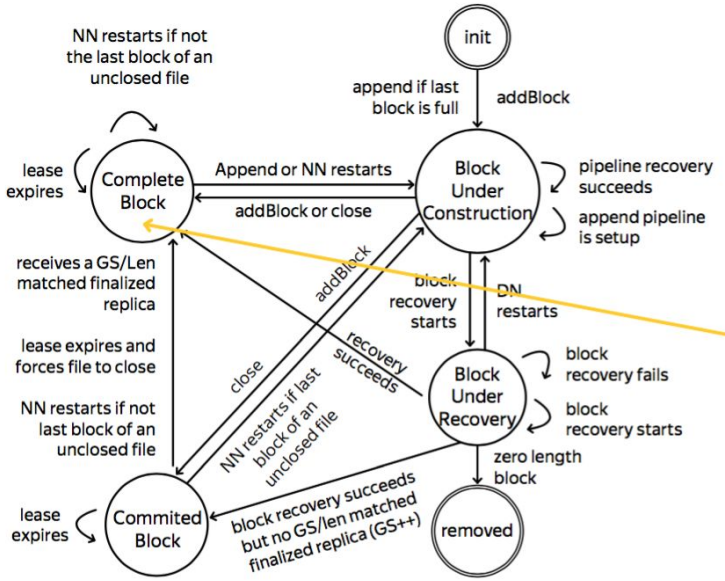
Simplified Block State Transition

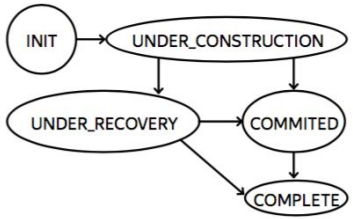




Simplified Block State Transition

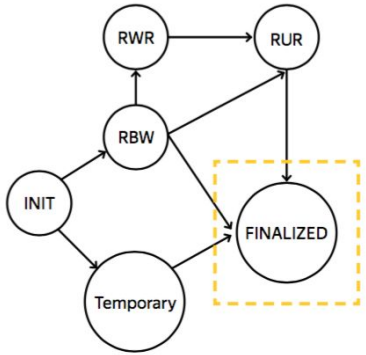
Block & Replica States Matching





Simplified Block State Transition

complete

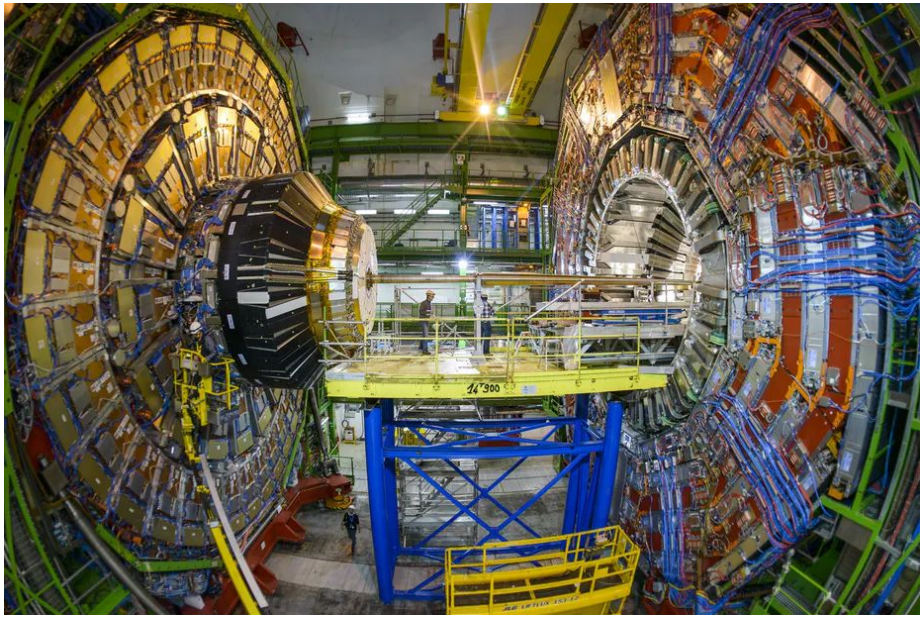


Tea / Coffee Break

Hadoop Sizing

BIGDATA TEAM

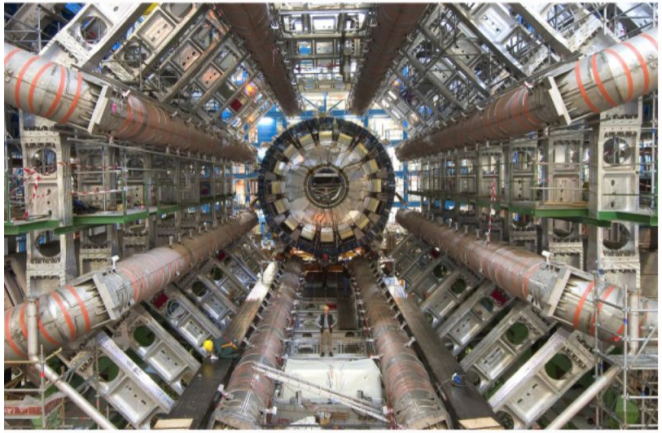
Запрос на консультацию

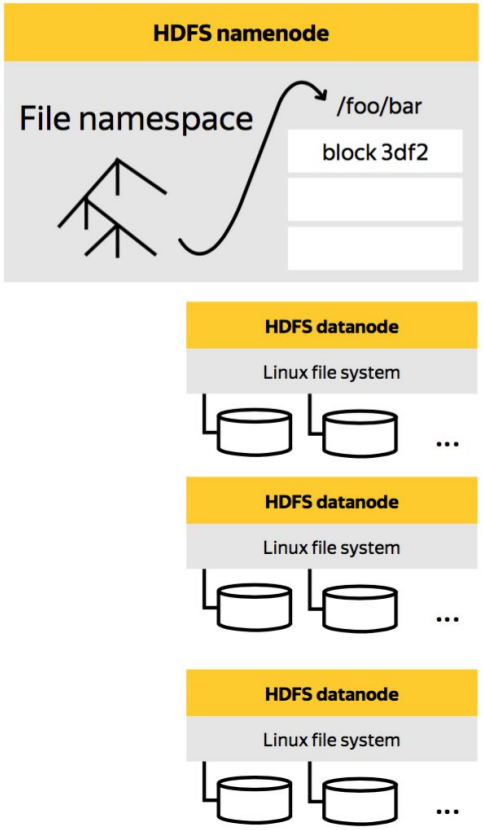


- The CERN Data Centre passed a major milestone on 29 June 2017 with more than **200 petabytes** of data now archived on tape

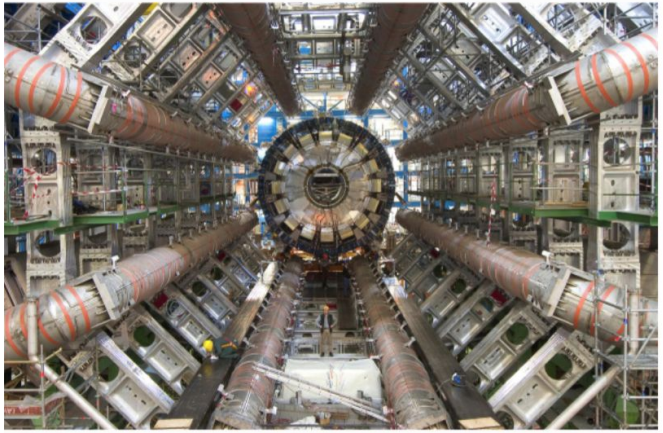
source: home.cern/news/...-200-petabyte-milestone

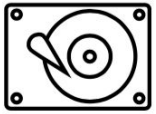
Hadoop Sizing



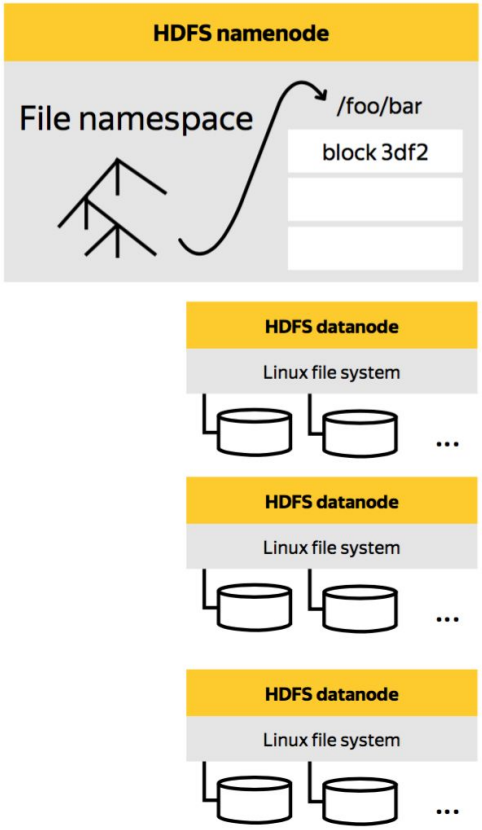


Hadoop Sizing





10 PB / 2 TB * 3 ~ 15 k



Считаем сбои

BIGDATA TEAM

Главное - не нарваться

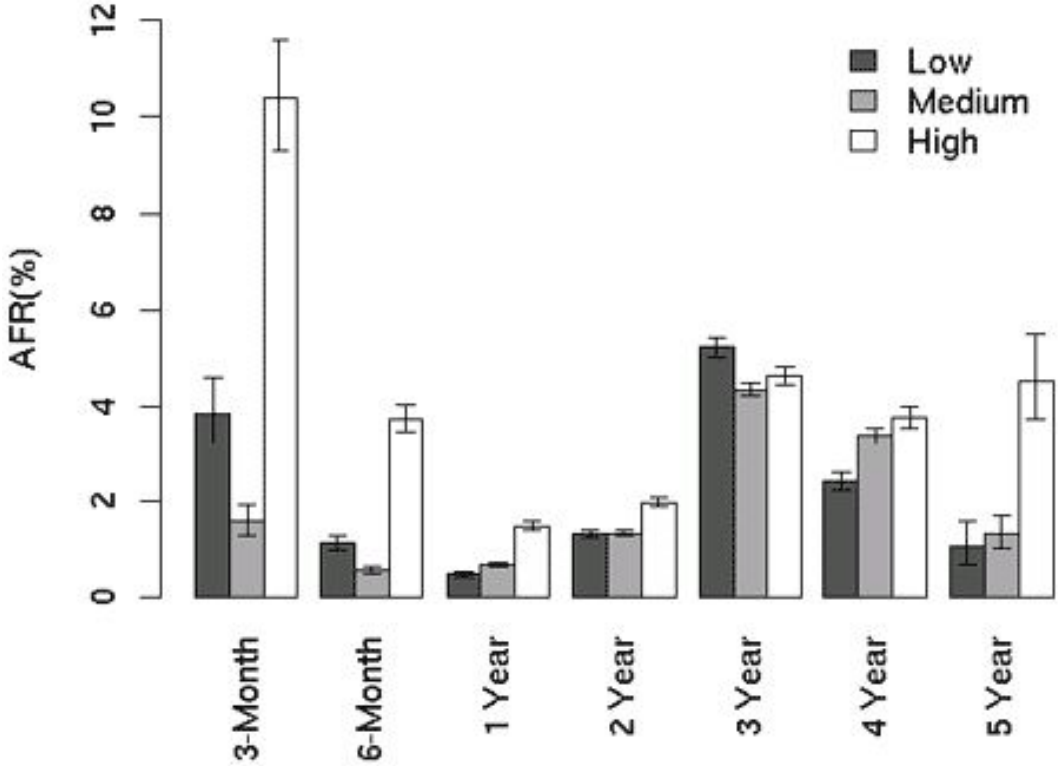
Backblaze Hard Drive Failure Rates for 2020

Reporting period 1/1/2020 - 12/31/2020 inclusive

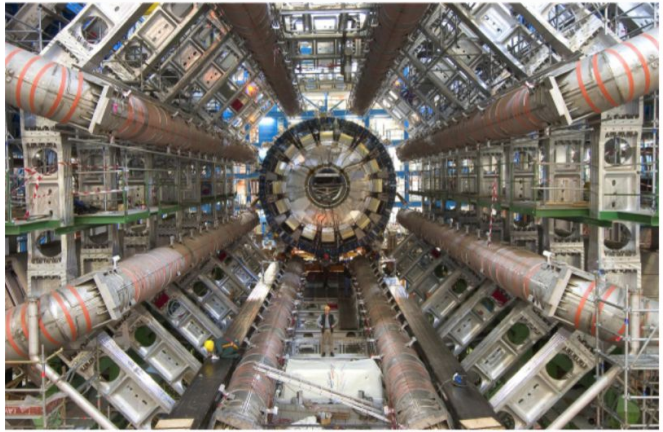
MFG	Model	Drive Size	Drive Count	Avg Age (months)	Drive Days	Drive Failures	AFR
HGST	HMS5C4040ALE640	4TB	3,100	56.65	1,083,774	8	0.27%
HGST	HMS5C4040BLE640	4TB	12,744	50.43	4,663,049	34	0.27%
HGST	HUH728080ALE600	8TB	1,075	34.85	372,000	3	0.29%
Seagate	ST14000NM001G	14TB	5,987	2.89	454,090	13	1.04%
Seagate	ST14000NM0138	14TB	360	1.56	5,784	0	0.00%
Seagate	ST16000NM001G	16TB	59	12.93	21,323	1	1.71%
Seagate	ST18000NM000J	18TB	60	3.27	5,820	2	12.54%

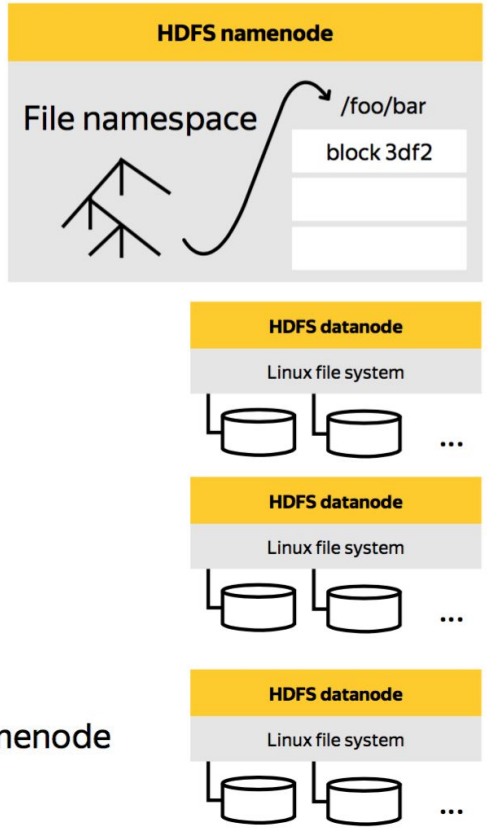
BIGDATA TEAM

Вероятность сбоя от перегрузки



Hadoop Sizing





10 PB / 2 TB * 3 ~ 15 k

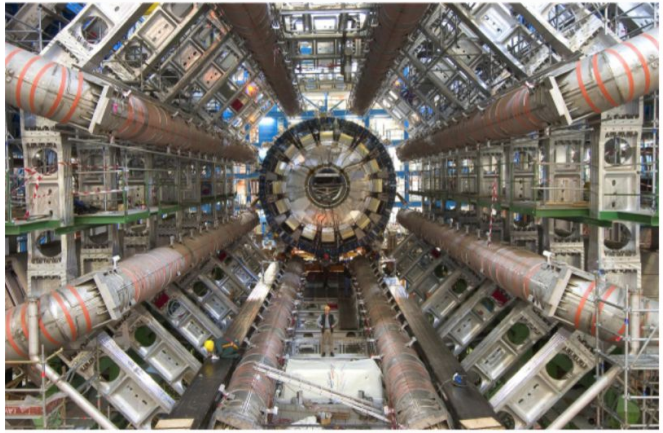
RAM

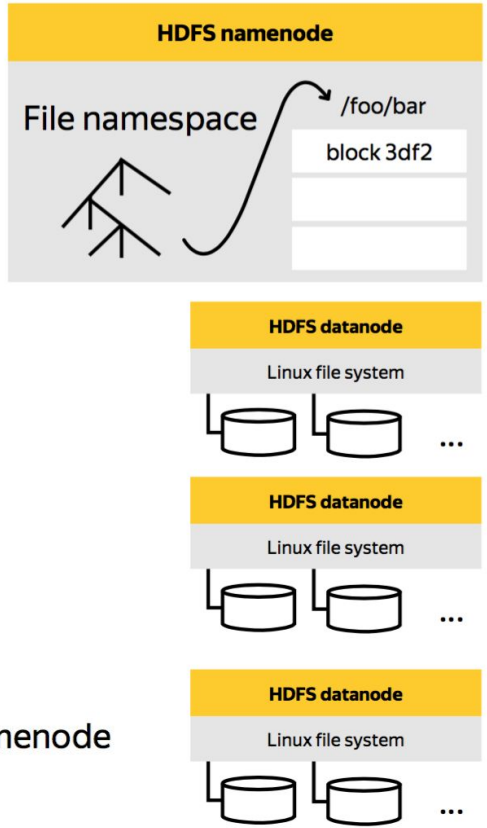
150 B - average block size on Namenode

https://issues.apache.org/jira/browse/HADOOP-1687

BIGDATA TEAM

Hadoop Sizing





Input interpretation:

$$\frac{10\ \text{PB (petabytes)}}{128\ \text{MB (megabytes)}} \times 150\ \text{bytes}$$

Result:

1.172×10^{10} bytes

Unit conversions:

11.72 GB (gigabytes)

10 PB / 2 TB * 3 ~ 15 k

RAM

150 B - average block size on Namenode

https://issues.apache.org/jira/browse/HADOOP-1687

Hadoop “Small Files Problem”

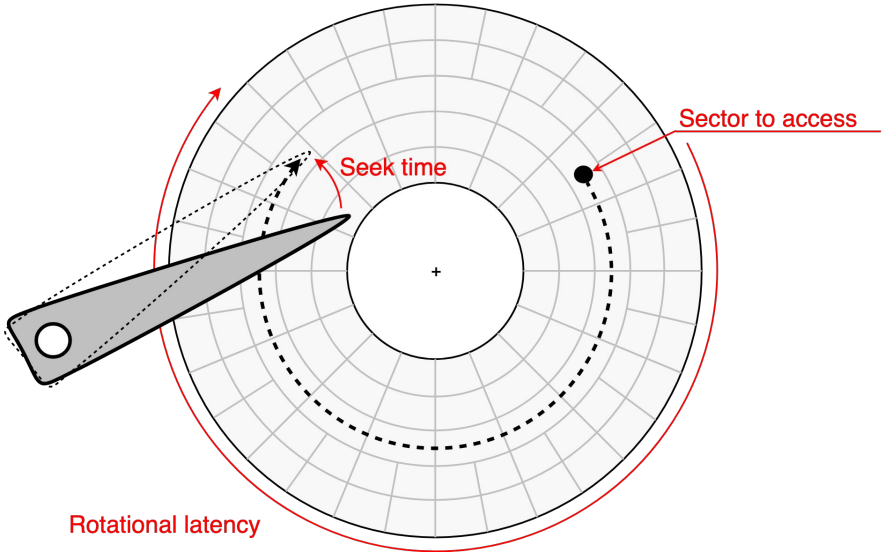
- reading speed - 600 MB/sec
- 10 PB → **207** дней

- reading speed - 600 MB/sec
- 10 PB → **103.5** дней

Выбор Block Size

RAM: 16GB, block size: **32MB**, replication factor: 3

- максимальный размер хранилища: 16GB / 150B x **32MB** / 3 = 1.138PB

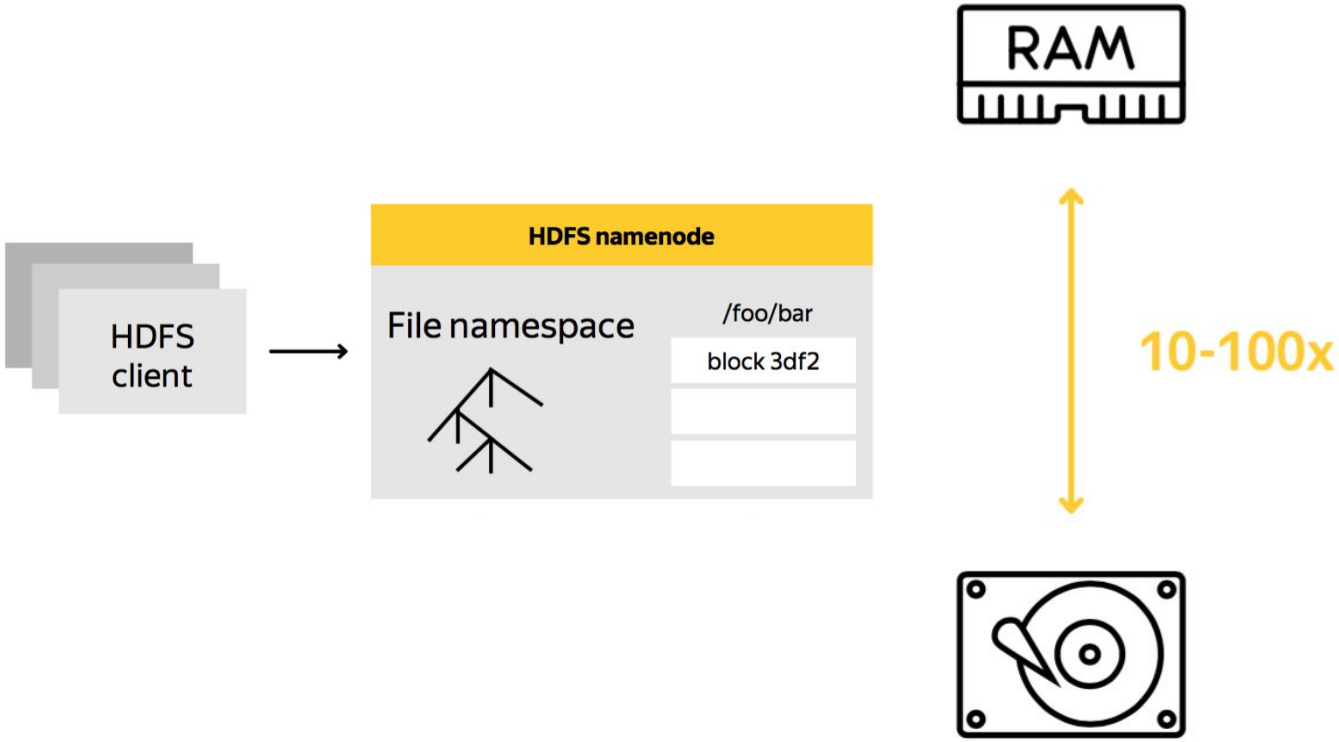


- HDD seek time: 0.2-0.8 мс
- SSD ~seek time: 0.08-0.16 мс

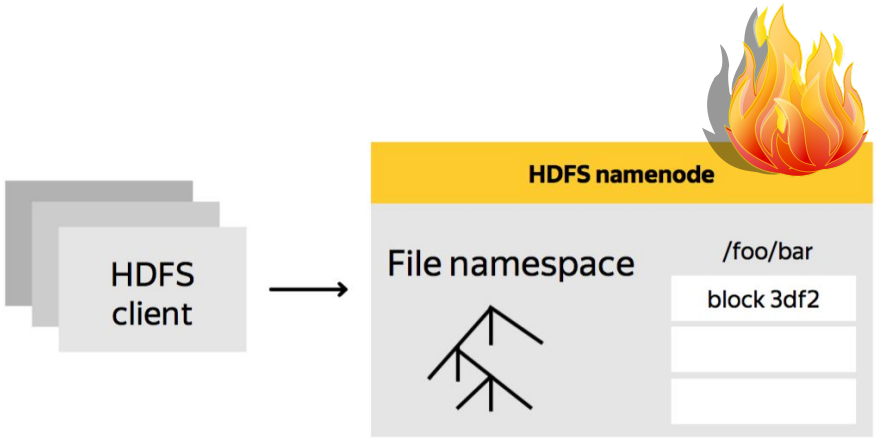
- reading speed - 600 MB/сек
- 32 MB ~ 50 мс

Архитектура Namenode

Быстрый доступ

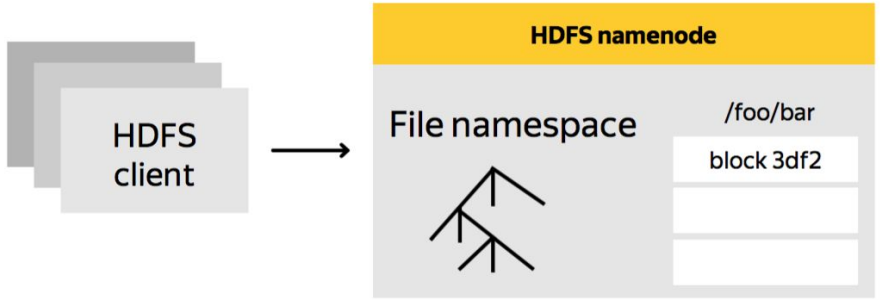


Namenode - это SPoF*



*SPoF = Single Point of Failure

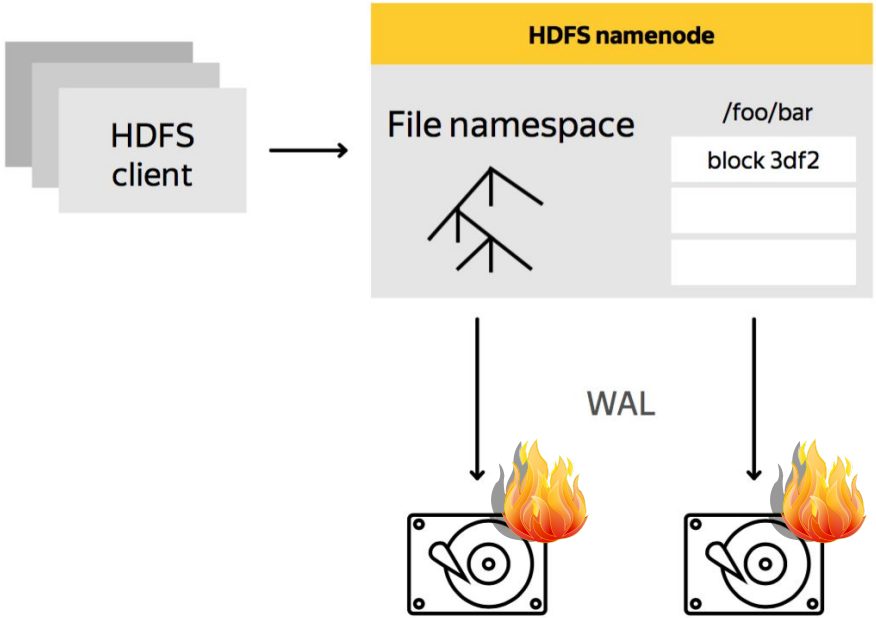
Борьба со сбоями Namenode



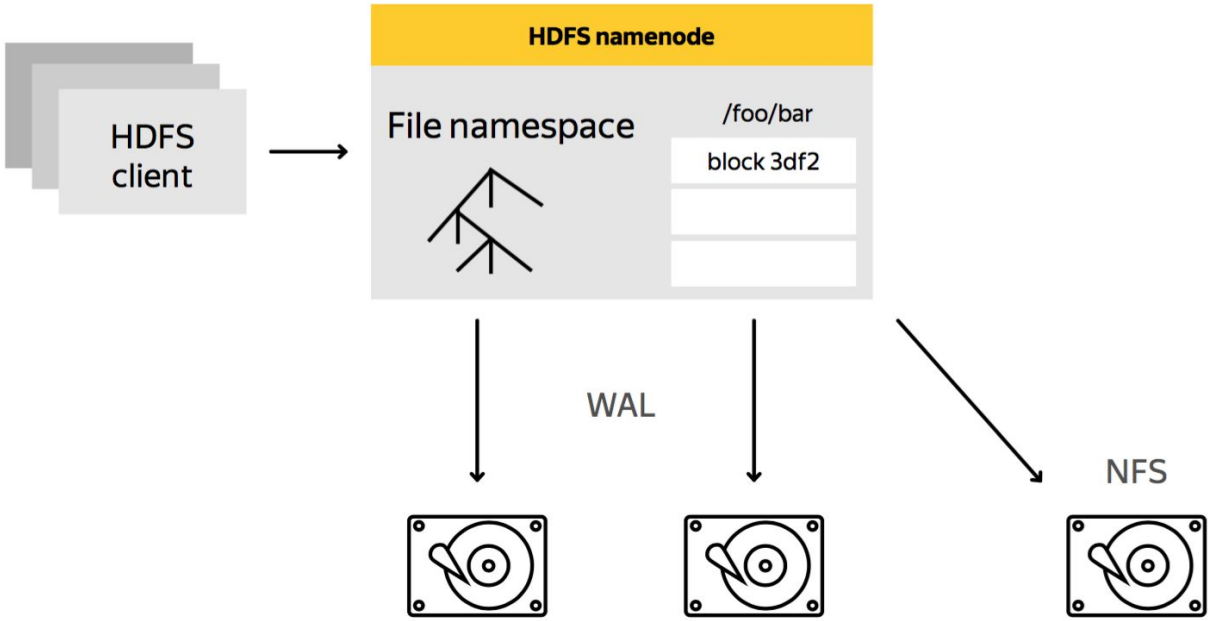

```
edits
...
mv a.txt b.txt
rm c.txt
...
```

*WAL = Write-Ahead Logging

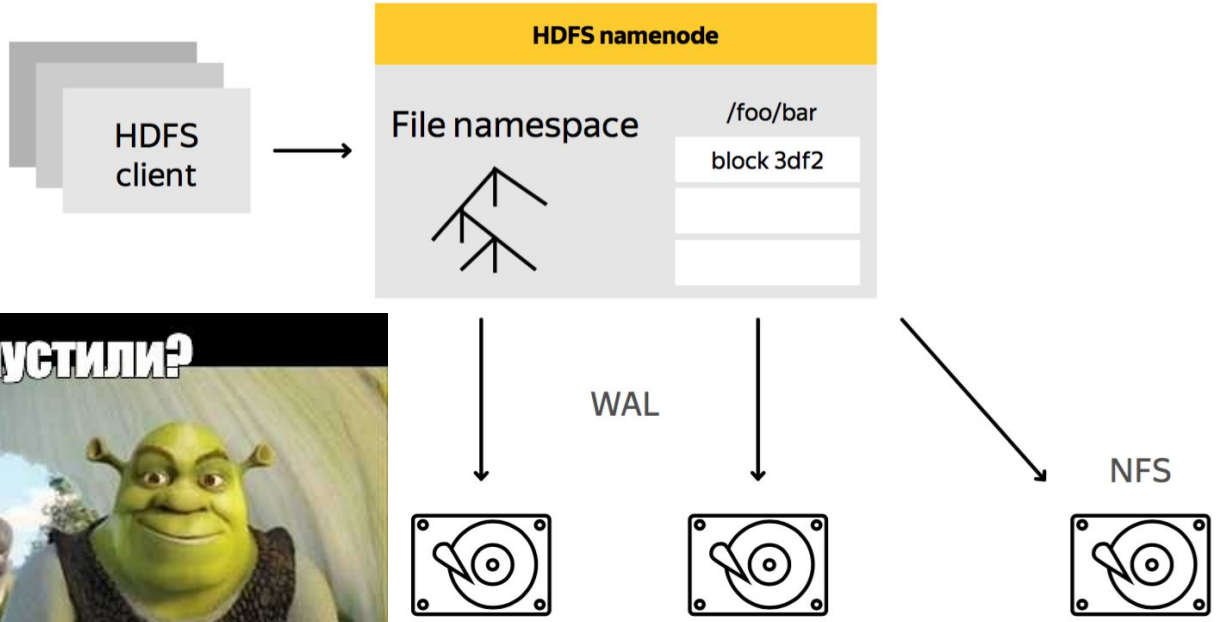
А что если...


HDFS namenode
HDFS client
File namespace
/foo/bar
block 3df2
WAL

А вот что

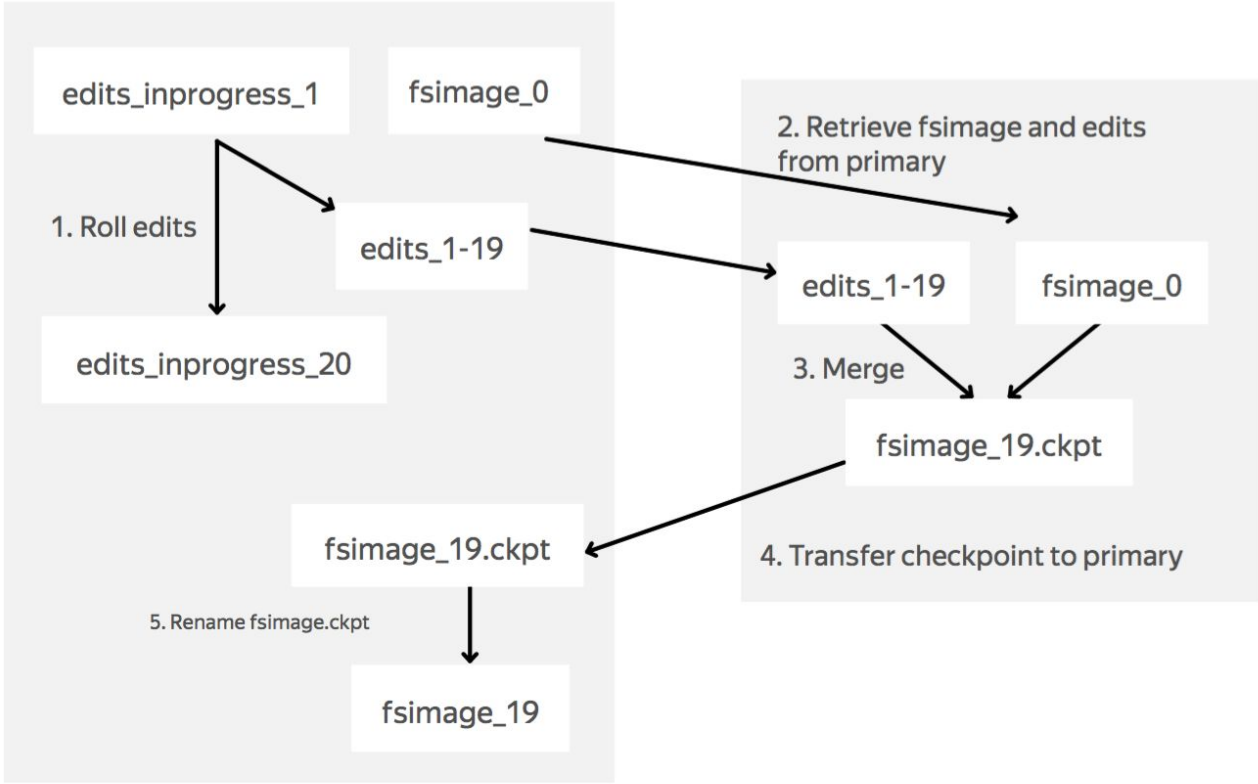

HDFS namenode
HDFS client
File namespace
/foo/bar
block 3df2
WAL
NFS

Но


HDFS namenode
HDFS client
File namespace
/foo/bar
block 3df2
WAL
NFS


NN запустили?
А сейчас?

Checkpoint Namenode



Подводные камни

Hadoop Sizing:

- **2x** ресурсов на Namenode**s**
- NFS

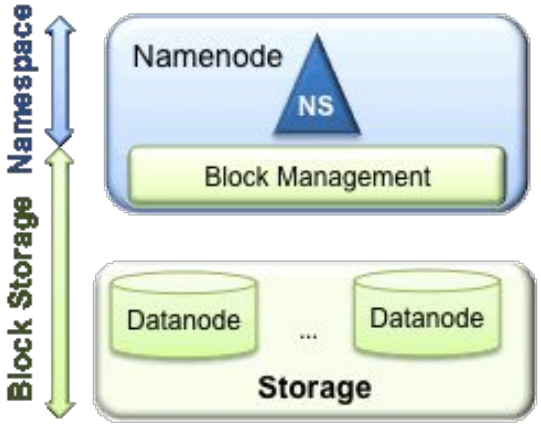
Избегать:

- Secondary Namenode
- Backup Namenode

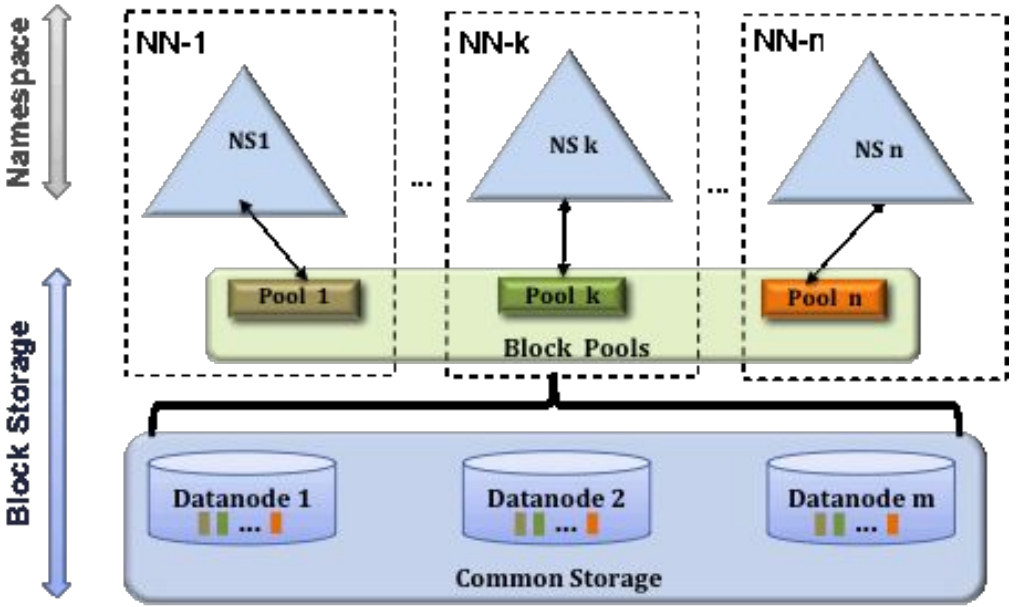
Использовать:

- Checkpoint Namenode

HDFS 2.0: HDFS Federation

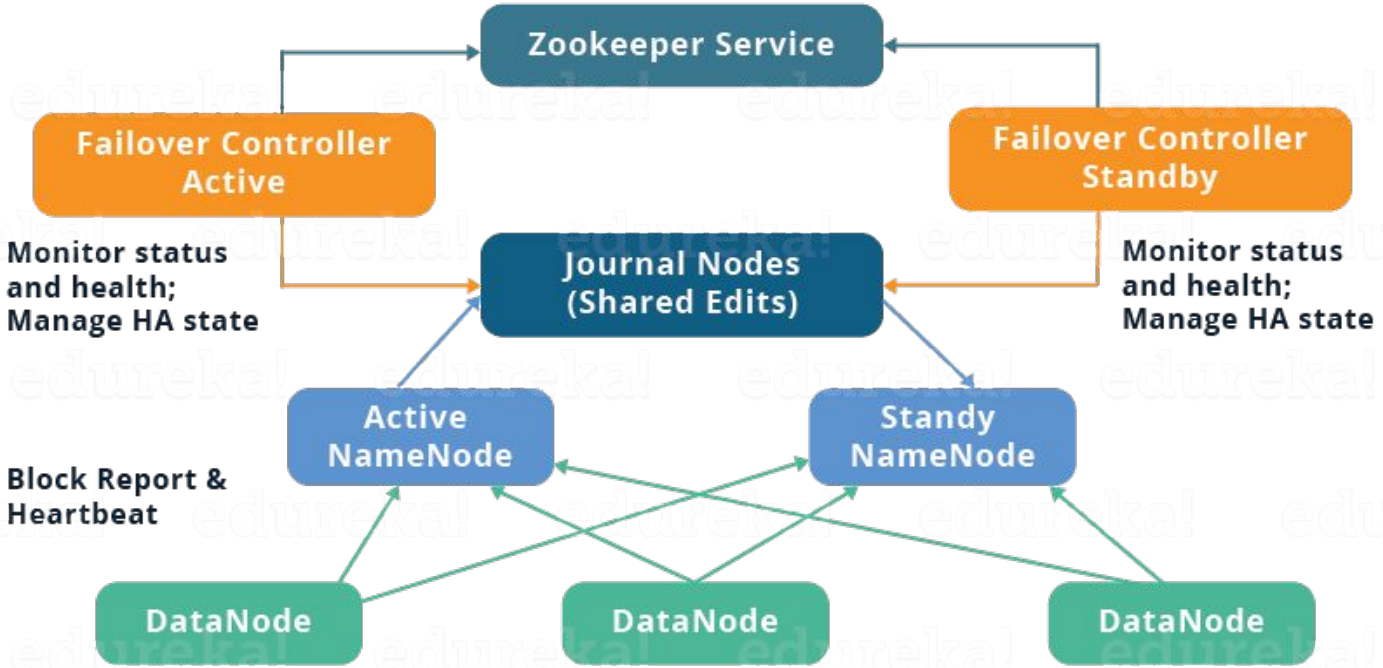


было



стало

Синхронизация через Journal Nodes



Checkpoint Namenode

Hadoop and NoSQL Downfall Parody
https://www.youtube.com/watch?v=hEqQMLSXQIY

Summary

- Вы можете перечислить 3 типа Многопроцессорных Вычислительных Систем, а также их типовые приложения
- Вы можете нарисовать диаграмму переходов состояний для блока и реплики
- Вы можете объяснить “design choice” архитектуры Namenode, а также указать на разницу между различными типами Namenode: Primary / Secondary / Checkpoint / Backup.
- Вы можете оценить ресурсы, необходимые для Hadoop-кластера (Hadoop sizing) для решения вашей задачи, а также можете объяснить что такое small files problem

Спасибо! Вопросы?

Feedback: http://rebrand.ly/x5bd2021q1_feedback_01_hdfs

Алексей Драль, aadral@bigdatateam.org
CEO at BigData Team, http://bigdatateam.org
https://www.linkedin.com/in/alexey-dral
https://www.facebook.com/bigdatateam