НИКИТА КАЗЕЕВ

KAZEEVN@YANDEX-TEAM.RU

СОДЕРЖАНИЕ

1 ХРАНЕНИЕ ДАННЫХ

- Без данных исследование не может быть (вос)произведено.
- Mutable data ruins reproducibility

2 LESS THAN RAM SIZE

- Не проблема
- файл на облачном диске, web-сервере, и т.д.

2.1 ФОРМАТ

- .csv.gz, .json.gz: кросс-платформенные, гибкие, инетрпретируемые, медленные
- HDF5, ROOT: бинарные, кросс-платформенные
- Формат Вашего инструмента (если Вы уверенны, что сможете прочесть его в будущих версиях)

2.2 GIT LARGE FILE STORAGE

- Версионирование больших файлов
- Интеграция с git без неудобного увеличения размера репозитория
- Open Source https://git-lfs.github.com/

3 MORE THAN RAM SIZE

- "Big Data", "NoSQL"
- Распределённые системы
- Горизонтальная масштабируемость
- Выбор зависит от задачи

3.1 ОТКУДА НАЧАТЬ?

- Mainframes -> supercomputers (not part of the talk)
- http://indico.cern.ch/getFile.py/access?contribId=521&sessionId=21&resId=0&materialId=slides&confId
- Google MapReduce (2004), Google File System (2003)

3.2 УРОКИ ИЗ ЦЕРНА

- Данные можно фильтровать. ЦЕРН выбрасывает 99.99%.
- Обработка должна быть возможно более парралельной. В ЦЕРНе алгоритмы применяются к болшому количеству независимых объектов одной природы. В результате возможно георафическая распределённость - задача выполняется сразу в нескольких странах.
- Локальность. Стоит обрабатывать данные там же, где они хранятся.

3.3 CERN GRID

- 30 Пб в год
- 170 дата-центров, 42 страны

3.4 HADOOP

- Масштабируем
- Стар и отлажен
- Устойчив к сбоям
- Open Source
- Заточен под парадигму MapReduce

3.5 CONTEMPORARY NOSQL

- Выбор зависит от задачи
- Общий прицип - чем проще операции, тем быстрее и надёжнее система
- Redis, HBase, Cassandra, Elastic, ...

4 ПОДЕЛИТЬСЯ ДАННЫМИ

- Будут ли они доступны через 30 лет?
- Как заинтересованные их найдут?
- Как сослаться на их использование?

4.1 НАУЧНЫЙ ЗООПАРК

- https://data.mendeley.com/
- https://www.dataone.org/
- http://www.openml.org/
- https://zenodo.org/