СИСТЕМЫ ОБРАБОТКИ БОЛЬШИХ ОБЪЕМОВ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ APACHE HADOOP (“SYSTEMS FOR PROCESSING LARGE AMOUNTS OF DATA USING APACHE HADOOP PLATFORM”)

Автор: студент 4-го курса, группы ДА-21 УНК «ИПСА» НТУУ «КПИ»
Загороднюк Андрей Александрович

Руководитель: Свирин Павел Владимирович

ЦЕЛЬ

- ✓ РЕШЕНИЕ ЗАДАЧИ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ ИСПОЛЬЗУЯ APACHE HADOOP ДЛЯ РАСПРЕДЕЛЁННЫХ ВЫЧИСЛЕНИЙ

ПРЕДМЕТ ИССЛЕДОВАНИЯ

- ✓ ИСПОЛЬЗОВАНИЕ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ HADOOP ДЛЯ ПРИКЛАДНИХ ВЫСОКОПРОДУКТИВНЫХ КЛАСТЕРНЫХ ВЫЧИСЛЕНИЙ

АКТУАЛЬНІСТЬ ЗАДАЧІ

- ✓ РАСПРЕДЕЛЕНИЕ ВЫЧИСЛЕНИЙ СНИЖАЕТ ЗАТРАТЫ И ВРЕМЯ, НЕОБХОДИМЫЕ ДЛЯ ВЫПОЛНЕНИЯ ЗАДАНИЯ

BIG DATA

- Серия подходов, инструментов и методов обработки данных
- Информация, которая не помещается на одном компьютере

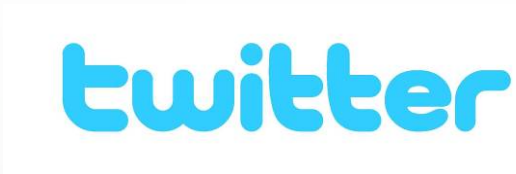
РАСПРЕДЕЛЁННЫЕ ВЫЧИСЛЕНИЯ

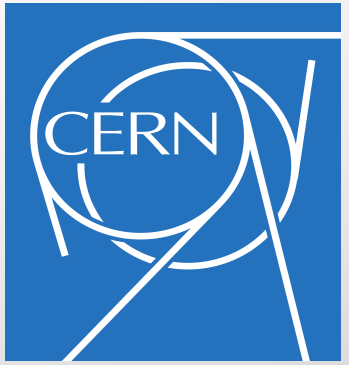
- География вычислительной среды
- Требование к наращиванию мощностей
- Общее использование ресурсов
- Отказоустойчивость

ДИСТРИБУТИВЫ HADOOP

cloudera

КТО ИСПОЛЬЗУЕТ HADOOP

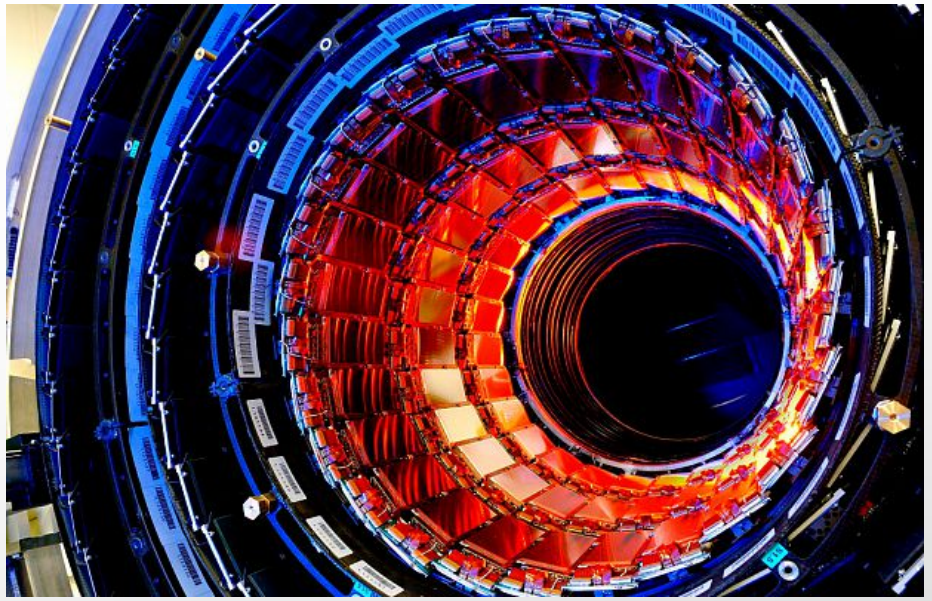




- Самый большой кластер Hadoop в Yahoo!:
  - 4500 серверов
  - Используется для поисковой системы и подбора рекламных объявлений

ЦЕРН

- Каждый год – 50 петабайт данных
- Три кластера:
  - CASTOR Cluster ~ 10 серверов(приблизительно 108 терабайт логов)
  - ATLAS Cluster ~20 серверов(работает с индексами каталогов экспериментальных данных)
  - Monitoring Cluster with ~10 серверов(логирует события с CERN Computer Center)



ОСНОВНЫЕ ТЕХНОЛОГИИ HADOOP

- HDFS (HADOOP DISTRIBUTED FILE SYSTEM) – Хранение данных
- MapReduce – Обработка данных

HDFS

- HDFS – специализированная файловая система, оптимизированная для параллельной потоковой работы с большими файлами
- Модель write once read many:
  - Нельзя изменять файл, можно только добавлять в конец
- Большой размер блока:
  - По-молчанию 64 МБ (часто 128 или 256 МБ)
  - Не эффективен произвольный доступ (базы данных и т.п.)

MapReduce

- MapReduce – технология распределенных вычислений
- Цель mapReduce – разделить логику приложения и организацию распределенного взаимодействия:
  - Программист реализует только логику приложения
  - Распределенная работа в кластере обеспечивается автоматически
- MapReduce работает с данными как с парами «ключ:значение»:
  - Смещение в файле: текст
  - Идентификатор пользователя: профиль
  - Пользователь: список друзей
  - Временная метка: событие в журнале

ТЕСТОВЫЕ ПРИМЕРЫ ИССЛЕДОВАНИЯ

- Генерация текстовых файлов
- Подсчёт количества слов (wordcount)
- Подсчёт максимальной температуры метеостанций
- Terasoft тест
- Решение судоку

РЕЗУЛЬТАТЫ

- Была исследована и проанализирована архитектура компонентов Apache Hadoop
- Было реализовано 5 прикладных задач, на примере которых были исследованы сильные и слабые стороны Apache Hadoop
- Получены результаты для дальнейших исследований

ЗАКЛЮЧЕНИЕ

- Система Apache Hadoop хорошо себя чувствует в среде больших данных
- Легко справляется с сортировкой огромного количества данных
- Возникают трудности при записи данных в файл