Как сегодня работает поисковая система

Поисковый кластер

Два дешевых сервера вычислят больше запросов и проиндексируют больше web-страниц, чем один дорогой.

Критерии выбора серверов:

- Стоимость
- Производительность
- Размер
- Потребление электроэнергии и тепловыделение

Много дешевых машин

Плюсы:

- Высокая производительность
- Низкая стоимость
- Простота изготовления, отсутствие «загадочных болезней»

Минус:

- Высокая частота отказов оборудования

И как же с этим бороться?

Программное обеспечение:

- Исключение сбойных серверов из кластера, перераспределение нагрузки на оставшиеся в строю машины;
- Хранение данных в нескольких экземплярах
- Непрерывный контроль целостности данных

Примеры

- RADIST: распределенное хранилище данных Рамблера;
- HICS: система для распределенного хранения и быстрой обработки сверхбольших массивов данных;
- Автоматическое «голодание» поисковых модулей.

Что в результате?

- Из ненадежного «железа» и специального программного обеспечения можно построить надежную и производительную систему.

Что хранится в кластере?

- Полный комплект веб-страниц, которые скачивал робот + частично хранится история изменения страниц
- Архив поисковых запросов
- Метаинформация
- Данные о посещаемости страниц Сети

Как объем данных помогает улучшить поиск

- Робот научился распознавать и удалять из URL необязательные параметры
- Индексатор стал лучше понимать естественный язык (повышение качества лингвистического анализа)
- Выявление «горячих» запросов и специальное ранжирование. Эврика!
- Разделение веб-страниц на смысловую часть и элементы навигации/дизайна.

Что ищут на Рамблере:

Авария Николая Караченцова:

- Небольшой всплеск перед публикациями в СМИ
- Резкое увеличение запросов сразу после первых сообщений
- Расширение тематики (номер машины, супруга, дилер, нейрохирурги и т. д.)
- Спад интереса

GTA San Andreas:

- Лавина запросов «коды gta san andreas»
- Горячий кофе

Сейчас:

- Зимняя резина, убийство Нуркадилова, пожар в сетуньском проезде, Т. Качарава и М. Згибай, IPS-19

Как это выглядит?

Как это выглядит?

Как это выглядит?