



ФОРС
ДИСТРИБУЦИЯ

Как начать использовать технологии BigData. Опыт и рекомендации.

Cloudera search
Внутренний проект
ФОРС Дистрибуция



Андрей Тамбовский
Директор по технологиям
ФОРС Дистрибуция

Предпосылки проекта

1. Желание получить опыт практического применения технологий BigData
2. Наличие 20+ опыта продаж лицензий ORACLE (прямых и через партнеров)
3. Изменение требований ORACLE → Variety
4. Наличие скрытых проблем – монотонная работа по поиску нужных данных

Как было: исходные данные

1. Много информации (GB-ТВ данных)
2. Различные форматы (doc, docx, pdf, ppt, xls и т.д.)
3. Запутанная структура (много папок с разными названиями, разные базы данных)
4. Несколько источников (файлы на диске, базы данных, сайты в интернете, почта...)

Как было: организация поиска

Как пользователь ищет информацию:

- a) Вспоминает где *примерно* лежит файл, дату создания, автора
- b) Запускает стандартный поиск Windows по параметрам (или Total Commander)
- c) *Ждет* время необходимое для обработки всех данных
- d) Получает список результатов

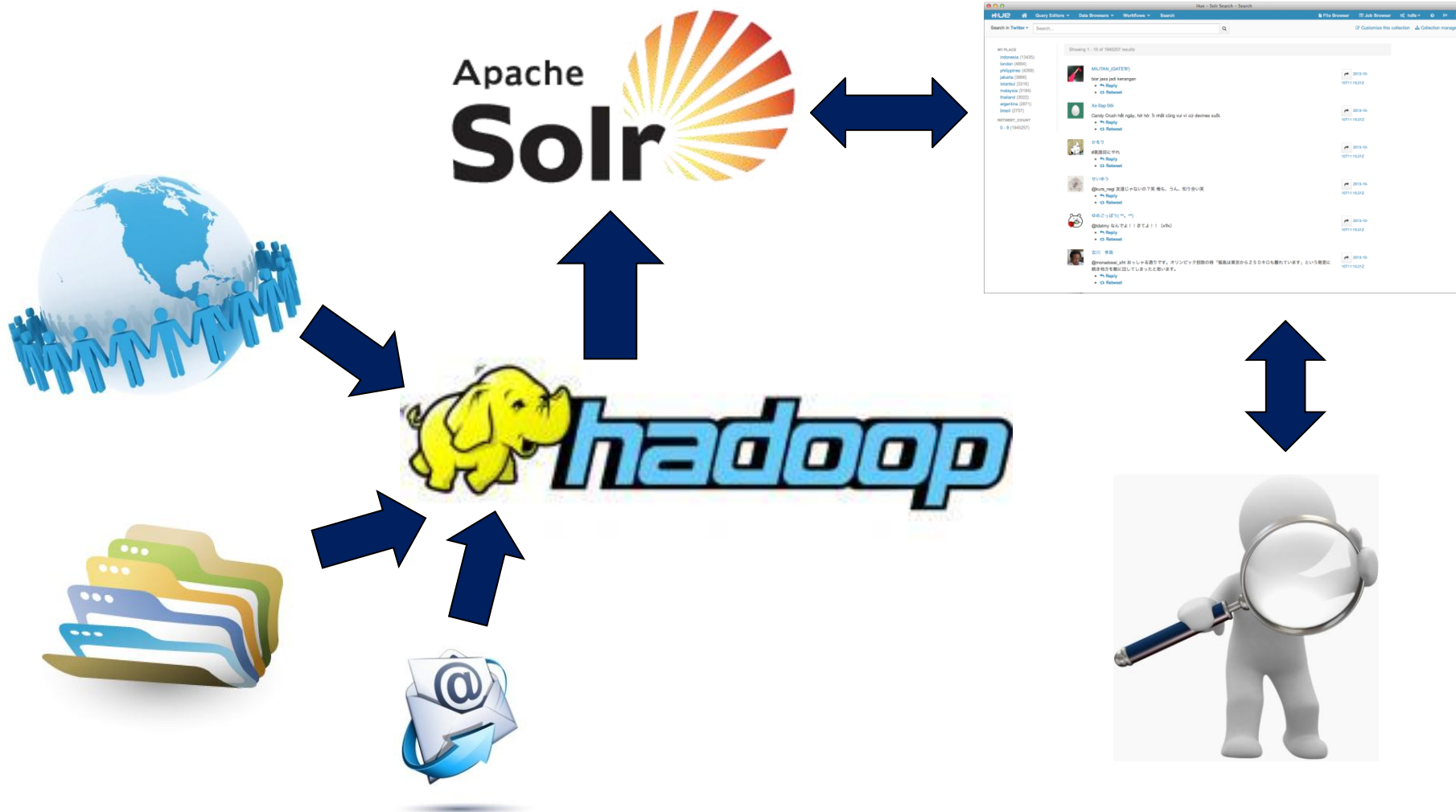
Как было: недостатки

- Длительное время ожидания, перед получением информации (*для поиска по 15GB данных при помощи Total Commander требуется более 2.5 часов*)
- Чем больше данных, тем дольше ждать
- Нахождение не всей информации по интересующей теме
- Возможность анализа очень ограниченного набора форматов файлов
- Необходимость установки специального ПО на ПК пользователя

Cloudera Search - преимущества

- Предоставляем пользователю *простой* поиск по текстам внутри файлов, схожий с поиском Yandex, Google, Yahoo
- Возможность настройки поисковика под свои нужды
- Возможность использования OCR
- *Безопасность* конфиденциальной информации. Различные уровни прав доступа
- Поиск без установки дополнительного ПО на машину пользователя. Нужен только браузер

Cloudera Search - как это работает?



Cloudera Search – выгоды

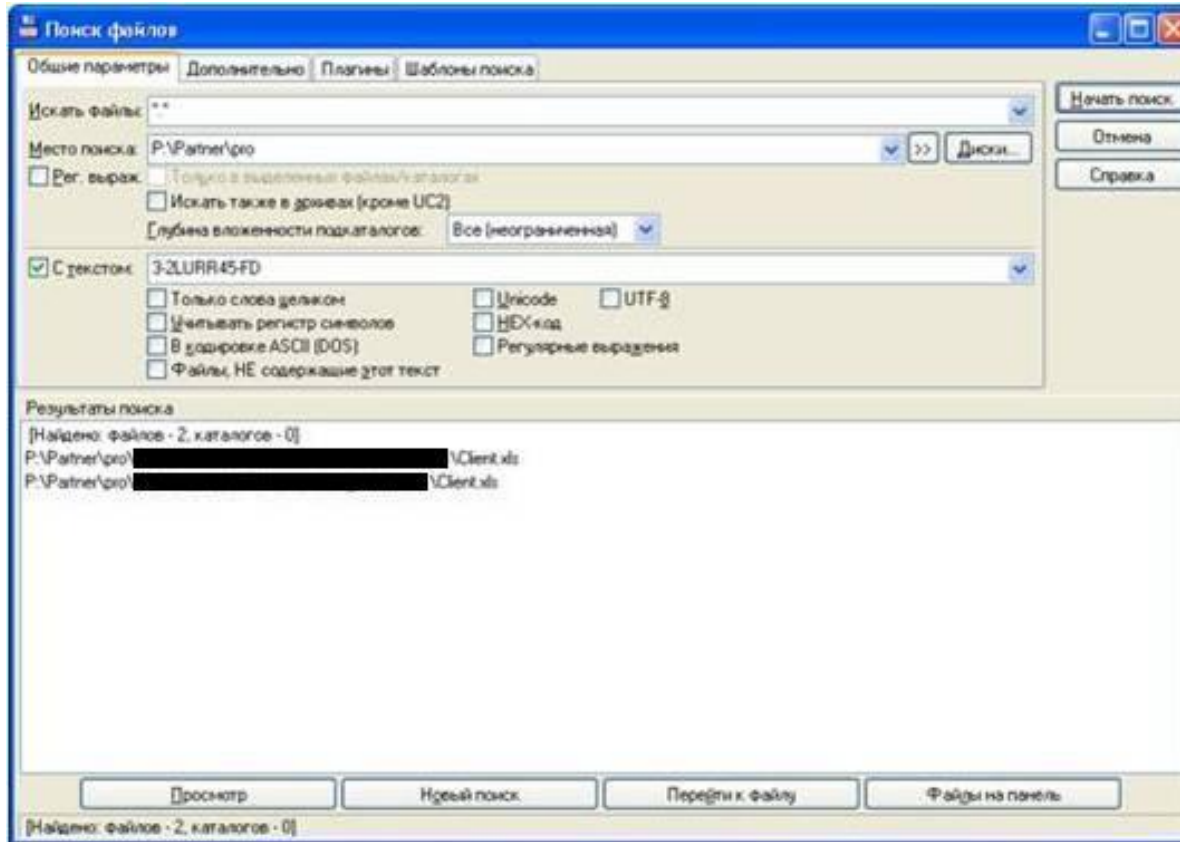
- Быстрый поиск нужной информации
- Возможность настройки фильтров, для сужения результатов поиска
- Быстрый доступ к файлу с интересующим его содержимым
- Поиск по тексту внутри файла
- Возможность поиска по метаданным (автору, дате создания)

Cloudera Search vs Старый поиск

Поиск лицензии 3-2LURR45-FD:

	Старый поиск	Cloudera search
Общий объем файлов	15 GB	15 GB
Затраченное время	~2 часа 30 минут	~1 минута
Количество найденных файлов	2 файла, не найдены файлы *.xlsx	4 файла
Средство поиска	Total commander	Любой браузер
Возможность поиска по картинкам	нет	Возможность подключить OCR

Внешний вид - Старый поиск



Внешний вид - Cloudera Search

The screenshot shows a web browser window with the Cloudera Search interface. The search bar contains the query '3-2LURR45-FD'. The results section displays four items, each with a title, date, author, and a link to the document. The first two items are 'Client.xls' files from 'Internal Systems' dated 2012-01-10T05:52:55Z. The last two items are 'VAD_OD_v092311-SW_(EN-RU_v01-OCT-11)_[redacted].xlsx' files from 'dmerkul' dated 2012-01-10T05:53:02Z. The sidebar on the left shows filters for file types (xls, xlsx), authors (Internal Systems, dmerkul), and dates (2 years ago - 2 years ago).

Showing 1 - 4 of 4 results

Название	Дата	Автор
Client.xls аказа: 3-2LURR45-FD ООО "[redacted]" Название и номер Договора: 117246, Россия, г. М /dc1.distr.fors.ru/share/partner/pro/[redacted]/Client.xls	2012-01-10T05:52:55Z	Internal Systems
Client.xls аказа: 3-2LURR45-FD ООО "[redacted]" Название и номер Договора: 117246, Россия, г. М /dc1.distr.fors.ru/share/partner/pro/[redacted]/Client.xls	2012-01-10T05:52:55Z	Internal Systems
VAD_OD_v092311-SW_(EN-RU_v01-OCT-11)_[redacted].xlsx аказа Дистрибьютора (VAD) (Гос. Сектор) Oracle ECE Ltd Oracle License №: 3-2LURR45-FD E /dc1.distr.fors.ru/share/partner/pro/[redacted].xlsx	2012-01-10T05:53:02Z	dmerkul
VAD_OD_v092311-SW_(EN-RU_v01-OCT-11)_[redacted].xlsx аказа Дистрибьютора (VAD) (Гос. Сектор) Oracle ECE Ltd Oracle License №: 3-2LURR45-FD E /dc1.distr.fors.ru/share/partner/pro/[redacted].xlsx	2012-01-10T05:53:02Z	dmerkul

Выводы

- Big Data – не маркетинг, а технология
- Big Data позволяет решать сложные задачи в короткие сроки
- Big Data позволяет решать сложные задачи при небольшом бюджете проекта
- Поиск применения технологий Big Data требует креативного мышления
- Самая большая проблема в проектах Big Data – определение Value

Контакты

«ФОРС Дистрибуция»

<http://www.partner.fors.ru>

129626, Москва, Графский
переулок, 14, стр. 2

Телефон: +7 (495) 913-3-913





ФОРС
ДИСТРИБУЦИЯ