

Разработка алгоритмов обеспечения качества
распределенного поисковго робота для сети
Интернет

Волков Сергей

19 мая 2011 г.

Глава 1

Введение

— В то время, когда наши
корабли бороздят просторы
Вселенной. . .

На текущем этапе развития, когда общество осуществляет переход от постиндустриальной эпохи к информационной, требования к системам хранения и обработки информации непрерывно растут. Традиционные подходы не справляются с ростом количества данных. Трудно оценить общий объем данных, однако, по оценкам IDC (International Data Corporation) в данный момент хранится порядка $1.8 \cdot 10^{21}$ байт, что в 10 раз больше чем в 2006 году.

К значительному количеству данных можно получить доступ через Всемирную Паутину (www). При таких объемах остро стоит задача организации эффективного поиска. Уже в 2009 году Google Search обработал более 109,5 миллионов сайтов, и более 10^{12} уникальных URL. На данный момент их индекс содержит $4 \cdot 10^{10}$ документов.

Одной из специфических областей поиска является поиск по новостным ресурсам. Для документов с новостных сайтов характерна привязка к дате, региону и тематике. Таким образом такие документы легко классифицировать, что позволяет производить более качественный поиск и анализ. В качественном инструменте для анализа СМИ заинтересованы различные консалтинговые и pr агентства, пресс-службы, маркетинговые отделы крупных компаний.

Одна из задач поисковой системы - нахождение и загрузка документов (Web crawling), за которую отвечает поисковый робот (Spider, Crawler). Web crawling весьма ресур-

соемкий процесс. Основные проблемы связаны с большим количеством данных, отсутствием контроля над данными, постоянным изменением структуры ресурсов, динамическим созданием страниц и низким качеством некоторых ресурсов. Однако, специализация на определенной узкой области web позволяет существенно повысить производительность web crawler'a.

Конечной целью работы является создание системы способной эффективно индексировать новости в рунете.

Глава 2

Обзор области

2.1 Web Search

Поисковая система — система, разработанная для поиска информации в www. Результаты поиска которой, как правило, представлены в виде списка “попадений”. Информация может состоять из веб страниц, изображений, мультимедийной информации. Одной из первых поисковых систем стал проект Archie, разработанный в 1990 году студентами McGill University. Программа скачивала списки файлов с открытых ftp серверов, и добавляла их в базу с возможностью поиска по названию.

Поисковая система состоит из трех основных компонент:

- **поисковый робот** — программа, предназначенная для перебора документов и занесения данных о них в базу.
- **индексатор** — программа, создающая на основе полученных с помощью робота данных индекс.
- **поисковик** — программа, осуществляющая поиск в полученном индексе на основе поискового запроса.

В условиях постоянно расширяющегося и изменяющегося www, непрерывно возрастают требования к поисковым системам.

Системы общего поиска нацелены на охват большей части данных доступных в www. Такие системы предназначены для поиска наиболее релевантных документов относящихся к объекту поиска.

Системы тематического поиска более разнообразны, и требования к ним более специфичны. Например Google Microblogging Search Engine, ориентированный на поиск по записям в микроблогах, где крайне важна задержка между созданием записи, и ее попаданием в индекс.

2.2 Поиск по новостям

Основные источники новостей в www — это электронные СМИ и блоги. По данным liveinternet на 2008 год, рунет насчитывает 4392 сайта СМИ, а число блогов значительно больше - по данным Яндекс за 2009 год в русскоязычной блогосфере насчитывается порядка 840000 активных блогов, на которых ежедневно публикуется порядка 300000 постов.¹ Очевидно, за прошедшее время количество таких сайтов значительно увеличилось. За сутки каждое из подобных изданий публикует до 100 документов(lenta.ru). Таким образом, можно говорить о десятках миллионов создаваемых документов в год.

Под новостью понимается документ содержащий текст, заголовок и дату. Для СМИ и блогов характерно:

- большое количество посторонних страниц, не содержащих новостей;
- схожая структура (как именован url, так и самого html);
- наличие rss ленты.

К новостным поисковым системам предъявляются следующие требования:

- минимальное время между публикацией статьи на новостном ресурсе и ее предоставление в поисковой выдаче;
- поиск должен осуществлять не по всей HTML-странице, а только по ее существенным частям.

2.3 Постановка Задачи

Конечной целью работы является создание поискового робота способного эффективно индексировать новости в рунете.

¹<http://mediarevolution.ru/audience/1962.html>

Поисковый робот (Web crawler) — программа для поиска веб-страниц в сети[1]. Грубо говоря поисковый робот начинает с URL для начальной страницы p_0 . Он скачивает p_0 , выделяет все URL которые в ней находятся, и добавляет их в очередь URL (*crawling frontier*). Затем робот в некотором порядке выбирает URL из очереди и повторяет процесс.

Каждая скачанная страница передается клиенту, который затем создает индекс по страницам.

2.3.1 Условия

Важным фактором, влияющим на качество поиска, является идентификация страницы содержащей новость и выделение её содержательной части. В данной работе предполагается наличие базы данных, содержащей правила на основе регулярных выражений, которые по URL определяют содержит ли данная страница новость, а так же правил по которым из веб-страницы выделяется содержательная часть.

Далее под *документом* понимаются выделенные по этим правилам данные.

2.3.2 Требования

- поддержка десятков миллионов документов;
- скорость роста базы документов — более 50 тысяч в день;
- попадание в индекс документов из rss лент не позже чем через 12 часов после их публикации;
- попадание в индекс старых документов из архива — некоторые документы могут находиться достаточно “глубоко” (например что-бы получить новости месячной давности на ресурсе fontanka.ru необходимо сделать 5 переходов).
- отсутствие дубликатов в пределах одного домена — под дубликатом понимается документ у которого совпадает URL или текст с другим документом.

В качестве основного показателя эффективности используется количество полученных документов за сутки. Поскольку объем данных непрерывно увеличиваться, неизбежна деградация производительности. Поэтому также в качестве метрики используется отношение количества полученных за сутки документов к общему числу документов.

2.4 Результаты

- Сделан сравнительный анализ различных open source поисковые роботы (DataparkSearch, AppSeek, mnlGoSearch, Nutch, Hounder, Heritix) и выбрать наиболее подходящий для решения задачи
- Изменено поведение ядра nutch для более эффективной работы с индексом большого объема.
- Проанализированы различные key-value хранилища (Memcached, MongoDB, Project Voldemort, Tokyo Cabinet) и выбрано MongoDB в качестве хранилища для системы удаления дубликатов из индекса
- Разработан и реализован плагин к Nutch для раннего удаления дубликатов
- Разработан и реализован плагин для более эффективного ранжирования ссылок для новостных сайтов
- Разработана и реализована система для автоматического создания url фильтров
- Измененная система протестирована на реальных данных.

Глава 3

Обзор средств

Большинство популярных поисковых сервисов предоставляют возможность поиска по новостям (google, yandex, yahoo!), однако они пользуются закрытыми алгоритмами и не предоставляют доступа непосредственно к индексу

3.1 Сравнение open source поисковых роботов

Существует достаточно много open source поисковых роботов. Для успешного решения задачи робот должен справляться с нагрузкой (порядка 100000 документов в день, база ссылок порядка 10^9 и порядка 10^7 документов в индексе), быть легко изменяем и расширяем. Поскольку предполагается коммерческое использование робота не протяжении долгого времени, проект должен быть достаточно зрелым и развивающимся.

Метрики

- язык
- поддержка robots.txt
- распределенность системы
- тип хранения индекса
- тип хранилища url
- поддержка

Роботы

- DataparkSearch — поисковая система разработанная для поиска по локальным файлам, группам сайтов и интранету
- AspSeek — поисковая система оптимизированная для работы с многими сайтами, и средней загрузкой — до нескольких миллионов страниц
- Nutch — поисковая система основанная на Lucene
- Hounder — поисковая система онованная на nutch
- Heritix — еще одна java поисковая система

Сравнение

3.1.1 Описание роботов

DataparkSearch ¹ — предназначен для работы с небольшой группой сайтов или интранета, написан на С. Состоит из двух частей — индексатора и CGI фронтенда. DataparkSearch отделился в 2003 году от mnoGoSearch. Имеет встроенные парсеры для html, xml, есть возможность написания собственных парсеров для других форматов. Данные по ссылкам хрянятся в SQL базе данных. Можно запустить сразу несколько процессов индексации работающих с одной базой. Данные по документам могут храниться как в бд, так и в собственном формате на диске (cache mode), который эффективно работает с несколькими миллионами документов.

AspSeek ² — поисковая система написанная на C++ и оптимизированная для работы с множеством сайтов. Состоит из индексирующего робота, поискового демона и CGI фронтенда. Данные поискового сервера хранятся в SQL базе данных и бинарных файлах (delta files), рассчитан для работы с несколькими миллионами документов.

Nutch ³ — поисковый робот написанный на java, работающий поверх системы Hadoop⁴. Изначально Nutch разрабатывался в рамках проекта Lucene⁵, однако в 2005 году отделился как отдельный проект. Благодаря работе поверх Hadoop обладает хорошей

¹<http://www.dataparksearch.org/>

²<http://www.aspseek.org/>

³<http://nutch.apache.org/>

⁴<http://hadoop.apache.org/>

⁵<http://lucene.apache.org/>

масштабируемостью (до 100 машин в кластере). Nutch отличается гибкой системой плагинов, через которые осуществляется поддержка множества протоколов (http, ftp, file) и форматов (от html до msexcel и swf).

Hounder ⁶ — поисковая система на java, робот которой основан на Nutch. Из дополнительного функционала следует отметить фильтр Байеса для разбиения документов по категориям.

3.1.2 Выбор

В качестве основы системы был выбран Nutch, так как он полностью удовлетворяет требованиям:

- нагрузка — Nutch использовался в качестве основы для Sapphire Web Crawler⁷, с помощью которого было скачано более 10⁹ документов со средней скоростью в 431 документ в секунду.
- расширяемость — благодаря модульности и гибкой системе плагинов можно достаточно легко изменять поведение системы.
- поддержка — проект разрабатывается более 7 лет, текущая стабильная версия проекта 1.2 была выпущена в сентябре 2010. Проект поддерживается “Yahoo! Research Labs”.

3.2 Архитектура Nutch

Высокая масштабируемость робота достигается за счет работы поверх MapReduce фреймворка *Hadoop*[3]. *Hadoop* на данный момент представляет набор подпроектов Apache Software Foundation, среди которых находятся Hadoop MapReduce и HDFS.

MapReduce — модель программирования для обработки больших объемов данных, впервые опубликованная[2] Google в 2004 году. При данном подходе логика программы реализуется в функциях *map*, которая преобразует пары ключ/значение

⁶<http://hounder.org/>

⁷<http://boston.lti.cs.cmu.edu/crawler/index.html>

в набор промежуточных пар ключ/значение, и *reduce*, которая обрабатывает все значения связанные с одним промежуточным ключом 3.1.

$$\begin{aligned} \text{map} : \langle \text{key}_{in}, \text{value}_{in} \rangle &\rightarrow \langle \text{key}_{int}, \text{value}_{int} \rangle^* \\ \text{reduce} : \langle \text{key}_{int}, \text{value}_{int}^+ \rangle &\rightarrow \langle \text{key}_{out}, \text{value}_{out} \rangle^* \end{aligned} \quad (3.1)$$

Написанная таким образом программа может автоматически параллельно выполняться на кластере машин, программное обеспечение которых брало бы на себя распределение данных, управление выполнением задач, поддержку отказов и управление взаимодействием между узлами кластера.

HDFS (Hadoop Distributed File System) — распределенная система хранения данных, основанная на модели GFS[4] — распределенной файловой системы используемой Google. *HDFS* предназначена для:

- больших файлов — имеются ввиду файлы от нескольких сотен мегабайт до нескольких терабайт;
- потокового доступа к данным — предполагается что данные записываются один раз, и программа в процессе работы использует большую часть из записанного набора данных;
- дешевого железа — риск отказа оборудования достаточно высок.

Файл в *HDFS* представляет из себя последовательность достаточно больших блоков (по умолчанию 64Mb), которые в нескольких экземплярах (обычно используется 3 реплики) хранятся на различных узлах — *DataNode*. Последовательность блоков в файле и их расположение управляется через *NameNode*. Таким образом нагрузка на передачу и запись данных распределяется между набором *DataNode*, а структура папок и файлов находится в *NameNode*.

3.2.1 Основные Этапы

Nutch является средством инкрементальной сборки, на каждом этапе выполняются следующие действия:

- *inject* — добавление списка URL в базу ссылок *crawldb* (используется при инициализации сборки или при добавлении новых доменов);

- *generate* — из *crawldb* выбирается фиксированное число ссылок для их последующего скачивания;
- *fetch* — скачивание документов по выбранным ссылкам;
- *parse* — документы парсятся, выделяются ссылки;
- *invertlinks* — обновляется база обратных ссылок;
- *index* — создается индекс по сегменту;
- *merge index* — индекс по сегменту объединяется с основным;
- *update* — обновляется *crawldb*

3.2.2 Система Плагинов

Особого внимания заслуживает система плагинов в *Nutch*, именно через плагины реализована основная функциональность. Плагины выполняют разбор документов, индексацию, поиск, ранжирование, фильтрацию ссылок и т.д.

Каждый плагин предоставляет одно или несколько *расширений* (*extensions*) для *точек расширений* (*extension points*), причем сами по себе *точки расширений* определены в плагине.

Таблица 3.1: Сравнение поисковых роботов.

Название документов	Язык	Распределенность	robots.txt	Индекс	Хранилище url	Количество
DataparkSearch	C		+	SQL database/собственный формат	SQL database	10 ⁶
AspSeek	C++	??	+	SQL database	SQL database	10 ⁶
Nutch	Java	+	+	Lucene index	распределенный файл	10 ⁹
Hounder	Java	+	+	Lucene index	распределенный файл	???

Глава 4

Алгоритмы

Подход к решению задачи:

1. запустить систему без модификаций;
2. найти самое слабое место в системе;
3. принять меры по его устранению;
4. перейти к пункту 2.

4.1 Ранжирование

При работе системы в стандартной конфигурации было замечено, что только из 10% скачиваемых веб-страниц выделяются документы. Это происходит из-за неоптимального упорядочивания ссылок из *crawldb*.

Определение порядка выбора URL для скачивания существенно сказывается на эффективности работы робота.[1][5][6] Порядок неважен только в том случае, если робот нацелен на одноразовое скачивание всего Web, и нагрузка создаваемая роботом на целевые сайты не важна, так как тогда каждая известная URL будет в конце концов загружена. Однако большинство роботов не способны посетить каждый URL по трем основным причинам:

- Ограничение по ресурсам — размер хранилища, ширина канала, CPU time для обработки страниц.
- Сбор документов занимает время, поэтому в определенный момент робот вынужден заново посещать некоторые страницы для нахождения изменений.

- Динамическое создание страниц — сейчас большинство сайтов раздают не статический контент, а создают его динамически при помощи скриптов обрабатывающих URL и возвращающих результат, таким образом количество страниц на сайте может быть неограничено.

Во всех остальных случаях существенно что бы робот сначала посещал “важные” страницы. Ранжирование отвечает за определение того, насколько URL “важна”. В Nutch порядок выбора URL определяется в плагинах подключенных к точке расширения *ScoringFilter*, в результате работы которых каждой URL сопоставляется *score*. На каждой итерации для скачивания выбираются URL с наибольшим *score*.

В стандартной конфигурации Nutch для выбора ссылок используется *OPIC Score* плагин.

4.1.1 OPIC

OPIC — On-Line Page Importance Computation[6]. Данный алгоритм рассчитывает важность страницы на основе важности страниц на него ссылающихся. В отличие от off-line методов, когда сначала скачивается часть Web, а потом на основе полученного графа ссылок вычисляется важность страниц, этот метод позволяет работать когда большая часть сетевого графа еще не известна.

В Nutch реализована упрощенная версия данного алгоритма, при котором в момент скачивания страницы к *score* каждой из ссылок с нее (*outlink*) добавляется S_0/n где S_0 — *score* скачанной страницы, а n — число исходящих ссылок. Данный способ плохо подходит для поиска новостей, поскольку для новостей основным признаком “важности” веб-страницы является присутствие в нем текста новости, что на практике не связано с количеством входящих ссылок.

Поскольку в системе идентификация новостей производится только по URL возникает желание скачивать только те URL которые подходят под правила.

4.1.2 Предложенный алгоритм

Так как в Nutch можно использовать одновременно сразу несколько *ScoringFilter*’ов (полученные из различных алгоритмов *score* просто перемножаются), вместо изменения стандартного метода следует создать отдельный плагин, который всем новым поступающим в систему URL которые подходят под правила разбора устанавливать $score = s_{fit}$ и s_n для остальных.

4.1.3 Особенности реализации

Данная функциональность была реализована при помощи двух плагинов. Первый проверяет является ли URL “полезной”, и добавляет необходимую информацию в *crawldatum* (в *crawldb* хранятся записи вида $\langle URL, crawldatum \rangle$, где *crawldatum* — различная дополнительная информация об URL), а второй выставляет ранг согласно *crawldatum* и настроек приложения. Такое разбиение было сделано потому что информация о том, что url является “полезной”, представляет собой интерес и без использования данного ранжирования (например это было использовано при получении статистики).

SkaiScoringMeta

SkaiScoringMeta — плагин являющийся расширением к точке ScoringFilter, выделяющий “полезные” URL. Для хранения дополнительной информации в *crawldatum* используется поле *metaData*, которое представляет из себя набор пар вида $\langle key, value \rangle$. Для сохранения индикатора “полезности” в мета данные *crawldatum* добавляется пара $\langle skai.scoring.fit, true \rangle$.

Для определения “полезности” URL используется кэш с регулярными выражениями, загружаемыми из базы данных при инициализации. Поскольку таких регулярных выражений может быть много (хотя бы по одному на сайт) и они разбиты по доменам, для сокращения времени проверки каждой URL используется хэш вида $\langle domainname, regex^+ \rangle$, таким образом время проверки URL практически не зависит от числа доменов.

Так же для уменьшения числа проверок, метаданные добавляются не в момент первого обнаружения ссылки системой (сразу после разбора веб-страницы), а только в момент добавления уникальной URL в *crawldb*, что уменьшает число проверок приблизительно в 10 раз, так как в среднем только 9% исходящих со страницы ссылок не были ранее известны.

SkaiScoring

SkaiScoring — плагин являющийся расширением к точке ScoringFilter, выставляющий *score*. Для ссылок только что попавших в систему, зависимости от наличия мета тега, выставляет значения *score* либо *s_{fit}*, либо *s_n*, где коэффициенты *s_{fit}* и *s_n* получаются из файла конфигурации *nutch*.

4.1.4 Результат

При использовании стандартных ScoringFilter'ов в среднем за цикл, из всех скачиваемых ссылок “полезных” скачивалось порядка 10% (В качестве пространства поиска использовалось 40 крупнейших СМИ, на каждой итерации выбиралось 20000 ссылок, “полезными” признавались непосредственно новости). А при использовании данных плагинов уже на ранней стадии сборки порядка 80% ссылок были “полезными” (Рис. 4.1). Благодаря предложенным оптимизациям время цикла практически не изменилось по сравнению с временем работы в стандартной конфигурации. Таким образом было получено увеличение эффективности в восемь раз по сравнению с базовой реализацией nutch.

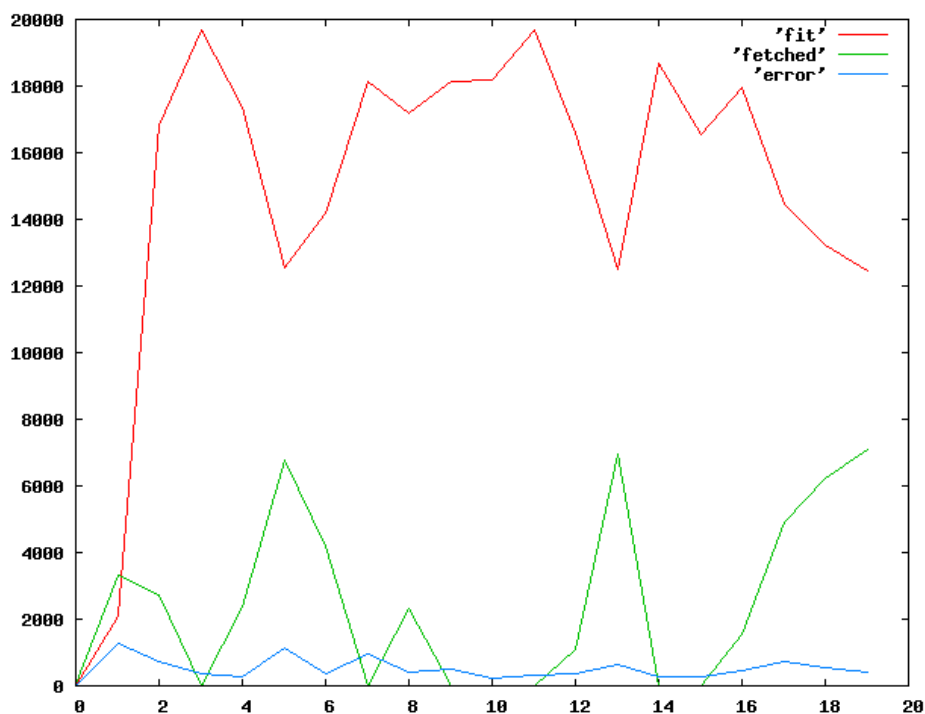


Рис. 4.1: График зависимости числа ссылок от номера итерации. fit—число полезных ссылок, fetched—число скачанных ссылок не являющихся полезными, error—число ошибок при скачивании

4.2 Работа с индексом

При росте размера индекса прямопропорционально растет время используемое для слияния полученного из сегмента индекса с основным. Этот этап не работает через MapReduce и фактически выполняется на одной машине, что приводит к простоям

кластера. Для индекса содержащего 2 миллиона документов (70GB) только копирование индекса из HDFS в локальную файловую систему и обратно занимает больше двух часов (при скорости локальной сети 20Mb/s), а общее время работы достигает 11 часов (2×2.5 GHz 2007 Xeon, 1.7GB RAM, скорость последовательного доступа к диску 50MB/s). При размере *crawldb* в 80 миллионов ссылок и генерации сегмента в 80 тысяч URL на слияние индекса уходит 68% от общего времени цикла (на кластере из трех узлов).

4.2.1 Lucene

Lucene — библиотека полнотекстового поиска. Индекс *Lucene* построен по принципу обратного файла (инвертированного индекса). Сам индекс состоит из сегментов, каждый из которых грубо говоря представляет отсортированный список термов. Для ускорения поиска сегменты “склеиваются” в один. Операция “склеивания” представляет из себя этап слияния из сортировки слиянием, и производится за время прямопропорциональное сумме размеров склеиваемых сегментов. По-умолчанию *Nutch* хранит индекс в одном сегменте, к которому на каждой итерации “приклеиваются” новые сегменты индекса, таким образом за каждую итерацию весь индекс перезаписывается.

Было рассмотрено несколько подходов к решению данной проблемы:

- **Асинхронное слияние** — выделяется отдельный сервер для хранения индекса, который по находит все новые сегменты, скачивает их и независимо от процесса основной сборки создает обновленный индекс. При этом время на слияние индексов не влияет на скорость скачивания документов, однако, уже при размере индекса в 5.3 миллионов документов индекс будет обновляться только один раз за сутки, из-за чего новости за последние 30 часов никогда не появятся в индексе (время на слияние+время цикла).
- **Изменение политики слияния** — поскольку *Lucene* может искать по нескольким сегментам можно не склеивать их в один, но при этом при большом количестве сегментов скорость поиска существенно падает т.к. возрастает время на поиск термина во всех словарях. Так что тут надо действовать хитро...

Литература

- [1] Junghoo Cho, Hector Garcia-Molina, Lawrence Page: Efficient Crawling Through URL Ordering, 1998.
- [2] Jeffrey Dean and Sanjay Ghemawat: MapReduce: Simplified Data Processing on Large Clusters, 2004.
- [3] Tom White: Hadoop: The Definitive Guide, 2009
- [4] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung: The Google File System, 2003.
- [5] Martin Ester, Matthias Groß, Hans-Peter Kriegel: Focused Web Crawling: A Generic Framework for Specifying the User Interest and for Adaptive Crawling Strategies, 2001.
- [6] Serge Abiteboul, Mihai Preda, Grégory Cobena: Adaptive On-Line Page Importance Computation, 2003
- [7] Пименов Александр, Hadoop Nutch и Lucene v3, 2009.
- [8] D Cutting, Nutch: an Open-Source Platform for Web Search
- [9] E Hatcher, O Gospodnetic, Lucene in action
- [10] R Khare, D Cutting, K Sitaker, A Rifkin, Nutch: A flexible and scalable open-source web search engine
- [11] T White, Hadoop: The Definitive Guide