

Check Up Master Datasheet 4/26/2004

Introduction

The Verity K2 configuration Check-Up module evaluates your search engine configuration files and logs to identify problems. It checks for common configuration problems that may occur in document spidering or text extraction, filtering, indexing, and search term expansion. The K2 Check-Up also provides a detailed analysis of your configuration and a suggested list of configuration changes that can improve quality of results and search engine performance. The general outline of the document is:

What you get	What information you will receive as a result of the check-up
The process	The things we will do with the content
Getting Started	The things we need from you, or the things we need access to, in order to start the check-up

1. What you get from New Idea Engineering

Reports

Once the tools and utilities have analyzed your style files, collections, system configuration and indexing logs, New Idea Engineering prepares a site report for you. Often, much of your installation is working correctly and small adjustments can make K2 work fine for you; other times, significant work and even redesign are indicated. In either case, the reports and expert analysis will give you the information you need to move forward to correct the problem.

Your users will be able to find the content they need; your content creators will know what users are looking for and can be confident that the right answers are showing up; and your management team will know that the investment in Verity K2 is paying off in lower support costs and better levels of service.

Collection:	support	corpcoll	web_site
Overall Health	Good	Check	Check
Last Update	4/12/2004	4/2/2003	12/19/2003
Currency (days)	13	389	128
Partition Health	Good	Check	Good
Document Count	3357	9428	396
Verity Broker	corp1	corp1	n/a
Verity Server	sup_server	hq_server	www
Alias	support	corporate	website
System	supportSite	corpSite	webSite
Gateway	Unknown	Web	FileSystem
Style Set	manual	manual	Def_FileSys
Doc Access	Public	Public	Public
Locale	default	default	english
Partition Count	1	9	1
Partition Doc Count	3414	9432	401
BIF Recs	3684	n/a	n/a
BIF Date	2/22/2004	4/10/2004	n/a
Index Tool	mkvdk	mkvdk	vspider
Build Script	support_build.bat	corp_coll_full.bat	nightly.txt
Update Script	support_update.bat	corp_coll_delta.bat	
Supporting Files	prep_support.pl		
Output Logfile	none	corp_status.log	none
start url	web.ideaeng.com/support	\\hq\content\	http://www.ideaeng.com
Collection Version	K2 4.51	K2 4.51	K2 4.51
Index By	schedule	schedule	manual
Notes	bulk file updated by SQLtable		
Page 1 of 3	(Dates as of 4/15/2004)		New Idea Engineering, Inc.

Sample Collection Report

Collection Report

One of the most comprehensive reports generated as a result of the Check Up is the Collection Status report. The meaning of each row follows.

Row Title	Row Description
Overall Health	General rating of the collection optimization: Good/Check/Poor
Last Update	Date the collection was last indexed
Currency (days)	How many days have passed since the last index
Partition Health	General rating of collection efficiency: Good/Check/Poor
Document Count	The number of documents indexed in the collection

Verity Broker	The system name of the Broker(s) that searches this collection
Verity Server	The system name of the Server(s) that searches this collection
Alias	The collection alias within the K2 Broker/Server network
System	The system DNS name or IP address of the platform where the collection is physically located
Gateway	The K2 gateway to the indexed documents: FileSys, Web, Outlook, Database, etc
Style Set	The name of the StyleSet editor style files used for this collection or “manual” is the style files were created externally
Doc Access	The type of access to the documents through K2: Secure/Public/Anonymous
Locale	The K2 locale that defines character set and stemming rules
Partition Count	The number of active partitions in the collection
Partition Doc Count	The total number of active and deleted documents within all active partitions
BIF Records	If bulk insert files are used to index the collection, the number of records that were submitted for indexing
BIF Date	The date of the latest bulk insert file used to create a collection
Index Tool	The K2 index tool used to index the collection: mkvdk/vspider/k2spider/other
Build Script	The batch, shell, or spider script used to index the collection
Update Script	The batch, shell, or spider script used to update the collection if incremental updates are used
Supporting Files	Any supporting files used in the index and update process

Output Logfile	The name of the log file created by the indexing tool during its most recent run
Start URL	The starting URLs or file UNC/paths for initial indexing
Collection Version	The version of the K2 system that created the collection
Index By	The method used to start periodic indexing: scheduled/manual/other
Notes	Additional notes relevant to the collection

Collection Report Categories

Other Reports

The K2 Check-Up includes a number of other reports including:

collection style file analysis	Confirms consistency within all style files
word/spanning word summary	Checks for excess or meaningless words which might indicate a bad style setting
document type summary	Verifies all indexed document types and counts

Other Included Findings Reports

Recommendations

Included with the K2 Check-Up and reports is a list of recommendations. Some of them are typically quite easy to implement, while others can be complex and require more time. Some of typical recommendations are shown below.

Using FIELD operators in Searches

Symptom: Slow searches

The search script uses field operators <CONTAINS>.

We recommend you define zones for those fields which you will use in searches, and utilize the more efficient zone operator <IN>.

Collection Document Count inconsistent with Bulk Insert File

Symptom: Missing documents

The number of records included in the bulk insert file does not match the number of records in the indexed collection.

We suggest you perform a collection audit. Analyze the log files to identify the documents that were not indexed, correct the problem, and re-index the collection.

Low Document to Partition Ratio

Symptom: Slow searches; large number of partitions in parts directory

The collection has a small average number of documents indexed in each partition, which could indicate inefficient indexing or high document turnover.

Schedule a complete re-index of all documents or perform an optimize merge after backing up the entire collection. Note that submitting bulk insert files with less than 64000 records is inefficient, and can lead to poorly optimized collections.

Review your existing index and update scripts and your approach to incremental updates to maintain optimum partition size, and regularly optimize or perform a full index.

Synonym Operator without Custom Thesaurus

Symptom: Many Irrelevant Results

The search script uses THESAURUS operator with no custom thesaurus defined.

Generally, use of the THESAURUS operator will return large number of unwanted and irrelevant documents when using the standard English thesaurus. Review your results and if this is the case, create a customer thesaurus based on your user search activity and your site-specific vocabulary.

Incorrect Document Dates

Symptom: Cannot find newest documents using date operations.

When results are ordered by date, the dates are wrong and the newest documents don't show up in the right date order.

Date problems can reflect server or gateway configuration problems, document parsing problems, or other problems in a bulk insert file. Reconfigure your web or gateway server to return accurate dates, or process the documents so the meta-data contains the document actual date. It is also possible to extract dates from existing meta-data such as datelines or dates included within the document for inclusion in bulk insert files.

Bad Summaries

Symptom: Summaries are inaccurate or contain JavaScript and HTML tags

When you review search results produced by K2, they are not very good or have JavaScript or HTML tags instead of valid content.

Two style files control the summaries: STYLE.PRM and STYLE.FSX. Review them for optimum settings. If your content has meta-tag descriptions, verify you are using the document description rather than the Verity calculate summary.

Bad Document Titles

Symptom: Document titles are empty or meaningless, or are all identical

When you review search results, the document titles are short and meaningless, or all documents have the same title.

Bad titles is often a result of office format documents saved without any properties, or HTML documents based on the same template. Process source documents to extract likely titles from content depending on your corporate guidelines – for example, perhaps the initial H1 tag provides a reasonable title.

We are careful to give recommendations that have the most impact on your problem with the least impact on your operations, so you can judge the return you will see on your investment and effort.

2. The Process

New Idea Engineering and the K2 Check-Up can get you on track to producing better search results, potentially reducing your support costs while increasing your user satisfaction. We have a tested methodology to performing the Check-Up, described here.

A. Pre-Check-Up

Prior to the Check-Up, there are a number of details that need to be addressed. These include the following.

1. Mutual Non-Disclosure Agreement

Because New Idea Engineering will be accessing your systems and data, and because you will have access to proprietary NIE tools and utilities, we suggest that we each execute a mutual non disclosure agreement. We have a standard form, or we can use your company's mutual NDA form.

2. Identify Primary Contacts

Both New Idea Engineering and your company will identify primary project managers for business and for technical matters during the check-up.

3. Arrange for Access

The K2 Check Up requires access to command line and web tools on the server where an instance of Verity K2 is installed. Access can be handled in three different ways, depending on your needs and policies:

- New Idea staff comes on-site to perform the check-up
- New Idea staff uses remote access via VPN/Terminal Services/SSH Telnet
- NIE receives copies of files and collections via ftp or CD/DVD

During this process, we will also identify the hardware and software we will be using in the project.

4. Issue a Purchase Order for the Check Up

Your company purchase order can be faxed or emailed to our facilities, as well as any other forms your company may require such as contractor policies, tax id numbers and other administrative paperwork.

5. Schedule the Start Time

Once the purchase order is issued, the project contacts will select a mutually acceptable date to start the project. Depending on availability, the lead time can be as little as a week once all other dependencies are resolved.

6. Verify Software Environment

- Java 1.3 or above run-time
- Oracle 8/9, SQL Server 2000 or higher or PostgreSQL
- Search 97, Information Server, K2 V4.01 or above

B. On-Site or Remote Access

Once we have a purchase order and have agreed to a start date, the check-up can begin. Over the five days required for the check-up, the process will follow this typical schedule.

Day 1

- Arrive on site or confirm remote access
- System interview with your technical staff to identify locations of files, HTML forms and scripts, and system topology.
- Install software and tools for search analytics, data discovery, and content check.
- Configure search tracking system
- Begin search capture and analytics

Day 2

- Gather indexing scripts, check log levels
- Gather index log files
- Create test area for collections
- Verify scripts work as expected with correct output level
- Run capture tools on collections

Day 3

- Begin integrating data into reports
- Verify no collections/servers/brokers have been missed

Day 4

- Process data into reports

Day 5

- Deliver report findings and recommendations
- Brief presentation of findings
- Identify recommendation pages for top 20 queries

3. How We Get Started

For the check-up, we will need access to the Verity environment including command line access to:

- tools and files in the Verity home directory
- collection directories
- index scripts and output logs

We will also use web browser access to

- the K2 Dashboard
- simple and advanced search forms

A. Verity Configurations and Assist Files

In addition to collection-related files, the K2 Check-Up will examine files in the Verity home directory. These change from version to version, but generally include:

- xml configuration files
- thesaurus and synonym files
- license and features
- collection properties and attributes
- app server, services and other required packages

B. Index Scripts and Log Files

The scripts or jobs that create and update collections are critical to the success of any K2 installation. While the names may change at different installations, the K2 Check-Up examines all of the index scripts and index log files to examine details including:

- proper collection creation, maintenance and management
- error-free indexing
- accurate options including include and exclude patterns
- vspider and K2spider control files
- bulk insert files

C. Collection Files

The K2 Check-Up will examine all of the files and scripts that define your collections, as well as inspecting some of the collection contents. The files and content that we access includes:

- style files
- collection pdd and parts directories
- word analysis from 'did' files within partitions
- partition record count from ddd files within partitions

D. Search Forms

Your HTML search forms and the scripts those forms access are also part of the K2 Check-Up. While the names of these forms and files change from site to site, what we are looking for includes:

- The form element of your HTML search page
- JavaScript code included in the search form
- The fields that are passed from HTML forms to the search scripts
- The set up your search scripts perform prior to search using the VSearch object
- Processing of the results document using Result and Document objects

E. Advanced Functionality

While the K2 Check-Up focuses on basic search functionality, it also verifies the operation of advanced features including:

- Parametric Search
- Knowledge Bases and Trees
- Recommendation Engine

Common Problems

This section addresses some common problems with K2 installations, the kinds of things you might hear from your users that indicate a problem exists, and the actions we take during the check-up to address the issue.

Poor result quality

Verity K2 is a powerful search engine that uses the latest in state-of-the-art technology to identify likely relevant documents based on statistical and other methods. But with users entering one and two word searches, the best technology can't always find the right answer.

On the other hand, you know how your site is structured, and what fields and metadata you use, so using the Verity Query Language, you can customize your search scripts to produce better results.

If your results are poor, you will hear:

“Can't find the right document”

“Nothing looks good on the first result page”

“The right document doesn't even show up until the third page of results”

“Search isn't working right”

Some of the files we examine in the K2 Check-Up to improve search result quality include:

ASP or JSP search scripts

build scripts and logs

style.zon

style.ddd and the files it includes

Poor or mangled summaries

When you look at search results, the summaries just don't look good. They may have snippets of JavaScript or encoded characters within the summary, and they may even contain HTML that is disrupting your results list formatting.

Controlling the type of summary that Verity extracts can help with bad summary results, but you can also simply remove unwanted characters from the summary when you display the result list, producing better looking results.

If your summaries are poor or mangled, you may hear:

“Summaries don’t reflect the real content of the page”

“The summaries have all these garbage characters in them”

“Summaries look ugly”

To address poor summaries, the K2 Check-Up includes examination of:

style.prm

The check-up may also suggest changes to your ASP or JSP search scripts.

Duplicate document in result list

Often, the same document turns up two or more times in a result list, which makes it really difficult to be confident that your indexing is working right, and confuses users with duplicate recommendations.

Sometimes you will find documents duplicated in your file system or web server, but usually the problem is one of inconsistent link usage or in how your web server creates multiple aliases for the same page. Checking why you are seeing duplicate documents can put you on the right track to correcting the situation.

If your searches return duplicate documents you will hear:

“Documents occur more than once in the result list”

“The first two documents in the results are the same document”

To address duplicate documents, the K2 Check-Up examines:

index scripts

index log files

ASP or JSP search scripts

HTML forms

Web Server behavior

Irrelevant docs in result list

Because of Verity's advanced relevance tools and automated assists like soundex and synonym expansion, sometimes documents that seems to be completely irrelevant will show up in your results. As with any result tuning, you may need to use your knowledge of the document structure, content, and metadata to improve the indexing and to tune the user query.

When searches return irrelevant documents, you may hear:

“These documents are not right for my query”

“The results don't answer my question”

To address irrelevant documents in result lists, the K2 Check-Up examines:

ASP or JSP search scripts,
style.ddd and included files
index scripts

Missing documents

Occasionally, the indexing process will miss documents – either because of bad links or because of a problem indexing the full document. If the missing page is a ‘landing page’ for a large section of your site, you might find many pages are missing.

Whenever the indexer runs, examine both the index log file and automate a ‘sanity check’ on your collection – preferably before you roll it out into production.

When you are missing documents, you will hear:

“My document isn't in the result list”

“The CEO's bio page is missing”

To address the issue of missing documents, the K2 Check-Up checks:

ASP or JSP search scripts
index scripts
build scripts
index log files

Index errors

Sometimes, problems like bad file system permissions or missing URLs can prevent the indexer from running at all. If you are using incremental indexing, you might not notice the problem for months, unless you regularly examine the index log file for errors.

Always check the index log file for obvious errors – missing documents, for example. But also add date checks, and check the collection to see if the number of documents seems right.

Index errors mean you'll be hearing feedback like:

“The index didn't work right”

“Last night's run failed”

“The search engine is showing no results!”

To address index errors, the K2 Check-Up examines:

index scripts

index log files

Slow Search

Users are impatient, and if your search takes a long time to return any results, users are unhappy. And for an average user, two or three seconds with no response is slow. You need to make sure that your search indexes are optimized and your search scripts are efficient.

When your search is unacceptably slow, you may hear:

“Searches take forever to complete”

“I never use search because it’s too slow”

To address slow search performance, the K2 Check-Up examines:

index scripts

ASP or JSP search scripts for field and zone operators

style.ddd and included files for fields and zones

Some field values are not working

Extracting fielded information – titles, authors, dates and other metadata – is an important step in indexing. And when you search, the performance you see can be based on how effectively you have designed your collections. Zone search is generally much faster than field search, but do you know which operators in your search script work on fields and which work on zones?

Bad field values mean you will hear:

“Not all documents have a title”

“The author fields are wrong”

“The dates are wrong on these documents”

“It takes forever to search for authors”

The K2 Check-Up addresses field and zone values by examining:

style.ddd and included files

style.tde

index scripts

bulk insert file values

gateway configuration and database if using DBMS Gateway

Date search isn't working right

Returning a current policy or meeting document is often one of the hardest things for a search engine to do. Dates are often wrong – sometimes web servers provide bad dates, other times the file system date is wrong – yet few people go to the trouble of improving their document dates. Sometimes you can extract a date from the URL, other times from the document itself – but your indexing and search scripts need to find the right data to deliver quality results.

When your date search is not working properly, you will hear:

“I can't get the newest documents to show up first”

“The search engine is returning old documents”

“This year's vacation schedule should show up first”

To address document dates, the K2 Check-Up examines:

style.tde

index scripts

web server configuration