International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V68I10P105

Volume 68 Issue 10, 25-28, October 2020
©2020 Seventh Sense Research Group®

Original Article

Website Search Engine Migration - Triggers, Plan, and Execution

Parameswara Rao Kandregula

IT Consultant, Cognizant Technology Solutions, Houston, USA.

Received Date: 16 September 2020
Revised Date: 19 October 2020
Accepted Date: 20 October 2020

***Abstract** - This paper provides the various aspects that should be considered while migrating search engines for websites from one product to another. It provides approaches on how to handle common aspects of a search engine and make the migration successful with the desired results.*

***Keywords** - Search Engine, Search Engine Optimization (SEO), Search Rules, Page Ranking.*

I. INTRODUCTION

Search Engines are a key part of any website, which not only helps users find content on a website but also improves the brand value of the website and conversion to profits for commercial websites. Popular website search engine products are Solr, Elasticsearch, Endeca, etc. Due to various business or technical needs like website re-architecture or cost savings, there could be a need to migrate existing search engines of a website to another product like Solr to ElasticSearch or vice versa. Detailed baselining, planning, execution, metrics comparison, and monitoring are needed for a successful migration with a positive impact.

II. SEARCH ENGINE MIGRATION

The journey of migration, like any other IT migration project, should be properly planned and measured; else, it can go off track or result in a broken search experience, which can turn away users, impact page ranking and hamper brand value also.

A. Trigger for Migration

Any migration initiative is initiated by the analysis of existing systems, which calls for the replacement of current products to help fill the gaps. But migrations are costly and time-consuming. Thus as shown in Fig. 1, all triggers to migration have to be properly analyzed to ensure that migration is the best option to take the business goals ahead and do not end with neutral or negative progress.

a) Capture Request Triggers

There can be various factors that can trigger the necessity to consider replacing the existing search on a website. It can come from constant user feedback, business feedback, or even from the IT department to align with the technology roadmap. Some common triggers are declining metrics, feature availability, cost savings, performance bottlenecks, etc. For example:

Declining search metrics: Persistent increase in search exits, search refinements, and other undesired search metrics [1] are strong indicators that the current search feature is not helping users find the right content and potentially driving them to competitor's websites for good.

1) Feature Availability: IT might be looking for features to fulfil business needs, but the availability of required features would force the team to check for alternatives to satisfy the business's long-term vision.

2) Cost: Keeping the IT run costs low within the targeted budget is a common goal. Commercial search products can be costly, and moving to cheaper or open-source products becomes motivational to decrease IT run costs.

3) Performance: If the search response time is not up to the desired threshold or unable to scale to cope with website traffic, then it becomes a deterrent for growth and imperative to replace it.

b) Analyze the Triggers

There can be various factors that can contribute to the triggers. Each trigger has to be analyzed to ensure the cause of each of them is the search feature. There could be false alarms often, and each trigger request has to be thoroughly analyzed to confirm such scenarios are avoided. Otherwise, a lot of time and money would be spent and would still end up with the same results. For Example:

Declining Search Metrics: Though search engines contribute to search metrics, there are multiple other points contributing to success or failure. One such main player is Data. Data has a huge hand in making or breaking search behaviour. Firstly, proper metadata for records should have been identified as per the requirements of the site. Secondly, it has to be ensured the records have the metadata with proper maintenance. Lastly, it has to be ensured that the metadata of records matches the search engine schema mapping and search rules. For example, if we are expecting users to be able to find products on a website based on colour. Then, 'colour' has to be identified as a property for records; all records with colour should have proper colour, and search engines should be mapped to use that property and allow search on the colour property. If the data factor is ruled out, then it has to be checked that all features of the existing search product have been leveraged or tuned and monitored.

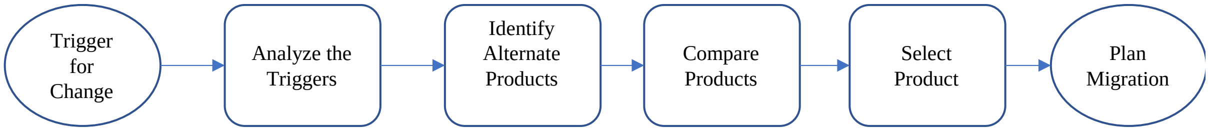

Fig 1: Analyse the Triggers and Alternatives

1)Feature Availability: Awareness of the search product capabilities is required to satisfy the business requests as needed. Alternatively, there should be responsive support from the product vendor to address product capability questions. Even if the product does not have the required features, most of the requests can be delivered with minor to major customizations. The level of customization depends on the business appetite to consume related cost, time, and also the team's skillset to deliver it effectively. But too many customizations can lead to increased dependency on IT and hinder product upgrades in the future. The wrong use of features can also result in negative results and hinder the usage of other features. For example, if the search boost is used for all searches, then the relevancy for top returned results will not always be the same. All channels of awareness of existing product features should be evaluated before concluding on the limited capability of the product.

***2)Cost*:** The cost of any IT product mainly depends on the cost of infrastructure to host it, the team to maintain it, and licensing cost. If servers are over capacity, then there could be options to reduce and save some money. Team cost depends on the skills needed to support the product and the scarcity or abundance of resources in the market. Often good search resources are scarce and need to depend on consultants, or existing team members can also be trained to cross skills. Switching to a different product will not save much on resource costs. License cost, combined with the feature set, is often a key factor to motivate to check on alternate products. It is recommended to have meetings with the existing product vendor regarding licensing and feature costs. The vendor might give a licensing discount and feature roadmap or custom feature support.

3)Performance: Search can be used not just for the search functionality but also for driving the content data on most of the pages of the website. Thus, search infrastructure gets a lot of traffic, and the logs of most page requests of websites would have search-related logs also. This can give false alarms that anyone's slow page could be due to a slow search. To confirm it, the search should be tested in isolation and also with data variations. Improper data flow, in the correct configuration, bad customizations, improper caching, under capacity can be common causes of performance degradation. The right caching strategy with proper SLA can give tremendous performance benefits. A meeting with the existing product vendor can provide recommendations on how to tune performance based on traffic and the existing composition of requests to search.If all possible analyses and remediations still do not give the desired results, it is best to check for alternative products.

c) Identify the alternatives

Once it has been decided that the existing product has to be replaced, the IT and business teams can start checking for alternate products. The market for search products is limited [2]. Thus finding the list of alternatives is easy. This can be done through a simple check on the internet to find the latest products. Alternatively, if there is an IT vendor supporting the website, their help can be used to suggest alternatives and contacts for each of the suggestions.

d) Compare the Products

All the identified products have to be compared to know each of the product's strengths and weaknesses. Also, it is very important to check that it solves the problems documented as triggers for initiating the change. Another important factor to consider is the process and timeline to seamlessly switch to a new product with minimum interruption to business. Most of the products can be contacted to conduct a demo of the product, and during the connection, the intentions of the change can be mentioned. The product company might give licensing discounts, commitment to metrics improvements, and also custom solutions to gain a new client.

e) Select the Product

Based on the data collected on each of the identified products, the IT and business teams should collectively select the product that solves the existing issues, meet and also aligns with the future goals and growth. The teams should make a strong case based on estimated migration cost, timeline to migrate, running cost, targeted metrics improvement, and present to the leadership executives to get budget approval for migration.

f) Plan the Migration Before starting the migration, and even while budget approval is pending, it is recommended to start the planning for migration. Generally, the search is core to any website, and it is connected to several interfaces. Thus, if all pieces of the integration are not documented, it would help to document the integrations and data flow. The main parts of the search are schema, data inputs, data transformation, search configuration, search rules, index jobs, SEO,

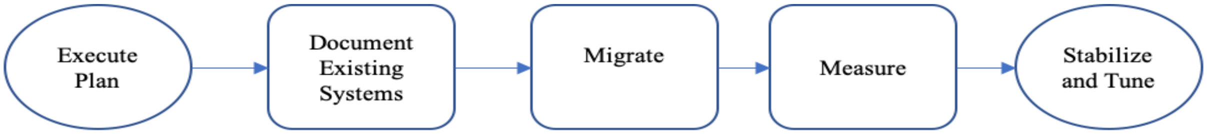

Fig 2: Execute the Migration

Consumers of the services and environments. The details on what to plan for each of them and how to execute it will be detailed in the next sections. Also, capturing and baseline search metrics for both business and performance would help to compare the impact of post-migration.

B. Execute the Migration

To execute the migration, the prerequisite of documentation of each component has to be completed. Each of the components has to be evaluated by IT and business together for either migrating as is or making changes and migrating or complete recreation in a new product or abandonment of the feature. There should be a structured approach to migration, as shown in Fig. 2.

a) Schema

Data fields of records are the crux of any search behaviour. All fields in the existing system should be documented and recreated in the schema of the new environment. Datatypes should match. Also, the field names should be kept the same to minimize changes in upstream and downstream systems. Often, schema also has field level transformation like stripping special characters, converting to ASCII, etc. These schema level transformations have to be evaluated in-depth to match both systems. Else there could be undesired search and frontend behaviour. Any changes in schema would likely need changes in interfacing systems.

b) Data Inputs

There would be multiple channels of input data to the search system. All those channels have to be reconnected to the new system. Asynchronous channels like queues or files in file share are relatively easy to manage as they can be paused or built new ones temporarily or permanently to switch to new systems. Synchronous channels like streaming logs, etc., could be challenging to avoid any loss of data during test and transition. It can be checked with the input systems if they can duplicate the data stream or at least make a backup of files to restore it in case of issues.

c) Data Transformations

Once data from multiple data sources is consumed, there are multiple levels of data transformation done before feeding it to the search engine. For example, sales data from past months for products have to be combined as a single sale value or profit weighing ratio mapping based on business needs, etc. Sometimes the transformation layer is outside the search engine as a pure ETL layer. In such cases, there would be minimal changes as ETL systems just need to be connected to the new system and fields mapped. In case the transformations are part of the existing search, then they have to be built again in the new systems or also a good opportunity to move to the ETL layer. Schema level cleaning like stripping special characters, etc., should also be done as is or implemented new to keep the data in systems clean and consistent.

d) Search Configuration

There would be a good amount of configuration in search systems that would define the response to search queries. All of those have to be documented and recreated in the new system. For example, search algorithms to apply, the fields on which search should happen, synonyms, stemming, stop words, a grouping of fields for different search types, the sequence of data filtering or transformation triggered by search, caching configurations, etc. Though most of it can be reconfigured, search algorithms cannot be. Search algorithms define how documents would be searched and ranked for a search query. Search algorithms are unique to each product, and some don't provide options to change or select one except for changing configurations like relevancy weights, etc. It would need a good amount of trial and error or rounds of testing to ensure the results match the desired settings before selecting the final relevancy algorithm and/or configuration.

e) Search Rules

Search rules are the rules that define how the behaviours of search or search results should be for specific search terms or patterns of search keywords. Boost, bury, landing pages are some of the common rules. Depending on how old the existing system has been used, these rules can be from tens to thousands. All have to be documented and recreated in the new system. If the existing format and future format are in known formats like JSON or XML, then it could be worth writing a script or program to migrate them. Else it would be a time-consuming process. Also, it is an opportunity to clean up old rules which might not be in use or actually contribute to erratic search results.

f) Index Jobs

Index jobs are the jobs that transform the data into documents that are stored and indexed in search engines. The existing job scripts, schedule, and timings have to be documented prior to starting the migration step. The script has to be rewritten as the different products will have different commands or api or scripts to trigger indexing. The schedule of new jobs has to be kept the same as existing ones to have minimal impact on dependent business flow. All new jobs have to be tested for the timing to ensure they are faster or at least match the existing job timings. If the jobs are taking time, then changing the storage type or creating indexes on tables or sharding can help to alleviate the timings.

g) SEO and Features

SEO helps public search engines like google, yahoo, and bing, find websites and improve the page ranking of websites. Most search products provide inbuilt SEO features. If there are any public URLs coming to the search engine directly or website pages are using the SEO feature of the product, then the URLs have to be captured, and redirect rules have to be put in place to redirect to URLs of the new product. In addition to SEO, search products also provide features like content management, product recommendations, etc. All those have to be recreated in the new product, and if the new product does not support them, then an alternate standalone product for that feature has to be found, or the feature has to be abandoned.

h) The consumer of Services

The website is not the only potential consumer of search services. Other applications, like mobile apps, etc., can also be using it. Thus, the dependent website and applications have to be updated to change their endpoints to use the new service. If there are changes in the contract like data format or changes in features, the corresponding has to be updated and thoroughly tested.

i) Environments

Search infrastructure normally involves multiple servers for each environment based on the topology of master-slave, shard distributed index, and varies from product to product. All dependent upstream and downstream systems have to coordinate the timing of changes, testing, and go-live plan. When moving to a new product, creating new environments in parallel to existing ones can be beneficial in multiple ways. Firstly, it helps to compare behaviour and performance. It helps to go live in phased manners with a canary deployment approach [3]. Also, it can help to fall back in case the migration is not going as planned or the results are not as desired. Based on the intended overlap, the licensing and infrastructure of the old product should be maintained.

C. Measure the Migration

Like any other migration, each phase of the migration should be measured to stay on track and reap the benefits of post-migration. Baselines captured during the planning phase should be used extensively to measure and compare the new product. Business metrics, performance metrics, page rankings, and any other indirect indicators should be constantly monitored. From an SEO perspective, there can be errors in the public search engine console. They should be constantly monitored, fixed, and cross verified to check the error count is coming down. The search servers' logs and application logs should be scanned for any new exceptions and acted upon to fix them. The cause of migration should be revisited to validate the problem is solved or show a trend to alleviate the problem with eventual elimination. Once the system has stabilized and the metrics are showing a healthy trend, the old systems can be decommissioned. But search engines should continue to be in a constant cycle of tuning, implementing, monitoring, and tuning.

III. CONCLUSION

Search engines have many moving parts and dependent systems. Existing search engine products should be completely explored to solve existing issues. If all options are exhausted, then the option for migration should be considered. With proper documentation of existing systems, planning, and constant measurement, there can be smooth and successful migration.

REFERENCES

[1] Google support analytics answer 10322321. [Online]. Available: https://support.google.com/analytics/answer/1032321

[2] Apache Solr market share page. [Online]. Available: https://www.datanyze.com/market-share/enterprise-search--287/apache-solr-market-share

[3] Colourful deployments page. [Online]. Available: https://opensource.com/article/17/5/colorful-deployments