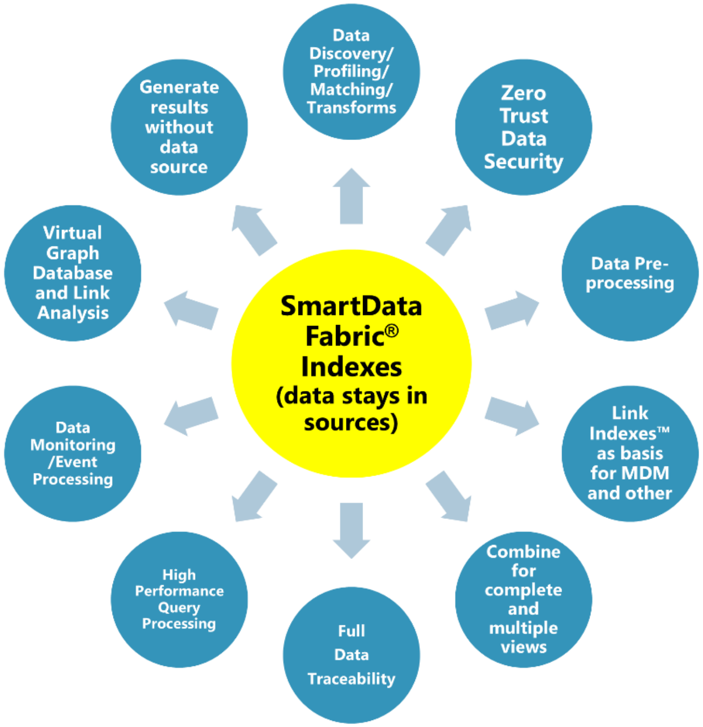
SMARTDATA FABRIC® VS. LEADING SEARCH ENGINES

REVISION 1.4

Introduction

There are a number of search engines on the market, both open source and commercial, some of which offer a limited free version. This document compares WhamTech SmartData Fabric®, an index-based security-centric distributed virtual data, master data and graph data management, and analytics software "platform", for want of a better term, with the top four search engines.

SmartData Fabric® leverages indexes for almost all capabilities, including data discovery, profiling, classification, security, cleansing, transformation, standardization, metadata and governance, high-performance query processing, link/relationship mapping, master data management, and virtual graph database and link analysis. Indexes are not just for high performance queries, but provide insight into and value from data and the relationships among data. See diagram below:


SmartData Fabric® Indexes (data stays in sources)
Data Discovery/ Profiling/ Matching/ Transforms
Zero Trust Data Security
Data Pre-processing
Link Indexes™ as basis for MDM and other
Combine for complete and multiple views
Full Data Traceability
High Performance Query Processing
Data Monitoring /Event Processing
Virtual Graph Database and Link Analysis
Generate results without data source

1. Use raw indexes for **DATA DISCOVERY** (metadata), build and maintain **DATA PROFILING, DATA MATCHING** within and across data sources, and **DEVELOPING AND TESTING DATA TRANSFORMS**
2. Support **FORRESTER ZERO TRUST DATA SECURITY FRAMEWORK** – discover, INDEX, classify and secure – CCPA/CCPR, GDPR, PCI, PHI, PII, etc.
3. **PRE-PROCESS DATA** while building and maintaining production indexes to address data management fundamentals, e.g., cleansing, transformation, standardization and security – data is usually discarded
4. Use **LINK INDEXES™ AS BASIS FOR MDM AND OTHER CAPABILITIES** – future development to use indexes exclusively for MDM match and merge
5. Provide **COMPLETE AND MULTIPLE VIEWS OF DATA** through queries on combined content, link and master data indexes
6. Provide **FULL DATA TRACEABILITY** as indexes and results contain unique pointers to data in sources – data lineage, governance and audit
7. Enable **HIGH PERFORMANCE, DISTRIBUTED PARALLEL QUERY PROCESSSING** through standard drivers, APIs, Web/data services, SQL and other query languages
8. **MONITOR DATA SOURCES** for content and relationships in near real-time, and support **EVENT PROCESSING**
9. Enable **VIRTUAL GRAPH DATABASE**, link analysis and graph/link visualization
10. **GENERATE RESULTS WITHOUT DATA SOURCE** when source is unavailable, for query optimization, or as storage, e.g., for IoT devices, as indexes are columnar and can be inverted and combined

Indexes are associated with data sources and federated adapters, and federation servers are associated with federated adapters and other federation servers. These distributed layers form the basis of SmartData Fabric®.

The website DB-Engines (www.db-engines.com) is an ad-based neutral provider of information about databases of all types and was used to provide most of the comparisons through the link: https://db-engines.com/en/system/Elasticsearch%3BMarkLogic%3BSolr%3BSplunk.

Summary

There are fourteen important benefits of SmartData Fabric® over the top four search engines:

1. **DOES NOT MAKE ANOTHER COPY OF DATA – LEAVES DATA WHERE IT RESIDES**
   This could be on-premise, in a data center, in a Cloud/multiple Clouds or third-parties such as Salesforce.com.
2. **DOES NOT CONVERT DATA TO A FILE/DOCUMENT FORMAT – LEAVES DATA IN ITS SOURCE FORMAT**
   This allows for regulatory compliance and avoids resource intensive schema transformation in particular. Examples include:
   - Indexing and querying original or copies of data sources.
   - Using copies of mainframe data files (MDFs) as data sources, which are usually large hierarchical files, and viewed as if they are modern RDBMSs in a standardized data format.
   - Exporting and storing CSV files from an RDBMS and working with them as though an active RDBMS, not flat files or documents.
3. **DISTRIBUTED AND INDEPENDENTLY CONFIGURABLE ON MULTIPLE PLATFORMS IN MULTIPLE LOCATIONS**
   No centralization, centralized sharding or centralized management.
4. **VIRTUAL DATA MANAGEMENT**
   Data discovery, classification, security, quality (cleansing, transformation and standardization), link/relationship mapping and governance, graph database, link analysis and interactive graph/link visualization.
5. **VIRTUAL/HYBRID MASTER DATA MANAGEMENT**
   Seamless, automatic and can be maintained in near real-time. Essential to good reporting, BI, analytics, CRM, GDPR, BPM workflows and almost any application. Provides data warehouse type views. Examples are single patient views/longitudinal patient records, single customer views and other master data-centric views.
6. **VIRTUAL RDF/TRIPLE STORE**
   Not physically stored, but using SQL and a combination of content, link and master data indexes, can query multiple types of data sources simultaneously as though a physically stored RDF/triple store. MarkLogic has a separate physical RDF/triple store.

7. **SQL SUPPORT**
   Extensive ANSI SQL support and basic PL/SQL. MarkLogic also has extensive SQL support.

8. **JOIN SUPPORT**
   Extensive join support within the indexes for a data source at the adapter level and across multiple data sources at the federation server level. MarkLogic has extensive join support within its data storage.

9. **EVENT PROCESSING SUPPORT**
   Can support internal and external/BPM workflows for both indexes and indexed views. Also, partnered with Oracle® to support for Oracle® Event Processing (OEP).

10. **INDEX UPDATES**
    Twelve different ways to update indexes. MarkLogic can also update indexes. Other search engines can typically add to, but not update, existing data and associated indexes.

11. **FOREIGN KEYS, I.E. REFERENTIAL INTEGRITY**
    As most search engines convert data to flat files/documents (for the most part), they do not retain or support relational or referential schemas and, therefore, foreign keys. WhamTech believes that converting data from canonical source formats, particularly, relational or referential, removes information. Retained foreign keys provide direct connections/relationships that can be lost when converting data to flat file/document formats. Primary key and foreign keys can be found through content indexes and the relationships retained in WhamTech Link Indexes™.

12. **SELF-JOINS**
    These retain links to similar data within the same tables and across multiple tables in an RDBMS or file system that are not captured by foreign keys. This is important for deduplication, master data management and other processes. Self-joins can be found through content indexes and the relationships retained in WhamTech Link Indexes™.

13. **EXTERNAL KEYS**
    These retain links to similar data across data sources, typically, at the entity level. Multiple algorithms can be used as match criteria for external keys. Self-joins can be found through content indexes and the relationships retained in WhamTech Link Indexes™.

14. **WRITE BACK TO/UPDATE DATA SOURCES**
    In many situations, there is a need to write back to/update data sources, e.g., a phone app allows a patient to review and select an appointment with a doctor or lab based on proximity and urgency, an update to a customer's email address has to be propagated back to operational systems, and true interoperability of a healthcare application interacting with one or more data sources. MarkLogic has ACID transaction processing to data copied and stored to its storage.

Comparison Table

The following table listing approximately follows the order listed in the above-referenced DB-Engines link:

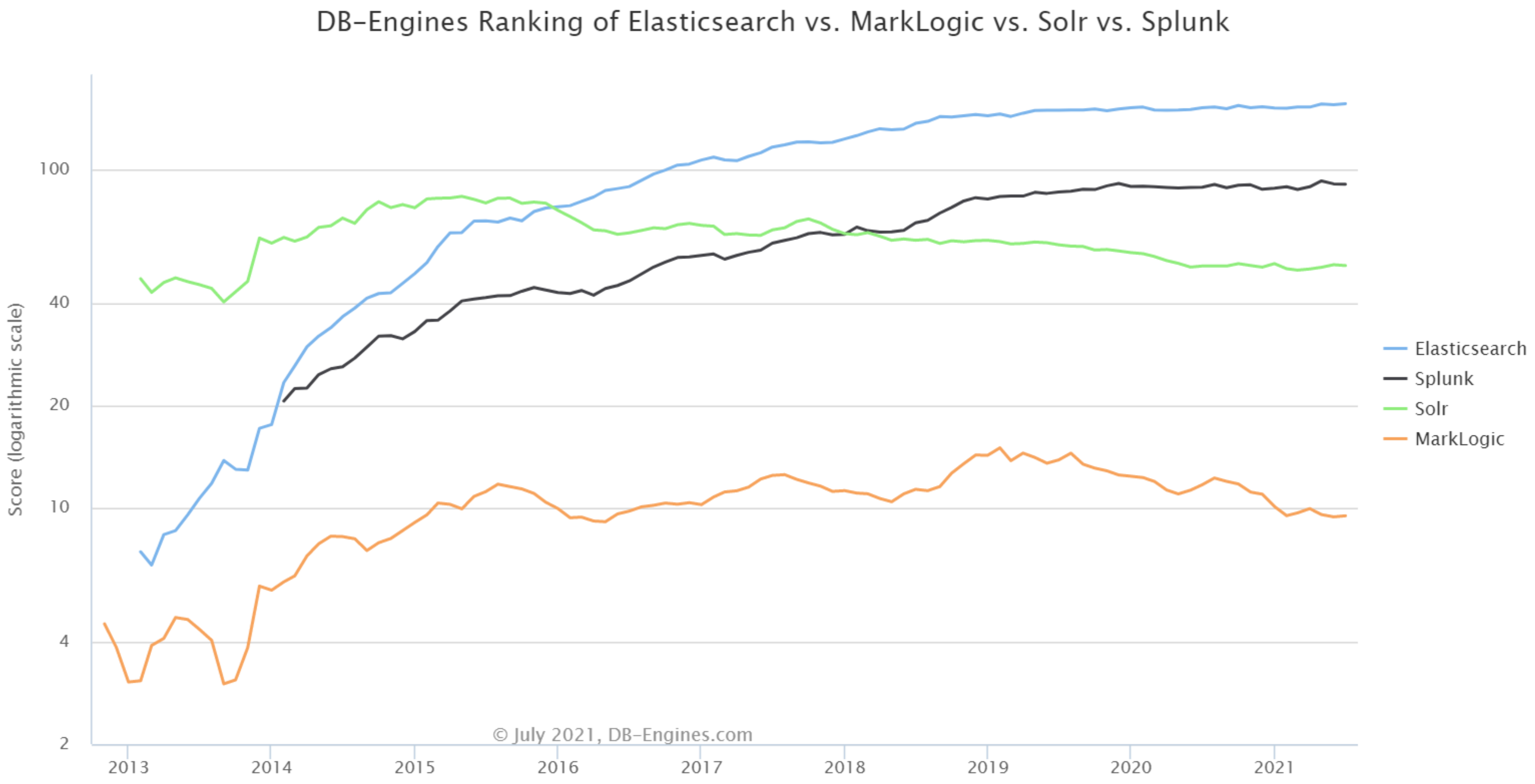
Feature – Yellow represents a SmartData Fabric benefit over the other vendors	**SmartData Fabric®**	**Elasticsearch**	**MarkLogic**	**Solr**	**Splunk**
Description	Index-based distributed data, graph data and master data management, analytics and security	A distributed, RESTful modern search and analytics engine based on Apache Lucene	Operational and transactional enterprise NoSQL database	A widely used enterprise search engine based on Apache Lucene	Analytics platform for Big Data
Primary database model	Virtual standard data view based on industry data models and search engine	Search engine	Document store Native XML DBMS RDF store Search engine	Search engine	Search engine
DB-Engines Ranking Trend Chart:	Score: N/A Rank: N/A	Score 155.76 Rank: #8 Overall #1 Search engines	Score 9.45 Rank: #48 Overall #8 Documents #1 Native XML DBMS #1 RDF stores #4 Search engines	Score 51.79 Rank: #20 Overall #3 Search engines	Score 90.05 Rank: #13 Overall #2 Search engines
Website	www.whamtech.com	www.elastic.co/-products/elasticsearch	www.marklogic.com	lucene.apache.org/solr	www.splunk.com
Technical documentation	http://www.whamtech.com/eiq_product_suite_help/	www.elastic.co/guide/-en/elasticsearch/-reference/current/-index.html	docs.marklogic.com	lucene.apache.org/solr/-resources.html	docs.splunk.com/-Documentation/-Splunk
Developer	WhamTech, Inc.	Elastic	MarkLogic Corp.	Apache Software Foundation	Splunk Inc.
Initial release	2003, predecessor product	2010	2001	2006	2003
Current release	7.8.0, June 2020	9.0, 2017	8.9.0, June 2021	7.8.0, June 2020	

Feature – Yellow represents a SmartData Fabric benefit over the other vendors	**SmartData Fabric®**	**Elasticsearch**	**MarkLogic**	**Solr**	**Splunk**
License	Commercial for production Unlimited evaluation and development free	Open Source	Commercial Limited version free	Open Source	Commercial Limited and development version free
Stores data	**No, usually leaves data in source, but option to store**	**Yes**	**Yes**	**Yes**	**Yes**
Modifies stored data	**No**	**Yes**	**Yes, except XML**	**Yes**	**Yes**
Centralized or distributed	**Centralization an option, but usually distributed**	**Centrally managed sharding**	**Centrally managed sharding**	**Centrally managed sharding**	**Centrally managed sharding**
Data management (discovery, profiling, classification, security, cleansing, transformation, standardization, metadata and governance)	**Yes**	**No**	**No**	**No**	**No, generally no need, as works with machine-generated data**
Master Data Management	**Yes, seamless and automatic**	**No**	**No**	**No**	**No, generally no need, as works with machine-generated data**
Cloud-based only	No	No	No	No	No
DBaaS offerings (sponsored)		Elasticsearch Service on Elastic Cloud: Try out the official hosted Elasticsearch and Kibana offering available on Amazon Web Services, Google Cloud and Microsoft Azure that's powered by the creators of Elasticsearch.			
Implementation language	C/C++	Java	C++	Java	
Server operating systems	Linux Windows	All OS with a Java VM	Linux OS X Windows	All OS with a Java VM and a servlet container	Linux OS X Solaris Windows
Data schema	Indexes same as sources = no schema transformation Option, indexes mapped to a standard data view	Schema-free	Schema-free	Yes	Yes

Feature – Yellow represents a SmartData Fabric benefit over the other vendors	**SmartData Fabric®**	**Elasticsearch**	**MarkLogic**	**Solr**	**Splunk**
Multiple data types	Yes, and can also transform	Yes	Yes	Yes	Yes
XML support	Yes	No	Yes	Yes	Yes
Triple store	Yes, virtual	No	Yes, physical RDFs	No	No
Indexes	Yes, multiple and indexed views, extensive	Yes	Yes	Yes	Yes
SQL	Yes, extensive	SQL-like query language	Yes	Solr Parallel SQL Interface	No, proprietary
Join support	Yes, extensive	Limited	Yes	Limited	No
Event processing support	Yes	No	No	No	No
Index updates	Yes, twelve ways	No	Yes	No	No
APIs and other access methods	Java API JDBC Native API ODBC RESTful API	Java API RESTful HTTP/JSON API	Java API Node.js Client API ODBC Proprietary Optic API RESTful HTTP API SPARQL WebDAV XDBC XQuery XSLT	Java API RESTful HTTP/ JSON API	HTTP REST
Supported programming languages	C C++ C# Java JavaScript Perl -needs testing PHP Python Ruby -needs testing	.Net Groovy Community Contributed Clients Java JavaScript Perl PHP Python Ruby	C C# C++ Java JavaScript (Node.js) Perl PHP Python Ruby	.Net Erlang Java JavaScript any language that supports sockets and either XML or JSON Perl PHP Python Ruby Scala	C# Java JavaScript PHP Python Ruby
Server-side scripts	Yes, also can invoke in data sources	Yes	Yes	Java plugins	Yes
Triggers	Yes	Yes, through "percolate"	Yes	Yes	Yes

Feature – Yellow represents a SmartData Fabric benefit over the other vendors	SmartData Fabric®	Elasticsearch	MarkLogic	Solr	Splunk
Partitioning methods	Each data source has own index-based adapters, plus, can segment based on sharding	Sharding, centrally managed	Sharding, centrally managed	Sharding, centrally managed	Sharding, centrally managed
Replication methods	Yes, multi-source	Yes	Yes	Yes	Yes, multi-source
MapReduce	Not per-se, but offers parallel distributed query processing	ES – Hadoop Connector	Yes	No, but alternate spark-solr: github.com/-lucidworks /-spark-solr and streaming expressions to reduce	Yes
Consistency concepts	Eventual consistency	Eventual consistency	Immediate consistency	Eventual consistency	Eventual consistency
Foreign keys	Yes	No	No	No	No
External keys	Yes	No	No	No	No
Self-joins	Yes	No	No	No	No
Transaction concepts	No	No	ACID	Optimistic locking	No
Concurrency	Yes	Yes	Yes	Yes	Yes
Durability	Yes	Yes	Yes	Yes	Yes
In-memory capabilities	Yes	Through Memcached and Redis integration	Yes, with range indexes	Yes	No
Access control	Role-based Access Control (RBAC), and support for Active Directory (AD), Identity Authentication Management (IAM), Attribute-based Access Control/Row-level Security (ABAC/RLS), Column-level Security (CLS) and Single Sign-on (SSO) and Multi-classification-based.	RBAC with Shield add-on	RBAC at the document and subdocument levels	RBAC with add-ons	RBAC with enterprise version
Write back to/update data sources	Yes	No	No, can update stored data	No	No

The trend chart for the leading four search engines is available through the link in the comparison table above, as follows:



For sales and other opportunities, please contact: Mark Armstrong, President, mark.armstrong@whamtech.com, (972) 991-5700 x708.

For technical matters, please contact: Gavin Robertson, CTO and Sr. VP, gavin.robertson@whamtech.com, (972) 991-5700 x706.

Information on WhamTech solutions, sales and services, and partnership and investment opportunities can be obtained through whamtech.com.