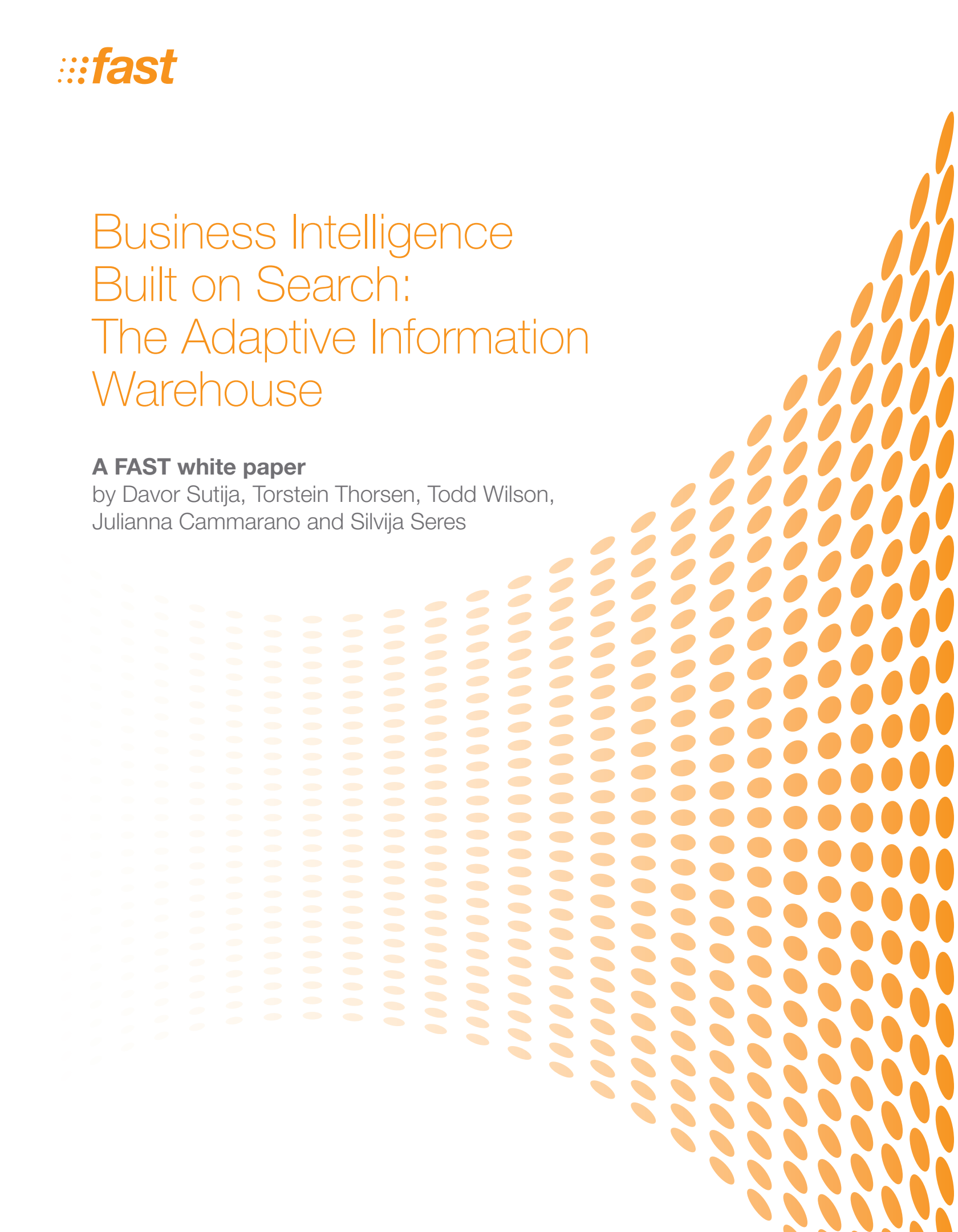
fast

Business Intelligence Built on Search: The Adaptive Information Warehouse

A FAST white paper

by Davor Sutija, Torstein Thorsen, Todd Wilson, Julianna Cammarano and Silvija Seres



The Adaptive Information Warehouse is a single platform, offering Data Integration, Linguistic Data Cleansing, Alerting, and Ad-hoc Query and Report generation. This vision is based on a business intelligence solution built on modern enterprise search. In this article we explore the benefits of this approach and explain how it is enabled by search.

Introduction

The convergence of Search with BI revolves not only about the incorporation of text and news into decision-making, but also the ability to interact with data directly without a business analyst intermediary. Business owners want information to flow naturally and in two directions: they want to receive important alerts including analytics, as well as being able to demand answers to ad-hoc queries through a portal browser without mastering the intricacies of being a 'power user' of legacy BI tools.

The new BI built on search makes possible extreme performance in ad-hoc analysis of historical data at arbitrary levels of granularity. This type of analysis is the fundamental core of what BI was meant to deliver, whether the COPA module within SAP or Sales Analysis and Forecasting in standard BI platforms.

The advantage of Search is that queries can range from the analysis and display of sets of individual transactions to global yearly totals for products based on any of hundreds of attributes in any n-dimensional combination: Contribution to Margin, Product Code, Channel, Territory, Revenue, Discount, Customer, Market, etc. It is no longer necessary to build a global Data Warehouse to offload dozens data sources, with the ensuing complexity and recourse to aggregations to limit data explosion.

Modern search installations measure data volumes in G, units of 120TB equivalent to the estimated size of the world-wide web, rather than in TB. With no barrier to the amount of data that can be linearly indexed, and made retrievable with 100% sub-second latency, search has already become the technology of choice in information archival solutions requiring extreme data storage volumes.

Now, by combining this scaling ability with the charting, analysis and data enrichment found in the Adaptive Information Warehouse, a new set of BI offerings are possible within a single platform: Data Integration, Linguistic Data Cleansing, Business Activity Monitoring and Alerting, and Ad-hoc Query and Report generation.

Thus, search is not only the portal, but also the vehicle by which data from previously separate applications and repositories is unified, mined, and made accessible for monitoring, alerting and reporting, in short, Business Intelligence. An end-to-end solution that redefines how the enterprise accesses and interacts with Intelligence, creating a more efficient and effective workforce where business users make better business decisions based on enriched, cleansed, relevant and actionable data.

The Search Advantage

Modern enterprise search platforms are built to combine the ability to extract value from the scale and information and user inconsistency of the internet and the information richness and strategic importance in enterprises. The following abilities

'Search is The Portal' is the new mantra to financial statistics aggregators such as Reuters and the Financial Times, who have replaced DB-centric solutions with Search for their top-of-the-line trading platforms and premium content on-line offerings. Business publishers, whether Hoovers in the US or the exceptionally innovative Schibsted in Scandinavia, are pushing the frontiers of how to integrate data from multiple (20-100+) separate databases with radically different data models into a semantically uniform representation where unique objects such as Companies, Individuals, Products are created with the unioned and cleansed information from data residing in these previously separate silos.

differentiate such search platform-based BI solutions from other offerings:

Performance, performance, performance: The use of complex ad-hoc queries traditionally necessitated the creation of separate data marts to shield the data warehouse from loads that can impact performance. AIW changes this by consistently delivering sub-second query response to a broad range of SQL queries that represent typical use in a BI environment. This is true irregardless of data model, and without prior optimization of either queries or schema.

In general AIW provides linear scalability on load-balanced distributed architecture optimized for low-cost commodity hardware. Users and enterprises demand these systems to handle terabytes of data, hundreds of data updates per second, thousands of queries per second, while maintaining sub-second query response, all at the same time. And they demand high speed analysis across all of these parameters – so ad-hoc query performance is a unique differentiator and makes possible all the following search-enabled differentiators.

Object Association and Dynamic Profiling: Extracting and analysing classes of variables in each data source, such as names, addresses and transactional metadata, ranging to hundreds of attributes typically stored in CRM system and data marts, allows the creation of new associated objects for the purposes of dynamic relationship discovery. Profiled dimensions are discovered on the fly for each result set, and regardless of their cardinality their component elements are numerically displayed to the user. These allow the user to properly create complex queries through intuitive contextual navigation.

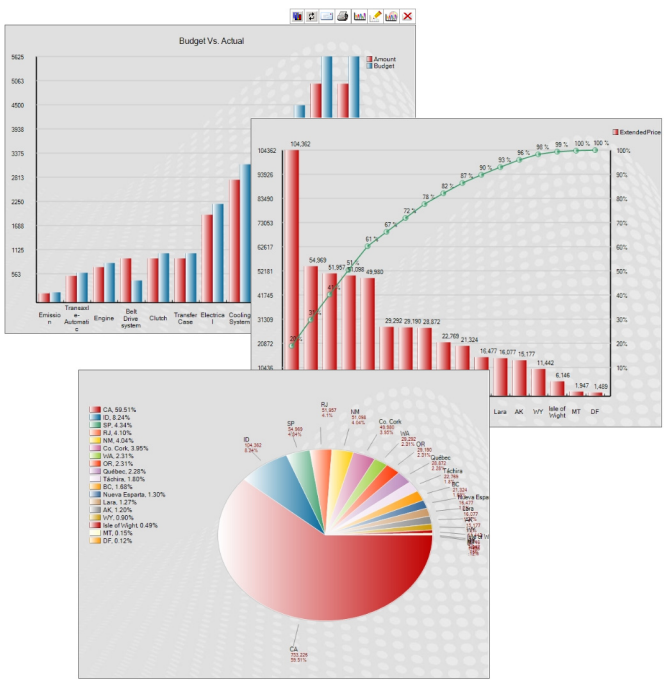
This provides BI developers with a short time to information distribution, because it removes the need to tag and classify the data. The time that traditional BI systems need to build new information cubes is one of the major pain points for information freshness and accessibility. Easy access and a complete view of supporting data provides business users with the tools to unearth new levels of associated information that support better decisions as well as the means to detect trends and relationships that were previously never know or explored.

Linguistic Data Cleansing: Leveraging the combination of ETL tools and modern enterprise search technology including text processing, fuzzy grouping and similarity measurement, structured data from repositories can be merged to create a clean master index in a matter of day or weeks rather than months. For instance, getting a single point of view of a vendor within an enterprise is difficult, since subsidiaries of the vendor in different countries are probably represented by different company codes in the ERP system. With fuzzy matching, similarities on vendor name can be utilized to get an aggregated view. But advanced linguistic analysis goes far beyond precision keyword matching and includes Levenshtein edit distance to ignore minor errors, phonetic search to match on vocal similarity, character-set normalization (e' and 'e = e), scope search for better accuracy, support for proper spelling dictionaries and support for synonyms to cover know variants. Superior levels of cleansed data throughout the enterprise increases return of investment by providing results in records time for expedited project deliver, enables cleansed results to be written back to transactional databases for complete synchronization throughout the enterprise, as well as providing a de-duplicated cleansed master for BI projects – a solution that ensures data governance is addressed at each level.

Combining structured and unstructured information: AIW leverages the advancement of XML technologies, a de facto standard document structuring framework that allows authors to define their own sets of meta data. Retrieval systems dealing with a large number of sources need the same flexibility—they must be "schema independent." Contextual search engines provide this independence by replacing predefined index layouts with a nested structure that has scopes and tags. With this, new types of precise queries can be asked that combine structure and content, imposing contextual constraints on the content. Scalability and consistency of the system are the very foundation for heavy data crunching behind

the scenes, seamlessly filtering and improving structured, unstructured and rich media content, queries and results, with no performance penalty for the users. Once again, this power depends on simultaneous scalability in several dimensions: data volume, query traffic, data and query complexity, fault tolerance, real time capabilities, etc. This enriched high-performance level of data access provides each user with information as it changes or as it's developing from countless sources, providing a complete 360 degree view. Users are now equipped with the most up-to-date information providing a competitive edge that is unsurpassed by traditional static approaches to business intelligence.

Convergence Becomes Reality



Information access in traditional BI systems is inflexible, costly and slow. Metadata maintenance is complex, expensive and imprecise; new information services are major undertakings; the system architecture is scattered and expensive to maintain; coordination of work is nontrivial. There is duplication of work; there is duplication of content; and still, important information gets lost. The information is often untimely.

Imagine this BI system in a graphical representation of cost and effort related to each element in its value chain: you get a bottom heavy pyramid. At the bottom of the pyramid are data sources, which in a typical enterprise include corporate, regional and local documents, in several languages and many formats. To manage this information and its associated applications, the enterprise has large teams of database architects, programmers and managers, plenty of software licenses and numerous high-end servers. At the top of the chain are the information consumers: employees, analysts, management and customers looking for product information.

If we introduce an information architecture based on a modern search platform, our bottom-heavy pyramid changes dramatically. A good search platform turns it on its head, and thereby improves information access, reduces total cost of ownership and increases overall enterprise performance.

There are several major examples in production where a modern search based solution takes the complexity out of Business Intelligence and puts intelligence into the hands of every decision maker. It allows business owners or any entitled user, with no specific BI tool training, to interact with data. They can download reports in any format, with conditional formatting for easy identification. Second, they can consequently make better decisions with this personally relevant information. Third, this "personalized portal" provides them with actionable information at a glance, where they can drill-down for detailed exploration, and dynamically change the nature of the reports. Finally, this system reduces IT costs and efforts, through a simple and intuitive interface for rapid adoption, and greatly reduced IT support requirements and number of tools needed to support users. Also manual report delivery is replaced with automated frameworks.

Such a search-based BI solution enables a proactive business model, provides actionable information at a glance and increases ROI with a simple and intuitive approach.

Data Quality

Integrating and merging data across multiple silos is a costly but necessary business for most large organizations. There are many reasons: re-organizations, mergers, compliancy requirements, re-use of information, cutting operating costs, or simply a need for a richer picture of customers or situations. As systems grow and their intended use changes over time, there is a regular need to join partially overlapping data sources, such as multiple customer registers without a common key.

Such unified picture of data across a number of silos requires significant data normalization with cleansing and matching capabilities. No matter how disciplined and organized the content providers are, data in different silos often evolves in different directions. Over time, the business-critical data becomes increasingly ambiguous: sources from different parts of the organization vary in formats, models, spelling, or local conventions. Data supplied by customers is even less clean, with broad variances caused by lack of knowledge about the accepted formats, expected keys or simply proper spelling. A traditional manual "normalization" effort takes a long time, has high associated costs, and is itself ridden with the same imperfection problems it is trying to solve. The alternative is to turn to modern search engines for rescue, with approximate matching capabilities and weighted multi-field comparison.

FAST has multiple live implementations in this area. Examples include several major telcos with phone-book databases that need cleansing of 'yellow pages' and 'white pages', and multinational banks with databases that need cleansing of customer information after mergers, or black-listing and cleansing of undesirable or dangerous customers.

Search-based multi-field comparisons are used together with matching logic. This automatic approach removes the need for costly, slow and often incorrect manual corrections that are used in traditional solutions. The ambiguities that are automatically resolved occur in two separate problem domains:

Parsing of imprecise input – data to be inserted in the system, that is either incorrect and should be deleted, or is mis-spelled and should be corrected.

Mini case study

A Scandinavian media house Schibsted shows how search is enabling a revolution both within the Front Office of information consumption, as well as the Back Office of data integration, cleansing and refinement. They were the crowned winners of the 2005 Data Warehouse of the Year in Norway for the process of building a Scandinavian directory listing of businesses and individuals by combining twenty-two publicly-available registries of telephone listings, tax and corporate registry information, web mining of corporate sites and soon, even updated shareholder information.

It is now possible to find all the officers of any given corporate entity in Norway, find their role in other registered companies, as well as their shareholdings and mortgaged property (release pending regulatory approval). Rather than allowing a small group of credit scoring agencies to determine worthiness, one can now build a nexus of information from publicly-available documents to determine net worth and personal financial involvement in specific companies.

The original Schibsted solution was built from scratch in twelve weeks, and even at launch, scored higher in precision than all other on-line existing information services, even from vendors with 20+ years of data banks – information often deep in history, but difficult to maintain, update, and keep relevant.

Matching of ambiguous results – consistent results when faced with choices among alternative answers

Together, the solutions to these two problems enable rich reuse of imperfect information – a pragmatic, practical and cast-effective alternative to the utopian vision of perfect data, enforced through perfect discipline and precision among all data producers, data classifiers, and data integrators. The search-enabled solution works, and provides

rich and correct information in a world where such imperfect and overlapping data needs to put in context, improved by technology, and revealed in all its strategic power.

Toolbox for Data Cleansing

The search-based data cleansing solution has been constructed as a Service Oriented Architecture, with the following groups of capabilities across multiple languages:

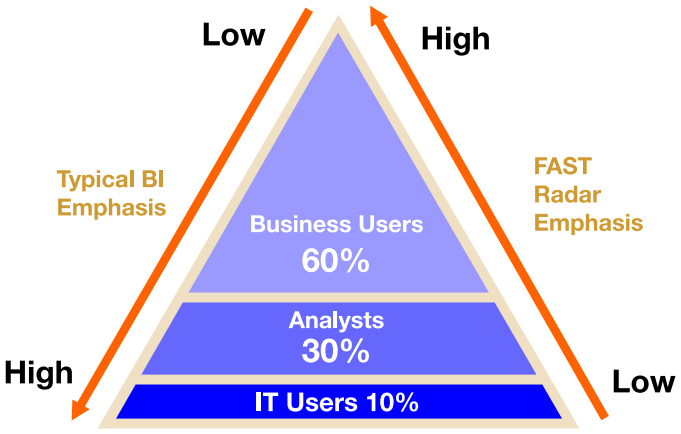
1. Structured data search, where data is extracted from a database in a model-sensitive way, so that the fields are properly tagged and can be used in a biased way, depending on the matching logics. This basically entails contextual awareness for the search engine, where localities are differentiated from street names and postal addresses are differentiated from physical addresses. Addresses are typically hierarchical, for example a part of the address may be <primary> 525 Collin Street <sub> Level 27, South Tower <sub> <primary>, and it is useful be able to weight the main primary part more heavily, or to be able to search for a South Tower on the Collin Street. In effect, the advanced search engine is able to model the data structure in its memory, rather than normalize it too early and thus lose valuable contextual info.

2. Dictionaries of names, phrases, or other key concepts are used to detect spelling errors and variations. Known variants of names are used by the search engine to ensure matching in case a name has many alternative spellings. An advanced search engine will take this into account when performing matching. In addition to spelling variations, typing errors might also occur in the original documents and in the user queries. These are handled by the spelling correction mechanism to ensure correct matching.

3. Character normalization is used to solve the challenge caused by diacritical characters (á, à, ä, å, etc.). Diacritical characters are often typed wrong or omitted. By performing character normalization, all diacritical characters are mapped to one normalized form on both query and content side, ensuring that one can obtain exact matches despite any differences in the use of diacritics.

4. Fuzzy and phonetic matching is used at times where character normalization and dictionary approaches do not give us 100% match. Fuzzy matching applies functions such as spell checking, lemmatization, stemming and synonyms to ensure that the search engine recognizes words and terminology that are considers alike. On the other hand, phonetic matching will detect words and names that sound similar although spelled differently (e.g. “Cheap” and “Sheep”).

5. Data Corrections, where the power of advanced linguistics of free-text search is used to define and apply rules that fix addresses and names. The linguistic tools in use include spellchecking, phonetics, bi-grams, n-grams and wild-carding, proximity boosting and forward or-boosting. In short, these tools first identify ambiguous overlapping data, they subsequently introduce a controlled



level of linguistic chaos (by breaking the words up in parts, for example, or dropping letters) in this data, finally, they use search engine’s fuzzy matching and flexible relevancy ranking to tidy up and find the correct normalized version of the data.

To ensure secure identification, these matching mechanisms should be employed across multiple fields of the merging systems. This is used to help decide if two similar looking records are the same person/company, for example: let us say that the last names are identical and the first names close

(Pam vs. Pamela), then matching the address-field might help determine if the two records are in fact the same person/company.

Other tools that have proved central in this data improvement exercise involve easy management of synonym dictionaries through web-based tools for business-managers, and reporting tools to detail the types of address errors that are not being detected and corrected by the system. On the query side, automated functionality to try alternative queries is being used. It has proved useful that the system allows a 'query completion' feature, where possible results are proposed or "auto-completed" rather than the operator having to type the entire word. Use of this feature is desirable as it improves the accuracy and decreases the time required to enter an address by an operator, however each key press generates a query to the underlying engine, thus increasing query rates and scalability requirements. Given the potentially large number of operators with such auto-complete capability, it is important that the system scales linearly and inexpensively.

Intelligence Based on Aggregate and Real-time Data

The power and speed of search-based intelligence tools can be best leveraged by use of a ROLAP Pyramid process. It provides aggregation and multi-dimensional analysis, and thus leads to increased user adoption.

A ROLAP Pyramid process provides a more efficient method of aggregating required data and allows users to predefine the navigation process. This definition is then used to optimize the aggregation process and group only the dimensions necessary at each level of the Pyramid. It is this patented technology that provides a more cost effective solution to end-user and IT organizations alike. For establishments that are already fully invested in multi-dimensional data stores or data warehouse environments, AIW and ROLAP provide extensions that will fully leverage these existing technology investments.

What this leaves us with is a fully integrated browser-based Business Intelligence framework that delivers aggregate and real-time data via drill down reporting, dash boarding, alerting and structured reporting from multiple data sources. It is a solution designed to take the complexity out of BI and put intelligence into the hands of each and every knowledge worker.

Ease-of-use is also experienced through familiar user interfaces. A BI tool built on search delivers personally relevant, actionable information to popular interfaces such as a browser, Microsoft Excel, Word, E-mail, Adobe .PDF, a PDA or a cell phone. It accommodates the needs of each user by presenting information in a manner in which they are most comfortable. One user may want a set of tabular reports while another wants to see data viewed in a chart or graph.

Conclusions

Studies have shown that Business Intelligence initiatives have over a 60% failure rate. This rate of failure is significant not only in terms of dollars spent on procurement of solutions but also on the time and effort spent to design and develop data warehouses and reporting structures, execute end-user training, and maintain a very expensive system. Failure rates increase as decision makers abandon initiatives when hours are spent sorting through vast amounts of data only to discover it's out-of-date and not relevant to their needs. As a cost-effective alternative, a BI platform based on search provides users with an integrated architecture for increased ROI and Rapid User Adoption.

All business information consumers need information that is relevant to their operational unit, information that provides aggregate data as well as up-to-date operational data. Search-based BI eliminates the hours spent each day sorting though thousands of static records – and it puts aggregate and real-time information at decision makers' finger tips and allows them to quickly create ad-hoc exception reports, as well as interactively investigate trends that deliver specific information that requires attention.

BI built on search can address the needs of organizations across all markets, from financial institutions dealing with rates, customers and branch perform-

ance analysis to government agencies dealing with military operation readiness. It maximizes the availability and usability of information without the complexity, time, and systems overhead associated with many traditional BI applications, helping companies across the world to align intelligence initiatives with business objectives for a proactive business strategy.

About FAST

FAST is the leading developer of enterprise search technologies and solutions that are behind the scenes at the world's best known companies with the most demanding search problems. FAST's solutions are installed in more than 3500 locations.

FAST is headquartered in Oslo, Norway and Needham, Massachusetts and is publicly traded under the ticker symbol 'FAST' on the Oslo Stock Exchange. The FAST Group operates globally with presence in Europe, North America, the Asia/Pacific region, South America, the Middle East and Africa. For further information about FAST, please visit www.fastsearch.com.

For any feedback or questions related to this paper, please contact us at feedback@fastsearch.com.

FAST™
www.fastsearch.com
info@fastsearch.com

Regional Headquarters

The Americas
+1 781 304 2400

Europe, Middle East & Africa (EMEA)
+47 23 01 12 00

Japan
+81 3 5511 4343

Asia Pacific
+612 9929 7725

FST1000277-01