Vývoj moderních technologií při vyhledávání

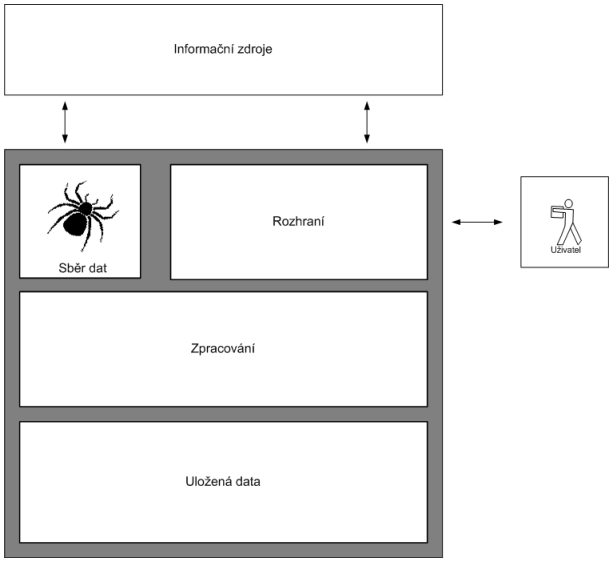
Patrik Plachý
SEFIRA spol. s.r.o.
plachy@sefira.cz

INFORUM 2007: 13. konference o profesionálních informačních zdrojích
Praha, 22. - 24.5. 2007

Abstrakt
Vzhledem k existenci řešení pro zajištění vysoce dostupných databází a robustních webových aplikací vzniká prostor pro vývoj dalších nadstavbových řešení, jako například inteligentní metody pro prohledávání. Zajímavé je sledovat vývoj ve vyhledávacích technologiích z rodiny Oracle, hlavně tedy Oracle Database a Oracle Aplikační Server a doplňující nadstavby, jako je Oracle Text a Secured Enterprise Search. Zkušenost s fulltextovými systémy získala společnost SEFIRA tím, že vytvořila nadstavbu nad produktem Oracle Text, která řeší velmi pokročile problematiku našich jazyků při fulltextovém vyhledávání a má již mnoho úspěšných instalací a implementací. Společnost SEFIRA se mnoho let zabývá fulltextovými systémy, implementací fulltextových nástrojů od společnosti Oracle, jejich vylepšováním a tvorbou nadstaveb. V článku se zabývám produkty Oracle Secured Enterprise Search a Oracle Text, a to z pohledu použití při vyhledávání.

1 Úvod

V dnešní době má mnoho organizací potřebu vytvářet ucelený zdroj informací a umožnit svým pracovníkům rychle vyhledávat relevantní informace z rozmanitých datových zdrojů a zároveň chránit citlivé informace. Architektura takového zdroje informací vychází z klasické představy úložiště dat, vhodného indexování a v neposlední řadě vyhledávacího nástroje. Dále se objevuje tlak na bezpečné uložení informací, inteligentní výsledky vyhledávání a řeší se, jak sbírat zdrojové informace.



Obrázek 1: Schéma systému poskytující informace

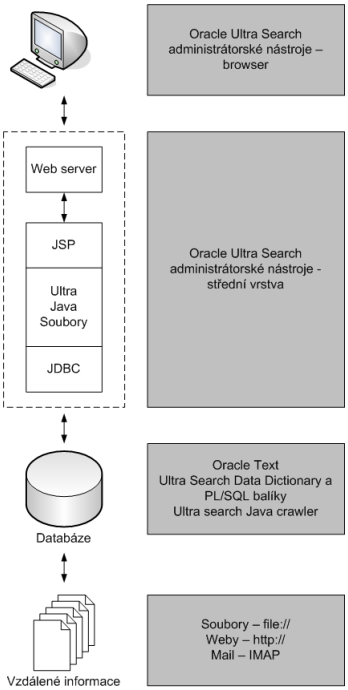
Do hry vstupuje kromě výkonu a způsobu práce s takovýmto zdrojem informací, také zpracování nestrukturovaných dat, fulltextové vyhledávání a přizpůsobení informací tak, aby zobrazovala uživatelům pouze ty výsledky vyhledávání, k nimž jsou oprávněni přistupovat.

2 Secure Search

Organizace často řeší problematiku komplexního a bezpečného prohledávání veškerého intelektuálního kapitálu, který má uložen v aplikacích, na souborovém systému a dalších vlastních zdrojích,např. i na webu. Způsob, jak dostat informace do úložiště má několik následujících možností.

- Aplikace, kterou vytvoříme na míru pro prostředí organizace.
- Využijeme nástroje úložiště, které např. může číst data z externích datových zdrojů.
- Použijeme produkt, který umí napojit různé zdroje, respektuje bezpečnostní politiky a také je pružný při konfiguraci na vnitřní infrastruktuře organizace.

Vytváření aplikace s sebou nese kromě míry chybovosti také komplikovaný způsob vývoje a nutnosti udržování znalostí o celém prostředí. Co se týká možnosti využití doplňujícího nástroje úložiště, disponuje Oracle databáze komponentou Ultra Search, která umožňuje prohledávat webové stránky, další databáze, mail servery a HTML. Architektura nástroje Ultra Search je postavena tak, že část řešení je uložena v databázi a část samostatně v aplikačním kontejneru.



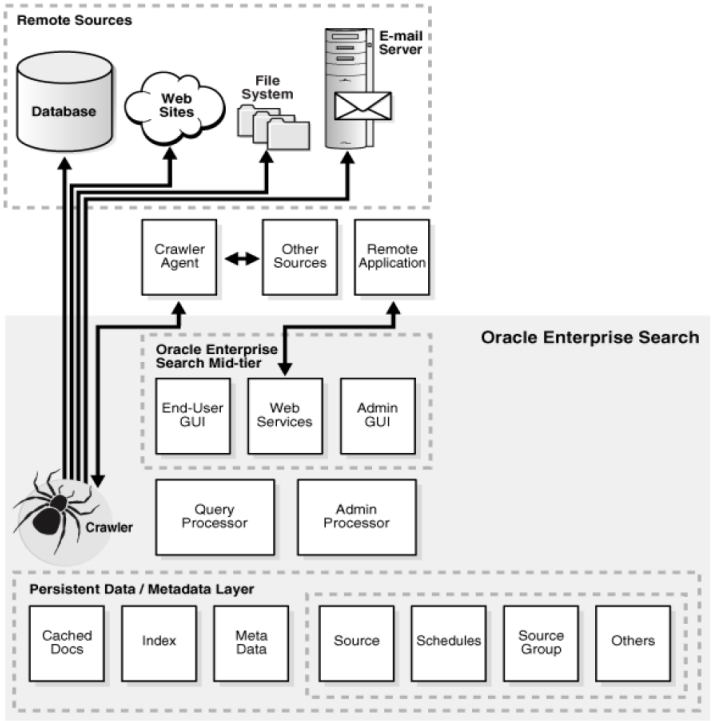
Obrázek 2: Ultra Search

Společnost Oracle přichází v současnosti s dalším produktem Secure Enterprise Search 10g, jenž poskytuje uživatelům rychlý přístup k podnikovým informacím a zároveň zajišťuje dodržování podnikových zásad zabezpečení. Secure Enterprise Search 10g indexuje a vyhledává veřejný, soukromý a sdílený obsah na interních i externích webových serverech, v databázích, na souborových serverech, v úložištích dokumentů, v systémech pro správu podnikového obsahu, v aplikacích a na portálech. Jeho uživatelsky přístupná webová rozhraní vracejí vysoce relevantní výsledky s rychlou dobou odezvy. Secure Enterprise Search 10g se

přímo integruje s různými systémy ověřování uživatelů, uchovává index ve zvlášť odolném úložišti a pro přizpůsobení možností vyhledávání používá bezpečné programovací rozhraní.

Obrázek 3: Použití Secure Search v organizaci

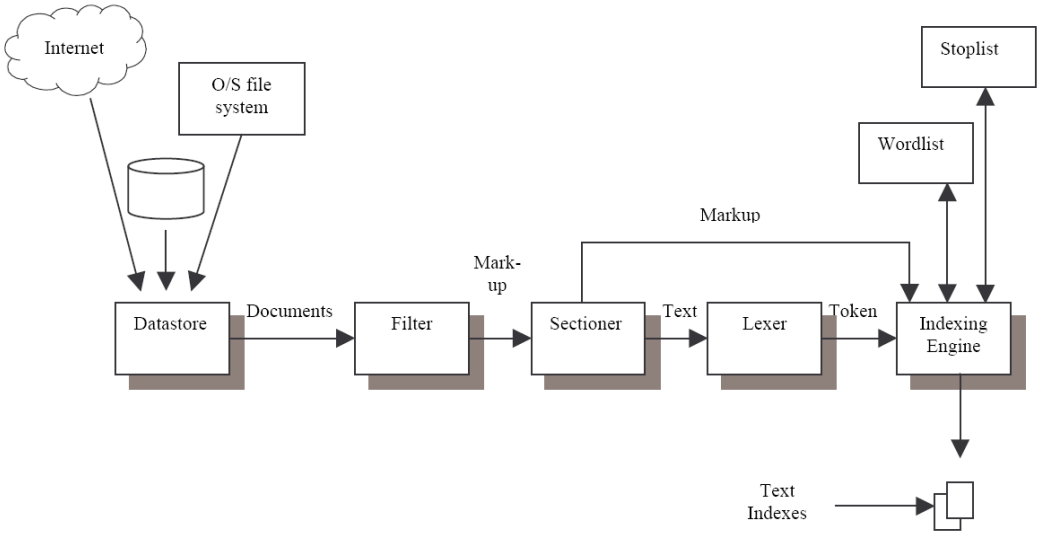
Secure Search disponuje vyhledávací technologií, jež poskytuje výsledky přizpůsobené roli uživatele v organizaci. Téměř každá organizace udržuje svoji znalostní základnu, mohou to být technické dokumenty, smlouvy, webové portály, ale informace mohou být uloženy i v databázích. Bezpečný přístup k různým úložištím, formátům souborů, datům a obsahu je nezbytnou součástí. Stejně tak je důležité jednoduché ovládání, jediný bod přístupu a integrace s dalšími podnikovými aplikacemi. Organizace, které intenzivně pracují s informacemi, budou s rostoucím významem vyhledávání podnikových informací požadovat od vyhledávačů stejnou úroveň škálovatelnosti, spolehlivosti a globální podpory, jakou nyní očekávají od zbytku své infrastruktury IT.



Obrázek 4: Architektura Secure Search

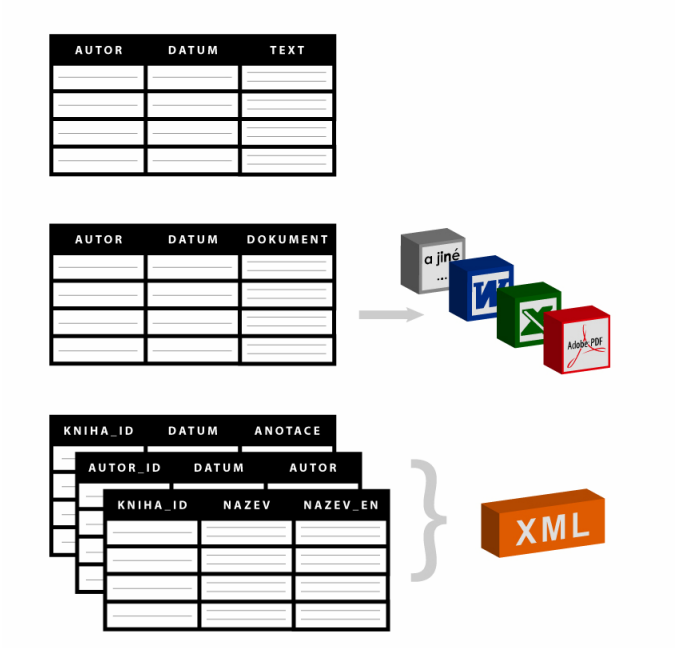
3 Oracle Text

Společnost SEFIRA používá pro fulltextové systémy od svého počátku komponentu Oracle Text, která je součástí Oracle databáze. Ta funguje tak, že využívá standardní SQL dotazy pro práci s nestrukturovaným textem. Vyhledávací nástroje postavené na Oracle technologii, využívají pro analýzu textu komponentu Oracle Text, která může vykonávat lingvistickou analýzu na dokumentech, právě tak používání různé strategie včetně klíčových slov, kontextové dotazy, booleovské operace a porovnávací operátory, apod. Princip architektury Oracle Text je založen na vytváření fulltextových indexů, které se vytvářejí pomocí speciálních filtrů pro daný formát dat.



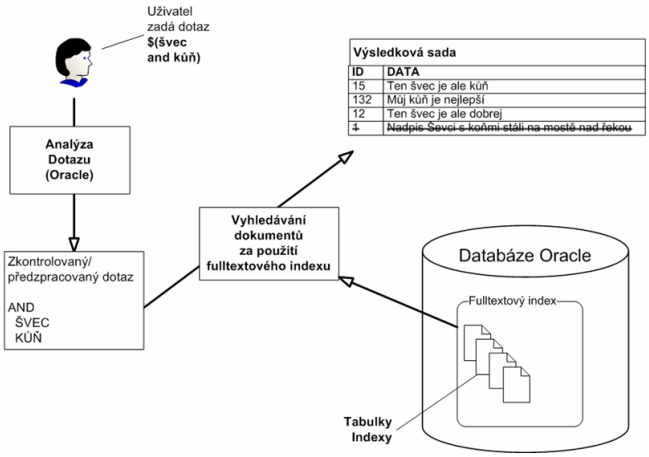
Obrázek 5: Architektura Oracle Text

Komponenta Oracle Text disponuje připravenou řadou filtrů pro nejběžnější formáty, tedy nejběžněji Microsoft Office, Adobe PDF, ale i HTML a XML. Pokud je to nutné lze využít volání standardních „autorecognize“ filtrů.



Obrázek 6: Způsoby uložení textů pro indexaci

Výhodou při používání Oracle Text je uložení všech dat v jednom datovém úložišti, jak strukturovaných i nestrukturovaných. Je zde uložen i index, dále úložiště poskytuje API pro vývoj aplikací a SQL pro práci. Jsou podporovány národní znakové sady.



Obrázek 7: Vyhledávání pomocí Oracle Text

Velkou výhodou Oracle Text je přizpůsobení uživatelským potřebám, ať už přizpůsobením indexů, CONTEXT, CTXCAT, CTXRULE, nebo také vkládáním uživatelských funkcí, tzv. klasifikace. Jedná se o silné mechanismy, kdy můžeme označovat důležité pasáže v nestrukturovaném dokumentu, např. témata, tituly, vybírat části dokumentu a zpracovávat je námi požadovaným způsobem. V poslední verzi 10g je kladen důraz na výkon, kde Oracle Text převyšuje výkonem a možnostmi jiné specializované nástroje pro vyhledávání.

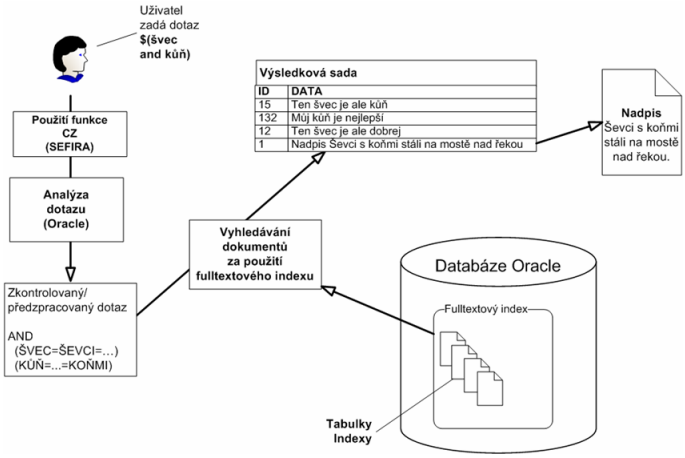
5 ConText

ConText je produkt společnosti SEFIRA a poskytuje řešení pro integrovanou správu nestrukturovaných informací. Umožňuje třídit textové informační zdroje (dokumenty) stejně efektivně jako strukturovaná data. ConText výrazně rozšiřuje schopnosti databáze Oracle podporovat všechny typy uživatelů a všechny typy dat. Umožňuje zpracovávat informace rychleji a efektivněji.

Co tedy ConText ve spolupráci s Oracle Text nabízí:

- vyhledání českých slov v dokumentech i strukturovaných datech
- vyhledávání českých slov ve všech časech, resp. pádech
- vyhledávání podle stejného slovního základu nebo kmene
- vyhledávání s funkcí STEM
    - *příklad*: po zadání vyhledání slova ***jíst*** bude výsledková sada obsahovat dokumenty obsahující toto slovo i ve tvarech ***jedla***, ***jez***, ***jíme*** atd.
    - *příklad*: po zadání vyhledání slova ***kůň*** bude výsledková sada obsahovat dokumenty obsahující toto slovo i ve tvarech ***koně***, ***koňmi*** atd.
- vyhledávání s různými operátory
    - AND *slovo1* AND *slovo2:* dokument obsahuje obě slova
    - OR *slovo1* OR *slovo2:* dokument obsahuje alespoň jedno ze slov
    - NEAR *slovo1* NEAR *slovo2:* hledaná slova jsou v dokumentu blízko sebe

- NOT *slovo1* NOT *slovo2:* dokument neobsahuje *slovo2*



Obrázek 8: Vyhledávání pomocí Oracle Text s ConText